# Benefits and risks of emphasis adaptation in study workflows

Nava Tintarev[1], Matt Green[1], Judith Masthoff[1], and Frouke Hermens[2]

[1] Department of Computing Science, University of Aberdeen,
[2] School of Psychology, University of Lincoln,
n.tintarev@abdn.ac.uk, matt@mjglab.org, j.masthoff@abdn.ac.uk,
frouke.hermens@gmail.com

**Abstract.** This paper looks at the effect of highlighting in a study plan, represented as a workflow with prerequisites. We compare the effectiveness of highlighting when the adaptation was correct (participants responded quicker and more correctly), and when it did not highlight the most relevant tasks (detrimental effect). False statements took longer to process than positive statements (deciding about things that were not in the plan), but also surprisingly had lower error rates than positive statements. These findings imply that when the system makes errors in the adaptation this is harmful, and may cause students to incorrectly believe that they do not need to do certain tasks.

**Key words:** Visualization · Plan presentation · Study workflows · User-centered evaluation · Highlighting · Emphasis adaptation

## 1 Introduction

In adaptive learning systems, methods such as link annotation and hiding have been used to help learners navigate learning materials [1]. One of the challenges has been to consider pre-requisites for learning modules, guiding students and supporting them in identifying which materials they should study next. One such approach is the traffic light metaphor ([2, 3]) which indicates differences between recommended reading and material the student is not yet ready for.

The approaches used in such systems (e.g., ISIS-tutor [4], ELM-ART [2], KnowledgeSea [5]) are often non-sequential (e.g., they jump between subjects) and for this reason may not give users an overview of, and an understanding of the pre-requisites, in the study plan. The visual information seeking mantra states: "Overview first, zoom and filter, then details-on-demand." [6]. Supplying an overview may help students to plan their study, and such overviews have been found to improve the efficiency of hypertext [7–9].

For this reason, this paper investigates the presentation of study plans. A study plan can be seen as a workflow with each step representing a study task, and the edges between these tasks representing the transition that occurs once each task is complete. At times several tasks, or prerequisites, must be completed before proceeding to the next step. The path through the workflow can be personalized for each student, and adapted as their goals change.

Previous work on visualizing plans has looked at filtering graphs by content [10], and applying fish-eye views to grow or shrink parts of a graph [11]. There is also research on verbalizing and explaining plans generated by A.I. planning systems [12, 13].

This paper studies the use of emphasis of relevant paths through a workflow as a means to improve the effectiveness of information presentation. This personalized path emphasizes all of the relevant tasks, including all prerequisites.

## 2 Experiment

In previous (unpublished) studies we found no significant difference in cognitive load (measured in a dual-task paradigm) between adaptations that included highlighting and those that did not. It is possible that the type of adaptation of plans is simply not effective. The current experiment investigates if an emphasis of dependent tasks, using border highlighting, affects participant performance. Since an adaptive system may sometimes adapt to an incorrect inferred goal, we also investigate the effect of such 'unhelpful' highlighting as well, in relation to correct adaptation in 'helpful' highlighting.

We investigate a) whether highlighting had an effect on errors and response times; and b) if so, whether performance was improved by the mere presence of highlighting or if there was a difference when highlighting was for a different path through the plan than for the current learning goal (unhelpful highlighting). In the current experiment we compare the performance (response time and accuracy) for plans with no highlighting, with helpful and unhelpful highlighting.

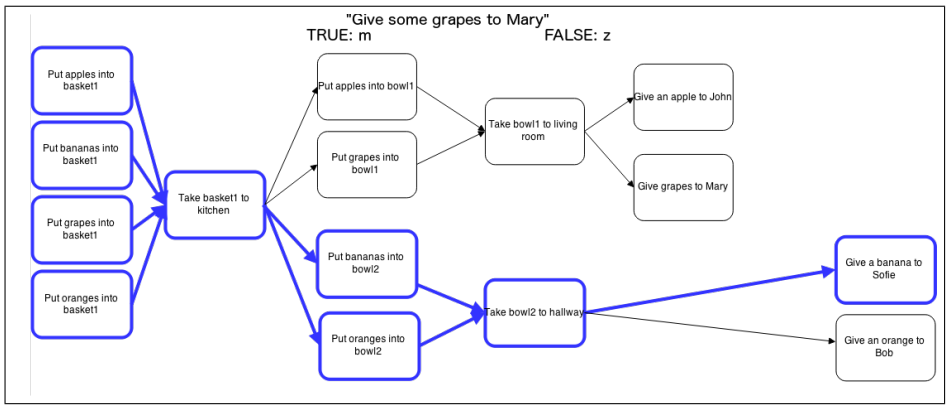


Fig. 1: Material from one experimental trial: plan and statement. The highlighting is *unhelpful* for a statement about grapes, while the highlighting is for bananas. The statement ("Give some grapes to Mary") is *true* since the step with grapes nevertheless is present in the plan.

### 2.1 Experimental design

The experiment employs a full within-participants design, with all of the participants seeing all of the variants, in randomized order.

The independent variables are: i) **htype** - whether the components of the plan that are highlighted constitute no highlighting, helpful highlighting, or unhelpful highlighting; and ii) **true value** - whether the statement (e.g., "You should study course x" or "Give some grapes to Mary") is true or false in relation to the plan. The dependent variables are: a) **Response time** - the time taken to respond to the statement about the plan; and b) **Errors** - the proportion of incorrect responses.

In the introduction screen participants were given the following instructions: *"On each screen you will be shown a plan and statement about the plan. For now, press any key to start a short practice session. This experiment studies different ways of presenting sequences of actions, or plans. You will be asked to press [true_key] if the statement is true and [false_key] if the statement is false."*

In each trial participants saw a statement and a plan (see Figure 1), and pressed a key to respond whether the statement was true or false for that plan. The keys for true/false were randomly assigned to either 'm' or 'z'. After each statement, participants were given quick feedback as a red or green dot with feedback text (either "correct" or "incorrect") before going on to the next trial.

Participants first completed a practice session (6 trials) before going on to the experimental trials (144). In addition to the independent variables we also included 6 different categories of items (farm, groceries, sports, stationery, furniture (filler), tableware (filler)), with 4 items in each (e.g., apple, grape, banana and orange). This gave a total of 144 trials: 6 categories * 4 items * 3 types of highlighting * 2 truth values. A break was inserted half way through to avoid participant fatigue.

### 2.2 Materials

**Plans.** The experiment uses an algorithm introduced and implemented in [14] that selects which steps to highlight, including prerequisite, or intermediate tasks that are required to reach an outcome. Given a study concept, the algorithm first selects all tasks that are related to a learning outcome. The algorithm then finds all paths between each pair of the selected tasks. All tasks on these paths are then added into the list of selected tasks. Lastly, the algorithm inspects all the selected tasks and checks if any of them require completion of other tasks.

While the system supports filtering by multiple items (e.g., apple, and banana) or object types (e.g., fruit), in this experiment it is applied to filtering by one object at a time (e.g., apple). The algorithm selects all the steps an item is directly involved in, as well as any prerequisite steps that may be required to achieve the final learning goal.

The plans were all of the same shape as Figure 1, and thus balanced in terms of width and number of steps, with only the names of the tasks replaced.

The categories used in the experimental trials were: farm, groceries, sports, stationery, furniture (filler), tableware (filler). For each trial and plan four objects were described, for example in the fruit category plans the following items were described: apple, pear, grapes, and banana. The range of domains was selected to minimize the effects of prior knowledge, and to ensure the generalizability of results.

**Statements.** The statements used in the experiment had four properties: category (e.g., fruit), item (e.g., apple), and the type of highlighting they were associated with (e.g., helpful, unhelpful, no highlighting) a truth value for the statement (i.e., whether or not the statement is true according to the plan). Figure 1 gives an example of a statement for the fruit category. The plan is highlighted for bananas, but the statement is about grapes, so this is unhelpful highlighting. The statement and its truth value are *true*; this is in the plan, but not for the current learning goal.

### 2.3 Hypotheses

H1: Helpful highlighting stimuli lead to faster response times than the no highlighting and unhelpful highlighting conditions.
H2: Helpful highlighting stimuli lead to fewer errors than the no highlighting and unhelpful highlighting conditions.
H3: True statements will lead to faster response times than false statements.
H4: True statements will lead to fewer errors than the false statements.

### 2.4 Results

The statistical analyses reported below were carried out in the mixed effects regression framework using the R package *lme4* [15]. This method is well suited for studying repeated measures (several trials per participant), it also allows us to model individual variations between subjects as might be expected by variation in working visual memory [16]. [17] and [18] describe the analysis method and its relationship to ANOVA. Items in the filler categories were excluded from analysis.

*Participants.* Participants were thirty-seven psychology undergraduate students, participating in a psychology experiment as part of their coursework. Data from two participants were removed because their average response times or error rates were more than 3 SDs away from the mean across participants.

**H1: Helpful highlighting stimuli lead to faster response times than the no highlighting and unhelpful highlighting conditions.** Table 1 summarizes the results, means are calculated by participant and response times were log normalized. The trend is for helpful highlighting to result in quicker response

| htype | times | times.sd | errors | errors.sd |
|---|---|---|---|---|
| unhelpful | 8.00 | 0.29 | 0.08 | 0.10 |
| no | 8.02 | 0.27 | 0.05 | 0.07 |
| helpful | 7.86 | 0.33 | 0.05 | 0.08 |

Table 1: Response times in log(ms), and error rates by subject average.

times than both unhelpful and no highlighting, as predicted by H1. Three models were built for complete two-way comparisons: helpful-unhelpful (Table 2), no-helpful (Table 3), no-unhelpful (Table 4) highlighting. There is a significant difference between helpful highlighting and the other two conditions ($p <= 0.01$), but no significant difference between unhelpful and no highlighting[3]. *H1 is supported - helpful highlighting decreases response times.*

| | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.08 | 0.05 | 169.81 | 0.00 |
| htype | -0.14 | 0.04 | -3.28 | 0.01 |
| true value | -0.17 | 0.03 | -4.84 | 0.00 |
| htype*true value | -0.01 | 0.05 | -0.27 | 0.79 |

Table 2: Model for response times in log(ms) comparing unhelpful and helpful highlighting.

| | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.08 | 0.05 | 190.68 | 0.00 |
| htype | -0.13 | 0.04 | -3.16 | 0.01 |
| true value | -0.12 | 0.03 | -3.55 | 0.00 |
| htype*true value | -0.06 | 0.05 | -1.28 | 0.20 |

Table 3: Model for response times in log(ms) comparing no and helpful highlighting.

**H2: Helpful highlighting stimuli lead to fewer errors than the no highlighting and unhelpful highlighting conditions.** Table 1 also summarizes the mean error rates. Overall, the error rates are very low, with only 5-8% errors on average. There are most errors in the unhelpful condition. Three models were built for complete two-way comparisons: helpful-unhelpful (Table 6), no-helpful highlighting (Table 7), no-unhelpful (Table 8). There is a significant difference

[3] Significance levels given using R package lmerTest, `http://cran.r-project.org/web/packages/lmerTest/index.html`, retrieved April 2015

| | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.08 | 0.05 | 189.15 | 0.00 |
| htype | 0.00 | 0.06 | 0.03 | 0.98 |
| true value | -0.12 | 0.03 | -3.51 | 0.00 |
| htype*true value | -0.05 | 0.05 | -0.98 | 0.33 |

Table 4: Model for response times in log(ms) comparing no and unhelpful highlighting.

between the helpful highlighting and the other two conditions ($p <= 0.01$), but not between the no and unhelpful highlighting conditions. *H2 is supported, relevant highlighting leads to fewer errors.*

**H3: True statements will lead to faster response times than false statements.** Table 5 summarizes the response times for true and false statements, with faster responses for true trials compared to false ones. In Tables 2, 3, and 4 we also see a significant difference for each type of highlighting ($p << 0.01$). *H3 is supported: response times are reliably faster for true statements compared to false statements.*

| true value | times | times.sd | errors | errors.sd |
|---|---|---|---|---|
| false | 8.04 | 0.31 | 0.05 | 0.08 |
| true | 7.88 | 0.27 | 0.07 | 0.09 |

Table 5: Response times as log(ms) and error rates by true value.

**H4: True statements will lead to fewer errors than the false statements.** Table 5 summarizes the error rates for true and false statements, with *more* errors for true statements. Tables 6, 7, and 8 show that this difference is significant at $p << 0.01$ for all types of highlighting. Further, we found a significant interaction between type of highlighting and truth value in the comparison between unhelpful and no highlighting ($p < 0.01$). *H4 is not supported: statements that are true led to **more** errors compared to false statements.*

| | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.08 | 0.05 | 169.81 | 0.00 |
| htype | -0.14 | 0.04 | -3.28 | 0.01 |
| true value | -0.17 | 0.03 | -4.84 | 0.00 |
| htype*true value | -0.01 | 0.05 | -0.27 | 0.79 |

Table 6: Model for errors comparing unhelpful and helpful highlighting.

| | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.08 | 0.05 | 190.68 | 0.00 |
| htype | -0.13 | 0.04 | -3.16 | 0.01 |
| true value | -0.12 | 0.03 | -3.55 | 0.00 |
| htype*true value | -0.06 | 0.05 | -1.28 | 0.20 |

Table 7: Model for errors comparing no and helpful highlighting.

| | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | 1.94 | 0.02 | 121.22 | 0.00 |
| htype | 0.00 | 0.01 | 0.29 | 0.77 |
| true value | -0.05 | 0.01 | -3.46 | 0.00 |
| htype*true value | 0.05 | 0.02 | 2.53 | 0.01 |

Table 8: Model for errors comparing no and unhelpful highlighting

### 2.5 Discussion

As predicted we found the unhelpful highlighting increased errors and response times compared to helpful highlighting (or to even no highlighting at all). However, contrary to expectations (H4), we found that statements that are true led to *more errors* compared to false statements even if these evaluations were quicker. This suggests that participants "learn" to rely on the highlighting and anticipate the relevant parts of the plan to be highlighted, when in fact this is only true some of the time. This is further corroborated by a significant interaction between type of highlighting and truth value in the comparison between unhelpful and no highlighting. That is, participants made most errors when the statement was true, but the highlighting of the plan was unhelpful. If participants learned to rely on the highlighting this could also explain the longer response times for false statements, as participants may first look for confirmation in the highlighted parts of the plan before performing a more thorough search.

## 3 Conclusion and future work

Border highlighting of prerequisite steps is an automatic adaptation in the system we are currently designing. The study described in this paper identified this adaptation as helpful, and confirmed the importance of getting the adaptation right: incorrect highlighting decreased effectiveness. We also found that creating a reliance on highlighting could have particularly adverse effects when learners are trying to answer statements that are true, but the highlighting is incorrect. These findings imply that when the system makes errors in the adaptation this is harmful, and may cause students to incorrectly believe that they do not need to do certain tasks.

The next step in this research is to compare *hiding* with *highlighting*, and investigate if individual differences in visual working memory affect which of

the adaptations is more effective. We also plan to study the value of highlighting adaptation in other visual representations of educational content such as graphs.

## References


1. Brusilovsky, P.: Adaptive navigation support: From adaptive hypermedia to the adaptive web and beyond. PsychNology Journal **2** (2004) 7–23
2. Brusilovsky, P., Schwarz, E., Weber, G.: ELM-ART: An intelligent tutoring system on world wide web. In: Intelligent Tutoring Systems. (1996)
3. Weber, G., Kuhl, H.C., Weibelzahl, S.: Developing Adaptive Internet Based Courses with the Authoring System NetCoach. In: Hypermedia: Openness, Structural Awareness, and Adaptivity. Springer Berlin Heidelberg (2002) 226–238
4. Brusilovsky, P., Pesin, L.: Adaptive navigation support in educational hypermedia: An evaluation of the isis-tutor. Journal of computing and Information Technology **6(1)** (1998) 27–38
5. Jae-Kyung, K., Farzan, R., Brusilovsky., P.: Social navigation and annotation for electronic books. In: Research advances in large digital book repositories. (2008) pp. 25–28
6. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: Symposium on Visual Languages. (1996) 336–343
7. Chen, C., Rada, R.: Interacting with hypertext: A meta-analysis of experimental studies. Human-Computer Interaction **11** (1996) 125–156
8. McDonald, S., Stevenson, R.J.: Disorientation in hypertext: the effects of three text structures on navigation performance. Applied Ergonomics **27** (1996) 61–68
9. Monk, A.F., Walsh, P., Dix, A.J.: A Comparison of Hypertext, Scrolling and Folding as Mechanisms for Program Browsing. In: People and Computers IV. Cambridge University Press (1988) 421–435
10. Henry, T.R.: Interactive graph layout: The exploration of large graphs. PhD thesis, The University of Arizona (1992)
11. Sarkar, M., Brown, M.H.: Graphical fisheye views of graphs. In: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM (1992) 83–91
12. Bercher, P., Biundo, S., Geier, T., Hoernie, T., Ritcher, F., Schattenberg, B., Nothdurft, F.: Plan, repair, execute, explain - how planning helps to assemble your home theatre. In: AAAI. (2014)
13. Mellish, C., Evans, R.: Natural language generation from plans. Computational Linguistics **15** (1989) 233–249
14. Tintarev, N., Kutlak, R., Masthoff, J., van Deemter, K., Oren, N., Vasconcelos, W.: Adaptive visualization of plans. In: UMAP'14 demo track. (2014)
15. Bates, D., Maechler, M., Bolker, B., Walker, S.: lme4: Linear mixed-effects models using Eigen and S4. (2013) R package version 1.0-4.
16. Conati, C., Merten, C.: Eye-tracking for user modeling in exploratory learning environments: An empirical evaluation. Know.-Based Syst. **20(6)** (2007) 557–574
17. Jaeger, T.F.: Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of memory and language **59**(4) (2008) 434–446
18. Baayen, R., Davidson, D., Bates, D.: Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language **59**(4) (2008) 390–412 Special Issue: Emerging Data Analysis.