Comparing Interface Layouts for the Presentation of Multimodal Search Results

Wolfgang Gritz*
TIB – Leibniz Information Centre for Science and Technology
Hannover, Germany
wolfgang.gritz@tib.eu

Christian Otto
L3S Research Center, Leibniz University Hannover
Hannover, Germany
christian.otto@tib.eu

Anett Hoppe†
TIB – Leibniz Information Centre for Science and Technology
Hannover, Germany
anett.hoppe@tib.eu

Georg Pardi
Leibniz-Institut für Wissensmedien
Tübingen, Germany
g.pardi@iwm-tuebingen.de

Yvonne Kammerer
Stuttgart Media University
Stuttgart, Germany
kammerer@hdm-stuttgart.de

Ralph Ewerth‡
TIB – Leibniz Information Centre for Science and Technology
Hannover, Germany
ralph.ewerth@tib.eu

ABSTRACT

Today's search engines allow users to discover relevant information in different types of modalities or media, e.g., web pages, text documents, images, or videos. It is, however, a challenging task to present mixed-modality result lists in an effective and easy-to-skim form. The two most commonly used approaches are to present the modalities side-by-side, each in a separate column of the result page; or to separate the modalities into multiple tabs. However, the field lacks a structured investigation on how the *column* or *tab* layout influence the users' perception and usage of multimodal resources in an academic search task. In this paper, we present a user study (N=50) where the participants were asked to accomplish a search task for a fictive computer science seminar at the university. We evaluate the influence of the different layouts on (1) user search behavior (e.g., time until first resource is saved) and (2) the relevance of the selected resources for the task at hand. Finally, we discuss the results and possible implications for the design of multimodal search result presentation.

CCS CONCEPTS

• **Information systems** → *Information retrieval*; *Web search engines.*

KEYWORDS

SERP layouts, scientific search engine, search behavior, web search

*Also with L3S Research Center, Leibniz University Hannover.
†Also with L3S Research Center, Leibniz University Hannover.
‡Also with L3S Research Center, Leibniz University Hannover.

CHIIR '23, March 19–23, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0035-4/23/03.
https://doi.org/10.1145/3576840.3578335

ACM Reference Format:
Wolfgang Gritz, Christian Otto, Anett Hoppe, Georg Pardi, Yvonne Kammerer, and Ralph Ewerth. 2023. Comparing Interface Layouts for the Presentation of Multimodal Search Results. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '23), March 19–23, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3576840.3578335

1 INTRODUCTION

Nowadays, learning or information acquisition often starts with a search engine. The interdisciplinary research area *Search as Learning* considers aspects from education, social science, psychology, and computer science (CS) to improve web-based learning, and investigates how the design of new search platforms can support learning [5]. Recent research shows that learners rely on different media (e.g., text, video, images) while learning with the help of the web and search engines [15]. It is generally accepted in learning research that a mix of different modalities can potentially be helpful for the learning process [14]. Search engine result pages (SERP) consist of several search results in either homogeneous or heterogeneous interfaces with various influences on user behavior [17]. Regarding homogeneous search interfaces, research has already considered different layout types [6, 16], for example, list, grid, and tabular formats. The list layout is the classic interface and consists of search results snippets displayed in vertical order. The grid layout also uses the horizontal axis for the snippets (e.g., Google Image Search). The *tab* layout shows dedicated tabs for different categories of search results like images, videos, or news. Finally, the tabular layout is similar to the grid layout, but the columns have different semantics (e.g., objective, subjective, and commercial [6]).

To the best of our knowledge, there is no work that has studied layouts for multimodal search results in an academic search task. In this paper, we investigate how users interact with two different SERP layouts in a study with $N = 50$ participants. To this end, we have developed the *Sally* search platform, which can present search results for scientific and educational articles in a tabular layout (i.e., one column per modality, referred to as *column* layout in this work) and in grid layouts (i.e., one tab per modality, referred to as *tab* layout in this work). The participants, CS students, were asked to bookmark resources useful for a fictive research task, being

randomly assigned either the *column* or the *tab* layout of result representation. We evaluate the influence of the different layouts on (1) user search behavior and (2) the relevance of the selected resources.

2 RELATED WORK

Roy et al. [17] investigated four different interfaces: (1) a heterogeneous grid, (2) a heterogeneous list, (3) a simple grid, and (4) a simple list. Their extensive experiments revealed a number of observations regarding user interactions for the different SERPs and task complexity. Khan et al. [8] proposed a search interface, which mixes the different modalities and additionally provides a graph visualization. They observed an improvement in user satisfaction, engagement and knowledge acquisition compared to traditional search platforms. Kuhar et al. [10] analyzed eye tracking data and a *User Experience Questionnaire (UEQ)* to determine emotions, impressions, and stimulation evoked by interaction with two digital library portals. They noticed significant differences in learners' interactions concerning search position and intuitiveness of the homepages. Arguello et al. [3] compared a blended and a non-blended interface variant. In the non-blended variant, only results of one type (e.g., web, image, videos, news, etc.) were displayed and tabs could be used to switch between the types. In the blended interface, an additional tab existed in which all types could be displayed. Rele et al. [16] and Kammerer and Gerjets [6] introduced a tabular layout that is divided into three columns. Kammerer and Gerjets [6] divided the search results into the categories of objective, subjective, or commercial information. They found that university students selected objective search results more often in the tabular layout than in the list layout. Kammerer and Gerjets [7] found that a 3*x*3 grid layout has advantages compared to a traditional Google-like SERP in terms of the trustworthiness of the selected search results. Siu and Chaparro (2014) compared a 3*x*3 grid and a list interface, when participants completed a set of informational and navigational search tasks. They found indicators that participants in the grid layout viewed the top-left result the most, but were often not sure in which order of relevance the search results were shown. Lewandowski and Kammerer [12] concluded in their review that a grid layout seems to support a rather balanced exploration of all search results and that a tabular interface has the potential to guide users to focus on specific kinds of search results or parts of search results, respectively. Lastly, Homte et al. [4] gave an overview of search engines in learning contexts and the different tasks related to this research field.

3 USER STUDY

This section explains our study, starting with information about the participants and the task in Section 3.1. In Section 3.2, we describe the study procedure. Finally, Section 3.3 outlines the technical environment specifically designed for this task: our scientific search engine *Sally*.

3.1 Participants and Task

We recruited $N = 50$ (male: 47, female: 2, non-binary: 1) Bachelor students from different CS seminar courses over the period of eight months, with group sizes of up to 15. Their average age was 23.0 ± 2.8 years in the range of 20 to 31 and they were currently enrolled in semester 6.1 ± 2.1 in CS. The participants did not receive compensation. They were asked to solve the following task: 'Find five resources suitable for a hypothetical 20-minute presentation on the topic of *types of neural networks and their applications* in front of colleagues'. This topic was chosen since the participants had a basic knowledge of CS but minimal prior knowledge of neural networks (according to the local curriculum). This was confirmed as participants rated their prior knowledge of neural networks on a seven-point Likert scale ranging from 1 (*not at all*) to 7 (*expert*) as 2.7 ± 1.4, with no one reporting expert level.

3.2 Procedure

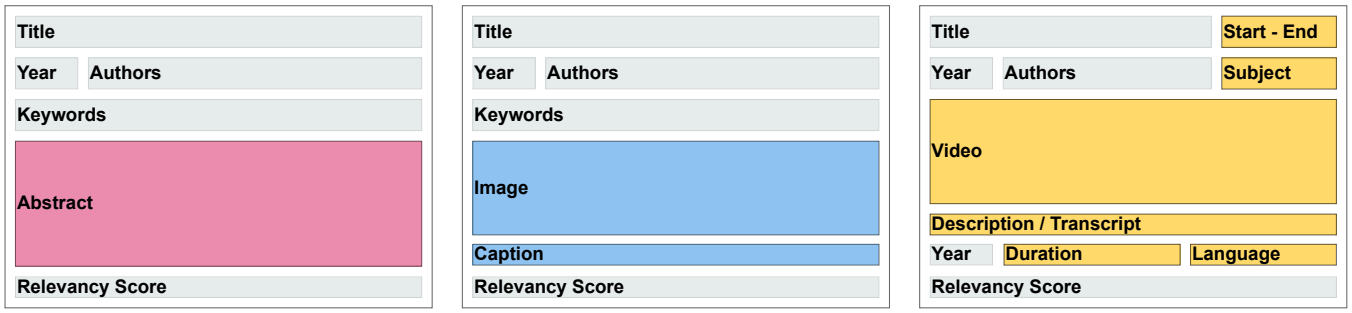
Participants were recruited in five different courses and asked to bring their laptops. To be as consistent as possible, the study instructor only introduced himself briefly and provided the link to the study web page. The study web page first collects demographic information, such as sex, age, English language proficiency, semester, and a self-assessment of prior knowledge on neural networks. Subsequently, the task scenario and instructions are introduced. This information remains available for reference throughout the entire study phase. Since we investigated two layout conditions, 25 participants per condition were randomly assigned to either the *column* or *tab* layout. In addition, as literature emphasizes the influence of the position of snippets [18], the six possible arrangements of modalities in both layouts is randomly (but equally) distributed. For example, in the *column* layout the left column (respectively *tab*) was text, image or video for different participants. Finally, since the study was conducted in seminar rooms, we can not exclude mild interactions between the students, even though we clearly advised against it.

After the participants completed the demographics questionnaire and read the instructions, they started the study by clicking on a button. The participants had 20 minutes to complete the search task (the remaining time was displayed on the bottom left of the screen) but could finish earlier at any time by clicking on a button. During the 20 minutes, the participants could enter queries, save resources, and remove them from the list. After completion, participants were redirected to a page with a short *UEQ* [11]. As we were primarily interested in the usability aspects, we used a shortened version of the original questionnaire that is focused on pragmatic quality. In addition to the questionnaire, there was also a text box for comments and to indicate why they chose the resources with their particular modality.

3.3 Technical Environment: Scientific Search Platform Sally

To have full control over all aspects of the study, from data presentation, over available learning resources, to the feature recording process, we implemented our own scientific search engine called *Sally*. In this section, we describe the data acquisition process and the components of the search engine *Sally* that are relevant to the study.

3.3.1 Frontend. Three different modalities are presented to the learner in *Sally*: research papers (Text), all figures from these papers



(a) Text-based Snippet **(b) Image-based Snippet** **(c) Video-based Snippet**

Figure 1: Schematic illustration of the search result snippets used in *Sally* for the modalities (a) text, (b) image, and (c) video. Red color symbolizes the entities used in text snippets only, blue those in image snippets only and blue those in video snippets only.

(Image), and video segments from a scientific video platform (Video). All resources are open access. The layouts of the three different snippets are detailed in Figure 1. The snippets are displayed in the same way in both variants. The only difference is whether snippets of multiple modalities (column layout) or only one modality (tab layout) are displayed at once (see Figure 2). The snippets can have different heights within a modality (e.g., images in various sizes) and between different modalities. We decided to define the top-left point of a snippet as an anchor point in both layout variants and align the results according to these. This can lead to white spaces between the snippets, however, it was more important in our opinion to maintain the order of relevance. For an entered search query, 25 results have been presented for each of the modalities, top-down in *column* and left to right in *tab* layout, sorted by relevance score retrieved by *Elasticsearch* [1]. Figure 3 shows a screenshot of an example output of the platform in the *column* layout.

3.3.2 Database. For the **text** modality, we downloaded an arXiv dump (timestamp: August 7^{th} 2021) and extracted titles, authors, abstracts, keywords, and submission date of each paper. To extract the figures for the **image** modality, we utilized Python's *fitz* library (version 0.0.1) to convert the PDF files into XHTML. Then, we scanned each document line-by-line for the `""` keyword to find base64-encoded images. The subsequent line usually contains the corresponding image caption, which we extracted as well. The **video** snippets are obtained from the TIB AV-Portal [2], that provides, for example, conference talks and lecture recordings. The platform's database is available in RDF (Resource Description Framework) format. Via Python's *rdflib* library (version 5.0.0) we gathered metadata like title, duration, authors, and publication date. The platform also provides video segment timestamps and automatically generated keywords based on visual concept detection and optical character recognition. We utilized these keywords as tag-like metadata for our video snippets. Finally, we implemented a simple web crawler to obtain the segmented speech transcript of each video as it was not present in the RDF tree.

3.3.3 Backend. *Sally* utilizes Elasticsearch [1] as its full-text search engine and database. Our website interacts with a web service implemented in *Flask* (version 1.1.2), which itself communicates with the database via Python's *Elasticsearch* library (version 7.13.2). For the experiments, the search queries were matched with the resource titles in our database and the respective description text. In the case of papers, this description was the abstract; for the images, it contained the caption, and for the video snippets, it contained the part of the speech transcript associated with that video segment.

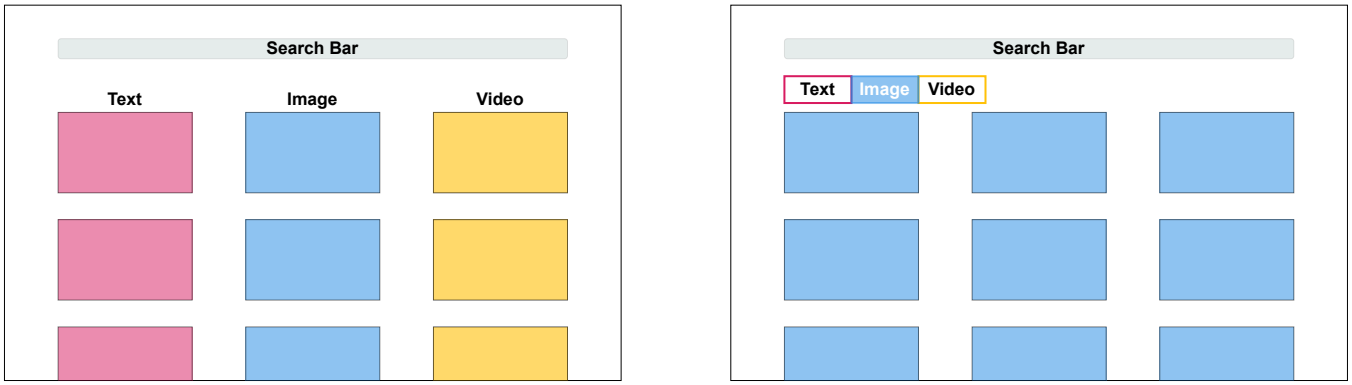
3.3.4 Data Tracking. In the study, we captured a variety of features with respect to demographics, search queries, bookmarked resources, tab changes, and visibility. This feature set is shown in Table 1.
The session time was measured by using the timestamp of the JavaScript events when the start button was clicked, or when it was confirmed that the study should end. For each subsequent event, the time was tracked relative to the start time.

4 EVALUATION

In this section, we describe the comparison of the *column* and *tab* layout. Section 4.1 describes the annotation process of our selected resources. In Section 4.2, we investigate the differences of user search behavior in both layouts. Similarly, in Section 4.3, we examine two additional search efficiency measures. Finally, in Sections 4.4 and 4.5 we compare the results of the *UEQ* as well as qualitative feedback from participants.

4.1 Data Annotation

We manually annotated the usefulness of 93 unique resources our 50 participants selected during the study, distributed over all three modalities. Two workgroup members annotated 20% of the dataset and achieved an inter-coder agreement of 0.74 according to Krippendorff's alpha [9]. The remaining 80% were annotated by one of the co-authors. The annotators were asked to assess whether a resource is relevant (yes/no), i.e., a) the resource describes a type of neural network *or* one application of neural networks, and b) it contains enough information for approximately three minutes in a fictive presentation (or approximately three slides).



(a) *Column* layout (b) *Tab* layout

Figure 2: Schematic representation of the two layouts *column* (left) and *tab* (right) with the three different snippet modalities text (red), image (blue) and video (yellow) are displayed.

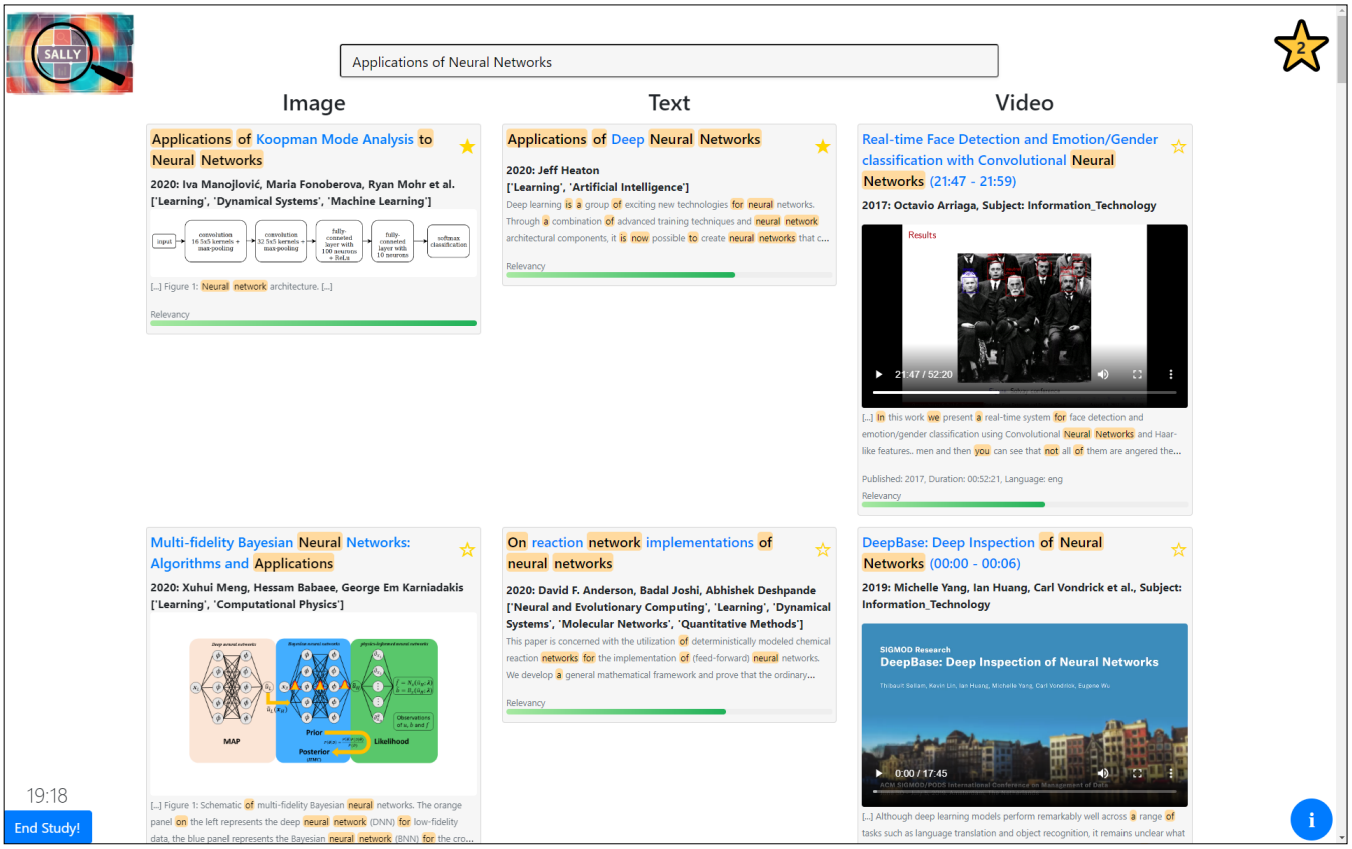


Figure 3: The output of the search engine for the query "Applications of Neural Networks" in (a) *column* or (b) *tab* layout is displayed. On the star icon (top right) the users can toggle a menu to see their saved resources, on "i" (bottom right) they can open a box with the task description and on "End Study!" (bottom left) the users can end the study early.

4.2 User Search Behavior

In this section, we describe a set of 10 user search behavior metrics derived from our recorded features. Our goal is to quantify the user search behavior of our participants with the system and subsequently, investigate the influence of the *column* and *tab* layout. Table 2 shows the corresponding features.

Event	Variable Type	Description	Record Type
Demographics		Sex, Age, Semester, English Skills, Pre Knowledge	Formular
Search Queries	Text	entered Search Queries after pressing Enter	Flask
Saved/Removed Documents	Binary	click on Star icon to save/remove from favorites list	Flask
Tab Change	nominal	(only tab) record if user switches tab with timestamp	Flask
Visibility of Sally	binary	whether the Sally page is currently visible or not	JavaScript

Table 1: This table shows the behavioral features tracked during the user study, the variable type and how they are recorded.

Feature	Description
query_count	absolute number of queries per user
query_length_mean	average length of all queries per user
count_selected_resource	number of resources saved during the search session
count_deselected_resource	number of resources removed during the search session
time_until_first_selected_resource	time in minutes between study start and first document saved
time_until_first_deselected_resource	time in minutes between study start and first document removed
time_until_last_selected_resource	time in minutes between study start and last document saved
time_until_last_deselected_resource	time in minutes between study start and last document removed
selected_resources_timespan	time in minutes between first and last time a resource was saved
session_duration	duration of the search session in minutes

Table 2: This table shows the user search behavior features recorded by *Sally* during the user study.

		median		mean ± std	
Search behavior measure	**p-value**	**col**	**tab**	**col**	**tab**
query_count	0.539	7	8	8.16 ± 4.94	9.24 ± 5.63
query_length_mean	0.383	22.0	21.9	23.77 ± 8.11	22.03 ± 8.95
count_selected_resource	0.359	6	6	6.44 ± 1.96	6.88 ± 2.22
count_deselected_resource	0.359	1	1	1.44 ± 1.96	1.88 ± 2.22
time_until_first_selected_resource	0.662	3.2	3.2	3.50 ± 2.53	4.08 ± 2.95
time_until_first_deselected_resource	0.275	4.5	8.5	6.69 ± 4.78	8.86 ± 5.50
time_until_last_selected_resource	0.221	11.6	13.7	12.43 ± 3.93	13.54 ± 3.89
time_until_last_deselected_resource	0.153	10.5	12.7	9.74 ± 4.80	12.55 ± 4.83
selected_resources_timespan	0.426	8.90	10.7	8.92 ± 3.64	9.48 ± 3.90
session_duration	0.438	14.8	15.8	14.29 ± 3.56	14.92 ± 3.99

Table 3: In this table the *median*, *mean* and standard deviation (*std*) for several user search behavior measures for the column (*col*) and *tab* layout are displayed. Significant differences including p-value between the layouts based on a Mann-Whitney U test are highlighted in bold.

To calculate whether there are significant differences between the two layout settings, we used the Mann-Whitney U test [13]. We found no significant differences between the layouts for any of the variables, as seen in Table 3. The query count and especially the query length are almost identical under both conditions. The amount of saved resources is also close to even. Also, the remaining metrics did not reveal significant differences between the two layouts.

4.3 Search Efficiency

To quantify the task-related search efficiency of the participants, we defined the two measures described in Table 4, which additionally consider our assessment of resource relevance.

Table 5 shows the results; the Mann-Whitney U test was used to check for statistical significance. Again, there is no significant difference for the time until the participants saved their first relevant resource between the *column* and *tab* layout.

4.4 Usability Evaluation

We used 12 items from the *User Experience Questionnaire (UEQ)* [11] to measure the *perceived usability* (i.e., pragmatic quality) of our

Feature	Description
time_until_first_relevant_resource	time in minutes until the first relevant resource saved
time_until_first_irrelevant_resource	time in minutes until the first relevant resource saved

Table 4: This table shows the search efficiency features recorded by *Sally* during the user study.

		median		mean ± std	
Search efficiency measure	**p-value**	**col**	**tab**	**col**	**tab**
time_until_first_relevant_resource	0.616	5.00	4.45	6.07 ± 3.19	5.90 ± 3.85
time_until_first_irrelevant_resource	0.294	3.60	4.20	3.99 ± 2.96	4.92 ± 3.14

Table 5: In this table the *median, mean* and standard deviation (*std*) for several search efficiency measures for the column (*col*) and *tab* layout are displayed. Significant differences including p-value between the layouts based on a Mann-Whitney U test are highlighted in bold.

Question			median		mean ± std	
from	**to**	**p-value**	**col**	**tab**	**col**	**tab**
1: not understandable	→ 7: understandable	0.592	5	5	5.56 ± 0.77	5.24 ± 1.20
1: difficult to learn	→ 7: easy to learn	0.763	6	6	5.48 ± 1.12	5.40 ± 1.58
1: unpredictable	**→ 7: predictable**	**0.042**	**4**	**5**	**4.20 ± 1.04**	**4.76 ± 1.42**
1: slow	→ 7: fast	0.773	5	5	4.68 ± 1.80	4.60 ± 1.55
1: obstructive	**→ 7: supportive**	**0.045**	**5**	**4**	**4.92 ± 1.08**	**4.12 ± 1.48**
1: complicated	→ 7: easy	0.332	6	6	5.64 ± 1.35	5.20 ± 1.66
1: not secure	→ 7: secure	0.415	4	5	4.76 ± 1.05	4.96 ± 1.40
1: does not meet expectations	→ 7: meets expectations	0.683	5	5	4.36 ± 1.25	4.52 ± 1.58
1: inefficient	→ 7: efficient	0.834	5	5	4.72 ± 1.28	4.64 ± 1.19
1: confusing	→ 7: clear	0.451	5	5	4.60 ± 1.61	4.92 ± 1.58
1: impractical	→ 7: practical	0.795	5	5	4.84 ± 1.31	4.88 ± 1.39
1: cluttered	→ 7: organized	0.722	5	5	4.13 ± 2.07	4.29 ± 2.02

Table 6: In this table the *median, mean* and standard deviation (*std*) of the questionnaire of the participants regarding the usability of the search platform for the column (*col*) and *tab* layout are displayed. Significant differences including p-value between the layouts based on a Mann-Whitney U test are highlighted in bold.

platform. The results are shown in Table 6. On the one hand, the *tab* layout was recognized as significantly more predictable than the *column* layout. The reason could be the higher level of control for the participants, i.e., they could focus on the modality tab they preferred. On the other hand, the *column* layout was rated to be more supportive. An explanation for this finding could be, that since the most relevant resources for every modality were presented simultaneously, no additional effort from the user was required.

4.5 Qualitative Feedback

In addition to the 12 items in the *UEQ*, we provided the participants with the opportunity to write down why they chose these modalities in their list of selected resources and to give additional feedback. The answers from a total of 35 participants showed a preference towards textual search results and seem to be independent of the layout type. This observation is underlined by a highly similar distribution of the selected resources with respect to their modality: 71/76 text, 41/40 image, and 13/9 video results in the *columns* and *tabs* layout, respectively. The preference for text over video could be due to participants' aversion to playing audio in a shared space. However, 45 of 50 participants chose at least one non-textual resource. Interestingly, two participants wrote that they chose a video to get a basic understanding of the topic, an image as an illustration for the presentation, and textual resources for the content. Six participants reported confusion about highly similar image search results being displayed and the same video being displayed multiple times. This was possible, because a paper can contain several figures. Further, we have considered individual segments from the videos as independent to provide more fine-grained results.

We will clarify that in the future or make it more obvious by design. Finally, several participants requested additional functionalities such as filtering and sorting functions.

5 CONCLUSIONS

In this paper, we have investigated the influence of two interface layouts for multimodal retrieval results in an academic task on user search behavior, search efficiency, and usability. To this end, we implemented a new search engine *Sally*, and let a group of 50 students gather resources for a hypothetical presentation by means of a user study. We found no significant differences between the two layouts with respect to the search behavior of the participants. However, the subjective usability evaluation revealed significant differences in terms of the supportiveness of the interface, which was rated significantly higher in the *column* layout. Conversely, the *tab* layout was rated higher in terms of predictability, indicating that there is a tradeoff between the layouts.

Generally, the presented study a) has given preliminary insights on possible differences in terms of predictability and supportiveness of the layouts, while it b) also emphasizes the need for further comparison and research on SERP designs. Naturally, the study poses some limitations regarding generalizability of the results: 1) we examined only one search scenario with 2) a certain demographic as participants and 3) a rather small sample size.

For the future, our plans are to improve on the limitations of our study design and enhance our tracking features with respect to touchscreen users, video interactions like pause, play, resume etc., and browsing behavior in other browser tabs.
Future work could also compare the proposed layouts with some kind of universal search (e.g., in a *list* layout). However, this comparison is not trivial, since, for example, ranking effects may occur. For instance, if many papers were displayed first, and videos and images were listed further down, how would that influence the user search behavior?

ACKNOWLEDGMENTS

Part of this work was financially supported by the Leibniz Association, Germany (Leibniz Competition 2018, funding line "Collaborative Excellence", project SALIENT [K68/2017]).

REFERENCES

[1] 2022. Elasticsearch Homepage. https://www.elastic.co/de/ Accessed: 21 October 2022.
[2] 2022. Homepage of the TIB AV-Portal. https://av.tib.eu/ Accessed: 21 October 2022.
[3] Jaime Arguello, Wan-Ching Wu, Diane Kelly, and Ashlee Edwards. 2012. Task complexity, vertical display and user interaction in aggregated search. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, 2012*. ACM, 435–444. https://doi.org/10.1145/2348283.2348343
[4] Jaurès Kameni Homte, Bernabé Batchakui, and Roger Nkambou. 2022. Search Engines in Learning Contexts: A Literature Review. *International Journal of Emerging Technologies in Learning (iJET)* 17, 2 (2022), 254–272. https://www.learntechlib.org/p/220473
[5] Anett Hoppe, Peter Holtz, Yvonne Kammerer, Ran Yu, Stefan Dietze, and Ralph Ewerth. 2018. Current Challenges for Studying Search as Learning Processes. In *7th Workshop on Learning & Education with Web Data (LILE2018), in conjunction with ACM Web Science*.
[6] Yvonne Kammerer and Peter Gerjets. 2012. Effects of search interface and Internet-specific epistemic beliefs on source evaluations during Web search for medical information: an eye-tracking study. *Behaviour & Information Technology* 31, 1 (2012), 83–97. https://doi.org/10.1080/0144929X.2011.599040
[7] Yvonne Kammerer and Peter Gerjets. 2014. The Role of Search Result Position and Source Trustworthiness in the Selection of Web Search Results When Using a List or a Grid Interface. *International Journal of Human–Computer Interaction* 30, 3 (2014), 177–191. https://doi.org/10.1080/10447318.2013.846790
[8] Abdur Rehman Khan, Umer Rashid, and Naveed Ahmed. 2022. An Explanatory Study on User Behavior in Discovering Aggregated Multimedia Web Content. *IEEE Access* 10 (2022), 56316–56330. https://doi.org/10.1109/ACCESS.2022.3177597
[9] Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement* 30, 1 (1970), 61–70.
[10] Maja Kuhar and Tanja Merčun. 2022. Exploring user experience in digital libraries through questionnaire and eye-tracking data. *Library & Information Science Research* 44, 3 (2022), 101175. https://doi.org/10.1016/j.lisr.2022.101175
[11] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work, 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria (Lecture Notes in Computer Science, Vol. 5298)*. Springer, 63–76. https://doi.org/10.1007/978-3-540-89350-9_6
[12] Dirk Lewandowski and Yvonne Kammerer. 2021. Factors influencing viewing behaviour on search engine results pages: a review of eye-tracking research. *Behaviour & Information Technology* 40, 14 (2021), 1485–1515. https://doi.org/10.1080/0144929X.2020.1761450
[13] Henry B. Mann and Donald R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1 (1947), 50 – 60. https://doi.org/10.1214/aoms/1177730491
[14] Richard E Mayer and Roxana Moreno. 2003. Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist* 38, 1 (2003), 43–52. https://doi.org/10.1207/S15326985EP3801_6
[15] Georg Pardi, Daniel Hienert, and Yvonne Kammerer. 2022. Examining the use of text and video resources during web-search based learning—a new methodological approach. *New Review of Hypermedia and Multimedia* 28, 1-2 (2022), 39–67. https://doi.org/10.1080/13614568.2022.2099583 arXiv:https://doi.org/10.1080/13614568.2022.2099583
[16] Rachana S. Rele and Andrew T. Duchowski. 2005. Using Eye Tracking to Evaluate Alternative Search Results Interfaces. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 49, 15 (2005), 1459–1463. https://doi.org/10.1177/154193120504901508
[17] Nirmal Roy, David Maxwell, and Claudia Hauff. 2022. Users and Contemporary SERPs: A (Re-)Investigation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain*. ACM, 2765–2775. https://doi.org/10.1145/3477495.3531719
[18] Christina Siu and Barbara S. Chaparro. 2014. First Look: Examining the Horizontal Grid Layout using Eye-tracking. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58, 1 (2014), 1119–1123. https://doi.org/10.1177/1541931214581234 arXiv:https://doi.org/10.1177/1541931214581234