MedienPädagogik
Zeitschrift für Theorie und Praxis der Medienbildung
www.medienpaed.com **ISSN 1424-3636**

Themenheft Nr. 54:
Forschungssynthesen in der Mediendidaktik. Ansätze und Herausforderungen
Herausgegeben von Svenja Bedenlier, Katja Buntins, Annika Wilmers und Michael Kerres

Cognitive Biases in Screening Processes – Search Strategies in Educational Technology Research

A Systematic Review on Learning with Virtual Reality

Katja Buntins[1], Miriam Mulders[1], and Nadine Schröder[2]

[1] Universität Duisburg-Essen

[2] Hochschule Bielefeld

Abstract
The aim of this article is to empirically investigate the advantages and disadvantages of different search strategies in synthesizing research papers that use Virtual Reality (VR) educational technology. The aim is to identify cognitive biases on the part of the reviewers through different concrete searches. Using two search strategies, the study identifies the extremes between an AND search that finds as few irrelevant studies as possible but overlooks relevant ones, and an OR search that searches as broadly as possible but picks up many irrelevant studies. The article aims to show how systematic searches in educational research should be designed to adequately address the typical challenges of systematic analyses (e.g., recall-precision problem, cognitive load). The search strategies were developed based on a Google Scholar search for existing systematic reviews on VR. Here, the two search strategies differed only in terms of their linkage between a technological (VR) and a pedagogical search term. The two elements were linked with either an AND or an OR. The search items were screened in a two-person cross design and evaluated on different measures of precision and recall. There was no evidence that the more comprehensive search (OR) is superior to the narrower search (AND), but slight evidence of cognitive biases in the screening or search process in the more comprehensive search (OR). These results should be further evaluated, investigated, and, above all, replicated in further studies.

Buntins, Katja, Miriam Mulders, and Nadine Schröder. 2023. «Cognitive Biases in Screening Processes – Search Strategies in Educational Technology Research. A Systematic Review on Learning with Virtual Reality». *MedienPädagogik* 54 (Research Syntheses): 103–124. https://doi.org/10.21240/mpaed/54/2023.11.19.X.

Kognitive Prozesse in Screening-Prozessen – Suchstrategien in der Bildungstechnologieforschung. Systematische Übersichtsarbeiten zum Lernen mit virtueller Realität

Zusammenfassung

Ziel dieser Studie ist es, die Vor- und Nachteile verschiedener Suchstrategien bei der Synthese von Forschungsarbeiten, die die Bildungstechnologie der virtuellen Realität (VR) nutzen, empirisch zu untersuchen. Hierbei ist das Ziel, kognitive Verzerrungen seitens der Reviewer:innen durch verschieden konkrete Suchen zu identifizieren. Mittels zweier Suchstrategien sollen die Extrema zwischen einer Suche (AND), die möglichst wenig irrelevante Studien findet, aber dafür auch relevante übersieht und einer Suche (OR), die möglichst breit sucht, aber hierbei viele irrelevante aufnimmt, dargestellt werden. Die Studie will aufzeigen, wie systematische Suchen in der Bildungsforschung gestaltet sein sollten, um die typischen Herausforderungen systematischer Analysen (z.B. Recall-Precision-Problem, kognitive Belastung) adäquat zu berücksichtigen. Die Suchstrategien wurden auf der Grundlage einer vorangegangenen Google Scholar-Suche nach bereits durchgeführten systematischen Übersichten zur VR entwickelt. Hierbei unterschieden sich die zwei verschiedenen Suchstrategien nur in Bezug auf ihre Verknüpfung zwischen einem technologischem (VR) und einem pädagogischen Suchterm. Die beiden Elemente wurden entweder mit einer AND oder einer OR Verbindung verknüpft. Die Suchbeitrage wurden in einem Kreuzdesign von zwei Personen gescreent und in Bezug auf verschiedene Präzisions- und Recallmaße evaluiert. Es fanden sich keine Hinweise dafür, dass die umfangreichere Suche (OR) der engeren Suche (AND) überlegen ist und jedoch leichte Hinweise auf kognitive Verzerrungen im Screening bzw. Suchprozess bei der umfangreicheren Suche (OR). Diese sollten in weiteren Studien weiter evaluiert, untersucht und vor allem repliziert werden.

1. Introduction

Research syntheses aim to comprehensively aggregate results on a specific research question (Gough, Oliver, and Thomas 2017) and then, depending on the type of research review, to describe, summarize, and quantify findings or derive new models or theories from the current state of research (Grant and Booth 2009). When creating such syntheses, researchers need an appropriate and efficient search strategy to find the most relevant research results on the question at hand. Ideally, a search strategy should find all relevant research and exclude all irrelevant research (Sampson and McGowan 2006). However, researchers often fail to plan and think through their search strategies sufficiently, to orient themselves toward previously successful syntheses of research, to involve information scientists, or even to test their search strategies in advance. Accordingly, many research syntheses are

carried out with inadequate scientific support under high time pressure to produce and publish results. Researchers rarely consider that such decisions affect their research and that different search strategies lead to different research results. What is more, they often do not consider the potentially far-reaching implications of different search strategies. The output, that is, the number of research articles, may vary greatly depending on the search strategy. This has further implications for the time needed to screen the papers, the cognitive load for screeners, and the likelihood of overlooking relevant criteria when there are too many criteria. There are few evidence-based contributions that examine how different search strategies affect further activities in syntheses of research (Geersing et al. 2012; Rogers, Bethel, and Boddy 2017).

Despite the lack of attention to how search strategies may affect research syntheses, researchers in many disciplines are quick to summarize the existing body of research on their topic (Andrews and Farris 1972; Lasser et al. 2020). The time pressure researchers are under can affect the psychological processes of those synthesizing research, impairing their ability to concentrate and increasing their cognitive load. Using the educational technology of Virtual Reality (VR) as an example, we aim to investigate the advantages and disadvantages of two different search strategies in the context of a specific synthesis of research.

In fast-growing research fields with a large body of literature (Larsen and Ins 2010), research syntheses play an important role (Eden 2002). This is particularly true of fields like educational technology that are developing quickly and must be translated quickly into practice because of their high practical relevance (Radianti et al. 2020; Wu, Yu, and Gu 2020). It is important that existing knowledge in the field, best practice examples, effects, mediators, and moderators are made available in a readable and condensed form, which is the aim of meta-analyses and other types of systematic reviews (Liu et al. 2017; Makransky and Petersen 2021).

This article is divided into the following parts. Part 2 presents the theoretical background. Part 3 discusses challenges of developing adequate search strategies in research syntheses. The focus here is on research about learning with VR, a topic in the field of educational technology that is currently attracting substantial research interest. Part 4 outlines the research questions and methodological approach, including the sample, procedure, and data analysis. Part 5 discusses the results, and Part 6 concludes by exploring the implications and limitations of the work.

2. Theoretical background

Depending on the type of research synthesis, researchers proceed systematically through a series of steps (Grant and Booth 2009). These include:

1. developing a research question,
2. defining inclusion and exclusion criteria,
3. searching,
4. screening,
5. quality checking,
6. coding,
7. extracting the results,
8. and synthesizing.

The methodological issues in research syntheses can be simplified by considering the relevant studies as a population, like the populations that are the subject of primary research. By looking at them in this way, approaches from primary research can be applied to secondary research. The population of the searched studies is defined by the research question, the operationalization of the population in the search (search query), and the specification of the inclusion and exclusion criteria. The population here refers to all studies about which the researcher wants to make statements, regardless of whether they are found. It can also be considered a true population. The literature search is a very important part of any research synthesis. In addition to deciding on the search terms, the literature search also includes the selection of databases and other search strategies, such as snowball searching or searching for all literature published on a topic to date (Gehanno, Rollin, and Darmoni 2013; Gusenbauer and Haddaway 2020; Shaffril et al. 2021).

3. Challenges of literature search in the context of research syntheses

3.1 The recall-precision problem

The proportion of studies found by a search is called **coverage** (Martín-Martín et al. 2021). If the found part of the population of studies does not fit the searched population, this is called a **coverage error** (Alvarez and VanBeselaere 2005). A coverage error is therefore a non-sampling error. Two different types of coverage errors may occur (see Figure 1). On the one hand, the search results may contain studies that are not relevant according to the criteria used in the synthesis: This is called **over-coverage**. On the other hand, the results may contain studies that are not covered by the search strategy: This is called **under-coverage**.

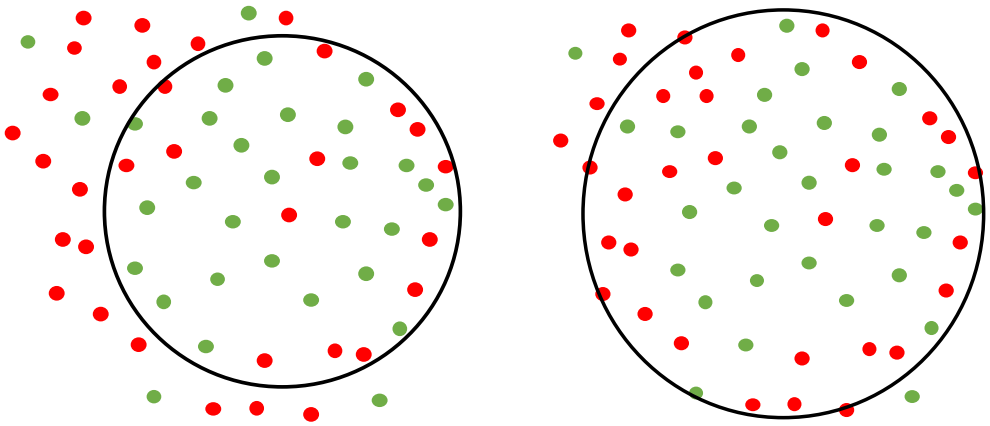

Fig. 1: Under- and over-coverage biases (red dots = not part of the population of relevant studies; green dots = part of the population of relevant studies, circle = found results).

In most literature syntheses, both forms of biases occur. This is described as the **recall-precision problem** (Bramer, Giustini, and Kramer 2016). Although there are a number of different formal definitions of recall, recall is defined here as the studies included in the search (n_{inc}) relative to all available and relevant studies (N) (Eysenbach, Tuische, and Diepgen 2001). Formally, then:

$$Recall = \frac{n_{inc}}{N}$$

This means that the closer the **recall** value is to one, the greater the proportion of relevant literature found in the search and the lower the under-coverage bias. However, if one tries to calculate the recall empirically, one faces the problem that the true population cannot be determined. Otherwise there would be no research on the advantages and disadvantages of different search strategies (Evans 2002). For this reason, recall must be estimated. Various comparison variables can be used for this purpose, such as the results of a database search. This produces a relative estimator for the quality of a subpopulation search (Straube et al. 2021). If one assumes that the included studies contain only correctly screened studies (i.e., positive examples), then the recall can also be described as sensitivity (Cooper and Varley-Campbell et al. 2018). **Precision**, on the other hand, describes the proportion of included studies in the review out of all studies (n_{inc}) that were found in the search (s_{all}). Formally, then:

$$Precision = \frac{n_{inc}}{s_{all}}$$

Again, of course, it should be noted that there are various formal definitions of precision. However, all these definitions agree that the more accurately the search string captures the true population, the closer the precision value is to one. The lower the precision value, the greater the over-coverage bias and the greater the number of irrelevant studies. A way to deal with the over-coverage bias is to screen the abstracts based on inclusion and exclusion criteria (Lee et al. 2012; Sampson, Tetzlaff, and Urquhart 2011). The aim of every research synthesis is to achieve a perfect balance between sensitivity and precision.

The recall-precision problem occurs when one factor is increased and the other simultaneously decreases (Stock and Stock 2013). Thus, if the recall is to be increased to obtain a more complete set of relevant documents, the precision decreases, so that more irrelevant documents are found. Conversely, if precision is to be increased to produce less work, this will decrease the number of relevant documents. To address the trade-off between work load (Bedenlier et al. 2020) and completeness, it is important to consider the details of the search strategy. These include the match between the research question and the search strategy as well as the quality of the search strategy.

There are several empirical research approaches and studies in medicine and other natural sciences that have dealt with the question of what changes in search strategy have what effect on various result parameters (e.g., number of hits, recall, precision). For example, there are a large number of studies dealing with the recall and precision of various databases. Bramer, Giustini and Kramer (2016) looked at the quality of Google Scholar compared to other databases in the context of liver cancer, referring to previous studies that had asked the same question before in other biomedical and medical contexts (Bramer et al. 2013; Gehanno, Rollin, and Darmoni 2013). Other studies looked at whether a non-standardized data-based search leads to more results than a standardized database search. Cooper and Lovell et al. (2018) found in the context of a study with very few matching studies that a non-standardized search leads to more matching hits. Furlan, Irvin and Bombardier (2006) asked to what extent a standardized database search can help with very specialized questions or lead to not finding relevant literature. They found that a standardized search was possible for their research objective, which was to identify nonrandomized studies. These findings indicate that the search strategy has an important influence on recall and precision. For this reason, search strategy standards are being developed. For example, Cooke, Smith and Booth (2012) explored whether the **PICO** (Population, Intervention, Comparison, Outcome) or **SPIDER** (Sample, Phenomenon of Interest, Design, Evaluation, Research Type) frameworks lead to more efficient exploitation of data. These frameworks help to conduct a search that is as sensitive as possible but still prescient (Methley et al. 2014). An addition to the PICO framework was “study type” or “study design”, so the acronym became

PICOS (Akers, Aguiar-Ibáñez, and Baba-Akbari 2009). In some studies, the S in PICOS stands for setting (Robinson, Saldanha, and Mckoy 2011). Studies evaluating these approaches have come to different conclusions. This inconsistency in the findings suggests that the choice of the right search framework depends on the research topic (Cooke, Smith, and Booth 2012; Methley et al. 2014; Rehman 2021).

The quality of systematic reviews may be affected at various points in the screening process. The search is an important step that affects all of the steps that follow. In this study, we use the example of VR to address the question of the extent to which the (im)precision of the search affects cognitive processes. The following section deals with researchers' cognitive processing when screening articles.

3.2 Cognitive load during screening processes

In the literature cited above, it is often assumed that a complex and nearly complete literature search is usually not possible, primarily due to the time involved, which would significantly limit the timeliness of the study. In addition, the retrieval of larger quantities of literature through more imprecise search strings leads to more literature that must be screened. However, Sampson, Tetzlaff and Urquhart (2011) showed that researcher errors already occur in the development of search strings. Shepperd, Bowes and Hall (2014) examined the extent of researcher bias and concluded that researchers are by far the greatest source of error in literature reviews. Biases are associated with several concepts, including prior knowledge, statistical and data processing skills, interests, and opportunities. Additionally, Süß and Schmiedek (2000) showed that in psychometric studies of cognitive performance that last several hours, fatigue and loss of motivation lead to a drop in performance, whereas effects of practice only help to improve the results slightly. Konig, Buhner and Murling (2005) revealed that working memory and cognitive processing were the most important predictors of multitasking performance in addition to attention and fluid intelligence. Yet there appears to be little research on cognitive resources of researchers during the review processes.

A simple, common theory for describing the limited cognitive processing capacity is that of cognitive load (Sweller 1988; Sweller, van Merriënboer, and Paas 2019). This empirically validated theory assumes that working memory is limited in its resources. When these limited resources are exceeded, no further input can be processed. Originally developed for teaching and learning settings (Buchner, Buntins, and Kerres 2021), it now has a variety of applications (Engström et al. 2017; Zhang et al. 2017). In cognitive load theory, three types of cognitive load are distinguished (Schnotz and Kürschner 2007; Sweller 2003): intrinsic cognitive load, extraneous cognitive load, and germane cognitive load. Intrinsic cognitive load means that the cognitive load increases with complex and novel tasks. Extraneous cognitive load

refers to the form of the material, with material that contains a large amount of irrelevant information increasing cognitive load. Germane cognitive load refers to the load of the learning itself.

If this is applied to the screening of articles for research syntheses, the following assumptions can be derived:

1. The better one knows the research question and the inclusion and exclusion criteria, the lower the cognitive load. In other words, the more prior knowledge about the research topic is available and the better the criteria are remembered, the less the extraneous load and the better the performance in screening and in including and excluding articles correctly.
2. The more irrelevant articles there are in a search query, the higher the cognitive load. This is because as the number of articles increases, processing time increases, and fatigue and errors are more likely. The more irrelevant articles there are, or the more articles are generally included in a search query, the higher the cognitive load. This is because as the number of articles increases, processing time increases, and fatigue and errors are more likely.
3. The less specific the abstracts are, the higher the cognitive load. This is because if the concepts mentioned in the abstract remain unclear and are not sufficiently described, this often requires the reviewer to read the abstract more than once, to combine information, or even to look at the full text. Overall, the processing time per article and the cognitive load increases.
4. The vaguer the constructs in the research question are, the higher the cognitive load. This is because the less specifically the constructs were defined in advance, the more uncertain the reviewer becomes about when to include or exclude articles. The missing information creates uncertainty and increases the cognitive load. In addition, nonspecific constructs in the research question increase the likelihood that different reviewers will have different understandings of the concepts and potentially decrease interrater reliability. These assumptions are probably not exhaustive and, more importantly, they are not empirically tested. Rather, they are meant to indicate that it is important to pay more attention to cognitive resources and the resulting researcher biases.

3.3 Search strategies in the field of Virtual Reality

Research syntheses in the medical field have been studied extensively with regard to the choice of the search string, and there are large numbers of studies dealing with how a search string should be constructed (Salvador-Oliván, Marco-Cuenca, and Arquero-Avilés 2019). However, little research of this kind has been done in the

field of educational technology (Bedenlier et al. 2020). The aim of this article is to use the example of VR to look at which search string best considers the needed time and cognitive resources.

With educational technologies such as VR that allow a new form of access to learning content, it is important to summarize existing study findings in a readable and condensed form. Two examples are the syntheses of Radianti et al. (2020) and Wu, Yu and Gu (2020), which dealt with the effectiveness of VR in various educational settings and forms of technology and tried to draw conclusions regarding VR in educational scenarios in general. Systematic analyses that address more specific questions concerning VR are difficult to find. With the rapid advances in VR technology, more and more studies are being published, but most focus on the technology itself rather than on instructional parameters (Allcoat and Mühlenen 2018; Mulders, Buchner, and Kerres 2020). There is a lack of meta-studies that aggregate relevant primary studies to answer questions for research and practice.

In the field of educational technology, learning in VR is gaining importance. As a result of increasingly cost-effective software and hardware, more open-source programs, and educational institutions‘ improved technical equipment, teachers and facilities managers face the question of how they can use technologies such as VR in teaching and learning scenarios in a meaningful and value-generating way. Moreover, there are various VR visualization technologies. For example (1) ears-wrapped VR Head Mounted Displays (HMD) that cover the eyes or (2) VR with three-dimensional content displayed on two-dimensional smartphone or tablet screens. Therefore, not only general research syntheses are needed, but also more precise syntheses that are targeted to specific VR technologies and groups of learners. These kinds of studies are rare compared to studies in which the use of VR is compared with analogue formats (Buchner 2022).

Methodologically, however, this article aims to investigate and highlight the differences between two search strategies. It deals with the challenge of choosing a search strategy that is specific enough to find as many relevant studies as possible, but that does not return too many irrelevant hits. This means identifying all articles that investigate and compare the use of HMD- and desktop-based VR.

This article explores the advantages and disadvantages of two different search strategies:

1. a search strategy that is as specific as possible with the result of a manageable number of studies, but with the risk of some relevant studies not being found (AND search)
2. a wider search strategy that results in many studies being found, with some of them being irrelevant (OR search).

The two search strings differ only in how the strings were connected. Whereas in the AND search, all search strings of (1) HMD- and (2) desktop-based VR were combined with an AND, in the OR search, these strings were connected with an OR. As a result, the AND search finds only those studies that name both VR technologies in the abstract and/or in the title or keywords. The OR search finds all studies that name at least one of the two technologies. It is worth noting that the aim is not to compare the quality of different search terms but to see what dangers arise from the broader (OR) and narrower (AND) search strings.

4. Methods

4.1 Research questions

The main research question investigated in this article is: "What is the impact of two different search strategies on the screening process in a research review?". Here, the aim is to place a special focus on the cognitive-psychological factors in screening. As described above, there is little research on fatigue processes during the screening process in reviews. By choosing an AND or an OR connection, we can address the following research questions:

1. How much larger is the data set in the OR search?
2. How many studies are not found in the AND search?
3. Is there evidence of cognitive biases in the two searches?

The two search designs were examined in the context of the content question, "Does desktop-based VR differ from HMD-based VR in terms of student learning and outcomes?".

The search strategies were developed based on a prior Google Scholar search for previously conducted systematic reviews using the search terms "virtual reality", "systematic review", "education", "technology", "learning", which identified the following meta-analyses, on which our own search strategies were based: Kavanagh et al. (2017), Radianti et al. (2020), Jensen and Konradsen (2018), Merchant et al. (2014).

From these search strings and the related research, as well as our prior knowledge regarding VR in educational settings, two search strings were created that included the technologies (HMD vs. desktop) to be compared (see Table 1). These were combined with two different educational context search strings, one related to the institutional context and one to the person. In addition, after screening some of the hits, we excluded some words using a NOT term. The search string was subsequently reflected upon, revised, and improved in dialogue with information scientists.

	Topic	Search String
Virtual reality	HMD-based VR	“head mounted display*” OR “head-mounted-display*” OR “HMD” OR “Oculus” OR “Samsung Gear” OR “Samsung Odyssey” OR “Google Cardboard” OR “Pimax” OR “Playstation VR” OR “Google Daydream” OR “HTC” OR “Pico” OR “Vive” OR “HP Reverb” OR “Valve Index” OR “Lenovo Mirage Solo” OR “immersive VR” OR “I-VR” OR “IVR” OR “immersive virtual realit*” OR “degrees of freedom VR” OR “degrees of freedom virtual realit*” OR “Kokoda VR” OR “immersive”
	AND/OR	
	Desktop-based VR	“desktop virtual realit*” OR “desktop VR” OR “desktop-VR” OR “desktop-3D” OR “DVR” OR “D-VR” OR “desktop 3D” OR “360 degrees video*” OR “360 degrees-video*” OR “360°-video*” OR “360° video*” OR “mobile VR” OR “laptop VR” OR “laptop 3D” OR “non-immersive” OR “low immersive” OR “less immersive”
	AND	
Education context	Institution	“school*” OR “gymnasium*” OR “youth organization*” OR “youth organisation*” OR “youth center*” OR “youth centre*” OR “leisure facility*” OR “higher education” OR “kindergarten*” OR “universit*” OR “colleg*”OR “apprenticeship*” OR “education*” OR “training*” OR “academic institut*” OR “academic context*” OR “learning institute*” OR “learning context*” OR “K-12*” OR “K12” OR “P-12” OR “P12” OR “museum*” OR “galler*” OR “librar*” OR “academ*” OR “tutor*” OR “class” OR “classes” OR “learning center*” OR “learning centre*”
	OR	
	Person	learner* OR student* OR trainee* OR graduate* OR teacher*
	NOT	
	Exclusion	“dynamic voltage restorer” OR “hard-to-cook” OR “hydrothermal carbonization” OR “heat transfer coefficients” OR “maximal inspiratory pressure” OR “high-temperature combustion” OR “Pico hydropower” OR “systematic review” OR “meta analysis” OR “heat transfer coefficient” OR “HIV testing and counseling” OR “desktop 3D printer” OR “integrated voltage regulators”

Tab. 1: Search strings.

The searches were then carried out on 16 October 2022 in *Scopus, Web of Science, PubMed,* and *ERIC*. These databases were selected because, according to studies on subject coverage, they offered broad coverage of this research topic (Gusenbauer and Haddaway 2020; Köstler 2023). Articles and conference proceedings were included in the search.

To answer the research question of which search strategy is more efficient, two searches were carried out. As presented in Table 1, we varied whether the search string had to contain (1) at least one HMD-based VR search term AND one desktop-based VR search term each or (2) at least one of the HMD-based VR search terms OR one of the desktop-based VR search terms. The two searches were cleaned of duplicates in their respective search queries. In theory, all search results from the AND search should also be found with the OR search.

4.2 Search results and research processes

As described above, two data sets were deliberately created that differ in the number of papers found. Consequently, the AND search results in 251 studies, while the OR Search results in 6403 studies. To ensure that the results are comparable, the analysis was carried out with 251 articles. For the AND linkage, all 251 were screened. The size of the sample led to a confidence interval of 95% and a margin of error of 6% (Kupper and Hafner 1989).

To evaluate both searches in terms of coverage, effort, and cognitive bias processes, both authors screened both searches. One author started with the OR link search and the other with the AND link search. Afterwards, they switched. This was done to distribute sequence effects and cognitive fatigue evenly between the two searches.

4.3 Inclusion and exclusion criteria

Inclusion and exclusion criteria were defined for the description of the population of studies, that is, the number of studies about which one wants to make statements (see Table 2). These include, above all, that there is a comparison of desktop-based and HMD-based VR. However, there are also other exclusion criteria, such as no educational context or no original data being collected. The exclusion criteria were applied hierarchically. This means that “1. no English language” was assigned if the study was in a language other than English, regardless of what other criteria would fit the study. Similarly, the exclusion criterion “6. secondary research” was only assigned if none of the previous criteria fit. It is therefore enough if one exclusion criterion is fulfilled for an article to be excluded. But all inclusion criteria must be fulfilled for the articles to be included.

Inclusion criteria	Exclusion criteria (in hierarchical order)
• English language • published between 2019 and 2022 • original empirical data • articles & conference proceedings • comparison of desktop-based and HMD-based VR • educational context • primary search	1. no English language 2. published before 2019 3. no original empirical data 4. other publication type than articles and conference proceedings 5. no comparison of desktop-based and HMD-based VR 6. no educational context 7. secondary research

Tab. 2: Inclusion and exclusion criteria.

4.4 Evaluation criteria

Once again, as a reminder, the study was conducted on 251 articles. The articles in the OR link search were randomly selected. These 2.251 studies were analyzed using different evaluation criteria. These are shown in Table 3.

There were three different goals here. The evaluation criteria "total number of studies" and "precision in the search" are used to determine the effort. The evaluation criteria "estimation of recall" and "studies not found" were used to determine coverage. The two inter-rater reliabilities (IRR) were used to represent biases due to cognitive processes. These differ, as described in the table, in that the inter-rater reliability (criteria compliance) is about the hierarchy of the exclusion criteria. In other words, here we check whether the same reason for exclusion was mentioned—which is an indicator of accuracy.

Evaluation criteria	Description	Formalization	Interpretation
Total number of studies	Total number of studies found in each search	$n_{fA \vee O}$	The more studies there are, the more work is required.
Estimation of recall	Proportion of included articles after screening from one search in all included articles after screening from both (To avoid biases, the included number of articles in the OR-link sample is estimated. The estimation takes place by scaling up.)	$R_{A \vee O} = \frac{n_{inc_{A \vee O_E}}}{(n_{inc_A} + n_{inc_E})}$	The higher the recall value, the more studies in total are found by the search strategy.
Precision in the search	Proportion of included articles after screening in the total number of studies	$P_{A \vee O} = \frac{n_{inc_{A \vee O}}}{n_{fA \vee O_S}}$	The higher the precision value, the more precisely the search string describes what is being searched for.
Inter-rater reliability (IRR)	Proportion of matches regarding inclusion and exclusion decisions after screening on the search	$IRR_{IE} = \frac{M_{inc \wedge exc_{A \vee O_S}}}{n_{fA \vee O_S}}$	The higher the IRR is, the more accurate reviewer's work.
Inter-rater reliability (criteria compliance)	Proportion of the search query matched with respect to the inclusion and exclusion criteria	$IRR_{IE} = \frac{M_{cri_{A \vee O_S}}}{n_{fA \vee O_S}}$	The higher the IRR is, the more accurate reviewer's work.
Studies not found	Number of missing studies in the AND-Link search found in the OR-link search		The higher the value, the more studies are overlooked in the narrower search string.
Included studies (unique)	Number of studies found in the one search only.	$n_A \neq \{O\} \vee n_O \neq \{A\}$	Study was only found with one search. The OR search refers to the entire sample of 6543 studies.

Legend:

n: number of studies *f:* found studies	*inc:* included studies *exc:* excluded studies	*M:* matching *cri:* criterion	*A:* AND-link search *O:* OR-link search	*E:* estimated *s:* sampled studies

Tab. 3: Evaluation criteria.

5. Results

The aim of this study was to determine which of the search strategies (AND vs. OR) is best suited to generate a search query that contains as many relevant studies as possible, excludes as many irrelevant studies as possible, and overlooks as few relevant studies as possible in the screening process. Altogether, using the example of the educational technology VR, we have tried to find out what suitable search strategies include that adequately address typical challenges of systematic analyses (e.g., recall-precision problem, cognitive load). For this purpose, two search strategies were evaluated according to the different evaluation criteria described above. The results are shown in Table 4. Here it can be seen that the OR search leads to 6152 more studies than the AND search. It took each of the two scientific researchers about 2.5 hours to screen the 251 articles. The articles in the OR search were randomly sampled. One of us started with the 251 articles in the AND-link Search. The other one started with the OR search. Taken together, the additional effort screening all articles found in the OR search can therefore be estimated at about 60 hours per person (2.5 per 251 articles). Here, the percentage given in precision is significant higher in the AND search. The recall estimates are about the same but are somewhat lower in the AND search. Looking at the interrater reliability, there are no relevant differences in either measure. There is no difference when looking at IRR inclusion and exclusion decisions. In the interrater reliability (criteria compliance) (see Table 2), there is slightly higher reliability for the OR search.

Evaluation criteria	AND search	OR search
Total number of studies	251	6403
Included numbers	49 of 251	2 of 251
Estimated numbers of included studies	-	51 (2*(6403/251))
Included studies	2 of 251	1 of 251
Recall estimation in the search	0.92	0.96
Precision in the search	0.195	0.008
Inter-rater reliability	0.964	0.964
Inter-rater reliability (criteria compliance)	0.773	0.800

Tab. 4: Results of the AND and OR search.

6. Discussion

The aim of this study was to empirically investigate the advantages and disadvantages of different search strategies using the example of a systematic literature search in the context of VR. The peculiarities of the field are, on the one hand, the fast-growing pace of the research field and, on the other, the multitude of studies comparing VR with other technologies or analogue formats instead of two VR realizations.

To this end, a broader and a narrower search were conducted and examined in terms of various recall, precision, and accuracy parameters. In the searches, there were no relevant differences in the accuracy of the screening (described using the two IRR evaluation criteria and none regarding a relevant low recall). On the contrary, more unique studies were found in the narrower (AND) search: studies that did not appear in the data set of the broader (OR) search. Since this is technically impossible, we suspect that human errors in removing duplicates occurred during this search. These were identified using the Eppi-Reviewer application and then automated for more than 98%, the rest manually. Subsequently, duplicates were searched for manually. This can lead to both technical and human errors. Another source of human error is the download limitation in *Scopus* and *ERIC*. In these databases, only a certain number of articles can be downloaded at the same time, so several files must be downloaded due to further filtering algorithms. These must be determined by means of logical closure.

There are very clear differences in precision – for example, there are 25 times more hits than in the broad OR search, and the precision value for the broader search is 0.008 compared to 0.195 for the narrower search (AND search). Hence, the workload is much higher for the broader search.

It should be noted that the inclusion and exclusion criteria in this search are manifest variables, meaning that they are directly observable. If we take a closer look at the missing matches, the cases where the same exclusion criteria were not selected, we see that they are all due to a lack of definition of the educational context. On the one hand, this could be due to a lack of concentration. On the other hand, however, there is a blurred distinction between psychoeducation and learning within the framework of rehabilitation measures. For example, the authors did not agree on whether a study on driving under the influence of alcohol among college students is education or still psychoeducation (Madigan and Romano 2020). To understand the processes here precisely, qualitative studies would be needed that apply the think aloud method, for example.

In the current search, we could not find any evidence of a greater cognitive load in the broader search (OR search). This means that the IRR is not lower here and is even higher in the exact code coding (criteria compliance). This can be interpreted as an indication that the cognitive load does not differ significantly. However, we believe that more studies are needed in this area. These vary further in terms of the fuzziness of the criteria, the experience of the reviewers, and the quantity of hits. It should also be noted that there is still a great deal of variance in the sample of 251 studies. In a further step, it would be interesting to see whether the values from the OR search can be confirmed in this way in other samples.

7. Conclusion

Our analysis showed that a narrow search string (AND search) does not lead to less relevant results, such as significantly fewer studies, than the broader search string (OR search). We will therefore continue to work with this in our upcoming meta-analysis. By doing so, we hope to make our work more efficient by systematically reducing the number of irrelevant articles.

Our finding is in line with Cooke's (2012) findings. However, it should be noted that the search string must indeed be as precise as possible. As mentioned above, cooperation with librarians and information scientists is crucial here. Studies have shown that they are not consulted in most review processes, although they would be willing to help (Grossetta Nardini et al. 2019). They can, for instance, help to refine a search string and point out missing elements in a search string.

Furthermore, an additional manual search should be attempted if necessary. On the one hand, this offers the possibility of finding literature outside of large databases and also grey literature or pre-print documents (Haddaway et al. 2015; Haddaway and Bayliss 2015).

7.1 Further research

In the field of VR, where there are already many studies with more being added all the time, it is crucial to design research syntheses efficiently. In this field, to systematically aggregate current technological trends, a systematic review is a state-of-the-art method providing profound results. Nevertheless, more research is needed on the question of how researchers can find relevant literature, also on other topics and research questions dealing with cognitive load and other cognitive variables.

A special focus should be placed on questions with fuzzy constructs, such as motivation, performance, or student engagement (Bond et al. 2020). There is hardly any research in this area to date. In the future, the influence of latent and thus fuzzy constructs on fatigue during screening and coding should be investigated.

Another question that has not been addressed adequately in the literature is how long articles can be screened before concentration decreases and fatigue appears, and to what extent this depends on the research question and the quality of the search string.

References

Akers, Jo, R. Aguiar-Ibáñez, and A. Baba-Akbari. 2009. *Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care*. York, UK: Centre for Reviews and Dissemination, University of York.

Allcoat, Devon, and Adrian von Mühlenen. 2018. "Learning in Virtual Reality: Effects on Performance, Emotion and Engagement". *Research in Learning Technology* 26. https://doi.org/10.25304/rlt.v26.2140.

Alvarez, R. Michael, and Carla VanBeselaere. 2005. "Web-Based Survey". In *Encyclopedia of Social Measurement*. Edited by Kimberly Kempf-Leonard, 955–62. New York: Elsevier. https://doi.org/10.1016/B0-12-369398-5/00390-X.

Andrews, Frank M., and George F. Farris. 1972. "Time Pressure and Performance of Scientists and Engineers: A Five-Year Panel Study". *Organizational Behavior and Human Performance* 8 (2): 185–200. https://doi.org/10.1016/0030-5073(72)90045-1.

Bedenlier, Svenja, Melissa Bond, Katja Buntins, Olaf Zawacki-Richter, and Michael Kerres. 2020. "Learning by Doing? Reflections on Conducting a Systematic Review in the Field of Educational Technology". In *Systematic Reviews in Educational Research: Methodology, Perspectives and Application*, edited by Olaf Zawacki-Richter et al., 111–27. Wiesbaden: Springer Fachmedien.

Bond, Melissa, Katja Buntins, Svenja Bedenlier, Olaf Zawacki-Richter, and Michael Kerres. 2020. "Mapping Research in Student Engagement and Educational Technology in Higher Education: A Systematic Evidence Map". *International Journal of Educational Technology in Higher Education* 17 (1): 2. https://doi.org/10.1186/s41239-019-0176-8.

Bramer, Wichor M., Dean Giustini, Bianca Kramer, and P. F. Anderson. 2013. "The Comparative Recall of Google Scholar Versus PubMed in Identical Searches for Biomedical Systematic Reviews: A Review of Searches Used in Systematic Reviews". *Systematic reviews* 2 (1): 115. https://doi.org/10.1186/2046-4053-2-115.

Bramer, Wichor M., Dean Giustini, and Bianca M. R. Kramer. 2016. "Comparing the Coverage, Recall, and Precision of Searches for 120 Systematic Reviews in Embase, MEDLINE, and Google Scholar: A Prospective Study". *Systematic reviews* 5 (1): 39. https://doi.org/10.1186/s13643-016-0215-7.

Buchner, Josef. 2022. "Systematic Reviews Als Analyseinstrument Der Forschungspraxis in Educational Technology Studien". Presentation at DGfE-Kongress 2022. https://doi.org/10.13140/RG.2.2.19953.35687.

Buchner, Josef, Katja Buntins, and Michael Kerres. 2021. "A Systematic Map of Research Characteristics in Studies on Augmented Reality and Cognitive Load: A Systematic Map of Research Characteristics". *Computers and Education Open* 2: 100036. https://doi.org/10.1016/j.caeo.2021.100036.

Cooke, Alison, Debbie Smith, and Andrew Booth. 2012. "Beyond PICO: The SPIDER Tool for Qualitative Evidence Synthesis". *Qualitative Health Research* 22 (10): 1435–43. https://doi.org/10.1177/1049732312452938.

Cooper, Chris, Rebecca Lovell, Kerryn Husk, Andrew Booth, and Ruth Garside. 2018. "Supplementary Search Methods Were More Effective and Offered Better Value Than Bibliographic Database Searching: A Case Study from Public Health and Environmental Enhancement". *Research synthesis methods* 9 (2): 195–223. https://doi.org/10.1002/jrsm.1286.

Cooper, Chris, Joanna Varley-Campbell, Andrew Booth, Nicky Britten, and Ruth Garside. 2018. "Systematic Review Identifies Six Metrics and One Method for Assessing Literature Search Effectiveness but No Consensus on Appropriate Use". *Journal of Clinical Epidemiology* 99: 53–63. https://doi.org/10.1016/j.jclinepi.2018.02.025.

Eden, Dov. 2002. "From the Editors". *Academy of Management Journal* 45 (5): 841–46. https://doi.org/10.5465/amj.2002.7718946.

Engström, Johan, Gustav Markkula, Trent Victor, and Natasha Merat. 2017. "Effects of Cognitive Load on Driving Performance: The Cognitive Control Hypothesis". *Human Factors* 59 (5): 734–64. https://doi.org/10.1177/0018720817690639.

Evans, David. 2002. "Database Searches for Qualitative Research". *Journal of the Medical Library Association JMLA* 90 (3): 290–93. https://pubmed.ncbi.nlm.nih.gov/12113512.

Eysenbach, G., J. Tuische, and T. L. Diepgen. 2001. "Evaluation of the Usefulness of Internet Searches to Identify Unpublished Clinical Trials for Systematic Reviews". *Medical informatics and the Internet in medicine* 26 (3): 203–18. https://doi.org/10.1080/14639230110075459.

Furlan, Andrea D., Emma Irvin, and Claire Bombardier. 2006. "Limited Search Strategies Were Effective in Finding Relevant Nonrandomized Studies". *Journal of Clinical Epidemiology* 59 (12): 1303–11. https://doi.org/10.1016/j.jclinepi.2006.03.004.

Geersing, Geert-Jan, Walter Bouwmeester, Peter Zuithoff, Rene Spijker, Mariska Leeflang, and Karel Moons. 2012. "Search Filters for Finding Prognostic and Diagnostic Prediction Studies in Medline to Enhance Systematic Reviews". *PloS one* 7 (2): e32844. https://doi.org/10.1371/journal.pone.0032844.

Gehanno, Jean-François, Laetitia Rollin, and Stefan Darmoni. 2013. "Is the Coverage of Google Scholar Enough to Be Used Alone for Systematic Reviews". *BMC Medical Informatics and Decision Making* 13 (1): 7. https://doi.org/10.1186/1472-6947-13-7.

Gough, David, Sandy Oliver, and James Thomas. 2017. *An Introduction to Systematic Reviews*. 2nd revised edition. Los Angeles: SAGE.

Grant, Maria J., and Andrew Booth. 2009. "A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies". *Health information and libraries journal* 26 (2): 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x.

Grossetta Nardini, Holly K., Janene Batten, Melissa C. Funaro, Rolando Garcia-Milian, Kate Nyhan, Judy M. Spak, Lei Wang, and Janis G. Glover. 2019. "Librarians as Methodological Peer Reviewers for Systematic Reviews: Results of an Online Survey". *Research Integrity and Peer Review* 4 (1): 23. https://doi.org/10.1186/s41073-019-0083-5.

Gusenbauer, Michael, and Neal R. Haddaway. 2020. "Which Academic Search Systems Are Suitable for Systematic Reviews or Meta-Analyses? Evaluating Retrieval Qualities of Google Scholar, PubMed, and 26 Other Resources". *Research synthesis methods* 11 (2): 181–217. https://doi.org/10.1002/jrsm.1378.

Haddaway, Neal R., and Helen R. Bayliss. 2015. "Shades of Grey: Two Forms of Grey Literature Important for Reviews in Conservation". *Biological Conservation* 191: 827–29. https://doi.org/10.1016/j.biocon.2015.08.018.

Haddaway, Neal R., Alexandra Mary Collins, Deborah Coughlin, and Stuart Kirk. 2015. "The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching". *PloS one* 10 (9): e0138237. https://doi.org/10.1371/journal.pone.0138237.

Jensen, Lasse, and Flemming Konradsen. 2018. "A Review of the Use of Virtual Reality Head-Mounted Displays in Education and Training". *Education and Information Technologies* 23 (4): 1515–29. https://doi.org/10.1007/s10639-017-9676-0.

Kavanagh, Sam, Andrew Luxton-Reilly, Burkhard Wuensche, and Beryl Plimmer. 2017. "A Systematic Review of Virtual Reality in Education". *Themes in Science and Technology Education* 10 (2): 85–119. https://www.learntechlib.org/p/182115/.

Konig, Cornelius J., Markus Buhner, and Gesine Murling. 2005. "Working Memory, Fluid Intelligence, and Attention Are Predictors of Multitasking Performance, but Polychronicity and Extraversion Are Not". *Human Performance* 18 (3): 243–66. https://doi.org/10.1207/s15327043hup1803_3.

Köstler, Verena. 2023. "Zwischen Präzision und Sensitivität: Generierung eines Studienkorpus am Beispiel einer Fragestellung zu Künstlicher Intelligenz (KI) in Bildungsprozessen". *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung* 54 (Research Syntheses): 1–27. https://doi.org/10.21240/mpaed/54/2023.08.10.X.

Kupper, Lawrence L., and Kerry B. Hafner. 1989. "How Appropriate Are Popular Sample Size Formulas?". *The American Statistician* 43 (2): 101. https://doi.org/10.2307/2684511.

Larsen, Peder Olesen, and Markus von Ins. 2010. "The Rate of Growth in Scientific Publication and the Decline in Coverage Provided by Science Citation Index". *Scientometrics* 84 (3): 575–603. https://doi.org/10.1007/s11192-010-0202-z.

Lasser, Jana, Verena Ahne, Georg Heiler, Peter Klimek, Hannah Metzler, Tobias Reisch, Martin Sprenger, Stefan Thurner, and Johannes Sorger. 2020. "Complexity, transparency and time pressure: practical insights into science communication in times of crisis". *Journal of Science Communication* 19 (5): N01. https://doi.org/10.22323/2.19050801.

Lee, Edwin, Maureen Dobbins, Kara DeCorby, Lyndsey McRae, Daiva Tirilis, and Heather Husson. 2012. "An Optimal Search Filter for Retrieving Systematic Reviews and Meta-Analyses". *BMC medical research methodology* 12 (1): 51. https://doi.org/10.1186/1471-2288-12-51.

Liu, Dejian, Kaushal Kumar Bhagat, Yuan Gao, Ting-Wen Chang, and Ronghuai Huang. 2017. "The Potentials and Trends of Virtual Reality in Education". In *Virtual, Augmented, and Mixed Realities in Education*, 105–30. Springer, Singapore. https://doi.org/10.1007/978-981-10-5490-7_7.

Madigan, Ruth, and Richard Romano. 2020. "Does the Use of a Head Mounted Display Increase the Success of Risk Awareness and Perception Training (RAPT) For Drivers?" *Applied ergonomics* 85: 103076. https://doi.org/10.1016/j.apergo.2020.103076.

Makransky, Guido, and Gustav B. Petersen. 2021. "The Cognitive Affective Model of Immersive Learning (CAMIL): A Theoretical Research-Based Model of Learning in Immersive Virtual Reality". *Educational Psychology Review* 33 (3): 937–58. https://doi.org/10.1007/s10648-020-09586-2.

Martín-Martín, Alberto, Mike Thelwall, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. 2021. "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A Multidisciplinary Comparison of Coverage via Citations". *Scientometrics* 126 (1): 871–906. https://doi.org/10.1007/s11192-020-03690-4.

Merchant, Zahira, Ernest T. Goetz, Lauren Cifuentes, Wendy Keeney-Kennicutt, and Trina J. Davis. 2014. "Effectiveness of Virtual Reality-Based Instruction on Students' Learning Outcomes in K-12 and Higher Education: A Meta-Analysis". *Computers & Education* 70: 29–40. https://doi.org/10.1016/j.compedu.2013.07.033.

Methley, Abigail M., Stephen Campbell, Carolyn Chew-Graham, Rosalind McNally, and Sudeh Cheraghi-Sohi. 2014. "PICO, PICOS and SPIDER: A Comparison Study of Specificity and Sensitivity in Three Search Tools for Qualitative Systematic Reviews". *BMC Health Services Research* 14 (1): 579. https://doi.org/10.1186/s12913-014-0579-0.

Mulders, Miriam, Josef Buchner, and Michael Kerres. 2020. "A Framework for the Use of Immersive Virtual Reality in Learning Environments". *International Journal of Emerging Technologies in Learning (iJET)* 15 (24): 208–24. https://doi.org/10.3991/ijet.v15i24.16615.

Radianti, Jaziar, Tim A. Majchrzak, Jennifer Fromm, and Isabell Wohlgenannt. 2020. "A Systematic Review of Immersive Virtual Reality Applications for Higher Education: Design Elements, Lessons Learned, and Research Agenda". *Computers & Education* 147: 103778. https://doi.org/10.1016/j.compedu.2019.103778.

Rehman, Yasir. 2021. "Difference Between Quantitative and Qualitative Research Question-PICO Vs. SPIDER". *American Academic Scientific Research Journal for Engineering, Technology, and Sciences* 77 (1): 188–99. https://asrjetsjournal.org/index.php/American_Scientific_Journal/article/view/6730.

Robinson, Karen A., Ian J. Saldanha, and Naomi A. Mckoy. 2011. "Development of a Framework to Identify Research Gaps from Systematic Reviews". *Journal of Clinical Epidemiology* 64 (12): 1325–30. https://doi.org/10.1016/j.jclinepi.2011.06.009.

Rogers, Morwenna, Alison Bethel, and Kate Boddy. 2017. "Development and Testing of a Medline Search Filter for Identifying Patient and Public Involvement in Health Research". *Health Information & Libraries Journal* 34 (2): 125–33. https://doi.org/10.1111/hir.12157.

Salvador-Oliván, José Antonio, Gonzalo Marco-Cuenca, and Rosario Arquero-Avilés. 2019. "Errors in Search Strategies Used in Systematic Reviews and Their Effects on Information Retrieval". *Journal of the Medical Library Association JMLA* 107 (2): 210–21. https://doi.org/10.5195/jmla.2019.567.

Sampson, Margaret, and Jessie McGowan. 2006. "Errors in Search Strategies Were Identified by Type and Frequency". *Journal of Clinical Epidemiology* 59 (10): 1057.e1-1057.e9. https://doi.org/10.1016/j.jclinepi.2006.01.007.

Sampson, Margaret, Jennifer Tetzlaff, and Christine Urquhart. 2011. "Precision of Healthcare Systematic Review Searches in a Cross-Sectional Sample". *Research synthesis methods* 2 (2): 119–25. https://doi.org/10.1002/jrsm.42.

Schnotz, Wolfgang, and Christian Kürschner. 2007. "A Reconsideration of Cognitive Load Theory". *Educational Psychology Review* 19 (4): 469–508. https://doi.org/10.1007/s10648-007-9053-4.

Shaffril, Mohamed, Hayrol Azril, Samsul Farid Samsuddin, and Asnarulkhadi Abu Samah. 2021. "The ABC of Systematic Literature Review: The Basic Methodological Guidance for Beginners". *Quality & Quantity* 55 (4): 1319–46. https://doi.org/10.1007/s11135-020-01059-6.

Shepperd, Martin, David Bowes, and Tracy Hall. 2014. "Researcher Bias: The Use of Machine Learning in Software Defect Prediction". *IEEE Transactions on Software Engineering* 40 (6): 603–16. https://doi.org/10.1109/TSE.2014.2322358.

Stock, Wolfgang G., and Mechtild Stock. 2013. *Handbook of Information Science*. Berlin, Boston: De Gruyter Saur. https://doi.org/10.1515/9783110235005.

Straube, S., J. Heinz, P. Landsvogt, and T. Friede. 2021. "Recall, Precision, and Coverage of Literature Searches in Systematic Reviews in Occupational Medicine: An Overview of Cochrane Reviews Recall, Precision Und Coverage Von Literatursuchen in Systematischen Reviews Aus Dem Bereich Arbeitsmedizin: Ein Überblick Über Cochrane Reviews". *GMS Medizinische Informatik, Biometrie und Epidemiologie* 17 (1). https://doi.org/10.3205/mibe000216.

Süß, Heinz-Martin, and Florian Schmiedek. 2000. "Ermüdungs- Und Übungseffekte Bei Mehrstündiger Kognitiver Beanspruchung". *Experimental Psychology* 47 (3): 162–79. https://doi.org/10.1026//0949-3964.47.3.162.

Sweller, John. 1988. "Cognitive Load During Problem Solving: Effects on Learning". *Cognitive Science* 12 (2): 257–85. https://doi.org/10.1016/0364-0213(88)90023-7.

Sweller, John. 2003. "Evolution of Human Cognitive Architecture". *Psychology of learning and motivation* 43: 216–66.

Sweller, John, Jeroen J. G. van Merriënboer, and Fred Paas. 2019. "Cognitive Architecture and Instructional Design: 20 Years Later". *Educational Psychology Review* 31 (2): 261–92. https://doi.org/10.1007/s10648-019-09465-5.

Wu, Bian, Xiaoxue Yu, and Xiaoqing Gu. 2020. "Effectiveness of Immersive Virtual Reality Using Head-mounted Displays on Learning Performance: A Meta-analysis". *British Journal of Educational Technology* 51 (6): 1991–2005. https://doi.org/10.1111/bjet.13023.

Zhang, Fan, Shamila Haddad, Bahareh Nakisa, Mohammad Naim Rastgoo, Christhina Candido, Dian Tjondronegoro, and Richard de Dear. 2017. "The Effects of Higher Temperature Setpoints During Summer on Office Workers' Cognitive Load and Thermal Comfort". *Building and Environment* 123: 176–88. https://doi.org/10.1016/j.buildenv.2017.06.048.