information

Article

Application and Investigation of Multimedia Design Principles in Augmented Reality Learning Environments

Jule M. Krüger * and Daniel Bodemer

Research Methods in Psychology—Media-Based Knowledge Construction, University of Duisburg-Essen, 47057 Duisburg, Germany; bodemer@uni-due.de
* Correspondence: jule.krueger@uni-due.de

Abstract: Digital media have changed the way educational instructions are designed. Learning environments addressing different presentation modes, sensory modalities and realities have evolved, with augmented reality (AR) as one of the latest developments in which multiple aspects of all three dimensions can be united. Multimedia learning principles can generally be applied to AR scenarios that combine physical environments and virtual elements, but their AR-specific effectiveness is unclear so far. In the current paper, we describe two studies examining AR-specific occurrences of two basic multimedia learning principles: (1) the spatial contiguity principle with visual learning material, leveraging AR-specific spatiality potentials, and (2) the coherence principle with audiovisual learning material, leveraging AR-specific contextuality potentials. Both studies use video-based implementations of AR experiences combining textual and pictorial representation modes as well as virtual and physical visuals. We examine the effects of integrated and separated visual presentations of virtual and physical elements (study one, $N = 80$) in addition to the effects of the omission of or the addition of matching or non-matching sounds (study two, $N = 130$) on cognitive load, task load and knowledge. We find only few significant effects and interesting descriptive results. We discuss the results and the implementations based on theory and make suggestions for future research.

Keywords: augmented reality; education; instructional design; multimedia design principles; multimodality; technology-enhanced learning

check for updates

Citation: Krüger, J.M.; Bodemer, D. Application and Investigation of Multimedia Design Principles in Augmented Reality Learning Environments. *Information* **2022**, *13*, 74. https://doi.org/10.3390/info13020074

Academic Editors: Ramon Fabregat, Jorge Bacca-Acosta and N. D. Duque-Mendez

Received: 5 January 2022
Accepted: 2 February 2022
Published: 4 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The possibilities for instructional materials have changed a lot over the past decades due to technological developments in the field of digital media. While pictorial presentations can provide additional information, complementing textual elements in traditional media such as books, the addition of auditory narrations and sounds can provide further sensory input. Richard E. Mayer's cognitive theory of multimedia learning (CTML) is based on the suggestion that humans can use two processing channels when learning, the auditory–verbal and the visual–pictorial channels, which both have capacity limitations and are used in the active processing of information [1]. In more recent years, another dimension was added to the mix: combining virtual elements with physical, real-world elements through the possibilities of augmented reality (AR) technologies (e.g., [2–4]). Instructional materials can thus include different representation modes, such as text and graphics, different sensory modalities, such as visual and auditive, and different realities, such as physical and virtual. Due to these vast possibilities it is important to examine how combinations of representations can be used most effectively and efficiently to support learning. The multimedia design principles that were described by Mayer [5], based on CTML, are one framework that has often been used in this regard. The different principles describe how textual and pictorial representations should be combined in order to best facilitate learning. In the current paper, we take a closer look at how the different features of instructional design can be combined based on multimedia principles in the specific case

of AR-based learning environments and how this might play a role in supporting learning processes and outcomes. We report two studies that implement and investigate two basic multimedia principles in AR scenarios: (1) the spatial contiguity principle, concerning spatially integrated physical and virtual elements, and (2) the coherence principle, with combined visual and auditory representations of contextually integrated physical and virtual elements.

1.1. Multimedia Learning

One of the most influential theories of multimedia-based learning and instruction is the cognitive theory of multimedia learning (CTML) by Mayer [1]. The theory describes cognitive information processing when learning with multi-modal, multi-representational learning material and is based on three assumptions: dual channels, limited capacity and active processing. The dual-channel assumption describes that the learning process includes two paths for processing information through sensory, working and long-term memory [1]. One path describes the auditory–verbal channel and the other the visual–pictorial channel. These channels are based on an integration of two dual-channel memory theories, namely Baddeley's model of working memory, with the distinction between visual sketchpad and phonological loop (e.g., [6]), and Paivio's dual-coding theory, with the distinction between the verbal system and the non-verbal, imagery system (e.g., [7]). The limited-capacity assumption describes that there is a limit to how much information can be processed simultaneously within one channel, such that instructional material should be designed in a way as to not overload the channels [1]. This is closely related to assumptions in cognitive load theory (CLT), which also postulates duration and capacity limitations of working memory [8,9]. The active-processing assumption describes that learning is not a passive process of information absorption but that learners need to actively process and make sense of perceived information [1]. This includes the selection, organization and integration of information, and should be supported by the instructional design. Three kinds of cognitive processing in learning postulated by CTML are extraneous, essential and generative processing. Extraneous processing describes cognitive processing that is not inherent to the learning goal and not directly induced by the content of the learning material but by its design. Essential processing describes the cognitive processing and representation of the learning content in working memory, including the active process of the selection of relevant information. Generative processing describes deeper learning processes, including the organization and integration of the learning content and prior knowledge into a coherent mental representation, which is proposed to be dependent on learners' motivation [1]. These types of processing can be mapped to the three types of cognitive load in CLT: extraneous, intrinsic and germane cognitive load. In the following sections, based on the assumptions of CTML, we will describe potential instructional differences concerning sensory modalities, the application of multimedia principles and the connection to AR characteristics and implementations.

1.2. Sensory Modalities

The sensory modality perspective in CTML distinguishes between material presented visually and material presented auditorily [1]. Based on Baddeley's model of working memory (e.g., [6]) it is described that visual and auditory materials are received through different sensory input channels and processed differently in the visuo-spatial sketchpad and the phonological loop. CTML proposes that the two channels should be leveraged in multimedia instruction, such that the independent limited resources in each channel can be used to process more information in total. The modality principle, for example, describes that learning is increased when spoken words and pictures are combined than when the same words in written form are combined with pictures [10]. Due to the digital nature of most AR-capable systems it is possible to visualize pictorial and textual information, but it is also possible to add auditory sounds and narrations. In the field of AR soundscapes, for example, spatial audio is recorded and reproduced with the goal of designing a naturalistic

and authentic sound experience for AR environments, with virtual sounds adapting to listeners' movements and real-world sounds [11]. Concerning AR experiences, adding this factor of different realities to the factor of modalities leads to four potential information origins: real visual elements, virtual visual elements, real auditory elements and virtual auditory elements. This may provide more complex information, such that a systematic analysis of the learning processes and outcomes under consideration of these different forms of informational elements may be necessary. In the current paper we describe two studies concerning two potential implementations of AR experiences. Concerning modalities, the first application focuses on the combination of real and virtual visual representations without auditory elements and the study's manipulation involves visuo-spatial integration. In the second application, all four kinds of elements are implemented, with the addition of virtual auditory elements being the focus of the study's manipulation.

1.3. Multimedia Design Principles

The representation mode perspective in CTML distinguishes between material presented verbally and material presented non-verbally [1]. Based on Paivio's dual-coding theory (e.g., [7]) it is described that information is processed differently in the verbal system and the non-verbal, imagery system. The general idea behind multimedia learning is that learning from a combination of words and pictures leads to better results than learning from words alone. In order to not overload one or both information-processing channels within their limited capacity, different multimedia principles, which should be considered when designing instructions, were formulated based on many years of empirical research [1]. Two principles for decreasing extraneous cognitive processing and thus increasing resources available for essential and generative processing are the spatial contiguity principle and the coherence principle. The *spatial contiguity principle* describes that corresponding pictures and words in multimedia presentations should be presented in a visuo-spatially integrated way instead of a separated presentation [12]. It is assumed that when material is presented in a separated way, more visual searching is necessary and cognitive resources need to be used to keep the individual elements in working memory before being able to integrate them mentally. This increases extraneous processing, using up resources that are then not available for essential and generative processing. The same idea is described in the split-attention effect in CLT [13]. Many empirical studies have reported a positive effect on learning outcomes when following the principle in instructional design [14]. Specifically in AR, this principle can be followed for combinations of virtual and physical pictorial as well as textual representations, which can be displayed in an integrated way, e.g., through video or optical see-through technology in AR systems. The *coherence principle* describes that extraneous elements that might disturb learning, such as interesting but irrelevant or unnecessarily detailed visual or auditory elements, should be excluded from multimedia presentations [15]. The addition of irrelevant but interesting elements, also called seductive details, can divert learners' attention away from, lead to difficulties in organization within and mislead the integration of the relevant learning content. Adding extraneous material can thus lead to an increase in extraneous processing, depleting cognitive working memory resources that cannot be used for essential and generative processing. Specifically in AR, this extraneous material can include real and virtual elements, which can contain both visual and auditory elements that can be more or less coherent considering the learning goal. In the two potential implementations of AR learning experiences described in the current paper, we examine the spatial contiguity principle with purely visual physical and virtual material in the first study and take a look at the coherence principle with audiovisual real and virtual material in the second study.

1.4. AR Characteristics

AR as a form of combining virtual and physical information has already been implemented in diverse topical areas, suggesting a general significance for formal and informal educationally relevant fields. For example, wearable AR has been used to provide

additional textual information about artists and paintings in art galleries [16], mobile AR has been used to train working memory in elderly people through an interactive serious game [17] and AR as well as virtual reality (VR) materials combined in a mixed reality (MR) application have been used to teach mathematical foundations to architecture students [18]. In various reviews of research on AR in formal and informal educational settings, its positive effects on learning outcomes, motivation, engagement, attitudes and cognitive load in comparison to non-AR implementations has been established (e.g., [2,4,19–22]). Still, more research on the effectivity and efficiency of specific design decisions in educational AR is necessary.

AR-based learning experiences have different attributes or affordances that are enabled through the features of AR technologies. Bower and colleagues (2014), for example, identified the rescaling of virtual objects and overlaying contextually relevant information as key affordances of AR [23]. Wu and colleagues (2013) described 3D-based, situated, ubiquitous and collaborative learning, the senses of immersion and presence, visualization of the invisible and the bridging of informal and formal learning in that respect [20]. Contextuality, interactivity and spatiality are three characteristics of AR experiences identified by Krüger, Buchholz and Bodemer (2019) [3]. The identification of AR-specific characteristics provides researchers and designers with another structure to conceptualize relevant research on and implementations of AR. Two of the characteristics that we will focus on in the current paper are *spatiality*, which includes the potential to place virtual objects in spatial proximity to corresponding physical objects, and *contextuality*, which includes the potential of AR to enable learning supported by virtual elements inside a relevant real-world environment [3]. While this starting point focusing on the technology's and experience's capabilities is a more technology-centered approach to multimedia design, the multimedia principles described in Section 1.2 come from a learner-centered approach that focuses on how multimedia technology can be designed to facilitate cognitive processing [24].

With a technologically enabled experience such as AR, it is important to step out of a technological perspective and connect capabilities with how they can be adapted based on what we know about human cognition. Mystakidis and colleagues executed a systematic mapping review of AR applications in the specific context of STEM learning in higher education and identified five instructional strategies and five instructional techniques often used in AR [25]. They clustered these into a taxonomy including five categories ranging from passive, teacher-centered information presentation to autonomous, student-controlled project work. In other reviews, CTML has been described as a relevant approach in AR-based learning, which should be and already is used as a basis for AR design. Sommerauer and Müller, for example, explicitly suggest on the content layer of their conceptual design framework that CTML should be used in the instructional design of AR applications and that any combination of Mayer's multimedia principles should be implemented [26]. In a review by da Silva et al. CTML is described as one of the most used theories in studies evaluating AR-based educational technology [27], which is supported by the results of a review on pedagogical approaches in AR-based education in which Garzón and colleagues found that CTML is a very popular approach that has been used in various content areas and levels of education [28]. In another systematic review with a focus on cognitive load and performance in AR-based learning research, Buchner and colleagues also describe CTML as an important theory, specifically mentioning the necessity for research examining the effect of either following or violating multimedia principles in AR [29]. Concerning the above-mentioned multimedia principles, the spatially close placement of real and virtual information enabled through spatiality is important for following the spatial contiguity principle. Spatial and temporal contiguity have specifically been mentioned as relevant principles that can be applied through AR [29]. In a study on a tablet-based AR implementation in STEM laboratory courses, the integration of real and virtual information following the spatial contiguity principle led to better acquisition of conceptual knowledge, although cognitive load was not rated differently [30]. When placing information in a relevant context or enriching a situation with contextually relevant information, which

is enabled through contextuality, the above-mentioned coherence principle needs to be considered. Mayer's immersion principle states that highly immersing features, which are also apparent in AR, can be seen as similar to seductive details [31], such that the coherence principle may be very significant for AR environments.

As described above in Section 1.3, implementing multimedia principles in instructional design has an influence on learners' cognitive processing of information. In research on AR in education, cognitive load has been included in various studies, most often using subjective measures such as the NASA TLX questionnaire for data collection but very rarely using measures of the three types of cognitive load described in CTML and CLT [32]. The specification of different types of cognitive load may be valuable in research on the implementation of multimedia principles in AR in particular. In general, the results concerning cognitive load in educational AR environments are inconclusive and both the potential decrease and potential increase in load are postulated by different researchers [29]. It is apparent that the elicitation of cognitive load through the specific design of applications combining virtual and physical elements for learning purposes needs to be further examined.

The following two studies will describe potential implementations of AR experiences using the different above-mentioned variables concerning modalities, multimedia principles and AR characteristics. Study one focusses on an application based on purely visual representations implementing the spatial contiguity principle concerning spatially integrated physical and virtual elements. Study two focusses on an application based on combined visual and auditory representations implementing the coherence principle concerning contextually coherent physical and virtual elements. Because active cognitive processing is at the center of CTML, both studies examine how cognitive processing resources are used, with a focus on cognitive load, task load and the resulting knowledge.

2. Study One: Spatial Contiguity Principle

In the first study, the goal is to examine the spatial integration of virtual and physical elements. The learning material is purely visual, and it is focused on the implementation of the spatial contiguity principle in AR. In AR-specific implementations, the principle can be applied to the spatially integrated visualization of virtual elements and physical elements in the real-world environment. In the study, virtual textual information is integrated into a real pictorial environment. Based on the spatial contiguity principle, we want to examine if this implementation of the principle in particular has a positive influence on cognitive factors, including cognitive load, task load and knowledge.

We hypothesize that complying with the spatial contiguity principle through AR leads to a decrease in extraneous processing and thus in extraneous cognitive load (H1.1a) by reducing visual search processes and decreasing the time that the individual elements need to be held in working memory for mental integration. In turn, the working memory capacities that are made available can be used for generative processing, thus increasing germane cognitive load when material is integrated instead of separated (H1.1b).

Furthermore, we hypothesize that the spatial integration of the learning material has an influence on task load. We expect that through decreasing the necessity of holding individual elements in working memory for a longer time when the visualization is integrated, mental demand is decreased (H1.2a). The decreased necessity for visual search processes leads to fewer necessary eye movements and thus a decrease in physical demand (H1.2b). We furthermore expect that temporal demand is decreased when the presentation is integrated (H1.2c), because fewer search and processing steps need to be taken within the same time. We propose that easier processing of the content with the integrated presentation leads to feelings of higher performance (H1.2d) in addition to lower effort (H1.2e) and frustration (H1.2f).

Through the decrease in extraneous cognitive load and the task-load-related factors as well as the resulting increase in germane cognitive load, we would also expect increased

resulting knowledge when information is spatially integrated (H1.3). All hypotheses of study one are summarized in Table 1.

Table 1. Hypotheses in study one.

Hypotheses in Study One	
H1.1a:	learning with an integrated presentation of real and virtual information leads to lower *extraneous cognitive load* than learning with a separated presentation.
H1.1b:	learning with an integrated presentation of real and virtual information leads to higher *germane cognitive load* than learning with a separated presentation.
H1.2a:	learning with an integrated presentation of real and virtual information leads to lower *mental demand* than learning with a separated presentation.
H1.2b:	learning with an integrated presentation of real and virtual information leads to lower *physical demand* than learning with a separated presentation.
H1.2c:	learning with an integrated presentation of real and virtual information leads to lower *temporal demand* than learning with a separated presentation.
H1.2d:	learning with an integrated presentation of real and virtual information leads to higher perceived *performance* than learning with a separated presentation.
H1.2e:	learning with an integrated presentation of real and virtual information leads to lower *effort* than learning with a separated presentation.
H1.2f:	learning with an integrated presentation of real and virtual information leads to lower *frustration* than learning with a separated presentation.
H1.3:	learning with an integrated presentation of real and virtual information leads to higher resulting *knowledge* than learning with a separated presentation.

2.1. Methods

In a between-subjects design with two conditions, the integration of the visual information was manipulated in a video-based simulation of a location-based informational AR application. One group received an integrated design, which resembled a see-through AR application because the virtual, textual information was placed as an overlay of the relevant pictorial information in a recorded video. The other group received a separated design, in which the information was displayed on a tablet in the video, such that the information was separated from the respective pictorial real information. Dependent variables are cognitive load, task load and the resulting knowledge.

2.1.1. Participants

The participants were reached through online platforms for participant sampling of the department and convenience sampling. Students could receive participant hours for taking part. The final dataset included $N = 80$ people after one outlier was filtered out based on high age. Primarily (95%) undergraduate students took part, of which most were in the study programs of applied cognitive and media science (84%) and psychology (14%), in which there are no classes related to the learning topic of the study. They were aged 17 to 33 ($M = 22.21$, $SD = 3.14$) and 20 indicated being male, 60 being female. On average, the participants did not indicate high prior knowledge beliefs concerning the focal learning topic of plants in a subjective rating ($M = 1.91$, $SD = 0.73$; 5-point response format from 1, low to 5, high). The participants on average indicated having rarely used general AR applications ($M = 2.23$, $SD = 0.98$) and AR learning applications ($M = 1.69$, $SD = 0.96$; both measured in 5-point response format, with 1—"never", 2—"rarely", 3—"now and then", 4—"often" and 5—"regularly"). They were randomly distributed into the two groups. In Table 2, the number of participants, gender, age, prior knowledge beliefs and prior usage of AR applications per condition are shown. The distribution is quite balanced for all variables. All subjects gave their informed consent for inclusion before they participated in

the study. The study with the ID psychmeth_2020_AR13_29 was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the department's Ethics Committee (vote ID: 2011PFBS7216).

Table 2. Distribution of number of participants, gender, age, prior knowledge beliefs and prior usage of AR applications split by condition in study one.

Condition	n	Gender		Age	Prior Knowledge	Usage AR Applications	
						General	Learning
		Male	Female	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Integrated	39	8	31	21.72 (2.76)	1.93 (0.60)	2.31 (0.92)	1.82 (1.00)
Separated	41	12	29	22.68 (3.43)	1.89 (0.85)	2.15 (1.04)	1.56 (0.92)

2.1.2. Materials

The independent variable was manipulated by showing simulated AR experiences to the participants through two different videos during the learning phase. In both videos, plants in a botanical garden were filmed. Additionally, in both videos, textual information about different plants were added: their common and scientific name, their height, the color of their blossoms and where they usually grow. The simulated AR experience in this study thus included the video of the botanical garden in the background as the real-world environment in which the participants should imagine themselves to be in. The additional textual information about the different plants were included as the virtual elements in the AR experience which the participants should imagine viewing while walking through the garden. In the video in the integrated visualization condition, information was placed as an overlay directly in front of the plants, resembling a visualization that may be possible with an AR glasses or other see-through version of an application (left picture in Figure 1). In the video in the separated visualization condition a tablet was held by the person filming the video and the textual information was displayed on that tablet, such that it was spatially separated from the real plants (right picture in Figure 1). In both conditions, the participants were asked to imagine that they were walking around in the botanical garden and using the application in the real world themselves. The videos were around three minutes long.



Figure 1. Screenshots from the videos used in study one: integrated (AR) visualization on the **left**; separated (non-AR) visualization on the **right**.

To measure subjective prior knowledge of the participants for the sample description, three questions from the ability belief subscale of the expectancy–value questionnaire by Wigfield and Eccles [33] were used in a reframed version, inquiring knowledge beliefs with translated items adapted to the content area. The 5-point response format from 1 (low) to 5 (high) had different anchor phrases. Cronbach's alpha was good for this scale ($\alpha = 0.82$).

To measure cognitive load the questionnaire by Klepsch et al. [34] was used, measuring both extraneous and germane cognitive load with three items each. The 7-point response format ranged from "not at all true" (1, low) to "completely true" (7, high), and means were

calculated for each subscale. Cronbach's alpha was acceptable for the germane cognitive load subscale (α = 0.78) but questionable for the extraneous cognitive load (α = 0.61) subscale, which we nonetheless kept as it was because of the already low number of three items.

Task load was measured through a translated, German version of the NASA TLX by Hart and Staveland [35,36] including six one-item scales for the six variables mental demand, physical demand, temporal demand, performance, effort and frustration. Most items were answered in a 21-point response format from "low" (1) to "high" (21). The performance item was measured from "good" (1) to "bad" (21), but we inverted the scores for the analysis. This provides an easier interpretability of the scores so that in the reported results high scores mean high perceived performance and low scores mean low perceived performance. Only the scores on the individual subscales were used for the analyses, no summarized version for general task load.

Learning outcomes were measured through a knowledge test that included different kinds of questions. In total, ten questions were administered in the form of multiple-choice questions with five possible answers (one correct answer). In five of these questions the participants had to match textual information about a plant to the picture of the plant (e.g., "Which family does this plant belong to?") and in five other questions they had to match textual information to the name of the plant (e.g., "Where does the ox-eye grow?"). One point was given for a correct answer. In addition to the multiple-choice questions, at the end a picture of each of the five plants had to be matched to its respective name, which could earn one point per correct answer. In total, 15 points could be achieved.

2.1.3. Procedure

The study took place during social distancing measures due to the COVID-19 pandemic in December 2020; therefore, it was fully online with a researcher supervising each participant through synchronous (voice/video) chat. After the researcher welcomed the participants, they read the conditions and were asked for their consent. They answered the questions concerning their prior knowledge on plants. After that, the participants viewed the video showing either the integrated AR or the separated tablet-view of real plants in the botanical garden and virtual textual information. The participants were asked to imagine that they were in a real-world situation using the respective application that was displayed in the video. Afterwards, the cognitive load questionnaire and NASA TLX were administered, followed by the knowledge test. In the end, demographic data were requested, the participants were debriefed and the session was completed by the researcher. This procedure can also be seen in Figure 2.

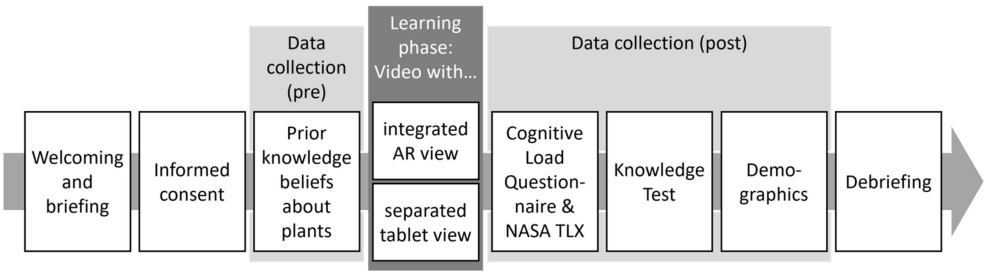


Figure 2. Procedure of study one.

2.2. Results

All hypotheses were statistically tested through independent one-sided *t*-tests, with the integration of material (integrated vs. separated) as the grouping variable and the appropriate score as the outcome variable. As suggested by Delacre and colleagues [37] we used Welch's *t*-test for all analyses, although Levene's test indicated the homogeneity of variances for all variables except physical demand (see Appendix A). The means and standard deviations of all variables can be seen in Table 3.

Table 3. Means and standard deviations of different variables in study one.

Mean and SD per Condition [a]	Possible Range	Integrated *M* (*SD*)	Separated *M* (*SD*)
H1.1a: extraneous CL	1–7	2.57 (1.19)	**2.90 (1.13)**
H1.1b: germane CL	1–7	**4.91 (1.29)**	4.46 (1.44)
H1.2a: mental demand	1–21	9.62 (4.42)	**10.51 (4.88)**
H1.2b: physical demand	1–21	1.82 (1.50)	**2.68 (3.38)**
H1.2c: temporal demand	1–21	6.90 (4.85)	**8.85 (5.44)**
H1.2d: performance	1–21	**13.31 (4.75)**	10.68 (4.99)
H1.2e: effort	1–21	8.77 (4.49)	**10.24 (5.05)**
H1.2f: frustration	1–21	5.97 (4.81)	**6.95 (5.79)**
H1.3: knowledge	0–15	**10.51 (2.81)**	10.49 (3.53)

[a.] Highest mean per subscale in bold.

2.2.1. H1.1: Cognitive Load

To test hypotheses H1.1a and H1.1b, concerning the influence of the integration of the material on extraneous cognitive load, the *t*-tests included the extraneous cognitive load and germane cognitive load subscale scores as outcome variables. H1.1a, concerning extraneous cognitive load, was tested with a one-sided *t*-test proposing lower scores for the integrated than the separated condition, while H1.1b, concerning germane cognitive load, was conversely tested proposing higher scores for the integrated than the separated condition. Boxplots showing the data for both types of cognitive load can be seen in Figure 3.

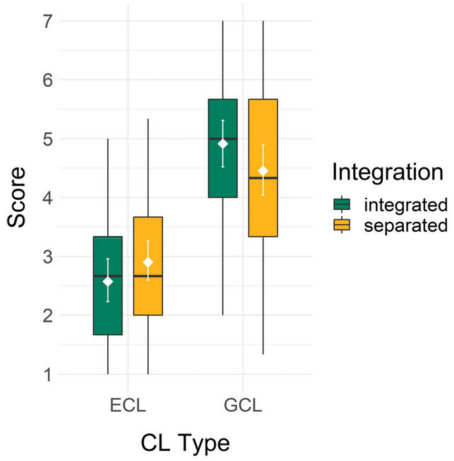


Figure 3. Distribution of extraneous cognitive load and germane cognitive load scores split by group in study one (boxplot with IQR (black), mean with bootstrapped 95% confidence interval (white)).

For H1.1a, concerning the lowering influence of the integration of the material on extraneous cognitive load, the score was indeed descriptively lower for the integrated ($M = 2.57$, $SD = 1.19$) than the separated ($M = 2.90$, $SD = 1.13$) condition. However, this effect with a small effect size was not significant, $t(77.14) = -1.27$, $p = 0.104$ and $d = -0.28$.

For H1.1b, concerning the positive influence of the integration of the material on germane cognitive load, a descriptively higher score was indeed found in the integrated ($M = 4.91$, $SD = 1.29$) than the separated ($M = 4.46$, $SD = 1.44$) condition. Again, this effect with a small effect size was not significant, $t(77.72) = 1.50$, $p = 0.069$ and $d = 0.34$.

Although descriptively the data match our expectations with small effects, H1.1a and H1.1b were not completely supported due to the non-significance of the effects: no significant advantages of the integration of the material concerning extraneous and germane cognitive load were found.

2.2.2. H1.2: Task Load

In order to test hypotheses H1.2a, H1.2b, H1.2c, H1.2d, H1.2e and H1.2f, concerning the influence of the integration of the material on task load, the *t*-tests included mental demand, physical demand, temporal demand, performance, effort and frustration subscale scores as outcome variables. Most hypotheses were tested with one-sided *t*-tests proposing lower scores for the integrated than the separated condition, except for H1.2d, concerning performance, which was tested proposing higher scores for the integrated than the separated condition. Boxplots showing the data for all six subconstructs of task load can be seen in Figure 4.

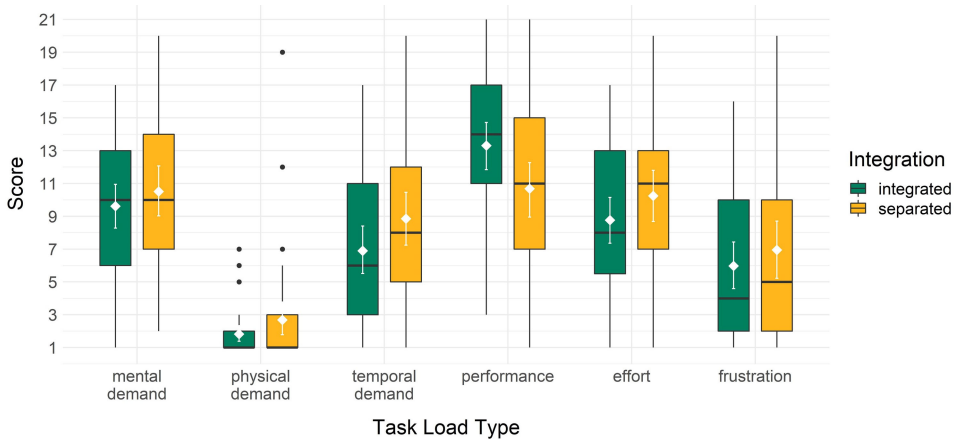


Figure 4. Distribution of task load subscale scores split by group in study one (boxplot with IQR (black), mean with bootstrapped 95% confidence interval (white)).

For H1.2a, concerning the lowering influence of the integration of the material on mental demand, the score was indeed descriptively lower for the integrated ($M = 9.62$, $SD = 4.42$) than the separated ($M = 10.51$, $SD = 4.88$) group, although this difference was not significant, $t(77.81) = -0.86$, $p = 0.195$ and $d = -0.19$. For H1.2b, concerning the lowering influence of the integration of the material on physical demand, the score was indeed descriptively lower for the integrated ($M = 1.82$, $SD = 1.50$) than the separated ($M = 2.68$, $SD = 3.38$) group. Although Cohen's d shows a small effect size, no significant difference was found between the two groups, $t(55.81) = -1.49$, $p = 0.071$ and $d = -0.33$. For H1.2c, concerning the lowering influence of the integration of the material on temporal demand, the score was indeed descriptively lower for the integrated ($M = 6.90$, $SD = 4.85$) than the separated ($M = 8.85$, $SD = 5.44$) group. This difference was significant with a small effect size, $t(77.68) = -1.70$, $p = 0.047$ and $d = -0.38$.

For H1.2d, concerning the positive influence of the integration of the material on perceived performance, the score was indeed descriptively higher for the integrated ($M = 13.31$, $SD = 4.75$) than the separated ($M = 10.68$, $SD = 4.99$) group. This difference was significant with a medium effect size, $t(78) = 2.41$, $p = 0.009$ and $d = 0.54$. For H1.2e, concerning the lowering influence of the integration of the material on effort, the score was indeed descriptively lower for the integrated ($M = 8.77$, $SD = 4.49$) than the separated ($M = 10.24$, $SD = 5.05$) group. No significant difference was found between the two groups, although a small effect size is apparent, $t(77.64) = -1.38$, $p = 0.086$ and $d = -0.31$. For H1.2f, concerning the lowering influence of the integration of the material on frustration, the score was indeed descriptively lower for the integrated ($M = 5.97$, $SD = 4.81$) than the separated ($M = 6.95$, $SD = 5.79$) group. No significant difference was found between the two groups, $t(76.66) = -0.82$, $p = 0.207$ and $d = -0.18$.

We thus only found a significant difference concerning temporal demand and performance in the expected directions, supporting H1.2c and H1.2d: temporal demand was perceived as lower while performance was perceived as higher when the material was integrated instead of separated. For the other variables no significant effects were found,

although the directions of the mean differences were as expected, with at least small effects for physical demand (H1.2b) and effort (H1.2e). Not even small effects were found for mental demand (H1.2a) and frustration (H1.2f).

2.2.3. H1.3: Knowledge

To test H1.3 on the positive influence of the integration of the material on knowledge, a one-sided *t*-test proposing higher scores for the integrated condition included knowledge test score as the outcome variable. Descriptively, the knowledge test scores in the integrated (M = 10.51, SD = 2.81) and separated (M = 10.49, SD = 3.53) groups barely differed and no significant difference was found, $t(75.68) = 0.04$, $p = 0.486$ and $d = 0.01$. Hypothesis H1.3 was thus not supported: no significant advantages of the integration of the material concerning knowledge were found.

As different kinds of items were used in the knowledge test, we also took a closer exploratory look at these. Concerning the six items with ten possible points, in which the picture of a plant had to be matched with a textual characteristic or name, the pattern was the same as in the complete knowledge test results: higher score for the integrated (M = 6.90, SD = 1.90) compared to the separated (M = 6.80, SD = 2.51) visualization. This difference was descriptively only minimally bigger than the general knowledge difference and not significant, $t(74.34) = 0.19$, $p = 0.426$ and $d = 0.04$. Concerning the five items in which only text where included, thus matching the name of a plant with a characteristic, this pattern was the opposite. The separated condition had a higher score (M = 3.68, SD = 1.40) than the integrated (M = 3.62, SD = 1.23) condition. This difference was also very small and not significant, $t(77.46) = -0.23$, $p = 0.819$ and $d = -0.05$ (two-sided *t*-test due to the descriptive direction of difference being opposite than expected for knowledge in general). Although these differences between the groups are very small, the items in which pictures and text needed to be matched thus descriptively showed an advantage for the integrated presentation, while the items in which different textual elements needed to be matched descriptively showed an advantage for the separated conditions. Boxplots showing the data split by item type and for the complete knowledge test can also be see in Figure 5.

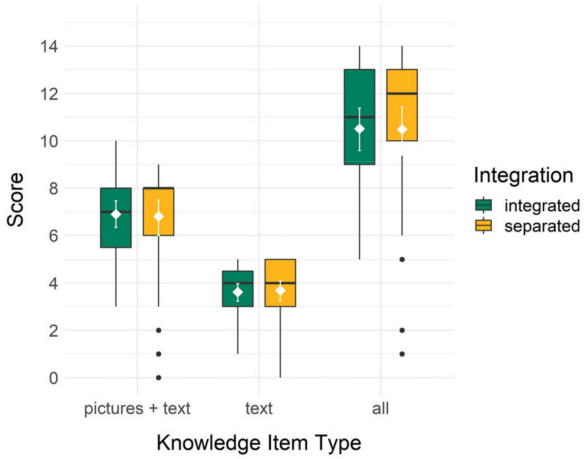


Figure 5. Distribution of knowledge test scores by item types split by group in study one (boxplot with IQR (black), mean with bootstrapped 95% confidence interval (white)).

2.3. Discussion

The goal of study one was to implement and examine the spatial contiguity principle with a combination of visual real and virtual elements, leveraging the AR characteristic spatiality. We examined the influence of the integration of the material on cognitive load, task load and knowledge. Only two of the hypotheses were supported by the data, namely H1.2c, concerning the decrease in temporal demand, and H1.2d, concerning the increase in perceived performance through the integrated presentation of the real and virtual elements.

All other hypotheses were not fully supported, although the tendencies in the data showed expected differences. Descriptively, the cognitive load scores were as expected, with small effects describing decreased extraneous cognitive load (H1.1a) and increased germane cognitive load (H1.1b) in the integrated presentation. Further non-significant effects with small effect sizes were found for the task load subconstructs physical demand (H1.2b) and effort (H1.2e), and even smaller, non-significant differences for mental demand (H1.2a) and frustration (H1.2f), all showing a decrease in the integrated presentation. Additionally, concerning H1.3, no effects on resulting knowledge were found, although interestingly different patterns were revealed for knowledge items in which pictures and text had to be combined and in which only text was included, with descriptively better results for the integrated presentation in the former and better results for the separated presentation in the latter. These results suggest that the spatial integration of pictorial real-world elements and textual virtual elements may have slightly strengthened the building of connections between the pictures and texts, while it may have even more slightly weakened the building of connections within the added texts. This differentiation between representational connections is in accordance with Seufert [38], who described connections within a representation as intra-representational and connections between representations as inter-representational coherence formation, and with Seufert and Brünken [39], who described these concepts as local and global coherence formation. They argue that for deeper understanding both forms of relating need to be achieved [38], and that learners can be supported in their coherence formation on a surface feature level or a deep structure level [39]. Following the spatial contiguity principle may be a form of support on the surface feature level, as the goal is to directly show learners which information belongs together. In future studies it may be interesting to also look at guidance that supports coherence formation in AR-based multimedia representations on a deep structure level, potentially extending the results of the current study.

Different factors may have led to these results showing mainly small, non-significant effects, although all with tendencies in the expected direction. The elements in the presentation could be understood independent from each other and did not have to be mentally integrated to learn the content, such that this pre-requisite for the spatial contiguity principle [12] and the split-attention effect [13] was not given. However, considering the learning goal of integrating the real-world elements and virtual elements, their combination was at least necessary for the knowledge test items in which pictures and textual information needed to be matched. Descriptively, the difference between the pattern in these items and the purely textual items supports this idea, but the simple connection between one picture and few textual characteristics for each plant may not be enough to lead to a strong effect.

Another factor leading to only small effects may have been that the learning situation was not demanding enough for the spatial contiguity principle to show effects on subjective cognitive load and task load. A boundary condition for the spatial contiguity principle is that the material needs to be complex for a strong effect to occur [12], and the split-attention effect is said to mainly appear in materials with higher element interactivity [13]. The learning materials concerning the different plants in a botanical garden were not very complex, including only real-world pictures and a few characteristics of said plants. More complex materials with higher element interactivity may lead to larger effects concerning the two types of cognitive load and the subconstructs of task load. This should be tested in future studies with a focus on material in which virtual textual elements need to be mentally combined into a coherent representation with the physical objects in order to understand their structure and function.

The simplicity of the materials due to the video-based implementation of the AR environment may also have been the reason for the missing effects. The video-based implementation provided different opportunities, including the decrease in individual differences that may usually be caused by non-standardized interaction with AR-based learning applications. This can, for example, include the duration that learners spend engaging with specific components or the exact view of the spatially integrated or separated

materials, which could be standardized through the videos. This way, the focus was on the implementation of the spatial contiguity principle decreasing potentially confounding variables and increasing comparability between participants. A limitation is that the usage of the video-based implementation also removed some factors of complexity that real, location-based AR experiences have. In a real-world AR experience, more cognitive load and potentially overload might have been evoked through a combination of all impressions and necessary skills [40], requiring a decrease in cognitive load through the instructional design. The goal of the study was to examine the specific case of the spatial integration of real and virtual elements, which may have been negatively influenced by the video implementation. Participants may not have perceived the real-world plants in the video as real because they saw them on a screen compared to seeing them physically in front of them. Although it was a video recording of the real plants and the participants were asked to imagine seeing the plants in the real world, it is not sure if they achieved this or if the real elements were perceived as virtual. It is thus not clear how well the results from this study can be transferred onto the usage of real AR applications. A future step would be to implement the spatial contiguity principle in a field study in a location-based AR environment, including a more authentic, naturalistic setting. The advantages and limitations of the video-based implementations in both studies will be discussed more broadly in the general discussion.

Other limitations of the study were that, although a complete standardization of the experiences would have been possible through the video-based implementation, there were some differences between the two videos that were not controlled. Because the videos were filmed on different days, it was sunny in the AR-view video and cloudy in the tablet-view video. Furthermore, the individual plants and information were shown for around 25 s in the tablet-view video and around 30 s in the AR-view video. While there was only little textual information that all participants should have been able to read in 25 s, the time difference and the weather may have had a systematic influence on the results, which should be taken into account.

In conclusion, study one showed (descriptive) results suggesting a potential decrease in extraneous cognitive load as well as task load and a simultaneous increase in germane cognitive load and knowledge concerning the linking of real and virtual elements when learning with an integrated instead of a separated presentation. This suggests that the spatial contiguity principle could be transferable onto AR environments, although the results need to be supported in additional research with more complex material and real AR applications.

3. Study Two: Coherence Principle

In the second study the goal is to examine the contextual coherence of virtual information. The learning material is audiovisual, and it is focused on the implementation of the coherence principle in AR. In AR-specific implementations the principle can be applied to real and virtual, visual and auditory elements. In the study, virtual sounds that either match or do not match the topic of the learning material are added into an application that also includes virtual texts and pictures in addition to real environmental sounds. These sounds are no direct part of the learning task and are compared to the omission of virtual auditory elements. Based on the coherence principle, we want to examine if the implementation of the principle in particular has a positive influence on cognitive factors, including cognitive load, task load and knowledge.

We hypothesize that complying with the coherence principle in AR and thus not adding any sounds leads to a decrease in extraneous processing and thus in extraneous cognitive load (H2.1a) by reducing the number of elements that have to be processed. In turn, the working memory capacities that are made available can be used for generative processing, thus increasing germane cognitive load when no sounds are added, although we expect matching sounds to also increase germane cognitive load within the limits due to motivational effects, which non-matching sounds do not elicit (H2.1b).

Furthermore, we hypothesize that following the coherence principle has an influence on task load. We expect that through the necessity to attend to less sensory input and the explicit usage of fewer sensory organs, both mental (H2.2a) and physical demand (H2.2b) are decreased when no sounds are added. We also expect temporal demand to be decreased with the omission of additional sounds (H2.2c), because less sensory input can be processed within the same time. We propose that the decreased potential for distraction through additional sensory input when no additional sounds are presented leads to feelings of higher performance (H2.2d) and lower effort (H2.2e). Frustration is on one hand expected to be smaller when no distracting sounds are added at all, but on the other hand is expected to be smaller for the addition of matching sounds compared to non-matching sounds because the reason for adding these is not apparent and might lead to even higher frustration (H2.2f).

Through the decrease in extraneous cognitive load as well as the task-load related factors, and the resulting increase in germane cognitive load, we would also expect increased resulting knowledge when no sounds are added, although through motivating effects and decreased frustration we would further expect that matching sounds lead to higher resulting knowledge than non-matching sounds (H2.3). All hypotheses of study two are summarized in Table 4.

Table 4. Hypotheses in study two.

Hypotheses in Study Two	
H2.1a:	learning with material including real and virtual information without additional virtual sounds leads to lower *extraneous cognitive load* than when virtual sounds are added.
H2.1b:	learning with material including real and virtual information without additional virtual sounds leads to higher *germane cognitive load* than when virtual sounds are added, where adding matching sounds leads to higher *germane cognitive load* than adding non-matching sounds.
H2.2a:	learning with material including real and virtual information without additional virtual sounds leads to lower *mental demand* than when virtual sounds are added.
H2.2b:	learning with material including real and virtual information without additional virtual sounds leads to lower *physical demand* than when virtual sounds are added.
H2.2c:	learning with material including real and virtual information without additional virtual sounds leads to lower *temporal demand* than when virtual sounds are added.
H2.2d:	learning with material including real and virtual information without additional virtual sounds leads to higher perceived *performance* than when virtual sounds are added.
H2.2e:	learning with material including real and virtual information without additional virtual sounds leads to lower *effort* than when virtual sounds are added.
H2.2f:	learning with material including real and virtual information without additional virtual sounds leads to lower *frustration* than when virtual sounds are added, where adding matching sounds leads to lower *frustration* than adding non-matching sounds.
H2.3:	learning with material including real and virtual information without additional virtual sounds leads to higher *knowledge* than when virtual sounds are added, where adding matching sounds leads to higher *knowledge* than adding non-matching sounds.

3.1. Methods

In a between-subjects design with three conditions, the addition or omission of matching or non-matching virtual sounds was manipulated in a video-based simulation of a location-based informational AR application. One group heard no additional sounds, one heard sounds matching the learning topic and one heard sounds that did not match the topic. Dependent variables are cognitive load, task load and knowledge.

3.1.1. Participants

The participants were reached through the same online platforms for participant sampling of the department as in study one and convenience sampling. Students could receive participant hours for taking part. The final dataset included N = 130 people after two outliers were filtered out based on very long study duration. Primarily (86%) students took part, of which most were in the study programs of applied cognitive and media science (76%) and psychology (17%), in which there are no classes related to the learning topic of the study. They were aged 18 to 61 (M = 23.72, SD = 7.95) and 34 indicated being male, 96 being female. On average, the participants did not indicate high prior knowledge beliefs concerning the focal learning topic of regional birds in a subjective rating (M = 1.59, SD = 0.50; 5-point response format from 1—low to 5—high). The participants on average indicated having very rarely used general AR applications (M = 1.81, SD = 0.81) and AR learning applications (M = 1.31, SD = 0.68; both measured in a 5-point response format, with 1—"never", 2—"rarely", 3—"now and then", 4—"often" and 5—"regularly"). The participants were randomly distributed into the three groups. In Table 5 the number of participants, gender, age, prior knowledge beliefs and prior usage of AR applications per condition are shown. The distribution is quite balanced for all variables except for age, with a descriptively higher age in the matching group. All subjects gave their informed consent for inclusion before they participated in the study. The study with the ID psychmeth_2020_AR14_30 was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the department's Ethics Committee (vote ID: 2012PFKL8474).

Table 5. Distribution of the number of participants, gender, age, prior knowledge beliefs and prior usage of AR applications split by condition in study two.

Condition	*n*	Gender		Age	Prior Knowledge	Usage AR Applications	
						General	Learning
		Male	Female	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
No sounds	43	11	32	22.30 (6.13)	1.54 (0.42)	1.84 (0.72)	1.35 (0.81)
Matching	44	13	31	26.52 (11.10)	1.61 (0.45)	1.73 (0.79)	1.36 (0.72)
Non-matching	43	10	33	22.28 (4.23)	1.61 (0.62)	1.86 (0.91)	1.21 (0.47)

3.1.2. Materials

The independent variable was manipulated by showing simulated AR experiences to the participants through three different videos during the learning phase. The visual material and real environmental sounds were the same for all videos. The scene was filmed in a forest, in which six different locations were walked towards and focused on with the camera, where a picture of and additional textual information (common name, scientific name, size and food) about a different bird then appeared in each location. The simulated AR experience in this study thus included the video of the forest including sounds in the background as the real-world environment which the participants should imagine themselves to be in. The additional pictorial (i.e., the bird) and textual (i.e., bird characteristics) information was included as virtual elements in the AR experience which the participants should imagine as being viewed through AR glasses. The videos differed in the additional virtual sounds that were played when focusing on a bird: either no additional sound (no sounds), the chirping of that bird (matching sounds) or another unrelated sound, such as a bell (non-matching sounds). In Figure 6, screenshots of two frames including birds from the videos are shown. In all conditions, the participants were asked to imagine that they were walking through the forest and using the AR application in the real world themselves. The videos took 6 min.



Figure 6. Two screenshots from the video used in study two.

Subjective prior knowledge was measured with the same scale by Wigfield and Eccles [33] used in study one, although the content area was adapted. Cronbach's alpha was questionable ($\alpha = 0.58$), but the scale was kept the same because it only has three items.

Extraneous and germane cognitive load were measured as in study one with the cognitive load scale by Klepsch and colleagues [34]. Cronbach's alpha was questionable for extraneous cognitive load at $\alpha = 0.59$ and germane cognitive load at $\alpha = 0.62$, but we kept the scales as they were because of the already low number of three items per scale.

Task load including mental demand, physical demand, temporal demand, performance, effort and frustration was also measured with the six one-item scales from the NASA TLX used in study one [35,36]. Again, the score for the performance item was inverted for easier interpretation of the scores, with high scores meaning high perceived performance and low scores meaning low perceived performance.

Learning outcomes were measured through a knowledge test (8 items) that included different kinds of questions, all multiple-choice questions with four possible answers (one correct answer). In four questions the participants had to match textual information about a bird to the picture of the bird (e.g., "What is the name of this bird?"), in four other questions they had to match textual information to the name of the bird (e.g., "What does the bullfinch eat?"). One point was given for a correct answer, so that in total 8 points could be reached.

For exploratory analyses, a sound–picture matching test was also administered in the two groups that received additional sounds in the learning test. In two items the participants were asked to match a sound to the bird's name and in two other items to the bird's picture.

3.1.3. Procedure

As in study one, the study took place during social distancing measures due to the COVID-19 pandemic in December 2020, such that it was fully online with a researcher supervising each participant through synchronous (voice/video) chat. The procedure is very similar to that of study one. After the researcher welcomed the participants, they read the conditions and were asked for their consent. They answered the questions concerning their prior knowledge on regional birds. After that, the participants viewed the video showing the virtual birds and textual information in the real forest and adding either no, matching or non-matching sounds. The participants were asked to imagine that they were in the real-world situation using the respective application that was displayed in the video. Afterwards, the cognitive load questionnaire and NASA TLX were administered, followed by the knowledge test. In the end, demographic data were requested, the participants were debriefed, and the session was completed by the researcher. This procedure can also be seen in Figure 7.

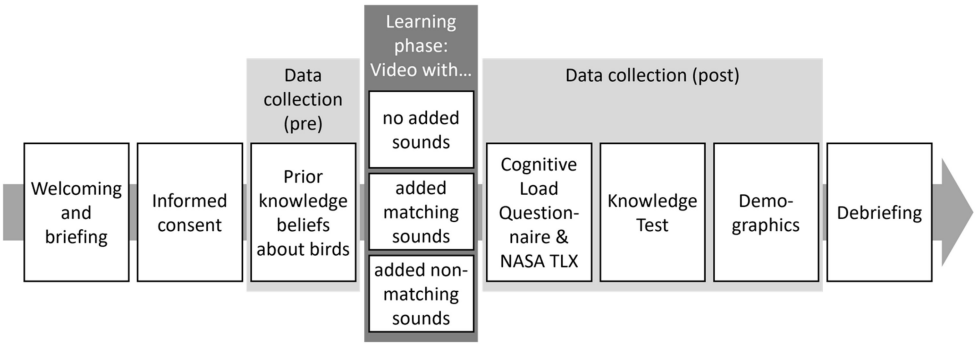


Figure 7. Procedure in study two.

3.2. Results

All hypotheses were statistically tested through one-way ANOVAs with type of sounds (no vs. matching vs. non-matching) as the predictor and the appropriate score as the outcome variable. Levene's test indicated the homogeneity of variances for all variables (see Appendix A). A-priori-determined contrasts based on the individual hypotheses are used for the analyses. The first comparison in the contrast analysis always focuses on no added sounds compared to added sounds, while the second comparison focuses on matching sounds compared to non-matching sounds. When the focus of the hypothesis is on the first comparison in the contrast, the second comparison is explored for a more complete picture. Means and standard deviations of all variables can be seen in Table 6.

Table 6. Means and standard deviations of different variables in study two.

Mean and SD per Condition [a]	Possible Range	No Sound *M* (*SD*)	Matching *M* (*SD*)	Non-Matching *M* (*SD*)
H2.1a: extraneous CL	1–7	*2.24 (1.08)*	**2.55 (1.00)**	2.53 (1.16)
H2.1b: germane CL	1–7	4.81 (1.41)	4.76 (1.26)	**5.12 (1.22)**
H2.2a: mental demand	1–21	*8.60 (4.47)*	**9.98 (4.71)**	9.74 (4.33)
H2.2b: physical demand	1–21	**4.12 (4.23)**	3.93 (3.39)	*3.40 (3.58)*
H2.2c: temporal demand	1–21	*6.84 (4.57)*	**7.64 (4.92)**	7.47 (4.67)
H2.2d: performance	1–21	11.95 (4.89)	*10.95 (5.25)*	**13.00 (4.72)**
H2.2e: effort	1–21	*7.95 (4.54)*	**9.70 (4.52)**	8.21 (4.38)
H2.2f: frustration	1–21	5.65 (4.83)	**6.11 (4.94)**	*5.00 (4.89)*
H2.3: knowledge	0–8	*3.70 (1.74)*	3.86 (1.77)	**4.02 (1.57)**

[a.] Highest mean per subscale in bold, lowest mean per subscale in italics.

3.2.1. H2.1: Cognitive Load

To test hypotheses H2.1a and H2.1b on the influence of added sounds on cognitive load, the one-way ANOVAs included extraneous cognitive load and germane cognitive load subscale scores as outcome variables. Boxplots showing the data for both types of cognitive load can be seen in Figure 8.

In H2.1a, extraneous cognitive load was hypothesized to be lower when no sounds were added than when either matching or non-matching sounds were added. This pattern was indicated descriptively in the group means, showing a lower extraneous cognitive load in the no sounds (M = 2.24, SD = 1.08) than the very similar matching sounds (M = 2.55, SD = 1.00) and non-matching sounds (M = 2.53, SD = 1.16) conditions. The overall model of the one-way ANOVA was not significant, $F(2, 127) = 1.14$, $p = 0.324$ and $\omega^2 < 0.01$. Additionally, no significant differences were found in the first comparison of the Helmert contrast analysis, comparing no sounds and added sounds, $t = -1.51$, $p = 0.135$, or in the second comparison, comparing matching and non-matching sounds, $t = 0.08$, $p = 0.938$.

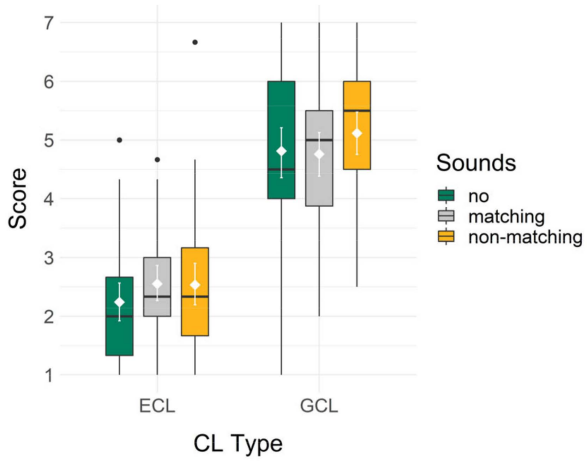


Figure 8. Distribution of extraneous cognitive load and germane cognitive load scores split by group in study two (boxplot with IQR (black), mean with bootstrapped 95% confidence interval (white)).

In H2.1b, germane cognitive load was hypothesized to be higher when no sounds were added than when either matching or non-matching sounds were added, and higher when matching than when non-matching sounds were added. Descriptively, a different pattern was shown with the highest germane cognitive load in the non-matching (M = 5.12, SD = 1.22) condition, then the no sounds (M = 4.81, SD = 1.41) and then the matching sounds (M = 4.76, SD = 1.26) conditions. The overall model of the one-way ANOVA was not significant, $F(2, 127) = 0.94$, $p = 0.393$ and $\omega^2 < 0.00$. Additionally, no significant differences were found in the first comparison of the Helmert contrast analysis, $t = -0.52$, $p = 0.607$, or in the second comparison, $t = -1.27$, $p = 0.205$.

We thus found no support for H2.1a and H2.1b: no significant advantages of leaving out additional sounds were found concerning extraneous or germane cognitive load.

3.2.2. H2.2: Task Load

For testing H2.2a, H2.2b, H2.2c, H2.2d, H2.2e and H2.2f on the influence of added sounds on cognitive load, the one-way ANOVAs included mental demand, physical demand, temporal demand, performance, effort and frustration subscale scores as outcome variables. Boxplots showing the data for all six subconstructs of task load can be seen in Figure 9.

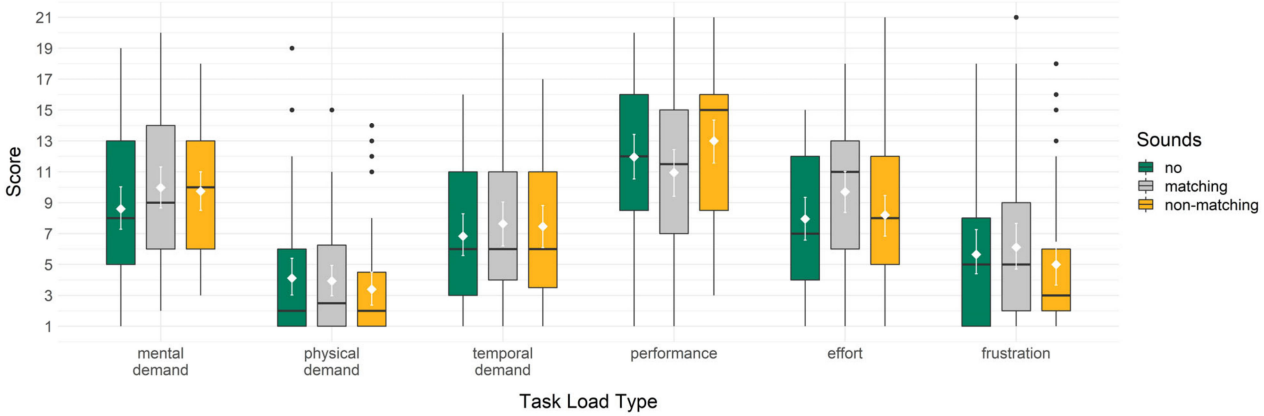


Figure 9. Distribution of task load subscale scores split by group in study two (boxplot with IQR (black), mean with bootstrapped 95% confidence interval (white)).

In H2.2a, mental demand was hypothesized to be lower when no sounds were added than when either matching or non-matching sounds were added, which was indicated descriptively in the group means, showing a lower mental demand in the no sounds ($M = 8.60$, $SD = 4.47$) than the matching sounds ($M = 9.98$, $SD = 4.71$) and the non-matching sounds ($M = 9.74$, $SD = 4.33$) conditions. The overall model of the one-way ANOVA was not significant, $F(2, 127) = 1.15$, $p = 0.320$ and $\omega^2 < 0.01$. Additionally, no significant differences were found in the first comparison of the Helmert contrast analysis, $t = -1.50$, $p = 0.137$, or in the second comparison, $t = 0.24$, $p = 0.810$.

In H2.2b, physical demand was hypothesized to be lower when no sounds were added than when either matching or non-matching sounds were added. An opposite pattern was indicated descriptively in the group means, showing the lowest physical demand in the non-matching sounds ($M = 3.40$, $SD = 3.58$), higher physical demand in the matching sounds ($M = 3.93$, $SD = 3.39$), and the highest physical demand in the no sounds ($M = 4.12$, $SD = 4.23$) condition. The overall model of the one-way ANOVA was not significant, $F(2, 127) = 0.45$, $p = 0.638$, $\omega^2 = -0.01$. Additionally, no significant differences were found in the first comparison of the Helmert contrast analysis, $t = 0.66$, $p = 0.512$, or in the second comparison, $t = 0.69$, $p = 0.494$.

In H2.2c, temporal demand was hypothesized to be lower when no sounds were added than when either matching or non-matching sounds were added, which was indicated descriptively in the group means, showing a lower temporal demand in the no sounds ($M = 6.84$, $SD = 4.57$) than the matching sounds ($M = 7.64$, $SD = 4.92$) and the non-matching sounds ($M = 7.47$, $SD = 4.67$) condition. The overall model of the one-way ANOVA was not significant, $F(2, 127) = 0.34$, $p = 0.710$, $\omega^2 = -0.01$. Additionally, no significant differences were found in the first comparison of the Helmert contrast analysis, $t = -0.81$, $p = 0.419$, or in the second comparison, $t = 0.17$, $p = 0.866$.

In H2.2d, performance was hypothesized to be perceived as higher when no sounds were added than when either matching or non-matching sounds were added. A different pattern was indicated descriptively in the group means, showing the highest perceived performance in the non-matching sounds ($M = 13.00$, $SD = 4.72$), lower perceived performance in the no sounds ($M = 11.95$, $SD = 4.89$) and lowest in the matching sounds ($M = 10.95$, $SD = 5.25$) conditions. The overall model of the one-way ANOVA was not significant, $F(2, 127) = 1.85$, $p = 0.162$ and $\omega^2 = 0.01$. Additionally, no significant differences were found in the first comparison of the Helmert contrast analysis, $t = -0.03$, $p = 0.980$, or in the second comparison, $t = -1.92$, $p = 0.057$.

In H2.2e, effort was hypothesized to be lower when no sounds were added than when either matching or non-matching sounds were added, which was indicated descriptively in the group means, showing a lower effort in the no sounds ($M = 7.95$, $SD = 4.54$) than the matching sounds ($M = 9.70$, $SD = 4.52$) and the non-matching sounds ($M = 8.21$, $SD = 4.38$) conditions. The overall model of the one-way ANOVA was not significant, $F(2, 127) = 1.95$, $p = 0.147$ and $\omega^2 = 0.01$. Additionally, no significant differences were found in the first comparison of the Helmert contrast analysis, $t = -1.20$, $p = 0.232$, or in the second comparison, $t = 1.56$, $p = 0.122$.

In H2.2f, frustration was hypothesized to be lower when no sounds were added than when either matching or non-matching sounds were added, and lower when matching than when non-matching sounds were added. A different pattern was indicated descriptively in the group means, showing the lowest frustration in the non-matching sounds ($M = 5.00$, $SD = 4.89$), higher frustration in the no sounds ($M = 5.65$, $SD = 4.83$) and the highest frustration in the matching sounds ($M = 6.11$, $SD = 4.94$) conditions. The overall model of the one-way ANOVA was not significant, $F(2, 127) = 0.57$, $p = 0.568$ and $\omega^2 = -0.01$. Additionally, no significant differences were found in the first comparison of the Helmert contrast analysis, $t = 0.10$, $p = 0.918$, or in the second comparison, $t = 1.06$, $p = 0.290$.

We thus did not find support for hypotheses H2.2a, H2.2b, H2.2c, H2.2d, H2.2e and H2.2f, showing no significant advantage for not adding sounds concerning mental demand, physical demand, temporal demand, perceived performance, effort and frustration.

3.2.3. H2.3: Knowledge

To test H2.3 on the influence of added sounds on knowledge, the one-way ANOVA included the knowledge test score as the outcome variable. Knowledge was hypothesized to be higher when no sounds were added than when either matching or non-matching sounds were added, and higher when matching than when non-matching sounds were added. Descriptively, an opposite pattern was shown with the highest knowledge in the non-matching sounds ($M = 4.02$, $SD = 1.57$) condition, then the matching sounds ($M = 3.86$, $SD = 1.77$) and the no sounds ($M = 3.70$, $SD = 1.74$) conditions. The overall model of the one-way ANOVA was not significant, $F(2, 127) = 0.40$, $p = 0.674$ and $\omega^2 = -0.01$. Additionally, no significant differences were found in the first comparison of the Helmert contrast analysis, $t = -0.78$, $p = 0.438$, or in the second comparison, $t = -0.44$, $p = 0.661$. Hypothesis H2.3 was thus not supported: no significant advantages of following the coherence principle by leaving out additional sounds were found concerning knowledge outcomes.

As different kinds of items were used in the knowledge test, we also took a closer exploratory look at these. Concerning the four items in which the picture of a bird had to be matched with a textual characteristic or name, the pattern was the same as in the complete knowledge test results: highest score for the non-matching ($M = 2.30$, $SD = 1.15$) compared to the matching ($M = 2.09$, $SD = 1.22$) and no sounds ($M = 2.09$, $SD = 1.19$) conditions. The full model was not significant, $F(2, 127) = 0.45$, $p = 0.636$ and $\omega^2 = -0.01$. Additionally, in the contrast analysis neither the first comparison, $t = -0.47$, $p = 0.640$, nor the second comparison, $t = -0.83$, $p = 0.407$, were significant. Concerning the four items in which only text was included, thus matching the name of a bird with a characteristic, this pattern was a little different. Here, the matching condition had the highest score ($M = 1.77$, $SD = 1.10$), then the non-matching ($M = 1.72$, $SD = 1.08$) and then the no sounds ($M = 1.60$, $SD = 1.03$) conditions. Again, the full model was not significant, $F(2, 127) = 0.28$, $p = 0.755$ and $\omega^2 = -0.01$, and in the contrast analysis neither the first comparison, $t = -0.71$, $p = 0.476$, nor the second comparison, $t = 0.23$, $p = 0.821$, were significant. Although these differences between the groups are very small, the items in which pictures and text needed to be matched thus descriptively showed an advantage for the non-matching sounds condition, while the items in which different textual elements needed to be matched descriptively showed an advantage for the matching sounds condition. Boxplots showing the data split by item type and for the complete knowledge test can also be seen in Figure 10.

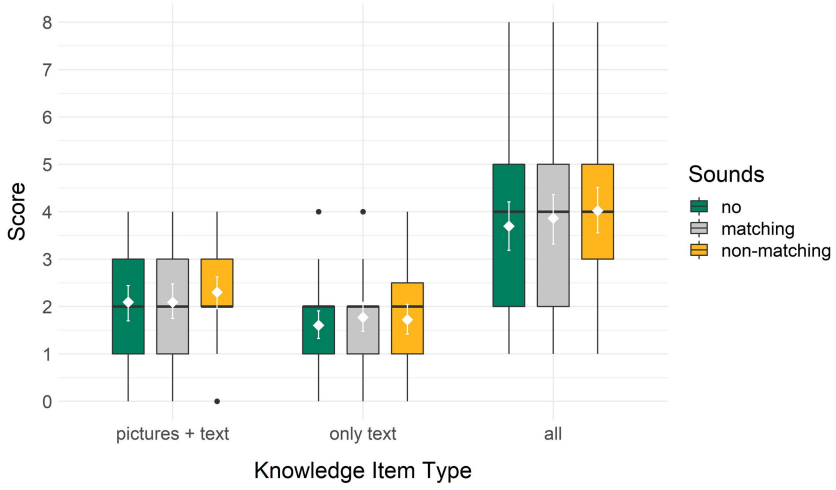


Figure 10. Distribution of knowledge test scores by item types split by group in study two (boxplot with IQR (black), mean with bootstrapped 95% confidence interval (white)).

In additional exploratory analyses, we compared how much the sounds were connected to the pictures and names of the birds in the participants' memory. In this analysis, we thus only compared the groups that received sounds, thus the matching and non-matching group, in a Welch's *t*-test with the score in the audio-matching knowledge test as the outcome variable. From eight possible points, the mean was higher in the non-matching ($M = 1.79$, $SD = 1.10$) than the matching ($M = 1.39$, $SD = 0.84$) condition. Descriptively, the data thus show that the condition with non-matching sounds remembered the connection between sound and bird more correctly than the matching group. The effect has a small effect size and is not significant, $t(78.53) = -1.92$, $p = 0.059$ and $d = -0.41$.

3.3. Discussion

The goal of study two was to implement and examine the coherence principle in a learning environment including a combination of (contextually matched) real and virtual, visual and auditory elements, leveraging the AR characteristic contextuality. We examined the influence of omitting or adding either matching or non-matching sounds on cognitive load, task load and the resulting knowledge. None of the hypotheses were supported, although the tendencies in the data partly showed expected effects. Descriptively, the extraneous cognitive load (H2.1a), mental demand (H2.2a), temporal demand (H2.2c) and effort (H2.2e) scores were as expected, with lower scores when no sounds were added in comparison to matching and non-matching added sounds. For all these scores, matching sounds led to the highest scores, which may suggest that learners tried to remember the sounds, which may have been more demanding and effortful. By adding sounds that were relevant for the learning material, learners may have been more distracted because they may have thought that the sounds were important to listen to and remember, which was probably not the case for the non-matching sounds. This is in accordance with the boundary conditions for the coherence principle specified by Mayer, which describe that the principle is more important when the added extraneous material is more interesting for learners [15]. While participants may also have tried to remember the sounds in the non-matching condition, the difference may have been that those were easier to distinguish and remember. We descriptively found that participants who received the non-matching sounds could better connect them to the birds that appeared simultaneously with the sound. This may have mainly been due to the recognizability of those sounds and that they were easily distinguishable in comparison to the different bird sounds. In the future, research could look at different kinds of sound, especially more and less distinguishable in addition to more and less familiar sounds.

Further non-significant effects that also had different descriptive patterns than those expected show that germane cognitive load (H2.1b) and perceived performance (H2.2d) were highest while physical demand (H2.2b) and frustration (H2.2f) were lowest when non-matching sounds were added. Maybe the non-matching sounds provided the learners with a way to stay alert when the visual information was shown. Additionally, they could have been able to tune these non-relevant sounds out after some time. Concerning the knowledge outcomes, hypothesis H2.3, describing a positive effect of sound omission in comparison to sound addition and a positive effect of matching in comparison to non-matching sounds, was also not supported. Still, interesting descriptive results were found when comparing the different forms of test items. While for the picture–text items (matching the pictures of birds to their characteristics) the highest mean score was in the non-matching sounds condition, for the completely textual items (matching the names of birds to their characteristics) the highest mean score was in the matching sounds condition. Although none of the differences between the groups are significant, this may indicate a difference for different kinds of test items. In general, the lack of significant differences concerning all variables may indicate that the difference between the conditions through the addition of only small sounds such as the ones used in the study is not big enough to completely disrupt learning and confirm an effect based on the coherence principle.

In the current study, we only examined cognitive factors, not affective or motivational aspects. In his description of the theoretical basis for the coherence principle, Mayer contrasts the idea that interesting elements can lead to higher motivation in learners with the active processing assumption of CTML, describing that humans need to actively select and process information for learning, such that no extraneous information should be added to disturb these processes [15]. Still, he further writes that based on the cognitive affective model of learning with media (see [41]) and the integrated cognitive affective model of learning with multimedia (see [42]), cognitive and affective processing can influence each other, describing that the first interest may be attracted through seductive details but that further interest must come from personal value that learners individually find in the material [15]. The cognitive in comparison to the affective factors of seductive details have been discussed in the literature, and it has been found that in high-load situations specifically information that cannot be completely ignored, such as auditory narrations, may interfere with learning [43]. While the information in the current study was not a narration, it was auditory. As expected, the bird sounds may have had a distracting effect, which added load onto the mental tasks of immersing oneself in the AR experience, watching the video and integrating pictorial and textual information, which is more mentally demanding than the combination of pictures and auditory narration (modality principle, [10]).

The descriptive results in the current study seem to at least partly and descriptively support the cognitive detriments of adding contextually relevant and interesting auditory elements to AR-based learning experiences. Still, research on AR often focuses on affective and motivational aspects, which have been identified as important factors in AR-based instructional environments (e.g., [4,19,21]). It may thus be interesting for future research to take a closer look at how the addition of atmospheric, relevant sounds in AR experiences have an influence on the enjoyment of the experience, and for example on learners' wishes to use the application more often, while still keeping the cognitive aspects in mind. The usage of environmental sound in the form of spatial AR soundscapes for realistic experiences through specific recording and reproduction techniques, taking into account movement and an interaction of virtual with real sounds [11], might be considered for this. A learning-specific focus might be on the relevance of atmospheric and specific sounds for motivational and cognitive learning goals within a specific context.

Again, the study was executed with a video-based implementation simulating an AR-based learning experience. In addition to the real and virtual visual elements presented in study one, in study two there were also additional real environmental sounds, and in two of the three conditions virtual sounds were added. This video-based implementation might thus have provided a more sensorily immersive experience than in study one. An advantage of using videos again was being able to keep the conditions the same except for the manipulated variable. The participants all received exactly the same visual material with the same real sounds in the background; only the added virtual sounds were different. In a study in a real forest, factors such as the weather, background sounds and present animals may differ between participants and lead to different coherence-related circumstances and distractions. Additionally, the interaction and duration might differ when people have control over the usage, where they may focus on different aspects of the material and environment. Through the video-based implementation, confounding factors are thus decreased and (descriptive) differences can be attributed to the omission or addition of sounds. The limitations due to this implementation are again mainly based on the question of the transferability of the results onto real AR-based learning experiences. Especially concerning coherence, the interaction with, movement in and presence within an immersive environment may play a big role and can be both motivating and distracting, as stated in Mayer's immersion principle [31]. These factors could not be considered in the same way in the present study, because the immersion in a video-based presentation is different than physically being in the environment. The transferability should be tested in future studies. The advantages and challenges of the video-based implementation will be further discussed in the general discussion.

Other limitations of the study were that the real-world background sound of the forest included bird sounds from the environment that may have been distracting, especially in the matching sounds condition. As this is realistic, it increases comparability to a setting in the real world but cannot be excluded as a confounding factor and should be taken into account when interpreting the results. Furthermore, some participants had technical difficulties with the video playback, although this was not systematic in one condition but randomly distributed.

In conclusion, it can be said that the descriptive results are not conclusive in supporting the coherence principle in the case of omitting or adding virtual matching or non-matching sounds in an environment with real and virtual elements. The addition of small sounds might not have that big of an influence, and future research should test for the transferability of the results onto more complex materials and real AR-based learning environments.

4. General Discussion

The goal of the paper is to apply multimedia design principles to the specific case of AR and evaluate them in experimental studies concerning their effects on cognitive load, task load and the resulting knowledge. Only a few of the hypotheses that we tested in the two studies were supported, although interesting and mostly expected descriptive tendencies emerged. In study one, the data did not fully support the expected positive effect of following the spatial contiguity principle through the integration of visual real and virtual elements on cognitive load, task load, and knowledge. Only the hypotheses describing decreased temporal demand and increased perceived performance through the integration of the material were supported. Descriptively, we also found the expected results of integration of the material concerning decreased extraneous cognitive load, mental demand, physical demand, effort, and frustration, in addition to increased germane cognitive load and positive results for the picture–text knowledge test items. In study two, the data did also not fully support the expected positive effects of following the coherence principle by omitting virtual sounds that are matching or non-matching to the learning material on cognitive load, task load, and knowledge. We did find descriptively lowest extraneous cognitive load, mental demand, temporal demand, and effort when no sounds were added, but opposed to what we expected found lowest physical demand and highest germane cognitive load and perceived performance when non-matching sounds were added. Again, different kinds of knowledge test items showed different patterns, with textual items revealing the highest score when matching sounds were added, and picture–text items revealing the highest score when non-matching sounds were added.

4.1. Methodological Approach

As already described in the study-specific discussions, there are several advantages but also some limitations due to the video-based, simulated implementation of the AR environment. While the implementation of the video format in a fully online study was a great opportunity for us to conduct AR-related research during social distancing measures in the COVID-19 pandemic, beyond this there are additional advantages for experimental research. Due to the possibility of standardizing the experience for all participants, individual differences in the usage of real, interactive AR applications (e.g., duration of engagement with specific components, focus on other aspects of the environment) and subsequently evoked random or systematic error can be decreased. Furthermore, the novelty effect that is often reported as a confounding variable in research on new technologies, such as AR [44], can probably be decreased when participants just imagine using real AR. Chang and colleagues, for example, found that an AR application increased learners' motivation in comparison to a learning video [45]. With a focus on cognitive aspects, removing potential motivational effects may increase the interpretability of the data. Additionally, disruptions or distractions that the operation of unknown technologies might bring are decreased. Through these advantages, effects can more securely be attributed to the studies' manipulations.

There are also apparent limitations of the approach. While the above-mentioned advantages all help with increased internal validity and the possibility of attributing effects to the experimental manipulations, those advantages also bring limitations when it comes to the question of the transferability of the results to the real usage of AR. The immersion principle by Mayer [31] states that immersive virtual environments may not always lead to better learning results because distractions of the environment may evoke an increase in necessary cognitive processing and eliminate motivational effects. This may also be the case for immersion in real-world environments that might provide motivational and atmospheric advantages but also increased cognitive demands due to many distractions. While removing these distracting and motivational aspects for research may help in decreasing confounding variables, those factors are not removed when using real AR applications. Furthermore, it is not clear if participants could really imagine being in the situations and if the perception of the real elements in the videos is comparable with the perception of those elements in physical reality. We did not ask the participants about their experience of imagining the situation taking place in the real world. Interaction and self-direction are key features in AR and every learner can have a different, individual experience when learning with AR, such that, for example, coherence-related factors may differ for each learner. In the video-based implementations the pace of the influx of information could not be regulated by the learners, although we did try to give them a lot of time to process information. This may have been more time than necessary for many participants, which may have led to boredom. While we do not know of cases in which this approach has already been used to evaluate AR, there has been research on general instructional videos in which virtual elements are added for extra information or guidance. In a study by Tsiatsos and colleagues, for example, virtual footprint symbols were added onto recorded videos of dancers for the teaching of Greek traditional dances [46]. These kinds of video overlays were classified as design patterns for video annotation and described as video augmentation through synchronized information overlays emphasizing the contents, linking to related content or containing additional relevant information themselves [47]. This may, in general, be an interesting alternative to fully location-based AR environments when their development or implementation is not possible, but the combination of real and virtual elements is an important focus of a learning area.

Both studies took place online with a supervising researcher over a (video/voice) call. The researcher was in direct contact at the beginning as well as the end and was available for questions during the study. Furthermore, participants were asked to minimize all distractions to account for the online situation. Still, especially concerning the learning phase, it is not completely clear how seriously the participants watched the videos, and external distractions could not be controlled for. In study two it might even be the case that participants did not listen to the sounds. They were asked to use headphones and adjust their volume based on a sound before the video, but they might still have turned sounds off or removed their headphones afterwards. In a laboratory those factors could have been better controlled.

Another limitation is based on the usage of questionnaires for measuring cognitive load and task load. While this method provided us with differentiated outcomes based on the different subconcepts, such as the different types of cognitive load, for future research the addition of objective, continuous measures of cognitive load may be considered. Retrospective measures cannot capture changes in load during the learning task, which may be particularly interesting for learning environments in which the focus on learning materials shifts over time due to navigational periods when looking for the next material (e.g., walking from one to the next plant or bird) or when environmental distracting variables change. Physical, continuous measures, such as heart rate or eye movement, may be able to catch those changes without disrupting learning processes and should be considered as complementary methods in the future. Objective measures can be valuable for verifying subjective results, which are sometimes biased because not all learners may rate cognitive load in an expected way [48].

4.2. Future Research

Implementing different multimedia design principles in AR adds new features to the mix, including the dimension of reality. As seen in the presented studies, following the principles of spatial contiguity and coherence was possible with combined presentations of virtual and real elements, and at least descriptively led to partly expected results. Based on the limitations described above, evaluating the implementations in more complex, naturalistic environments with real AR-based learning material and with more complex learning topics is necessary for their confirmation. In general, as already suggested in meta-reviews by [26–29], Mayer's CTML and following its multimedia design principles are important approaches used in multiple studies on AR in education. Although we focused on those in the current study, spatial contiguity and coherence are not the only principles that may play an important role in AR-based learning and instruction.

The *multimedia principle*, which states that text and pictures should be combined in instructional material for better learning [49], is in general important due to the multiple representations that can be used in AR. Elements of the real world are often pictorial, but both textual and pictorial virtual elements can be added, as seen in the presented studies. The combination of virtual texts and pictures with different real-world locations and elements might be interesting for research concerning specific multimedia mechanisms. In study two, the focus was on the *coherence principle* with additional more or less relevant sounds. While sounds are part of the third version of the coherence principle, the first version focuses more on visual aspects that distract learners' attention away from relevant visual elements [15]. In AR, additional information that may not be necessary for learning but can lead to a different context or atmosphere is often part of learning experiences due to the environment in which the learning takes place. Relevant visual backgrounds give a context and should thus increase immersion and motivation, but may distract from the learning material and use additional resources for unnecessary processing. A potentially detrimental seductive details effect concerning visual aspects in AR experiences should be examined in future studies.

When pictorial and textual information are combined, it is always good to take a look at the *modality principle* [50], and thus the combination of pictures and spoken words. AR can be used with a lot of different media, such that the addition of narration to pictures may bring a great potential to follow the modality principle without a lot of effort here. While AR is often described as a visual experience, the inclusion of auditory elements is possible and should be examined more closely. When narrations are added to AR-based experiences, following the *temporal contiguity principle* [51] also seems quite straightforward for the combination of virtual elements with real-world environments. The spatial integration of information could be easily achieved in the representations in study one. The temporal contiguity principle expands this idea towards narrations, suggesting that spoken text is provided exactly when corresponding pictorial virtual elements are shown and, in an AR-specific case, exactly when looking at corresponding physical elements in the real world. Here, it should be distinguished between information where one part of the elements is only understandable when the other part is also available (e.g., verbally describing processes in the real world that are not apparent or visible) and information where each part is understandable on its own, but where adding the other part may give more information (e.g., naming and characterizing objects in the real world through spoken text).

The focus of the current studies is on very specific features of instructional design of AR applications. In reality, AR applications are often more complex and incorporate more features than just a short presentation of spatially and contextually embedded information. In the taxonomy that Mystakidis and colleagues developed based on a systematic review, five clusters of instructional methods in AR-based learning interventions are described and the approaches in the current studies best fit into the most passive, teacher-centered category where presentation and observation are in focus [25]. The other four clusters describe methods that focus on activities that are more or completely learner-controlled, including project-based and experiential learning. In order to establish an empirical basis for the

integration of multimedia principles into those more complex and interactive instructional methods, research in more authentic, naturalistic settings in which learners can interact with the representations is necessary. An important aspect that distinguishes AR settings from non-AR multimedia learning is the focus on being immersed and feeling present within a specific environment as an important affordance of the experience [20]. While Mayer's immersion principle states that immersion does not necessarily lead to improved learning, the motivational aspects of immersion are recognized [31]. Its interaction with the application of the other multimedia principles may provide an interesting research area. The feeling of presence is also relevant in VR settings, and maybe even more so in remote collaboration through VR, where it was suggested to be included into a taxonomy of computer-supported cooperative work (CSCW) [52]. In further steps, collaborative or cooperative learning settings, which often come naturally with more interactive and project-based settings, may be the focus of future research on the implementation of multimedia principles in AR-based learning.

Concerning the methods used, for future examinations of implementations of multimedia design principles in AR, it seems to be important to further take into account different types of cognitive load (e.g., extraneous and germane cognitive load), which compared to each other descriptively showed different patterns in both studies. Additionally, the different subconstructs of task load either supported each other, as in study one, or also showed different, more differentiated patterns, as in study two. Adding this measure of the NASA TLX, which is often used in research on AR [32], can provide researchers with more detailed information on potential effects. Furthermore, both studies showed that different kinds of knowledge and thus different test items (i.e., combining pictures and text; text only) should be considered for learning outcome examination. When learning with AR in the outside world, the real-world environment is often a pictorial and not a textual representation. Additional visual virtual materials can be both pictures and text, and multimedia effects should be further examined including the factor of virtual and physical realities in AR.

In general, we can conclude that AR brings many new opportunities concerning the application of multimedia design principles in the context of physical and virtual elements. The characteristics of AR, including spatiality, interactivity and contextuality, can help in considering AR-specific potentials when implementing those principles. Spatial contiguity and coherence, as two well-studied and -supported principles, were the focus of the current paper, although more should be implemented and investigated in future studies.

Author Contributions: Conceptualization, J.M.K. and D.B.; Methodology, J.M.K.; Formal Analysis, J.M.K.; Investigation, J.M.K.; Data Curation, J.M.K.; Writing—Original Draft Preparation, J.M.K. and D.B.; Writing—Review and Editing, J.M.K. and D.B.; Visualization, J.M.K.; Supervision, D.B. and J.M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The studies were conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the Department of Computer Science and Applied Cognitive Science of the University of Duisburg-Essen (study one: ID 2011PFBS7216, 26.11.2020; study two: ID 2012PFKL8474, 7 December 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the studies.

Data Availability Statement: Publicly available datasets were analyzed in this study. These datasets can be found here: study one: osf.io/5gzyv/; study two: osf.io/qtnr7/.

Acknowledgments: We thank Laura Kusber and Stefan Baudeck for creating the video-based material and collecting the data in the studies.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Result of Levene's test for all variables.

Study One—Levene's Test	*F*-Value	df1	df2	*p*
H1.1a: extraneous cognitive load	0.09	1	78	0.764
H1.1b: germane cognitive load	1.00	1	78	0.320
H1.2a: mental demand	0.64	1	78	0.427
H1.2b: physical demand	4.07	1	78	0.047 *
H1.2c: temporal demand	0.53	1	78	0.469
H1.2d: performance	0.18	1	78	0.671
H1.2e: effort	0.15	1	78	0.696
H1.2f: frustration	1.25	1	78	0.266
H1.3: knowledge	0.78	1	78	0.379
Study two—Levene's Test	***F*-Value**	**df1**	**df2**	***p***
H2.1a: extraneous cognitive load	0.53	2	127	0.592
H2.1b: germane cognitive load	0.54	2	127	0.585
H2.2a: mental demand	0.20	2	127	0.819
H2.2b: physical demand	0.93	2	127	0.397
H2.2c: temporal demand	0.46	2	127	0.633
H2.2d: performance	0.06	2	127	0.944
H2.2e: effort	0.46	2	127	0.633
H2.2f: frustration	0.05	2	127	0.952
H2.3: knowledge	1.48	2	127	0.232

Note. * $p < 0.05$.

References

1. Mayer, R. 2 Science of Learning: Determining How Multimedia Learning Works. In *Multimedia Learning*; Cambridge University Press: Cambridge, UK, 2020; pp. 29–62, ISBN 978-1-316-94135-5.
2. Radu, I. Augmented Reality in Education: A Meta-Review and Cross-Media Analysis. *Pers. Ubiquitous Comput.* **2014**, *18*, 1533–1543. [CrossRef]
3. Krüger, J.M.; Buchholz, A.; Bodemer, D. Augmented Reality in Education: Three Unique Characteristics from a User's Perspective. In Proceedings of the 27th International Conference on Computers in Education, Kenting, Taiwan, 2–6 December 2019; Chang, M., So, H.-J., Wong, L.-H., Yu, F.-Y., Shih, J.L., Eds.; Asia-Pacific Society for Computers in Education: Kenting, Taiwan, 2019; pp. 412–422.
4. Garzón, J. An Overview of Twenty-Five Years of Augmented Reality in Education. *Multimodal Technol. Interact.* **2021**, *5*, 37. [CrossRef]
5. Mayer, R.E. *Multimedia Learning*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009; ISBN 978-0-511-81167-8.
6. Baddeley, A.D. *Essentials of Human Memory*; Cognitive psychology; Psychology Press: Hove, UK, 1999; ISBN 978-0-86377-545-1.
7. Paivio, A. *Mental Representations: A Dual Coding Approach*; Oxford Psychology Series; Oxford University Press: Oxford, UK; Clarendon Press: New York, NY, USA, 1986; ISBN 978-0-19-503936-8.
8. Sweller, J.; van Merrienboer, J.J.G.; Paas, F.G.W.C. Cognitive Architecture and Instructional Design. *Educ. Psychol. Rev.* **1998**, *10*, 251–296. [CrossRef]
9. Sweller, J.; van Merriënboer, J.J.G.; Paas, F.G.W.C. Cognitive Architecture and Instructional Design: 20 Years Later. *Educ. Psychol. Rev.* **2019**, *31*, 261–292. [CrossRef]
10. Mayer, R. 13 Modality Principle. In *Multimedia Learning*; Cambridge University Press: Cambridge, UK, 2020; pp. 281–300. ISBN 978-1-316-94135-5.
11. Hong, J.Y.; He, J.; Lam, B.; Gupta, R.; Gan, W.-S. Spatial Audio for Soundscape Design: Recording and Reproduction. *Appl. Sci.* **2017**, *7*, 627. [CrossRef]
12. Mayer, R. 9 Spatial Contiguity Principle. In *Multimedia Learning*; Cambridge University Press: Cambridge, UK, 2020; pp. 207–226. ISBN 978-1-316-94135-5.
13. Ayres, P.; Sweller, J. The Split-Attention Principle in Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning*; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 206–226. ISBN 978-1-139-54736-9.
14. Schroeder, N.L.; Cenkci, A.T. Spatial Contiguity and Spatial Split-Attention Effects in Multimedia Learning Environments: A Meta-Analysis. *Educ. Psychol. Rev.* **2018**, *30*, 679–701. [CrossRef]
15. Mayer, R. 6 Coherence Principle. In *Multimedia Learning*; Cambridge University Press: Cambridge, UK, 2020; pp. 143–165, ISBN 978-1-316-94135-5.

16. Jung, T.; tom Dieck, M.C.; Lee, H.; Chung, N. Relationships among Beliefs, Attitudes, Time Resources, Subjective Norms, and Intentions to Use Wearable Augmented Reality in Art Galleries. *Sustainability* **2020**, *12*, 8628. [CrossRef]
17. Han, K.; Park, K.; Choi, K.-H.; Lee, J. Mobile Augmented Reality Serious Game for Improving Old Adults' Working Memory. *Appl. Sci.* **2021**, *11*, 7843. [CrossRef]
18. Cabero-Almenara, J.; Barroso-Osuna, J.; Martinez-Roig, R. Mixed, Augmented and Virtual, Reality Applied to the Teaching of Mathematics for Architects. *Appl. Sci.* **2021**, *11*, 7125. [CrossRef]
19. Akçayır, M.; Akçayır, G. Advantages and Challenges Associated with Augmented Reality for Education: A Systematic Review of the Literature. *Educ. Res. Rev.* **2017**, *20*, 1–11. [CrossRef]
20. Wu, H.-K.; Lee, S.W.-Y.; Chang, H.-Y.; Liang, J.-C. Current Status, Opportunities and Challenges of Augmented Reality in Education. *Comput. Educ.* **2013**, *62*, 41–49. [CrossRef]
21. Bacca, J.; Baldiris, S.; Fabregat, R.; Graf, S. Kinshuk Augmented Reality Trends in Education: A Systematic Review of Research and Applications. *Educ. Technol. Soc.* **2014**, *17*, 133–149.
22. Goff, E.E.; Mulvey, K.L.; Irvin, M.J.; Hartstone-Rose, A. Applications of Augmented Reality in Informal Science Learning Sites: A Review. *J. Sci. Educ. Technol.* **2018**, *27*, 433–447. [CrossRef]
23. Bower, M.; Howe, C.; McCredie, N.; Robinson, A.; Grover, D. Augmented Reality in Education–Cases, Places and Potentials. *Educ. Media Int.* **2014**, *51*, 1–15. [CrossRef]
24. Mayer, R. 1 The Promise of Multimedia Learning. In *Multimedia Learning*; Cambridge University Press: Cambridge, UK, 2020; pp. 3–28, ISBN 978-1-316-94135-5.
25. Mystakidis, S.; Christopoulos, A.; Pellas, N. A Systematic Mapping Review of Augmented Reality Applications to Support STEM Learning in Higher Education. *Educ. Inf. Technol.* **2021**. [CrossRef]
26. Sommerauer, P.; Müller, O. Augmented Reality for Teaching and Learning-a Literature Review on Theoretical and Empirical Foundations. In Proceedings of the Twenty-Sixth European Conference on Information Systems (ECIS2018), Portsmouth, UK, 23–28 June 2018.
27. Da Silva, M.M.O.; Teixeira, J.M.X.N.; Cavalcante, P.S.; Teichrieb, V. Perspectives on How to Evaluate Augmented Reality Technology Tools for Education: A Systematic Review. *J. Braz. Comput. Soc.* **2019**, *25*, 3. [CrossRef]
28. Garzón, J.; Baldiris, S.; Gutiérrez, J.; Pavón, J. How Do Pedagogical Approaches Affect the Impact of Augmented Reality on Education? A Meta-Analysis and Research Synthesis. *Educ. Res. Rev.* **2020**, *31*, 100334. [CrossRef]
29. Buchner, J.; Buntins, K.; Kerres, M. The Impact of Augmented Reality on Cognitive Load and Performance: A Systematic Review. *J. Comput. Assist. Learn.* **2021**. [CrossRef]
30. Altmeyer, K.; Kapp, S.; Thees, M.; Malone, S.; Kuhn, J.; Brünken, R. The Use of Augmented Reality to Foster Conceptual Knowledge Acquisition in STEM Laboratory Courses—Theoretical Background and Empirical Results. *Br. J. Educ. Technol.* **2020**. [CrossRef]
31. Mayer, R. 18 Immersion Principle. In *Multimedia Learning*; Cambridge University Press: Cambridge, UK, 2020; ISBN 978-1-316-94135-5.
32. Buchner, J.; Buntins, K.; Kerres, M. A Systematic Map of Research Characteristics in Studies on Augmented Reality and Cognitive Load. *Comput. Educ. Open* **2021**, *2*, 100036. [CrossRef]
33. Wigfield, A.; Eccles, J.S. Expectancy–Value Theory of Achievement Motivation. *Contemp. Educ. Psychol.* **2000**, *25*, 68–81. [CrossRef] [PubMed]
34. Klepsch, M.; Schmitz, F.; Seufert, T. Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. *Front. Psychol.* **2017**, *8*, 1997. [CrossRef] [PubMed]
35. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*; Elsevier: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183, ISBN 978-0-444-70388-0.
36. Hart, S.G. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2006**, *50*, 904–908. [CrossRef]
37. Delacre, M.; Lakens, D.; Leys, C. Why Psychologists Should by Default Use Welch's *t*-Test Instead of Student's *t*-Test. *Rips* **2017**, *30*, 92–101. [CrossRef]
38. Seufert, T. Supporting Coherence Formation in Learning from Multiple Representations. *Learn. Instr.* **2003**, *13*, 227–237. [CrossRef]
39. Seufert, T.; Brünken, R. Supporting Coherence Formation in Multimedia Learning. In Proceedings of the Instructional design for effective and enjoyablecomputer-supported learning. In Proceedings of the First Joint Meeting of the EARLI SIGs Instructional Design and Learning and Instruction with Computers, Tübingen, Germany, 7–9 July 2004; Gerjets, P., Kirschner, P., Elen, J., Joiner, R., Eds.; Knowledge Media Research Center: Tübingen, Germany, 2004; pp. 138–147.
40. Dunleavy, M.; Dede, C.; Mitchell, R. Affordances and Limitations of Immersive Participatory Augmented Reality Simulations for Teaching and Learning. *J. Sci. Educ. Technol.* **2009**, *18*, 7–22. [CrossRef]
41. Moreno, R.; Mayer, R.E. Interactive Multimodal Learning Environments. *Educ. Psychol. Rev.* **2007**, *19*, 309–326. [CrossRef]
42. Plass, J.L.; Kaplan, U. Chapter 7-Emotional Design in Digital Media for Learning. In *Emotions, Technology, Design, and Learning*; Tettegah, S.Y., Gartmeier, M., Eds.; Emotions and Technology; Academic Press: San Diego, CA, USA, 2016; pp. 131–161, ISBN 978-0-12-801856-9.
43. Park, B.; Flowerday, T.; Brünken, R. Cognitive and Affective Effects of Seductive Details in Multimedia Learning. *Comput. Hum. Behav.* **2015**, *44*, 267–278. [CrossRef]

44. Derby, J.L.; Chaparro, B.S. The Challenges of Evaluating the Usability of Augmented Reality (AR). *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2021**, *65*, 994–998. [CrossRef]
45. Chang, R.-C.; Chung, L.-Y.; Huang, Y.-M. Developing an Interactive Augmented Reality System as a Complement to Plant Education and Comparing Its Effectiveness with Video Learning. *Interact. Learn. Environ.* **2016**, *24*, 1245–1264. [CrossRef]
46. Tsiatsos, T.; Stavridou, E.; Grammatikopoulou, A.; Douka, S.; Sofianidis, G. Exploiting Annotated Video to Support Dance Education. In Proceedings of the 2010 Sixth Advanced International Conference on Telecommunications, Barcelona, Spain, 9–15 May 2010; pp. 100–105.
47. Seidel, N. Interaction Design Patterns for Spatio-Temporal Annotations in Video Learning Environments. In *Proceedings of the Proceedings of the 20th European Conference on Pattern Languages of Programs*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 1–21.
48. Zu, T.; Munsell, J.; Rebello, N.S. Subjective Measure of Cognitive Load Depends on Participants' Content Knowledge Level. *Front. Educ.* **2021**, *6*, 647097. [CrossRef]
49. Mayer, R. 5 Multimedia Principle. In *Multimedia Learning*; Cambridge University Press: Cambridge, UK, 2020; pp. 117–138, ISBN 978-1-316-94135-5.
50. Mayer, R.E.; Pilegard, C. Principles for Managing Essential Processing in Multimedia Learning: Segmenting, Pre-Training, and Modality Principles. In *The Cambridge Handbook of Multimedia Learning*; Mayer, R., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 316–344, ISBN 978-1-139-54736-9.
51. Mayer, R.E. Temporal Contiguity Principle. In *Multimedia Learning*; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2009; pp. 153–169, ISBN 978-0-511-81167-8.
52. Cruz, A.; Paredes, H.; Morgado, L.; Martins, P. Non-Verbal Aspects of Collaboration in Virtual Worlds: A CSCW Taxonomy-Development Proposal Integrating the Presence Dimension. *JUCS-J. Univers. Comput. Sci.* **2021**, *27*, 913–954. [CrossRef]