



Comparing human-made and AI-generated teaching videos: An experimental study on learning effects

Torbjørn Netland^{*}, Oliver von Dzengelevski, Katalin Tesch, Daniel Kwasnitschka

Department of Management, Technology, and Economics, ETH Zurich, Weinbergstrasse 56/58, 8092, Zurich, Switzerland

ARTICLE INFO

Keywords:

Data science applications in education
Distance education and online learning
Applications in subject areas

ABSTRACT

In the age of generative AI, can teaching videos be efficiently and effectively generated by large language models? In this study, the authors used generative AI tools to develop four short teaching videos for a management course and then compared them with human-generated videos on the same subjects in an online experiment. In an across-subject experimental design, 447 participants completed two treatment conditions presenting different mixes of AI-generated and human-made videos. The participants were asked to rate their learning experiences after each video and had their learning outcomes tested in a multiple-choice exam at the end of the session ($N = 1788$ video treatments). The findings show that human-generated videos provided a statistically significant but small advantage to participants in terms of learning experience, indicating that the participants still prefer to be taught by human teachers. However, a comparison of exam results between the experimental groups implies that the participants eventually acquired knowledge about the content to a similar degree. Given these findings and the ease with which AI-generated teaching videos can be created, this study concludes that AI-generated teaching videos will likely proliferate.

1. Introduction

The release of OpenAI's chatbot ChatGPT 3.5 (Chat Generative Pre-trained Transformer) on November 30, 2022, has left technology reporters and scholars awed by the new opportunities. Such AI-powered chatbot technologies have developed fast over the past years, including Google's LaMDA (Language Model for Dialogue Applications). Yet, it was the first time this technology was available to the general public in the quality of ChatGPT 3.5. Large language models can generate illustrations, images, avatars, and slideware materials and read texts with human-like performance. These tools can potentially disrupt current teaching models by taking over many of the tasks that teachers do today (Alam, 2021; Wang et al., 2023). In this study, we continue a recent stream of research that studies the opportunities of generative AI to create teaching videos efficiently and effectively (Leiker et al., 2023a, 2023b). Our research objective is to test the learning experience and the learning outcome of human-made versus AI-generated teaching videos.

We take advantage of having 45 teaching videos developed during the Covid-19 pandemic. The videos have been publicly available on an educational YouTube channel since the fall semester of 2021¹ and are used in a course on operations management at our university. These 45 videos were entirely made by the teaching professor and staff; no form of AI was used to generate any content or material. From this database, we selected four different topics and developed comparable videos with generative AI. The videos cover

^{*} Corresponding author.

E-mail address: tnetland@ethz.ch (T. Netland).

¹ All videos can be accessed at the following URL: <https://www.youtube.com/hashtag/pom4all>.

four classic operations management concepts: the workplace organization method 5S (video 1), the theory of inventive problem-solving (TRIZ) (video 2), just in time (JIT) and Kanban (video 3), and economic order quantity (EOQ) (video 4). To strengthen the external validity of our study, the videos are independent of each other and cover both quantitative and qualitative concepts taught by four different teachers. We used ChatGPT3.5 to generate the video script, Midjourney and Dall-E to generate visuals, and Movio (now HeyGen) to generate a teaching avatar and an audio clip. The teaching team provided prompts to the different AI tools, as presented in the method section of this paper. It is important to note that the teaching team did not *create* content but *curated* the AI-generated content into teaching videos of length and design comparable to the human-made videos.

To test the effectiveness of AI-generated teaching videos versus manually created videos, we administered an online experiment with 447 participants on Amazon Mechanical Turk (MTurk). In the experiment, participants were assigned randomly into two conditions: the first condition presents the four videos in the sequence AI-generated video 1, human-made video 2, AI-generated video 3, and human-made video 4, while the second condition presents the four videos in the sequence human-made video 1, AI-generated video 2, human-made video 3 and AI-generated video 4. The participants were surveyed upon completing each video concerning their learning experience and, finally, took an exam to test learning outcomes, including equally many questions from each of the four topics. The survey and exam questions were the same for both treatment conditions. In total, we collected $N = 1341$ assessments of learning experience and learning outcome and analyzed the data from our experiment econometrically.

2. Background and hypothesis development

In recent years, there has been a gradual shift toward online and asynchronous learning (Beege et al., 2022; Sablić et al., 2021; Tondeur et al., 2023). This trend can be attributed to several factors, such as the increased accessibility and convenience of online platforms, potential changes in education systems through the entry of digital natives into schools and universities, and the experiences gained during the Covid-19 pandemic when most education had to shift from classroom to online teaching in a matter of only a few weeks (Alon et al., 2023; Anders et al., 2024). Popular online university platforms like Khan Academy, Coursera, Udacity, and edX, which offer massive open online courses (MOOCs) on various topics, have showcased the effectiveness of online teaching with videos (Ruipérez-Valiente et al., 2022). Also, other platforms like YouTube and TikTok have entered the education market (Nguyen et al., 2023; Orús et al., 2016). Many universities and schools have already started offering online courses and degree programs, with video-based instruction playing a pivotal role.

Educational videos have traditionally been produced by human teachers and education technology experts. However, the latest developments in generative AI suggest that AI may be capable of taking over this and other educational tasks in the future (Jeon et al., 2023; Lim et al., 2023). There were only 15 months between the launch of Open AI's ChatGPT 3.5 in November 2022 and the launch of Open AI's text-to-video tool Sora on February 15, 2024. By leveraging advanced algorithms that analyze and learn from vast amounts of data, generative AI systems can potentially produce highly engaging and effective teaching materials tailored to the needs of learners and teachers. These AI-driven systems can identify complex patterns and relationships within content domains, allowing for synthesizing novel and coherent educational videos. Generative AI can integrate multimodal data such as text, audio, and visuals in the generation process, accommodating diverse learning preferences. Ultimately, the use of generative AI in creating educational teaching videos has the potential to transform pedagogical practices, improve learning, and democratize access to high-quality educational resources worldwide (Dao et al., 2021; Leiker, Finnigan, et al., 2023).

We build on the pedagogy research that separates *learning experience* and *learning outcome* (Allan, 1996). Learning experience encompasses the student's felt experience during the acquisition of knowledge, including their emotional, cognitive, and social engagement with the content, teaching methods, and interactions. In contrast, learning outcomes focus on the measurable knowledge, skills, and abilities that students gain directly from their involvement in the learning experience. This choice mirrors the recent research of Leiker, Gyllen, et al. (2023) and Pi et al. (2022), who studied the effects of replacing a human teacher with an AI-generated avatar on students' learning experiences and outcomes.

2.1. Learning experience for human-made versus AI-generated teaching videos

Because learning occurs within the learner rather than on the side of the teacher, learning experience is argued to be a robust proxy for learning (e.g., Nehari et al., 1978; Zerihun et al., 2012). When learners enjoy the teaching video, they are more likely to pay attention and internalize the content. Our study furthers the scholarly research on whether generative AI can replace humans as content creators (Lawson et al., 2021; Leiker, Finnigan, et al., 2023; Liu et al., 2024; Pinkwart et al., 2020; Schroeder et al., 2022; Tang et al., 2021) and student-facing lecturers (e.g., Dao et al., 2021; Leiker, Gyllen, et al., 2023; Pataranutaporn et al., 2021; Pi et al., 2022) without infringing on their learning experience.

Recent research by Leiker, Gyllen, et al. (2023) suggests that replacing human teachers with an AI avatar does not reduce students' learning experience. However, their study is limited by the use of a single human teacher in a talking head video compared to a distinct AI avatar (cf., Beege et al., 2022). To the best of our knowledge, no study has explored the difference in learning experience and outcomes between human-made teaching videos and comparable AI-generated videos.

Arguably, human-made videos should offer a more relatable and empathetic approach, with educators drawing from their own experiences and emotions to create content that resonates with students (Schneider et al., 2022; Tack et al., 2022). Chu et al. (2017) discuss the importance of emotional arcs in storytelling and capturing attention. Though they argue that AI may assist humans in enhancing their storytelling, it is likely that humans are still better at making content compelling for other humans. This human touch should positively influence the learning experience.

A growing branch of literature is concerned with the detection of AI content, analyzing differences between AI-generated and human-generated content (e.g., Rana et al., 2022; Weber-Wulff et al., 2023). Markowitz et al. (2024) argue that though AI-generated content is increasingly difficult to distinguish from human-generated content, “AI-generated text had a more analytic style and was more affective, more descriptive, and less readable than human-generated text.” (p. 63) Less readable scripts for educational videos would translate to less logically coherent teaching material, thus negatively affecting learners’ perception of the respective video and reducing their learning experience.

Chu and Liu (2023) find that AI-generated narratives reduce counterarguing and promote story-consistent emotional responses and intentions compared to human-generated content. This may render AI-generated teaching content more compelling for viewers and, hence, improve their learning experience. However, Chu and Liu (2023) also find that labeling content as AI leads to the opposite reaction—less effective narrative transportation, higher counterarguing and less story-consistent responses. As participants may recognize the use of AI due to AI-specific features, their skepticism may outweigh the potential advantages in the crafting of the narrative of the presented content.

Considering the arguments above, we will test the following hypothesis.

H1. *Learners rate their learning experience higher after watching teaching videos made by humans than after watching comparable videos generated by AI.*

2.2. Learning outcomes for human-made versus AI-generated teaching videos

In terms of learning outcomes, research has documented mixed results. In a meta-review of studies on digital pedagogical agents—on-screen characters that facilitate instruction for the learner—Schroeder et al. (2013, p. 20) conclude that such tools have “small, yet positive effects on learning.” Their finding is also supported by more recent studies that concluded that AI-generated teaching content can be more effective than human teachers (Leddo et al., 2021; Pi et al., 2022). However, the study by Leiker, Gyllen, et al. (2023) finds no statistically significant difference in learning outcomes between a group of students who watched teaching videos featuring a human teacher compared to a group that watched a teaching video delivered by an AI-avatar.

While the presence of human teachers can make the learning experience more relatable, it can also distract from the core content of the video (Sondermann et al., 2024). AI-generated videos may reduce typical human flaws, such as speech inconsistencies, distracting physical movements and emotional biases, and this way help learners focus on content rather than delivery, thereby potentially improving comprehension and retention of learning material. However, human imperfections may also raise viewers’ attention, thus aiding the internalization of the learning material. For example, research by Beege et al. (2022) suggests that reducing the distance between the learner and teacher improves learning effects, implying that more approachable delivery may be beneficial.

As discussed above, in the context of learning experience, the content generated by AI tools may have features that affect learning outcomes. A more analytical and descriptive style (Markowitz et al., 2024) may aid the retention of facts. Likewise, narratives that elicit less counterarguing (Chu & Liu, 2023) may help viewers internalize the presented content to a greater extent. However, this effect may be counteracted by human skepticism towards AI-generated content once participants suspect the origin of the scripts to be an AI tool. Moreover, potential advantages in the logical and descriptive assembly of content on the side of AI tools may be counterbalanced by participants’ greater familiarity with narratives with human features and imperfections. Being more used to content crafted by humans, involving small redundancies, logical jumps and errors, participants may be more inclined to closely follow the presentation of human-generated teaching videos compared to AI-generated content that may appear sleek, bland and ‘too perfect’ in comparison.

While both humans and AI have unique strengths with the potential to influence learning outcomes positively, we hypothesize that the benefits of a human connection and content generation outweigh the benefits of AI-generated content in terms of learning outcomes. We test the following hypothesis.

H2. *Learning outcomes from watching teaching videos made by humans are superior to learning outcomes from watching comparable videos generated by AI.*

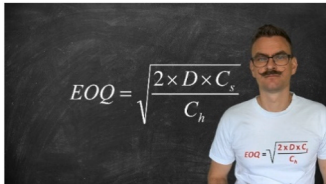
3. Methods

3.1. Production of teaching videos

For our experiment, we produced videos with and without AI. The teaching team produced the first set of videos entirely without AI support, including scripting, recording, curation, and production. These videos were self-produced during the fall semester of 2020 when teaching shifted to online delivery due to the Covid-19 pandemic. The teachers were present in all videos, which is generally preferred vis-à-vis voice-over lectures (Chen & Wu, 2015; van Wermeskerken et al., 2017). We refer to these videos as *human-made* teaching videos. The second set of videos was created during the spring semester of 2023 using a combination of generative AI tools. We refer to these videos as *AI-generated* teaching videos.

For fair comparisons of learning effects, the AI-generated videos cover the same topics as the human-generated videos and are taught by an AI-generated replica of the human teacher. To reduce information and selection bias, we selected four different topics taught by four different teachers originating from different countries. From our collection of 45 teaching videos on production management, we selected the four topics “5S” (a workplace organization method), the theory of inventive problem-solving “TRIZ” (a method to solve technical problems), “Just-in-Time (JIT) and Kanban” (an intra-logistics concept) and the Economic Order Quantity

Table 1
The human-made versus AI-generated teaching videos.

Video content	Human-made teaching video	AI-generated teaching video
5S	 Duration: 3.23 URL: https://youtu.be/5VEKrpPTiel	 Duration: 3.07 URL: https://youtu.be/Q-2X_pbQj80
TRIZ	 Duration: 3.17 URL: https://youtu.be/UrsqA4KV1Ew	 Duration: 3.21 URL: https://youtu.be/4-wmcU4TdNg
JIT and Kanban	 Duration: 2.43 URL: https://youtu.be/h1IDE11UhuY	 Duration: 2.29 URL: https://youtu.be/iHlXrzYvtEU
EOQ	 Duration: 2.54 URL: https://youtu.be/gxsdn7_1r6E	 Duration: 2.46 URL: https://youtu.be/xE0bBV6OLaM

(“EOQ”) (an equation for calculating the optimal amount of goods to purchase for inventory holding). These videos span from quantitative topics (EOQ) to qualitative topics (5S) and from methods (5S and TRIZ) to concepts (JIT and Kanban). To avoid instructor-presence-effects bias, which suggests that learners rate their learning experience more highly when the instructor is present in instructional videos (Beege et al., 2023), we ensured that both the AI-generated and human-made videos had the instructor present for approximately the same duration. The videos ranged from 2 min 29 s to 3 min 23 s. Table 1 gives an overview of the teaching videos.

The approach we employed in creating human-made videos is similar to the production of a wide range of contemporary video content. We aimed for short informative lecture videos, each focusing on one operations management concept. The teaching staff created the script and storyboard, shot the video A-roll in front of a green screen, and produced the video. We used Adobe Creative Suite for the production, primarily Adobe Premiere and Adobe After Effects. For B-rolls and animations, we used both our materials and stock libraries. On average, the scripting took 3 h, the recording of videos took 1 h, and the production took approximately 12–24 h per video. During the Covid-19 fall semester of 2020, we produced 3–5 videos weekly in a makeshift recording studio equipped with standard recording technology costing less than EUR 1000.

The production of AI-generated videos was considerably faster than that of human-generated videos—even when accounting for curation by the video producer and quality assurance by the teaching staff. On average, it only took 3–4 h to create a video, and never

Table 2
Generative AI tools used to develop AI-generated teaching videos.

Name	Functionality	Payment model	URL
ChatGPT 3.5	Text-to-text	Monthly subscription	https://chat.openai.com
Midjourney	Text-to-image	Image-based monthly subscription	https://midjourney.com
Dall-E 2 by OpenAI	Text-to-image	Image pay-as-you-go	https://openai.com/dall-e-2/
Movio (HeyGen)	Text-to-avatar & simple video editor	Minute-based monthly subscription	https://www.heygen.com/avatars?from=moviola

more than 8 h. We acknowledge that our AI-generated videos involve human curation and, thus, are not fully automated.² However, human involvement was minimized as it did not involve the core parts of video production, such as script generation, avatar generation, or suggestions for supporting materials. Human curation only involved promoting content, checking for AI hallucinations, and assembling the AI-generated material.

We started with prompting ChatGPT 3.5 with a relevant subject, i.e., “Can you write a script for an about 5-min-long educational YouTube video about economic order quantity?”. In most cases, ChatGPT could immediately generate suitable text and provide suggestions for visual illustrations. Keeping our intervention to a minimum, we sometimes had to continue prompting ChatGPT to adjust content, tone, and style if necessary. For example, ChatGPT suggested two speakers for one video, but we wanted only one to have a comparable video to the human-made one. Within a few minutes, the script was useable (see Appendix A for an example of how the 5S video was created).

We then fed the AI-generated text into Movio together with one photo of the teacher who taught the corresponding human-generated video. Movio then generated an AI avatar of the human teacher with an AI voice as a basis for the video. Simultaneously, we prompted Midjourney and Dall-E using relevant parts of the script generated by ChatGPT for background illustrations. Midjourney and Dall-E provided different and often unsuitable results, which needed more curation and prompting than ChatGPT. For example, homonyms such as “operations” in operations management resulted in photos of medical surgery. More curation was needed to provide fair competition between AI-generated and human-made videos. However, even with occasional struggles, this part of the video-making process was significantly shorter than searching stock material for a specific image, as we did for the human-generated videos.

Once the AI-generated material was ready, we assembled and edited the video, following a similar procedure as for our human-generated videos. For that purpose, we used Premiere from Adobe again. Table 2 summarizes the generative AI tools used in our study.

3.2. Design of the online experiment

Our study relies on an online experiment using experimental vignettes and a survey (Aguinis et al., 2014). Online experiments offer the advantage of combining the controlled conditions of a lab experiment with the ability to draw conclusions that apply to a broader population when conducted on representative samples (Wulff et al., 2020).

Before conducting the online experiment in July 2023, we pre-registered our hypotheses by submitting them to an academic conference in January 2023.³ We set up the online experiment using MTurk, a crowd-sourcing platform for work tasks offered by Amazon. As illustrated in Fig. 1, we randomly assign participants to two conditions in the experiment: The first condition presents four teaching videos in the sequence human-AI-human-AI. The second condition presents the same four videos in the opposite sequence, AI-human-AI-human. We considered this alteration treatment better than the alternative of providing only AI-generated respectively human-made teaching videos to different participants because it calibrates the learning experience scales for each respondent. The sequence of the videos is unimportant as the topics are independent. The sequence was chosen randomly and kept constant throughout the experiment.

To capture the *learning experiences*, we constructed an index based on nine survey questions that had to be answered after each video (see Table 3). The survey questions (provided in Appendix B) aim to cover essential dimensions of the learning experience associated with the consumption of teaching videos, including the use of visuals, pace, and understandability of the content, among others. We used a 5-point Likert scale ranging from 1 “strongly disagree” to 5 “strongly agree.” From these items, we compute an unweighted average that captures the self-reported quality of participants’ learning experiences. Considering the novelty of the investigated topic, no survey questions suitable for the specific purpose of this paper could be identified. The focus of prior literature on learning experience is the evaluation of courses (e.g., Wilson et al., 1997), not teaching videos. To ensure the robustness of the learning experience index designed by the authors, a battery of robustness checks is reported in the result section.

To measure *learning outcomes*, we implement a final exam with 12 multiple-choice questions. We chose the questions in a way that neither the AI nor human version of the videos was favored, and the necessary answers were contained in both types of videos. The

² At the time of our research, some AI vendors claimed to be able to fully automate video creation from prompts. We tested several of these services but none were remotely close to creating videos that looked and felt similar to the human-made videos—an important condition for our research.

³ The hypotheses and research design were submitted as an extended abstract on January 17, 2023, to the 30th EurOMA conference, July 1–6, 2023, in Leuven, Belgium. EurOMA does not publish abstracts online. The AI-generated videos were made in March–April 2023. The main data collection was conducted in July–August 2023.

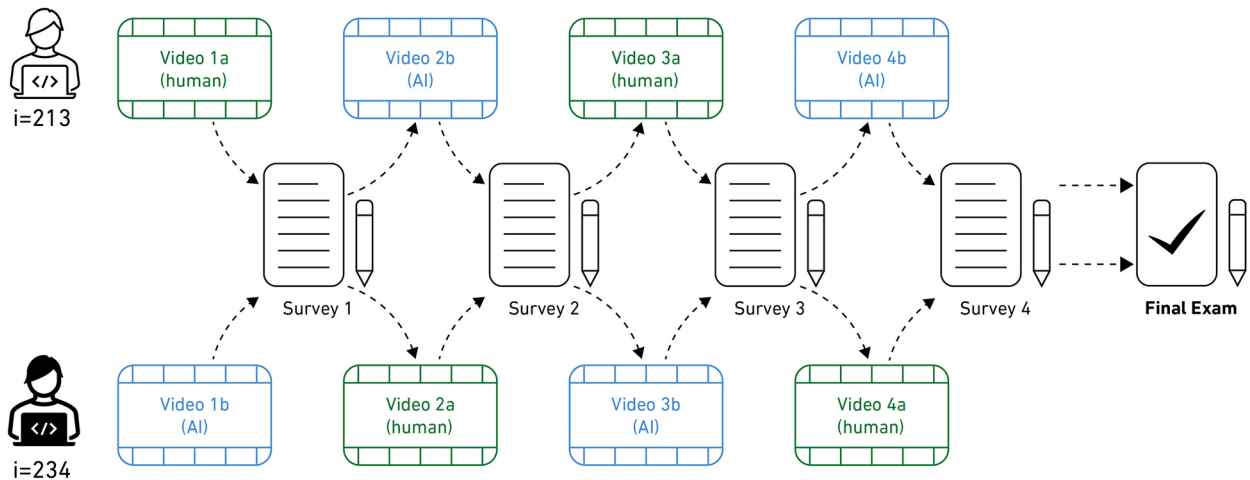


Fig. 1. Experimental research design.

Table 3

Descriptive statistics by human-made vs AI-generated videos.

Variable	Human-made			AI-generated		
	Mean/Mode	Sd.	Sample	Mean/Mode	Sd.	Sample
Learning outcome	1.24	0.82	894	1.23	0.77	894
Learning experience	4.17	0.60	858	4.03	0.75	859
Optimism about AI	4.20	0.76	894	4.20	0.76	894
Topic familiarity	3.05	1.32	894	3.05	1.32	894
Education	3.00	–	894	3.00	–	894
Learning type	3.00	–	894	3.00	–	894

Note(s): The table above showcases the descriptive statistics differentiated by the treatment condition. For each variable, the mean or mode is provided alongside the standard deviation (Sd.), except where the standard deviation is not applicable, as in the case of categorical/ordinal variables (indicated with ‘–’). For continuous variables, the mean is reported. For categorical/ordinal variables, the mode is reported.

exam questions are realistic as they represent typical multiple-choice questions used in production management courses. Furthermore, we implement five attention-check questions throughout the survey and the exam to enhance data quality and exclude poor-effort answers (Hauser et al., 2016). We designed our attention checks following suggestions from Huang et al. (2012). The survey and exam questions were the same for both treatment conditions. Additionally, our survey captures relevant control variables such as educational level, learning type, and previous knowledge about the topics taught in the videos. The details are presented in Appendix B.

3.3. Sample characteristics

Using a random sampling strategy, we recruited US-based participants in this study through MTurk. Volunteer participants were offered a base salary of USD 8.00 for participating in the experiment, which took an average of 28.4 min to complete for the included participants (a maximum of 45 min allowed). We recruited participants for the first condition and, two weeks later, for the second condition. We implemented a mechanism to prevent participants from re-registering for the experiment and receiving both treatments. We sampled laypeople, which ensures a realistic setting for teaching videos about introductory topics such as ours. To control out the influence of incidental prior exposure to the topics, a control variable (*Topic familiarity*) was included.

Our initial sample comprised 472 participants for the first and 483 for the second condition. After eliminating incomplete responses and participants failing the attention checks, we received a final sample of 447 participants, including 213 for the first and 234 for the second condition. This yields a total of 1788 video treatments. This is two and a half times higher than the minimum sample size we calculated to be necessary to detect a small effect, with 80% statistical power, in our planned model.

Table 3 presents the descriptive statistics for human-made versus AI-generated teaching videos across the total sample. The statistics summarize the learning outcome, learning experience, and control variables. The descriptive statistics for individual teaching videos are included in Appendix C.

Most of the survey participants had completed a higher education program (85.9%). Regarding those who had not, 8.3% had completed high school or less, 5.6% had completed technical or vocational training, and the remaining 0.2% had completed another type of education. Regarding learning types, participants' preferences were relatively equally distributed: 27.5% preferred visual learning, 17.5% preferred reading, 36.0% preferred listening, and 19.0% preferred hands-on experiences.

Table 4
Spearman Correlations between continuous and ordinal variables.

	Learning outcome	Learning experience	Optimism about AI	Topic familiarity	Education
Learning outcome	1.00				
Learning experience	−0.02	1.00			
Optimism about AI	−0.03	0.26***	1.00		
Topic familiarity	−0.18***	0.29***	0.16***	1.00	
Education	−0.14***	0.17***	0.06*	0.38***	1.00

Note(s): The table above presents the Spearman correlation coefficients for relationships between various continuous and ordinal variables. Coefficients close to 1 or -1 indicate a strong positive or negative correlation, respectively, while values closer to 0 suggest a weak or no correlation. Significance levels are reported as follows: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$).

To check if the participants enjoyed the AI-generated videos, we use a one-sample t -test to test if the learning experience ratings for AI videos are statistically significantly higher than “3” (“neutral”) for all learning experience items. All tests have been statistically significant at the 5% level. (In the order of the items listing in Table C-1 in Appendix C: (1) $t = 39.06$, $p < 0.001$; (2) $t = 32.35$, $p < 0.001$; (3) $t = 21.71$, $p < 0.001$; (4) $t = 28.78$, $p < 0.001$; (5) $t = 36.70$, $p < 0.001$; (6) 30.14 , $p < 0.001$; (7) 35.10 , $p < 0.001$; (8) $t = 42.62$, $p < 0.001$; (9) $t = 26.27$, $p < 0.001$). This suggests that the AI-generated videos in our experiment are generally well-perceived.

The correlation table is presented in Table 4. We report the Spearman correlation instead of the Pearson correlation to allow for the inclusion of an ordinal control variable (Education). The results indicate that the variables are correlated at a relatively low level. To ensure that multicollinearity was of no concern, we inspected variance inflation factors, which were low for all reported regression models. Our data shows no correlation between learning experience (how much students enjoy a course) and learning outcome (how well the students do on the exam). Unsurprisingly, the level of education (weakly) correlates with the familiarity of the content in the videos (0.38).

4. Results

4.1. Effect of AI-generated videos on learning experience

To test the effect of AI-generated teaching videos versus human-made videos, we use ordinary least squares regression and report three models: Model 1 reports the main effect only, Model 2 the controls only, and Model 3 the full model. Table 5 presents the effects on participants' learning experience.

Table 5, Model 3 shows that the participants' learning experience of the videos was statistically significantly lower for AI-generated videos (-0.16 , $p < 0.001$). This evidence aligns with Hypothesis 1, which says that human-made videos provide a better learning experience than AI-generated videos. However, considering the average score of 4.10, a reduction of *Learning Experience* by -0.16 is relatively small in practical terms, representing a decrease of about 4% compared to the average (or 0.233 standard deviations).

Table 5
Experimental effects on learning experience.

Variables	Learning Experience (self-reported)		
	Model 1 – Effect only	Model 2 – Controls only	Model 3 – Full model
AI-generated [Baseline: Human-made]	−0.15***		−0.16***
Optimism about AI		0.19***	0.19***
Topic familiarity		0.13***	0.13***
Video topic [Baseline: 5S]			
TRIZ		−0.27***	−0.28***
JIT and Kanban		0.03	0.03
EOQ		−0.19***	−0.19***
Educational level [Baseline: High school]			
Technical/vocational training		0.11	0.11
Higher Education		0.30*	0.30*
Other		0.80***	0.79***
Learning type [Baseline: Visual]			
Reading		−0.11	−0.12
Listening		0.03	0.03
Hands-on		−0.12	−0.12
Random effects (Participants)	YES	YES	YES
Constant	4.18***	2.80***	2.89***
N	1733	1717	1717
R^2	0.01	0.22	0.24
R^2 -Adj	0.01	0.22	0.23

Note(s): Significance levels are reported as follows based on Heteroskedasticity-robust Huber-White standard errors: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Comparing Models 1 and 3, the negative correlation we find concerning the AI-generated videos and learning experience is of similar magnitude with and without control variables and carries the same sign, strengthening the evidence. Compared to Model 2, which only includes control variables, Model 3 explains another 2% of the variation in the dependent variable, as evidenced by the R^2 score. With an R^2 score of 0.24, Model 3 explains about one-quarter of the overall variance in the dependent variable. Compared to this, the model explains a higher share of the between-subject variance (Between-subject $R^2 = 0.30$), and a lower share of within-subject variance (Within-subject $R^2 = 0.13$).

4.2. Effect of control variables on learning experience

We also observe differences in learning experience depending on the video topic (independent of AI-generated or human-made). The videos on the topic *5S* scored a better learning experience than the videos on *TRIZ* ($-0.28, p < 0.001$) and *EOQ* ($-0.19, p < 0.001$) but approximately equal to the video on *JIT and Kanban* ($0.03, p = 0.271$). Furthermore, we find that participants with higher education ($0.30, p = 0.01$) or “other” education ($0.79, p < 0.001$) had a higher learning experience. We also found that the learning experience of the educational videos differed with the respondents’ learning type. As expected, compared to respondents preferring visual learning, respondents with a predominantly reading-oriented learning type had a poorer learning experience ($-0.12, p = 0.79$). The difference to respondents preferring hands-on learning is insignificant ($-0.12, p = 0.10$).

4.3. Effect of AI-generated videos on learning outcomes

Table 6 presents the effects on participants’ learning outcomes—their exam scores. As evident in Table 6, Model 6, the influence of AI-generated content on participants’ learning—is marginal, exhibiting a decrease of -0.01 , which is statistically insignificant ($p = 0.69$). Juxtaposing Models 4 and 6, this finding remains consistent across models. Comparing Model 5 and Model 6 in terms of their R^2 value, we find that including the explanatory variable *AI generation* does not increase the R^2 ($\Delta R^2 = 0.00$), implying that it does not elevate the explanatory capability of the full model beyond that of Model 5, including only controls. The overall R^2 of Model 6 is 0.17, which is similar to the within-subject and between-subject R^2 scores (0.18, respectively 0.16). Our findings contradict Hypothesis 2, which we therefore reject. We do not find that human-made videos are more effective teaching tools than AI-generated videos.

4.4. Effect of control variables on learning outcomes

Again, we find that control variables are linked to variations in the outcome variable. Compared to videos on the topic of *5S*, participants’ scores were lower for videos on *TRIZ* ($-0.18, p < 0.001$), considerably lower for *JIT and Kanban* ($-0.72, p < 0.001$), and also lower for *EOQ* ($-0.46, p < 0.001$). Considering the educational level (using “High school or lower” as the reference category), somewhat surprisingly, participants with a higher education qualification demonstrated marginally insignificantly lower exam scores ($-0.14, p = 0.06$). Also, participants with “Other” education scored lower on the exam than participants in the reference category ($-0.81, p < 0.001$). Turning to the learning type (using visual learners as the baseline), we find that participants with a preference for

Table 6
Experimental effects on learning outcome.

Variables	Learning Outcome (exam score)		
	Model 4 – Effect only	Model 5 – Controls only	Model 6 – Full model
AI-generated [Baseline: Human-made]	−0.02		−0.01
Optimism about AI		−0.02	−0.02
Topic familiarity		−0.08***	−0.08***
Video topic [Baseline: 5S]			
TRIZ		−0.18***	−0.18***
JIT & Kanban		−0.72***	−0.72***
EOQ		−0.46***	−0.46***
Educational level [Baseline: High school]			
Technical/vocational training		0.01	−0.02
Higher Education		−0.15	−0.14
Other		0.81***	−0.81***
Learning type [Baseline: Visual]			
Reading		−0.02	−0.02
Listening		−0.16**	−0.16**
Hands-on		0.10	0.10
Random effects (Participants)	YES	YES	YES
Constant	1.24***	2.05***	2.06***
N	1788	1788	1788
R^2	0.00	0.17	0.17
R^2 -adj	0.00	0.17	0.17

Note(s): Significance levels are reported as follows based on Heteroskedasticity-robust Huber-White standard errors: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

learning by listening showed significantly lower scores ($0.16, p = 0.01$). No significant difference was observed for participants who prefer hands-on learning or reading.

4.5. Robustness checks

We conduct a battery of checks to assess the robustness of our findings. First, we strengthen the requirements for sample inclusion. To be included in the sample for our main analysis reported above, respondents needed to have selected answer options “4” or “5” for the self-assessed attention check (“I paid attention when answering the questions in this survey”) and an ethical promise (“I felt comfortable providing my honest opinions and experiences in this survey”). Only including respondents who scored “5” on either question reduces the sample to 1044 for learning experience and 1084 for learning outcome. Analyzing the constrained samples, we find nearly identical results with unchanged levels of statistical significance concerning our hypothesis tests.

Second, we further constrain our sample based on respondents’ responses to the statement, “I felt rushed while completing this experiment.” Excluding neutral answers (“3”) and responses signaling agreement (“4” or “5”), the sample sizes are reduced to 842 for learning experience and 868 for learning outcome. Refitting our full models in the further-reduced sample yields consistent results with unchanged significance levels. The same holds if further constraining the sample to all respondents but those who strongly disagreed with the statement (i.e., responding with “1”), which leaves a sample size of 670, respectively 692 respondents.

Third, we assess the results of our full models when refitting them with fixed effects instead of random effects. We find no notable differences between the fixed effects and random effects model, which implies that the between- and within-effects are congruent (Ketokivi et al., 2021) and further testifies to the robustness of our findings.

Fourth, considering the time required to respond to the survey leaves our results unchanged. Controlling for the time (in seconds) required to answer the survey yields nearly identical results with unchanged interpretations and levels of significance concerning our hypotheses. Our findings also remain robust when excluding respondents with response times exceeding the 10th, 25th, or 50th percentile. With a view to the continuum of resistance model (Lin & Schaeffer, 1995), this allows a sound judgment about non-response bias due to the similarity between late responses and non-responses.

We also perform a correction of our perceptual dependent variable *learning experience* using a common method bias marker variable, which is available for 736 respondents in the sample. For this purpose, we included the statement, “I think clicking on the ‘like’ or ‘dislike’ button of videos I watch on the internet is important,” which is theoretically unrelated to any of the questions included in the survey⁴ but is subject to similar sources of bias (esp., acquiescence bias, central tendency bias, social desirability bias) also considering that it is in the same perceived topical area (online videos). Like the items used to construct the dependent variable, the statement for the common method bias marker variable also incites participants to self-reflect about a certain preference. We regress the dependent variable on the theoretically unrelated marker variable and calculate the residual, which represents the scores for the dependent variable, adjusted for the method bias common to the two theoretically unrelated variables. Replacing the scores of the dependent variable with these scores for the available sample and refitting the full model, we find an unchanged coefficient ($-0.16, p < 0.001$).

Moreover, we conducted a series of robustness checks to ensure that our results were not affected by the assumption of approximate equidistance between the different points of the Likert scale items. We computed measures of relative frequency of (1) strong agreement, (2) strong agreement or agreement, (3) strong disagreement, as well as (4) disagreement or strong disagreement across the nine items used for learning experience (e.g., such that a strong agreement to three out of nine items would result in a score of 0.33; and strong agreement to all nine items would yield a score of 1.00). These measures do not assume equidistance as they either utilize only the extreme points of scales or binarize agreement (disagreement) by grouping different levels of the scales. Using these alternative measures of learning experience yields highly significant and consistent results concerning the variable *AI generated* ($p < 0.001$ in each regression); affirming the robustness of our results.

We also run a regression on *learning experience* that does not require the computation of an index, or a factor based on multiple items. In this regression, our chosen dependent variable is in agreement with the item “If I studied this subject, I would recommend this video to others,” as recommendability of a learning experience arguably implies a positive perception of it. Using this single item as our dependent variable yields consistent results with unchanged statistical significance.

As an additional robustness check, we also refitted our regression using bootstrapped standard errors (with 10,000 replications each) to assess whether potential violations of the assumption of normality in the regression residuals affected our conclusions concerning our hypotheses. These results confirmed our original findings as the effect of AI-generated videos on learning outcomes remains statistically insignificant ($p = 0.685$), whereas the effect on learning experience remained highly statistically significant ($p < 0.001$).

Lastly, we refit our model for *learning experience* not using an equally weighted index but the output of an exploratory factor analysis using the nine survey items relating to each video as input. We find a Cronbach alpha score greater than 0.70 and items for each video loading on a common factor with all factor loadings greater than 0.50. We predict the underlying latent factor using the regression method and use them as dependent variables in our regression. We find a nearly unchanged coefficient of the variable *AI generated* significantly at the same level as in the main analysis ($-0.23, p < 0.001$).

We note that common method bias is no major concern in our study design because the dependent variable of the first regression is

⁴ The videos included in the survey did neither have a like or dislike button. The common method bias marker variable was included at the end of the survey to avoid anchoring the attention of respondents.

a perceptual measure (*Learning Experience*) while the explanatory variables are objective measures (*AI-generated*, *Video topic*, *Educational level*, *Learning type*). In the second regression, also the dependent variable (*Learning Outcome*) is an objective variable, scored on the basis of participants' answers to the presented exam questions. In neither of the two regressions are the findings sensitive to the inclusion of the two perceptual control variables (*Optimism about AI*, *Topic familiarity*).

4.6. Subsample analyses

We also perform subsample analyses in which we individually consider the four operations management concepts that our teaching videos cover: *5S*, *TRIZ*, *JIT and Kanban*, and *EOQ*. These analyses provide additional insights into the variance of the respondents' answers contingent on video topics. In our main models presented above, variance relating to individual video topics has been controlled out by including a respective variable. Contrary to our main models, the video-specific models cannot account for participant-specific random effects, as each participant has only been exposed to either a human- or an AI-generated video concerning a given topic.

As regards learning experience (see Table D-1 in Appendix D), for which our main analysis found an overall small negative effect, our subsample analyses show notable differences across topics. For two of the four video topics (*5S* and *JIT and Kanban*), we find positive effects of AI generation on the learning experience, significant at the 1%-level (with coefficient sizes of 0.17, respectively 0.15). For the other two video topics (*TRIZ* and *EOQ*), we find negative effects of AI generation, also significant at the 1%-level, yet with effect sizes approximately twice as large (-0.30 , respectively -0.36). In other words, for half of the taught topics, participants mildly preferred AI-generated videos. For the video topics for which participants did not prefer AI-generated videos, they relatively more strongly disliked the AI-generated content.

Concerning learning outcomes, as measured by exam scores (see Table D-2), for which our main analysis found no significant effect of AI-generated videos, we find no effects of the variable for two video topics (*5S* and *JIT and Kanban*) and opposed effect directions for the other two video topics. For videos on *TRIZ*, we find a negative effect (-0.14) yet marginally insignificant at the 5%-level, while for the *EOQ*-themed videos, we find a similarly sized positive effect (0.16) of AI generation, statistically significant at the 5%-level.

5. Discussion

Our study investigated the effectiveness of human-made versus AI-generated teaching videos in management education. While examining the participants' learning experiences, we observed a statistically significant difference in favor of human-made videos, which gives support in favor of hypothesis 1. However, the size of the associated effect and the amount of variance it explains in the dependent variable is relatively small. This implies the relative effectiveness of AI-generated teaching videos in our experiment, which scored lower in terms of learning experience than human-made videos but without lacking far behind. Considering the rapid progress of generative AI, it is likely that this small gap can soon be bridged and AI-generated content can provide learning experiences to students that are similar to traditional human-made content. We also found that AI-generated videos were well-received by the participants, and, to our surprise, there was no statistically significant evidence in favor of human-made or AI-generated videos regarding the learning outcome, leading to the rejection of hypothesis 2.

Our results can find explanations in the learning science literature that emphasizes the importance of human teachers for students' learning experience, such as mimics (Schneider et al., 2022) or professional appearance (Beege et al., 2022), which helps students relate to the teacher. However, the non-significant differences between human-made versus AI-generated videos regarding learning outcomes yield interesting avenues for future research. One possible explanation for the absence of a difference is the relative simplicity of the concepts taught in the videos. We restrained the videos to clear and well-defined single concepts and produced short videos of only a few minutes, also known as micro-teaching (cf. Nguyen et al., 2023). The advantage of this design is that we could create human-made and AI-generated videos that capture almost the exact same content and run an online experiment with several content videos resource-efficiently. The drawback is that we cannot tell if the results would hold for more complex content taught during longer teaching videos.

Our study contributes to the existing literature on AI-supported education (e.g., Darvishi et al., 2024; Lim et al., 2023; Pataranutaporn et al., 2021; Pinkwart et al., 2020) by empirically investigating the impact of AI-generated versus human-generated teaching videos in operations management. The results highlight the need for educators to exercise caution when utilizing AI-generated videos in the syllabus and suggest that human-generated videos may still hold an advantage in promoting student learning due to offering a better experience. However, research on AI avatars shows that students can now recognize emotions in AI avatars (Lawson et al., 2021), hinting that the human advantage is being reduced. Furthermore, our results show that AI-generated content is on par with human-made videos in terms of learning outcomes and is much quicker and cheaper to produce, which bears the potential to make education more accessible using new technologies.

Although not tested directly in this study, the middle way, where humans work with AI to create teaching content, is a potentially promising path forward. In a blended model, human educators can collaborate with AI systems to design and curate content, leveraging the strengths of both parties. This collaboration can result in highly engaging and adaptive learning experiences that cater to individual needs while preserving the human connection, which is vital for effective teaching and learning. Ultimately, the integration of human-made and AI-generated educational videos should have the potential to enhance students' learning experiences and outcomes by providing personalized, engaging, and accessible resources that meet diverse learning needs.

6. Limitations and future research

We acknowledge several limitations of this study, which could have influenced the findings. The sample size and population demographics may not represent the broader population, potentially limiting the generalizability of the results. Factors such as the appearance of instructors and avatars may bias learner's perceptions to some extent (e.g., Beege et al., 2022). It is important to acknowledge that learning experiences are multifaceted and influenced by various factors such as individual learning preferences, prior knowledge, and motivation.

Furthermore, the specific subject matter and content used in the study could also have impacted the findings, as different topics may yield different outcomes in terms of learning experiences. We note that we have only tested very short and concise teaching videos covering four topics in operations management. The "consensus" on these basic topics is relatively clear and non-controversial. It could be that fields with clear definitions and rules, such as mathematics and engineering, are easier to teach via AI-generated content than topics such as philosophy and politics.

It is also clear that the current AI-generated videos have not yet reached their full potential in offering unique learning experiences (Lim et al., 2023). Our AI videos were limited by the AI technology available during spring 2023. As AI technology continues to evolve, we may see more advanced AI-generated content that surpasses human-made content in terms of adaptability, personalization, and engagement. Perhaps the quality of the content delivery, whether created by humans or AI, is more important for the learning experience than the source of the content itself. Questions like this may arguably be more difficult to answer in an experimental setting like ours. Thus, future research could expand on such questions using qualitative research methods such as interviews (Vallis et al., 2023).

Furthermore, our chosen cross-subject experimental design has limitations for learning-related research that is important to acknowledge. Individual differences existing between study participants may affect the learning experience and outcomes. Many of our variables are self-reported by participants rather than objectively measured. We sought to reduce concerns by introducing participant-specific control variables and modeling participant-specific unobserved heterogeneity by including random effects in our panel regression. Yet, our results need to be interpreted with the recognition that participants' learning processes may be unique in complex ways, introducing unsystematic noise into the analyses presented.

Future research could consider experiments with different comparisons than presented in this study, for example, comparing videos created by research assistants based on instructions comparable to the prompts inputted into AI tools. In this study, we compared AI-generated and AI-delivered content with human-generated and human-made content. Holding the delivery mode constant, future studies could seek to investigate whether viewers can detect whether the content was generated by humans or AI and which video they prefer. We expect that with progress in generative AI, viewers will increasingly prefer AI-generated content.

Finally, we point out potential limitations concerning data quality from our online experiment on Amazon MTurk. Although we took steps to ensure a high-quality experimental design, including a series of robustness tests, we cannot entirely rule out the opportunity that participants took the experiment lightly. Unlike students following a course, online participants do not have the same incentives to learn the content and score well on an exam. Future research could conduct similar experiments with students as participants.

7. Conclusion

We present one of the first studies exploring the effects of AI-generated teaching material in business education. In conclusion, this study found non-significant effects in learning outcomes between human-made and AI-generated videos but significant, albeit small differences in learning experiences in favor of human-made videos. As AI technology continues to advance, it will be critical for researchers and educators to explore new ways of leveraging AI-generated content to enhance learning experiences while also remaining mindful of the potential limitations and ethical considerations associated with its use. Our paper gives an example of what the future may hold for teaching and learning. As we type up the conclusions of this study, OpenAI recently launched its text-to-video tool Sora. We encourage other scholars to continue investigating the potential benefits of such advancements in AI-generated content for learning experiences.

CRedit authorship contribution statement

Torbjørn Netland: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Conceptualization. **Oliver von Dzengelevski:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation. **Katalin Tesch:** Visualization, Software, Resources, Data curation. **Daniel Kwasnitschka:** Writing – review & editing, Software, Methodology, Formal analysis, Data curation.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used several generative AI tools to generate four educational videos, as fully described in the Materials and Methods section. During the writing process, the authors used MS Word's spellcheck, Grammarly, and ChatGPT 4.0 to improve readability and language. The paper is written by the authors, who take full responsibility for its content.

Data availability

Data will be made available on request.

Appendix A. Example of AI video generation for the 5S video

This section provides an example of how the AI videos were generated using the 5S video. The full scripts and details of all videos are available on request to the corresponding author.

We used three generative AI tools: (1) ChatGPT for the script, (2) Movio for the video with a teaching avatar in front of a green screen, and (3) DALL-E 2 and Midjourney for background illustrations. We used Adobe Premier to curate the video and illustration materials into a finished video.

Step 1: Script generation with ChatGPT 3.5

For the 5S video, we needed two prompts before we had a script, as shown below:

1st prompt: Can you make a script for a 6–7 min educational video on 5S?

ChatGPT 3.5:

Here's an example of a script for a 6–7 minute educational video on 5S:

[Opening shot of a cluttered and disorganized work area. Cut to the host standing in front of a whiteboard or screen, introducing the topic of 5S.]

Host: Welcome to our video on 5S. Today, we're going to explore how organizations can use the 5S method to improve efficiency, reduce waste, and increase productivity in the workplace.

[Cut to animation or examples of different types of work environments, such as a manufacturing plant, an office, or a hospital.]

(...) [parts removed for sample brevity]

[Closing shot of the host summarizing the key points of the video, and encouraging viewers to learn more about 5S.]

Host: So there you have it, a brief overview of the 5S method. I hope you found it informative.

2nd prompt: Do you know the original, Japanese terms for the 5S? Can you write the script using those?

ChatGPT 3.5:

Sure! The original Japanese terms for the 5S are:

(...) [parts removed for sample brevity]

Host: Welcome to our video on 5S. Today, we're going to explore how organizations can use the 5S method, Seiri, Seiton, Seiso, Seiketsu, and Shitsuke, to improve efficiency, reduce waste, and increase productivity in the workplace.

(...) [parts removed for sample brevity]

[Closing shot of the host summarizing the key points of the video, and encouraging viewers to learn more about 5S.]

Host: So there you have it, a brief overview of the 5S method, Seiri, Seiton, Seiso, Seiketsu, and Shitsuke. I hope you found it informative. Remember, implementing 5S can have a significant impact on organizational performance and efficiency. If you want to learn more, be sure to check out our other videos, and stay tuned for more on this topic.

End of the script. Please note that this is an example, and you may need to adjust the script to suit your specific needs and requirements.

As for B-roll footage, you can use a combination of footage such as.

- Time-lapse footage of a work area before and after implementing 5S
- Footage of employees sorting, organizing, cleaning, and maintaining the work area
- Footage of employees training and conducting audits to ensure compliance with 5S
- Graphs and charts that visually represent the improvement in efficiency and productivity after implementing 5S.

Step 2: Video creation with avatar

We uploaded a photo of the 5S teacher (D. Kwasnitschka) to Movio together with the script of the narration suggested by ChatGPT 3.5. Movio generated a video with a talking AI-generated synthetic teacher resembling the original teacher.

Step 3: Generation of background illustrations

For generating the illustration, we started with the prompts included in the scripts generated by ChatGPT. For example, as shown above, we prompted “a cluttered and disorganized work area” to generate an illustration for the 5S video. However, more prompts were needed for illustrations than for scripts. Gradually adding more details to the prompt (e.g., “make it a work area in manufacturing”) aided us in getting a useable image. For our videos, we generally found Midjourney to be a better tool than Dall-E 2.

Step 4: Video production and post-processing in Adobe Premier

Finally, the video material and illustrations were placed during production and post-processing in Adobe Premier. This process was manual and done by a video producer on the author team of this article. Finally, the video was uploaded to a non-public page on YouTube.

Appendix B. Online vignette experiment with survey and exam questions

Dear participant,

Welcome to our online experiment on the learning effects of teaching videos. This experiment is anonymous. Your data will only be handled by the research team and not shared with third parties. We do not know your identity.

You will watch four very short teaching videos and be asked to rate your learning experience after each of them by scoring ten short statements on a scale ranging from “Strongly disagree” to “Strongly agree.” Note that some of the videos are developed and generated partly by using Artificial Intelligence (AI) technology.

You will also be asked attention check questions, which you need to answer correctly to be able to proceed and **earn your base salary**. Hence, it is important to pay full attention to all content. If you fail the attention check questions, we reserve the right to reject payment.

At the end, you will also conduct a short exam that tests if you learned the concepts you were taught. The exam consists of 12 multiple-choice questions with four answer alternatives.

You pass through this teaching experiment step-by-step and cannot go back. This means you will watch each video **only once**. The entire experiment takes less than 1 h to complete.

Thank you for participating in our online experiment and for taking it seriously. We value your participation!

The Research Team.

Part 1 Videos and learning experience questions

[The below questions were asked after each of the four videos]

Please rate the following statements about video no. [1, 2, 3, 4].

[1 – Strongly disagree, 2 – Disagree, 3 – Neither agree nor disagree, 4 – Agree – 5 Strongly agree]

1. The content of the video was easy to understand.
2. The lecturer provided enough details.
3. The video was engaging.
4. The video was interesting.
5. The lecturer used relevant examples to support the content.
6. The letter after A in the alphabet is B.
7. The visuals used in the video were helpful.
8. The video was well-paced and structured.
9. The video was informative.
10. If I studied this subject, I would recommend this video to others.

Part 2 Controls

[Asked after the learning experience questions of the last video]

What is the highest education degree you have completed?

1. High school or lower
2. Technical/vocational training
3. Higher Education (e.g., Bachelor's, Master's, Doctorate)
4. Other

When learning new information, I prefer to ...

- a. See it visually presented, such as in diagrams or pictures.

- b. Read about it in books, articles, or other written materials.
- c. Hear and listen to information, such as lectures, podcasts, and conversations.
- d. Have hands-on activities and physical experiences, such as experiments, field trips, and simulations.

Please rate the following statements:

[1 – Strongly disagree, 2 – Disagree, 3 – Neither agree nor disagree, 4 – Agree – 5 Strongly agree]

- 1. I was already well familiar with the lecture content before this experiment.
- 2. I paid attention when answering the questions in this survey.
- 3. I felt rushed while completing this experiment.
- 4. I felt comfortable providing my honest opinions and experiences in this survey.
- 5. I believe that artificial intelligence (AI) has the potential to improve many aspects of our lives.
- 6. I think clicking on the “like” or “dislike” button of videos I watch on the internet is important [common method bias marker variable]

Part 3 Exam questions (learning outcome)

[Asked at the end]

Welcome to your final exam, which consists of 12 short multiple-choice questions. There is always only one correct answer. It is important to us that you do your best on the exam and read all questions and answers. Please note that there is a time limit. Do not navigate away from this page as the exam goes on. Note that you cannot navigate backward in this survey.

- 1. What is kanban?
 - a. A method for continuous improvement
 - b. A method for material replenishment
 - c. A method for inventive problem-solving
 - d. A method for calculating inventory levels
- 2. What is the main purpose of 5S?
 - a. Keep an organized workplace
 - b. Never pass on defective products
 - c. Improve problem-solving capabilities
 - d. Reduce the amount of scrap
- 3. Who is credited with the invention of TRIZ?
 - a. Henrik Neubauer
 - b. Taiichi Ohno
 - c. Jim Wormack
 - d. Genrikh Altshuller
- 4. What does the EOQ formula calculate?
 - a. The optimal time between two orders
 - b. The optimal level of inventory
 - c. The optimal number of kanban cards to be used
 - d. The optimal number of items to be ordered
- 5. What is NOT a problem with “push production”?
 - a. It builds unnecessary inventory
 - b. It hides quality problems
 - c. It increases the number of transports needed
 - d. It increases the time a part spends in production
- 6. What does 5S stand for?
 - a. Synchronized & standardized storage system solution
 - b. See, sense, smell, sight, and sound
 - c. Seiri, seiton, seisō, seiketsu, and shitsuke
 - d. Safety, sustainability, simplicity, standardization, and system
- 7. What is the key idea in TRIZ?
 - a. Start with a blank-sheet approach to innovation
 - b. Reuse innovative elements from previous innovations
 - c. Engage in problem-solving at the production shop floor
 - d. Use an agile methodology
- 8. What happens if the ordering cost is cut in half (cheaper to put in an order)?
 - a. The EOQ goes down
 - b. The EOQ goes up
 - c. The EOQ remains unchanged

- d. The EOQ is unrelated to ordering costs
9. What does “pull production” mean?
 - a. All parts are moved as soon as they are finished in a process
 - b. Parts are only delivered when they are needed
 - c. A process produces a large amount of parts to benefit from economy of scale
 - d. Operators pull a cord if they need help with production
10. What is NOT a direct benefit of 5S?
 - a. Less searching for tools and parts
 - b. A clean and orderly workstation
 - c. Less clutter and tools not needed
 - d. Improves leadership capabilities
11. According to TRIZ, what represents a constraint in innovation processes?
 - a. Creativity
 - b. Capabilities
 - c. Contradictions
 - d. Cleverness
12. JIT was developed in Japan. What is the capital of Japan? [Attention marker question]
 - a. Toyota
 - b. Tokyo
 - c. Inventory
 - d. Kanban
13. What is NOT affecting the economic order quantity (EOQ)?
 - a. Inventory holding cost
 - b. Ordering cost
 - c. Market demand
 - d. Market competition

Your unique completion code is: [XXXXX]

Copy this value to paste into MTurk.

When you have copied this ID, please click the next button to submit your survey.

Appendix C. Descriptive statistics for individual variables

Table C-1

AI-generated versus human-made teaching videos

	Video 1 (5S)			Video 2 (TRIZ)			Video 3 (JIT and Kanban)			Video 4 (EOQ)		
	AI	Human	<i>t</i> -statistic	AI	Human	<i>t</i> -statistic	AI	Human	<i>t</i> -statistic	AI	Human	<i>t</i> -statistic
Learning Experience												
(1) The content of the video was easy to understand	4.47	4.06	5.82***	3.81	4.25	−5.16***	4.45	4.17	3.89***	3.88	4.35	−5.38***
(2) The video provided enough details.	4.38	4.07	4.26***	3.60	4.16	−5.68***	4.30	4.21	1.26	3.84	4.26	−5.06***
(3) The video was engaging	4.12	3.96	1.76	3.43	4.01	−5.64***	4.14	3.95	2.28*	3.50	4.13	−6.32***
(4) The video was interesting	4.36	4.29	0.92	3.66	4.13	−5.04***	4.37	4.13	3.04**	3.66	4.33	−6.77***
(5) The lecturer used relevant examples to support the content	4.16	4.12	0.63	3.75	4.08	−3.95***	4.29	4.26	0.42	3.87	4.20	−4.04***
(6) The visuals used in the video were helpful	4.21	4.08	1.91	3.67	4.17	−5.24***	4.28	4.22	0.83	3.67	4.22	−6.38***
(7) The video was well-paced and structured	4.26	4.14	1.54	3.82	4.15	−3.82***	4.26	4.15	1.43	3.87	4.29	−4.89***
(8) The video was informative	4.32	4.30	0.32	3.94	4.30	−4.64***	4.33	4.20	2.02*	4.02	4.36	−4.45***
(9) If I studied this subject, I would recommend this video to others	4.35	4.02	4.09***	3.54	4.22	−6.33***	4.36	4.12	3.11**	3.57	4.26	−7.09***
Learning Outcome												
Exam results	1.51	1.65	−2.53*	1.35	1.44	−1.36	0.79	0.93	−1.95	1.29	0.97	3.99***

Note(s): The respondents were randomly selected into two treatment groups: Group 1 got the sequential 1–4 teaching videos in the following sequence AI – Human – AI – Human. Group 2 got the videos as Human – AI – Human – AI. T-tests are provided and significance levels for two-sided hypotheses tests are reported as follows: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Appendix D. Subsample analyses

Table D-1

Regressions on learning experience for individual video topics

Variables	Model A1 – 5S	Model B1 – TRIZ	Model C1 – JIT & Kanban	Model D1 – EOQ
AI generated	0.17**	−0.30***	0.15**	−0.36***
Optimism about AI	0.16***	0.22***	0.20***	0.19***
Topic familiarity	0.04*	0.19***	0.07**	0.18***
Educational level [Baseline: High school]				
Technical/vocational training	0.27	0.14	−0.22	0.20
Higher Education	0.21	0.42*	−0.09	0.46**
Other	0.17	1.15***	0.12	0.99***
Learning type [Baseline: Visual]				
Reading	−0.15	−0.09	−0.02	−0.24*
Listening	−0.12	0.12	−0.01	0.02
Hands-on	−0.06	−0.21	0.01	−0.19
Constant	3.21***	2.23***	3.19***	2.50***
N	428	431	428	430
R ²	0.12	0.35	0.15	0.37
R ² -adj	0.11	0.34	0.13	0.36

Note(s): Significance levels are reported as follows based on Heteroskedasticity-robust Huber-White standard errors: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table D-2

Regressions on learning outcome (exam scores) for individual video topics

Variables	Model A2 – 5S	Model B2 – TRIZ	Model C2 – JIT & Kanban	Model D2 – EOQ
AI generated	−0.09	−0.14	−0.02	0.16*
Optimism about AI	−0.07	0.03	−0.04	0.00
Topic familiarity	−0.07**	−0.02	−0.06*	−0.16***
Educational level [Baseline: High school]				
Technical/vocational training	−0.11	0.12	−0.02	0.08
Higher Education	−0.13	0.04	−0.30*	−0.14
Other	0.15	−0.45**	−1.25***	−1.59***
Learning type [Baseline: Visual]				
Reading	−0.11	0.03	0.13	−0.11
Listening	−0.09	−0.08	−0.19**	−0.25*
Hands-on	−0.01	0.23*	0.15	0.03
Constant	2.29***	1.34***	1.48***	1.76***
N	447	447	447	447
R ²	0.07	0.03	0.10	0.14
R ² -adj	0.05	0.01	0.08	0.12

Note(s): Significance levels are reported as follows based on Heteroskedasticity-robust Huber-White standard errors: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

References

- Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, 17(4), 351–371.
- Alam, A. (2021). Should robots replace teachers? Mobilisation of AI and learning analytics in education. In *2021 International Conference on Advances in Computing, Communication, and Control (ICAC3)* (pp. 1–12). IEEE.
- Allan, J. (1996). Learning outcomes in higher education. *Studies in Higher Education*, 21(1), 93–108.
- Alon, L., Sung, S., Cho, J., & Kizilcec, R. F. (2023). From emergency to sustainable online learning: Changes and disparities in undergraduate course grades and experiences in the context of COVID-19. *Computers & Education*, 203, Article 104870.
- Anders, G., Buder, J., Merkt, M., Egger, E., & Huff, M. (2024). Associations between mind wandering, viewer interactions, and the meaningful structure of educational videos. *Computers & Education*, 212, Article 104996.
- Beege, M., Kriegelstein, F., & Arnold, C. (2022). How instructors influence learning with instructional videos - the importance of professional appearance and communication. *Computers & Education*, 185, Article 104531.
- Beege, M., Schroeder, N. L., Heidig, S., Rey, G. D., & Schneider, S. (2023). The instructor presence effect and its moderators in instructional video: A series of meta-analyses. *Educational Research Review*, 100564.
- Chen, C.-M., & Wu, C.-H. (2015). Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance. *Computers & Education*, 80, 108–121.
- Chu, E., Dunn, J., Roy, D., Sands, G., & Stevens, R. (2017). *AI in storytelling: Machines as cocreators*. McKinsey & Company Media & Entertainment.
- Chu, H., & Liu, S. (2023). *C. An AI tell good stories? Narrative Transportation and Persuasion with ChatGPT*.
- Dao, X.-Q., Le, N.-B., & Nguyen, T.-M.-T. (2021). AI-powered moocs: Video lecture generation. In *2021 3rd international conference on image, video and signal processing* (pp. 95–102).
- Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education*, 210, Article 104967.

- Hauser, D. J., & Schwarz, N. (2016). Attentive Turks: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48, 400–407.
- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27, 99–114.
- Jeon, J., Lee, S., & Choe, H. (2023). Beyond ChatGPT: A conceptual framework and systematic review of speech-recognition chatbots for language learning. *Computers & Education*, 206, Article 104898.
- Ketokivi, M., Bromiley, P., & Awaysheh, A. (2021). Making theoretically informed choices in specifying panel-data models. *Production and Operations Management*, 30(7), 2069–2076.
- Lawson, A. P., Mayer, R. E., Adamo-Villani, N., Benes, B., Lei, X., & Cheng, J. (2021). Do learners recognize and relate to the emotions displayed by virtual instructors? *International Journal of Artificial Intelligence in Education*, 31(1), 134–153.
- Leddo, J., & Garg, K. (2021). Comparing the effectiveness of AI-powered educational software to human teachers. *International Journal of Social Science and Economic Research*, 6(3), 953–963.
- Leiker, D., Finnigan, S., Gyllen, A. R., & Cukurova, M. (2023a). Prototyping the use of Large Language Models (LLMs) for adult learning content creation at scale. *arXiv preprint arXiv:2306.01815*.
- Leiker, D., Gyllen, A. R., Eldesouky, I., & Cukurova, M. (2023b). Generative AI for learning: Investigating the potential of learning videos with synthetic virtual instructors. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (pp. 523–529). Cham: Springer Nature Switzerland.
- Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *International Journal of Management in Education*, 21(2), Article 100790.
- Lin, I.-F., & Schaeffer, N. C. (1995). Using survey participants to estimate the impact of nonparticipation. *Public Opinion Quarterly*, 59(2), 236–258.
- Liu, M., Zhang, L. J., & Biebricher, C. (2024). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education*, 211, Article 104977.
- Markowitz, D. M., Hancock, J. T., & Bailenson, J. N. (2024). Linguistic markers of inherently false AI communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, 43(1), 63–82.
- Nehari, M., & Bender, H. (1978). Meaningfulness of a learning experience: A measure for educational outcomes in higher education. *Higher Education*, 7(1), 1–11.
- Nguyen, H., & Diederich, M. (2023). Facilitating knowledge construction in informal learning: A study of TikTok scientific, educational videos. *Computers & Education*, 205, Article 104896.
- Orús, C., Barlés, M. J., Belanche, D., Casaló, L., Fraj, E., & Gurrea, R. (2016). The effects of learner-generated videos for YouTube on learning outcomes and satisfaction. *Computers & Education*, 95, 254–269.
- Pataranutaporn, P., Danry, V., Leong, J., Punpongsonan, P., Novy, D., Maes, P., & Sra, M. (2021). AI-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12), 1013–1022.
- Pi, Z., Deng, L., Wang, X., Guo, P., Xu, T., & Zhou, Y. (2022). The influences of a virtual instructor's voice and appearance on learning from video lectures. *Journal of Computer Assisted Learning*, 38(6), 1703–1713.
- Pinkwart, N., & Liu, S. (2020). *Artificial intelligence supported educational technologies*. Springer.
- Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, 10, 25494–25513.
- Ruipérez-Valiente, J. A., Staubitz, T., Jenner, M., Halawa, S., Zhang, J., Despujol, I., Maldonado-Mahauad, J., Montoro, G., Pepper, M., Rohloff, T., Lane, J., Turro, C., Li, X., Pérez-Sanagustín, M., & Reich, J. (2022). Large scale analytics of global and regional MOOC providers: Differences in learners' demographics, preferences, and perceptions. *Computers & Education*, 180, Article 104426.
- Sablić, M., Miroslavljević, A., & Skugor, A. (2021). Video-based learning (VBL)—past, present and future: An overview of the research published from 2008 to 2019. *Technology, Knowledge and Learning*, 26(4), 1061–1077.
- Schneider, S., Kriegelstein, F., Beege, M., & Rey, G. D. (2022). The impact of video lecturers' nonverbal communication on learning – an experiment on gestures and facial expressions of pedagogical agents. *Computers & Education*, 176, Article 104350.
- Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research*, 49(1), 1–39.
- Schroeder, K. T., Hubertz, M., Van Campenhout, R., & Johnson, B. G. (2022). Teaching and learning with AI-generated courseware: Lessons from the classroom. *Online Learning*, 26(3).
- Sondermann, C., Huff, M., & Merkt, M. (2024). Distracted by a talking head? An eye tracking study on the effects of instructor presence in learning videos with animated graphic slides. *Learning and Instruction*, 91, Article 101878.
- Tack, A., & Piech, C. (2022). The AI Teacher test: Measuring the pedagogical ability of Blender and GPT-3 in educational dialogues. *arXiv preprint arXiv:2205.07540*.
- Tang, K.-Y., Chang, C.-Y., & Hwang, G.-J. (2021). Trends in artificial intelligence-supported e-learning: A systematic review and co-citation network analysis (1998–2019). *Interactive Learning Environments*, 1–19.
- Tondeur, J., Howard, S. K., Scherer, R., & Siddiq, F. (2023). Untangling the great online transition: A network model of teachers' experiences with online practices. *Computers & Education*, 203, Article 104866.
- Vallis, C., Wilson, S., Gozman, D., & Buchanan, J. (2023). Student perceptions of AI-generated avatars in teaching business ethics: We might not be impressed. *Postdigital Science and Education*, 1–19.
- van Wermeskerken, M., & van Gog, T. (2017). Seeing the instructor's face and gaze in demonstration video examples affects attention allocation but not learning. *Computers & Education*, 113, 98–107.
- Wang, X., Liu, Q., Pang, H., Tan, S. C., Lei, J., Wallace, M. P., & Li, L. (2023). What matters in AI-supported learning: A study of human-AI interactions in language learning using cluster analysis and epistemic network analysis. *Computers & Education*, 194, Article 104703.
- Weber-Wulff, D., Anohina-Naumecca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1), 26.
- Wilson, K. L., Lizzio, A., & Ramsden, P. (1997). The development, validation and application of the Course Experience Questionnaire. *Studies in Higher Education*, 22(1), 33–53.
- Wulff, J. N., & Villadsen, A. R. (2020). Are survey experiments as valid as field experiments in management research? An empirical comparison using the case of ethnic employment discrimination. *European Management Review*, 17(1), 347–356.
- Zerihun, Z., Beishuizen, J., & Van Os, W. (2012). Student learning experience as indicator of teaching quality. *Educational Assessment, Evaluation and Accountability*, 24, 99–111.