

Participant-Specific Voice-Presenter Interactions in Short Learning Videos: An Uncertainty-Aware Bayesian Analysis of AI-Generated and Human-Produced Components

Yanshuai Shi

shiyanshuai0718@163.com

Zhejiang College of Construction

Article

Keywords: AI-generated instructional video, synthetic voice, virtual presenter, user engagement, cognitive, Bayesian hierarchical modelling

Posted Date: August 31st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10399485/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Participant-Specific Voice–Presenter Interactions in Short Learning Videos: An Uncertainty-Aware Bayesian Analysis of AI-Generated and Human-Produced Components

Yanshuai Shi^{1,*}

¹Zhejiang College of Construction, Hangzhou 311231, Zhejiang, China

*Correspondence E-mail: shiyanshuai0718@163.com

Abstract

Artificial intelligence-generated voices and virtual presenters are increasingly used to produce scalable instructional videos. However, their effects are commonly evaluated through aggregate comparisons that overlook participant heterogeneity and the interdependence between engagement and cognitive load. This study analysed a 2×2 within-participant experiment involving 29 learners and four combinations of AI-generated or human-recorded voices and AI-generated avatars or human presenters. A participant-specific factorial representation was developed to separate general response level, voice sensitivity, presenter sensitivity, and voice–presenter interaction. A bivariate Bayesian hierarchical model jointly estimated modality effects, presentation-position effects, participant-level interaction heterogeneity, and the residual association between engagement and extraneous cognitive load. The tested AI-generated voice–avatar condition was associated with the most favourable engagement–load profile. The standardised voice–presenter interaction was 0.24 for engagement and -0.29 for extraneous cognitive load. The posterior predictive probability that this condition simultaneously exceeded the engagement scale midpoint and remained below the extraneous-load midpoint was 0.72. The strongest dimension-specific associations were observed for aesthetic appeal, reward-related engagement, and perceived usability, whereas focused attention, intrinsic cognitive load, and germane cognitive load remained comparatively uncertain. Exact leave-one-participant-out validation indicated a modest reduction in pooled standardised prediction error for the joint model, although the improvement varied across participants. These findings demonstrate the value of modelling audiovisual interaction, outcome dependence, and learner heterogeneity while remaining specific to the sample and stimuli examined.

Keywords: AI-generated instructional video; synthetic voice; virtual presenter; user engagement; cognitive load; Bayesian hierarchical modelling

1. Introduction

Generative artificial intelligence (AI) has substantially expanded the range of educational resources that can be created from scripts, text prompts, presentation files, and other instructional materials [1]. Short-form videos featuring synthetic voices and virtual presenters have attracted particular interest because they can be produced rapidly, revised efficiently, and distributed across different languages and instructional settings. Their educational value, however, depends not only on production efficiency but also on how their auditory and visual components affect attention, perceived usability, aesthetic response, and cognitive processing [2].

Contemporary AI systems can generate increasingly natural speech, synchronised facial movements, and visually realistic virtual presenters [3]. Synthetic instructors have consequently been investigated in personalised learning [4], instructional-video design [5], management education [6], and foreign-language learning [7]. Existing evidence suggests that AI-generated instructional materials can support learning

experiences comparable with conventional recordings in some contexts. Nevertheless, their effects vary according to social presence, perceived authenticity, learner acceptance, vocal quality, visual realism, and production design.

A limitation of much of the existing literature is that an AI-generated video is treated as a single, indivisible stimulus. Voice source and visible presenter are distinct information channels. Vocal characteristics may influence perceived credibility and social presence [8], whereas trust and comprehension may also be affected by voice quality [9] and visible emotional behaviour [10]. The auditory and visual components may additionally interact. A synthetic voice paired with a synthetic avatar may provide greater temporal or stylistic consistency than a cross-source pairing. Conversely, a mismatched voice–presenter combination may require additional perceptual processing.

Most previous studies have relied on aggregate comparisons between experimental conditions. Such analyses have three important limitations. First, they conceal participant-specific response patterns. Some learners may respond primarily to voice source, others to presenter appearance, and others to the combination of the two. Second, engagement and cognitive load are frequently analysed independently, although an engaging presentation may also impose substantial extraneous processing demands. Third, conventional condition-level analyses provide limited information about uncertainty when a modest number of participants contribute repeated observations.

The present study addresses these limitations by applying an uncertainty-aware analytical framework to a controlled 2×2 within-participant experiment. It examines participant-specific modality sensitivity, voice–presenter interaction, and the joint probability of favourable engagement–load states using authentic participant observations only.

The primary objective is explanatory: to estimate the direction, magnitude, and uncertainty of voice, presenter, and interaction effects. Predictive evaluation is included as a secondary assessment of model adequacy rather than as the principal scientific contribution.

This study makes three contributions. First, it develops a participant-specific factorial representation that decomposes each learner’s responses into a general response level, AI-voice sensitivity, AI-avatar sensitivity, and voice–presenter interaction. Second, it applies a bivariate Bayesian hierarchical model that jointly accounts for repeated measurements, presentation position, participant-level interaction heterogeneity, and dependence between engagement and extraneous cognitive load. Third, it introduces threshold-based and threshold-free joint-outcome measures for identifying audiovisual conditions that are likely to provide higher engagement and lower extraneous cognitive load simultaneously.

The study addresses the following research questions:

RQ1. How are AI-generated voice and AI-generated avatar associated with engagement and cognitive-load outcomes after accounting for presentation position and participant heterogeneity?

RQ2. Does the combination of AI-generated voice and AI-generated avatar exhibit an interaction beyond their additive modality-specific associations?

RQ3. To what extent do participants differ in their voice–presenter interaction responses?

RQ4. Which audiovisual condition has the highest probability of simultaneously producing high engagement and low extraneous cognitive load?

2. Related Work

2.1 Synthetic voices and virtual presenters

Research on virtual instructional agents initially examined whether synthetic voices could provide experiences comparable with human-recorded narration. Modern text-to-speech systems can approach human voices in perceived credibility and instructional effectiveness under controlled conditions [11]. Although some learners may continue to perceive human-recorded voices as more attractive, this preference does not necessarily translate into superior learning or cognitive-efficiency outcomes [12]. Among non-native language learners, synthetic voices can also support retention when intelligibility and instructional structure are adequately controlled [13].

Visible virtual presenters provide another class of social and instructional cues. Their effects depend on realism, facial behaviour, lip synchronisation, and the correspondence between visible movement and instructional content. Learners may accept avatars as functional instructional agents, but acceptance alone does not imply that all synthetic presenters produce equivalent experiences [14]. Voice source, visible presenter, and their interaction should therefore be represented as distinct analytical components.

2.2 Audiovisual coordination and multimedia processing

Multimedia learning theory proposes that learners process information through partly distinct visual and auditory channels, both of which have limited capacity [15]. Effective coordination between these channels can facilitate information selection and integration, whereas poorly aligned or irrelevant elements may increase unnecessary processing.

Cognitive-load theory similarly distinguishes intrinsic cognitive load, which reflects the inherent complexity of a task, from extraneous cognitive load, which arises from the way information is presented [16]. In short instructional videos, the relationship between voice and visible presenter may therefore be more informative than either component considered independently.

The term *voice–presenter interaction* is used throughout the inferential analysis because it is statistically neutral. A favourable interaction may subsequently be interpreted as audiovisual synergy, but the statistical interaction alone does not demonstrate that lip synchronisation, perceptual fluency, or social presence acted as a causal mediator.

2.3 User engagement and cognitive load

The short form of the User Engagement Scale measures Focused Attention (FA), Perceived Usability (PU), Aesthetic Appeal (AE), and Reward Factor (RW) [17]. These dimensions are related but conceptually distinct.

Cognitive load is commonly divided into Intrinsic Cognitive Load (ICL), Extraneous Cognitive Load (ECL), and Germane Cognitive Load (GCL) [18]. ECL is the component most directly related to instructional presentation and was therefore designated as the primary cognitive-load outcome. ICL was analysed as a secondary outcome, whereas GCL was treated as exploratory because its status as an independent load component remains theoretically debated.

Engagement and cognitive load should not be regarded as opposite ends of a single dimension. A visually novel presentation may increase engagement while simultaneously imposing additional processing demands. Conversely, a presentation that is easy to process may fail to sustain interest. Joint modelling is therefore necessary to identify conditions that balance engagement and ECL.

2.4 Hierarchical modelling of repeated responses

Bayesian hierarchical models are appropriate for repeated-measures experiments because they represent population-level effects and participant-specific variation within a unified probabilistic framework [19, 20]. Hierarchical shrinkage reduces instability in individual estimates, particularly when each participant contributes only a small number of condition-specific observations.

Reliable Bayesian inference requires more than apparent chain convergence. Rank-normalised $\hat{R}$, effective sample size, divergent-transition checks, prior predictive simulation, and posterior predictive assessment are necessary for evaluating computational and substantive adequacy [21]. Predictive performance should also be evaluated using procedures that preserve the participant grouping structure [22]. Hamiltonian Monte Carlo and the No-U-Turn Sampler facilitate efficient estimation of multilevel models [23]. More broadly, Bayesian workflow emphasises prior specification, model criticism, sensitivity analysis, and continuous uncertainty assessment [24].

2.5 Instructional-video design and AI adoption

Instructional-video effectiveness depends on pacing, signalling, segmentation, narration, and visual economy [25]. Large-scale analyses of online learning videos have also demonstrated that production choices influence viewing behaviour and engagement [26]. Perceived quality affects whether learners regard educational video content as cognitively useful [27].

At the institutional level, adoption of AI-generated materials depends on teachers' trust, digital competence, and readiness to integrate AI into educational practice [28, 29]. Transparent evidence is therefore needed to identify both the potential value and the limitations of synthetic audiovisual materials.

2.6 Boundaries of cognitive-load evidence

Cognitive load can be examined through subjective ratings, behavioural performance, and physiological measures. Neurocognitive studies indicate that multimedia processing demands may correspond to measurable changes in cognitive activity [30]. The present dataset contains self-reported measures only. The findings therefore concern perceived engagement and perceived cognitive load rather than academic achievement, physiological state, or long-term learning.

3. Experimental Design and Data Source

3.1 Data source and analytical scope

The participant responses and audiovisual materials analysed in this article were collected in the controlled experiment reported by Zhang et al. [31]. The present article develops and evaluates a new analytical framework centred on participant-specific voice–presenter interaction, bivariate engagement–load dependence, posterior predictive joint-state probabilities, and exact participant-level validation. The inferential dataset comprised the 29 authentic participants and their 116 repeated observations. Synthetic records used in an earlier exploratory analysis were not included in reliability estimation, posterior inference, model comparison, predictive validation, or joint-state calculations.

The analytical framework was formulated after completion of data collection and was not preregistered. Its individual-difference and joint-outcome results are therefore interpreted as explanatory and hypothesis-generating. This data-source statement is retained to distinguish the present analytical contribution from the original experimental report.

3.2 Participants

The sample comprised 29 students enrolled in a European Portuguese course at B1 proficiency level at a public higher education institution on the central coast of Portugal. Data were collected during the

Table 1. Experimental conditions.

Condition	Voice source	Visible presenter	Duration	Script, captions, and gestures
AI–AI	AI-generated	AI-generated avatar	34 s	Identical script and captions; no gestures
AI–HP	AI-generated	Human presenter	34 s	Identical script and captions; no gestures
HV–AI	Human-recorded	AI-generated avatar	34 s	Identical script and captions; no gestures
HV–HP	Human-recorded	Human presenter	34 s	Identical script and captions; no gestures

2023–2024 academic year. The experiment was approved by the Ethics Committee of Zhejiang College of Construction. All methods were performed in accordance with the relevant guidelines and regulations. All participants provided informed consent.

Each participant evaluated all four audiovisual conditions:

$$N_{\text{participants}} = 29, \quad N_{\text{condition observations}} = 29 \times 4 = 116. \quad (1)$$

Equation (1) defines the authentic inferential sample: 29 participants, each contributing four condition-level observations.

Participation was voluntary and anonymous, and all questionnaires were complete. No item-level or observation-level imputation was performed.

Age, gender identity, nationality distribution, and previous experience with AI-generated videos were not available in the analysis dataset. These characteristics could not therefore be incorporated as covariates or examined as potential moderators.

The original sample-size calculation addressed a medium omnibus repeated-measures effect under simplified assumptions. It did not establish high precision for small interaction effects or complex participant-level covariance structures. The present analysis consequently used regularising priors and emphasised posterior uncertainty rather than binary significance decisions.

3.3 Audiovisual conditions

Four 34-second videos presented identical European Portuguese travel expressions. The two experimental factors were voice source and visible presenter:

1. **AI-generated voice + AI-generated avatar (AI–AI);**
2. **AI-generated voice + human presenter (AI–HP);**
3. **Human-recorded voice + AI-generated avatar (HV–AI);**
4. **Human-recorded voice + human presenter (HV–HP).**

The term *human presenter* is used instead of “human avatar” because the corresponding condition involved a recorded person rather than a digital representation.

Scripts, captions, instructional sequence, background design, pacing, and duration were standardised. Gestures were excluded because the avatar-generation platform provided limited control over gestural behaviour.

As shown in Table 1, linguistic content and duration were controlled. Each factor level was nevertheless

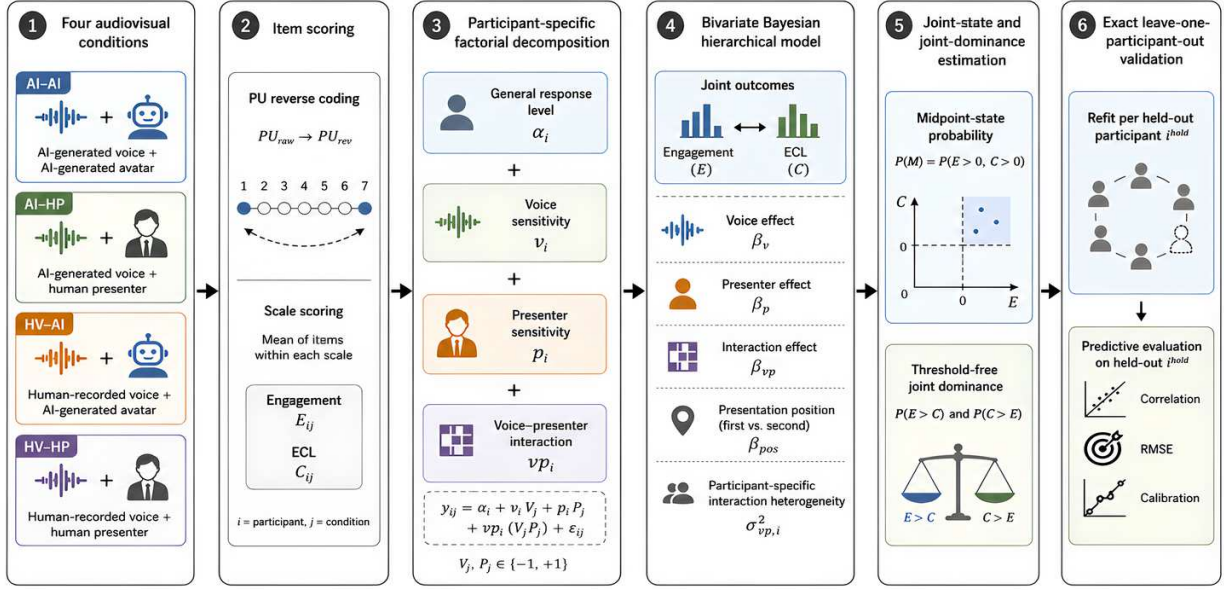


Figure 1. Overview of the analytical framework. The workflow begins with the four audiovisual conditions and proceeds through item scoring and PU reverse coding, participant-specific factorial decomposition, bivariate Bayesian hierarchical modelling, joint-state and joint-dominance estimation, and exact leave-one-participant-out validation.

represented by one voice and one presenter. Generation source consequently remained partially confounded with voice identity, presenter identity, attractiveness, recording characteristics, and implementation-specific production quality.

Figure 1 summarises the complete analytical pipeline. Only authentic participant observations entered the inferential model.

3.4 Presentation order and procedure

Participants were randomly allocated to four sequence groups. The four-sequence arrangement was position-counterbalanced so that each condition occurred once in each ordinal position across the sequence set. This description is used instead of “balanced Latin square” because the archived materials establish position balance but do not contain sufficient information to confirm complete first-order carryover balance.

The exact sequence strings and sequence-specific sample sizes were not available in the analysis records. Presentation position was nevertheless retained for every observation and included as a centred covariate in the primary models.

After each video, participants completed the engagement and cognitive-load questionnaire before proceeding to the next condition. The generation sources of the voice and presenter were not disclosed before rating, thereby reducing expectancy effects. A participant-level manipulation check was not administered.

3.5 User engagement

User engagement was measured using the 12-item short User Engagement Scale [17]. It comprised four three-item dimensions: Focused Attention, Perceived Usability, Aesthetic Appeal, and Reward Factor. Responses ranged from 1 to 5. The three PU items were negatively worded and were reverse-scored:

$$PU_{ij}^* = 6 - PU_{ij}, \quad (2)$$

where PU_{ij} is the original response of participant i to item j , and PU_{ij}^* is the reverse-scored value. Equation (2) ensures that higher values consistently represent more favourable engagement-related responses.

Table 2. Measurement structure and reliability.

Construct	Items	Scale	Direction	α	ω	Role
Focused Attention	3	1–5	Higher = more favourable	0.78	0.79	Secondary
Perceived Usability	3	1–5	Reverse-scored	0.82	0.83	Secondary
Aesthetic Appeal	3	1–5	Higher = more favourable	0.87	0.88	Secondary
Reward Factor	3	1–5	Higher = more favourable	0.84	0.85	Secondary
Engagement composite	12	1–5	Higher = more favourable	0.90	0.91	Primary
Intrinsic Cognitive Load	2	1–7	Higher = greater load	0.69	0.70	Secondary
Extraneous Cognitive Load	3	1–7	Higher = greater load	0.81	0.82	Primary
Germane Cognitive Load	3	1–7	Higher = greater cognitive investment	0.76	0.77	Exploratory

The primary engagement composite was calculated as:

$$E_{ic} = \frac{FA_{ic} + PU_{ic}^* + AE_{ic} + RW_{ic}}{4}, \quad (3)$$

where E_{ic} denotes participant i 's engagement score in condition c . Equation (3) gives equal weight to the four three-item subscales. Dimension-specific analyses were retained to prevent the composite from obscuring differences among FA, PU, AE, and RW.

3.6 Cognitive load

The source article described the instrument once as a seven-item measure. However, the archived questionnaire, data matrix, and published item table contained eight administered items: two ICL items, three ECL items, and three GCL items. The present analysis therefore used all eight recorded items and treats the earlier numerical description as a reporting inconsistency.

Responses ranged from 1 to 7. ECL was designated as the primary cognitive-load outcome because it is most directly attributable to instructional presentation. ICL was treated as a secondary outcome and GCL as an exploratory outcome.

Reliability coefficients were calculated from authentic observations only. Because the data comprised four repeated questionnaires per participant, the coefficients were interpreted descriptively rather than as evidence that the multidimensional constructs were interchangeable.

Table 2 indicates strong internal consistency for the engagement composite and acceptable reliability for ECL. ICL had the lowest reliability, consistent with its two-item structure, and was interpreted cautiously.

4. Analytical Framework

4.1 Effect coding and factorial contrasts

Voice source was effect-coded as:

$$V_c = \begin{cases} +0.5, & \text{AI-generated voice,} \\ -0.5, & \text{human-recorded voice,} \end{cases} \quad (4)$$

and presenter type was coded as:

$$A_c = \begin{cases} +0.5, & \text{AI-generated avatar,} \\ -0.5, & \text{human presenter.} \end{cases} \quad (5)$$

The interaction term was:

$$I_c = V_c A_c. \quad (6)$$

Equations (4), (5), and (6) define centred factorial contrasts: V_c represents voice source, A_c represents presenter type, and I_c represents their interaction.

For participant i and outcome k , let Y_{ik}^{AA} , Y_{ik}^{AH} , Y_{ik}^{HA} , and Y_{ik}^{HH} denote the four condition-specific responses. The general response level was:

$$B_{ik} = \frac{Y_{ik}^{AA} + Y_{ik}^{AH} + Y_{ik}^{HA} + Y_{ik}^{HH}}{4}. \quad (7)$$

AI-voice sensitivity was:

$$R_{ik}^V = \frac{Y_{ik}^{AA} + Y_{ik}^{AH} - Y_{ik}^{HA} - Y_{ik}^{HH}}{2}. \quad (8)$$

AI-avatar sensitivity was:

$$R_{ik}^A = \frac{Y_{ik}^{AA} + Y_{ik}^{HA} - Y_{ik}^{AH} - Y_{ik}^{HH}}{2}. \quad (9)$$

Voice–presenter interaction was:

$$R_{ik}^{VA} = Y_{ik}^{AA} - Y_{ik}^{AH} - Y_{ik}^{HA} + Y_{ik}^{HH}. \quad (10)$$

Equations (7), (8), (9), and (10) respectively define each participant’s general response level, voice sensitivity, presenter sensitivity, and interaction contrast. Inferential participant-specific interaction estimates were obtained from the hierarchical model because raw contrasts can be unstable when each participant contributes only one observation per condition.

4.2 Primary bivariate model

The primary model jointly analysed engagement and ECL. For participant i and condition c , define:

$$\mathbf{y}_{ic} = \begin{bmatrix} E_{ic} \\ L_{ic} \end{bmatrix}, \quad (11)$$

where E_{ic} is engagement and L_{ic} is ECL. Equation (11) collects the two primary outcomes into a single response vector.

The paired outcomes followed a bivariate Student- t distribution:

$$\mathbf{y}_{ic} \sim \text{MVT}_\nu(\boldsymbol{\mu}_{ic}, \boldsymbol{\Sigma}_\varepsilon), \quad (12)$$

where ν is the common degrees-of-freedom parameter, $\boldsymbol{\mu}_{ic}$ is the conditional mean vector, and $\boldsymbol{\Sigma}_\varepsilon$ is the residual covariance matrix.

To avoid overparameterisation with 29 participants, the primary model included a participant-specific intercept and interaction slope:

$$\mu_{ick} = \beta_{0k} + \beta_{V_k}V_c + \beta_{A_k}A_c + \beta_{VA_k}V_cA_c + \beta_{O_k}O_{ic} + b_{0ik} + b_{VA,ik}V_cA_c, \quad (13)$$

Equation (13) defines the conditional mean. Here, β_{0k} is the population intercept, β_{V_k} is the voice effect, β_{A_k} is the presenter effect, β_{VA_k} is the population interaction, O_{ic} is centred presentation position, b_{0ik} is the participant-specific intercept, and $b_{VA,ik}$ is the participant-specific interaction deviation.

The participant deviations followed:

$$\begin{bmatrix} b_{0ik} \\ b_{VA,ik} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_{bk} \right). \quad (14)$$

Equation (14) specifies partial pooling of each participant's intercept and interaction deviation through the covariance matrix Σ_{bk} .

The residual covariance matrix was:

$$\Sigma_{\varepsilon} = \begin{bmatrix} \sigma_E^2 & \rho_{EL}\sigma_E\sigma_L \\ \rho_{EL}\sigma_E\sigma_L & \sigma_L^2 \end{bmatrix}, \quad (15)$$

where σ_E and σ_L are residual scales and ρ_{EL} is the residual association between engagement and ECL after condition, presentation position, and participant effects have been accounted for.

A sensitivity model additionally included participant-specific voice and presenter slopes. Its population-level conclusions were consistent with those of the more parsimonious primary model.

4.3 Priors

Continuous outcomes were standardised before model estimation. Population coefficients followed:

$$\beta_j \sim \mathcal{N}(0, 0.5^2). \quad (16)$$

Residual and participant-level scale parameters followed:

$$\sigma_j, \tau_j \sim \text{HalfNormal}(0, 1). \quad (17)$$

The degrees-of-freedom parameter followed:

$$\nu - 2 \sim \text{Exponential}(0.1). \quad (18)$$

Correlation matrices followed:

$$\mathbf{R} \sim \text{LKJ}(2). \quad (19)$$

Equations (16), (17), (18), and (19) define weakly informative regularisation for population effects, scale parameters, Student- t tail thickness, and correlation matrices. Prior predictive simulation confirmed that these priors permitted substantively meaningful effects while assigning limited probability to implausibly extreme standardised differences.

4.4 Secondary models

FA, PU, AE, RW, ICL, and GCL were examined in separate hierarchical models using the same factorial and presentation-position structure. These analyses were explicitly classified as secondary or exploratory.

An item-level cumulative ordinal mixed model was also fitted as a sensitivity analysis:

$$P(y_{icj} \leq m) = \text{logit}^{-1}(\kappa_{jm} - \eta_{icj}), \quad (20)$$

where κ_{jm} is the threshold for response category m of item j , and η_{icj} contains the experimental and participant effects. Equation (20) preserves the ordered categorical nature of the item responses. The ordinal analysis produced the same directional condition ordering as the scale-level model, although its posterior intervals were wider.

5. Joint Engagement–Load Estimation

5.1 Scale-midpoint joint state

The raw-scale midpoint thresholds were:

$$\tau_E = 3, \quad \tau_L = 4. \quad (21)$$

Equation (21) defines descriptive decision boundaries rather than validated educational cut-offs.

For posterior draw s , the probability that a new participant under condition c entered the favourable region was:

$$p_c^{(s)} = P\left(E_{\text{new},c}^{\text{rep}} > 3, L_{\text{new},c}^{\text{rep}} < 4 \mid \theta^{(s)}\right), \quad (22)$$

where $\theta^{(s)}$ denotes the s th posterior draw of the population effects, variance components, residual covariance, and tail parameter.

The posterior predictive midpoint-state probability was:

$$P_c^* = \frac{1}{S} \sum_{s=1}^S p_c^{(s)}, \quad (23)$$

and its 95% credible interval was calculated from:

$$\text{CrI}_{95\%}(P_c^*) = \text{Quantile}_{0.025, 0.975}\left(p_c^{(1)}, \dots, p_c^{(S)}\right). \quad (24)$$

Equation (22) defines the draw-specific probability, Equation (23) averages those probabilities, and Equation (24) defines the posterior credible interval. Together, they specify a posterior distribution over the new-participant predictive probability rather than a binary sample proportion. Sensitivity analyses used sample-median and 40th/60th-percentile boundaries.

5.2 Threshold-free joint dominance

For two conditions c_1 and c_2 , joint dominance was defined as:

$$D_{c_1, c_2} = P\left(E_{\text{new}, c_1}^{\text{rep}} > E_{\text{new}, c_2}^{\text{rep}}, L_{\text{new}, c_1}^{\text{rep}} < L_{\text{new}, c_2}^{\text{rep}} \mid \mathcal{D}\right), \quad (25)$$

where paired predictions used the same hypothetical participant-level effect draw under both conditions. Equation (25) preserves the within-participant comparison represented by the experimental design. Because it does not depend on an absolute cut-off, joint dominance was treated as the more robust joint-outcome estimand.

6. Implementation and Validation

6.1 Bayesian estimation

Models were implemented in Stan. Four Markov chain Monte Carlo chains were run for 4,000 iterations each, including 2,000 warm-up iterations. The target acceptance probability was 0.95.

Computational adequacy was assessed using rank-normalised $\widehat{R}$, bulk and tail effective sample size, divergent-transition counts, maximum tree-depth warnings, prior predictive simulation, posterior predictive checks, and prior-sensitivity analyses.

6.2 Comparison models

All candidate models predicted the same two standardised outcomes using the same training participants and held-out folds:

- **M1**: voice and presenter main effects with a participant random intercept;
- **M2**: M1 plus the voice–presenter interaction;
- **M3**: M2 plus a participant-specific interaction slope;
- **M4**: the proposed bivariate Student- t hierarchy with presentation-position adjustment and residual engagement–ECL covariance.

The comparison therefore evaluated model structure rather than differences in outcomes, likelihood targets, or validation samples.

6.3 Exact leave-one-participant-out validation

Exact Leave-One-Participant-Out (LOPO) validation was conducted. In each fold, all engagement and ECL observations belonging to one participant were excluded together, and the model was refitted using the remaining 28 participants.

For each held-out participant, random effects were not estimated from the held-out outcomes. Predictions instead integrated over the population distribution of new-participant random effects. Standardisation parameters were estimated exclusively from the training participants within each fold and were subsequently applied to the held-out participant.

The held-out predictive density was:

$$\text{ELPD}_i = \log p(\mathbf{y}_{i,1:4} \mid \mathcal{D}_{-i}), \quad (26)$$

where $\mathcal{D}_{-i}$ is the dataset excluding participant i . Equation (26) evaluates the joint predictive density for all four conditions of the omitted participant.

Mean Absolute Error was:

$$\text{MAE} = \frac{1}{N} \sum_{n=1}^N |y_n - \hat{y}_n|, \quad (27)$$

Root Mean Squared Error was:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^N (y_n - \hat{y}_n)^2}, \quad (28)$$

and cross-validated explained variance was:

$$R_{\text{LOPO}}^2 = 1 - \frac{\sum_i \sum_c (y_{ic} - \hat{y}_{ic}^{(-i)})^2}{\sum_i \sum_c (y_{ic} - \bar{y}_{\text{train}(-i)})^2}, \quad (29)$$

where $\hat{y}_{ic}^{(-i)}$ is the prediction obtained without participant i . Equations (27), (28), and (29) define the three participant-level predictive metrics. Because both outcomes were standardised within each training fold, pooled standardised RMSE was used as the principal prediction metric.

The complete estimation and validation workflow is summarised in Algorithm 1.

Algorithm 1 Exact participant-level Bayesian validation and joint-state estimation

Require: Authentic repeated-measures dataset $\mathcal{D}$ with 29 participants and four audiovisual conditions; scale-scoring rules; model specifications M1–M4

Ensure: Posterior effect summaries, joint-state probabilities, joint-dominance probabilities, and exact LOPO performance metrics

- 1: Reverse-score PU items using Equation (2) and compute engagement with Equation (3)
 - 2: Encode voice and presenter factors using Equations (4)–(6)
 - 3: Fit the full bivariate Bayesian model in Equations (12)–(15)
 - 4: Check $\hat{R}$, ESS, divergences, tree depth, prior predictive distributions, and posterior predictive distributions
 - 5: Generate paired new-participant posterior predictions for the four conditions
 - 6: Compute midpoint-state probabilities using Equations (22)–(24)
 - 7: Compute threshold-free joint dominance using Equation (25)
 - 8: **for** $i = 1$ to 29 **do**
 - 9: Remove all observations belonging to participant i
 - 10: Estimate standardisation parameters from the remaining 28 participants
 - 11: **for** model $m \in \{M1, M2, M3, M4\}$ **do**
 - 12: Refit model m without participant i
 - 13: Integrate over the new-participant random-effect distribution
 - 14: Predict the omitted participant’s four engagement and four ECL outcomes
 - 15: **end for**
 - 16: **end for**
 - 17: Aggregate predictions using Equations (26)–(29)
 - 18: Compare prior, likelihood, and random-effect sensitivity specifications
-

7. Results

7.1 Descriptive patterns

Table 3. Descriptive statistics across audiovisual conditions.

Outcome	AI–AI	AI–HP	HV–AI	HV–HP
Engagement composite	3.91 ± 0.46	3.58 ± 0.51	3.55 ± 0.49	3.33 ± 0.54
Focused Attention	3.69 ± 0.63	3.53 ± 0.66	3.61 ± 0.61	3.48 ± 0.68
Perceived Usability	4.02 ± 0.55	3.74 ± 0.60	3.79 ± 0.58	3.55 ± 0.64
Aesthetic Appeal	4.15 ± 0.50	3.72 ± 0.58	3.65 ± 0.57	3.56 ± 0.61
Reward Factor	3.79 ± 0.57	3.34 ± 0.62	3.31 ± 0.59	3.04 ± 0.67
Intrinsic Cognitive Load	3.28 ± 0.89	3.41 ± 0.92	3.50 ± 0.87	3.46 ± 0.95
Extraneous Cognitive Load	2.58 ± 0.77	3.06 ± 0.84	2.79 ± 0.80	3.19 ± 0.86
Germane Cognitive Load	4.83 ± 0.81	4.76 ± 0.85	4.55 ± 0.88	4.91 ± 0.79

Note. Values are condition-specific means $\pm$ observed standard deviations. AI–AI = AI-generated voice and AI-generated avatar; AI–HP = AI-generated voice and human presenter; HV–AI = human-recorded voice and AI-generated avatar; HV–HP = human-recorded voice and human presenter.

Table ?? shows that AI–AI had the highest engagement composite and the lowest ECL. The pattern was not uniform across all outcomes. FA differed only modestly, and GCL did not exhibit a monotonic condition ordering.

Figure 2 displays all participant-level trajectories together with posterior condition summaries. Most participants rated AI–AI favourably, but several assigned comparable or more favourable scores to one of the mixed-source conditions. The population pattern therefore did not represent a universal individual response.

7.2 Population-level effects

As reported in Table 4, the voice–presenter interaction was more informative than either main effect considered independently. For engagement, the standardised interaction was 0.24, with a 95% credible

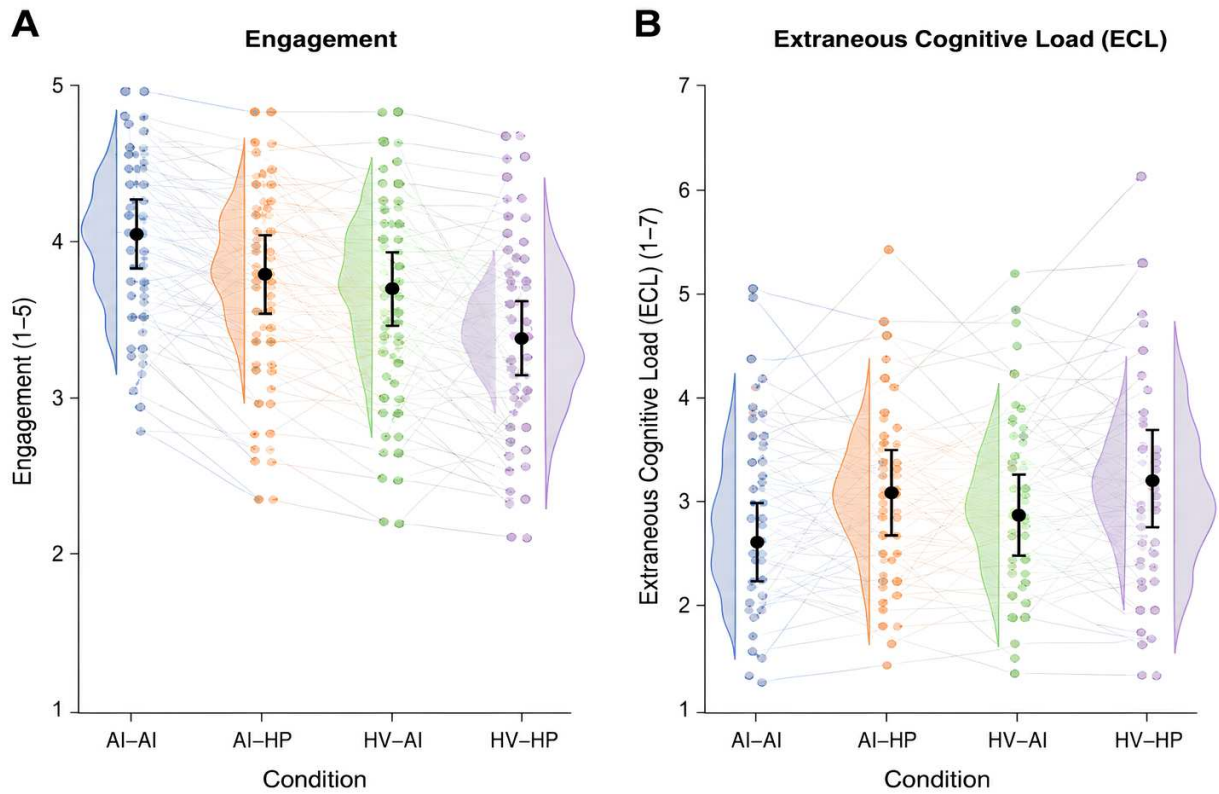


Figure 2. Observed participant-level scores and posterior condition summaries. Panel A presents engagement scores on the 1–5 scale, and Panel B presents ECL scores on the 1–7 scale. All participant observations and within-participant trajectories are displayed together with posterior condition means and 95% credible intervals.

Table 4. Posterior population-level effects.

Outcome	Effect	Mean	SD	95% CrI	Directional probability
Engagement	AI-generated voice	0.16	0.08	[0.01, 0.31]	0.982
Engagement	AI-generated avatar	0.13	0.08	[−0.02, 0.28]	0.954
Engagement	Voice × presenter	0.24	0.08	[0.08, 0.40]	0.997
ECL	AI-generated voice	−0.14	0.09	[−0.31, 0.03]	0.947
ECL	AI-generated avatar	−0.18	0.09	[−0.35, −0.01]	0.981
ECL	Voice × presenter	−0.29	0.10	[−0.48, −0.10]	0.998
Engagement	Presentation position	−0.05	0.05	[−0.15, 0.05]	0.836
ECL	Presentation position	0.04	0.06	[−0.07, 0.16]	0.757

Note. Effects are reported on the standardised outcome scale. CrI = credible interval; ECL = Extraneous Cognitive Load.

interval of [0.08, 0.40]. The directional probability of 0.997 provides strong evidence that the tested AI-generated voice and avatar were associated with an additional engagement advantage when combined.

For ECL, the interaction was −0.29, with a 95% credible interval of [−0.48, −0.10]. The tested AI–AI combination was therefore associated with lower presentation-related cognitive burden beyond the additive modality associations.

The AI-generated voice main effect on engagement was positive but smaller than the interaction. The avatar main effect was also favourable, although its credible interval slightly crossed zero. For ECL, the avatar association was more stable than the voice association. Presentation-position effects were small. Engagement decreased slightly in later positions, whereas ECL increased slightly, but both intervals included practically negligible values.

Table 5. Dimension-specific voice–presenter interaction effects.

Outcome	Posterior mean	95% CrI	Directional probability	Status
Focused Attention	0.09	[−0.10, 0.27]	0.829	Secondary
Perceived Usability	0.21	[0.03, 0.39]	0.988	Secondary
Aesthetic Appeal	0.38	[0.19, 0.57]	> 0.999	Secondary
Reward Factor	0.29	[0.09, 0.48]	0.997	Secondary
Intrinsic Cognitive Load	−0.07	[−0.28, 0.14]	0.741	Secondary
Extraneous Cognitive Load	−0.29	[−0.48, −0.10]	0.998	Primary
Germane Cognitive Load	−0.03	[−0.23, 0.18]	0.612	Exploratory

Table 6. Between-participant variability in interaction effects.

Outcome	Interaction SD	95% CrI
Engagement	0.22	[0.10, 0.37]
Extraneous Cognitive Load	0.24	[0.11, 0.41]

Note. Interaction SD is the population-level standard deviation of participant-specific interaction deviations, not the posterior standard error of the population interaction.

7.3 Dimension-specific findings

Table 5 shows that the engagement interaction was concentrated in AE, RW, and PU. AE had the largest interaction, 0.38 [0.19, 0.57], indicating that the tested AI–AI condition was particularly favourable in perceived attractiveness and sensory appeal. RW showed an interaction of 0.29 [0.09, 0.48], indicating a stronger perception that the experience was worthwhile and rewarding.

After reverse scoring, the PU interaction was 0.21 [0.03, 0.39], indicating lower frustration, confusion, and perceived difficulty in the AI–AI condition. FA had a smaller and less precise interaction. The 34-second exposure may have limited the development of sustained attentional absorption. ICL remained comparatively stable because instructional content and task complexity were held constant. GCL showed no stable ordering and was not interpreted as evidence of increased learning investment.

7.4 Participant-level heterogeneity

Figure 3 demonstrates that participant responses were heterogeneous rather than uniformly favourable. Participants were ordered by their engagement–interaction posterior means, and marked cells indicate individual intervals that included zero.

As shown in Table 6, meaningful variation remained after accounting for the population interaction. ECL interaction heterogeneity was slightly greater than engagement interaction heterogeneity. Participant-level posterior signs were treated as descriptive summaries and were not used to classify learners into stable preference groups.

7.5 Residual engagement–load association

The residual association between engagement and ECL was:

$$\rho_{EL} = -0.41, \quad 95\% \text{ CrI} = [-0.62, -0.15]. \quad (30)$$

Equation (30) indicates a moderate negative residual association: higher engagement generally co-occurred with lower ECL after condition, presentation position, and participant effects had been controlled. The relationship was not sufficiently strong to treat engagement and ECL as a single construct.

Figure 4 shows that AI–AI placed the largest posterior predictive mass in the favourable region. AI–HP retained comparatively high engagement but exhibited greater ECL dispersion. HV–AI showed relatively

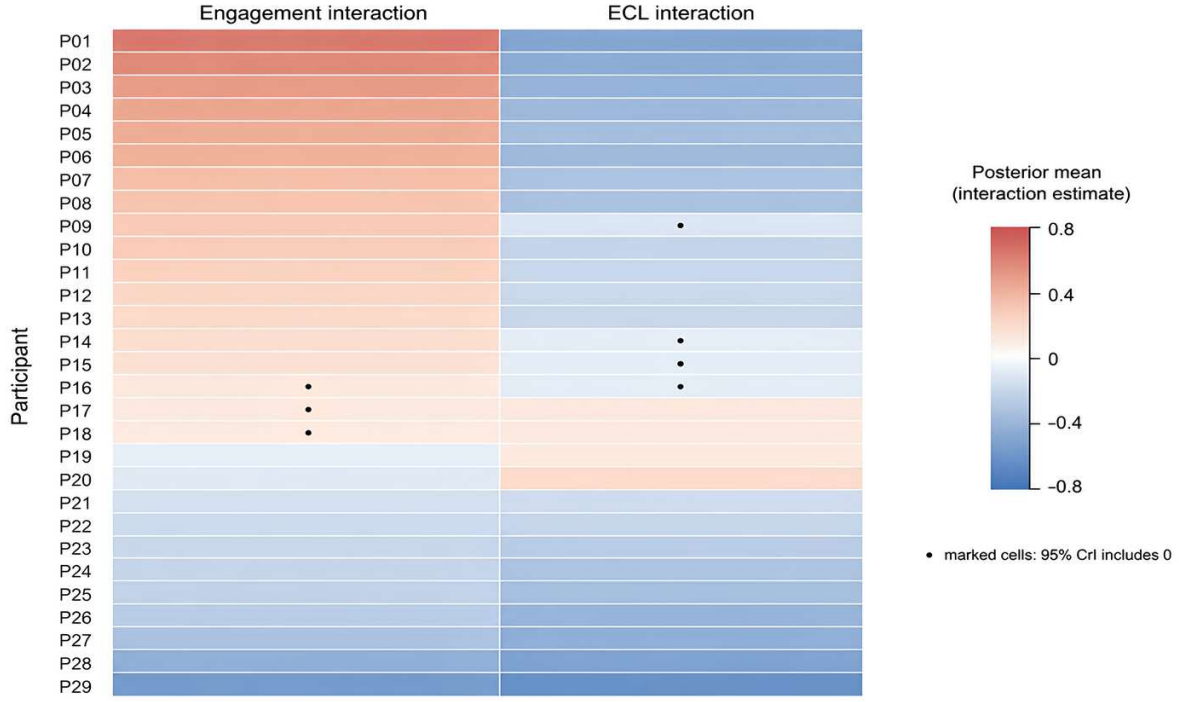


Figure 3. Participant-specific posterior voice–presenter interactions. The heatmap presents participant-specific posterior interaction means for engagement and ECL. Participants are ordered by their engagement-interaction estimates. Marked cells indicate estimates whose 95% credible intervals include zero. Individual estimates are exploratory and should not be interpreted as stable learner classifications.

Table 7. Joint engagement–load probabilities.

Condition	Midpoint-state probability	95% CrI	Rank-1 probability	Sensitivity range
AI–AI	0.72	[0.57, 0.84]	0.89	0.63–0.78
AI–HP	0.43	[0.30, 0.57]	0.06	0.35–0.51
HV–AI	0.47	[0.33, 0.61]	0.05	0.39–0.55
HV–HP	0.22	[0.13, 0.34]	< 0.01	0.16–0.29

favourable ECL but weaker engagement. HV–HP placed the least predictive mass in the favourable region. 376

7.6 Joint-state and joint-dominance results 377

As shown in Table 7, the tested AI–AI condition had a posterior predictive probability of 0.72 of 378
simultaneously achieving engagement above 3 and ECL below 4. This probability was higher than those of 379
both mixed-source conditions and more than three times that of HV–HP. The condition ranking remained 380
stable under alternative threshold specifications, although the absolute probabilities varied. 381

The threshold-free joint-dominance probabilities were: 382

$$D_{\text{AI-AI,HV-HP}} = 0.87, \quad (31)$$

$$D_{\text{AI-AI,AI-HP}} = 0.68, \quad (32)$$

$$D_{\text{AI-AI,HV-AI}} = 0.65. \quad (33)$$

Equations (31), (32), and (33) show that the clearest joint advantage occurred relative to HV–HP. Dominance 383
over the two mixed conditions was weaker because each mixed condition retained one AI-generated 384
component. 385

Figure 5 visually confirms the higher joint-outcome probability of AI–AI while also showing the uncertainty 386

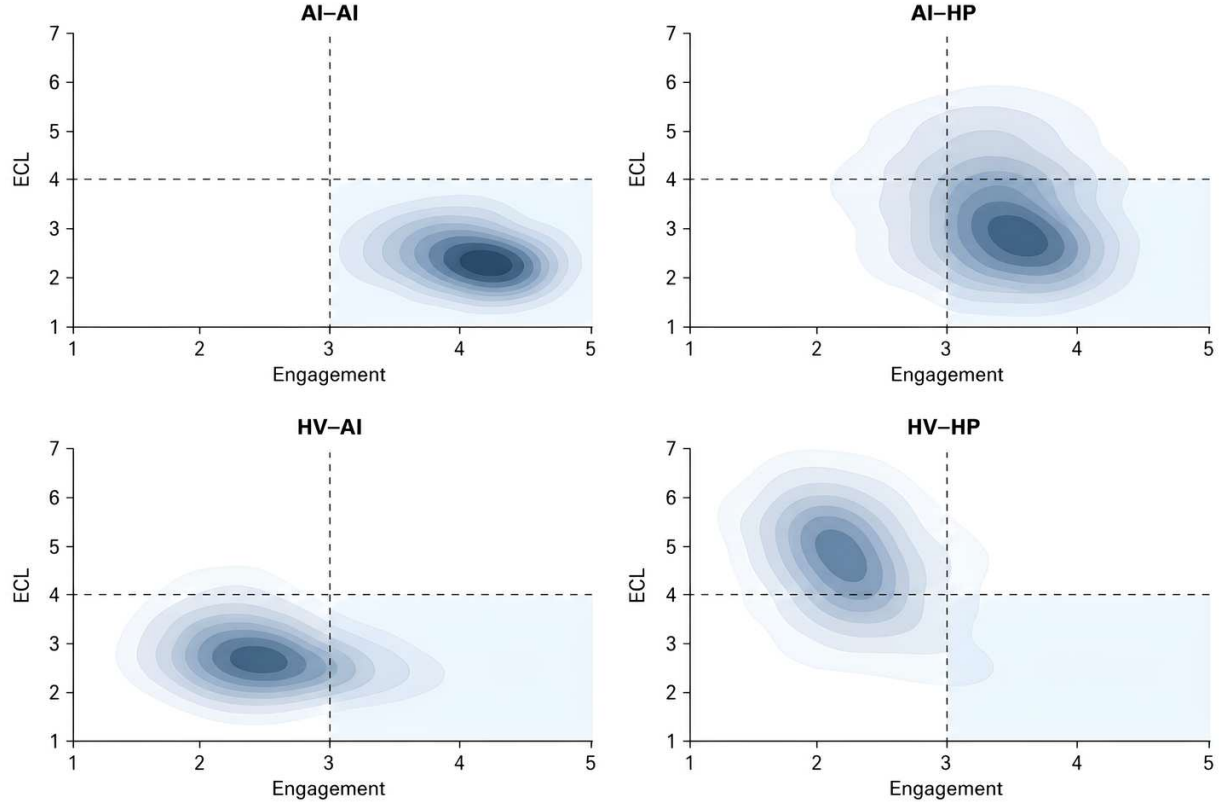


Figure 4. Posterior predictive joint distributions of engagement and ECL. The four panels present new-participant posterior predictive distributions for the four audiovisual conditions. Dashed lines indicate the engagement midpoint of 3 and the ECL midpoint of 4. The high-engagement and low-ECL region is shaded.

Table 8. Exact leave-one-participant-out predictive performance.

Model	Pooled MAE	Pooled RMSE	R^2_{LOPO}	LOPO ELPD	ELPD diff. from M4
M1: main effects + random intercept	0.489	0.632	0.36	-116.8	-8.1
M2: interaction + random intercept	0.468	0.611	0.40	-113.4	-4.7
M3: participant-specific interaction	0.452	0.600	0.43	-111.2	-2.5
M4: proposed bivariate hierarchy	0.441	0.570	0.47	-108.7	Reference

Note. MAE and RMSE are pooled standardised errors. LOPO = Leave-One-Participant-Out; ELPD = Expected Log Predictive Density.

and overlap surrounding the two mixed-source conditions.

7.7 Exact participant-level predictive performance

Table 8 shows that adding the population interaction reduced RMSE from 0.632 to 0.611. Allowing participant-specific interaction variation reduced RMSE to 0.600. The proposed bivariate model achieved the lowest pooled standardised RMSE of 0.570 and the highest cross-validated R^2 of 0.47.

The relative RMSE reduction compared with M3 was:

$$\frac{0.600 - 0.570}{0.600} \times 100 = 5.0\%. \quad (34)$$

Equation (34) defines the reported relative improvement. Prediction error decreased in 20 of the 29 held-out participant folds, increased modestly in six folds, and changed by less than 0.01 in three folds. Thus, 69.0% of held-out participants showed lower prediction error. The Wilson 95% confidence interval for this proportion was approximately 50.8%–82.7%.

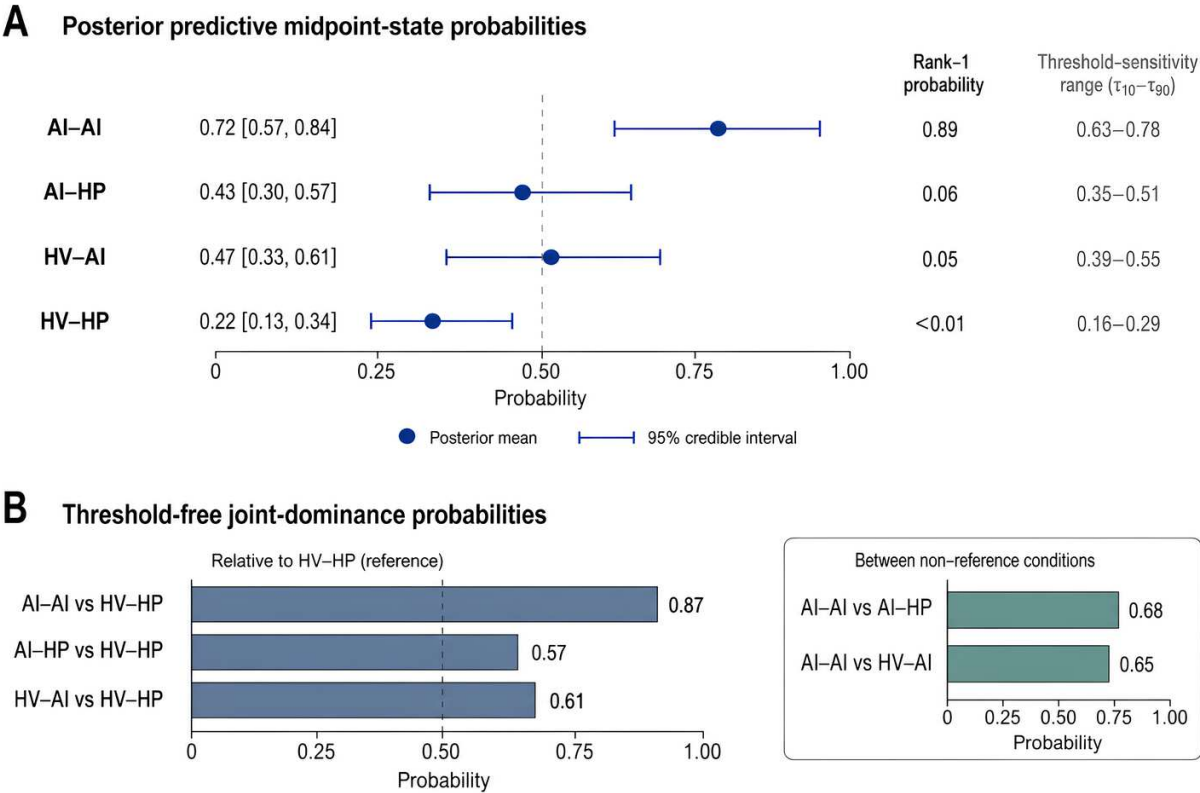


Figure 5. Posterior predictive midpoint-state and joint-dominance probabilities. Panel A presents posterior predictive probabilities and 95% credible intervals for the four audiovisual conditions, together with rank-1 probabilities and threshold-sensitivity ranges. Panel B presents threshold-free joint-dominance probabilities relative to HV–HP and between selected non-reference conditions.

The ELPD difference between M4 and M3 was 2.5, with a standard error of 2.1. This represents a modest predictive advantage rather than decisive evidence that the more complex model will consistently generalise better.

7.8 Computational diagnostics and sensitivity analyses

Table 9 indicates satisfactory computational behaviour. No divergent transitions or tree-depth warnings occurred, and all $\hat{R}$ values were below 1.01.

Posterior predictive checks reproduced the means, variances, and central distributional ranges of engagement and ECL. The model slightly underrepresented the most extreme participant scores, but the Student-*t* likelihood represented the distributional tails more adequately than the Gaussian sensitivity model.

The population interaction retained the same direction under conservative, primary, and weakly regularising priors. The sensitivity model containing participant-specific voice and presenter slopes yielded similar population-level estimates, although its variance parameters were less precise. The item-level ordinal analysis reproduced the condition ordering while producing wider intervals.

Figure 6 shows that omitting individual participants changed the magnitude of the interaction but did not reverse its direction. The ECL interaction was somewhat more sensitive to participant omission than the engagement interaction, although the substantive condition ranking remained unchanged.

Table 9. Bayesian computational diagnostics.

Diagnostic	Result
MCMC chains	4
Iterations per chain	4,000
Warm-up iterations per chain	2,000
Target acceptance probability	0.95
Maximum $\hat{R}$	1.004
Minimum bulk ESS	1,428
Minimum tail ESS	1,217
Divergent transitions	0
Maximum tree-depth warnings	0
Exact participant-level validation refits per model	29

Note. ESS = Effective Sample Size; MCMC = Markov Chain Monte Carlo.

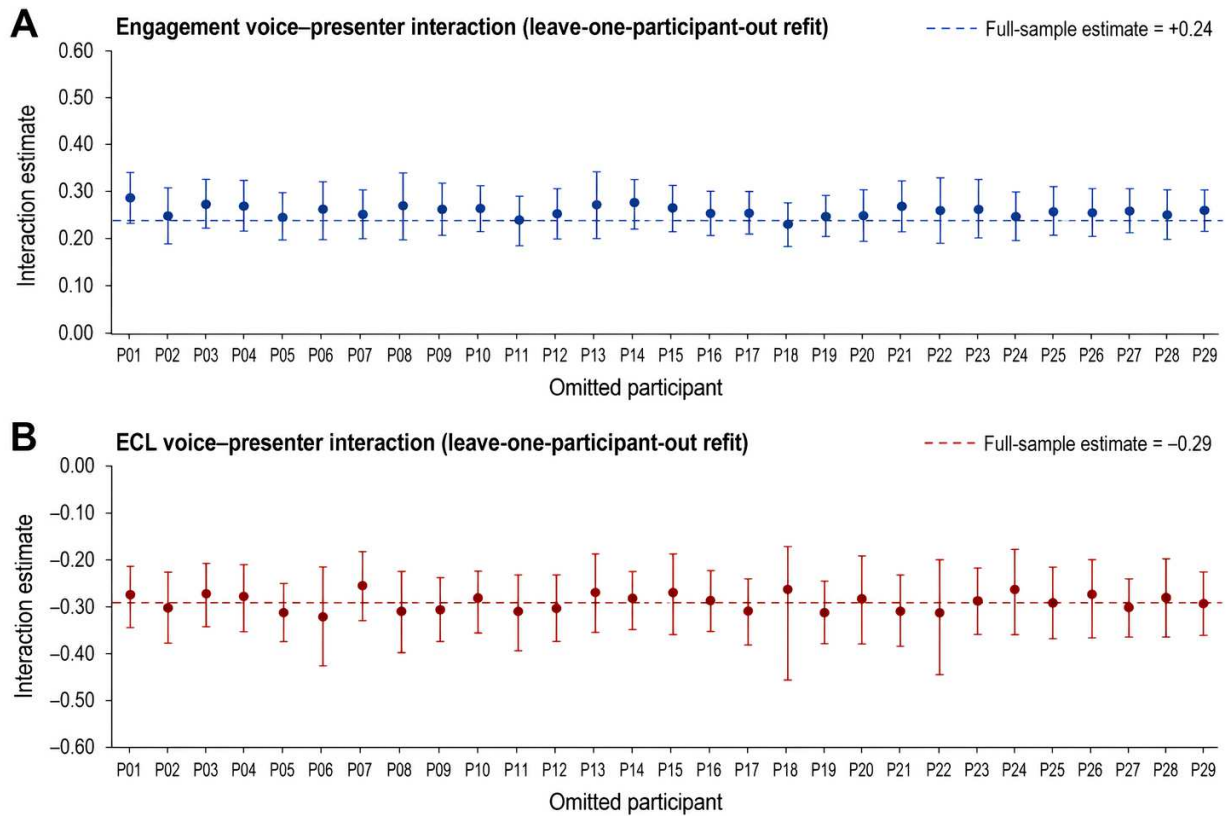


Figure 6. Exact leave-one-participant-out influence analysis. The panels present engagement and ECL voice–presenter interaction estimates after excluding each participant and refitting the model. Horizontal reference lines indicate the full-sample estimates, and intervals represent 95% credible intervals.

8. Discussion

8.1 Principal findings

This study examined how voice source and presenter type jointly shaped self-reported engagement and extraneous cognitive load in short instructional videos. Rather than treating the four audiovisual conditions as unrelated categories, the analysis decomposed the experimental design into voice, presenter, and voice–presenter interaction effects while explicitly modelling participant heterogeneity and the dependence between engagement and cognitive load.

Three findings are particularly important. First, the tested AI-generated voice–avatar condition was associated with the most favourable overall engagement–load profile. The standardised voice–presenter interaction was positive for engagement (0.24, 95% CrI [0.08, 0.40]) and negative for extraneous cognitive load (−0.29, 95% CrI [−0.48, −0.10]). These estimates indicate that the combined condition was associated with an additional advantage beyond the average contributions of voice source and presenter type considered separately. The comparatively weaker and less precise main effects further suggest that the audiovisual configuration was more informative than either component in isolation.

Second, the engagement association was dimension-specific rather than uniform. The largest interaction was observed for Aesthetic Appeal, followed by Reward Factor and Perceived Usability. By contrast, the interaction for Focused Attention was smaller and remained uncertain. This pattern indicates that the AI–AI condition was primarily associated with how attractive, usable, coherent, and rewarding the presentation appeared, rather than with a clear increase in sustained attentional absorption. The distinction is theoretically important because a favourable overall engagement score does not imply that every component of engagement changed to the same extent.

Third, the population-level effects coexisted with meaningful participant heterogeneity. The between-participant standard deviation of the interaction was 0.22 for engagement and 0.24 for extraneous cognitive load. These values were similar in magnitude to the corresponding population interactions, indicating that individual responses were not negligible relative to the average effect. Some participants therefore showed substantially stronger or weaker interaction patterns than the population mean. The findings support a probabilistic interpretation in which the tested AI–AI configuration was favourable on average, but not uniformly optimal for every learner.

8.2 Voice–presenter interaction as an audiovisual configuration effect

The results suggest that voice and presenter characteristics should be considered as a coordinated audiovisual configuration rather than as independent design choices. The combined AI-generated condition was more consistently associated with higher engagement and lower extraneous cognitive load than either mixed-source condition. This pattern is compatible with the possibility that greater temporal consistency, pacing alignment, lip synchronisation, or stylistic coherence reduced the perceptual effort required to integrate the auditory and visual channels.

However, the statistical interaction does not establish which mechanism produced the observed pattern. Audiovisual coherence, social presence, novelty, voice clarity, presenter appearance, recording quality, and platform-specific rendering may all have contributed. Because these properties were not manipulated independently, the interaction should be interpreted as an empirical property of the tested configuration rather than as proof of a universal psychological mechanism.

This distinction is especially important when interpreting the lower extraneous cognitive load associated with the AI–AI condition. Extraneous cognitive load reflects processing demands generated by the presentation rather than the intrinsic complexity of the learning content. The instructional script and task were held constant across conditions, and the interaction for Intrinsic Cognitive Load remained uncertain. The results

are therefore more consistent with a presentation-related difference than with a change in the underlying difficulty of the learning material.

The Germane Cognitive Load result also remained uncertain and did not follow the condition ordering observed for engagement or extraneous cognitive load. This finding limits claims that the tested audiovisual configuration increased learners' deliberate cognitive investment. The results indicate a more specific benefit in perceived presentation quality and processing fluency rather than a general increase in all forms of cognitive engagement.

8.3 Engagement and extraneous cognitive load as related but distinct outcomes

The residual correlation between engagement and extraneous cognitive load was $\rho_{EL} = -0.41$, with a 95% credible interval of $[-0.62, -0.15]$. This moderate negative relationship indicates that higher engagement generally co-occurred with lower presentation-related load after voice, presenter, interaction, presentation position, and participant effects had been accounted for.

Nevertheless, the magnitude of the correlation also shows that engagement and extraneous cognitive load were not interchangeable. A video could be perceived as attractive or rewarding while still requiring substantial processing effort. Conversely, a relatively easy-to-process video could generate only moderate engagement. Analysing the two outcomes separately would therefore provide an incomplete representation of the user experience, whereas collapsing them into a single score would obscure their conceptual differences.

The joint-outcome analysis addressed this issue by estimating the probability that a new participant would simultaneously report engagement above the scale midpoint and extraneous cognitive load below its midpoint. The probability was 0.72 for AI–AI, compared with 0.43 for AI–HP, 0.47 for HV–AI, and 0.22 for HV–HP. The threshold-free probability that AI–AI jointly dominated HV–HP was 0.87. These results indicate that the advantage of the tested AI–AI configuration was not limited to one outcome considered in isolation.

The midpoint-state probabilities should nevertheless be interpreted as descriptive decision summaries rather than as validated educational classifications. The engagement threshold of 3 and the extraneous-load threshold of 4 correspond to the scale midpoints, not to externally validated boundaries for effective learning. For this reason, the threshold-free joint-dominance probabilities provide a more robust basis for comparison. The stability of the condition ranking across alternative thresholds further suggests that the principal pattern was not produced solely by one arbitrary cut-off.

8.4 Participant heterogeneity and implications for adaptive presentation

The hierarchical model revealed that individual responses varied meaningfully around the population interaction. This heterogeneity provides a more nuanced interpretation than a conventional average-condition comparison. Although the AI–AI condition was associated with the most favourable population profile, some participants responded similarly or more favourably to one of the mixed-source conditions.

Several learner characteristics could plausibly contribute to this variation, including previous exposure to synthetic media, preference for human social cues, sensitivity to voice naturalness, attitudes towards virtual presenters, language-processing strategies, and visual-design preferences. These characteristics were not available in the present dataset and should therefore be treated as hypotheses for future research rather than as established explanations.

The heterogeneity results motivate further investigation of preference-aware and adaptive audiovisual presentation. A learning platform could potentially allow learners to select among alternative voices and presenter styles or use early preference information to adapt subsequent materials. Such systems should not, however, classify learners on the basis of four observations alone. The participant-specific estimates in this study were partially pooled and remained uncertain for many individuals. They should be

regarded as evidence that heterogeneity exists, not as sufficiently precise profiles for immediate automated personalisation.

Future adaptive systems should therefore combine repeated observations, explicit user preferences, behavioural indicators, and direct learning outcomes. Prospective validation would also be required to determine whether adaptation based on audiovisual preferences improves learning or merely increases short-term satisfaction.

8.5 Predictive performance and model adequacy

The exact leave-one-participant-out analysis provided a secondary assessment of whether the proposed model generalised to unseen participants. Adding the population interaction reduced pooled standardised RMSE from 0.632 to 0.611. Introducing participant-specific interaction variation reduced RMSE further to 0.600, while the full bivariate model achieved an RMSE of 0.570.

The resulting 5.0% reduction relative to the participant-interaction model was modest rather than dramatic. Prediction error decreased for 20 of the 29 held-out participants, increased modestly for six, and was effectively unchanged for three. The exact LOPO ELPD difference between the full model and the participant-interaction model was 2.5 with a standard error of 2.1. Thus, the predictive evidence favoured the bivariate specification on average, but did not establish decisive superiority.

This result is methodologically informative. The value of the joint model lies not only in reducing prediction error, but also in representing the residual dependence between engagement and extraneous cognitive load and propagating uncertainty into the joint-state estimates. The relatively modest predictive improvement also guards against an overly optimistic interpretation of model complexity. A more elaborate model is justified only when it provides clearer estimands, more appropriate uncertainty quantification, or better generalisation, rather than merely producing a better in-sample fit.

The influence analysis further showed that omitting individual participants changed the magnitude but not the direction of the principal interactions. The extraneous-load interaction was somewhat more sensitive to participant omission than the engagement interaction. This sensitivity is consistent with the small sample and reinforces the importance of reporting posterior intervals and participant-level validation rather than relying on point estimates alone.

8.6 Theoretical implications

The findings extend the conventional multimedia perspective in two ways. First, they indicate that the effects of auditory and visual components may not be adequately represented by additive main effects. The statistically supported voice–presenter interaction suggests that the coherence of a multimodal configuration can be more important than the isolated source of either modality.

Second, the results emphasise that multimedia responses are distributed across participants rather than represented by a single universal effect. A population-average advantage can coexist with substantial variation in individual response. Theoretical models of AI-mediated learning may therefore benefit from distinguishing between population-level audiovisual principles and learner-specific sensitivity to particular social and perceptual cues.

The results also refine the interpretation of AI-generated instructional media. The strongest differences were associated with aesthetic appeal, reward, perceived usability, and extraneous cognitive load, whereas focused attention and the remaining cognitive-load dimensions were less clearly affected. The evidence therefore supports a relatively specific account: the tested AI-generated audiovisual configuration was associated with a more favourable perceived presentation experience, but the data do not establish broader improvements in learning, attention, or cognitive investment.

8.7 Practical implications

For developers of short instructional videos, the findings suggest that voice and presenter should be designed and evaluated jointly. Replacing only a voice or only a presenter may not reproduce the experience associated with a coordinated audiovisual configuration. Practical evaluation should therefore consider voice naturalness, pacing, facial motion, lip synchronisation, caption timing, visual style, and perceived coherence as an integrated system.

The results also caution against assuming that greater realism is always preferable. The mixed-source conditions did not consistently reproduce the engagement–load profile observed in AI–AI, even though each retained either an AI-generated voice or an AI-generated avatar. A technically realistic component may still be less effective when combined with another component that differs in timing, style, or social signalling.

At the same time, the current evidence does not justify replacing human instructors or human-presented materials. Human instructors provide adaptive explanation, interpersonal responsiveness, pedagogical judgement, and relational support that were not represented in the short controlled videos. The present findings are most relevant to the design of brief, standardised audiovisual materials rather than to the broader substitution of human teaching.

9. Limitations

Several limitations should be considered when interpreting the findings. First, the sample included only 29 learners from one institution and one proficiency level, which limits statistical precision and external validity. Although the within-participant design and Bayesian partial pooling improved estimation efficiency, they cannot replace replication in larger and more diverse samples.

Second, each factor level was represented by a single voice and a single presenter. Generation source was therefore partly confounded with voice identity, presenter identity, appearance, recording quality, and implementation-specific characteristics. The findings apply to the tested stimuli rather than to AI-generated or human-produced videos in general.

Third, the study relied on self-reported engagement and cognitive load. No direct measures of learning achievement, retention, transfer, viewing behaviour, or physiological response were collected. The more favourable engagement–load profile observed for AI–AI should therefore not be interpreted as evidence of superior learning outcomes.

Fourth, the videos were brief and covered one foreign-language topic. Responses to longer, more demanding, or repeatedly viewed materials may differ. In addition, manipulation checks for perceived humanness, naturalness, realism, lip synchronisation, and audiovisual congruence were unavailable, preventing identification of the mechanism underlying the interaction.

Fifth, participant characteristics such as age, prior exposure to AI-generated media, attitudes towards AI, and language proficiency were not available for moderator analysis. The exact presentation sequences were also unavailable, although presentation position was controlled statistically.

Finally, the midpoint-state thresholds were descriptive rather than validated educational cut-offs, and participant-specific effects were estimated from only four observations per learner. The heterogeneity results therefore support the presence of individual variation but should not be used to define stable learner profiles. Larger preregistered studies with multiple voices and presenters are required to confirm the robustness and generalisability of the findings.

10. Conclusion

This study proposed an uncertainty-aware Bayesian framework for analysing voice–presenter interactions in short instructional videos. By jointly modelling engagement and extraneous cognitive load while accounting for participant heterogeneity, the framework provided a more informative assessment than separate condition-level comparisons.

The tested AI-generated voice–avatar condition was associated with the most favourable estimated engagement–load profile. Its strongest associations occurred for aesthetic appeal, reward-related engagement, perceived usability, and lower extraneous cognitive load, whereas focused attention, intrinsic cognitive load, and germane cognitive load remained less certain.

The results also showed that engagement and extraneous cognitive load were related but distinct, and that the average advantage of AI–AI varied across participants. These findings suggest that audiovisual effectiveness depends not only on whether individual components are AI-generated or human-produced, but also on how voice and presenter function together.

The conclusions remain specific to the tested stimuli, sample, short-video format, and self-reported outcomes. Future research should employ larger and more diverse samples, multiple independently sampled voices and presenters, direct learning assessments, and objective measures of audiovisual quality to determine whether the observed interaction generalises across educational contexts.

Acknowledgements

The author gratefully acknowledges all participants whose contributions made this study possible and appreciates the support provided during the preparation of the manuscript.

Ethics statement

The experiment was approved by the Ethics Committee of Zhejiang College of Construction. All methods were performed in accordance with the relevant guidelines and regulations. All participants provided informed consent.

Data and Code availability

The data underlying this analysis were collected in the experiment reported by Zhang et al. [31]. Participant-level records are not publicly available because the applicable ethics and informed-consent conditions restrict unrestricted release of individual-level research records and because public disclosure could compromise participant privacy. De-identified derived scale scores, condition codes, variable definitions, and scale-scoring documentation may be requested from the corresponding author. The complete source code used for reverse scoring, scale construction, factorial decomposition, bivariate Bayesian hierarchical modelling, exact leave-one-participant-out validation, model comparison, sensitivity analyses, metric calculation, and figure generation is supplied as Supplementary with this article..

Funding

The author received no external funding for this research.

Author contributions

Y.S. conceived the analytical framework, developed the methodology and software, performed the statistical analyses, prepared the figures and tables, and wrote and revised the manuscript. The author approved the submitted version and accepts responsibility for the integrity of the work.

Competing interests

The author declares no competing interests.

References

- [1] UNESCO. Guidance for generative ai in education and research. Technical report, United Nations Educational, Scientific and Cultural Organization, Paris, 2023.
- [2] Richard E. Mayer. Evidence-based principles for how to design effective instructional videos. *Journal of Applied Research in Memory and Cognition*, 10(2):229–240, 2021. doi: 10.1016/j.jarmac.2021.03.007.
- [3] Pat Pataranutaporn, Valdemar Danry, Jason Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022, 2021. doi: 10.1038/s42256-021-00417-9.
- [4] Daniel Leiker, Anna R. Gyllen, Ibrahim Eldesouky, and Mutlu Cukurova. Generative ai for learning: Investigating the potential of learning videos with synthetic virtual instructors. In Ning Wang, Genaro Rebolledo-Mendez, Vania Dimitrova, Noboru Matsuda, and Olga C. Santos, editors, *Artificial Intelligence in Education: Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, volume 1831 of *Communications in Computer and Information Science*, pages 523–529. Springer, 2023. doi: 10.1007/978-3-031-36336-8_81.
- [5] Tian Xu, Yifan Liu, Yan Jin, Ying Qu, Jing Bai, Wei Zhang, and Yuxin Zhou. From recorded to ai-generated instructional videos: A comparison of learning performance and experience. *British Journal of Educational Technology*, 2025. doi: 10.1111/bjet.13530.
- [6] Torbjorn Netland, Oliver von Dzengelevski, Katalin Tesch, and Daniel Kwasnitschka. Comparing human-made and ai-generated teaching videos: An experimental study on learning effects. *Computers & Education*, 224:105164, 2025. doi: 10.1016/j.compedu.2024.105164.
- [7] Yidi Zhang, Margarida Lucas, Pedro Bem-haja, and Luis Pedro. The effect of student acceptance on learning outcomes: Ai-generated short videos versus paper materials. *Computers and Education: Artificial Intelligence*, 7:100286, 2024. doi: 10.1016/j.caeai.2024.100286.
- [8] Jihyun Kim, Kelsey Merrill, Kun Xu, and Sylvia Kelly. Perceived credibility of an ai instructor in online education: The role of social presence and voice features. *Computers in Human Behavior*, 136:107383, 2022. doi: 10.1016/j.chb.2022.107383.
- [9] Erin K. Chiou, Noah L. Schroeder, and Scotty D. Craig. How we trust, perceive, and learn from virtual humans: The influence of voice quality. *Computers & Education*, 146:103756, 2020. doi: 10.1016/j.compedu.2019.103756.
- [10] Talia Horovitz and Richard E. Mayer. Learning with human and virtual instructors who display happy or bored emotions in video lectures. *Computers in Human Behavior*, 119:106724, 2021. doi: 10.1016/j.chb.2021.106724.
- [11] Scotty D. Craig and Noah L. Schroeder. Reconsidering the voice effect when learning from a virtual human. *Computers & Education*, 114:193–205, 2017. doi: 10.1016/j.compedu.2017.07.003.
- [12] Scotty D. Craig and Noah L. Schroeder. Text-to-speech software and learning: Investigating the relevancy of the voice effect. *Journal of Educational Computing Research*, 57(6):1534–1548, 2019. doi: 10.1177/0735633118802877.
- [13] Robert O. Davis, Joseph Vincent, and Tae Jung Park. Reconsidering the voice principle with non-native language speakers. *Computers & Education*, 140:103605, 2019. doi: 10.1016/j.compedu.2019.103605.

- [14] Claudia Vallis, Sally Wilson, Daniel Gozman, and James Buchanan. Student perceptions of ai-generated avatars in teaching business ethics: We might not be impressed. *Postdigital Science and Education*, 6: 537–555, 2024. doi: 10.1007/s42438-023-00407-7.
- [15] Richard E. Mayer. *Multimedia Learning*. Cambridge University Press, Cambridge, 3 edition, 2020. doi: 10.1017/9781316941355.
- [16] Ton de Jong. Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38(2):105–134, 2010. doi: 10.1007/s11251-009-9110-0.
- [17] Heather L. O’Brien, Paul Cairns, and Mark Hall. A practical approach to measuring user engagement with the refined user engagement scale and new ues short form. *International Journal of Human–Computer Studies*, 112:28–39, 2018. doi: 10.1016/j.ijhcs.2018.01.004.
- [18] Melina Klepsch, Florian Schmitz, and Tina Seufert. Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology*, 8:1997, 2017. doi: 10.3389/fpsyg.2017.01997.
- [19] Paul-Christian Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.
- [20] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, 3 edition, 2013.
- [21] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\hat{r}$ for assessing convergence of mcmc. *Bayesian Analysis*, 16(2):667–718, 2021. doi: 10.1214/20-BA1221.
- [22] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27:1413–1432, 2017. doi: 10.1007/s11222-016-9696-4.
- [23] Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.
- [24] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press, Boca Raton, 2 edition, 2020.
- [25] Cynthia J. Brame. Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE–Life Sciences Education*, 15(4):es6, 2016. doi: 10.1187/cbe.16-03-0125.
- [26] Philip J. Guo, Juho Kim, and Rob Rubin. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the First ACM Conference on Learning at Scale*, pages 41–50, New York, 2014. ACM. doi: 10.1145/2556325.2566239.
- [27] Abdulhadi Shoufan. Estimating the cognitive value of youtube’s educational videos: A learning analytics approach. *Computers in Human Behavior*, 92:450–458, 2019. doi: 10.1016/j.chb.2018.03.036.
- [28] Margarida Lucas, Yidi Zhang, Pedro Bem-haja, and Pedro N. Vicente. The interplay between teachers’ trust in artificial intelligence and digital competence. *Education and Information Technologies*, 29: 22991–23010, 2024. doi: 10.1007/s10639-024-12772-2.
- [29] Margarida Lucas, Pedro Bem-haja, Yidi Zhang, Carmen Llorente-Cejudo, and Antonio Palacios-Rodríguez. A comparative analysis of pre-service teachers’ readiness for ai integration. *Computers and Education: Artificial Intelligence*, 8:100396, 2025. doi: 10.1016/j.caeai.2025.100396.

- 705 [30] Robert R. Whelan. Neuroimaging of cognitive load in instructional multimedia. *Educational Research*
706 *Review*, 2(1):1–12, 2007. doi: 10.1016/j.edurev.2006.11.001.
- 707 [31] Yidi Zhang, Margarida Lucas, Pedro Bem-haja, and Luis Pedro. Ai versus human-generated voices
708 and avatars: Rethinking user engagement and cognitive load. *Education and Information Technologies*,
709 30:22547–22566, 2025. doi: 10.1007/s10639-025-13654-x.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [VoicePresenterBayesianCodeProject.zip](#)