

Redesigning Math Assessment Items as Video Stories is Associated With Reduced Physiological Stress

Fettah Kiran , Paul T. Cirino , and Ioannis Pavlidis , *Fellow, IEEE*

Abstract—Foundational mathematics assessments can trigger anxiety and acute stress that discourage persistence in quantitative pathways. We introduce a video-story item format that places skill-matched mathematical problems inside short, relatable narrative videos. This video format represents a total-format design, combining relatable narrative, enacted video delivery, and temporal presentation; therefore, its observed effects reflect this combined design rather than any individual component in isolation. We evaluate whether this redesigned test experience is associated with lower physiological stress during problem solving. In a controlled study of lower division undergraduates ($N = 50$), participants completed a 36-item instrument designed around 12 skill-matched triplets—Video (V), Word (W), and Abstract (A)—while we continuously recorded multimodal affective signals during problem solving. We operationalize acute test stress as elevated autonomic arousal. We, therefore, analyze baseline-normalized cardiac dynamics [heart rate and interbeat interval (IBI/RR)-derived timing] and facial perspiration, and use facial expressions to characterize displayed valence. Cross-classified mixed-effects models accounting for participant differences and item difficulty show convergent reductions in physiological arousal across cardiac and facial-perspiration channels for the video-story format relative to conventional formats (A, W). Separate valence models show a shift toward more positive displayed valence. These affective advantages remain after controlling for time-on-item and occur without measurable loss in overall test performance. Crucially, we provide a practical procedure to derive participant-relative stress labels directly from baseline-normalized multimodal physiology, avoiding per-item self-report and enabling scalable dataset construction. Data, code, and materials are openly shared.

Index Terms—Affective computing, assessment item format, educational assessment, learning analytics, learning technologies, math anxiety, multimodal physiology, stress detection.

I. INTRODUCTION

MATH anxiety is common by adolescence: in the 2012 Programme for International Student Assessment, about 30% of students across Organization for Economic Cooperation and Development (OECD) countries reported feeling helpless when doing mathematics problems [1]. Metaanalytic

evidence further shows that higher math anxiety is reliably associated with lower math achievement [2], and secondary-school samples consistently report higher levels among girls than boys [3]. The consequences extend beyond grades: math anxiety is also intertwined with engagement and intentions to use mathematics in the future, with potential downstream implications for course-taking choices and persistence in quantitative pathways [1].

These costs often become most visible in evaluative settings, where anxiety can spike precisely when performance matters most. In a large university sample, state math anxiety rose before and during a math test and fell afterward [4]. Related work emphasizes that state (situational) and trait (enduring) components of math anxiety can relate differently to performance depending on task demands [5]. Complementing self-report, neuroimaging findings indicate that higher math anxiety predicts greater activation in pain-related neural circuitry during the anticipation of math [6]. Evaluations can also provoke acute physiological arousal—a stress response—that has been associated with reduced performance even among students with adequate mathematical competence [7]. We, therefore, distinguish *anxiety* as subjective worry/anticipation (which can include both situational and enduring components) from *stress* as an acute arousal response during test-taking: anxiety can amplify stress reactivity, but elevated arousal can also arise even when chronic anxiety is low [8], [9].

This framing has also motivated a methodological shift toward objective measures of learners' affect during quantitative work. Researchers increasingly model affective and attentional states from physiological signals collected during problem solving. For example, Pizzie and Kraemer [10] showed that electrodermal activity during math problem solving relates to performance among individuals varying in math anxiety, with this relationship moderated by emotion-regulation strategies. Choksi et al. [11] demonstrated that smartwatch physiological signals can support real-time inference of emotion and attention in affective-learning contexts. Together, this work suggests that affective responses during math tasks are measurable and can be shaped by task conditions and regulation strategies [10], [11].

Recent academic-stress research has further expanded this perspective by integrating multiple physiological, behavioral, and psychological signals. Liu et al. [12] reviewed academic-stress detection studies from 2012 to 2024 and found growing use of physiological, behavioral, psychological, wearable, and nonwearable data streams, often combined with machine-learning or deep-learning methods. This situates this study

Received 19 January 2026; revised 12 June 2026 and 2 July 2026; accepted 9 July 2026. Date of publication 21 July 2026; date of current version 24 August 2026. This work was supported by the National Science Foundation under Grant 1760760. (Corresponding author: Ioannis Pavlidis.)

Fettah Kiran and Ioannis Pavlidis are with the Affective and Data Computing Laboratory, University of Houston, Houston, TX 77204 USA (e-mail: ipavlidis@uh.edu).

Paul T. Cirino is with the Department of Psychology, University of Houston, Houston, TX 77204 USA.

This article has supplementary downloadable material available at <https://doi.org/10.1109/TLT.2026.3715735>, provided by the authors.

Digital Object Identifier 10.1109/TLT.2026.3715735

within a broader effort to model academic stress from convergent multimodal signals rather than from a single indicator alone. Academic stress–performance relationships are also complex: Zhang et al. [13] show that academic stress can relate to performance through intervening psychological factors, including depression and learning motivation, with sex moderating parts of this relationship. Thus, lower physiological arousal during assessment should not be expected to translate mechanically into higher item accuracy.

Most existing interventions aim to improve test-taking by adding brief self-regulation steps around the assessment rather than changing the items themselves. Examples include short mindfulness exercises administered before tests [14] and brief breaks integrated into tests [15]. While such interventions can help, their effectiveness can depend on implementation and context, and they may not be easy to scale within standard testing workflows. A complementary approach is to reduce physiological stress by changing the *test experience itself*—specifically, by redesigning the *presentation format* of questions so that they feel less threatening while preserving the underlying mathematical demand. Because affective responses are multidimensional, we also examine whether such redesign is accompanied by more positive expressed valence.

This design perspective aligns with learning-science results showing that context and narrative can shape how learners enter and sustain problem solving. Research on contextualization and personalization finds that problem settings and narratives can influence interest and persistence even when the underlying content is unchanged [16]. Anchored-instruction approaches similarly use rich, storylike “anchors” to situate quantitative reasoning in everyday scenarios, emphasizing the role of context in motivating and organizing problem-solving activity [17]. In assessment settings, these results motivate a straightforward hypothesis. Presenting skill-matched questions as relatable stories may reduce physiological arousal during test-taking, and secondarily, shift expressed valence in a more positive direction.

A relatable story, however, can be delivered through multiple media—for example, as a comic strip, a textual narrative, or an audio vignette. We focus on *short video stories* because video-centric platforms are pervasive among U.S. teens and young adults (e.g., YouTube is widely used), making video a practical and familiar delivery medium [18]. Evidence from learning settings further suggests that video production choices—especially keeping segments concise to manage attention and cognitive load—are associated with higher viewer engagement [19], [20]. Accordingly, we adopt concise video vignettes for scalability and practical deployment, without implying that other story media cannot be effective.

In this study, we focus on foundational mathematics assessment, that is, lower division quantitative reasoning tasks involving arithmetic, percentages, ratios, probabilities, rates, and introductory growth reasoning. We test whether the video-story format is associated with lower physiological stress during foundational mathematics assessment, without compromising performance. Specifically, we compare a *video-story* question format, in which skill-matched items are presented as short, relatable narrative videos (V), against two conventional formats:

word-based text questions (W) and abstract formula-based questions (A). Because the V condition represents a *total-format design*, combining relatable narrative, enacted video delivery, and temporal presentation, we interpret its effects as reflecting this combined design rather than any single component in isolation.

A. Research Questions

We pose the following research questions.

- RQ1: During foundational mathematics assessment, relative to conventional formats (W, A), is the video-story format (V) associated with lower within-participant autonomic arousal, our operational measure of physiological stress? As a secondary affective outcome, is V also associated with a shift toward more positive expressed valence?
- RQ2: Does the video-story format affect question-level accuracy after accounting for between-participant differences in overall performance, item difficulty, and time-on-question?

B. Contributions

To address RQ1 and RQ2, we conducted a controlled experiment in which lower division students completed a foundational math test containing video-story and conventional questions while we monitored multimodal affective responses and performance. A preliminary pilot study using the first 20 participants was reported as an extended abstract at the 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos [21]; this article reports the full study with expanded sensing and analysis. Our contributions are as follows.

- 1) A novel *video-story* question format for foundational math tests, designed to make the test experience less physiologically stressful.
- 2) Multimodal evidence that the video-story format (V) is associated with lower physiological arousal relative to conventional formats (W, A) across multiple baseline-normalized channels, accompanied by detectable shifts toward more positive displayed valence.
- 3) Robustness results showing that this arousal association is not explained by time-on-question and holds across continuous-signal and participant-relative stress-label models, including stricter-threshold checks.
- 4) Evidence of heterogeneity in benefit: some participants show substantially larger arousal reductions than others, motivating future personalization.
- 5) After accounting for individual and question-level differences, we find no overall difference in accuracy by format, suggesting lower arousal *without sacrificing performance*.
- 6) An openly accessible multimodal dataset annotated with relative stress labels.¹

¹[Online]. Available: <https://github.com/UH-ACDC/affective-math-test-repo>

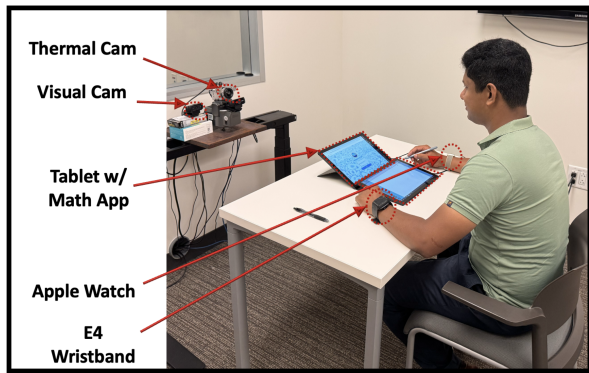


Fig. 1. Experimental setup. Shown is a subject preparing to complete the foundational math test. All data collection sensors are highlighted.

II. STUDY DESIGN

A. Participants and Procedure

We recruited lower division undergraduate students in the Houston, TX area ($N = 50$) under a protocol approved by the University of Houston Institutional Review Board (IRB). Recruitment was carried out through in-class announcements. To reduce the likelihood that unusually elevated academic stress would carry into the laboratory session, experimental sessions were not scheduled during formal course test periods. Upon providing informed consent, participants completed a four-stage procedure as follows:

- 1) a resting baseline session;
- 2) a pretest state anxiety assessment;
- 3) a foundational mathematics test administered via a tablet-based application; and
- 4) a posttest usability evaluation of the math application.

During the baseline and math test sessions, we recorded physiological signals and facial video (see Fig. 1). Physiological measures included facial perspiration signal (FP) from an FLIR Tau 640 thermal camera, heart rate via the Apple Watch (HR_{AW}), and heart rate via an Empatica E4 wristband (HR_{E4}), worn on the hand opposite the Apple Watch. The E4 also provided interbeat interval (IBI/RR) data. Facial video was recorded to estimate facially expressed emotion and derive displayed valence, while FP, HR_{AW}/HR_{E4} , and the IBI/RR-derived cardiac timing channel served as proxies for arousal.

These signals index autonomic nervous system activity through complementary pathways. Facial perspiration reflects sympathetic sudomotor activation, whereas heart rate reflects cardiac autonomic regulation shaped by both sympathetic and parasympathetic influences [22], [23]. The IBI/R-R-derived cardiac timing channel comes from the same inter-beat intervals used to compute heart rate variability (HRV) metrics, but we do not interpret it as a conventional window-based HRV estimate of vagal tone or autonomic balance. Using multiple channels therefore provides convergent evidence for arousal changes during test-taking and helps mitigate single-sensor artifacts or modality-specific confounds.

To increase measurement reliability and improve scalability, we employed two distinct heart-rate monitors. This redundancy served the following three purposes:

- 1) to mitigate technical failures;
- 2) to cross-validate heart-rate data via agreement across channels; and
- 3) to assess whether a widely available consumer device (Apple Watch) produces results comparable to those from a research-grade device (E4).

Time-on-question (duration logging): Because video narratives impose a minimum viewing time, exact duration-matching across formats would require altering the materials (e.g., adding filler text or truncating narratives). Such edits would change the test experience and reduce ecological validity. Instead, the tablet application logs item start and submission timestamps to compute time-on-question for every item. Here, time-on-question reflects the *total* time from item onset to final submission (including viewing/reading, deliberation, and response entry, as well as any in-app interactions), rather than a pure viewing/reading-time measure. Duration-based explanations are evaluated explicitly in the analysis using this covariate.

B. Resting Baseline and Pretest State Anxiety

The baseline session aimed to provide a consistent participant-specific reference level prior to the math test. Because participants could arrive with elevated arousal due to environmental novelty or preceding daily activities, baseline recording did not begin immediately upon arrival. Participants first completed study briefing, laboratory familiarization, wearable-sensor fitting, sensor testing, and signal-quality checks. This setup and acclimatization period typically lasted approximately 30 min before the formal baseline recording began. Participants then sat quietly and listened to soothing music for five minutes in the lab.

We, therefore, interpret the resulting baseline as a practical participant-specific no-task reference rather than as an absolute physiological ground state. Because physiological arousal during such a resting period can still vary substantially across individuals, absolute physiological values during task performance can be difficult to interpret; accordingly, all physiological analyses used participant-specific baseline-normalized signals [24].

Following the baseline session, participants completed the state anxiety scale (STAI Form Y-1) [25] to quantify pretest state anxiety.

C. Foundational Math Test Design and Administration

Participants completed a 36-item foundational mathematics test administered on a tablet. Here, “foundational mathematics” refers to lower division quantitative reasoning skills commonly expected in secondary-school completion and early undergraduate coursework, including arithmetic, percentages, ratios, probabilities, rates, proportional reasoning, and introductory compound-growth reasoning. The assessment targeted applied quantitative reasoning and moderate multistep calculation rather than advanced algebraic proof, formal calculus, or higher theoretical mathematics. The test compared three presentation formats—*Abstract (A)*, *Word (W)*, and *Video (V)*—with 12 items per format. For item design, the 36 questions were grouped into 12 skill-matched A/W/V triplets, each targeting the same underlying mathematical construct while varying representation

and delivery. For test administration, however, items were not presented as triplets; they were presented in format-specific 12-item blocks. Block order was counterbalanced across participants using a cyclic rotation: $A \rightarrow W \rightarrow V$, $W \rightarrow V \rightarrow A$, and $V \rightarrow A \rightarrow W$. These orders were assigned at random; in the realized sample, group sizes were unequal due to finite-sample randomness ($n = 23, 14, 13$). Within each block, question order was fixed across participants.

Abstract items presented symbolic or decontextualized quantitative tasks. Word items presented contextualized text-based problems. Video items presented analogous applied problems as short, protagonist-centered narrative videos. Both W and V items were contextualized, but V was designed to be more relatable through enacted micronarratives with identification cues, such as character, goal, and consequence. Contextualized items drew equally from two young-adult U.S. story domains: *health and medical* contexts and *personal-finance and consumer-choice* contexts, with six concepts in each domain.

To support authenticity, the videos were performed by student actors from the University of Houston School of Theater and Dance, with casting reflecting the racial and ethnic diversity typical of a major U.S. city. Full item text, video-scenario descriptions, and YouTube links are provided in Supplementary Table S2; the Abstract–Word–Video triplet mapping for the 12 skill-matched themes is provided in Supplementary Table S1.

The design aimed to preserve construct-level mathematical demand across formats while varying context, modality, and temporal presentation. A representative cross-format example is provided in Supplementary Section S1. The 12 skill-matched themes span three broad cognitive-demand categories—procedural fluency, applied quantitative interpretation, and multistep applied reasoning; Supplementary Table S1 shows the category assigned to each theme. The assessment therefore sampled foundational quantitative reasoning across these categories rather than advanced proof-based, graph-theoretic, or open-ended mathematical reasoning.

Although the A , W , and V variants were aligned at the construct level, Video items unfolded temporally and required integration of information presented over time.

D. Posttest Usability Evaluation

Upon completing the math test, participants rated the usability of the application using the system usability scale (SUS) [27], a validated ten-item Likert questionnaire yielding a score from 0 to 100 (with 68 commonly used as an “above-average” benchmark). The app was designed following the Apple Human Interface Guidelines for intuitive interfaces [28]; Fig. 2(a) illustrates the self-explanatory test interface. SUS scores were high overall [mean $\pm$ standard deviation (SD): 85.15 ± 15.14 ; Fig. 2(b)], with almost all scores exceeding 68, indicating above-average usability [26]. Moreover, the mean score exceeds the threshold commonly associated with *excellent* usability (e.g., > 80) [26], alleviating concern that usability issues in the app could confound the observed stress responses.

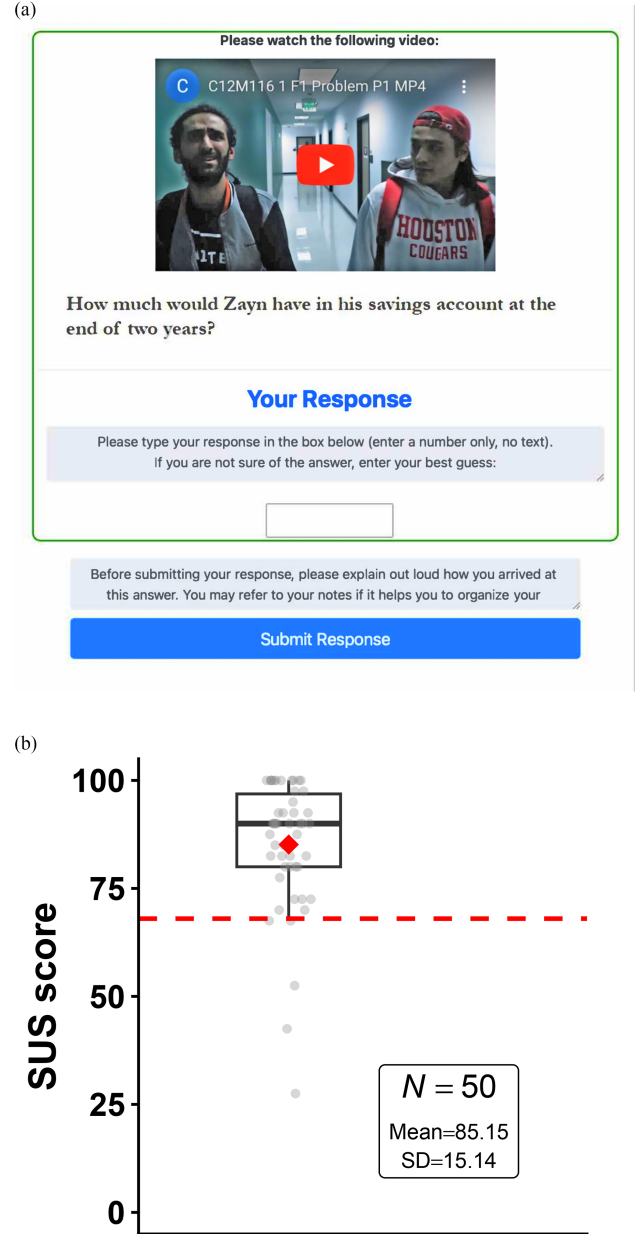


Fig. 2. Usability of math test app. (a) Example screen showing the tablet app delivering a V (Video) question. (b) Distribution of participants' SUS scores. The red dashed line marks the widely used usability threshold of 68 [26]; the red diamond marks the sample mean.

III. METHODS

A. Facial Perspiration Extraction From Thermal Video

To reduce motion-induced artifacts due to natural head movement, we tracked a fixed facial tissue region in the thermal stream using the tissue tracker of Zhou et al. [29]. Tracking was initialized by manually selecting the mandibular region of interest (ROI) in the first frame (or the lower nose ROI when the mandibular region was occluded). In subsequent frames, the tracker estimated the best-matching ROI using spatiotemporal smoothing [29]. When the facial ROI could not be reliably tracked because of severe head rotation, occlusion, or temporary

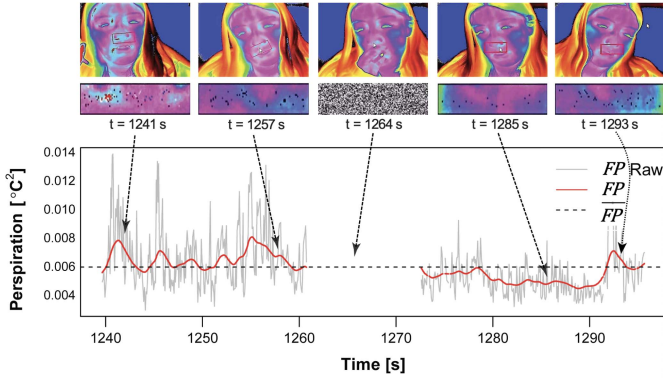


Fig. 3. Extraction of FP from the mandibular ROI of participant S10 during one math question. (Top) Thermal snapshots illustrating head motion and brief occlusions (e.g., hand covering the ROI). (Middle) Corresponding mandibular ROIs (red boxes). (Bottom) Raw FP (gray) and FFT-filtered FP (red). When reliable extraction is not possible (e.g., ROI occlusion around $t = 1262\text{--}1271$ s), the algorithm returns NA for those frames, preserving measurement validity.

loss of the mandibular/lower nose region, the corresponding FP samples were treated as missing. No time-series interpolation or imputation was applied to these missing FP samples, because filling frames with unreliable facial tracking could introduce artificial sudomotor dynamics. Question-level FP summaries were, therefore, computed from the valid FP samples available within each question window.

Within the tracked ROI, activated perspiration pores appear as cold (dark) spots against warmer surrounding tissue. Prior work demonstrated that thermal detection of facial perspiration in the mandibular/lower nose region provides a remote measure of sudomotor activity associated with sympathetic arousal [30]. We quantified this phenomenon using a morphology-based algorithm [30], producing the facial perspiration signal $FP(t)$ (see Fig. 3). To suppress high-frequency noise arising from residual tracking imperfections [29], we applied a low-pass fast Fourier transform (FFT) filter with cutoff $f = 0.4$ Hz [31]. Fig. 3 illustrates the raw and filtered FP signals for a representative participant.

B. Facially Displayed Emotions From RGB Video

We estimated *facially displayed* emotions from the red-green-blue (RGB) face video using a convolutional neural network (CNN) previously described in [32]. For participant i at time t , the CNN outputs a probability vector $\vec{\mathcal{E}}(i, t)$ over seven basic emotion categories {Neutral, Surprised, Sad, Happy, Afraid, Disgusted, Angry}, with components $e_{i,t,k} \in [0, 1]$ and $\sum_{k=1}^7 e_{i,t,k} = 1$.

1) *External Validation of the CNN*: As detailed in the supplementary materials of [32], we validated the displayed-emotion CNN against the Ryerson audio-visual database of emotional speech and song (RAVDESS) [33], a dynamic video dataset that better matches our time-series setting than still-image benchmarks. In that validation, the CNN was run on 64 708 RAVDESS facial frames (speech subset), spanning silent and speaking segments; “calm” clips were excluded to match the network’s

output taxonomy. Using the dataset’s frame-level ground truth, the CNN achieved an overall accuracy of 77.82%, supporting its use for labeling *displayed* emotion dynamics in our study context.

2) *Interpretation: Displayed Versus Felt Affect*: We emphasize that these outputs quantify *facial affective display*. Facial expressions do not map one-to-one onto experienced emotions, especially in social interaction where display-management pressures can shape “social faces” [34], [35]. In our protocol, participants completed the test alone, which reduces interpersonal audience effects and should make facial behavior less socially filtered. Nevertheless, visual emotion labels still require careful interpretation in high-cognitive-load mathematics tasks. Although the CNN used here was externally validated on RAVDESS, labels such as AFRAID, ANGRY, or SAD should be understood as classifications of facial configurations, not as literal reports of conscious fear, anger, or sadness. Concentration, confusion, uncertainty, and effortful problem solving can produce facial configurations that overlap with canonical negative-expression categories.

To contextualize this issue, Supplementary Fig. S2 compares facial-expression distributions produced by the same CNN in the present affective math-test study and in a separate affective essay-writing study reported in [36]. The essay-writing condition showed relatively higher SAD classifications, whereas the math-testing condition showed relatively higher AFRAID classifications. We interpret this contrast not as evidence that essay-writing participants were literally sad or that math-test participants were literally fearful, but as evidence that different academic tasks can evoke different facial-display morphologies under cognitive load.

This task-dependent interpretation is also consistent with established differences in facial morphology across negative-expression categories described in the facial action coding system. SAD displays are typically characterized by a more subdued configuration involving brow and lip-corner movements, whereas AFRAID displays involve a broader pattern including raised brows, widened eyes, eyelid tension, and lip stretching [37]. Accordingly, reflective essay writing may produce a more inwardly focused facial-display morphology, whereas mathematical problem solving may produce a more vigilant, tense, or uncertainty-monitoring morphology.

C. Time-Series Preprocessing

All channels were time-aligned during data acquisition using NetTime [38] but sampled at different rates. To enable integrated analysis, we resampled all channels to a common 1-Hz grid using mean aggregation within each 1-s bin. This resolution matches the dynamics of the cardiac channels and is sufficient for the slower physiological fluctuations of interest [39].

Concretely, FP (8 fps thermal) was averaged to 1 Hz; the RGB-based emotion vectors were produced at an effective 10 fps and mean-pooled element-wise to 1 Hz; HR_{E4} was already at 1 Hz; and the E4 IBI/RR series ($HRV.IBI$) was binned to 1 Hz for alignment with the common time grid. Apple Watch heart rate (HR_{AW}) arrives approximately every 5 s, so at 1 Hz, it

TABLE I
CHANNELWISE OBSERVATIONS ACROSS THE PROCESSING PIPELINE

Channel	Sec-Level	Cross-Channel	Cook's Filter	Q-Level
FP	152 311	NA	NA	1794
HR _{E4}	170 145	34 071	32 367	1738
HR _{AW}	34 505	34 071	32 367	1738
NHRV	95 682	NA	NA	1666
$\vec{E}$	172 321	NA	NA	1661

NHRV denotes the normalized IBI/RR-derived cardiac timing channel used in the analyses, not a conventional window-based HRV estimate. NA indicates that cross-channel matching or Cook's-distance filtering was not applicable to that channel.

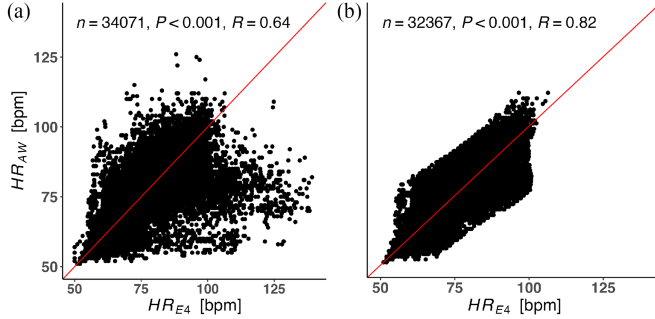


Fig. 4. HR quality control via cross-channel agreement. (a) HR_{AW} versus HR_{E4} for timestamps observed by both devices before Cook's-distance filtering ($R = 0.64$). (b) After Cook's-distance outlier removal from the matched HR pairs, agreement improves ($R = 0.82$).

contains intermittent valid readings with substantial missingness (see Table I). Importantly, this 1-Hz binning step did not compute conventional HRV metrics within each question window. Conventional HRV indices typically require longer windows to estimate parasympathetic or vagal modulation reliably. In contrast, our procedure used the available IBI/RR samples as a beat-to-beat cardiac timing signal, binned only for temporal alignment with the other channels. Therefore, the resulting channel should be interpreted as an IBI/RR-derived cardiac timing measure, not as a stable item-level estimate of cardiac vagal tone or autonomic balance.

We exploited dual HR sensing for quality control. Because HR_{AW} and HR_{E4} should agree under normal conditions, we paired the two streams only at timestamps where both devices reported heart rate (cross-channel in Table I; Fig. 4(a)). This timestamp-matching step explains most of the apparent reduction in HR sample count. The E4 reported heart rate at approximately 1 Hz, whereas the Apple Watch reported heart rate intermittently. Consequently, matching the E4 stream to Apple Watch observations reduced the paired HR set from 170 145 E4 samples to 34 071 synchronized pairs. The matched Apple Watch and E4 heart-rate streams showed strong agreement: Cook's-distance filtering removed only 1 704 of 34 071 matched pairs, leaving 32 367 paired observations (95.0% retained). We, therefore, treated the removed 5.0% as influential discordant observations and retained the remaining paired samples for downstream analysis (Cook's Filter in Table I; Fig. 4(b)). Inspection of the synchronized participant videos suggested that the removed discordant samples were driven primarily by transient

wrist-motion artifacts, such as abruptly moving the sensor-worn wrist, rather than by systematic device failure. This limited artifact filtering increases measurement reliability by reducing the influence of such transient artifacts.

For FP, missing samples caused by temporary ROI loss were excluded before 1-s aggregation and question-level averaging. These missing intervals were rare and brief: across question windows, the median proportion of missing FP samples was 2.55%. Because participants generally maintained visual attention on the tablet positioned near the thermal and RGB cameras, occasional tracking losses represented short interruptions relative to the longer question-level observation windows and were unlikely to materially affect the aggregated FP measures used in the continuous-signal and stress-label analyses.

Time-on-question: For each participant i and item j , the tablet application logged question start and submission timestamps, yielding time-on-question $QTIME(i, j)$. We retain $QTIME$ at the question level because the video format imposes a minimum viewing duration and may confound format comparisons if not controlled for in analysis.

D. Baseline Calibration and Normalization

Physiological stress is operationalized as elevated physiological arousal relative to an individual reference condition [40]. Because baseline levels vary substantially across individuals [41], we calibrated each channel relative to each participant's baseline session. This session served as a no-task reference state rather than an absolute physiological ground state. As described in Section II, recording followed laboratory acclimatization, sensor fitting, and signal-quality checks; subsequent physiological measurements are therefore interpreted relative to each participant's own no-task condition.

For facial perspiration and heart rate, we subtract participant-specific baseline means (computed over the baseline session). Specifically, for FP, HR_{AW}, and HR_{E4}, we compute baseline-calibrated signals by centering each time series at the corresponding baseline mean. For the cardiac timing channel, we use the E4 IBI/RR series ($HRV.IBI$) and normalize it by the concurrent HR_{E4}, following prior work that accounts for the dependence between cardiac timing variability and average heart rate [42] and our prior stress-testing pipeline [43]:

$$CFP(i, t) = FP(i, t) - \overline{FP}_{BL}(i) \quad (1)$$

$$CHR_{AW}(i, t) = HR_{AW}(i, t) - \overline{HR}_{AW,BL}(i) \quad (2)$$

$$CHR_{E4}(i, t) = HR_{E4}(i, t) - \overline{HR}_{E4,BL}(i) \quad (3)$$

$$CHRV(i, t) = HRV.IBI(i, t) / HR_{E4}(i, t) \quad (4)$$

The calibrated signals are not generally *normally distributed* (see Supplementary Fig. S1), so prior to modeling, we apply ordered quantile normalization (OQN) [44], yielding approximately normal channels NFP, NHR_{AW}, NHR_{E4}, and NHRV. We retain the notation NHRV for continuity with the analysis tables and figures, but interpret it as the normalized IBI/RR-derived cardiac timing channel rather than as a conventional window-based HRV estimate. Supplementary Fig. S1 shows the

corresponding QQ plots before and after OQN. OQN was used only as a distributional transformation for inferential modeling; it is not a temporal smoothing operation. It does not average neighboring samples, apply a moving window, or remove local fluctuations from the physiological trajectories. Instead, OQN preserves the empirical rank ordering of observations while mapping their marginal distribution onto approximately normal quantiles. Thus, observations that are high or low relative to the corresponding channel distribution remain high or low after transformation, although the numerical spacing among observations, especially in the tails, is rescaled.

This distinction is important because acute physiological peaks may carry meaningful descriptive information. We therefore interpret raw baseline-calibrated signals as the appropriate scale for visualizing physiological amplitudes, including local peaks, whereas OQN-transformed values are used for mixed-effects inference. Several calibrated physiological channels were skewed and contained extreme values; entering these variables directly into Gaussian mixed-effects models would make estimates, standard errors, confidence intervals, and hypothesis tests more vulnerable to tail-driven leverage and distributional misspecification. Effects that remain detectable after OQN, therefore, provide a more conservative test of the underlying association: the transformation reduces undue influence from extreme tail values while preserving the temporal structure and rank ordering of the observations.

For notational convenience, we refer to any one of these normalized arousal channels as $NPS_c \in \{NFP, NHR_{AW}, NHR_{E4}, NHRV\}$, and denote its question-level mean for participant i on item j by $\overline{NPS}_c(i, j)$. In particular, we compute the within-question means $\overline{NFP}(i, j)$, $\overline{NHR}_{AW}(i, j)$, $\overline{NHR}_{E4}(i, j)$, and $\overline{NHRV}(i, j)$ for downstream analysis.

E. Relative Stress Labeling

Stress labeling is necessary to summarize evolving physiological stress during testing. Per-question self-report would burden participants and disrupt the natural testing process, and self-report can be subjective [45], [46]. We therefore adopt a physiologically driven labeling strategy.

In prior naturalistic work (STREL), we placed stress thresholds at two standard deviations from the baseline [47]. Here, the study is a low-stakes laboratory task and physiological shifts are more moderate, as illustrated by the representative question-level physiological time series in Fig. 5; applying a $\pm 2\sigma$ rule would yield very few “stress” points. We therefore define *relative* stress within-participant by thresholding each participant’s question-level distribution at its mean: values above the mean are labeled relatively higher stress (RHS), and values at or below the mean are labeled relatively lower stress (RLS). For NHRV, lower values correspond to higher arousal, so the inequality is reversed; Fig. 6 illustrates the resulting within-participant dichotomization for one representative participant. Participant-level question-series plots for FP, HR_{AW} , and HR_{E4} (analogous to Fig. 5) are provided in Supplementary Section S10.

For each participant i , let $\overline{NFP}(i, j)$, $\overline{NHR}_{AW}(i, j)$, $\overline{NHR}_{E4}(i, j)$, and $\overline{NHRV}(i, j)$ denote question-level means,

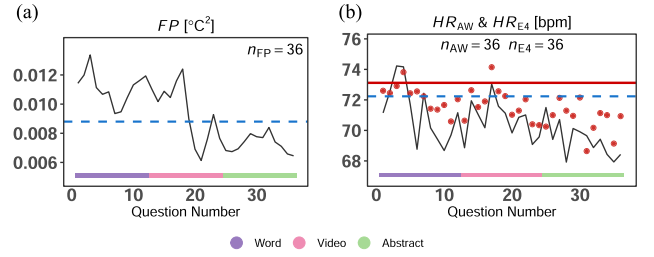


Fig. 5. Question-level physiological time series of participant S006 during the math test. (a) Facial perspiration across questions. (b) Heart rate across questions, with red points showing HR_{AW} and the black curve showing HR_{E4} . Horizontal reference lines indicate participant-specific baseline values obtained during the resting baseline session. Colored background bands identify the format block being completed (Word, Video, or Abstract). The figure illustrates the magnitude and temporal variability of the physiological signals observed during testing.

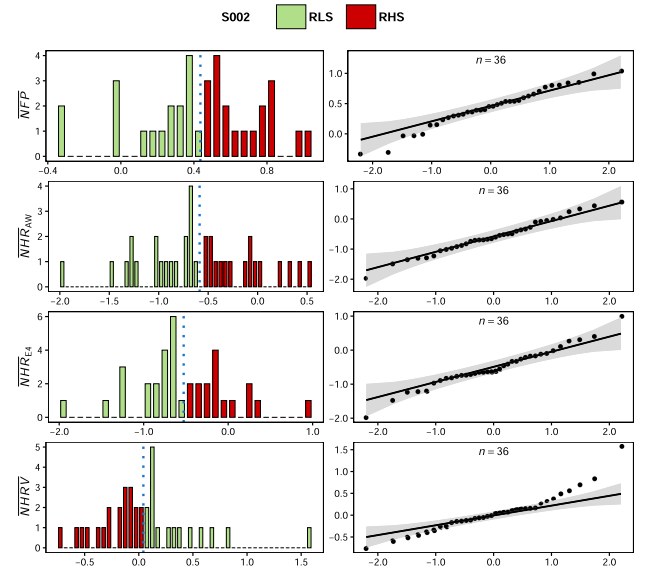


Fig. 6. Channel-specific relative stress labeling for participant S002. (Left) Histograms of question-level mean arousal values across the 36 questions. The dotted vertical line marks the participant’s mean for that channel, used as the RHS/RLS threshold. For NHRV, the labeling direction is reversed because lower values are treated as relatively higher arousal. (Right) QQ plots indicate that the OQN-transformed question-level aggregates are approximately normal, making the mean an appropriate within-participant reference point for RHS/RLS labeling.

and let $\overline{NFP}(i)$, $\overline{NHR}_{AW}(i)$, $\overline{NHR}_{E4}(i)$, and $\overline{NHRV}(i)$ denote participant-level means across all questions. We assign channel-specific relative stress labels $\mathcal{L}$ by

$$\mathcal{L}_{FP}(i, j) \equiv \begin{cases} \text{RHS}, & \overline{NFP}(i, j) > \overline{NFP}(i) \\ \text{RLS}, & \text{otherwise} \end{cases} \quad (5)$$

$$\mathcal{L}_{HR_{AW}}(i, j) \equiv \begin{cases} \text{RHS}, & \overline{NHR}_{AW}(i, j) > \overline{NHR}_{AW}(i) \\ \text{RLS}, & \text{otherwise} \end{cases} \quad (6)$$

$$\mathcal{L}_{HR_{E4}}(i, j) \equiv \begin{cases} \text{RHS}, & \overline{NHR}_{E4}(i, j) > \overline{NHR}_{E4}(i) \\ \text{RLS}, & \text{otherwise} \end{cases} \quad (7)$$

$$\mathcal{L}_{HRV}(i, j) \equiv \begin{cases} \text{RHS}, & \overline{NHRV}(i, j) < \overline{NHRV}(i) \\ \text{RLS}, & \text{otherwise} \end{cases} \quad (8)$$

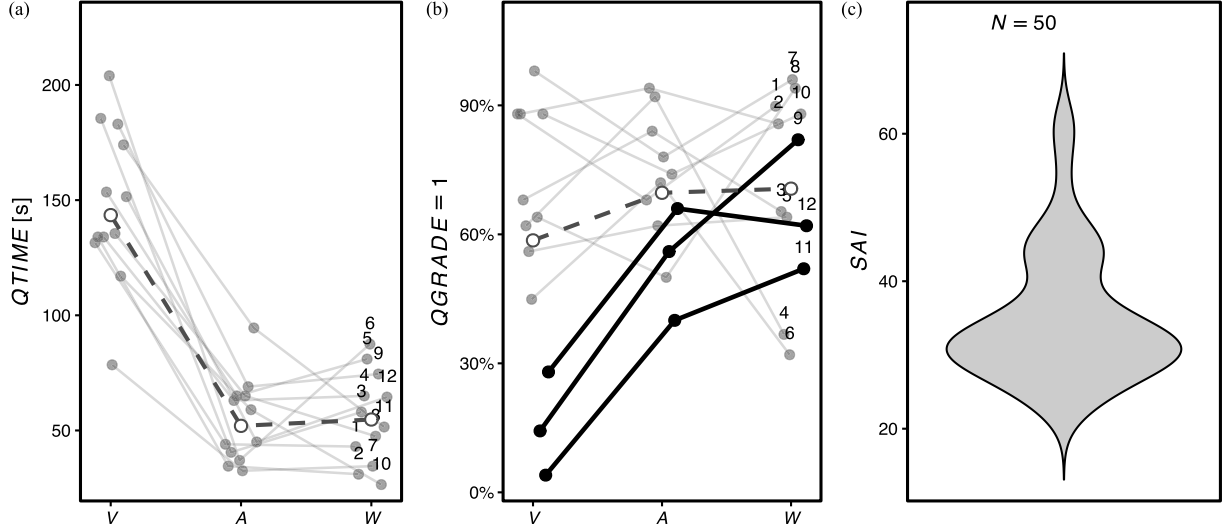


Fig. 7. Exploratory graphs. (a) Mean time-on-question (QTIME [s]) by format for each of the 12 matched triplets (gray polylines; labels 1–12); dashed line shows the overall format mean. (b) Accuracy (QGRADE = 1) by format for each triplet (gray); the three lowest performing Video triplets ($k = 9, 11, 12$) are shown in bold; dashed line shows the overall format mean. (c) Violin plot of SAI.

1) *Robustness. Stricter Percentile-Based Stress Labels:* Mean-threshold labeling yields a balanced within-participant split (approximately half RHS / half RLS per participant and channel), which is useful for statistical power in a low-stakes laboratory task where physiological shifts are moderate. However, one may prefer a stricter definition in which only the most elevated arousal observations are labeled as “higher stress” (addressing the concern that mean-splitting may classify mildly elevated arousal as RHS). We, therefore, report robustness checks using a percentile-based scheme computed *within participant and channel*.

Concretely, for each participant i and channel c , we assign the higher stress label only to the upper tail of that participant’s question-level distribution $\{\text{NPS}_c(i, q)\}_q$. For NFP, NHR_{AW}, and NHR_{E4}, observations at or above the 67th percentile (top third of questions for that channel) are labeled RHS, and the remainder are labeled RLS. For NHRV, where lower values correspond to higher arousal, the inequality is reversed: values at or below the 33rd percentile are labeled RHS, and values above are labeled RLS. All stress-label analyses are reestimated under this stricter scheme to verify that the main format effects are not an artifact of the mean-threshold choice.

IV. RESULTS

A. Collinearity Check

Before modeling, we examined the pairwise correlation matrix of all predictors to assess collinearity. We found no strong correlations among any pair ($|r| \leq 0.5$), so we retained the full predictor set.

B. Descriptive Statistics

The sample included 50 participants, balanced in terms of sex (24 males/26 females). We summarized the distributions of

time-on-question (QTIME), accuracy (QGRADE; correct = 1, incorrect = 0), pretest anxiety (SAI), and facially expressed emotions, and we reported formatwise descriptives using the skill-matched triplet structure of the test (see Fig. 7).

Question format is a categorical variable with levels $\text{QTYPE} \in \{V, A, W\}$. Ideally, 600 questions would be expected per format (V, A, and W), for a total of 1800. Six questions were lost early in the study due to a transient app malfunction that was promptly fixed. No other questions were lost in the FP channel, but the cardiovascular and valence channels include additional losses due to occasional sensor malfunctions. Question-level missingness ranges between 0.33% and 7.72% depending on the channel (Q-level in Table I) and does not materially affect the modeling results.

1) *Skill-Matched Triplets (Paired Comparisons):* We use the 12 matched triplets for within-triplet descriptive comparisons of time and accuracy; the triplet-to-item mapping and full prompts (including video links) are provided in Supplementary Tables S1 and S2. For interpretability, the 12 triplets can be grouped into three dominant cognitive-demand categories: procedural fluency ($k = 2, 3, 7, 10$), applied quantitative interpretation ($k = 1, 4, 5, 6$), and multistep applied reasoning ($k = 8, 9, 11, 12$); this classification is summarized in Supplementary Table S1.

2) *Time-On-Question Within Matched Triplets:* Fig. 7(a) shows mean time-on-question (QTIME [s]) for each format *within each triplet* ($k = 1, \dots, 12$). Across essentially all matched triplets, the Video variant requires more time than its Abstract and Word counterparts, largely because participants must first watch the video clip before responding.

3) *Accuracy Within Matched Triplets:* For each triplet k and format $F \in \{A, V, W\}$, we computed accuracy as

$$\text{Acc}_F(k) = \frac{1}{n_F(k)} \sum_{i=1}^{n_F(k)} \mathbb{I}[\text{QGRADE}_{i,k,F} = 1] \times 100\%$$

and summarize Video's within-triplet advantage/disadvantage by the *accuracy differential*

$$\Delta_V(k) = \text{Acc}_V(k) - \frac{\text{Acc}_A(k) + \text{Acc}_W(k)}{2}$$

where more negative values indicate that the Video item underperforms its matched Abstract/Word counterparts.

Fig. 7(b) plots the within-triplet accuracies $\text{Acc}_F(k)$ (percent correct) for each format, and Supplementary Table S3 reports all 12 matched triplets ranked from worst to best for Video by $\Delta_V(k)$. Three Video items [triplets $k = 9, 11, 12$; highlighted in Fig. 7(b)] stand out as clear underperformers, with very low Video accuracy (14.3%, 4.0%, and 28.0%, respectively) and large negative differentials [$\Delta_V = -54.7, -42.0$, and -36.0 percentage points (pp)]. In the remaining nine triplets, Video performance is broadly comparable to the conventional formats and sometimes better: $\Delta_V(k)$ is close to zero for most triplets and becomes positive for several cases (ranks 10–12). Overall, the paired descriptives suggest that Video performance depends on item construction: most Video items perform on par with (and sometimes better than) their matched counterparts, whereas a small subset yields markedly lower accuracy, highlighting where video-story design can be improved.

4) *Pretest Anxiety (SAI)*: Pretest anxiety scores were $\text{SAI} = 35.00 \pm 8.93$ and were largely in the typical range (30–40), with a small number of participants below 30 and above 40 [see Fig. 7(c)].

5) *Facial Valence (Expressed Emotions)*: We next summarize facially expressed emotions as a descriptive proxy for valence during problem solving. Supplementary Fig. S3a shows the proportions of facially displayed emotions for each participant by question format. Descriptively, NEUTRAL expressions are common, but emotional expressions are also present, notably AFRAID and HAPPY. We focus on these two displayed-expression categories because HAPPY is the predominant positive facial-display label, whereas AFRAID is relevant to the math-anxiety literature, which characterizes mathematics anxiety in terms of tension, apprehension, or fear related to manipulating numbers and solving mathematical problems [7], [48], [49]. We test whether the *V* format is associated with higher HAPPY display and lower AFRAID display in Section IV-C4; Supplementary Fig. S3b1–b4 provides illustrative high-confidence examples of these CNN-assigned display labels. Other emotions such as DISGUSTED and SURPRISED are much less prevalent in these distributions (see Supplementary Fig. S3a).

C. Models

1) *Continuous-Signal Mixed-Effects Models (RQ1)*: We report the predictors of question-level physiological responses recorded while participants solved the math test. Physiological arousal was quantified through four baseline-normalized question-level means, $\overline{\text{NPS}}_c(i, j)$, where $c \in \{\text{NFP}, \text{NHR}_{\text{AW}}, \text{NHR}_{\text{E4}}, \text{NHRV}\}$, representing facial perspiration from thermal imagery (NFP), heart rate recorded through the Apple Watch (NHR_{AW}), heart rate recorded through the E4 (NHR_{E4}), and the normalized IBI/RR-derived cardiac timing channel (NHRV), respectively. Here, “baseline-normalized”

indicates that each participant's question-level signal was expressed relative to that participant's corresponding baseline level. The overbar denotes the within-question temporal mean, i.e., each signal is averaged over the duration of question j . These four channels provide complementary correlates of arousal; in the context of math testing, elevated arousal is often associated with mental stress.

In implementation, we model the channel-standardized response

$$z\overline{\text{NPS}}_c(i, j) \equiv \frac{\overline{\text{NPS}}_c(i, j) - \mu_c}{\sigma_c} \quad (9)$$

where μ_c and σ_c are the sample mean and standard deviation of $\overline{\text{NPS}}_c$ over all (i, j) observations included in the analysis for channel c . Thus, all effects are expressed in within-channel standard deviation units.

To analyze these question-level physiological signals, we present the corresponding continuous-signal models. Unlike the stress-label (dichotomized) analyses, these models operate directly on the recorded signals (i.e., without thresholding), and thus, provide a complementary, less assumption-dependent view of question-format effects. For each channel c , we fit linear mixed-effects models to the *continuous* outcomes $z\overline{\text{NPS}}_c(i, j)$ defined in (9). All continuous covariates were z-scored; fixed-effect coefficients are, therefore, standardized effect sizes (SD change in the outcome per 1 SD change in a covariate, or the SD difference for each format contrast relative to Video). For each channel c , we fit the prespecified *full* model in (10) (with *V* as the reference level for question format) and report inference from this model.

$$\begin{aligned} z\overline{\text{NPS}}_c(i, j) = & b_0 \\ & + b_A \mathbb{I}[\text{QTYPE}(i, j) = A] + b_W \mathbb{I}[\text{QTYPE}(i, j) = W] \\ & + b_T z \log(\text{QTIME}(i, j)) \\ & + b_G \mathbb{I}[\text{QGRADE}(i, j) = 1] + b_S \mathbb{I}[\text{SEX}(i) = F] \\ & + b_P z\text{SAI}(i) + u_{\text{PID}(i)} + v_{\text{QID}(j)} + \varepsilon_{ij} \end{aligned} \quad (10)$$

where $u_{\text{PID}(i)} \sim \mathcal{N}(0, \sigma_u^2)$ and $v_{\text{QID}(j)} \sim \mathcal{N}(0, \sigma_v^2)$ are random intercepts for participant and question (item), respectively, and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is the residual term. Fig. 8(a1)–8(a4) visualizes the estimated format effects for each continuous physiological channel.

a) *Log-transforming time-on-question*: We use $\log(\text{QTIME})$ because response times are strictly positive and typically right-skewed; logging reduces leverage from unusually long durations and often yields a more nearly linear association on the model scale.

b) *Key predictors, potential confounders, and controls*: The *key predictors* for the research question are the question-format indicators $\mathbb{I}[\text{QTYPE}(i, j) = A]$ and $\mathbb{I}[\text{QTYPE}(i, j) = W]$, which quantify contrasts relative to the Video baseline. Time-on-question, $z \log(\text{QTIME}(i, j))$, is included as a *potential confounder*: item duration differs by format and can itself affect physiological arousal, so adjusting for time directly tests whether the format effect is an artifact of longer exposure. The remaining fixed effects in the full model, $\mathbb{I}[\text{QGRADE}(i, j) = 1]$, $\mathbb{I}[\text{SEX}(i) = F]$, and $z\text{SAI}(i)$, serve as *control covariates*

TABLE II
CONTINUOUS-SIGNAL LINEAR MIXED-EFFECTS MODELS (BASELINE-NORMALIZED, Z-SCORED OUTCOMES)

Predictor	NFP				NHR _{AW}				NHR _{E4}				NHRV			
	<i>b</i>	SE	<i>t</i>	<i>p</i>	<i>b</i>	SE	<i>t</i>	<i>p</i>	<i>b</i>	SE	<i>t</i>	<i>p</i>	<i>b</i>	SE	<i>t</i>	<i>p</i>
Intercept	−0.347	0.151	−2.288	0.026	−0.241	0.179	−1.349	0.183	−0.237	0.170	−1.393	0.169	0.320	0.182	1.754	0.085
QTYPE=A	0.048	0.051	0.948	0.343	0.357	0.038	9.323	<0.001	0.272	0.046	5.874	<0.001	−0.205	0.040	−5.175	<0.001
QTYPE=W	0.248	0.051	4.903	<0.001	0.433	0.039	11.240	<0.001	0.323	0.047	6.927	<0.001	−0.207	0.040	−5.190	<0.001
<i>z</i> log(QTIME)	−0.031	0.024	−1.296	0.195	−0.019	0.018	−1.036	0.301	0.002	0.022	0.088	0.930	0.012	0.019	0.600	0.549
QGRADE=1	−0.051	0.037	−1.374	0.170	0.057	0.028	2.056	0.040	0.048	0.034	1.416	0.157	−0.054	0.029	−1.858	0.063
SEX=F	0.538	0.208	2.580	0.013	−0.102	0.251	−0.407	0.686	0.021	0.237	0.089	0.929	−0.334	0.256	−1.305	0.198
<i>z</i> SAI	−0.033	0.104	−0.317	0.753	−0.087	0.125	−0.695	0.490	−0.038	0.118	−0.325	0.747	0.112	0.128	0.872	0.387
Observations	1794				1738				1738				1666			
R^2_m	0.083				0.058				0.023				0.040			
R^2_c	0.551				0.756				0.636				0.747			
$\hat{\sigma}_u$	0.682				0.830				0.781				0.845			
$\hat{\sigma}_v$	0.000				0.012				0.000				0.000			
$\hat{\sigma}_\varepsilon$	0.668				0.491				0.602				0.506			

Fixed effects are question format (contrasts QTYPE = *A* and QTYPE = *W* relative to Video), log-standardized time-on-question *z* log(QTIME), accuracy indicator QGRADE=1, sex indicator SEX=*F*, and standardized state anxiety *z*SAI. Random intercepts are included for participant ($u_{PID(i)}$) and question/item ($v_{QID(j)}$). Reported are coefficient *b*, standard error (SE), Satterthwaite *t*, and *p*, along with Nakagawa–Schielzeth R^2 . Variance components are standard deviations: $\hat{\sigma}_u$, $\hat{\sigma}_v$, and $\hat{\sigma}_\varepsilon$. NHRV denotes the normalized IBI/RR-derived cardiac timing channel, not a conventional window-based HRV estimate.

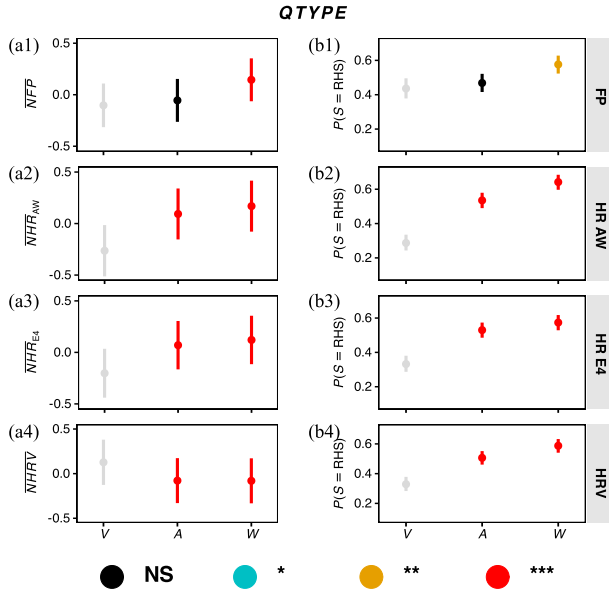


Fig. 8. Question-format effects from mixed-effects models. (a1)–(a4) Continuous-signal models. Model-estimated standardized means ($\pm 95\%$ CI) of the baseline-normalized physiological signals by question type (Video *V*, Abstract *A*, Word *W*) for NFP, NHR_{AW}, NHR_{E4}, and NHRV. Here, NHRV denotes the normalized IBI/RR-derived cardiac timing channel, not a conventional window-based HRV estimate. (b1)–(b4) Relative stress-label models. Model-predicted probability ($\pm 95\%$ CI) of RHS by question type for the same channels. Point colors encode the significance of each format contrast relative to Video (NS, *, **, *** for $p \geq 0.05$, $p < 0.05$, $p < 0.01$, $p < 0.001$). For NHRV, lower values are treated as relatively higher arousal.

capturing performance, sex, and baseline anxiety, respectively. Finally, $u_{PID(i)}$ and $v_{QID(j)}$ capture remaining participant-level and question-level shifts in the baseline-normalized signal, so format effects are estimated from within-participant contrasts after accounting for both.

c) Reported models: In the current analysis, we report the *full* continuous-signal specification in (10) for each physiological channel, without AIC-based reduction. This choice

keeps the inferential target consistent across channels and avoids channel-by-channel changes in covariate adjustment. Accordingly, Table II reports the full set of fixed effects—format contrasts (QTYPE), time-on-question *z* log(QTIME), and controls (QGRADE, SEX, *z*SAI)—together with uncertainty and the estimated variance components under (10).

d) Fixed effects (question format, time, and controls): Across the cardiac channels, nonvideo formats were consistently associated with *higher arousal* relative to Video. This pattern is evident in Table II and Fig. 8(a2)–8(a4). In NHR_{AW}, both Abstract and Word items were higher than Video (*A*: $b = 0.357$, $p < 0.001$; *W*: $b = 0.433$, $p < 0.001$). In NHR_{E4}, the same pattern held (*A*: $b = 0.272$, $p < 0.001$; *W*: $b = 0.323$, $p < 0.001$). In NHRV (where higher values indicate *lower* arousal), both Abstract and Word were *lower* than Video (*A*: $b = -0.205$, $p < 0.001$; *W*: $b = -0.207$, $p < 0.001$), again indicating greater arousal for conventional formats. For facial perspiration (NFP), Word items were higher than Video ($b = 0.248$, $p < 0.001$), whereas Abstract was not significantly different from Video ($b = 0.048$, $p = 0.343$).

Time-on-question did not show evidence of an association with any continuous physiological outcome ($p \geq 0.195$ across channels), supporting the interpretation that the format effects are not an artifact of longer exposure. Among the controls, question-level accuracy was small and statistically detectable only for NHR_{AW} ($b = 0.057$, $p = 0.040$); sex was detectable only for NFP ($b = 0.538$, $p = 0.013$); and state anxiety (*z*SAI) was not significantly associated with any continuous channel ($p \geq 0.387$). Overall, the dominant and consistent fixed-effect pattern in the continuous models is the elevation of arousal under conventional formats relative to Video (with the expected sign reversal for NHRV).

e) Variance decomposition (participant versus item): Variance components in Table II are reported as standard deviations corresponding to (10): $\hat{\sigma}_u$ for the participant random intercept $u_{PID(i)}$, $\hat{\sigma}_v$ for the question random intercept $v_{QID(j)}$,

TABLE III
RELATIVE STRESS-LABEL MIXED-EFFECTS LOGISTIC REGRESSION MODELS (PARTICIPANT-RELATIVE OUTCOMES)

Predictor	NFP				NHR _{AW}				NHR _{E4}				NHRV			
	<i>b</i>	SE	<i>z</i>	<i>p</i>	<i>b</i>	SE	<i>z</i>	<i>p</i>	<i>b</i>	SE	<i>z</i>	<i>p</i>	<i>b</i>	SE	<i>z</i>	<i>p</i>
Intercept	-0.251	0.149	-1.679	0.093	-0.882	0.146	-6.038	<0.001	-0.785	0.140	-5.609	<0.001	-0.801	0.147	-5.467	<0.001
QTYPE = A	0.130	0.174	0.750	0.453	1.050	0.159	6.589	<0.001	0.818	0.151	5.404	<0.001	0.738	0.155	4.767	<0.001
QTYPE = W	0.565	0.174	3.255	0.001	1.493	0.162	9.200	<0.001	0.996	0.152	6.539	<0.001	1.066	0.157	6.806	<0.001
<i>z</i> log(QTIME)	-0.118	0.068	-1.732	0.083	-0.046	0.071	-0.654	0.513	-0.051	0.066	-0.769	0.442	-0.107	0.070	-1.538	0.124
QGRADE=1	-0.057	0.109	-0.526	0.599	0.096	0.110	0.873	0.383	0.177	0.107	1.651	0.099	0.233	0.112	2.080	0.038
SEX=F	0.045	0.102	0.438	0.662	-0.150	0.108	-1.389	0.165	-0.003	0.105	-0.024	0.981	-0.045	0.109	-0.416	0.677
zSAI	0.076	0.052	1.455	0.146	0.042	0.055	0.762	0.446	-0.023	0.054	-0.431	0.666	-0.010	0.055	-0.182	0.855
Observations	1794				1738				1738				1666			

Outcomes are participant-relative dichotomized stress labels $S_c(i, j) \in \{\text{RHS}, \text{RLS}\}$ derived per physiological channel c . Fixed effects are question format (contrasts $QTYPE = A$ and $QTYPE = W$ relative to $QTYPE = V$ [Video]), log-standardized time-on-question $z \log(\text{QTIME})$, and controls ($QGRADE=1$, $SEX=F$, $zSAI$). Models include cross-classified random intercepts for participant and question/item. Reported are coefficient b , standard error (SE), Wald z , and p .

and $\hat{\sigma}_\varepsilon$ for the residual ε_{ij} . Across all signals, $\hat{\sigma}_u$ was substantial (0.682–0.845), indicating stable between-participant differences in baseline-normalized physiological response. In contrast, $\hat{\sigma}_v$ was zero (singular) for three channels and small when nonzero (for NHR_{AW} , $\hat{\sigma}_v = 0.012$). This suggests limited residual item-to-item heterogeneity in the continuous outcomes after accounting for format (and covariates), consistent with the study design goal of constructing closely matched items across formats.

f) Model explainability (R^2): For completeness, Table II reports Nakagawa–Schielzeth R^2 (marginal R_m^2 , fixed effects only; conditional R_c^2 , fixed+random effects). Marginal R_m^2 values are modest (0.023–0.083), whereas conditional R_c^2 values are much larger (0.551–0.756). This gap indicates that most variance in baseline-normalized physiology is attributable to stable participant baselines, with a smaller contribution from item-to-item differences. These values can be misread to imply that format effects are “statistically significant but practically small”; however, that conclusion uses the wrong yardstick. Our inferential target is a reliable *within-participant* shift in arousal by format after accounting for participant and item effects, not between-person prediction of physiological levels. This within-participant target—and its practically meaningful magnitude in higher stress risk units—is most directly expressed in the participant-relative stress-label models reported next (see Section IV-C2; Fig. 8(b1)–8(b4) and Table III).

2) Relative Stress-Label Mixed-Effects Models (RQ1): We report predictors of *relative stress* at the question level using dichotomized stress labels derived from four physiological channels, $c \in \{\text{NFP}, \text{NHR}_{AW}, \text{NHR}_{E4}, \text{NHRV}\}$. For each participant i and question j , we assign RHS when the within-question mean exceeds that participant’s test-wide mean in the same channel. For $NHRV$, however, which decreases with arousal, RHS is assigned when the within-question mean is *below* the participant’s test-wide mean. The complement is labeled RLS. This within-participant construction is intended to express a participant’s *internal state* more directly than continuous physiological markers.

For each stress-label channel $S_c(i, j) \in \{\text{RHS}, \text{RLS}\}$, we fit a mixed-effects logistic regression with cross-classified random intercepts for participant and question. The stress-label model

specification is

$$\begin{aligned}
 \text{logit}(P(S_c(i, j) = \text{RHS})) = & b_0 \\
 & + b_A \mathbb{I}[\text{QTYPE}(i, j) = A] + b_W \mathbb{I}[\text{QTYPE}(i, j) = W] \\
 & + b_T z \log(\text{QTIME}(i, j)) \\
 & + b_G \mathbb{I}[\text{QGRADE}(i, j) = 1] + b_S \mathbb{I}[\text{SEX}(i) = F] \\
 & + b_P zSAI(i) \\
 & + u_{\text{PID}(i)} + v_{\text{QID}(j)}
 \end{aligned} \tag{11}$$

where $u_{\text{PID}(i)} \sim \mathcal{N}(0, \sigma_u^2)$ and $v_{\text{QID}(j)} \sim \mathcal{N}(0, \sigma_v^2)$ are random intercepts for participant and item (question), respectively. As before, $QTYPE = V$ is the reference level. Fig. 8(b1)–8(b4) shows the corresponding predicted probabilities of RHS by format.

a) Key predictors, potential confounders, and controls: The format indicators $\mathbb{I}[\text{QTYPE}(i, j) = A]$ and $\mathbb{I}[\text{QTYPE}(i, j) = W]$ are the focal predictors, quantifying within-participant contrasts relative to Video (positive b indicates higher odds of RHS). We include $z \log(\text{QTIME}(i, j))$ as a potential confounder to address duration-based explanations of stress differences across formats. The remaining covariates—question-level accuracy $\mathbb{I}[\text{QGRADE}(i, j) = 1]$, sex $\mathbb{I}[\text{SEX}(i) = F]$, and pretest anxiety $zSAI(i)$ —serve as controls for performance, sex, and baseline anxiety.

b) Reported models: In the current analysis, we report the *full* stress-label specification in (11) for each channel, without AIC-based reduction. This keeps the inferential target consistent across channels and avoids channel-by-channel changes in covariate adjustment. Accordingly, Table III reports the full set of fixed effects (format contrasts, $z \log(\text{QTIME})$, and controls) with uncertainty and the number of observations for each channel.

c) Fixed effects (question format, time, and controls): Across the cardiovascular stress labels (NHR_{AW} , NHR_{E4} , $NHRV$), $QTYPE$ remains the dominant predictor: both Abstract and Word questions show significantly higher odds of RHS relative to Video (see Table III). The same monotone ordering (Video lowest, Word highest) is visible in Fig. 8(b2)–8(b4). In NFP, Word also shows higher odds of RHS than Video

($b = 0.565$, $p = 0.001$), whereas Abstract is not significantly different from Video ($p = 0.453$), consistent with NFP acting as a corroborating but comparatively noisier channel.

Time-on-question does not account for the format effects: $z \log(\text{QTIME})$ is not statistically significant in any stress-label model ($p \geq 0.083$; Table III). Among the controls, question-level accuracy is significant only for NHRV ($b = 0.233$, $p = 0.038$) and is not significant for the other channels ($p \geq 0.099$); sex and pretest anxiety are not significant in any channel ($p \geq 0.146$). Overall, the stress-label models reinforce the continuous-signal conclusion that conventional formats are associated with higher arousal than Video, and this difference is not explained by longer exposure time.

d) Format effects are large in stress-risk terms: To quantify magnitude, we convert log-odds coefficients to odds ratios ($\text{OR} = \exp(b)$). For example, in the NHR_{AW} stress-label model the $W-V$ contrast is $b = 1.493$ (see Table III), corresponding to $\text{OR}_{W:V} = \exp(1.493) = 4.45$, i.e., W questions have about $4.5\times$ higher odds of RHS than V questions. Equivalently, the reciprocal OR expresses the reduction under V : $\text{OR}_{V:W} = 1/4.45 \approx 0.23$ (about a 77% reduction in RHS odds). Under the approximately balanced mean-threshold labels ($p_0 \approx 0.50$), $\text{OR}_{W:V} = 4.45$ maps illustratively from $p(\text{RHS} | V) \approx 0.50$ to $p(\text{RHS} | W) \approx 0.82$. Across cardiac channels, A/W versus V odds ratios range from $\approx 2\times$ to $\approx 4.5\times$. This implies that V reduces RHS odds by roughly 50%–78% relative to A/W .

e) Participant-relative interpretation: Because RHS/RLS is defined within participant, stable between-participant baselines are removed by construction. Thus, Fig. 8(b1)–8(b4) should be read as format-driven shifts in *within-participant* relative stress probability, rather than between-person differences in stress propensity.

3) Robustness to the Stress-Label Threshold: As a robustness check on the dichotomization rule, we repeated the crossed-intercept stress-label mixed-effects logistic model [see (11)] using participant-specific tercile thresholds (“tercile labels”). For channels where higher values indicate higher arousal (NFP, NHR_{AW} , NHR_{E4}), we assign RHS only when the within-question value falls in the participant’s top third of their within-test distribution (above the 67th percentile). For NHRV, where lower values indicate higher arousal, we instead assign RHS only when the within-question value falls in the participant’s bottom third (below the 33rd percentile). As elsewhere in this article, we report the *full-model* specification including QTYPE, $z \log(\text{QTIME})$, QGRADE, SEX, and $z\text{SAI}$, while keeping the crossed random-intercept structure.

The results were largely unchanged. Question format remained the dominant predictor: across the cardiovascular labels (NHR_{AW} , NHR_{E4} , NHRV), both Abstract and Word items continued to exhibit higher odds of RHS relative to Video, with effect sizes comparable to the mean-rule labels (e.g., $b \approx 0.64$ – 1.08 ; Supplementary Table S4). Time-on-question was estimated in all models; it reached statistical significance only for NHR_{AW} ($b_T = -0.146$, $p = 0.035$) and was nonsignificant for NFP, NHR_{E4} , and NHRV (all $p > 0.05$). Question-level accuracy (QGRADE = 1), sex, and $z\text{SAI}$ were also estimated in all models and showed no strong evidence of association.

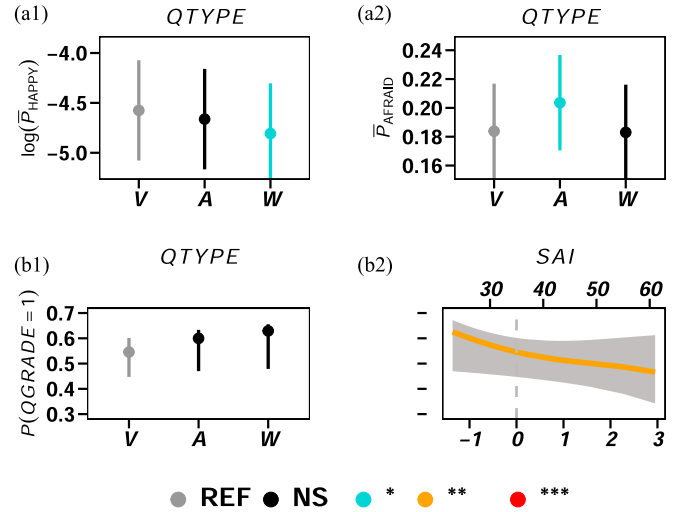


Fig. 9. Model-implied effects for valence and performance. (a1) $Y_{\text{HAPPY}} = \log(\bar{P}_{\text{HAPPY}})$ by question format. (a2) $Y_{\text{AFRAID}} = \bar{P}_{\text{AFRAID}}$ by question format. (b1) $P(\text{QGRADE} = 1)$ by question format from the full performance model (see Table IV). (b2) Effect of pretest state anxiety (SAI) on $P(\text{QGRADE} = 1)$ from the same model. In (a1), (a2), and (b1), points are estimated marginal means with 95% CIs; the gray point denotes the reference category (Video), and point styling indicates the significance of each format contrast relative to Video (NS, *, **, ***). In (b2), the solid curve and shaded band show the fitted mean and 95% CI; the bottom x-axis reports standardized anxiety ($z\text{SAI}$, used in the model), and the top x-axis shows the corresponding raw SAI scale for interpretability. The dashed vertical line marks the sample mean of SAI, and the curve annotation indicates a significant anxiety effect (**, $p < 0.01$).

Overall, the dichotomous format effects are robust to the choice of thresholding rule. Fig. 9 summarizes the model-implied valence and performance effects reported in the next two subsections.

4) Valence Mixed-Effects Models (RQ1): We examine whether the three question formats (V , A , W) are associated with systematic differences in facial-expression valence while participants solved the math test. From the CNN’s frame-level emotion probabilities, we focus on two basic emotions that offer a clear valence contrast: HAPPY (positive valence) and AFRAID (negative valence).

a) Outcomes and preprocessing: For each participant i and question j , we compute the within-question mean $\bar{P}_{\text{VAL}}(i, j)$, with $\text{VAL} \in \{\text{HAPPY}, \text{AFRAID}\}$, by averaging frame-level CNN probabilities over question j . Because $\bar{P}_{\text{HAPPY}}(i, j)$ is right-skewed, we analyze $\log(\bar{P}_{\text{HAPPY}}(i, j))$; $\bar{P}_{\text{AFRAID}}(i, j)$ is analyzed untransformed.

b) Model specification: We fit two cross-classified mixed-effects models (one per outcome) with V as the reference format. Because facial-affect display serves as a complementary affective outcome and is strongly participant-dependent, we use a parsimonious format-only fixed effect with random intercepts for participant and question. The model is

$$\begin{aligned}
 Y_{\text{VAL}}(i, j) = & b_0 \\
 & + b_A \mathbb{I}[\text{QTYPE}(i, j) = A] + b_W \mathbb{I}[\text{QTYPE}(i, j) = W] \\
 & + u_{\text{PID}(i)} + v_{\text{QID}(j)} + \varepsilon_{ij}
 \end{aligned} \tag{12}$$

where $Y_{\text{HAPPY}}(i, j) \equiv \log(\bar{\mathcal{P}}_{\text{HAPPY}}(i, j))$ and $Y_{\text{AFRAID}}(i, j) \equiv \bar{\mathcal{P}}_{\text{AFRAID}}(i, j)$. The participant and question random intercepts satisfy $u_{\text{PID}(i)} \sim \mathcal{N}(0, \sigma_{\text{PID}}^2)$ and $v_{\text{QID}(j)} \sim \mathcal{N}(0, \sigma_{\text{QID}}^2)$, and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is an observation-level residual. Fig. 9(a1) and 9(a2) visualizes the corresponding model-implied format effects for HAPPY and AFRAID. We interpret these valence outcomes as task-sensitive facial-display probabilities rather than direct measures of subjective affect: in particular, higher AFRAID probabilities may reflect vigilant, tense, uncertain, or effortful problem-solving facial configurations, not necessarily consciously experienced fear or math anxiety.

c) *Question format effects*: Supplementary Table S5 shows that $\log(\bar{\mathcal{P}}_{\text{HAPPY}})$ is significantly lower for W questions relative to V questions ($p=0.015$), while the $A-V$ contrast is not significant ($p=0.338$). For $\bar{\mathcal{P}}_{\text{AFRAID}}$, the probability is significantly higher for A questions relative to V questions ($p=0.015$), whereas the $W-V$ contrast is not significant ($p=0.915$). In combination with the stress-label results (see Section IV-C2), these findings indicate that video questions are associated with a more positive displayed-affect profile than conventional text-based formats, albeit in outcome-specific contrasts (higher HAPPY relative to W ; lower AFRAID relative to A).

d) *Variance explained*: Nakagawa–Schielzeth R^2 values show that question format contributes a statistically detectable but modest signal in both valence outcomes: marginal R_m^2 is small, whereas conditional R_c^2 is substantially larger (see Supplementary Table S5). The estimated variance components indicate that this gap is driven primarily by stable between-participant differences, with comparatively minor question-to-question heterogeneity so that format effects operate on top of strong participant-level baselines.

5) *Performance Model (RQ2)*: We report predictors of question-level performance during the math test. The response is item-level accuracy, $\text{QGRADE}(i, j) \in \{0, 1\}$, modeled on the log-odds scale for participant i on question j , i.e., $\text{logit}(P(\text{QGRADE}(i, j) = 1))$. Question format (QTYPE) is the focal predictor set, encoded by $\mathbb{I}[\text{QTYPE}(i, j) = A]$ and $\mathbb{I}[\text{QTYPE}(i, j) = W]$, with V (Video) as the reference category. As elsewhere in this article, we report the *full-model* specification, including time-on-question $z \log(\text{QTIME}(i, j))$, within-question physiological arousal (baseline-normalized facial perspiration, $z\overline{\text{NFP}}(i, j)$), sex (SEX), and pretest state anxiety ($z\text{SAI}$) as fixed effects.

To account for stable between-participant differences and item difficulty, we use cross-classified random intercepts for participant and question. The performance model is

$$\begin{aligned} \text{logit}(P(\text{QGRADE}(i, j) = 1)) = & b_0 \\ & + b_A \mathbb{I}[\text{QTYPE}(i, j) = A] + b_W \mathbb{I}[\text{QTYPE}(i, j) = W] \\ & + b_T z \log(\text{QTIME}(i, j)) + b_N z\overline{\text{NFP}}(i, j) \\ & + b_S \mathbb{I}[\text{SEX}(i) = F] + b_P z\text{SAI}(i) \\ & + u_{\text{PID}(i)} + v_{\text{QID}(j)}. \end{aligned} \quad (13)$$

TABLE IV
PERFORMANCE MIXED-EFFECTS LOGISTIC REGRESSION FOR QUESTION-LEVEL ACCURACY (QGRADE = 1; REFERENCE: V [VIDEO])

Predictor	b	SE	z	p
Intercept	0.644	0.518	1.243	0.214
QTYPE = A	0.426	0.652	0.654	0.513
QTYPE = W	0.630	0.654	0.963	0.336
$z \log(\text{QTIME})$	-0.151	0.109	-1.388	0.165
$z\overline{\text{NFP}}$	-0.114	0.091	-1.255	0.210
SEX = F	0.095	0.348	0.273	0.785
$z\text{SAI}$	-0.441	0.172	-2.561	0.010
Observations	1794			

Fixed effects are reported in log-odds units with Wald z statistics; question format is represented by the contrasts QTYPE=A and QTYPE=W relative to V . Continuous covariates were standardized prior to estimation: SAI as $z\text{SAI}$, facial perspiration $\overline{\text{NFP}}$ as $z\overline{\text{NFP}}$, and time-on-question QTIME as $z \log(\text{QTIME})$. The table reports fixed effects and observations.

Fig. 9(b1) and 9(b2) visualizes the corresponding model-implied accuracy by format and the fitted effect of $z\text{SAI}$ on $P(\text{QGRADE} = 1)$.

Fixed effects (format, time, physiology, and controls): Table IV shows no evidence that accuracy differs by question format after adjustment: neither Abstract nor Word differs significantly from Video (QTYPE=A: $p=0.513$; QTYPE=W: $p=0.336$). Time-on-question and facial perspiration are also not statistically associated with accuracy in the full model ($z \log(\text{QTIME})$: $p=0.165$; $z\overline{\text{NFP}}$: $p=0.210$), and the sex effect is not significant ($p=0.785$). In contrast, pretest state anxiety shows a significant negative association with accuracy ($z\text{SAI}$: $b = -0.441$, $p=0.010$), indicating lower odds of a correct response among participants with higher pretest anxiety.

D. Participant-Specific Video-Story Calmness Effects

Let $i \in \{1, \dots, N\}$ index participants and $j \in \{1, \dots, J\}$ index questions. For each physiological channel $c \in \{\text{NFP}, \text{NHR}_{\text{AW}}, \text{NHR}_{\text{E4}}, \text{NHRV}\}$, let $\overline{\text{NPS}}_c(i, j)$ denote the participant-baseline-normalized question-level mean response for channel c . We model the channel-standardized response $z\overline{\text{NPS}}_c(i, j)$ as defined in (9); thus, all coefficients and derived contrasts are expressed in within-channel standard deviation units.

Our goal is to estimate, for each participant i , their *typical* physiological difference between the video-story condition V and the text-based formats A and W , after controlling for time-on-question and question/item effects. A key motivation is that the calming benefit of the video-story format is not expected to be uniform: some participants may show a large reduction in arousal under V , whereas others may show little change (or even the opposite). To represent this heterogeneity, we allow the format effect to vary by participant—a mixed-model analogue of a participant $\times$ format interaction—by including participant-specific random slopes for the $A-V$ and $W-V$ contrasts. In addition, because some questions are intrinsically more arousing than others regardless of format, we include a question random intercept to absorb item-specific arousal. Concretely, for each

channel c , we fit

$$\begin{aligned} z\overline{\text{NPS}}_c(i, j) = & \beta_{0c} \\ & + \beta_{Ac} \mathbb{I}[\text{QTYPE}(i, j) = A] + \beta_{Wc} \mathbb{I}[\text{QTYPE}(i, j) = W] \\ & + \beta_{Tc} z \log(\text{QTIME}(i, j)) \\ & + u_{Ac,i} \mathbb{I}[\text{QTYPE}(i, j) = A] + u_{Wc,i} \mathbb{I}[\text{QTYPE}(i, j) = W] \\ & + u_{0c,i} + v_{c,j} + \varepsilon_{c,ij} \end{aligned} \quad (14)$$

where $\text{QTYPE} \in \{V, A, W\}$ is question format with V (video-story) as the reference level, and $z \log(\text{QTIME})$ is a log-standardized time-on-question covariate. The fixed effects β_{Ac} and β_{Wc} are the population-average contrasts $A-V$ and $W-V$. The random slopes $u_{Ac,i}$ and $u_{Wc,i}$ are participant-specific deviations from these contrasts, so participant i 's implied contrasts are $(\beta_{Ac} + u_{Ac,i})$ and $(\beta_{Wc} + u_{Wc,i})$. The term $u_{0c,i}$ is a participant random intercept capturing stable differences in overall arousal; $v_{c,j}$ is a question/item random intercept capturing item-specific arousal; and $\varepsilon_{c,ij}$ is residual noise.

1) *Counterfactual Fitted Means*: Using the fitted model, we form participant-specific *counterfactual* mean responses under each format by evaluating (14) at a common exposure time $z \log(\text{QTIME}) = 0$. To isolate format effects, we hold the item effect fixed (so it cancels in contrasts). Denote these fitted means by $\hat{\mu}_{c,i}^V$, $\hat{\mu}_{c,i}^A$, and $\hat{\mu}_{c,i}^W$. Under (14), these satisfy

$$\begin{aligned} \hat{\mu}_{c,i}^V &= \beta_{0c} + u_{0c,i} \\ \hat{\mu}_{c,i}^A &= \beta_{0c} + \beta_{Ac} + u_{0c,i} + u_{Ac,i} \\ \hat{\mu}_{c,i}^W &= \beta_{0c} + \beta_{Wc} + u_{0c,i} + u_{Wc,i} \end{aligned} \quad (15)$$

where we have suppressed the common fixed item term and set $z \log(\text{QTIME}) = 0$ for clarity. Importantly, the participant intercept $u_{0c,i}$ cancels when comparing formats, so individual differences in video-story benefit are captured by the participant-specific format effects $u_{Ac,i}$ and $u_{Wc,i}$.

2) *Video-Story-Text Contrasts and Calmness Direction*: We defined the participant-specific *video-story versus text difference* as the contrast between the fitted mean response under the video-story condition and the average of the two text-based formats

$$\Delta_{c,i} \equiv \hat{\mu}_{c,i}^V - \frac{1}{2} (\hat{\mu}_{c,i}^A + \hat{\mu}_{c,i}^W) . \quad (16)$$

Because lower values of NFP, NHR_{AW}, and NHR_{E4} correspond to reduced physiological arousal, whereas higher values of NHRV indicate longer normalized IBI/RR-derived cardiac timing intervals, we aligned directions so that *larger scores always mean more calmness under the video-story intervention*. Specifically, we defined the per-channel video-story calmness score $C_c(i)$ as

$$C_c(i) \equiv \begin{cases} -\Delta_{c,i}, & c \in \{\text{NFP}, \text{NHR}_{\text{AW}}, \text{NHR}_{\text{E4}}\}, \\ +\Delta_{c,i}, & c = \text{NHRV}. \end{cases} \quad (17)$$

Thus, $C_c(i) > 0$ indicates that participant i is estimated to exhibit, on average, lower arousal or longer normalized IBI/RR-derived timing under the video-story condition than under text-based formats, after accounting for time-on-question and cross-classified participant and question effects.

3) *Overall Video-Story Calmness Score and Ranking*: Finally, we aggregated across channels to obtain an *overall* video-story calmness score per participant

$$C_{\text{overall}}(i) \equiv \frac{1}{|C_i|} \sum_{c \in C_i} C_c(i) \quad (18)$$

where $C_i \subseteq \{\text{NFP}, \text{NHR}_{\text{AW}}, \text{NHR}_{\text{E4}}, \text{NHRV}\}$ is the set of channels available for participant i . Participants were then ranked within each channel by $C_c(i)$ and overall by $C_{\text{overall}}(i)$, enabling identification of individuals for whom the video-story intervention is associated with the greatest reduction in physiological arousal.

4) *Descriptive Summary and Interpretation*: In our cohort ($N = 50$), the overall calmness score $C_{\text{overall}}(i)$ was typically positive, indicating that most participants exhibited lower arousal or longer normalized IBI/RR-derived timing under the video-story format than under text-based formats after accounting for time-on-question and item effects. Specifically, C_{overall} had mean 0.250 (SD 0.313) and median 0.201, with an interquartile range from 0.074 to 0.411. Overall, 41/50 participants (82%) had $C_{\text{overall}}(i) > 0$, whereas 9/50 (18%) had $C_{\text{overall}}(i) < 0$, suggesting a minority for whom the video-story format was not calmer on average.

The magnitude of the heterogeneity is substantial: C_{overall} ranged from -0.501 to 1.213 (within-channel SD units, averaged across available channels). The mean exceeding the median (mean 0.250 versus median 0.201) suggests a right-skewed pattern, with a small number of individuals exhibiting very large positive calmness scores relative to the typical participant. For example, the top-ranked participants show pronounced effects (S045: 1.213; S026: 0.802; S042: 0.773), whereas most participants have more modest positive scores. The full participant ranking is provided in Supplementary Table S6.

V. DISCUSSION

Foundational mathematics assessments are frequently experienced as stressful, and this affective burden can undermine longer term engagement with quantitative coursework. Here, we asked whether stress during math testing is lower when the question experience is changed: presenting skill-matched items as short, relatable video stories (V) rather than as conventional Abstract (A) or Word (W) questions. Across $N = 50$ lower division undergraduates completing a 36-item test, our multi-modal evidence converges on a clear conclusion: the video-story format (V) is associated with reduced physiological arousal (acute stress) and a more favorable expressed-valence signature. These affective advantages persist after accounting for time-on-question. Critically, we observe no measurable degradation in overall test performance at the aggregate level once participant- and item-level heterogeneity are taken into account.

A. Video-Story is Associated With Lower Stress

In this work, we treat elevated autonomic arousal during test-taking—indexed by baseline-normalized heart rate, IBI/RR-derived cardiac timing, and facial perspiration—as an operational marker of *stress* in the arousal–valence framework. Across the cardiac channels, the format effect is strong and consistent: in the stress-label models, conventional formats (*A/W*) typically increase the odds of a participant being in the *RHS* state by about $2\text{--}4.5\times$ relative to video-story (e.g., $\text{OR} = 4.45$ for Word versus Video in NHR_{AW} ; Table III). The continuous-signal models agree: *A/W* raise heart rate and reduce the IBI/RR-derived timing channel relative to *V*. Facial perspiration provides convergent but weaker support: *W* is reliably higher than *V*, whereas the *A–V* contrast is not consistently distinguishable in this sudomotor channel. Taken together, these results indicate that the video-story format is associated with a calmer physiological signature during test-taking—most robustly evident in cardiac dynamics and corroborated by facial perspiration.

B. Not a Simple Time Artifact: Temporal Structure Matters

A natural concern is that videos are slower by design—participants must watch the clip—and that longer exposure could mechanically shift question-level physiological averages. We, therefore, adjusted for time-on-question directly using $z \log(\text{QTIME})$. In the continuous-signal models, $z \log(\text{QTIME})$ is small and nonsignificant across channels (see Table II). In the stress-label models, the time covariate is nonsignificant in the mean-threshold specifications (see Table III) and remains nonsignificant under the stricter tercile-label robustness check, with the exception of a single channel (NHR_{AW} ; Supplementary Table S4). Meanwhile, the *A* and *W* contrasts relative to *V* remain strong across the cardiovascular outcomes. These results argue against a simple duration-based explanation.

However, the relevant distinction is not only how long participants spent on an item, but how cognitive demand was organized within that time. In *A* and *W* items, the mathematical demand is available from the outset: participants can begin identifying quantities, selecting operations, and calculating immediately. In *V* items, by contrast, the item typically begins with a narrative-viewing interval before the mathematical question becomes fully specified. During this initial interval, participants watch the scenario and integrate the story context, which helps frame how the problem is understood; only later, when the relevant quantities and question become clear, does active mathematical calculation become focal.

This temporal structure may itself contribute to the lower average arousal observed for *V* items. We, therefore, do not interpret the *V* effect as evidence that calculation itself was necessarily less demanding once it began. Rather, the narrative-viewing interval may function as an affective and cognitive on-ramp: participants first understand the situation, identify the context, and orient to the problem before active calculation becomes focal. Under this interpretation, the mathematical demand is not removed; instead, participants may enter the calculation phase with better contextual grounding and lower anticipatory or evaluative arousal. Because our data do not identify the exact moment

at which participants shifted from viewing/comprehension to active calculation, this account remains an interpretation rather than a directly estimated mechanism. Even if this mechanism is only partially responsible for the observed arousal differences, it remains educationally meaningful: the video-story format may help students approach the mathematical work gradually, with contextual and affective preparation, rather than confronting the full evaluative demand immediately at item onset.

C. More Positive Affect

The facial-expression analyses complement the arousal results, but they should be interpreted as evidence about *displayed affective tone*, not as direct evidence of felt emotion. Relative to *V*, *W* questions are associated with lower *HAPPY* display, whereas *A* questions are associated with higher *AFRAID* display. As discussed in Section III-B and illustrated in Supplementary Fig. S2, negative-expression labels in academic testing contexts can reflect task-dependent facial-display morphology associated with cognitive effort and evaluative tension. This interpretation is consistent with prior neuroimaging evidence that math anxiety can engage pain-related neural circuitry during anticipation of mathematical activity [6], suggesting that mathematics may evoke a more threat-like anticipatory affective posture than some other academic tasks for some learners. We do not claim that the facial-expression classifier detects pain, fear, or anxiety directly. Rather, we interpret the relatively greater *AFRAID*-like display in the math-test condition than in the essay-writing comparison as a task-sensitive facial-display pattern that may reflect vigilance, tension, uncertainty monitoring, or evaluative concern during mathematical problem solving. In this interpretation, the lower *AFRAID* probabilities observed for video-story items may reflect relative relaxation of vigilant, tense, or uncertainty-monitoring facial configurations, rather than a literal reduction in consciously experienced fear. In the arousal–valence framework, this places *V* in a more favorable region of the observed affective-display space: lower physiological arousal coupled with modest shifts toward more positive and less negative facial display. This interpretation is compatible with the physiological results: the reduced *AFRAID* display under video-story items aligns with lower arousal in the cardiac and facial-perspiration channels.

D. Accuracy: Overall Parity

At the model level, we do not find a statistically significant difference in accuracy by format once cross-classified participant and item effects are modeled: neither *A* nor *W* differs significantly from *V* in the full performance model (see Table IV). This is practically important: the physiological and valence benefits of video-story presentation appear achievable without an overall cost to accuracy. However, the skill-matched triplet descriptives reveal an equally important design lesson: video questions are not automatically performance-neutral. We, therefore, treat this heterogeneity as a substantive design finding, not merely as a limitation: the affective benefit of video-story presentation depends on whether the item design keeps the mathematical information clear, stable, and available for step-by-step use. A

small subset of V items underperformed markedly relative to their matched A/W counterparts (see Fig. 7(b); Supplementary Table S3). Full prompts and video links for these items are provided in Supplementary Table S2, enabling inspection of the specific scenarios. This pattern suggests that the video medium introduces additional design degrees of freedom (pacing, narrative detail, and modality of information presentation) that can either support or hinder problem solving. A plausible mechanism is extraneous cognitive load: if quantitative givens are transient, split across modalities, or embedded in distracting narrative detail, the learner may miss or forget key information [50], [51]. Consistent with this interpretation, an exploratory physiological comparison suggested that the three markedly underperforming Video triplets $k = 9, 11, 12$ tended to elicit higher facial perspiration than the remaining Video items, with a directional paired Wilcoxon trend $p = 0.073$. We interpret this result cautiously because the subset contains only three items and the analysis was exploratory rather than confirmatory; Supplementary Fig. S4 shows the corresponding item-level pattern. These observations motivate concrete item-design safeguards for future V questions, such as persistent on-screen display of numerical parameters, concise scripts, clear separation between narrative flavor and mathematical givens, and optional replay/rewind or a brief on-screen text recap of required values.

E. Heterogeneity in Calmness Benefit

To summarize the video-format benefit at the individual level, we compute for each participant a *calmness score*, defined as that participant's contrast between baseline-normalized physiological responses under V and under the nonvideo formats. Positive values indicate that the participant is, on average, less aroused during video items; values near zero indicate little change; and negative values indicate higher arousal under V . Most participants exhibit a positive calmness score, but the distribution is wide and includes a minority with near-zero or negative values, showing that the calming benefit is not uniform across learners.

F. Evidence That the Test was Psychologically Consequential

Although the test was administered in a volunteer laboratory session rather than as a graded course assessment, the significant negative association between pretest state anxiety ($zSAI$) and accuracy indicates that the session still carried meaningful performance pressure: participants who arrived in a more anxious state performed worse, consistent with established anxiety–performance effects in evaluative settings. This relationship supports the interpretation that the test engaged an anxiety–performance coupling of the kind that motivates concern about stress in foundational mathematics testing [2], [8], strengthening the study's external validity.

G. Implications for Assessment Design in the Short-Video Era

From an educational-technology perspective, the findings support an affective redesign strategy that changes the *presentation* of assessment items rather than adding auxiliary interventions around the test. Short narrative videos are

already a dominant information-consumption medium for today's learners [18], and our results suggest that leveraging this medium—when paired with relatable storytelling—is associated with lower physiological stress during mathematics testing while preserving overall accuracy. If replicated in authentic educational settings, such designs could provide a scalable pathway to making foundational mathematics assessments less intimidating and more engaging.

H. Multimodal Stress Labeling Without Per-Question Self-Report

Methodologically, our key contribution is a practical alternative to burdensome per-question self-report for building stress-aware educational datasets. We show that participant-relative stress labels can be derived from baseline-normalized physiology and supported via *convergence* across autonomic channels from distinct pathways—cardiac dynamics (heart rate and IBI/RR-derived timing) and sudomotor activity (facial perspiration; cholinergic). Cross-channel agreement, together with consistent results from continuous-signal analyses and participant-relative stress labeling, strengthens the construct validity of physiological “grading” of arousal. This multimodal approach enables scalable dataset construction during test administration and supports reproducible comparisons of alternative stress-inference pipelines on the same signals and labels. By releasing data and code openly, we aim to accelerate follow-on work in multimodal educational measurement.

I. Limitations and Ecological Validity

Several limitations bound interpretation. First, the V condition intentionally bundles relatable storytelling, enacted video delivery, and temporal presentation. This design was chosen for ecological validity, but it means that the present study estimates a *total-format* effect rather than isolating a video-only or narrative-only mechanism. The observed effects may reflect the semantic relatability of the story, audiovisual delivery, social cues from enacted characters, temporal pacing, or their combination. While one might propose a factorial decomposition to isolate “video” from “story,” such manipulations are difficult to realize without creating artificial stimuli that alter how learners naturally engage with the item, thereby undermining the very validity such an experiment would seek to preserve. Second, our study was not embedded in a graded course test. Nevertheless, the observed negative association between pretest state anxiety (SAI) and overall test performance indicates that the session still carried meaningful performance pressure, supporting the relevance of the affective responses we measure.

J. Cultural Specificity and Transferability

Relatable stories are context dependent. Our video narratives were designed for a U.S. urban student population and used story domains (e.g., personal finance and health) and casting choices intended to reflect that context. Replications outside this setting should adapt narrative content, language, and representation to the target learner population, since what is “relatable”—and thus, potentially stress-reducing—may differ across cultures and regions.

K. Future Directions: Sensory Channels and Storytelling Beyond Video

The most immediate next step is to refine video-story item construction to eliminate the small number of underperforming *V* items while retaining the robust calming signature observed here. This includes varying the amount and form of mathematical support in multistep video-story items through features such as persistent numerical displays, annotations of key quantities, summary panels, or brief text recaps. More broadly, because the present *V* condition bundles story content with video delivery, future work should hold the *storytelling principle* constant while varying the *delivery medium* across formats that learners naturally consume—for example, text-based narratives, audio vignettes, static visual/comic presentations, and full video. This comparison would help clarify how different sensory channels contribute to physiological arousal during mathematical assessment: text may preserve the semantic and relatable story structure while minimizing audiovisual stimulation; audio may add tone of voice, pacing, emphasis, and human voice cues; static visual formats may provide contextual realism while keeping key quantities available; and full video may add social cues and embodiment, but also increase cognitive load through multimodal integration demands.

The optimal weighting of these sensory channels is also likely to vary across learners. The participant-specific calming scores in this study already show that some students matched better to the video-story format than others; analogous variation may occur for audio, text, static visual, or mixed-media narratives. Future systems could, therefore, estimate learner-specific sensory profiles, identifying which combination of semantic narrative, auditory cues, visual context, and social embodiment produces the strongest calming effect without compromising mathematical clarity.

Future time-resolved designs should further separate narrative exposure, calculation onset, and response execution within each modality, for example, by marking when the mathematical question and key quantitative information become available within each item.

VI. CONCLUSION

In summary, our findings provide convergent multimodal evidence that a video-story question format is associated with lower physiological stress during foundational math testing and is accompanied by improved expressed-valence signatures. These affective benefits persist after accounting for time-on-question and do not produce an overall accuracy penalty. With careful item design, storytelling-based assessment formats—delivered via video and potentially other media—offer a promising pathway toward less stressful quantitative assessment. Taken together, these findings answer the motivating questions of this study.

DATA AND CODE

The data, analysis code, and video-question materials used in this study are openly available in the project's GitHub repository.²

REFERENCES

- [1] Organisation for Economic Co-operation and Development, *PISA 2012 Results: Ready to Learn (Volume III): Students' Engagement, Drive and Self-Beliefs*. Paris, France: OECD Publishing, 2013.
- [2] C. Barroso, C. M. Ganley, A. L. McGraw, E. A. Geer, S. A. Hart, and M. C. Daucourt, "A meta-analysis of the relation between math anxiety and math achievement," *Psychol. Bull.*, vol. 147, no. 2, pp. 134–168, 2021.
- [3] A. Devine, K. Fawcett, D. Szűcs, and A. Dowker, "Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety," *Behav. Brain Functions*, vol. 8, pp. 1–9, 2012.
- [4] R. Conlon, A. Hicks, C. Barroso, and C. M. Ganley, "The effect of the timing of math anxiety measurement on math outcomes," *Learn. Individual Differences*, vol. 86, 2021, Art. no. 101962.
- [5] F. Demedts, B. Reynvoet, D. Sasanguie, and F. Depaepe, "Unraveling the role of math anxiety in students' math performance," *Front. Psychol.*, vol. 13, 2022, Art. no. 979113.
- [6] I. M. Lyons and S. L. Beilock, "When math hurts: Math anxiety predicts pain network activation in anticipation of doing math," *PLoS One*, vol. 7, no. 10, 2012, Art. no. e48076.
- [7] M. H. Ashcraft, "Math anxiety: Personal, educational, and cognitive consequences," *Curr. Directions Psychol. Sci.*, vol. 11, no. 5, pp. 181–185, 2002.
- [8] J. P. Jamieson, A. E. Black, L. E. Pelaia, and H. T. Reis, "The impact of mathematics anxiety on stress appraisals, neuroendocrine responses, and academic performance in a community college sample," *J. Educ. Psychol.*, vol. 113, no. 6, pp. 1164–1176, 2021.
- [9] R. Hembree, "The nature, effects, and relief of mathematics anxiety," *J. Res. Math. Educ.*, vol. 21, no. 1, pp. 33–46, 1990.
- [10] R. G. Pizzie and D. J. M. Kraemer, "The association between emotion regulation, physiological arousal, and performance in math anxiety," *Front. Psychol.*, vol. 12, 2021, Art. no. 639448.
- [11] K. Choksi, H. Chen, K. Joshi, S. Jade, S. Nirjon, and S. Lin, "SensEmo: Enabling affective learning through real-time emotion recognition with smartwatches," in *Proc. IEEE 21st Int. Conf. Mobile Ad-Hoc Smart Syst.*, 2024, pp. 453–459.
- [12] S. Liu, Y. Zhang, L. Zhao, and Z. Liu, "Academic stress detection based on multisource data: A systematic review from 2012 to 2024," *Interactive Learn. Environ.*, vol. 33, no. 3, pp. 1823–1849, 2025.
- [13] Y. Zhang, S. Liu, L. Zhao, and Z. Liu, "How academic stress affects academic performance: A chain-mediated model of sex moderation," *J. Psychol. Afr.*, vol. 34, no. 5, pp. 487–494, 2024.
- [14] H. Zuo and L. Wang, "The influences of mindfulness on high-stakes mathematics test achievement of middle school students," *Front. Psychol.*, vol. 14, 2023, Art. no. 1061027.
- [15] R. Toscano, M. A. Guerra, J. Ortega, and B. E. Moscoso, "Exploring the impact of coffee break in mathematics courses to improve academic performance and reducing exam stress," in *Proc. Int. Conf. Appl. Eng. Innov. Technol.*, 2024, pp. 471–477.
- [16] D. I. Cordova and M. R. Lepper, "Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice," *J. Educ. Psychol.*, vol. 88, no. 4, pp. 715–730, 1996.
- [17] Cognition and Technology Group at Vanderbilt, "Anchored instruction and its relationship to situated cognition," *Educ. Researcher*, vol. 19, no. 6, pp. 2–10, 1990.
- [18] M. Faverio and O. Sidoti, "Teens, social media and technology 2024," *Pew Res. Center, Tech. Rep.*, Dec. 2024. Accessed: Dec. 29, 2025. [Online]. Available: <https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/>
- [19] P. J. Guo, J. Kim, and R. Rubin, "How video production affects student engagement: An empirical study of MOOC videos," in *Proc. 1st ACM Conf. Learn. Scale Conf.* New York, NY, USA: ACM, 2014, pp. 41–50.
- [20] C. J. Brame, "Effective educational videos: Principles and guidelines for maximizing student learning from video content," *CBE—Life Sci. Educ.*, vol. 15, no. 4, 2016, Art. no. es6.
- [21] F. Kiran, A. Wesley, T. Tolar, P. T. Cirino, P. Tsiamyrtzis, and I. Pavlidis, "Relatable and humorous videos reduce hyperarousal in math exams," in *Proc. 11th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos*, 2023, pp. 1–4.
- [22] W. Boucsein, *Electrodermal Activity*, 2nd ed. New York, NY, USA: Springer, 2012.
- [23] G. G. Berntson et al., "Heart rate variability: Origins, methods, and interpretive caveats," *Psychophysiology*, vol. 34, no. 6, pp. 623–648, 1997.
- [24] S. R. Fishel, E. R. Muth, and A. W. Hoover, "Establishing appropriate physiological baseline procedures for real-time physiological measurement," *J. Cogn. Eng. Decis. Mak.*, vol. 1, no. 3, pp. 286–308, 2007.

²<https://github.com/UH-ACDC/affective-math-test-repo>

- [25] C. D. Spielberger, "State-trait anxiety inventory," in *The Corsini Encyclopedia of Psychology*. Hoboken, NJ, USA: Wiley, 2010. [Online]. Available: <https://doi.org/10.1002/9780470479216.corpsy0943>
- [26] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," *Int. J. Hum.-Comput. Interact.*, vol. 24, no. 6, pp. 574–594, 2008.
- [27] J. Brooke, "SUS—A quick and dirty usability scale," in *Usability Eval. in Industry*, P. W. Jordan, B. Thomas, I. L. McClelland, and B. Weerdmeester, Eds. New York, NY, USA: Taylor & Francis, 1996, pp. 189–194.
- [28] Apple Inc., "Human interface guidelines," 2024. Accessed: May 15, 2025. [Online]. Available: <https://developer.apple.com/design/human-interface-guidelines/>
- [29] Y. Zhou, P. Tsiamyrtzis, P. Lindner, I. Timofeyev, and I. Pavlidis, "Spatiotemporal smoothing as a basis for facial tissue tracking in thermal imaging," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 5, pp. 1280–1289, May 2013.
- [30] D. Shastri, M. Papadakis, P. Tsiamyrtzis, B. Bass, and I. Pavlidis, "Perinatal imaging of physiological stress and its affective potential," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 366–378, Mar. 2012.
- [31] H. Sorensen, D. Jones, M. Heideman, and C. Burrus, "Real-valued fast Fourier transform algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 6, pp. 849–863, Jun. 1987.
- [32] M. T. Hasan, S. Zaman, A. Wesley, P. Tsiamyrtzis, and I. Pavlidis, "Sympathetic activation in deadlines of deskbound research—A study in the wild," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, 2023, pp. 1–8.
- [33] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, 2018, Art. no. e0196391.
- [34] J. J. Gross, "The emerging field of emotion regulation: An integrative review," *Rev. Gen. Psychol.*, vol. 2, no. 3, pp. 271–299, 1998.
- [35] R. E. Kraut and R. E. Johnston, "Social and emotional messages of smiling: An ethological approach," *J. Pers. Social Psychol.*, vol. 37, no. 9, pp. 1539–1553, 1979.
- [36] A. Wesley, "A mixed emotions framework and its applications in affective computing," Ph.D. dissertation, Dept. Comput. Sci., Univ. Houston, Houston, TX, USA, May 2024.
- [37] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System: The Manual*. Salt Lake City, UT, USA: Research Nexus, 2002.
- [38] NetTime Project, NetTime, version 3.14. Accessed: May 15, 2025. [Online]. Available: <https://nettime.sourceforge.net>
- [39] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Front. Public Health*, vol. 5, 2017, Art. no. 258.
- [40] R. S. Lee, "The physiology of stress and the human body's response to stress," in *Epigenetics of Stress and Stress Disorders*. Amsterdam, The Netherlands: Elsevier, 2022, pp. 1–18.
- [41] A. Parchment, R. W. Wohleber, and L. Reinerman-Jones, "Psychophysiological baseline methods and usage," in *Foundations Augmented Cognition: Neuroergonomics Operational Neuroscience*. Cham, Switzerland: Springer, 2016, pp. 361–371.
- [42] J. Sacha, "Why should one normalize heart rate variability with respect to average heart rate?" *Front. Physiol.*, vol. 4, 2013, Art. no. 306.
- [43] S. Zaman et al., "Stress and productivity patterns of interrupted, synergistic, and antagonistic office activities," *Sci. Data*, vol. 6, no. 1, 2019, Art. no. 264.
- [44] R. A. Peterson and J. E. Cavanaugh, "Ordered quantile normalization: A semiparametric transformation built for the cross-validation era," *J. Appl. Statist.*, vol. 47, nos. 13–15, pp. 2312–2327, 2020.
- [45] G. Matthews, "Stress states, personality and cognitive functioning: A review of research with the Dundee Stress State Questionnaire," *Pers. Individual Differences*, vol. 169, 2021, Art. no. 110083.
- [46] N. Gao, M. Saiedur Rahaman, W. Shao, and F. D. Salim, "Investigating the reliability of self-report data in the wild: The quest for ground truth," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., ACM Int. Symp. Wearable Comput.*, 2021, pp. 237–242.
- [47] C. Rott, F. Kiran, M. Segers, P. Van den Bossche, and I. Pavlidis, "STREL—Naturalistic dataset and methods for studying stress and relaxation patterns in critical leading roles," *IEEE Trans. Affect. Comput.*, vol. 17, no. 1, pp. 334–347, Jan.–Mar. 2025.
- [48] F. C. Richardson and R. M. Suinn, "The mathematics anxiety rating scale: Psychometric data," *J. Counseling Psychol.*, vol. 19, no. 6, pp. 551–554, 1972.
- [49] OECD, "Does math make you anxious," OECD Publishing, Tech. Rep. 48, 2015.
- [50] P. Chandler and J. Sweller, "Cognitive load theory and the format of instruction," *Cogn. Instruct.*, vol. 8, no. 4, pp. 293–332, 1991.
- [51] R. E. Mayer and R. Moreno, "Nine ways to reduce cognitive load in multimedia learning," *Educ. Psychol.*, vol. 38, no. 1, pp. 43–52, 2003.



a Faculty Lecturer with research interests include human–computer interaction, affective computing, and naturalistic studies.

Fethah Kiran received the B.S. degree in computer and instructional technologies education from Ondokuz Mayıs University, Samsun, Türkiye, in 2013, the M.S. degree in computer science and engineering from Louisiana State University, Baton Rouge, LA, USA, in 2020, and the Ph.D. degree in computer science from the University of Houston, Houston, TX, USA, in 2025.

From 2013 to 2015, he worked as a Content Developer at the Distance Learning Center, Canakkale Onsekiz Mart University, Türkiye. He is currently Yildiz Technical University, Istanbul, Türkiye. His research interests include human–computer interaction, affective computing, and naturalistic studies.



Paul T. Cirino received the Ph.D. degree in clinical neuropsychology from the University of Houston, Houston, TX, USA, in 1996.

He completed his APA/Division 40 clinical internship at the Arizona Health Sciences Center, University of Arizona, Tucson, AZ, USA, in 1996. Subsequently, he was the Kirk Dornbush Postdoctoral Fellow in developmental neuropsychology at Georgia State University, followed by neuropsychology staff positions at the Kennedy Krieger Institute in Baltimore, MD, USA and at Georgia State University. He

returned to the University of Houston as a Faculty in 2002, where he is currently a Full Professor in psychology. He is also the Director of the Developmental Neuropsychology Clinic, which provides evaluations for children with cognitive, learning, and behavioral problems. His grants and published work reflect his interests in academic areas (reading and mathematics) and neurocognitive function (particularly executive function, attention, and processing speed), and his work examines both typical and neurodevelopmental populations.

Dr. Cirino has been continuously funded with grants from National Institutes of Health (NIH), Institute of Education Sciences (IES), and National Science Foundation (NSF) for more than 20 years.



Ioannis Pavlidis (Fellow, IEEE) received the B.E. degree in electrical engineering from the Democritus University of Thrace, Xanthi, Greece, in 1987, the M.S. degree in robotics from the University of London, London, U.K., in 1989, and the M.S. and Ph.D. degrees in computer science from the University of Minnesota, Minneapolis, MN, USA, in 1995 and 1996, respectively.

He is the Eckhard-Pfeiffer Distinguished Professor of computer science and the Director of the Affective and Data Computing Laboratory, University of Houston, Houston, TX, USA. He has led interdisciplinary teams spanning engineering, medicine, and behavioral science, and his group has released open datasets and tools to support reproducible affective computing research. He has published extensively in the areas of affective computing, data science, and science of science. He was the first to develop contact-free methods for measuring physiological variables, including electrodermal activity, breathing, and heart function, which he used to study stress in the wild.

Dr. Pavlidis has received funding by multiple sources, including the National Science Foundation, transportation agencies, and medical institutions, for his research.