Role of Multimodal Learning Systems in Technology-Enhanced Learning (TEL): A Scoping Review

Yoon Lee[1(✉)], Bibeg Limbu[2], Zoltan Rusak[3], and Marcus Specht[1]

[1] Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
y.lee@tudelft.nl
[2] Faculty of Engineering, University of Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany
[3] Faculty of Industrial Design Engineering, Landbergstraat 15, 2628 CE Delft, Netherlands

Abstract. Technology-enhanced learning systems, specifically multimodal learning technologies, use sensors to collect data from multiple modalities to provide personalized learning support beyond traditional learning settings. However, many studies surrounding such multimodal learning systems mostly focus on technical aspects concerning data collection and exploitation and therefore overlook theoretical and instructional design aspects such as feedback design in multimodal settings. This paper explores multimodal learning systems as a critical part of technology-enhanced learning used for capturing and analyzing the learning process to exploit the collected multimodal data to generate feedback in multimodal settings. By investigating various studies, we aim to reveal the roles of multimodality in technology-enhanced learning across various learning domains. Our scoping review outlines the conceptual landscape of multimodal learning systems, identifies potential gaps, and provides new perspectives on adaptive multimodal system design: intertwining learning data for meaningful insights into learning, designing effective feedback, and implementing them in diverse learning domains.

Keywords: Multimodal Learning Analytics (MMLA) · Sensor-based Technology · Learning Domains

1 Introduction

With the increasing application of Technology-Enhanced Learning (TEL), the educational roles of teachers and students are constantly changing [39]. The seismic shift was observed during the pandemic in the last few years, which forced the educational focus from traditional classroom learning to online and hybrid environments [39]. Owing to the proliferation of digital platforms and devices designed for educational purposes [25], TEL technologies have resulted

O. Viberg et al. (Eds.): EC-TEL 2023, LNCS 14200, pp. 164–182, 2023.
https://doi.org/10.1007/978-3-031-42682-7_12

in the availability of copious amounts of data on both the learner and their learning process. As a direct consequence, TEL technologies are being further enhanced with sophisticated Artificial Intelligence (AI), particularly Machine Learning (ML) techniques and Learning Analytics (LA).

Such technological advancements have reinforced the role of TEL, not only as a LA tool but also as a form of feedback agent in learning. For instance, the advent of ChatGPT[1] appears to bring transformative development in the field, as it has the potential to change the foundations of learning and education [27]. Although such interactions are currently limited only to text modality, information acquisition will become even more accessible via multiple sensory modalities, with the convergence of diverse speech-based conversational agents [31] and sensor technologies, in the form of multimodal interactions (e.g., Generative AI combined with VR agents) [27]. In this context, the importance of multimodality is not only confined to TEL as data input from the digital world [13], but also as outputs in both the physical and the virtual world, which can trigger cognitive, behavioral, and emotional changes in learners.

Multimodal learning systems, a subgroup of TEL, frequently employ multiple sensors and AI techniques to gather contextual learning data from diverse modalities to provide a comprehensive understanding of learning processes. This understanding can assist us, as practitioners and researchers, to reflect on the efficacy of the design of multimodal learning systems: how to digitize learning and learner information as data [25], how to process and intertwine multimodal data to best contextualize learning [21,25,44], and how to design and implement feedback and LA, also called multimodal learning analytics (MMLA) in learning systems to address students necessities [25,40].

The field of MMLA combines different types of data from multiple modalities and sources to gain contextual insights into the learning process. Di Mitri et al. [13], in their conceptual framework called "Multimodal Learning Analytics Model (MLeAM)", portrayed multimodality in learning systems as a series of steps involving sensor capturing, annotation, predictions, and feedback implementation, in a loop. Although their conceptual framework has precisely aligned the multimodal data stream in input space, the framework has yet to be extended to the dimensions of feedback design and its implications for learning domains. In order to get insights into the design of feedback and MMLA, a critical component of TEL, we examine through a review how previous studies utilize multimodality in their learning systems from data collection to feedback implementation, which has yet to be collectively understood in previous research. Therefore, we investigate multimodal learning systems in three primary stages: 1) data collection and integration, 2) design decisions for the design of multimodal feedback, and 3) implications for system implementation in diverse learning domains. The following three research questions will be tackled by reviewing and analyzing studies in the field.

– RQ1. How is multimodal data collected and processed to get insights about learning in MMLA?

[1] https://openai.com/.

– RQ2. How is learner feedback designed in the context of multimodal learning systems?
– RQ3. What are the considerations for implementing multimodal learning systems in various learning domains?

2 Methodology

The literature search was conducted with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) approach. The review itself was later adapted to a *scoping review* due to the erratic landscape of the multimodal learning systems we observed based on our preliminary searches, as our focus was on investigating the emerging topic, multimodality, as a critical component of TEL. Therefore, we adopted the five-stage approach of a scoping review of Arksey and O'Malley [3]: 1) identifying the research questions, 2) identifying relevant literature based on inclusion and exclusion criteria, 3) selecting studies, 4) analyzing and synthesizing the data, 5) and summarizing and reporting the results.

Through various search engines, such as Scopus and Web of Science, 1,794 search results were found based on keyword search (i.e. (multimodal OR multisensory) AND feedback AND (learning OR education)). Only results that included a description of learning systems designed for human users were selected, resulting in 274 papers. The results included papers from various subject areas, such as computer science, engineering, social sciences, psychology, and art and humanities. Six researchers further coded and filtered the remaining 274 papers in the eligibility check process with inclusion and exclusion criteria, such as having multimodal components as both the input and output of the system implementation. Using Cohen's Kappa coefficient by comparing observed and random probabilities, the inter-rater reliability among six coders' scores has been evaluated (Cohen's Kappa: good, $0.9 > 0.81 \geq 0.8$). The primary author solely proceeded with the rest of the overall review, and 27 papers were chosen for the final review. To compensate limitation of the PRISMA methodology caused by its strict application of inclusion and exclusion criteria, we applied the snowball method to extend the discussion with other relevant research in the field for a further scoping review, which resulted in 30 papers ranging from 2010 to 2023.

3 Results

The search conducted with the methodology described above resulted in 30 papers published between 2010 and 2023. Most of the systems in the resulting studies were based on a sensor-based approach and built for on-site learning than online learning. Similarly, they were often geared towards individual learning scenarios instead of collaborative learning. The majority of the selected studies' primary intervention was in the form of real-time feedback rather than post-hoc

feedback, and most targeted learners more than teachers. A significant proportion of studies were conducted in K-12 education and higher education. Table 1 provides an overview of all the selected studies and their learning domains, data inputs, and feedback modalities.

4 Discussion

4.1 RQ1. How Is Multimodal Data Collected and Processed to Get Insights About Learning in MMLA?

Multimodal data collection in MMLA is performed using a host of sensors that correspond to the five primary modalities used by humans (i.e., visual, auditory, tactile, taste, and smell [11]), with various information layers, such as data types, frequencies, and resolutions. Our literature search yielded no studies that addressed the modalities of taste and smell, indicating a dearth of technology capable of capturing them. Of the 30 papers, six studies (20.0%) collected visual and auditory data, two studies (6.7%) collected visual and tactile data, two studies (6.7%) collected auditory and tactile data, and two studies (6.7%) collected all three of them. Tactile data has been most frequently used as the major data stream in eleven studies (36.7%), while visual and auditory data have been used in five (16.7%) and two studies (6.7%), respectively.

Sensor-Based Data Collection *Visual and auditory sensors* are frequently used to collect audio and video data. Visual data is collected using different types of cameras (e.g., webcam [30], infrared camera [31], motion capture camera [52]), which consist of various information layers such as RGB [9], shapes, sizes, and textures. Visual data is further processed, often with AI and ML techniques, for various purposes such as image recognition [41], facial expression analysis [10], gaze and posture analysis [15,37], and trajectory tracking of the body [7,24] and objects.

Auditory data is captured through the microphone, having volume and frequency as essential features. The human voice is commonly captured as an auditory modality that is used for corpus analysis [1], speech analysis, voice trait analysis [15,47], and musical trait analysis [34,52].

Tactile sensors, such as Inertial Motion Units (IMUs), are used to capture learners' physical movement and orientation detection [20,47] while force trajectory is tracked via force sensors [7]. Additionally, *environmental sensors* collect information about the physical learning environments, such as temperature, humidity, noise level, and air quality [49]. More sophisticated *physiological sensors* technologies (e.g., eye tracker [31,32], electroencephalogram (EEG)) are also applied for more accurate and deeper insights into the physiological state of the learner. Such sensors collect physiological information such as heart rate and skin conductance, which are further interpreted as clues of learners' stress levels, arousal, and emotional states [12]. For example, one dominant tendency in MMLA is the inclusion of physiological sensors to evaluate learners' affective

Table 1. List of papers included in the scoping review.

Lit.	Learning Domains	Data Input	Feedback Modalities
[1]	**Language Learning (Foreign)**: Personalized foreign language training	**Auditory**: Vocal traits and sentence formulation from microphone	**Visual/Auditory** : Dashboard (analytic dashboards), Text (paraphrasing), Voice (model voice, dialogue simulation)
[4]	**Medical Education** : Gamified palpation training	**Tactile** : Pressure sensitivity and timing from wearable glove (ParsGlove)	**Visual/Auditory** : Graphics (gamified task), Sound Effects (rewarding purpose for successful task completion)
[5]	**Medical Education** : Palpation analytics for training	**Tactile** : Orientation, position, and pressure of hand from wearable glove (ParsGlove)	**Visual** : Dashboard (real-time monitoring panel, feedback panel, post-hoc examogram with tables, graphs, and scores)
[6]	**Conceptual Learning** : Teaching properties of graphs and diagrams for visually impaired students	**Tactile** : Stylus orientation from Phantom Omni	**Auditory/Tactile** : Voice (task guidance), Physical Movements (Phantom Omni)
[7]	**Language Learning (First)** : Teaching handwriting for children	**Tactile** : Characters written on touch screen	**Visual/Tactile** : Graphics (cartoon face, color trajectory for guidance), Text (message alerts), Vibrations (error), Physical Movements (haptic rendering of correct trajectory)
[9]	**Conceptual Learning** : Gamified interface for better knowledge gain in collaborative learning	**Visual / Auditory** : Learner poses from a camera, Group corpus from microphone	**Visual** : Graphics (virtual farmland getting mature with better participation in the class)
[10]	**Clear Communication** : Practicing public speaking	**Visual/Auditory** : Behavior (gaze, facial expressions) from webcam and speech, voice activity from a microphone (Multisense)	**Visual** : Graphics (interactive virtual audiences), Dashboard (after-action report with analysis, containing text and graphs, and video recording)
[15]	**Clear Communication** : Presentation training	**Visual/Auditory** : Speech rate and filled pauses from six microphones, Gaze detection from six cameras (HAAR Cascade), posture skeleton (Kinect)	**Visual** : Graphics (symbolic icons-thumbs up and down with color coding, video capture of presentation), Text (evaluation summary with statistics)

(*continued*)

Table 1. (*continued*)

Lit.	Learning Domains	Data Input	Feedback Modalities
[16]	**Conceptual Learning** : Immersive organic chemistry education with gamified VR	**Tactile** : Finger tracking via haptic glove	**Visual/Tactile** : Graphics (3D-gamified molecule graphic with color changes via VR headset), Vibrations (for approval or disapproval for task performance)
[19]	**Language Learning (First)** : Teaching reading to children with dyslexia	**Tactile** : Letters and spatial arrangements recognized by pogo pins	**Visual/Auditory** : 3D Physical Prototype alphabets with colored LED, Graphics (colored text graphics), Voice (associated sound for letters)
[20]	**Sports Education** : Dancing supports from expert performances	**Auditory/Tactile** : Movement features (timing, intensity) from a 3D accelerometer, electromyography sensor, and vocalization from a microphone	**Auditory** : Voice (Pre-recorded experts' vocalization mapped and synthesized with learner's movement)
[23]	**Conceptual Learning** : Haptic rendering for elementary science education	**Visual** : Images that student took using mobile phone cameras	**Tactile** : Vibrations (haptic rendering which represents the texture of photos taken)
[24]	**Sports Education** : Motion analysis for golf swing	**Visual** : Visual marker recognition from motion capture cameras	**Visual/Auditory** : Graphics (comparing color-coded trajectory traces with reference), Sound Effects (putting sound with different pitch, indicating performance levels)
[29]	**Medical Education** : Injection simulator for veterinary education	**Visual / Tactile** : Visual marker from AR tracking camera and orientation from gyro sensor on syringe simulator	**Visual/Auditory/Tactile** : Graphic (AR/VR graphic with simulator and vein with evaluative color coding), Sound Effects (bell chimes and dog barks as evaluation), Physical Movements (force from syringe simulator)
[30]	**Conceptual Learning** : Companion robot for e-reading in higher education	**Visual** : Webcam-based video recording on learners' behaviors	**Visual/Auditory** : Robot Gestures (facial expressions), Robot LED (approval, disapproval), Robot Speech (dialogue)
[32]	**Language Learning (First)** : Calligraphy trainer based on expert's handwriting	**Tactile** : Touch from tablet and muscle activities, gesture embedded accelerometer and gyroscope on Myo sensor, and eye tracker glasses from SMI	**Visual/Tactile** : Graphics (characters with colored trajectory guidance), Dashboard (summative evaluation with graphs, video recording), Physical Movements (pressure feedback with saturation)

(*continued*)

Table 1. (*continued*)

Lit.	Learning Domains	Data Input	Feedback Modalities
[34]	**Musical Education** : Musical ensemble between human and machine	**Visual/Auditory** : Motion of cueing data from the camera, evaluating tempo from the microphone	**Visual/Auditory** : Graphics (projection of shadow-like pianist, Music (violin, viola, cello, double bass play in coordination)
[35]	**Sports Education** : Indoor rowing training with VR	**Visual/Tactile** : Hand movements from cameras and ergometer on handles	**Visual/Tactile** : Graphics (VR simulation for hand movement trajectories and gauge bar with color coding), Physical Prototype (feedback for hand position based on haptic markers)
[37]	**Clear Communication** : Automatic feedback for oral presentation skills	**Visual/Auditory** : Head gaze, posture detection from the camera, usage of filled pauses and volume from the microphone, pitches from text, and image size/density recognition from presentation slides	**Visual/Auditory**: Dashboard (post-hoc report with statistical scores and advice for weaknesses), Graphics (video, image recording), Audio (audio recording)
[38]	**Medical Education** : VR Epidural Administration	**Tactile** : Interaction with 3D-syringe with Novint Falcon, Leap Motion, and Touch 3D	**Visual/Auditory** : Graphics (virtual patient in a mixed-reality environment with Oculus Rift DK2, Dashboard (score, time completion, and number of attempts performed), Sound Effects (needle injection)
[41]	**Conceptual Learning** : Mathematics education for children with visual impairments	**Visual** : Image recognition based on markers (TopCodes)	**Auditory/Tactile** : Sound Effects (drum and piano), Voice (number reading, verbal rewards), Tangible user interface with braille (iCETA)
[42]	**Language Learning (First)** : Teaching handwriting for blind children	**Auditory/Tactile** : Audio pan and pitch represents from microphone and stylus gestures from Phantom Omni	**Auditory/Tactile** : Sound Effects (Pitch differences for drawing in the appropriate trajectory), Physical Movements (force feedback to minimize the distance with trajectory and playback feedback from Phantom Omni)
[45]	**Sports Education** : Virtual training for rowing skills	**Tactile** : Velocity of rowing tracking from accelerators	**Visual/Auditory/Tactile** : Graphic (Immersive virtual reality for immersion and optimal trajectory guidance), Sound Effects (errors), Vibrations (errors)

(*continued*)

Table 1. (*continued*)

Lit.	Learning Domains	Data Input	Feedback Modalities
[46]	**Clear Communication** : Presentation trainer for public speaking	**Visual/Auditory** : Body posture, hand gesture analysis based on a visual marker recognition from Kinect, voice volume, buffer, and pauses from the Kinect microphone	**Visual/Tactile** : Dashboard (mirrored video of the users), Graphics (symbolic icons for approval/disapproval), Text (evaluation and guidance), Vibrations (correction alert. through wristband)
[48]	**Language Learning (Foreign)** : Mobile tutoring system for English pronunciation	**Auditory** : Phoneme recognition from the microphone on mobile device	**Visual** : Graphics (flash card, emoji-based mouth shape animation), Dashboard (list of mispronounced words), Text (word-to-text with color-coded highlights)
[50]	**Medical Education** : Ultrasonography training	**Visual** : Position recognition of simulator from Kinect	**Visual** : Graphics (3D organ models and simulations, suggestive trajectory in guided mode)
[51]	**Clear Communication** : Presentation trainer	**Visual/Auditory/Tactile** : Body posture and movements from Kinect and voice volume, voice modulation, and phasing of speaking from microphone	**Visual** : Dashboard (posture analysis with skeleton, color-coded symbolic icons, suggestive text)
[52]	**Musical Education** : Violin play training	**Visual/Auditory/Tactile** : Video with visual markers from motion capture camera and Kinect, ambient audio from the microphone, and physiological (EMG) data from Myo sensors	**Visual/Auditory** : Dashboard (audio, video, and motion-capture recording with data visualization as post-hoc feedback)
[53]	**Medical Education** : Gross anatomy education for undergraduate medical students	**Tactile** : Stylus orientation from Phantom Omni haptic stylus	**Visual/Tactile** : Graphics (3D human anatomical structures), Physical Movements (rotation, touch feedback from Phantom Omni haptic stylus)
[54]	**Conceptual Learning** : Geometrical concept teaching for children with virtual 3D space	**Tactile** : Stylus orientation from Phantom Omni haptic stylus	**Visual/Auditory/Tactile** : Graphics (3D graphics with gamification), Sound Effects (immersive object moving sound), Physical Movements from Phantom Omni haptic stylus

states (e.g., cognitive load from pupil dilation and blinks based on eye tracking data [30,32]). However, such applications are often criticized for their obstructiveness. To compensate for such limitations, remote detection technologies that are developed and implemented in affective computing can be utilized: assessing learners' bio-data based on vision-based detection (e.g., heart rate [55]) and behavior recognition algorithms [30], without having to have intrusive biosensors implemented, which allows more stealth monitoring of learner activities.

With the current advancement of deep learning technologies, a subset of ML, and increased computational capabilities, high-resolution sensor data, with an unstructured data form, has become the resource of deep neural network developments. Deep neural networks can be used to make a sophisticated prediction about the learners' performance in LA, such as their attention prediction during e-reading [30], and provide personalized support. Before the emergence of deep learning technologies, only structured data with statistical explainability, such as log data from learning management systems, had been the target resources of traditional ML and LA [36]. However, dynamic data with uninterpretable patterns, such as image, video, sound, and text [43], are now widely used for the various model developments for the classification, prediction, and detection tasks [31]. It is expected that such emerging models developed based on unstructured or semi-structured information with non-numeric organizations, such as data from eye trackers and EEG data, will expand the horizons of MMLA.

Log-Based Data Collection

Log data, collected through learners' interaction with learning management systems via mouse clicks, keyboard inputs, and touch, in quantified forms (e.g., number, frequency) [26], have been the traditional sources of data in LA. Such data collection is mainly done in online platforms, such as MOOCs [18], with bigger sample scales and broader demographics than the sensor-based data collection. Although the collection of log data is often more accessible due to the absence of complex hardware infrastructure and sensors and can be interpreted with relative ease [36], the insights from log data have been subject to criticism for its superficial interpretations [33]. Also, its smaller computation requirements make it suitable for extensive data collection at larger scales. With the emergence of generative AI, the log data, such as the discourse between learners and the system, will become much more valuable due to its potential for personalized chat-based learning assistants, which will become more common with current advancements in Transformer-based Natural Language Processing (NLP) models (e.g., GPT-4) [27].

Questionnaire Data Collection

The questionnaire is one traditional data collection method for learning analytics: evaluating learning on objective (e.g., knowledge gain) and subjective levels

(e.g., learning experience). One common approach has been a pre-post questionnaire to measure the objective learning outcomes (e.g., knowledge gain) through task performances. With increasing emphasis on the User Experience (UX) in computer-assisted systems, more measures are developed and implemented for gauging the UX of the system (e.g., System Usability Scale (SUS) [8], Attrakdiff questionnaire [22]). Most existing measures have been developed especially for Human-Computer Interaction (HCI), which focus primarily on computer-based artifacts [2]. However, with the expansion of the physical and virtual ecosystems of TEL, there are emerging necessities for more standardized measures for evaluating the UX of various peripheral devices (e.g., AR/VR, conversational agents) and agents (e.g., virtual robots) [31]. Deciding the timing of questionnaire-based data collection (e.g., real-time, post-hoc) is one challenge, where researchers should balance the timely aspect and obstructiveness of the data collection.

Observation-Based Data Collection

Observation-based data collection is typical where the targeted learners are not fully capable of expressing their own perspectives (e.g., children with intellectual disability) or experts' opinion takes an essential role in evaluation (e.g., evaluating collaborative learning [9]). In such cases, observers' evaluation of observable indicators becomes the means of gauging learners' learning progress and performances [25]. The evaluation objectives are often learners' internal states, such as affects, attention, and perceived experiences [14,30,31], that influence learning experiences and potential learning outcomes. Since the evaluation is dependent on third-person observation, having clear annotation standards and frameworks is essential for the validity of the data. However, in some cases, practitioners often design and execute the measures themselves without having solid standards or frameworks [2]. Another challenge comes from individual differences: behaviors occur differently due to cultural backgrounds and individual differences [14], such as behavioral or emotional expressiveness. Alternative methods of combining human annotations with other layers of ground truths are suggested to compensate for such limitations: implementation of biosensor data (e.g., eye tracker [31] electroencephalography (EEG) [14]) and collecting self-reported ground truths [30] from learners.

4.2 RQ2. How Is Learner Feedback Designed in the Context of Multimodal Learning Systems?

Multimodal learning feedback in the form of in situ real-time feedback has often been provided via physical components, such as touch-based devices, wearables, haptic devices, physical prototypes, and speakers. In our literature search, in situ real-time feedback was more predominant than post-hoc feedback, while some took the hybrid approach. Such feedback often employed intuitive pictograms, color-coding, sound effects, vibration, and force feedback to provide immediate responses as learning interventions. In the meantime, post-hoc feedback has often

Multimodal Feedback			
Feedback Modality	Visual	Auditory	Tactile
Feedback Characteristics	Spacial / Temporal	Temporal	Temporal
Feedback Timing	Post-hoc / Real-time		Real-time
Feedback Functions	Semantic / Intuitive		Intuitive
Feedback Types	-Graphics -Dashboards -Text	-Sound Effects -Music -Voice	-Physical Movements -Vibrations

Fig. 1. Multimodal feedback involves decision-making regarding the feedback modalities, characteristics, timing, functions, and specific types of feedback.

been provided as dashboards, narrative text, and personalized voice feedback. Learner dashboards continue to be a prominent tool for supporting self-regulated learning in LA and MMLA. Dashboards in LA provide an easy-to-understand visual representation of complex learning data in real-time, which allows educators and students to make informed decisions. Generally, the feedback design in MMLA includes the following design elements: feedback modalities, characteristics, timing, types, and functions (see Fig. 1). In the following sections, we investigate the feedback elements found in the previous studies concerning their modalities.

Visual

Graphics are the primary visual element with intuitive delivery. The realistic graphical features have often been combined with engaging virtual environments [16,45], gamification elements [4,9,16] for specific learning objectives and contexts with enhanced immersion [54]. The symbolic features of the graphic have been used to communicate complex constructs. Symbolic pictograms, icons, and emojis are used [15,46] to help reinforce or correct learners' behaviors. Visual effects, such as 2D/3D effects and color coding [7,15,19,24,48], are used for highlighting specific information in the visual message delivery. Additionally, motion graphics/animations are used to convey dynamic information, such as a reference for the model trajectory and movements [24,45].

Dashboards commonly use post-hoc visual language with extensive and collective information. For example, statistical analysis of learning progress and performances has been shown through data visualization via graphs [1,5,47], tables [1,5], and gauge bars [35]. Multimodal learning systems track more sophisticated data from sensors capable of monitoring latent constructs in learners. Video [10,32] and audio recordings [52] are used as feedback for summative evaluation via dashboards so learners can reflect on their learning.

Text is used for its descriptive nature, capable of delivering narratives and details. It is a distinctive feature compared to other visual languages since the text relies on its semantic nature and the meaning layer, while visual languages

mainly depend on intuitive understanding. Thanks to its clarity in message delivery, text feedback has often been used in dashboards for written descriptions [15,46] and message alerts [7]. To differentiate the information hierarchy, some visual traits, such as font size [19], highlighting [48], and colors [15,48], have partially been applied to texts.

Auditory

Sound effects refer to types of auditory stimuli that are artificially made. Sound effects are used for positive feedback in dashboards for showing approval and rewards (e.g., bell chime [29]), while alerting sound effects are used for intuitively signaling learners for behavior corrections (e.g., dog barks [29]). Sound effects are also used for in situ real-time feedback [32] in multimodal learning systems for better immersion in certain educational scenarios (e.g., golf putting sound with different pitches [24]).

Voice feedback has been commonly implemented for its semantic and phonetic features. Since lexical meaning can be delivered through voice messages, vocal instructions are given for the concept delivery [6], guidance [37], and dialogue simulations [1]. The acoustic features have been mainly emphasized for assisting pronunciations of second language learners and young learners [1]. Various tonal differences were applied to the vocal feedback to highlight specific information or certain sound units.

Music has been implemented for musical education [34,52] and context-giving for the immersions. Musical traits of learners' instrumental play, such as tempo, pitch, and timbre, have been corrected by providing specific parts of the musical recording as guidance. Music can also create a certain ambiance with immersive visual aids.

Tactile

Physical movements have often been used for providing feedback on the psychomotor aspects of complex skill learning. For instance, model movements have been demonstrated for sports training (e.g., rowing [35,45]) and delivering abstract concepts [6,16,41,54]. Fine motor movements were given for handwriting education with the trajectory (e.g., handwriting [7,32,42]). The syringe prototype provided the force feedback [29], intertwining with physical probes for more effective veterinary training. Movement feedback has often been helpful for learners with visual impairments for compensating their limitations in visual knowledge acquisition [6].

Vibrations constitute the majority of tactile feedback found in literature, often referred to as vibrotactile feedback, in the form of small vibrations and frictions.

Vibrations are simplistic and are not able to encode complex information. Vibrations have been implemented for concept delivery (e.g., texture rendering [23]), guidance (e.g., haptic trajectory [7]), and as corrective feedback (e.g., vibration buzzers [16,46]). All vibration feedback, and other tactile feedback, have mostly been adopted as real-time feedback due to their temporal context-specific nature and, therefore, not used in the dashboards.

4.3 RQ3. What Are the Considerations for Implementing Multimodal Learning Systems in Various Learning Domains?

To answer RQ3, we analyzed the multimodal learning systems found in our literature according to the three learning domains from revised Bloom's taxonomy [28]: cognitive, psychomotor, and affective domains. The *cognitive domain* involves the development of our mental skills and acquiring knowledge. The *Psychomotor domain* relates to discreet physical functions, reflex actions, and interpretive movements of the human body, while the *affective domain* involves our feelings, emotions, and attitudes. Furthermore, we also cluster the learning systems according to their specific application domain and learning goals and present some of the largest clusters. It should be noted that it is not our intention to present learning systems as exclusive to one domain, and we only seek to categorize the learning systems according to their primary learning objective.

Cognitive Domain

Conceptual Learning: Multimodal feedback loops for conceptual learning primarily focus on facilitating knowledge delivery and comprehension. Providing haptic feedback, in addition to visual-oriented course content, to demonstrate various physical phenomena has shown an enhanced understanding of the phenomena in learners [16,23]. The inclusion of additional modalities in instructions of conceptual learning can assist learners with visual impairments (e.g., haptic feedback from a stylus on Phantom Omni[2] [6]).

Language Learning: Multimodality is also beneficial in various aspects of language learning, which has traditionally been considered a predominately cognitive domain. Improving pronunciation and intonation in learning a foreign language has been a commonly targeted learning objective through various methods, such as visual aids with ideal mouth movements [48] and audio feedback with standard pronunciation [1]. Children and learners with disabilities have been the main end user of learning systems for first-language learning. For children, teaching how to read [19], write with characters [7], and improve handwriting skills [7] has been the main focus. For writing tasks, force feedback has been commonly given through stylus [42,53,54] and colored trajectory feedback [7] to indicate learners' errors intuitively.

[2] http://www.immersion.fr/.

Medical Education: Multimodal learning systems for medical education have been implemented to compensate for textbook-oriented education, aiming at more practice-based learning. Yeom et al. [53] suggested a 3D visual and tactile education system offering vivid visuals and tactile structures of human organs for gross anatomy class. Similarly, Palpation education tools [4], ultrasonography simulator [50], injection simulators for human patients [38], and animals [29] have been designed to promote authentic real-life practices of such complex skills with mock-ups. Those mock-ups have embedded collective sensors and software for learning analytics and feedback so that learners can receive real-time feedback during their learning practices [4,29]. Medical education systems also tend to involve physical props, mainly for tactile data collection and embedded performance assessment algorithms to provide real-time instructions and feedback.

Affective Domain

Clear Communication Skills: Systems have been developed to improve clear communication skills during learners' presentations. With real-time evaluation, learners were asked to reflect on their performances and improve their skills over practice [15,37,46,51]. Combining visual and auditory data collected from a webcam, microphone, and Kinect [15,46,51], learners' posture, gaze, facial expression, and voice traits for clear communication have been evaluated. Systems gave the correction in real-time, by short written descriptions [15,46], real-time posture analysis [46,51], and performance analysis on the dashboards as post-hoc feedback [15,37,46].

Psychomotor Domain

Sports Education: In sports education or training, learning goals are predominantly psychomotor. As such, during sporting activities, real-time physical features have been evaluated: posture and strike patterns [24] for the golf swing, body orientation and posture [35] for rowing, and body movement [20] for dancing. These learning systems aim to correct learner errors in real time and offer actionable plans to improve learners' motor skills. As motor-skills development demands conscious repetitive practices [17] where learners rely on apprenticeship-based education, systems can computationally model experts or mentors [32] and use ML to provide real-time feedback.

Musical Education: Systems in musical education were implemented to support human-machine ensemble [34] and violin play [52]. Based on visual and auditory indicators collected from a camera and a microphone, the phasing of violin play was analyzed on the dashboard [52]. To support the human-machine play [34], the shadow visual of the pianist and pre-recorded music piece has been played along with the learner's play during repetitive practices. Through visual aid, learners were taught to understand the current issues, correct errors, and internalize better techniques in an analytical and reflective manner.

4.4 Challenges and Opportunities

Generalization vs. Personalization of Multimodal Learning Systems. While most systems aim at the best generalizability in the application, more and more learning systems target feedback provisions with personalization since one system should be general enough to cover targeted user groups while it should effectively reflect individuals' critical learning necessities. In this regard, future studies for refining the generalizability and personalization of critical learning necessities, timing, frequencies, and effects in multimodal learning system design would greatly benefit the community.

Overarching MMLA Frameworks for Higher-Level Learning Objectives. Multimodal learning systems are often modeled as domain-specific and context-based since most systems aim to improve concrete learning activities with clear system goals. However, in many cases, such goals are set based on fragmentary frameworks, lacking overarching models for higher-level learning objectives that can be universally applied to general domains or even domain-specific instructional design. Having such an overarching framework could work as a common ground where practitioners and researchers can exchange and share their knowledge and grow as a community while defining learning features is often the biggest challenge in MMLA with advancements in a data-driven approach.

Closing the Feedback Loop in MMLA. Our findings suggest that despite the advances in AI and ML algorithms, multimodal learning systems often fail to close the feedback loop. Though the systems we examined in our study included feedback in the system loop, most MMLA systems in the field need to take the current analytics into the context of the next round of feedback provision. In this sense, closing the feedback loop based on various modalities and evaluating the effect of the feedback loop for further optimization seems to be an essential challenge in the field.

5 Conclusion

In this scoping review, we investigated multimodal learning systems, which is an integral extension of modern TEL systems. We investigate systems in three stages as an extension to the MLeAM framework: 1) multimodal data collection and processing, 2) multimodal feedback design decisions, and 3) multimodal system implementation for various learning domains. The result indicates the necessity of a more holistic understanding of the whole process in order to design effective systems and multimodal instruction patterns. We also identified critical challenges in multimodal learning systems, such as defining learning indicators, balancing the generalization and personalization of analytics and interventions, and closing the feedback loop in multimodal learning systems. Our paper provides an overview of the role multimodality plays in defining the potential of the next generations of TELs and outlines important considerations for data

collection, feedback design, and MMLA design for adaptive TEL system implementations.

With more evidence-based, data-driven approaches taken in LA, the quality of data is getting increasingly important, especially in the context of MMLA. Although data is becoming more accessible through sensors on commercialized devices (e.g., laptops, webcam [30]) and increasing public datasets, engineering competencies are becoming more critical in MMLA [44] as how data is collected and processed impacts the quality of data and therefore, the predictions it makes. Based on our analysis of RQ1, MMLA builds upon, rather than replacing, traditional LA but using data from multiple modalities. By doing so, MMLA is able to make more robust predictions about learners' performance across multiple domains, as evidenced by RQ3, and can also provide more personalized feedback. However, even with the increased roles of data engineering and advancements in AI, researchers' insights, experiences, and domain knowledge are still critical [44]. For example, the black-box nature of ML models makes decisions and predictions in MMLA often not explainable and requires human interpretations. Explainable AI (e.g., tree-based models) [43] can supplement the current MMLA in partially addressing this issue by providing better interpretability of such analysis. This also holds true for the LA dashboards, as evidenced by RQ2, with the majority of multimodal learning systems still relying on the affordances of traditional LA dashboards, which support necessities for the stronger MMLA and feedback design based on multimodalities.

Acknowledgements. This work was supported by Leiden-Delft-Erasmus Centre for Education and Learning (LDE-CEL).

References

1. Ai, R., et al.: Sprinter: language technologies for interactive and multimedia language learning. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 2733–2738 (2014)
2. Alenljung, B., Lindblom, J., Andreasson, R., Ziemke, T.: User experience in social human-robot interaction. In: Rapid automation: concepts, methodologies, tools, and applications, pp. 1468–1490. IGI Global (2019)
3. Arksey, H., O'Malley, L.: Scoping studies: towards a methodological framework. Int. J. Soc. Res. Methodol. **8**(1), 19–32 (2005)
4. Asadipour, A., Debattista, K., Chalmers, A.: Visuohaptic augmented feedback for enhancing motor skills acquisition. Vis. Comput. **33**(4), 401–411 (2017)
5. Asadipour, A., Debattista, K., Patel, V., Chalmers, A.: A technology-aided multimodal training approach to assist abdominal palpation training and its assessment in medical education. Int. J. Hum. Comput. Stud. **137**, 102394 (2020)
6. Bernareggi, C., Ahmetovic, D., Mascetti, S.: μgraph: haptic exploration and editing of 3D chemical diagrams. In: The 21st International ACM SIGACCESS Conference on Computers and Accessibility, pp. 312–317 (2019)
7. Brondi, R., Satler, M., Avizzano, C.A., Tripicchio, P.: A multimodal learning system for handwriting movements. In: 2014 International Conference on Intelligent Environments, pp. 256–259. IEEE (2014)

8. Brooke, J., et al.: SUS-A quick and dirty usability scale. Usability Eval. Ind. **189**(194), 4–7 (1996)
9. Chen, H., Tan, E., Lee, Y., Praharaj, S., Specht, M., Zhao, G.: Developing AI into explanatory supporting models: an explanation-visualized deep learning prototype. In: The International Conference of Learning Science (ICLS) (2020)
10. Chollet, M., Ghate, P., Neubauer, C., Scherer, S.: Influence of individual differences when training public speaking with virtual audiences. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents, pp. 1–7 (2018)
11. Cope, B., Kalantzis, M., Group, N.L.: Multiliteracies: Literacy Learning and the Design of Social Futures. Literacies (Routledge), Routledge (2000). https://books.google.nl/books?id=a6eBiUQLJu4C
12. Cukurova, M., Giannakos, M., Martinez-Maldonado, R.: The promise and challenges of multimodal learning analytics. Br. J. Educ. Technol. **51**(5), 1441–1449 (2020)
13. Di Mitri, D., Schneider, J., Specht, M., Drachsler, H.: From signals to knowledge: a conceptual model for multimodal learning analytics. J. Comput. Assist. Learn. **34**(4), 338–349 (2018)
14. D'Mello, S., Graesser, A.: Mining bodily patterns of affective experience during learning. In: Educational data mining 2010 (2010)
15. Domínguez, F., Chiluiza, K.: Towards a distributed framework to analyze multimodal data. In: Proceedings of Workshop Cross-LAK-held at LAK 2016, pp. 52–57 (2016)
16. Edwards, B.I., Bielawski, K.S., Prada, R., Cheok, A.D.: Haptic virtual reality and immersive learning for enhanced organic chemistry instruction. Virtual Reality **23**(4), 363–373 (2019)
17. Ericsson, K.A., Krampe, R.T., Tesch-Römer, C.: The role of deliberate practice in the acquisition of expert performance. Psychol. Rev. **100**(3), 363 (1993)
18. Ezen-Can, A., Boyer, K.E., Kellogg, S., Booth, S.: Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach. In: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, pp. 146–150 (2015)
19. Fan, M., Antle, A.N., Cramer, E.S.: Design rationale: opportunities and recommendations for tangible reading systems for children. In: Proceedings of the The 15th International Conference on Interaction Design and Children, pp. 101–112 (2016)
20. Françoise, J., Fdili Alaoui, S., Schiphorst, T., Bevilacqua, F.: Vocalizing dance movement for interactive sonification of Laban effort factors. In: Proceedings of the 2014 Conference on Designing Interactive Systems, pp. 1079–1082 (2014)
21. Frolova, E.V., Rogach, O.V., Ryabova, T.M.: Digitalization of education in modern scientific discourse: new trends and risks analysis. Eur. J. Contemp. Educ. **9**(2), 313–336 (2020)
22. Hassenzahl, M., Wiklund-Engblom, A., Bengs, A., Hägglund, S., Diefenbach, S.: Experience-oriented and product-oriented evaluation: psychological need fulfillment, positive affect, and product perception. Int. J. Hum. Comput. Interact. **31**(8), 530–544 (2015)
23. Hightower, B., Lovato, S., Davison, J., Wartella, E., Piper, A.M.: Haptic explorers: supporting science journaling through mobile haptic feedback displays. Int. J. Hum. Comput. Stud. **122**, 103–112 (2019)
24. Jakus, G., Stojmenova, K., Tomažič, S., Sodnik, J.: A system for efficient motor learning using multimodal augmented feedback. Multimedia Tools Appl. **76**(20), 20409–20421 (2017)

25. Järvelä, S., Malmberg, J., Haataja, E., Sobocinski, M., Kirschner, P.A.: What multimodal data can tell us about the students' regulation of their learning process? Learn. Instruct. **72**, 101203 (2021)
26. Jia, J., He, Y., Le, H.: A multimodal human computer interaction system and its application in smart learning environments. In: Cheung, S.K.S., Li, R., Phusavat, K., Paoprasert, N., Kwok, L.F. (eds.) ICBL 2020. LNCS, vol. 12218, pp. 3–14. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51968-1_1
27. Kasneci, E., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023)
28. Krathwohl, D.R.: A revision of bloom's taxonomy: an overview. Theor. Pract. **41**(4), 212–218 (2002)
29. Lee, J., et al.: An intravenous injection simulator using augmented reality for veterinary education and its evaluation. In: Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry, pp. 31–34 (2012)
30. Lee, Y., Chen, H., Zhao, G., Specht, M.: WEDAR: webcam-based attention analysis via attention regulator behavior recognition with a novel E-reading dataset. In: 24th ACM International Conference on Multimodal Interaction (ICMI), pp. 319–328. ACM (2022)
31. Lee, Y., Specht, M.: Can we empower attentive E-reading with a social robot? An introductory study with a novel multimodal dataset and deep learning approaches. In: LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023). ACM (2023)
32. Limbu, B.H., Jarodzka, H., Klemke, R., Specht, M.: Can you ink while you blink? Assessing mental effort in a sensor-based calligraphy trainer. Sensors **19**(14), 3244 (2019). https://doi.org/10.3390/s19143244
33. Liu, S., d'Aquin, M.: Unsupervised learning for understanding student achievement in a distance learning setting. In: 2017 IEEE Global Engineering Education Conference (EDUCON), pp. 1373–1377. IEEE (2017)
34. Maezawa, A., Yamamoto, K.: MuEns: a multimodal human-machine music ensemble for live concert performance. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 4290–4301 (2017)
35. Maria, K., Filippeschi, A., Ruffaldi, E., Shorr, Y., Gopher, D.: Evaluation of multimodal feedback effects on the time-course of motor learning in multimodal VR platform for rowing training. In: 2015 International Conference on Virtual Rehabilitation (ICVR), pp. 158–159. IEEE (2015)
36. Mathew, A., Amudha, P., Sivakumari, S.: Deep learning techniques: an overview. Adv. Mach. Learn. Technol. Appl.: Proc. AMLTA **2020**, 599–608 (2021)
37. Ochoa, X., Domínguez, F., Guamán, B., Maya, R., Falcones, G., Castells, J.: The rap system: automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. In: Proceedings of the 8th International Conference on Learning Analytics and Knowledge, pp. 360–364 (2018)
38. Ortegon, T., et al.: Prototyping interactive multimodal VR epidural administration. In: 2019 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–3. IEEE (2019)
39. Oyedotun, T.D.: Sudden change of pedagogy in education driven by COVID-19: perspectives and evaluation from a developing country. Res. Globalization **2**, 100029 (2020)
40. Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., Mirriahi, N.: Using learning analytics to scale the provision of personalised feedback. Br. J. Educ. Technol. **50**(1), 128–138 (2019)

41. Pires, A.C., et al.: Learning Maths with a tangible user interface: lessons learned through participatory design with children with visual impairments and their educators. Int. J. Child-Comput. Interact. **32**, 100382 (2022)
42. Plimmer, B., Reid, P., Blagojevic, R., Crossan, A., Brewster, S.: Signing on the tactile line: a multimodal system for teaching handwriting to blind children. ACM Trans. Comput. Hum. Interact. (TOCHI) **18**(3), 1–29 (2011)
43. Rahman, M.A., Brown, D.J., Shopland, N., Burton, A., Mahmud, M.: Explainable multimodal machine learning for engagement analysis by continuous performance Test. In: Antona, M., Stephanidis, C. (eds) Universal Access in Human-Computer Interaction. User and Context Diversity. HCII 2022. Lecture Notes in Computer Science. vol 13309. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-05039-8_28
44. Romero, C., Ventura, S.: Educational data mining and learning analytics: an updated survey. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **10**(3), e1355 (2020)
45. Ruffaldi, E., Filippeschi, A., Avizzano, C.A., Bardy, B., Gopher, D., Bergamasco, M.: Feedback, affordances, and accelerators for training sports in virtual environments. Presence: Teleoperators Virtual Environ. **20**(1), 33–46 (2011)
46. Schneider, J., Börner, D., Van Rosmalen, P., Specht, M.: Presentation trainer, your public speaking multimodal coach. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 539–546 (2015)
47. Sebkhi, N., Desai, D., Islam, M., Lu, J., Wilson, K., Ghovanloo, M.: Multimodal speech capture system for speech rehabilitation and learning. IEEE Trans. Biomed. Eng. **64**(11), 2639–2649 (2017)
48. Shukla, S., Shivakumar, A., Vasoya, M., Pei, Y., Lyon, A.F.: iLEAP: a human-AI teaming based mobile language learning solution for dual language learners in early and special educations. In: International Association for Development of the Information Society (2019)
49. Silva, M.J.: Children using electronic sensors to create and use knowledge on environmental health. First Monday (2020)
50. Sokolowski, J.A., Garcia, H.M., Richards, W., Banks, C.M.: Developing a low-cost multi-modal simulator for ultrasonography training. In: Proceedings of the Conference on Summer Computer Simulation, pp. 1–5 (2015)
51. Van Rosmalen, P., Börner, D., Schneider, J., Petukhova, O., Van Helvert, J.: Feedback design in multimodal dialogue systems. In: International Conference on Computer Supported Education (CSEDU). vol. 2, pp. 209–217 (2015)
52. Volpe, G., et al.: A multimodal corpus for technology-enhanced learning of violin playing. In: Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter, pp. 1–5 (2017)
53. Yeom, S., Choi-Lundberg, D.L., Fluck, A.E., Sale, A.: Factors influencing undergraduate students' acceptance of a haptic interface for learning gross anatomy. Interact. Technol. Smart Educ. **14**(1), 50–66 (2017)
54. Yiannoutsou, N., Johnson, R., Price, S.: Exploring how children interact with 3D shapes using haptic technologies. In: Proceedings of the 17th ACM Conference on Interaction Design and Children, pp. 533–538 (2018)
55. Yu, Z., Li, X., Niu, X., Shi, J., Zhao, G.: AutoHR: a strong end-to-end baseline for remote heart rate measurement with neural searching. IEEE Sig. Process. Lett. **27**, 1245–1249 (2020)