Humanities, Arts and Social Sciences Studies

Research Article

Humanities, Arts and Social Sciences Studies
https://so02.tci-thaijo.org/index.php/hasss
ISSN (Online): 2630-0079

MODALITY EFFECTS ON MULTIMEDIA LEARNING ACROSS PROFICIENCY LEVELS: EVIDENCE FROM THAI LEARNERS OF JAPANESE

Tewich Sawetaiyaram

School of Language and Communication, National Institute of Development Administration, Thailand

ABSTRACT

Mayer's (2009) multimedia learning theory introduces the Modality Principle, which holds that people learn better when visuals are paired with narration rather than text. While this principle has been widely supported in studies involving native speakers, findings from research on second language (L2) learners have been less consistent. This study examined how media type (Picture + Narration, PN vs. Picture + Text, PT) and Japanese proficiency (JLPT level) influenced comprehension and watch time (i.e., time-on-task) among Thai learners. A total of 116 learners were grouped by JLPT level (N1, N2, N3) and randomly assigned to one of two media types (PN or PT). After viewing the materials, participants completed a five-item comprehension test designed to reflect different levels of cognitive processing based on Bloom's revised taxonomy. The results of binary logistic regression analysis indicated that higher-proficiency learners performed better on items requiring higher levels of understanding and analysis, whereas no statistically significant effect of media type on comprehension was observed. Two-way ANOVA revealed both proficiency and media type affected watch time, as well as a significant interaction between the two variables. Lower-proficiency learners, particularly at the N3 level, spent more time on text-based media. These findings suggest that learner proficiency may play a more prominent role than media modality in shaping comprehension outcomes and processing behavior.

Keywords: Modality principle; multimedia learning; Picture and Narration; Picture and Text; Japanese language proficiency levels

Corresponding author:
Tewich Sawetaiyaram
sawetaiyaramtewich@gmail.com

Received: *6 October 2025*
Revised: *14 September 2026*
Accepted: *15 September 2026*
Published: *23 November 2026*

Citation:
Sawetaiyaram, T. (2026). Modality effects on multimedia learning across proficiency levels: Evidence from Thai learners of Japanese. Humanities, Arts and Social Sciences Studies, 26(3), 814–829. https://doi.org/10.69598/hasss.26.3.281569

1. INTRODUCTION

The COVID-19 pandemic ushered in a "new normal" in Thai society, prompting all sectors—including the field of education—to integrate digital technology into their practices. In particular, teaching and learning have seen rapid transformation through the adoption of digital tools, which have altered learner behavior, stimulated active learning, reduced boredom, and enhanced learning experiences both inside and outside the classroom (Boonphak, 2020; Malithong, 2000). The use of multimedia has been recognized as an effective approach to develop digital learning materials. Such media are typically developed in the form of computer-assisted instructional lessons that learners can study independently on computers. These media combine two

or more types of single-channel formats—for example, images and text or images and narration—offering greater motivation and improved comprehension compared to using a single medium alone (Malithong, 2000; Mayer, 2009).

As a result, the design of multimedia materials has become a critical concern for educators. Among the prominent design frameworks is Mayer's (2009) principles of multimedia learning, which propose several guidelines such as: learners learn better from media that include images and narration (Multimedia Principle); learning improves when irrelevant content is excluded (Coherence Principle); and media that pair images with narration are more effective than those that combine images and text (Modality Principle). Of the various concepts within Mayer's Cognitive Theory of Multimedia Learning (CTML), the Modality Principle is of particular interest because it involves presenting information through distinct perceptual channels—auditory and visual—which affect cognitive load and learning efficiency.

The Modality Principle posits that learning is more effective when images are paired with spoken narration rather than with written text. This assumption is based on the theory that short-term memory comprises separate channels for auditory and visual information. Using both channels simultaneously helps reduce cognitive load and facilitates learning.

A substantial body of research supports this principle among native English-speaking learners (Atkinson, 2002; Craig et al., 2002; Jeung et al., 1997; Kalyuga et al., 1999, 2000; Leahy et al., 2003; Mayer et al., 2003; Mayer & Moreno, 1998; Moreno & Mayer, 1999, 2002; Moreno et al., 2001; Mousavi et al., 1995; Tindall-Ford et al., 1997). However, findings from studies on second-language (L2) learners have been mixed, particularly with respect to content comprehension. Some studies have reported results consistent with the Modality Principle, showing that learners in the PN condition achieved better comprehension than those in the PT condition (e.g., Harskamp et al., 2007). In contrast, other studies have found no significant differences between the two media types (e.g., Hanrattanasakul, 2017), while some have reported partial or task-dependent effects. For example, Chen (2006) found advantages for PN in immediate vocabulary recall but not in content comprehension, and Tabbers et al. (2001) reported that modality effects depended on presentation conditions, emerging under system-paced but not user-paced environments.

Despite the growing body of research, studies on second-language multimedia remain limited, particularly those examining learners across multiple proficiency levels. Furthermore, it remains unclear how this principle affects content comprehension among learners with different language proficiency levels. Studies involving non-English second-language multimedia contexts remain relatively scarce.

Among the few is a study by Hanrattanasakul (2017), which examined Japanese-language multimedia. However, this study used only five comprehension questions post-viewing, all claimed to assess the understanding level of cognitive behavior and were analyzed as a composite score without item-level differentiation. A closer examination of the test items suggests that they may not uniformly represent the same cognitive level. While some questions appear to involve direct recall of information, others seem to require more complex processes, such as interpreting relationships or drawing inferences. This variation is important because, even within the same category in Bloom's revised taxonomy (Anderson & Krathwohl, 2001), different subskills (e.g., inferring versus interpreting) may place different demands on learners. In practice, tasks involving abstract concepts or unfamiliar terminology may be more challenging than those based on concrete and familiar content, despite being classified at the same level. Bloom's revised taxonomy (Anderson & Krathwohl, 2001) also emphasizes that cognitive categories do not necessarily correspond to levels of difficulty. For this reason, learner comprehension should be considered more broadly, rather than relying solely on taxonomy-based classifications. From this perspective, item-level analysis becomes essential for a more precise understanding of learner performance, as the lack of such analysis makes it difficult to determine how different types of comprehension are affected by media design, particularly across learners with varying proficiency levels.

In addition to these issues related to measurement and task design, another important but underexplored aspect concerns how learners engage with multimedia over time. In particular, quantitative learning behavior, such as watch time (i.e., time-on-task), may provide further insight into how learners process multimedia content. Prior research has suggested that the time spent engaging with instructional materials can influence learning outcomes (Hanrattanasakul, 2017; Harskamp et al., 2007; Tabbers et al., 2004). From the perspective of Cognitive Load Theory (Sweller, 1994), learners who lack well-developed schemas may require more time to process information, which can increase cognitive load and hinder learning. In this sense, longer watch time may reflect processing difficulty rather than successful comprehension, especially among lower-proficiency learners or when dealing with more complex content. However, previous studies have rarely examined this aspect in depth, particularly with respect to differences across proficiency levels and media types. As a result, it remains unclear how watch time varies among learners with different levels of language proficiency and how it interacts with instructional design. Examining this issue may provide

a more comprehensive understanding of learning behavior and support the development of more effective multimedia materials.

Building on these considerations, this study aims to address the limitations identified in previous research by examining both comprehension outcomes and learning behavior in a second-language multimedia context. Specifically, this study investigates whether multimedia that combines images with spoken narration leads to better comprehension than image-text presentations, and whether media type influences learners' watch time. In addition, the study explores how these effects vary across learners with different levels of Japanese language proficiency.

By incorporating item-level analysis and considering both comprehension and engagement measures, the study seeks to provide a more comprehensive understanding of how multimedia design interacts with learner characteristics. The findings are expected to offer insights that can inform the design of more effective multimedia materials for second-language learning. Given the practical importance of multimedia design in Japanese language teaching, it is necessary to examine the applicability of the Modality Principle in this context.

2. RESEARCH OBJECTIVES

1. To investigate whether different types of multimedia and varying levels of Japanese language proficiency affect learners' comprehension of content.
2. To examine whether different types of multimedia and varying levels of Japanese language proficiency affect learners' watch time, and whether there is an interaction effect between these variables on watch time.

3. RELATED LITERATURE

The Modality Principle is one of the key principles of multimedia design proposed by Mayer (2009). Mayer found that American learners performed better when studying materials consisting of images and narration (Picture-Narration: PN) than when using materials with images and on-screen text (Picture-Text: PT). This is because PT delivers information through the visual channel only, while PN divides the cognitive load between the visual and auditory channels, reducing sensory memory overload and enhancing the potential for new understanding (Mayer, 2009).

This principle is supported by 13 studies involving native English speakers. All studies divided learners into PN and PT groups and assessed comprehension after learning. The results consistently showed that PN materials significantly improved content comprehension (Atkinson, 2002; Craig et al., 2002; Jeung et al., 1997; Kalyuga et al., 1999, 2000; Leahy et al., 2003; Mayer et al., 2003; Mayer & Moreno, 1998; Moreno & Mayer, 1999, 2002; Moreno et al., 2001; Mousavi et al., 1995; Tindall-Ford et al., 1997). However, findings from studies using English-language multimedia with second-language (L2) learners have been inconsistent (Chen, 2006; Harskamp et al., 2007; Karbalaei et al., 2016; Tabbers et al., 2001).

Chen (2006) examined 78 intermediate adult ESL learners and compared audio-picture and text-picture annotations under intentional and incidental learning conditions. These annotation formats broadly correspond to the modality contrast represented by PN and PT, respectively. Audio-picture annotations produced significantly better immediate vocabulary recall than text-picture annotations. However, no significant overall difference between the two annotation types was found in reading comprehension, although the written-recall results varied by learning condition. Thus, the findings were consistent with the Modality Principle for immediate vocabulary recall but not consistently for reading comprehension. Similarly, Karbalaei et al. (2016) examined advanced Iranian EFL learners whose proficiency was assessed using the Preliminary English Test. They found that audio-picture annotations produced significantly better immediate vocabulary recall than text-picture annotations. However, their study did not assess content comprehension.

Harskamp et al. (2007) conducted two experiments with Dutch EFL learners (different samples in each experiment), dividing participants into PN and PT groups using automated English slide presentations. They found that PN groups comprehended better than PT groups. Notably, in the second experiment, this superiority was observed only when the learners studied for less than nine minutes—beyond that, the difference disappeared.

Tabbers et al. (2001) conducted two experiments using a 4C/ID-model lesson with Dutch (Experiment 1) and Belgian (Experiment 2) learners. In Experiment 1, learners were assigned to PN and PT groups with automated media. Experiment 2 used a 2 × 2 factorial design (Auto-paced/User-paced × PN/PT). In Experiment 1, no significant difference was found. However, in Experiment 2, PN outperformed PT significantly under automated conditions. No significant differences were found in test performance under user-paced conditions,

although the audio group reported lower mental effort. In addition, learners in the text-based condition spent more time engaging with the material than those in the narration condition.

Hanrattanasakul (2017) studied Thai learners of Japanese, dividing 80 participants into PN and PT groups (n = 40 each), with subgroups based on JLPT proficiency levels (N2 and N3, n = 20 each). After viewing the materials, participants answered five comprehension questions categorized as understanding per Bloom's taxonomy. The total score was analyzed using two-way ANOVA. No significant effects were found for either media type or proficiency level. However, the PT group spent significantly more time studying the material than the PN group ($p < .05$). Hanrattanasakul concluded that while media type did not affect comprehension, watch time might be an influential factor.

Previous research indicates that multimedia with images and narration (PN) improves vocabulary recall more effectively than media with images and text (PT) (Chen, 2006; Karbalaei et al., 2016), aligning with the Modality Principle. However, results regarding content comprehension are mixed. Some studies support the principle (e.g., Harskamp et al., 2007), others report contrary findings (e.g., Chen, 2006; Hanrattanasakul, 2017), and some yield conflicting findings within the same study (Tabbers et al., 2001). One reason for this inconsistency may be that most studies examine learners at only a single proficiency level, making direct comparisons difficult.

Hanrattanasakul (2017), who studied Japanese language learners—a group similar to that of the present research—raises several issues for consideration. For example, although Hanrattanasakul categorized all five comprehension questions as understanding level, a review of the materials revealed that some items required higher-order thinking than simple understanding. Analyzing only the total score may obscure item-level differences. Item-level analysis is likely to provide more meaningful insight into learner performance. Furthermore, Hanrattanasakul focused only on learners at N2 and N3 levels. However, higher proficiency may also influence learners' ability to comprehend multimedia content. It is therefore worth investigating whether learners with higher Japanese proficiency would yield different results.

In addition to these cognitive and learner-related explanations, multimedia learning may also be understood from a multimodal perspective. Kress and van Leeuwen (2001) argue that different modes, such as image, sound, and text, are not interchangeable, as each provides distinct affordances for meaning-making. Building on this view, Kress (2010) suggests that meaning emerges through the interaction and coordination of multiple semiotic resources rather than from any single mode. This perspective may help explain why findings related to the Modality Principle in second-language contexts have been inconsistent. While cognitive load theory predicts clear advantages for certain media formats, learners may in practice draw on multiple semiotic resources in flexible ways, particularly under learner-paced conditions. From this perspective, differences between media conditions may be less pronounced or may not consistently emerge in comprehension outcomes.

Given these limitations, the present study aims to broaden understanding of the Modality Principle by examining how different types of multimedia and varying Japanese language proficiency levels influence learners' comprehension of content and watch time.

4. METHODOLOGY

4.1 Participants

The participants in this study were 116 Thai learners who had passed the Japanese Language Proficiency Test (JLPT) within the past two years. This criterion was adopted to ensure that proficiency groupings reflected current language ability. As the JLPT is administered twice annually in Thailand, the available pool of eligible participants at each proficiency level within the recruitment period was limited. Based on their JLPT certifications, they were categorized into three proficiency levels: N1 (n = 39), N2 (n = 38), and N3 (n = 39). Within each proficiency group, participants were randomly assigned to one of two media conditions. The PN group (Picture + Narration) consisted of N1: 19, N2: 19, and N3: 20 learners. The PT group (Picture + Text) consisted of N1: 20, N2: 19, and N3: 19 learners.

4.2 Research instruments

The instrument developed by Hanrattanasakul (2017) was used in this study. It was originally designed for Thai learners of Japanese at the N2–N3 levels, which are broadly comparable to the participant group in the present study (N1–N3), particularly at the intermediate and advanced levels. This alignment reduces variability associated with differences in learner background. The original study reported expert review to support content appropriateness. However, detailed psychometric properties were not fully established. In the present study, additional procedures were conducted, including expert review, pilot testing, and reliability analysis, to further examine the suitability of the instrument for the current research context.

The instrument compares two multimedia formats, Picture + Narration (PN) and Picture + Text (PT), which correspond directly to the independent variable of this study. This structure ensures a clear alignment between the measurement and the research objectives, while also allowing comparison with previous studies without introducing additional variation from the development of a new instrument. The overall structure of the instrument was retained without modification. It consists of multimedia materials presented in two formats (PN and PT), followed by a comprehension test. The instrument was further examined through expert review and pilot testing, and details of each component are provided in the following sections.

4.2.1 Multimedia materials

The multimedia materials used in this study were based on those developed in Hanrattanasakul (2017). Two formats were created: PN (Picture + Narration) and PT (Picture + Text). The PN version consisted of images paired with Japanese audio narration delivered at a natural pace by a native Japanese speaker. The PT version displayed the same images accompanied by Japanese on-screen text.

The content in both versions was identical, differing only in the mode of presentation, either via narration or text. In both cases, the image appeared first, followed by either the narration or text after a one-second delay. No background music or irrelevant text was included in either version, in accordance with Mayer's (2009) Coherence Principle, which states that extraneous information may impose cognitive load that interferes with learning.

The multimedia content focused on the topic of hot springs in Japan. It began with a general introduction to hot springs and their cultural significance, followed by a concrete example of Japanese monkeys bathing in hot springs as a response to cold weather. The content then moved to more abstract explanations of why Japanese people value hot springs, including references to Shinto purification rituals (*misogi*) and Buddhist teachings suggesting that bathing in hot water brings merit. The content was organized in a progression from familiar and observable phenomena to more abstract cultural and religious interpretations.

To provide a clearer overview of the structure and content of the materials, the instructional sequence was organized into 15 scenes. Each scene represents a segment of the lesson with varying levels of conceptual complexity. An overview of the content and length of each scene is presented in Table 1.

Table 1: Overview of multimedia content by scene

Scene	Content summary	Length (characters)
1	Greeting and instructions to begin the lesson	54
2	Introduction to the topic of hot springs and activation of learners' prior experience	71
3	Overview of hot springs in Japan and their cultural significance	128
4	Introduction of monkeys bathing in hot springs as a real-world example	56
5	Location and seasonal behavior of monkeys visiting hot springs	52
6	Explanation of monkeys' behavior in relation to cold weather conditions	64
7	Origin of the monkey hot spring phenomenon and its development	104
8	Transition from animal examples to human cultural practice	38
9	Introduction of religious influences on Japanese bathing culture (Shinto)	60
10	Explanation of purification rituals (*misogi*) in Shinto practices	69
11	Description of purification practices using cold water in traditional contexts	48
12	Introduction of Buddhism and the concept of bathing in hot water	55
13	Explanation of the "*Unjitsu-gyō*" and its religious implications	67
14	Cultural acceptance of hot water bathing in relation to religious practices	89
15	Summary interpretation of hot spring preference as a fusion of religious beliefs	94

As shown in Table 1, the materials progress from concrete examples to more abstract cultural and religious explanations, which may impose increasing cognitive demands on learners. The full script of the multimedia materials is provided in Appendix A, with an English translation available in Appendix B.

Both multimedia formats were user-paced. Each screen featured a "next" button that allowed learners to proceed to the next scene once they had completed the current one. However, they were not allowed to return to previous scenes. In the PN condition, learners could replay the narration using a "repeat" button.

The only functional difference between the two versions was the mode of verbal information delivery, which was either spoken narration (PN) or written text (PT). Watch time was recorded as the total time each participant spent engaging with the multimedia materials, from the moment the "start" button was pressed until the lesson ended, including any replays of the narration.

4.2.2 Content of the multimedia

The multimedia content was presented in Japanese and focused on the topic of hot springs in Japan. It began with a description of monkeys bathing in hot springs, including the reasons for this behavior, and then moved to broader explanations of why hot springs are culturally significant to Japanese people. In particular, the content addressed three main factors: the Shinto purification ritual (*misogi*), Buddhist teachings (*Unjitsu-gyō*), and interpretations proposed by hot spring experts regarding Japanese preferences for hot spring bathing. The content was organized in a progression from concrete and observable phenomena to more abstract cultural and religious interpretations. The topic was selected because it represents a familiar and culturally meaningful aspect of Japanese society and allows the integration of both concrete and abstract information within a single narrative.

The multimedia used a conversational style intended to address learners directly. It employed polite forms (*masu*-form) and invitational phrases (e.g., *desu ne*, *mashō*), creating a sense of engagement. This design followed Mayer's (2009) Personalization Principle, which emphasizes that conversational-style media foster better learner comprehension. The linguistic accuracy of the content was verified by native speakers of Japanese.

The material was organized into 15 scenes. For the analysis, scene divisions were marked by a single line break, and Japanese punctuation was used consistently throughout the text.

jReadability (Lee et al., n.d.) was used to examine grammatical complexity and overall readability. The system applies the model developed by Lee and Hasebe (2019), using measures such as sentence length, vocabulary composition, and the distribution of verbs and auxiliary verbs. According to the analysis, the material comprised 34 sentences, 700 word tokens, 198 lexical types, and 1,065 characters, including symbols and white spaces. The mean sentence length was 20.59 words. The text received a readability score of 3.63 and was classified as Lower Intermediate (*chūkyū zenhan*, 中級前半).

Vocabulary levels were analyzed using Reading Tutor (Kawamura et al., 2013), which classifies words according to JLPT-based categories. Reading Tutor uses the pre-2010 JLPT Test Content Specifications, which divided proficiency into four levels (Levels 1–4). Because the former system did not include a separate N3 category, Reading Tutor reports the corresponding items under the combined label N2–N3. The system identified 548 units in total. Of these, 81 were punctuation marks and one was a numeral. All 82 non-lexical units were excluded, leaving 466 vocabulary occurrences for the vocabulary-level analysis. Repeated occurrences were counted separately. Of the 466 occurrences, 309 (66.3%) were classified as N5, 54 (11.6%) as N4, 69 (14.8%) as N2–N3, 6 (1.3%) as N1, and 28 (6.0%) as *kyūga*i (級外). The five categories accounted for all vocabulary occurrences, with N4 and N5 together representing 77.9% of the total.

The term *kyūgai* indicates only that an item was not included in the vocabulary list used by Reading Tutor; it does not mean that the item is more difficult than N1 vocabulary. This category included general-use words, such as 温かい (*atatakai*, "warm") and 多く (*ōku*, "many"), as well as more specialized or linguistically complex items, such as 禊 (*misogi*, "ritual purification") and 耐え難い (*taegatai*, "unbearable"). The *kyūgai* items were therefore not treated as a group of uniformly difficult vocabulary.

All participants held JLPT certification at N3 or above. Since N3 was the lowest certified level represented in the sample, the predominance of N4–N5 vocabulary indicates that most vocabulary occurrences were classified below the lowest certified JLPT level in the group. It is therefore reasonable to expect that much of the vocabulary fell within the participants' general proficiency range, although this does not imply that every participant knew every N4 or N5 word.

Reading Tutor and jReadability describe different features of the text. Reading Tutor provides a JLPT-based vocabulary profile, whereas jReadability estimates overall textual readability. The numerical outputs of the two systems are reported according to their respective analytical procedures and are not directly comparable. Neither analysis directly measures individual vocabulary knowledge or comprehension, and the results are not interpreted as establishing a percentage threshold for comprehension.

This progression also involved different levels of cognitive demand. In terms of the revised taxonomy proposed by Anderson and Krathwohl (2001), the earlier sections mainly required learners to remember and understand concrete information, such as the description of monkeys bathing in hot springs. The later sections required them to interpret relationships among cultural and religious ideas. For example, the statement「日本は世界の温泉大国で」(*Nihon wa sekai no onsen taikoku de*, "Japan is one of the world's leading hot-spring countries") presents factual information that can be understood directly. In contrast, the explanation「温泉に入る行為は宗教的に大変いい行いだということです」(*Onsen ni hairu kōi wa shūkyōteki ni taihen ii okonai da to iu koto desu*, "Bathing in hot springs is considered a highly virtuous act in religious terms") requires learners to interpret an event within a cultural context. The comprehension questions covered a corresponding range of cognitive processes.

The amount and content of verbal information were held constant across conditions to ensure a fair comparison between modalities. In addition, the overall appropriateness of the content was assessed through expert review and pilot testing, which indicated that the material was suitable for the target participants and that the accompanying task could be completed within a reasonable time without major difficulty.

4.2.3 Comprehension test

After viewing the multimedia materials, participants completed a comprehension test. The test items and response options were originally developed in Thai by Hanrattanasakul (2017) and were retained in their original form in the present study. The test was administered in Thai to ensure that comprehension was assessed independently of participants' ability to process test questions in Japanese. The items were translated into English for reporting purposes only, with permission obtained from the original author. The wording, structure, and content of the test were retained to ensure comparability with the original study.

In the original study, the instrument underwent expert review involving specialists in Japanese language teaching and educational technology to ensure content appropriateness. However, detailed psychometric properties were not fully established. In the present study, additional procedures were therefore conducted to examine the suitability of the instrument for the current research context.

To establish content validity, the instrument was reviewed by three experts with experience in teaching Japanese as a foreign language. The review focused on content relevance, clarity of instructions, item difficulty, and alignment with learner proficiency levels. The feedback confirmed that the items were appropriate for the intended population without requiring modification.

A pilot test was conducted with 10 participants across N1–N3 proficiency levels. The pilot examined the clarity of instructions, task feasibility, and time required to complete the test. The results indicated that the test could be completed within a reasonable time frame and did not present major difficulties for participants.

The test consisted of five multiple-choice items. Although the original study classified all items as measuring the understanding level based on Bloom's taxonomy, further examination suggested that the items varied in cognitive demands. Some items required direct recall of information (remembering), while others involved interpreting or evaluating relationships between ideas (understanding and analyzing). Therefore, the items were reclassified according to Bloom's revised taxonomy (Anderson & Krathwohl, 2001) to better reflect their cognitive characteristics, as presented in Table 2. No changes were made to the content of the test.

Table 2: Questions and response options

Questions	Response options	Bloom's revised taxonomy
1. Why do Japanese monkeys soak in hot springs?	A. To maintain the beauty of their fur B. To regulate their body temperature C. To relieve the cold D. To shed their fur	Remembering
2. What is the main purpose of building hot spring baths for monkeys?	A. To attract tourists B. To prevent monkeys from entering hot springs meant for humans C. To solve the problem of monkeys dying from the cold D. To create a pleasant atmosphere for the town	Understanding (Inferring)
3. What is the purpose of the *Misogi* ritual?	A. To purify the body and mind B. To practice self-discipline in ancient Japanese culture C. To ward off evil spirits D. To accumulate merit	Remembering
4. Which Buddhist teaching relates to the Japanese practice of soaking in hot springs?	A. People must wash their body and mind before important rituals B. Bathing in cold water is a sin C. Bathing in hot water brings good karma D. Both hot and cold water can purify the body and mind	Understanding (Interpreting)
5. Which of the following is *not* related to why Japanese people like hot springs?	A. The fusion of two religions B. Ritual cleansing of the body and mind before important ceremonies C. The belief that soaking in hot water brings merit D. Old wisdom to protect against cold weather	Analyzing

The test was conceptualized not as a unidimensional scale, but as a set of heterogeneous items designed to capture multiple levels of cognitive processing. This approach allows for a more nuanced assessment of comprehension across different cognitive demands, rather than assuming a single underlying construct. Cronbach's alpha ($\alpha = .374$) was calculated based on data collected in the main study for descriptive

purposes. However, this coefficient should be interpreted with caution. Given the limited number of items and their heterogeneous nature, internal consistency is not an appropriate primary indicator of measurement quality in this context. Instead, the study adopts an item-level analytical approach, allowing each item to function as an independent indicator of specific cognitive demands.

Item-level analysis was employed to examine learner performance across different cognitive levels. This approach provides a more refined interpretation of comprehension outcomes and aligns with the theoretical framework underlying the design of the instrument.

Nevertheless, the limited number of items and their heterogeneous nature may constrain the reliability of aggregated measures. Therefore, the findings should be interpreted with caution, particularly when extending beyond the specific context of the present study.

4.3 Data analysis

There were two independent variables: learner proficiency level (three groups: N1, N2, N3) and media type (two groups: PN, PT), forming a 2 × 3 factorial design.

To analyze content comprehension (Objective 1), Binary logistic regression was used to examine both main and interaction effects of proficiency level and media type on the binary outcome of each comprehension item (correct = 1, incorrect = 0). The analysis reported Exp(B) values (odds ratios), significance levels (p-values), Nagelkerke R^2, and model accuracy. The Hosmer–Lemeshow test was used to assess model fit. Statistical significance was set at $p < .05$.

To analyze watch time (Objective 2), two-way ANOVA was used to examine the effects of proficiency level, media type, and their interaction (Ability × Media) on mean watch time (in seconds), as recorded from the moment participants pressed the "start" button until the lesson ended, including any replays of the narration.

Where significant differences were found, post-hoc Tukey HSD tests were conducted to explore pairwise differences across proficiency levels. In addition, simple main effects analysis was conducted to explore interaction effects, identifying how media type influenced watch time within each proficiency level.

4.4 Ethical considerations

This study was approved by the Institutional Review Board of the National Institute of Development Administration (NIDA) (Approval No. COA2025/0092). All participants were informed about the purpose of the study and provided informed consent prior to participation. Participation was voluntary, and all responses were kept confidential and used solely for research purposes.

5. RESULTS

Table 3 presents descriptive statistics including the number of learners in each group (n), the percentage of correct responses for each comprehension question (Q1–Q5), and the mean and standard deviation of watch time (measured in seconds), disaggregated by JLPT level and media type.

Table 3: Descriptive statistics: Percentage of correct responses for each comprehension question, mean and standard deviation of watch time

JLPT level	Media type	N	Q1 (%)	Q2 (%)	Q3 (%)	Q4 (%)	Q5 (%)	Watch time (M, sec)	Watch time (SD, sec)
N1	PN	19	100	84.2	100	47.4	78.9	288.16	102.11
N1	PT	20	100	75	100	60	70	328.27	85.1
N2	PN	19	100	100	100	21.1	47.4	290.62	64.79
N2	PT	19	100	84.2	100	31.6	36.8	457.99	141.94
N3	PN	20	90	90	90	5	0	382.32	181.16
N3	PT	19	89.5	94.7	94.7	21.1	15.8	563.54	194.05

As shown in Table 3, Q1 and Q3 consistently showed the highest correct response rate across all groups, whereas Q4 and Q5, which involved more complex content, yielded lower accuracy. In terms of JLPT level, N1 learners (advanced level) had the highest accuracy, followed by N2 (pre-advanced) and N3 (intermediate) learners, particularly on the more complex Q4 and Q5.

Comparison across media types (PN vs. PT) revealed that comprehension performance was similar between the two formats across all questions. However, in terms of watch time, learners in the PT group consistently spent more time than those in the PN group across all proficiency levels. This trend was most prominent among N3 learners (PT: M = 563.54, SD = 194.05; PN: M = 382.32, SD = 181.16), indicating that lower-proficiency learners spent more time engaging with the materials.

5.1 Logistic regression analysis: Comprehension accuracy by JLPT level and media type

Table 4 presents the results of a binary logistic regression analysis, assessing the effects of learners' JLPT level (N2 vs. N3, N1 vs. N3) and media type (PT vs. PN) on the probability of answering each question correctly.

Table 4: Results of binary logistic regression analysis

Question	Exp (B) N2 vs. N3	Sig. N2 vs. N3	Exp (B) N1 vs. N3	Sig. N1 vs. N3	Exp (B) Media	Sig. Media	Nagelkerke R^2	Accuracy
Q1	184894594.4	1.0	0.0	0.9996	1.059	0.5571	0.292	0.966
Q2	0.324	0.1394	0.98	0.1772	1.95	0.2993	0.059	0.879
Q3	131736173.2	1.0	0.0	0.9996	0.5	0.7334	0.257	0.974
Q4	2.448	0.142	8.21	0.0	0.487	0.122	0.132	0.724
Q5	8.738	0.01	34.888	0.0	1.071	0.9236	0.255	0.75

Note: N3 served as the reference group. Media comparison contrasts PT against PN.

Table 4 reveals that JLPT level had a significant effect on Q4 and Q5 performance. Specifically, learners at N1 and N2 levels were more likely to answer correctly than those at the N3 level ($p < .05$), especially on Q5. For Q1–Q3, no statistically significant differences were found by JLPT level ($p > .05$). In some cases, such as Q1 and Q3, the near-perfect accuracy among N1 and N2 learners caused perfect separation, resulting in extremely high or zero Exp(B) values, which limited practical interpretation despite high model accuracy.

Regarding media type, no significant effects were found across any of the questions ($p > .05$), indicating that the type of media—PT or PN—did not significantly influence learners' likelihood of answering correctly. Nagelkerke R^2 values ranged from 0.059 to 0.292, indicating moderate model explanatory power. Accuracy ranged from 72.4% to 97.4%, particularly high for Q1 and Q3.

5.2 Two-way ANOVA: Watch time by JLPT level and media type

To examine the effects of JLPT level and media type on participants' watch time, a two-way ANOVA was conducted. A summary of the ANOVA results is presented in Table 5.

Table 5: Two-way ANOVA: Watch time by JLPT level and media type

Source	SS	Df	F	Sig.
JLPT level	537903.89	2	14.33	0.0000
Media Type	484248.79	1	25.80	0.0000
JLPT × Media Type	117534.82	2	3.13	0.0476
Error	2064749.14	110		

As summarized in Table 5, the two-way ANOVA revealed statistically significant main effects of both JLPT level and media type on watch time ($p < .001$), as well as a significant interaction effect between JLPT level and media type ($p < .05$). This indicates that the influence of media type on watch time varies depending on the learner's JLPT level, particularly among lower-proficiency learners.

Table 6: Post-hoc Tukey HSD pairwise comparison of watch time by JLPT level

Group 1	Group 2	Mean difference	p-value
N1	N2	65.58	0.1513
N1	N3	161.88	0.0000
N2	N3	96.30	0.0189

Table 6 confirms that N3 learners (intermediate) spent considerably more time studying the materials than both N2 and N1 learners. The difference between N1 and N2 was not statistically significant. This suggests that lower-proficiency learners require more time to process multimedia content.

A two-way ANOVA was conducted to examine the effects of learners' JLPT level, media type, and their interaction on watch time. The results revealed that JLPT level had a statistically significant effect on watch time, $F(2, 110) = 14.33$, $p < .001$. This indicates that learners with different JLPT levels spent different amounts of time engaging with the learning materials. Post-hoc pairwise comparisons using Tukey's HSD (Navarro & Foxcroft, 2025) showed a slight difference between the N1 and N2 groups ($p = .1513$), whereas more significant differences were observed between N1 and N3, and between N2 and N3. Specifically, learners in the N3 (intermediate) group spent more time viewing the materials than those in the N1 (advanced) group. The data

suggest that learners with lower proficiency levels (N3) tend to spend the most time studying the multimedia content, followed by those at the N2 and N1 levels. This pattern may reflect differences in processing demands across proficiency levels.

At the same time, media also had a statistically significant effect on watch time, $F(1, 110) = 25.80$, $p < .001$, indicating that the type of media influenced the duration of study. The mean comparison revealed that learners using the PT (Picture + Text) format spent more time viewing the content than those using the PN (Picture + Narration) format. This result may suggest that the PT format is associated with longer processing time than the PN format.

Table 7: Simple main effects analysis for watch time by media type and JLPT level

Comparison	F	p-value
Media type at N3	9.10	0.0046
Media type at N2	21.86	0.00004
Media type at N1	1.78	0.1899
JLPT level at PN	3.51	0.0367
JLPT level at PT	12.66	0.00003

As presented in Table 7, media type (PT vs. PN) had a statistically significant effect on watch time among learners in the N2 group ($p < .001$) and the N3 group ($p = .005$), but not in the N1 group ($p = .190$). Meanwhile, learners' JLPT level had a substantial effect on watch time within both the PN condition ($p < .05$) and the PT condition ($p < .001$). These results confirm that the observed interaction between JLPT level and Media Type was driven by the lower-proficiency learners (N3), who spent the most time studying the material when using the PT format.

6. DISCUSSION

This study examined the effects of media type (Picture + Text [PT] vs. Picture + Narration [PN]) and learner proficiency on comprehension and viewing time. No statistically significant differences were observed in comprehension outcomes between the two media conditions across the five test items (Q1–Q5). This finding is consistent with Tabbers et al. (2004), who reported that in learner-paced environments, differences between text-based and narration-based formats may diminish when learners are able to control the pace of processing. These findings suggest that modality effects may not consistently emerge under learner-paced conditions in the present context.

This result can be interpreted through the Cognitive Theory of Multimedia Learning (Mayer, 2009). The modality principle predicts an advantage for narration over on-screen text under conditions of high cognitive load, especially when learners are unable to control the pace of presentation. However, in learner-paced environments, this advantage may be reduced, as learners are able to allocate additional time to process information as needed. In such contexts, differences between media formats may become less pronounced, not because modality is irrelevant, but because learners can compensate for processing demands through self-regulation.

A more detailed analysis of the five test items provides further insight into these findings. Items Q1–Q3 showed consistently high accuracy across all groups, suggesting limited discriminative power and a possible ceiling effect. These items involved relatively simple and concrete content, which could be processed without substantial difficulty regardless of media type or proficiency level. In contrast, Q4 and Q5 were more demanding and showed clearer differences across proficiency levels. These items required the integration of multiple elements and involved more abstract and culturally specific concepts. The observed differences across proficiency levels were therefore most evident in these more complex items, suggesting that task difficulty, rather than media modality, played a central role in shaping learner performance in the present study.

A closer examination of item characteristics supports this interpretation. Although Q2 and Q4 were both classified at the understanding level in Bloom's revised taxonomy (Anderson & Krathwohl, 2001), they differed in their conceptual demands. Q2 required learners to infer intentions from a sequence of observable events, whereas Q4 required the interpretation of more abstract cultural and religious concepts. This suggests that classification within the same cognitive category does not necessarily imply equivalent levels of processing difficulty. In the present study, the lower performance observed in Q4 and Q5 may therefore be related not only to cognitive level, but also to the lexical and conceptual characteristics of the materials.

Although the descriptive results in Table 3 show a general tendency for the PT condition to be associated with longer processing time and lower accuracy across several cells, these differences were not

statistically significant. Therefore, the findings do not provide sufficient evidence to conclude that media type had a consistent effect on comprehension outcomes. Notably, no statistically significant differences were observed in comprehension outcomes between the PT and PN conditions, although differences in processing time were evident. Learners, particularly those at a lower proficiency level (N3), spent more time engaging with PT materials than PN materials. This pattern is consistent with the findings of Tabbers et al. (2001), who reported that, under user-paced conditions, learners in the text-based condition spent more time engaging with the material without demonstrating corresponding gains in learning outcomes. Taken together, these findings suggest that increased processing time does not necessarily translate into improved comprehension but may instead reflect greater processing demands associated with text-based input.

One possible explanation for this pattern can be found in Cognitive Load Theory (Sweller, 1994; Sweller et al., 2019). Learners with less developed schemas may require more time to process incoming information, particularly when tasks involve integrating multiple elements. In the present study, lower-proficiency learners spent more time engaging with the materials but were not significantly more likely to answer the questions correctly. This suggests that increased processing time alone may not be sufficient to support comprehension when cognitive demands remain high.

However, several factors should be considered when interpreting these findings. The present study did not directly measure variables such as reading speed or learners' familiarity with text-based or subtitle-like materials. Therefore, it is not possible to determine the extent to which these factors contributed to the observed differences in processing time. At the same time, previous studies suggest that learner-related variables and instructional conditions may influence whether modality effects emerge (Harskamp et al., 2007; Tabbers et al., 2001). It is also possible that differences in lexical and conceptual difficulty across items contributed to the observed patterns.

From a multimodal perspective, these findings may also be interpreted in terms of how learners construct meaning through the interaction of multiple semiotic modes. As Kress (2010) suggests, meaning is not derived from a single mode but emerges through the coordination of linguistic, visual, and auditory resources. In the present study, the absence of statistically significant differences in comprehension between the PT and PN conditions may reflect the possibility that learners were able to integrate available modes under learner-paced conditions, although this was not directly measured. The differences observed in processing time also suggest that each mode may impose different processing demands, even when overall comprehension outcomes remain similar.

In summary, the results indicate that learner proficiency is a key factor influencing both comprehension and learning behavior. While media type did not produce statistically significant differences in comprehension outcomes under the present conditions, differences in processing time suggest that the two formats may engage learners differently. These findings highlight the importance of considering the interaction between learner characteristics, task complexity, and instructional design, rather than focusing solely on media modality.

7. IMPLICATIONS

The findings of this study offer several implications for the design and use of multimedia materials in second-language learning contexts.

The results indicate that the effectiveness of a media format may depend on both learner proficiency and task demands. Although no significant differences in comprehension were observed between Picture + Narration (PN) and Picture + Text (PT) under learner-paced conditions, lower-proficiency learners tended to spend more time processing text-based materials. This pattern suggests that PT materials may impose greater processing demands, particularly for learners at lower proficiency levels. In situations where the content is conceptually complex, PN formats may therefore be more suitable, as they reduce the need for simultaneous reading and visual processing.

When PT materials are used with lower-proficiency learners, additional instructional support becomes important. For example, text can be segmented into shorter units, linguistic structures can be simplified, and key information can be highlighted to guide attention. Providing opportunities for repeated exposure may also help learners process the material more efficiently. These forms of scaffolding can reduce processing demands and support comprehension without fundamentally changing the content.

The findings also highlight the importance of aligning task difficulty with learner proficiency. Materials that involve concrete and familiar content appear to be accessible across proficiency levels, whereas tasks requiring the interpretation of abstract or culturally specific concepts place greater demands on learners. Sequencing instructional materials from concrete to more abstract content, and adjusting the complexity of comprehension tasks accordingly, may therefore support more effective learning.

Finally, the results suggest that different types of comprehension questions place varying demands on learners. Lower-level learners may benefit from a greater proportion of literal comprehension questions, while higher-level learners may be better able to engage with inferential and analytical tasks. Designing assessments that reflect this progression may provide a more accurate representation of learner ability.

From a broader theoretical perspective, the results suggest that the effectiveness of different media types may not be fully explained by cognitive load alone. While the Cognitive Theory of Multimedia Learning provides a useful framework for understanding processing demands, the results also indicate that learners may draw on multiple semiotic resources when interpreting multimedia input. This supports a more integrative view of multimedia learning, in which meaning is constructed through the coordination of visual, auditory, and textual modes. Accordingly, the findings highlight the importance of considering both cognitive and multimodal perspectives in the design and evaluation of multimedia learning materials.

8. LIMITATIONS

Several limitations of the present study should be acknowledged. The number of test items was limited to five, which may restrict the reliability and generalizability of the findings. Although item-level analysis provided detailed insights into performance patterns, a larger number of items would allow for more stable measurement across different levels of difficulty. In addition, the sample size within each subgroup was relatively small. While participants were distributed across proficiency levels (N1–N3) and media conditions (PN vs. PT), the number of participants in each group may limit the statistical power to detect subtle effects, particularly interaction effects. This constraint arose primarily from the eligibility criteria, which required participants to hold a valid JLPT certificate obtained within the preceding two years. Future research should aim to include a minimum of 30 participants per subgroup in order to support more stable estimates and increase sensitivity to smaller effect sizes.

Another limitation concerns the measurement of cognitive processes. Although the study examined both comprehension outcomes and processing time, it did not directly assess variables such as cognitive load, reading speed, or learners' familiarity with text-based or subtitle-like materials. As a result, the interpretations are based on observable performance patterns rather than direct evidence of cognitive processing.

Finally, variation in item difficulty may have influenced the results. As observed in the analysis, some items (Q1–Q3) were relatively easy, while others (Q4–Q5) required more complex processing. In addition, the presence of more abstract or less familiar lexical and conceptual elements in certain items may have contributed to differences in performance across proficiency levels. Future research may benefit from more systematic control of item difficulty and lexical distribution across test materials.

9. CONCLUSION

This study examined the effects of media type on multimedia learning across different proficiency levels. The findings revealed no statistically significant differences in comprehension were observed between Picture + Narration and Picture + Text conditions under learner-paced conditions, although differences in processing time were identified, particularly among lower-proficiency learners. These results suggest that the effectiveness of multimedia design may depend on the interaction between learner proficiency, task demands, and processing conditions. Rather than supporting a single optimal modality, the findings point to a more context-dependent view of multimedia learning.

From a theoretical perspective, the study contributes to a more integrated understanding of multimedia learning by highlighting the role of both cognitive processing and multimodal meaning-making. Overall, the findings provide a clearer picture of how learners engage with different media formats in second-language learning contexts.

DECLARATION

1. Conflict of interest

The author declares no conflict of interest.

2. Generative AI and AI-assisted technologies in the writing process

During the preparation and revision of this manuscript, the author used ChatGPT (OpenAI, GPT-5) to identify potential grammatical and typographical errors, omitted words, and punctuation issues, as well as to improve sentence flow. The author carefully reviewed and revised all AI-assisted output and

takes full responsibility for the final content. Generative AI was not used for idea generation, research data generation, statistical analysis, or production of the study findings.

3. Data availability statement
The data supporting the findings of this study are available from the corresponding author upon reasonable request.

4. Ethics statement
This study was approved by the Ethics Committee in Human Research of the National Institute of Development Administration (NIDA) on June 26, 2025 (COA No. 2025/0092; Protocol ID No. ECNIDA 2025/0107). All participants were informed of the purpose of the study and provided informed consent before participating. Participation was voluntary, and all responses were kept confidential and used solely for research purposes.

5. Funding
This research received no external funding.

6. Contributor Role Taxonomy (CRediT)
Tewich Sawetaiyaram: Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing.

ORCID

Tewich Sawetaiyaram: https://orcid.org/0009-0004-6022-4900

REFERENCES

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology, 94*(2), 416–427. https://doi.org/10.1037/0022-0663.94.2.416

Boonphak, K. (2020). Review article on learning management in the new normal. *Journal of Industrial Education, 19*(2), A1–A6. [in Thai]

Chen, Z. (2006). *The effects of multimedia annotations on L2 vocabulary immediate recall and reading comprehension: A comparative study of text-picture and audio-picture annotations under incidental and intentional learning conditions* [Doctoral dissertation, University of South Florida]. USF Tampa Graduate Theses and Dissertations. https://digitalcommons.usf.edu/etd/2478/

Craig, S. D., Gholson, B., & Driscoll, D. M. (2002). Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundancy. *Journal of Educational Psychology, 94*(2), 428–434. https://doi.org/10.1037/0022-0663.94.2.428

Hanrattanasakul, N. (2017). *Second language comprehension in multimedia learning material: A case study of redundancy principle and modality principle in Japanese multimedia material* [Unpublished master's thesis]. National Institute of Development Administration. [in Thai]

Harskamp, E. G., Mayer, R. E., & Suhre, C. (2007). Does the modality principle for multimedia learning apply to science classrooms? *Learning and Instruction, 17*(5), 465–477. https://doi.org/10.1016/j.learninstruc.2007.09.010

Jeung, H.-J., Chandler, P., & Sweller, J. (1997). The role of visual indicators in dual sensory mode instruction. *Educational Psychology, 17*(3), 329–345. https://doi.org/10.1080/0144341970170307

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*(4), 351–371. https://doi.org/10.1002/(SICI)1099-0720(199908)13:4%3C351::AID-ACP589%3E3.0.CO;2-6

Kalyuga, S., Chandler, P., & Sweller, J. (2000). Incorporating learner experience into the design of multimedia instruction. *Journal of Educational Psychology, 92*(1), 126–136. https://doi.org/10.1037/0022-0663.92.1.126

Karbalaei, A., Sattari, A., & Nezami, Z. (2016). A comparison of the effect of text-picture and audio-picture annotations in second language vocabulary recall among Iranian EFL learners. *GIST – Education and Learning Research Journal, 12*, 51–71. https://doi.org/10.26817/16925777.244

Kawamura, Y., Kitamura, T., & Hobara, R. (2013). *Reading tutor: Japanese language reading tutorial system*. Tokyo International University. https://chuta.cegloc.tsukuba.ac.jp/index_e.html

Kress, G. (2010). *Multimodality: A social semiotic approach to contemporary communication*. Routledge.

Kress, G., & van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Arnold.

Leahy, W., Chandler, P., & Sweller, J. (2003). When auditory presentations should and should not be a component of multimedia instruction. *Applied Cognitive Psychology, 17*(4), 401–418. https://doi.org/10.1002/acp.877

Lee, J.-H., & Hasebe, Y. (2019). Readability measurement of Japanese texts based on levelled corpora. In A. Bekeš & I. Srdanović (Eds.), *The Japanese language from an empirical perspective: Corpus-based studies and studies on discourse* (pp. 143–168). Ljubljana University Press. https://doi.org/10.4312/9789610602170

Lee, J.-H., Sunakawa, Y., Hori, K., & Hasebe, Y. (n.d.). *jReadability: A Japanese text readability measurement system*. jReadability Portal. https://jreadability.net/

Malithong, K. (2000). *Theknoloyi kan sueksa lae nawattakam* [Educational technology and innovation]. Arun Karnphim. [in Thai]

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.

Mayer, R. E., Dow, G. T., & Mayer, S. (2003). Multimedia learning in an interactive self-explaining environment: What works in the design of agent-based microworlds? *Journal of Educational Psychology, 95*(4), 806–812. https://doi.org/10.1037/0022-0663.95.4.806

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology, 90*(2), 312–320. https://doi.org/10.1037/0022-0663.90.2.312

Moreno, R., & Mayer, R. E. (1999). Multimedia-supported metaphors for meaning making in mathematics. *Cognition and Instruction, 17*(3), 215–248. https://doi.org/10.1207/S1532690XCI1703_1

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94*(1), 156–163. https://doi.org/10.1037/0022-0663.94.1.156

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction, 19*(2), 177–213. https://doi.org/10.1207/S1532690XCI1902_02

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology, 87*(2), 319–334. https://doi.org/10.1037/0022-0663.87.2.319

Navarro, D., & Foxcroft, D. (2025). *Learning statistics with jamovi: A tutorial for beginners in statistical analysis*. Open Book Publishers. https://doi.org/10.11647/OBP.0333

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Tabbers, H. K., Martens, R. L., & van Merriënboer, J. J. G. (2001). The modality effect in multimedia instructions. *Proceedings of Annual Meeting of the Cognitive Science Society, 23*, 1024–1029. https://escholarship.org/uc/item/8fd0d9x5

Tabbers, H. K., Martens, R. L., & van Merriënboer, J. J. G. (2004). Multimedia instructions and cognitive load theory: Effects of modality and cueing. *British Journal of Educational Psychology, 74*(1), 71–81. https://doi.org/10.1348/000709904322848824

Tindall-Ford, S., Chandler, P., & Sweller, J. (1997). When two sensory modes are better than one. *Journal of Experimental Psychology: Applied, 3*(4), 257–287. https://doi.org/10.1037/1076-898X.3.4.257

APPENDIX

Appendix A: Original Japanese script of the multimedia materials

Scene	Content	Length (characters)
1	みなさん、こんにちは。協力してくれてありがとうございます。準備ができたら、「次へ」のボタンを押してください。	54
2	みなさんは温泉に行ったり、温泉の湯に入ったりしたことがありますか。今日はその「温泉」についていろいろな話を、みなさんにご紹介したいと思います。	71
3	日本と言えば、真っ先に思い浮かぶのは「温泉」だと思います。日本は世界の「温泉大国」で、日本人は昔も今も、温泉に浸かることが大好きなようです。もちろん、温泉に入ることが好きな外国人もたくさんいるでしょう。温泉に入るために、わざわざ日本に行く人もいるようです。	128
4	温泉が好きなのは我々人間だけではなく、猿も温かい湯に浸かることが好きなようです。日本には猿専用の温泉があります。	56
5	猿専用の温泉は長野県の地獄谷野猿公苑にあります。１１月から３月あたりに多くの猿が温泉に入りにきています。	52
6	その理由は、厳しい寒さと関係しているようです。耐え難いその大変さをしのぐために、入るとみられます。夏には温泉に入らないですから。	64
7	でも、なぜ「猿の温泉」が作られたのでしょうか。きっかけは、一匹の子猿が近くの旅館の露天風呂に入ったことらしいです。それ以来、他の猿も真似まねするようになったので、猿のための温泉を作ることになったと言われています。	104
8	さて、動物の話はさておき、次は日本人が温泉が大好きな理由を見ていきましょう。	38
9	日本人の温泉好きの理由は、宗教からの影響だと思われています。昔の日本は「神道」という宗教の「禊」という儀式がありました。	60
10	「禊」というのは、水を浴びて体と心を清浄する儀式です。そのころ、重要な儀式に参加する際には、禊の儀式で心身を清めなければなりませんでした。	69
11	禊を行うには、冷たい川や滝、海の水などで心身を清浄しなければなりません。冬なら、大変そうですね。	48
12	そして、８世紀の中頃、日本に仏教が入ってきました。そのとき、仏教の「温室経」というお経も同時に入ってきました。	55
13	「温室経」を簡単に説明すると「湯に入ると功徳が得られる」ということです。つまり、温泉に入る行為は宗教的に大変いい行いだということです。	67
14	当時の日本では、冷たい水で心身を清める禊の儀式があったため、「温かい湯でも大丈夫」と教える「温室経」は、すごく歓迎されていたらしいです。浸かるなら、温かい湯のほうがいいですよね。	89
15	温泉学者の「松田忠徳」は、日本人の「温泉好き」という習慣は「禊」と「温室経」の融合で生まれたのではないかと推測しました。つまり、日本人の温泉好きは二つの宗教の融合で生まれたというわけです。	94

Appendix B: English translation of the multimedia materials

Scene	Content	Length (characters)
1	Hello everyone. Thank you very much for your cooperation. When you are ready, please press the “Next” button.	54
2	Have you ever been to a hot spring or bathed in hot spring water? Today, I would like to introduce various aspects of these “onsen” (hot springs) to you.	71
3	When one thinks of Japan, one of the first things that comes to mind is “onsen.” Japan is known as a “hot spring nation” of the world, and it seems that Japanese people, both in the past and in the present, have greatly enjoyed soaking in hot springs. Of course, there are also many foreigners who enjoy hot springs. Some people even travel all the way to Japan specifically to bathe in them.	128
4	It is not only humans who enjoy hot springs; monkeys also seem to like soaking in warm water. In Japan, there are hot springs exclusively for monkeys.	56
5	These monkey hot springs are located in Jigokudani Monkey Park in Nagano Prefecture. Many monkeys come to bathe there from around November to March.	52
6	This appears to be related to the severe cold. It is believed that they enter the hot springs to endure such harsh conditions. After all, they do not bathe in hot springs during the summer.	64
7	But why were “monkey hot springs” created in the first place? It is said that the origin lies in a young monkey that entered an open-air bath at a nearby inn. Since then, other monkeys began to imitate this behavior, which eventually led to the creation of hot springs specifically for them.	104
8	Now, setting aside the topic of animals, let us next consider why Japanese people are so fond of hot springs.	38
9	It is believed that their fondness is influenced by religion. In ancient Japan, there was a ritual called “misogi” in the religion of Shinto.	60

Appendix B: English translation of the multimedia materials (continued)

Scene	Content	Length (characters)
10	"Misogi" is a ritual in which one purifies both body and mind by bathing in water. At that time, before participating in important rituals, it was necessary to cleanse oneself through misogi.	69
11	To perform misogi, one had to purify the body and mind using cold water such as rivers, waterfalls, or the sea. In winter, this must have been quite difficult.	48
12	Then, in the mid-8th century, Buddhism was introduced to Japan. At that time, a Buddhist scripture called the "UnjitsuGyō" (Sutra of the Bathhouse) was also introduced.	55
13	Simply explained, this sutra teaches that "one gains merit by bathing in hot water." In other words, bathing in hot springs was considered a highly virtuous act in religious terms.	67
14	In Japan at that time, where the misogi ritual required purification with cold water, the teaching of the UnjitsuGyō —that warm water was also acceptable—was apparently very well received. After all, if one is to bathe, warm water is preferable.	89
15	Hot spring scholar Tadanori Matsuda has suggested that the Japanese custom of "fondness for hot springs" may have emerged from the fusion of "misogi" and the "UnjitsuGyō". In other words, the Japanese love of hot springs can be understood as having developed from the blending of these two religious traditions.	94