國立臺灣師範大學文學院英語學系

碩士論文

Department of English, College of Liberal Arts

National Taiwan Normal University

Master's Thesis

# 呈現不同視覺媒材對英語學習者多模態聽力理解與認知負荷之影響

# Assessing Multimodal Listening: Efficacy of Different Visual Formats on L2 Learners' Listening Comprehension and Cognitive Load

劉紹威

Liu, Shao-Wei

指導教授：劉宇挺 博士

Advisor：Liu, Yeu-Ting, Ph.D.

中華民國 113 年 7 月

July 2024

**ACKNOWLEDGEMENTS**

I would like to start by expressing my external gratitude to my advisor, Professor Yeu-Ting Liu, for guiding me with patience and kindness, for believing my potential when I doubted myself, and for motivating me pursuit higher achievements throughout the process of composing this thesis and the work preceding it. His valuable feedback and insightful ideas are pivotal for the completion of my work. His enthusiasm for teaching, research, and administrative work has always been a profound influence on my journey as a graduate student at NTNU. I am perpetually grateful for his devotion to my thesis—they are invaluable to me.

My sincere gratitude is extended to my committee members, Professor Mei-Zhen Wu and Professor Mae-Han Chen, for offering suggestions and for attending my oral defense. I particularly appreciate for their feedback on my discussion and pedagogical implication, which profoundly enriched this thesis. The validation of my work through review and recognition from these respected researchers is immensely rewarding

I am also deeply grateful to all the students who participated in the experiment. It was their support and efforts that made the thesis possible. In addition to the students, thanks must go to their English teacher and the Director of Academic Affairs, who gave all the necessary support for recruiting students and conducting the experiment in school.

Finally, my appreciation goes to my parents for giving me love and support for pursuing my master’s degree at NTNU, and for their trust in my potential to become a better English teacher after completing the master’s degree. This journey would not be fulfilled without their assistance.

## 摘要

視覺媒材與第二語言聽力表現的關係，長期以來備受關注。在眾多視覺媒材之中，以影片做為聽力測驗媒材最為普遍。然而，過往的研究結果呈現分歧。為深入瞭解分歧的起因，探究視覺媒材與第二語言學習者語言程度之間的複雜關係尤為重要。本研究旨在探討不同第二語言程度(高或低)的學習者，在接觸不同視覺媒材(動態影片或連續靜態影像)後，其聽力表現是否呈現差異。為豐富研究成果，本研究同時測量學習者的主觀認知負荷(內在、外在與增生)以及對視覺媒材的整體感受。研究結果顯示，在68名學習者中，學習者的第二語言程度與所接觸的視覺媒材皆對聽力表現產生直接影響。其中第二語言程度的影響尤為顯著。高成就學習者在不同視覺媒材條件下的聽力表現皆顯著高於低成就學習者。進一步分析，縱使低成就學習者的聽力表現顯著低於高成就學習者，低成就學習者仍對於視覺媒材展現正向態度且仔細留意不同非語言訊息之意涵。就視覺媒材而言，動態影片相較於連續靜態影像更能顯著提升聽力表現。此差異可能源於動態影片提供較細膩的非語言訊息且引發較低的外在認知負荷與較高的增生認知負荷所致。此外，結果亦可推論動態影片為較合適的多模聽力視覺素材。本研究結果再次呼應第二語言學習者之語言程度與所接觸之視覺媒材對於聽力表現的影響。同時，本研究為聽力教學與測驗中，多模態資源的運用提出建議，希冀透過教學與測驗方法，培養學習者面對真實世界的聽力技巧與策略。


**關鍵字：** 多模態聽力、視覺模態、第二語言能力、認知負荷

# ABSTRACT

The interaction between visual components and L2 listening performance has garnered considerable attention, with video as a prevalent assessment tool, albeit with discrepant results. The inconsistency necessitates a nuanced examination of the intricacies between diverse visual formats and learners' L2 proficiency. In light of the existing findings and assumptions, the present study investigated the interplay of L2 proficiency levels (i.e., high vs. low) and contrasting visual formats (i.e., dynamic video vs. sequential static imagery). The study also gauged learners' perceived cognitive load at different proficient levels while encountering various visual formats to further enrich the analysis. The manipulation of L2 proficiency and visual formats resulted in four listening conditions (i.e., high-proficient + dynamic video, low-proficient + dynamic video, high-proficient + sequential static imagery, and low-proficient + sequential static imagery). A cohort of 68 participants, stratified by proficiency and visual format, were assigned to one of the four listening conditions. They then completed the listening comprehension test and questionnaires measuring their perceived cognitive load (i.e., intrinsic, extraneous, and germane) and overall perception.

Results elucidated that learners' L2 proficiency exerted a more profound impact on listening performance than visual formats. Specifically, high-proficient learners consistently outperformed their low-proficiency counterparts, irrespective of visual formats. While low-proficient learners' listening performance was inferior to that of high-proficient learners, they generally reported positive attitudes toward visual inclusion, and paid close attention to various nonverbal information. Visual formats, on the other hand, demonstrated a comparatively limited impact on learners' listening performance; they nonetheless exhibited statistically significant effects, with dynamic video yielding superior outcomes compared to sequential static imagery. The enhanced performance was derived from the detailed nonverbal information in dynamic video, the low extraneous cognitive load, and the high germane cognitive load. Additionally, dynamic video might be an optimal visual material for eliciting learners' multimodal listening ability.

The present study underscores the role of learners' L2 proficiency and visual formats in shaping listening performance within a multimodal listening environment. The study's implications are far-reaching for L2 listening instruction and assessment. By judiciously considering the role of visual components and learners' proficiency levels, instructors can apply effective strategies to cultivate learners' listening skills in real-world, multimodal contexts.

**Keywords:** Multimodal listening, Visual mode, L2 proficiency, Cognitive load

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER ONE:

# INTRODUCTION

The advancement of technology has increased the accessibility and availability of multimodal resources, such as video in foreign language (L2) classrooms, particularly in teaching L2 listening. Given the ubiquity of videos for pedagogical purposes, existing assessments have yet to fully manifest the approach to measuring L2 learners' listening ability. Typically, L2 listening ability has been measured exclusively through audio-formatted assessments (Kang, Arvizu, Chaipuapae & Lesnov, 2019). These assessments might fail to capture most real-world interpersonal communication in which listeners could see the speakers (Jewitt, 2013; Wagner, 2007, 2010a). Unable to adequately simulate the real-world multimodal discourse (Jewitt, 2013), the assessment might lead to construct underrepresentation and threaten test validity (Wagner, 2008; 2010a; Ockey & Wagner, 2018). However, with the inclusion of visuals, listening assessment could align with most target language use domains (TLU) (Bachman & Palmer, 1996), enhance authenticity (Aryadoust, Foo & Ng, 2022; Ockey, 2007), and avoid construct underrepresentation (Lesnov, 2022; Messick, 1996; Wagner, 2010a ). In light of the significance of visuals in listening assessment, a plethora of studies have explored the effectiveness of visuals in listening assessment (Batty, 2014, 2020; Brett, 1997; Coniam, 2001; Ginther, 2001; Gruba, 1993; Ockey, 2007; Shin, 1998; Sueyoshi & Hardison, 2005; Suvorov, 2009; 2015; Wagner, 2010a, 2010b, 2013).

Existing studies have predominantly emphasized the audio-visual distinction to validate the effect of visuals, albeit with mixed results. The conflicting results have been categorized into three strands: visuals increased performance, visuals made no difference, and visuals debilitated performance (Ockey & Wagner, 2018). Studies in the first strand (Baltova, 1994; Parry & Meredith, 1984; Shin, 1998; Sueyoshi &

Hardison, 2005; Wagner, 2010a, 2013), which suggested videos lead to increased performance, have jointly indicated that nonverbal information (e.g. lip movement, gesture, or photographs) could enhance comprehension of the aural inputs and compensate for learners' inadequate linguistic knowledge (Baltova, 1994; Sueyoshi & Hardison, 2005; Wagner, 2010a). Although initial research within the first strand yielded positive findings regarding the benefits of video, the second strand found no discernible influence (Batty, 2014; Coniam, 2001; Gruba, 1993; Londe, 2009). The inherent limitations of spoken discourse in specific video formats, for instance, talk shows in which information was primarily conveyed through speech, could restrict the amount of contextual information available to the video group (Coniam, 2001). Given the lack of informative video content, learners in video groups may not experience a significant advantage, leading to performance comparable to that of the audio-only group. Findings from the first two strands suggested that including sufficient nonverbal cues within videos that corresponded to the auditory information may positively influence learners' listening. For the last strand, which indicated the adverse effects (Pusey & Lenz, 2014; Suvorov, 2009), several potential causes have been identified. The impact of video may vary depending on the audio content. For instance, video accompaniment might hinder comprehension during lectures, but it could benefit dialogues. (Suvorov, 2009).

In light of the mixed findings, the effectiveness of video in enhancing listening performance might hinge on the degree of congruence between nonverbal cues presented visually and auditory information. Additionally, videos could be more beneficial in certain spoken text types, such as dialogue (Suvorov, 2009). While these findings offered valuable insights, they were primarily derived from a limited framework that contrasted audio-only and video conditions. A comprehensive understanding of visual effectiveness requires examining formats beyond video, which necessitates further investigation of other arrays of visuals.

Existing research examining the influence of static visuals on listening performance employed single photographs or screenshots derived from videos (Suvorov, 2009; Ockey, 2007). Findings revealed that the initial presentation of static visuals might be beneficial, but the lack of additional information made the benefit outcome diminish over time and resulted in a low level of engagement on visuals (Ockey, 2007). In addition, learners' listening performance was incongruent with their preferred presentation mode. Interestingly, learners with a propensity for static visuals did not achieve significantly higher performance outcomes when exposed to the particular visual format they favored. The discrepancy implied that visual formats alone might not fully explain the findings (Pusey & Lenz, 2014; Suvorov, 2009). To be more specific, other potential moderating variables, including participants' language proficiency, should be addressed in future research (Liu & Aryadoust, 2023; Pusey & Lenz, 2014).

To consolidate the effect of other moderating variables, examining the interaction between input mode and L2 proficiency was of paramount importance (Suvorov & He, 2022). Learners' L2 proficiency might be a significant factor in determining the amount of benefit derived from video. For instance, high-proficient learners who possesses a rich linguistic repertoire and effective decoding skills might potentially remain unaffected by video and consistently exhibit outstanding performance in such conditions (Sueyoshi & Hardison, 2005). On the contrary, low-proficient learners seemed to evince more association and reliance on video. For example, they tend to compensate for their inadequate linguistic knowledge and facilitate meaning retrieval through video (Liu & Aryadoust, 2023; Sueyoshi &

Hardison, 2005). Similar observations can be made when considering the application of static visuals. Compared to high-proficient learners who demonstrated reasonable attentional control over static visuals, low-proficient learners found that the transition between static visuals might slightly distract them and negatively affect their performance (Ginther, 2002; Ockey, 2007). The above findings stipulated that learners with relatively high L2 proficiency might be less susceptible to different input modes. In other words, visuals might significantly influence on low-proficient learners, underlining the need to further explore how visuals might impact learners' listening comprehension across the L2 proficiency spectrum.

The cognitive demand attributed to processing multimodal inputs has been postulated in previous studies (Buck, 2001; Ockey, 2007; Suvorov, 2009), suggesting that the inclusion of visuals might increase learners' cognitive load and hinder their listening comprehension (Buck, 2001). However, an alternative interpretation could be that learners' emergent decoding abilities, rather than the presence of visuals, contribute to this phenomenon (Guichon & McLornan, 2008; Lee, Liu & Tseng, 2021). Processing multimodal inputs might be a laborious task for L2 learners, particularly for those whose decoding skills were not yet developed. When encountering visuals and auditory information simultaneously, these learners might experience an increased cognitive load. This phenomenon might compel a change in processing inputs among learners. They might restrict their attentional resources solely to auditory inputs in to gain more information (Cross, 2011; Gruba, 1993; Kamiya, 2022). Due to their emergent decoding skills, their cognitive resources might be drained by processing auditory inputs, leaving less capacity for processing visual information. Insufficient cognitive capacity might preclude their attention to visual information, depriving them of their opportunity to utilize visual inputs to mitigate their limited linguistic knowledge. Accordingly, to prevent learners from experiencing excessive cognitive strain, Mayer (2010) suggested that learners' cognitive load could be managed by selecting level-appropriate materials and presenting information clearly and concisely, minimizing the cognitive effort required for learners to process it

The incorporation of both visual and audio modes might be a better representation of real-life communication events, and could increase the authenticity and validity of the listening assessment (Ginther, 2002, Terao, 2024). Extensive research has also examined participants' listening comprehension in audio-only versus video-mediated contexts, albeit with discrepancies (Wagner, 2010a, 2013). Hitherto, limited attention has been devoted to compare the effects of diverse visual formats, such as dynamic video and static visuals (Suvorov, 2009, 2015; Ockey, 2007), particularly in exploring which visual formats could better elicit learners' real-life listening skills, and reduce their cognitive burden (Shin, 1998; Terao, 2024). To mitigate the insufficiency, it is necessary to compare the effectiveness of different visual formats on learners' listening performance and cognitive load. To optimize learners' listening comprehension with varying L2 proficiency, a comparative analysis of video and static visuals could inform the development of more effective teaching and testing formats. Concurrently, extant research suggested that visual presentation format may not be the sole determinant influencing L2 learners' diverse listening comprehension outcomes, underscoring the necessity to investigate the interplay between varied visual input and L2 proficiency levels. Though numerous studies have indicated participants' language proficiency in participants' demographic information, few have conducted a pretest to differentiate it. By manipulating learners' language proficiency in accompaniment with different input modes, researchers could scrutinize the interaction between the two variables (i.e., input mode or L2 proficiency) on learners' listening performance in the multimodal environment. Lastly, the cognitive demand attributed to the inclusion of visuals has been raised, but primarily based on speculation. To substantiate the postulate, the current study gauge learners' perceived cognitive load via implementing questionnaires.

To address the gaps, particularly on the dearth of the comparison between different visual formats, the interaction between visual formats and L2 proficiency, and the lack of evidence on test-takers perceived cognitive load in multimodal listening assessment, the present study would like to probe into two visual formats (i.e., dynamic video vs. sequential static imagery) on L2 learners with different listening abilities (i.e., high vs. low proficient). In addition, the present study would also examine test-takers perceived cognitive load (i.e., intrinsic, extraneous, and germane) and overall perception via questionnaire. Four research questions are as follows:

1. Do different visual modalities (i.e., dynamic video and sequential static imagery) affect L2 learners' listening comprehension? If so, to what extent?
2. Do different proficient L2 learners (i.e., high vs. low-proficient) perform differently in the listening assessment? If so, to what extent?
3. Do visual modalities (i.e., dynamic video and sequential static imagery) and learners' L2 proficiency (i.e., high vs. low-proficient) have an interaction effect on their listening comprehension?
4. What are the perceived cognitive load (i.e., intrinsic, extraneous, and germane) of different proficient learners (i.e., high vs. low-proficient) in different visual modalities (i.e., dynamic video and sequential static imagery)?

The findings of the present study are expected to shed light on the field of foreign language listening teaching, learning, and testing, particularly on the inclusion of multimodal resources in listening instruction and assessment. On such bases, test developers could design multimodal listening assessments that would not only avoid overburdening test-takers but also improve their motivation to comprehend the listening inputs. Instructors could design tailored listening materials that could satisfy the needs of learners at different proficiency levels and enhance their ability to decode nonverbal information within a conversation.

# CHAPTER TWO:
# LITERATURE REVIEW

## 2.1 The Role of Listening Construct and its Effects on Different Views of Listening

Developing robust listening assessment begins with a rigorous definition of the underlying listening constructs. This process entailed two sequential phases: conceptualization and operationalization (Buck, 2001). Conceptualization refers to test developers' knowledge of listening comprehension and their interpretation of the target language use (TLU) domain (Bachman & Palmer, 1996), while operationalization unfolds how test developers utilize their knowledge and experience. Two principles outlined in defining the listening construct offered a benchmark for researchers to elucidate the development of existing listening assessments.

L2 listening has been predominantly assessed via audio-only assessments (Kang et al., 2019; Suvorov & He, 2022), and the measurement was associated with how researchers have conceptualized listening comprehension. According to the early definition, listening was a sense-making process of the incoming streams of auditory information (Fries, 1945; Furness, 1952; Lado, 1964). Buck (2001) has proposed the definition, suggesting that the process might require listeners to actively derive meaning and employ their understanding of the auditory input (p.31). These early accounts have led to a widely recognized default of listening construct (Buck, 2001, p.114):

- process extended samples of realistic spoken language, automatically and in real time,
- understand the linguistic information that is unequivocally included in the text,
- make whatever inferences are unambiguously implicated by the content of the passage.

While the prevalent listening construct has been widely accepted, it has its limitations. Assessments have often been operationalized as processing aurally presented linguistic information (Lesnov, 2018). For instance, Kang et al. (2019) reviewed twenty academic listening tests, and found that listening ability was solely considered to be processing verbal acoustic information. This approach, based on existing studies, seemed to be the norm in evaluating L2 learners' listening comprehension. However, it fails to account for the diverse needs and abilities of learners, particularly those with limited linguistic knowledge and immature decoding skills.

Researchers who bolstered the concept of multimodal listening have raised concerns about the exclusive application of audio-only listening assessment. The audio-only listening assessment might be a laborious task for low-proficient L2 learners with limited linguistic knowledge and immature decoding skills (Goh, 2000; Vandergrift, 2007). They might not be able to grasp the idea of the entire verbal input since the majority of their cognitive resources have been drained (Vandergrift, 2007). When learners grappled with complex linguistic information and fast speech rates, their listening comprehension might be hindered, resulting in poor listening performance. Therefore, incorporating visual information might be a feasible approach to compensate for L2 learners' inadequate linguistic knowledge and mitigate their listening performance (Vandergrift, 2007).

Apart from the impact on the micro-level, incorporating visuals might be beneficial at the macro-level. As Wagner (2010a) suggested, in most target language use (TLU) domains, listeners could see the speakers. The concept has been elaborated by Jewitt (2013), indicating that real-life communication is a multimodal discourse, including both verbal (e.g., spoken text) and nonverbal inputs (e.g., gestures, facial expressions). If test developers aim to make inferences on learners' listening ability outside the assessment (Bachman & Palmer, 1996), the listening assessment should simulate the TLU domain. To reach this objective, researchers should design multimodal listening assessments that incorporate both verbal and nonverbal information. With the inclusion of visuals, listening assessment could also avoid construct-underrepresentation (Messick, 1996) and improve authenticity (Ockey, 2007). Accordingly, researchers urged for expanding existing listening constructs (i.e., audio-only) or for a new definition to be enacted (Batty, 2014; Ockey, 2007).

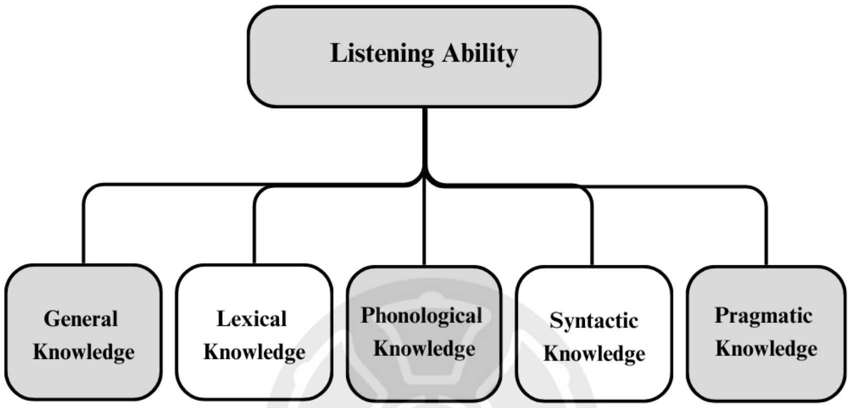
## 2.2 Theoretical Accounts of Incorporating Visual Inputs in Listening Assessment

Regarding to the construct of multimodal listening assessment, it is vital to understand how it has been conceptualized. Multimodal listening has been defined as "the process of receiving, and processing information through two or more communication modes, constructing meaning from co-present modes" (Campoy-Cubillo, 2019). By definition, listening involved multiple communication modes, and speech or aural input was only considered as a part of the entire communication event. Based on the premise of Campoy-Cubilo's (2019) definition, the following section will review Rost's listening assessment model (2016) to manifest the concept of multimodal listening further.

Rost's model for listening assessment (2016) comprised six testable components of listening abilities: general knowledge, interactive pragmatic knowledge, pragmatic knowledge, syntactic knowledge, lexical knowledge, and phonological knowledge (See Figure 1). As observed in the model, these components could be further categorized into linguistic (e.g., lexical, syntactic knowledge) and non-linguistic (e.g., general knowledge) knowledge. Linguistic knowledge was the building block of listening comprehension (Anderson, 2000; Goh & Vandergrift, 2021), while non-linguistic knowledge holistically shaped learners' comprehension. Both types of knowledge are essential to learners' listening comprehension and can be directly or indirectly examined via the multimodal listening assessment.

**Figure 1**

*Rost Listening Assessment Model*



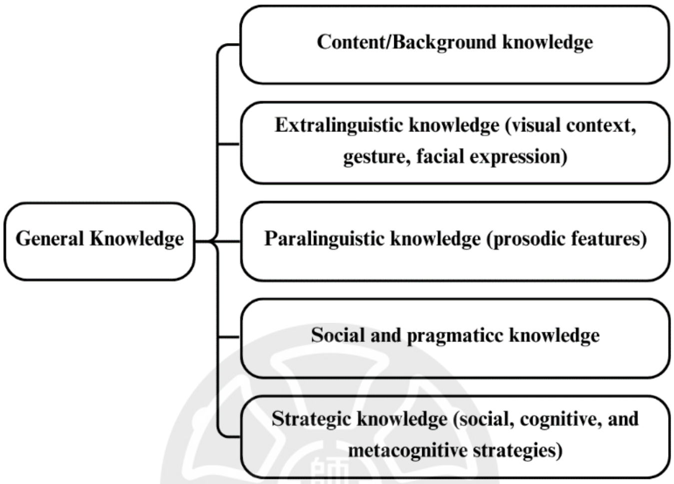

As the comprehension of linguistic information was generally considered a "cognitive process" (Field, 2008; Goh & Vandergrift, 2021), researchers tend to adopt the three stages of cognitive processing: decoding, parsing, and utilization (Anderson, 2000) in accompaniment with the two-way processing (i.e., bottom-up and top-down) (Goh & Vandergrift, 2021) to contextualize the internal mechanism of listening and how it has been examined indirectly. Typically, linguistic knowledge was delivered as verbal information and received by an aural channel, and the very first information received by listeners was a stream of sounds. The process of receiving and segmenting sound streams into words was called decoding. In this stage, learners' lexical and phonological knowledge would be measured while recognizing individual sound segments. These identified words might be further separated based on the syntactic structure in the next stage, parsing. Parsing requires learners' lexical, phonological, and syntactic knowledge to separate meaningful units and transmit them into mental representation. The aforementioned two stages constituted the bottom-up processing; in these stages, several linguistic knowledge would be evaluated indirectly.

Lastly, learners might draw on their mental representation to interpret the information in the utilization stage. The mental representation might be associated with their prior knowledge, which is encoded in their long-term memory. In order to form their mental representation and make sense of the linguistic information, learners need to utilize their general and pragmatic knowledge. General knowledge refers to the knowledge of the world, including the commonly discussed topics in daily communication and the way people interact with each other. It might be facilitative when dealing with information that was explicitly stated. For some implicit messages, pragmatic knowledge played a role. For example, given that the relationship between two speakers was not told, learners could still make inferences based on their understanding of the relationship between interlocutors within a specific context (Rost, 2016).

The cognitive process of listening has elucidated the majority of linguistic knowledge that would be measured based on Rost’s listening assessment model, albeit with little description of general knowledge. General knowledge (See Figure 2) distinguished Rost’s model from existing models. It comprised several distinct subsets: content and background, strategic, paralinguistic, and most importantly, extra-linguistic knowledge. The extra-linguistic knowledge, including visual context and other nonverbal information (e.g., gestures and facial expressions), has rationalized incorporating visuals in listening assessment. The incorporation of visuals might activate learners’ schemata, which might also facilitate them to exhibit top-down processing. As mentioned earlier, low-proficient learners might have problems decoding linguistic information using bottom-up processing. The inclusion of visuals might allow them to draw on their mental image to compensate for their limited linguistic knowledge. With such inclusion, learners' general knowledge could be examined simultaneously.

**Figure 2**

*The Subsets of General Knowledge*




Since extra-linguistic knowledge was integral to the listening model, incorporating visuals might not be recognized as construct irrelevant variables. Without validity concerns, test developers could incorporate visuals and other nonverbal information to mitigate learners' inadequate linguistic knowledge, measure their general knowledge, and enhance their listening comprehension. In addition, the inclusion of visual elements in listening assessments might facilitate more nuanced interpretations of test-takers performance within and outside the assessment settings.

Rost's listening assessment model has implied that some discrete listening abilities were the prerequisite for reaching listening comprehension, but learners' test performance might not directly observe them. Likewise, learners' problems in listening, particularly within the internal cognitive mechanism, might be difficult to determine and solve. Therefore, with the inclusion of visuals, researchers might be allowed to observe whether visual inputs could partially alleviate their problems regarding decoding and limited linguistic knowledge and if possible, could improve their listening comprehension.

### 2.3 The Effect of Visuals on L2 Learners' Listening Performance

Rost (2016) indicated that the medium of inputs (i.e., video or audio-only) might affect learners' listening performance. The assumption has been corroborated by a number of previous studies (Batty, 2014, 2021; Brett, 1997; Coniam, 2001; Cubillo & Julián, 2015; Lynch, 2011; Ockey, 2007; Suvorov, 2009, 2015; Wagner, 2008, 2010a, 2013), though the effects were inconsistent. In the following sections, the effects of two visual formats (i.e., dynamic video vs. static imagery) on learners' listening performance will be reviewed.

#### 2.3.1 Dynamic visuals as the input mode in listening assessment

Video constituted a prevalent input modality in the context of multimodal listening assessment. In a recent review by Suvorov and He (2022), approximately 30 studies emphasized the effect of video on listening performance, albeit with mixed results (Ockey & Wagner, 2018). The heterogeneous outcomes predominantly emerged from comparative analysis of L2 learners' listening performance between audio-only and video-mediated cohorts. These results can be further stratified into three categories: visual enhanced performance, video made no discernible impact, and video impaired performance. (Ockey & Wagner, 2018).

Existing studies that suggested videos lead to increased listening performance have stressed that the contextual information and nonverbal stimuli embedded in videos have made the content more comprehensible (Parry & Meredith, 1984; Shin, 1988; Wagner, 2010a, 2013). As Parry and Meredith (1984) reported, the redundant effects on nonverbal stimuli and aural information facilitated learners' sense-making process. Like Parry and Meredith (1984), Shin (1998) also found that video groups seemed to gain more contextual information than audio groups, yielding significantly better performance than audio-only groups. More recently, Wagner (2010a) specifically made a set of videos that integrated with pictures representing the subjects in the videos. The posttest performance revealed that video groups scored 6.5% higher in score than audio-only groups. Similar results have been found in his later study, in which the same set of videos were applied (Wagner, 2013). In line with Parry and Meredith (1984) and Shin (1988), Wagner further indicated the benefits of nonverbal cues. To put forward the impacts of nonverbal information, he probed into items in which video groups outperformed audio groups. Interestingly, he found that the congruence between verbal and nonverbal information might be critical to learners' outstanding performance within these items. In this line of research in which visuals might lead to increased performance, scholars have jointly maintained the benefits of nonverbal information to listening performance, particularly when nonverbal information is congruent with verbal information.

Unlike the aforementioned studies (Parry & Meredith, 1984; Shin, 1998; Wagner, 2010a, 2013), which explored the benefits of videos on listening performance, another line of research has asserted that videos made no difference in learners' listening performance (Baltova, 1994; Brett, 1997; Coniam, 2001; Ginther, 2002; Gruba, 1993; Londe, 2009). Although the research design in this strand was mainly identical to the previous one, learners' listening performance between groups fluctuated, suggesting that audio-only groups might sometimes exhibit better performance than video groups or vice versa. More importantly, the subtle difference between the two groups was not statistically significant. For example, Gruba (1993) compared participants' performance in audio versus video conditions after listening to or watching the simulated academic lecture. The result showed that the video group scored higher than audio group, but the observed difference did not achieve statistical significance. Different from Gruba (1993) and Brett (1997), who found video groups dominated audio groups, Coniam (2001) reported that the audio-only groups scored slightly higher than the video groups. Given the subtle difference, Coniam further postulated that the nature of the spoken text might influence the effect of video. In this study, a prolonged talk show video on the discussion of educational issues was used, but the inadequate nonverbal information offered in the talk show might yield less advantage to the video group. As a result, the audio group's performance might be superior to the video group.

Thus far, only two studies have found video debilitated participants' listening performance (Pusey & Lenz, 2014; Suvorov, 2009), but they jointly implied that the detrimental effects might not solely be attributed to video. Another way to phrase this would be, other factors, such as the spoken text (Suvorov, 2009) or learners' L2 proficiency (Pusey & Lenz, 2014), might play a critical role. In Suvorov's (2009) study, three listening conditions — audio-only, photograph, and video were compared using academic lectures and dialogue between students and professors. Results showed that the audio-only group scored significantly higher than the video groups, suggesting that videos may hinder test-takers' listening comprehension. Interestingly, the video's debilitated effects might be contingent on the spoken text. Learners might perform better in interactional dialogue than in transactional lectures under video conditions, suggesting that the negative impact might be alleviated within certain spoken text. Similar to Suvorov (2009), Pusey and Lenz (2014) found video groups scored significantly lower than audio-only groups. The results, though partially attributed to the unequal quality of the acoustic verbal inputs, Pusey and Lenz have implied that learners "pre-existing listening proficiency" might potentially affect their listening performance.

The discrepant findings in the previous studies could be generally summarized in the following. First of all, the effect of video on learners' listening performance might be interdependent with the congruence of verbal and nonverbal information (Wagner, 2010a). Videos might ameliorate learners' listening performance when the nonverbal information is interrelated with verbal information. Secondly, videos might be more beneficial in accompaniment with interactional discourse, such as dialogue than transactional discourse, such as academic lecture (Suvorov, 2009; Wagner, 2010a). Lastly, it is noteworthy that previous studies predominantly focused on comparing between video and audio-only conditions. A comprehensive understanding of the effect of visuals required further investigation of other visual formats, such as static visuals.

### 2.3.2 Static imagery as the input mode in listening assessment

The effect of static visuals (e.g., a single photograph or a series of still images) on learners' listening performance has received relatively scant attention in previous research (Ginther, 2002; Suvorov, 2009; Ockey, 2007). Suvorov (2009) contributed to a limited body of research examining the impact of audio-only, video, and a single photograph on test-takers' listening performance. One of the key findings of the research was that test-takers perform significantly better in a single photograph than in video conditions. The benefits of static visuals have been corroborated by Ginther (2002). He elicited that static visuals were particularly facilitated when congruent with audio information. More importantly, static visuals might promote the construction of mental representation and make the representation more tangible and long-lasting (Ginther, 2002).

To enrich the findings with different perspectives, Ockey (2007) applied the qualitative-oriented approach to explore learners' perception toward including video and a series of static images. With the aim of exploring learners' perceptions, a modified retrospective verbal report methodology was employed. For the visual conditions, the video was adapted from a lecture on politics, and the series of still images was derived from the videotape with a total of five images in each talk. The images would change every two minutes. Results showed that most test-takers perceived positively toward the use of a series of still images at the initial stage, but the effect seemed to gradually decline as the talk developed. They generally reported that still images could enhance their understanding of the context. Only two participants found themselves slightly distracted by the still images, particularly when they switched from one to the other.

As previous studies suggested, visual intervention might demonstrate different effects on learners. Some might perform better in static visuals than in audio-only or video versions (Suvorov, 2009), while others might be slightly distracted by static visuals (e.g., low-proficient learners) (Ockey, 2007). To understand factors contributing to the various outcomes, researchers discovered that learners' engagement in static imagery was limited to the initial stage (Ockey, 2007) since it entrenched their understanding of the contextual information. Once their knowledge of the context has been established, they might not perceive the information given afterward as valuable as the initial session. Additionally, their positive perception of static imagery might not ensure increased performance. For example, those who preferred a single photograph did not perform significantly better than others within their preferred mode. To elucidate such a phenomenon, Suvorov (2009) postulated that other modulating factors, such as L2 proficiency or cognitive load, might have more potential impact than their perceptions.

## 2.4 The Interaction between Learners' L2 Proficiency and Input Modes

A number of studies have asserted that visuals alone might not offer a comprehensive understanding of the relevant contributing factors involved in learners' listening performance (Ockey, 2007; Pusey & Lenz, 2014; Suvorov, 2009). In this regard, examining variables other than visuals might provide further insights into the modulating factors of listening performance. To examine the potential modulating variables in multimodal listening assessment, Suvorov and He (2022) reported that the interplay between visuals and other variables was the second most frequent research aim. Among different variables, L2 proficiency has garnered more attention in previous studies (Suvorov & He, 2022). Findings on the interaction between different visual formats and L2 proficiency have indicated that high-proficient learners consistently outperformed low-proficient learners, regardless of input mode (Ginther, 2002; Sueyoshi & Hardison, 2005). More specifically, these studies have posited that L2 proficiency might influence learners' listening performance more than input modes. As suggested by Sueyoshi and Hardison (2005), high-proficient learners performed significantly better than low-proficient learners in two visual-mediated conditions: audio-visual with face and audio-visual with face and gesture. For these groups of learners, their experienced decoding skills might facilitate their utilization of lip movements, which might lead to better listening performance. In the same vein, Ginther (2002) reported that high-proficient learners scored significantly higher in four listening conditions with different arrays of static visuals. These findings verified that learners with a comparatively higher level of L2 proficiency tend to demonstrate superior performance on tasks involving dynamic and static visuals. This is likely attributable to their extensive linguistic knowledge and well-developed decoding abilities. However, greater achievement might not be equivalent to significant influence. According to Liu

& Aryadoust's (2023) meta-analysis, the video might be more impactful to low-proficient learners due to their insufficient linguistic knowledge. The "impact" could be further elaborated as their reliance on video. These learners might rely on videos to compensate for their inadequate linguistic knowledge, and their listening performance might be contingent on the presence and the usefulness of videos (Liu & Aryadoust, 2023; Vandegrift, 2007). A similar phenomenon could also be observed in static visuals. For example, low-proficient learners tend to be distracted by static visuals (Ockey, 2007). High-proficient learners, on the other hand, might not be distracted by static visuals. Also, they might be unaffected in situations where static visuals are less informative and sometimes incongruent with aural inputs (Ginther, 2002; Ockey, 2007). As suggested above, high-proficient learners are prone to be more adaptive to different types of visuals due to their rich linguistic repertoire and better visual utilization ability. In contrast, low-proficient learners might be more susceptible to the modification of visuals due to their limited linguistic knowledge and reliance on visuals.

The modulating effects of L2 learners' proficiency on their listening performance within different input modes have been evidenced (Ginther, 2002; Liu & Aryadoust, 2023; Sueyoshi & Hardison, 2005; Vandegrift, 2007). The various effects were based on the differentiation of learners' L2 proficiency. Though their L2 proficiency has been identified in their demographic profiles, most of them were determined either by the language program they registered or a specific language test they took before participating in the classes (Suvorov & He, 2022). Recognizing the methodological shortcoming, Campoy-Cubilo and Querol-Julián (2015) stressed that learners' overall language and listening proficiency should be determined by pretests. By discriminating their L2 proficiency via pretests, researchers might be able to make a valid inference on the study's findings, particularly regarding the contributing factor of L2 proficiency. Additionally, initiating a pretest that aligned with the standardized language proficiency test might provide researchers with a validated measuring yardstick to contextualize the present study.

### 2.5 The Cognitive Demand of L2 Learners in Processing Multimodal Inputs

The cognitive load derived from processing multimodal inputs has been one of the major concerns for researchers, who revealed hesitancy on the inclusion of visuals in listening assessment (Buck, 2001; Ockey, 2007; Suvorov, 2015). Their concerns about the potential cognitive consequences, however, were merely based on speculation. Few studies have gauged learners' cognitive load, nor did they specify the type of cognitive load learners might encounter in multimodal listening assessment. Understanding learners' cognitive processing mechanisms is important to further examine learners' potential cognitive demand in a multimodal listening environment.

For L2 learners, processing multimodal inputs might be a cognitively demanding process due to their limited linguistic knowledge and immature decoding skills (Guichon & McLornan, 2008; Lee et al., 2021). The speed and accuracy of their processing might be contingent on the amount of information they could retain. Based on Mayer's (2001) limited capacity assumption, the amount of information humans can handle in each channel in a given time is limited. If the incoming information exceeds learners' cognitive capacity, their cognitive resources might be drained or overloaded. The constraints might compel learners to be selective on the incoming information, and most would choose to focus on auditory information (Cross, 2011; Gruba, 1993; Kamiya, 2022). Such a narrow focus on auditory information might deprive them of their opportunity to utilize nonverbal information to compensate for their limited linguistic knowledge (Vandergrift, 2007). As a result, their comprehension might be hindered, which could lead to poor listening performance.

To avoid overburdening L2 learners, examining their cognitive load while processing multimodal inputs is necessary. According to Sweller (1999), the type of cognitive load that learners might encounter could be categorized into two types: *Intrinsic* and *Extraneous cognitive load. Intrinsic cognitive load* refers to the inherent complexity of the materials. In the context of multimodal listening assessment, *Intrinsic cognitive load* usually stems from the difficulty and complexity of auditory information such as the lexical density, speech rate. To reduce learners' *Intrinsic cognitive load*, test developers should manage the difficulty level of the spoken text. *Extraneous cognitive load*, on the other hand, usually stems from instructors' design and presentation of the materials. In multimodal listening assessment, verbal and nonverbal information would be presented simultaneously. When learners were asked to process multimodal inputs simultaneously, their *Extraneous cognitive load* might increase. To manage their *Extraneous cognitive load*, incorporating visual input that could facilitate learners' listening comprehension without devoting unnecessary cognitive resources to the complex inputs might be important. In addition to the *Intrinsic* and *Extraneous cognitive load*, the *Germane cognitive load* was relevant to the mental efforts needed for the sense-making process (Mayer, 2005). It could indicate the outcome promoted by learners' mental efforts (Chi et al., 2023). High *Germane cognitive load* results when the materials could induce learners' mental efforts and boost their listening performance. To synthesize the above discussion on different measurements in managing learners' cognitive load, Mayer (2010, p.134) proposed an instructional guideline for designing multimodal materials. He stressed that task designers should manage *intrinsic cognitive load*, reduce *extraneous cognitive load*, and maximize *germane cognitive load*. In developing a multimodal listening assessment, the criteria could be achieved by selecting level-appropriate linguistic inputs and incorporating suitable visual formats that could facilitate listening comprehension without taxing learners' cognitive resources.

In brief, previous studies on the cognitive demand derived from multimodal listening assessment were primarily based on researchers' speculation. Few of them have indicated the type of cognitive load learners might encounter. Among the three types of cognitive load mentioned previously, *Intrinsic cognitive load* is associated with the complexity of aural information. Although it could be reduced by adding visuals, studies on visual inclusion were limited to listening performance. Few have explored the effect of visuals on reducing learners' *Intrinsic cognitive load.* As previous studies suggested, low-proficient learners might be more susceptible to different visual inputs. By monitoring their cognitive load derived from both linguistic and non-linguistic, researchers could gain more insight on the application of multimodal materials for testing and pedagogical purposes. Similarly, more research on the effect of visuals in increasing *Germane cognitive load* in assessment settings was needed to ensure whether multimodal listening assessment could promote learners' efforts and motivation in listening.

### 2.6 Summary

To explore the efficacy of different visual formats (i.e., dynamic video vs. sequential static imagery) on different proficient learners' listening comprehension and perceived cognitive load, the present study has reviewed Rost's listening assessment model (2016). Under the theoretical framework of Rost's model, the multimodal listening assessment construct has been conceptualized by incorporating linguistic and non-linguistic components. The linguistic elements constituted the prerequisite knowledge for listening comprehension, and non-linguistic elements operationalized the concept of visual inputs. Both verbal and nonverbal information co-construct the meaning of the listening assessment.

Aside from the theory underpinning the construct of multimodal listening assessment, the following sections will review the major findings of previous studies.

- Studies on the effects of video have found that video might lead to increased performance when the verbal and nonverbal information were congruent, particularly in interactional spoken text such as dialogue. While previous studies may report null or debilitated effects on videos, these negative impacts might not only be contingent on video.
- Evaluating the influence of visuals necessitates considering formats beyond video such as static visuals. Learners' might perform better in static visuals than video, and they generally found static visuals useful only at the initial stage. However, limited research on static visuals versus video might not allow researchers to generalize the findings.
- In addition to input modes, learners' L2 proficiency has been posited to be more influential to their listening performance. Some studies have further suggested that different proficient learners might exhibit different patterns within a diverse input mode.
- Existing studies on the cognitive demand derived from processing multimodal inputs were mostly based on speculation, and few have indicated the type of cognitive load learners might encounter. The effect of visuals in managing learners' *intrinsic cognitive load* has been examined, primarily in the learning settings, and few emphasized assessment scenarios.

Based on Rost's listening assessment model and the aforementioned findings, the present study would like to recruit participants with different L2 proficiency. In the study, learners' listening performance under the two visual formats would be examined in accompaniment with their perceived cognitive load while processing multimodal inputs. Details about the experimental design will be explained in the next chapter.

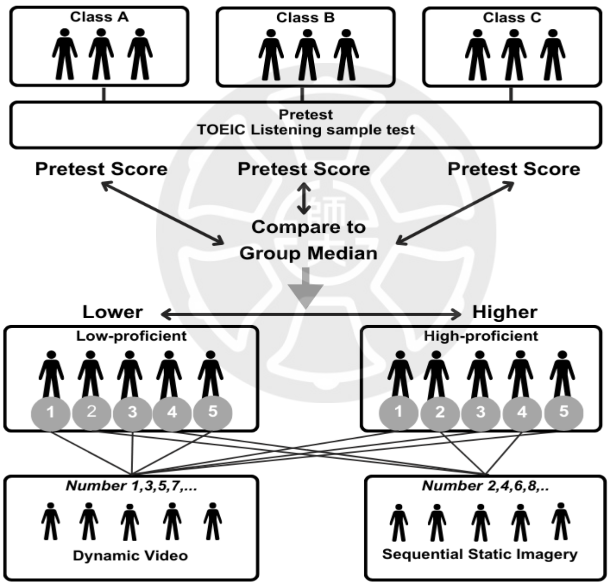
## CHAPTER THREE:
## METHODOLOGY

### 3.1 Research Design

A 2x2 factorial design (see Figure 3) was used in the present study to explore the effect of two visual formats (i.e., dynamic video vs. sequential static imagery) and different L2 proficiency (i.e., high vs. low proficiency) on listening comprehension performance, resulting in two between-subject independent variables and one dependent variable.

**Figure 3**

*Visual Schematization of the Grouping Procedure*




Participants from three classes, class A, B, and C, took the TOEIC official listening sample exam as the pretest to determine their listening proficiency. The pretest was conducted by their English teacher during the first few weeks of the semester. Namely, prior to their participation of the present study, all participants had already completed the pretest. After taking the pretest, their listening proficiency was further differentiated by their test scores. Participants' listening test scores in all three classes were calculated, and the median score of all participants was functioned as a reference for participants' listening proficiency. Participants who scored below the median were classified as low proficient learners and those who scored higher than the median were labeled as high-proficient learners. The number of high and low-proficient learners might vary from class to class, but the distribution of each proficiency among the cohort of participants was approximately equal.

After differentiating participants' listening proficiency, they were assigned into different groups. In the group division, each class was viewed as a single unit. In each unit, participants' pretest scores, both high and low-proficient, were arranged in ascending order. The lowest score in each group was marked as number 1, and the second lowest was number 2. Next, participants, both high and low-proficient, who were marked as odd numbers (e.g., 1, 3, 5, 7) were assigned to dynamic video conditions, and those marked as even numbers (e.g., 2, 4, 6, 8) were assigned into sequential static imagery condition.

Based on the assignment, half of the participants who were labeled as high-proficient in the class were designated to dynamic video condition, and another half were in sequential static imagery condition. On the other hand, half of the participants who were labeled as low- proficient in the class were also assigned in dynamic video condition and another half were assigned in sequential static imagery condition.

### 3.2 Participants

A cohort of 70 from a public senior high school in Northern Taiwan was recruited in the present study. All the participants were twelfth graders from three intact classes – Class A with 25 students, Class B with 20 students, and Class C with 25 students. Three classes were taught by the same English teacher in the English Aural-Oral Training class. Participants were between 17 to 18 years old and had been learning

English for about ten years in school. After analyzing participants' listening performances, 2 participants' performances deviated from the normality, and their performance data were excluded in the following analysis. The final number of participants was 68, with 24 students from Class A, 20 students from Class B, and 24 students from Class C.

## 3.3 Materials

### 3.3.1 Video selection

The visual inputs in the present study were a series of English learning videos made by Voice of America (VOA: https://learningenglish.voanews.com/), a state-owned news company and radio broadcaster in the United States. The video series, *Let's learn English*, has two levels (i.e., Level 1 & 2). In the present study, the videos in level 2 were adopted. The level 2 series included 30 episodes, each lasting for approximately 5 minutes. It included two sessions: a dialogue video and several explicit instructions on a specific grammar concept applied in the dialogue. In each video, the dialogue would be displayed first. During the dialogue, several short grammar instruction sessions were broadcasted to explain the key sentences and phrases used by the speakers. The setting of the dialogue was in the United States, and the dialogue had a wide range of topics, such as business or travel. In each episode, there was a conversation under a topic. The settings of the conversation would change throughout the development of the conversation. Though the dialogues were scripted, the actors still demonstrated some features of an unscripted spoken text in the conversation, for example, hesitation or filled and unfilled pauses. In addition to the linguistic aspects, extra-linguistic features such as facial expressions and gestures could also be observed in the dialogue.

The series of videos had 30 episodes, but not all were appropriate for incorporating in the listening assessment. To determine suitable videos for the listening assessment, researchers must establish essential selection criteria. First of all, the topic of the video should be aligned with real-life communication. Within such standard, episodes that contained fictional content, for instance, a heroine rescuing human beings, were excluded. Secondly, human talk should dedicate the majority of communicative events in the video, rather than background music or animations. For example, episodes included a session illustrating a series of exercises done by the actors with some background music. In this session, the audience could only gain much information from the conversation between the speakers. This type of video, thus, should not be included as the materials for listening assessment. Lastly, the conversation in the videos should include a variety of events happening to the actors or people surrounding them. Various communicative events could offer sufficient resources for constructing listening comprehension questions.

The introduction to the criteria provided a holistic point of view on the type of videos being selected in the listening assessment. An in-depth analysis of the video scripts might yield a deeper understanding of the lexical profile and the difficulty level of the spoken text. To conduct a text-analysis, all the video scripts were uploaded to the Oxford Text Checker. This online text analyzer allowed users to upload their text and analyze the sentence length, word count, and lexical profile. For the lexical profile, the Oxford Text Checker had several reference wordlists, and the one adopted in the present study was the 5,000 wordlist in the Oxford dictionary. It was expected that the 5,000 wordlist could cover as many as the words contained in the video scripts to avoid too many unlisted words. After the analysis, the platform automatically present a text analysis report, with detailed information about the video scripts. The following tables indicate the analysis report of the six video scripts. As Table 1 suggested, the average duration of the video was approximately 228 seconds, and the average word count was 421 words, with an average sentence length of 5.7 words. The average speech rate of the video was around 111 words per minute.

In addition, Table 2 indicates the lexical profile of the videos. The word level distribution was aligned with the CEFR framework, suggesting that B1-level vocabulary could reach approximately 90% of the video script coverage. Notably, the coverage was expected to increase with the inclusion of the non-listed words. The non-listed words were mostly the characters' names (e.g., Anna, Pete) or proper nouns (e.g., the White House) that the analyzer might not identified. These non-listed words might repeatedly appear in the spoken text and might not be too challenging for listeners. Therefore, the actual lexical coverage of B1 level vocabulary might be higher than 90% with the inclusion of proper nouns.

**Table 1**

*Results of the Video Scripts Text Analysis*

| **Videos** | **Duration (sec)** | **Word Count (words)** | **Average Sentence Length (words/min)** |
|---|---|---|---|
| 1 | 267 | 343 | 5.4 |
| 2 | 296 | 391 | 5.9 |
| 3 | 229 | 380 | 6.8 |
| 4 | 296 | 538 | 6 |
| 5 | 310 | 411 | 5.1 |
| 6 | 299 | 465 | 5.1 |
| Total | 282.8 | 421 | 5.7 |

**Table 2**
*The Lexical Profile of Video Scripts*

| Videos | CEFR level | | | | | Non-listed words |
|---|---|---|---|---|---|---|
| | A1 | A2 | B1 | B2 | C1 | |
| 1 | 83% | 5% | 2% | 4% | 1% | 5% |
| 2 | 84% | 4% | 2% | 2% | 0% | 8% |
| 3 | 85% | 6% | 1% | 1% | 0% | 6% |
| 4 | 77% | 6% | 7% | 0% | 1% | 9% |
| 5 | 82% | 5% | 1% | 0% | 0% | 11% |
| 6 | 82% | 6% | 1% | 1% | 0% | 9% |

### 3.3.2 Listening condition

In the present study, participants with different L2 proficiency were assigned to the dynamic video condition or the sequential static imagery condition. The detailed description of the design of each listening condition were as follows. All video formats were comprised of at least one episode. The episodes were selected and downloaded from the official VOA learning English website. During the editing process, the grammar instruction sessions were excluded, and the remaining dialogue sessions lasted for approximately 3.5 to 4 minutes. At the very beginning of each dialogue, there was a small heading and subheading indicating the topic and the focused grammar concept of the dialogue. To avoid confusing the listeners, the heading and subheading were blocked by stickers showing the ordinal number of the video, as shown in Figure 4.

**Figure 4**

*Example Screenshot of the Video Condition*


For the purpose of listening assessment, English captions were pixelated. Aside from the editing mentioned above, the rest of the video maintained its original states. To develop the sequential static imagery formats, researchers first extracted the audio soundtrack from the video format. After extracting the audio soundtrack, the static images were selected. The static images were a series of screenshots of video formats. The number of screenshots depended on the complexity of the conversation. For example, if the conversation mostly happened in just a few scenarios, and the actors rarely exhibited distinctive actions, fewer screenshots would be selected or vice versa. Once the screenshots were selected, they would be imported into the editing software called Power Director, an editing software that allowed editors to upload their own materials and conduct post-editing. No other special effects were applied when editing the static images since the screenshots were already post-edited in the video condition. Lastly, the images were inserted in the video track and arranged based on the development of the conversation. The image transition was not predetermined but followed the progression of the dialogue.

## 3.4 Instruments

### 3.4.1 Listening proficient pretest

To differentiate participants' L2 proficiency, particularly on their listening proficiency, a listening proficiency pretest was conducted. The listening proficiency pretest is the sample test in the Test of English for International Communication (TOEIC) Listening test official test preparation guide, 7th edition. The Educational Testing Service (ETS) published this sample listening test and has been recognized as a valid test battery for measuring test-takers overall listening comprehension ability. The entire listening test included 100 multiple choice questions with four different types of test items: photographs, question-response, conversation, and short talk (See Table 3). The listening test lasted for approximately 45 minutes. In the present study, participants were asked to take the listening test, and their test performance was evaluated. Each correct answer would get 1 point, and their scores would range from 0 to 100.

**Table 3**
*Item type and the number of test items in the TOEIC sample listening test (ETS, 2022)*

| TOEIC Listening Test | | |
|---|---|---|
| **Part** | **Item Type** | **Number of items** |
| Part 1 | Photograph | 6 |
| Part 2 | Question-Response | 25 |
| Part 3 | Conversation | 39 |
| Part 4 | Talks | 30 |

### 3.4.2 Listening comprehension items

The listening comprehension items included 30 four-option multiple-choice questions, all developed according to the content of the videos, with five questions for each video. Within the five questions, one assessed global comprehension, three measured local comprehension, and one gauged inferential comprehension. The entire six videos all followed the same arrangement. The arrangement was aligned with the educational purposes for learning languages (Nation & Newton, 2020; Weir & Vidaković, 2013). According to the definition, the purpose of listening varied, and could be generally summarized, but not limited to, the following statements:

- To understand the main message or important information bits within a recording (Global)
- To understand detailed information or to find specific data (Local)
- To understand the context and infer information (Inferential)

With comprehension questions covered these aspects, test-takers' listening ability could be precisely gauged and genuinely reflect on their listening test scores. To ensure equitable application of test items across both visual formats and to avoid advantaging either condition, researchers designed items that were congruent in both formats and answerable by participants in both conditions. The complete sampled listening comprehension items are attached in Appendix A.

Apart from the test design, the following section would emphasize the sequence of the test. The listening comprehension questions were presented immediately after the end of each video on the monitor in front of the classroom. It is worth noting that only questions were presented on the monitor. The four options, one correct answer, and three distractors, were only printed on the test booklet. In other words, participants were required to listen and read the questions on the monitor and choose a correct answer from options printed on the test booklet. For each item, participants had 15 seconds to answer, and the entire listening assessment lasted 45 minutes.

### 3.4.3 Questionnaire

After conducting the listening comprehension test, participants were required to complete a questionnaire which contained three sections. The first section intended to elicit participants' background information, with a total of four questions concerning their gender, past English learning experience, scores in college entrance listening examination, and their current English proficiency level, which was determined by the score of the most recent language proficiency test they have taken. The first three questions were checkbox questions which only required participants to check the option that matched their current situation. The last question, regarding their current language proficiency level, was a fill-in-blank question that required them to fill in the name of the language proficiency test and their exact score.

The second part of the questionnaire focused on participants' perceived cognitive load. The cognitive load questionnaire was adapted from Chi et al. (2023). The entire questionnaire had nine items, evaluating test-takers' *intrinsic, extrinsic, and germane cognitive load* (Paas, Rankl & Sweller, 2003). According to Chi et al. (2023), six items measuring intrinsic and extrinsic cognitive load were adapted from Leppink, Van der Vleuten, Van Gog & Van Merriënboer (2013), and the rest of the questions, measuring germane cognitive load were adapted from Klepsch, Schmitz & Seufert (2017). These items were proved to be valid instruments for gauging participants' perceived cognitive load. Thus, the present study adapted their questionnaire and rephrased some statements to match the current context (See Table 4).

**Table 4**

*Question statements of the cognitive load questionnaire*

| | | Original items | | Rephrased items |
|---|---|---|---|---|
| Measuring intrinsic cognitive load | 1 | The topic (Business clustering) covered in the video was very complex. | 1 | The main idea of the conversation covered in the video was difficult to understand. |
| | 2 | The story used to explain the topic was very complex for me. | 2 | The language use in the conversation was very difficult for me. |
| | 3 | The terms used to explain the topic were very complex for me. | 3 | The speech rate in the conversation was too fast to follow. |
| Measuring extraneous cognitive load | 4 | The concurrent presentation of captions and dynamic/static images was effective for learning this topic. | 4 | The visual inputs in the listening test were effective for understanding the conversation and answering the questions. |
| | 5 | The full/partial captions in the video was effective for learning this topic. | 5 | The concurrent presentation of visuals and audio inputs was effective for understanding the conversation and answering questions. |
| | 6 | The animation/images in the video was/were effective for learning this topic. | 6 | The nonverbal information such as facial expression and gestures was effective for understanding the conversation and answering the questions. |
| Measuring germane cognitive load | 7 | I made an effort to understand the details and overall context of the video. | 7 | I made an effort to understand the visuals as well as the audio inputs in the listening test. |
| | 8 | My point while watching the video was to understand everything correctly. | 8 | My point while listening and watching the video was to understand everything in the conversation and answer the questions correctly |
| | 9 | The video consisted of elements supporting my comprehension of the topic. | 9 | The video consisted of elements supporting my comprehension of the listening test |

The questionnaire of the present study was a nine-item-10-point Likert scale questionnaire in which participants were required to rate from zero to ten, with a zero score indicating *strongly disagree* and ten meaning *strongly agree.* The questionnaire demonstrated in the following table was written in English, but the actual questionnaire applied in the experiment was translated into Mandarin Chinese.

The last section emphasized participants' perceived usefulness and their perception of the inclusion of visuals in listening assessment. The questionnaire included five questions, three checkbox questions and two open-ended questions. These questions were primarily referred to Ockey's research (2007) on test-takers' perception after taking listening assessments in video-mediated or series of static image conditions. With participants' responses to the questionnaire, researchers could triangulate their performance and perception after receiving a multimodal listening assessment. The sample items of the questionnaire are attached in Appendix B.

### 3.5 Procedure

The present study was conducted in an elected course entitled English Aural-Oral Training. Prior to the study, participants from three different classes, Class A, B and C, were required to take the TOEIC sample listening pretest in order to evaluate their overall listening proficiency. Following the pretest, participants' pretest scores were calculated, and the median score of all participants was the benchmark to differentiate participants' listening proficiency. Each participant's pretest score was compared with the median score. Participants who scored equal to or above the median were assigned to highly proficient group, and those who scored below the median were assigned to the low proficient group. The assignment could ensure that the number of participants in high proficient group was tantamount to a low proficient group. After classifying into two proficient groups, participants' scores were further arranged in ascending order. In other words, the lowest score was labeled as number one, and the second lowest within the group was marked as number two. Once each participant was assigned a number, they were further distributed to dynamic video or sequential static imagery conditions. For those who were marked with odd numbers (e.g., 1, 3, 5) were assigned to dynamic video format, and those labeled as even numbers (e.g., 2, 4, 6) were assigned to sequential static imagery format. The arrangement could ensure that participants in both proficient groups were equally distributed to either visual format, and the rest of the two classes also followed this arrangement.

Participants from three classes were tested in different class times. Students in each class were distributed to either high or low-proficient groups, and they were further assigned to one of the two visual formats according to the procedure noted above. The listening tests were initiated three times in three different classrooms. For the first time, students in Classes A and B were separated into two classrooms based on their assigned visual formats. The second time, only one student in Class B was being tested. A make-up test was held since the student was absent in the first test. For the last time, students in Class C were separated into two classrooms based on their assigned visual formats. The classroom size varied, but the monitors in the three classrooms were the same size. Identical monitor size could ensure that the quality of the visual formats was consistent in all three classrooms.

In the first and the third test, the researcher and their English teacher were each in charge of one classroom to deliver the test materials, explain the procedure, and monitor participants' behavior. As for the second test, the researcher was in charge. Prior to the experiment, the procedure was explained to ensure that all participants had no further questions concerning the procedure. Once the explanation was completed, each participant would be required to check if they all got the test materials, including a test booklet, an answer sheet, and some blank papers for them to take notes while listening. After ensuring that they all got the necessary items for the test, they were required to pay attention to the direction of the assessment. The direction was presented on the monitor in front of the classroom in accompaniment with an audio broadcast. Participants could also refer to the direction in the test booklet.

Immediately after the direction, the listening assessment began. The first session of the assessment was a conversation presented in either video or static imagery condition. By the end of the conversation, directions for answering a set of listening comprehension questions would be displayed on the monitor and broadcasted via audio recording. When answering the comprehension questions, participants were required to read and listen to the comprehension questions presented on the monitor and select the correct answer from the four options written on the test booklet. It is worth noting that the options would only be printed on the test booklet. Participants had 15 seconds to answer the questions and the listening test lasted approximately 40 minutes. After the listening assessment, participants were required to complete a questionnaire. It took approximately 10 minutes for them to complete the questionnaire. Once they completed and handed in all the documents, they were allowed to leave the classroom.

### 3.6 Data Analysis

In order to answer the first three research questions regarding the role of participants' listening proficiency and the effect of different visual presentations on participants' listening performances, a two-way ANOVA was conducted to scrutinize whether participants' listening performance were significantly different among the four experimental groups. For the fourth research question, which investigated participants' perceived cognitive load in different listening conditions, the mean scores of the Likert scale were calculated.

# CHAPTER FOUR:
# RESULTS

The present study explored the role of participants' listening proficiency (i.e., high or low) and different types of visual formats (i.e., dynamic video or sequential static imagery) on their listening performance and perceived cognitive load. The following sections will first present the descriptive statistics of the listening comprehension score among four experimental groups. A two-way ANOVA analysis results will follow after the descriptive data to further address the modulating effects of the two variables (i.e., listening proficiency and visual formats). Considering the cognitive burden that multimodal inputs might cause, the results of participants' perceived cognitive load and other qualitative data about their perception of nonverbal inputs will be presented at the very end of this section.

## 4.1 Descriptive Statistics

Table 5 indicated the number of participants (*n*), mean scores (*M*), and standard deviation (*SD*) of each listening condition. The mean scores of the four experimental groups suggested that high-proficient learners in dynamic video condition gained the highest score ($M$ = 19.88), followed by another group of high-proficient learners in sequential static imagery condition ($M$ = 17.89), and then the low-proficient learners in the dynamic video condition ($M$ = 16.94). Low-proficient learners in the sequential static imagery condition yielded the lowest score ($M$ = 12.85).

In addition to a broad view of participants' listening performance among four listening conditions, Table 5 also showed the results by variables. High-proficient learners ($M$ = 18.86) attained higher mean scores than low-proficient learners ($M$ = 15.09), regardless of visual formats. As for the listening performance by visual formats, the dynamic video group ($M$ = 18.45) showed a higher mean score than the sequential static imagery group ($M$ = 15.75). Therefore, whether the differences between each variable have reached statistical significance should be further corroborated via ANOVA analysis

**Table 5**

*Descriptive Statistics of Listening Comprehension Score*

| **Proficiency** | **Visual Format** | ***n*** | ***M*** | ***SD*** |
|---|---|---|---|---|
| High Proficient | Dynamic Video | 18 | 19.88 | 3.37 |
| | Sequential Static Imagery | 19 | 17.89 | 3.49 |
| | Total | 37 | 18.86 | 3.53 |
| Low Proficient | Dynamic Video | 17 | 16.94 | 3.56 |
| | Sequential Static Imagery | 14 | 12.85 | 3.86 |
| | Total | 31 | 15.09 | 4.18 |
| Total | Dynamic Video | 35 | 18.45 | 3.72 |
| | Sequential Static Imagery | 33 | 15.75 | 4.39 |

## 4.2 Two-way ANOVA Analysis

The normality of participants' listening comprehension scores among the four experimental groups was measured via the Shapiro-Wilk test. The test result did not show a significant departure from normality, with a p-value greater than 0.05. In addition, Levene's test, which gauged the homogeneity of participants' listening performance among four groups, was insignificant, $F(3, 64) = .34$, $p = .793$. The normality and homogeneous test results satisfied the prerequisite assumption for ANOVA analysis.

**Table 6**

*Two-way ANOVA Results of L2 Proficiency and Visual Formats on Listening Comprehension*

| **Source** | **Dependent variable** | ***df*** | ***F*** | ***p*** | ***Partial eta-squared ($\eta^2$)*** |
|---|---|---|---|---|---|
| Proficiency | Listening Comprehension Test Score | 1 | 21.12 | **.000*** | .248 |
| Visual Format | | 1 | 12.23 | **.001*** | .161 |
| Proficiency × Visual Format | | 1 | 1.44 | .233 | .022 |

*Note. * p < 0.05*

Results of the 2×2 between-subject ANOVA were presented in Table 6, with participants' listening proficiency and visual formats as independent variables, and their listening comprehension score as the dependent variable. Table 6 indicated that both listening proficiency, $F(1, 64) = 21.12$, $p < .05$, with a large effect size ($\eta^2 = .248$), and visual format, $F(1, 64) = 12.23$, $p < .05$, with a large effect size ($\eta^2 = .161$) caused a significant main effect on their listening comprehension. The statistical results indicated that both language proficiency and visual formats have significant impact on learners' listening comprehension. Among these two variables, learners' L2 proficiency was more influential than visual formats on their listening performance due to its relatively larger effect size. However, the interaction effect between listening proficiency and visual format did not reach significance, $F(1, 64) = 1.44$, $p > .05$, suggesting that different proficient participants' listening performance were not dependent on the type of visual formats they received (See Figure 5).

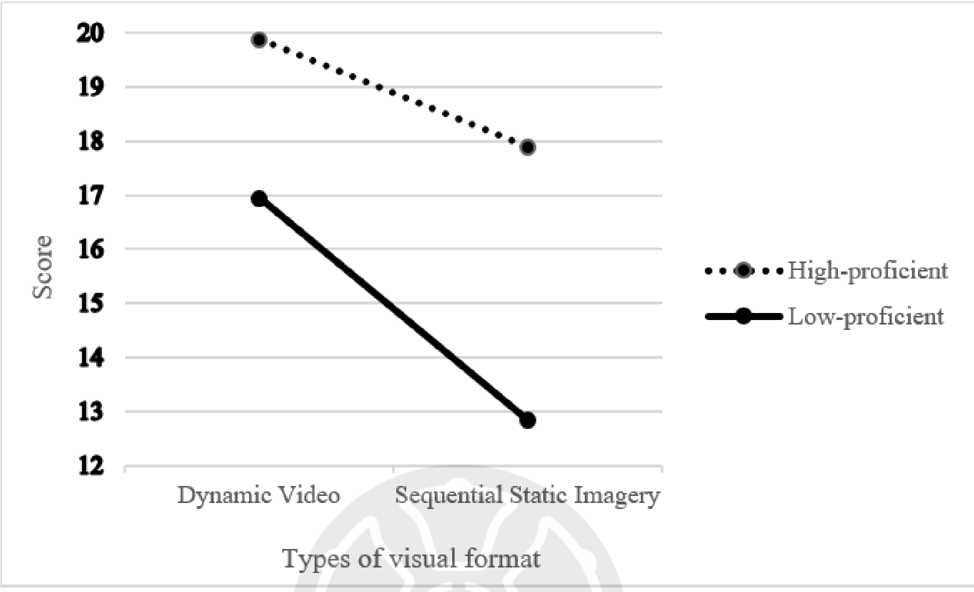



**Figure 5**
*The Interaction of L2 Proficiency and Visual Formats*

**4.3 Questionnaire Results**

As indicated in Table 7, three types of perceived cognitive load, intrinsic ($\alpha = .79$), extraneous ($\alpha = .93$), and germane ($\alpha = .86$), all reached high internal consistency according to Cronbach's alphas coefficient. On such a basis, the ensuing paragraph would highlight the major results of the questionnaire.

Intrinsic cognitive load, which is derived from the inherent complexity of a concept, is difficult for instructors to alleviate. Accordingly, low-proficient learners tend to report higher perceived intrinsic cognitive load than high-proficient learners, with low-proficient learners in the sequential static imagery group rated the highest perceived intrinsic cognitive load ($M = 4.88$). Extraneous cognitive load, on the other hand, was relevant to the design of the materials. This section is a reversed item; namely, higher participants' ratings referred to lower perceived extraneous cognitive load. As the table suggested, participants in the sequential static imagery group tend to report higher extraneous cognitive load than those in dynamic video groups (HSSI: $M$ = 6.31; LSSI: $M$ = 6.88). An ideal visual manipulation might reduce learners' extraneous cognitive load and avoid requiring them to spend extra attentional resource irrelevant to listening comprehension. The results of the extraneous cognitive load rating might indicate that sequential static imagery might not be an ideal approach to visual manipulation. Lastly, germane cognitive load refers to learners' mental exertion and motivation derived from the material design. According to the results, both visual formats have induced high germane cognitive load ($M$ > 7.00), with dynamic video yielding a slightly higher germane cognitive load (HDV: $M$ = 8.72; LDV: $M$ = 8.11).

Another part of the questionnaire aimed to probe participants' perceived usefulness of visuals, their attention to nonverbal information, and the information they could learn from viewing visuals. The questionnaire would have two types of questions: tick-box and short response.

The analysis of participants' responses will be presented in the following sections. Most participants considered visual information helpful (83%), and only a few reported being distracted by visuals. For those who reported being distracted by visuals, most of them were under sequential static conditions (5 out of 7). Based on the results of the survey, the inclusion of visuals was considered to be beneficial for most participants. Some were reported to be distracted, particularly from sequential static imagery conditions.

**Table 7**

*Descriptive Statistics of the Cognitive Load Questionnaire*

| | Statements | Cronbach's Alpha (α) | H+DV | H+SSI | L+DV | L+SSI |
|---|---|---|---|---|---|---|
| | | | *M* | | | |
| Intrinsic cognitive load | 1. The main idea of the conversation covered in the video was difficult to understand. | | 3.67 | 3.89 | 3.94 | 4.93 |
| | 2. The language use in the conversation was very difficult for me. | | 3.78 | 3.84 | 4.00 | 4.57 |
| | 3. The speech rate in the conversation was too fast to follow. | | 3.44 | 4.26 | 4.29 | 5.14 |
| | Total | .797 | 3.62 | 4.00 | 4.07 | 4.88 |
| Extraneous cognitive load | 4 The visual inputs in the listening test were effective for understanding the conversation and answering the questions. | | 8.06 | 6.37 | 8.00 | 6.86 |
| | 5. The concurrent presentation of visuals and audio inputs was effective for understanding the conversation and answering questions. | | 8.17 | 6.53 | 8.06 | 6.79 |
| | 6. The nonverbal information such as facial expression and gestures was effective for understanding the conversation and answering the questions. | | 8.06 | 6.05 | 8.00 | 7.00 |
| | Total | .937 | 8.09 | 6.31 | 8.01 | 6.88 |
| Germane cognitive load | 7. I made an effort to understand the visuals as well as the audio inputs in the listening test. | | 9.11 | 7.74 | 8.53 | 7.00 |
| | 8. My point while listening and watching the video was to understand everything in the conversation and answer the questions correctly | | 8.44 | 7.37 | 7.94 | 7.14 |
| | 9. The video consisted of elements supporting my comprehension of the listening test | | 8.61 | 6.74 | 7.88 | 7.07 |
| | Total | .869 | 8.72 | 7.28 | 8.11 | 7.07 |

*Note.* H = High-proficient; L = Low-proficient; DV = Dynamic Video; SSI = Sequential Static Imagery.

Another part of the questionnaire aimed to probe participants' perceived usefulness of visuals, their attention to nonverbal information, and the information they could learn from viewing visuals. The questionnaire would have two types of questions: tick-box and short response.

The analysis of participants' responses will be presented in the following sections. Most participants considered visual information helpful (83%), and only a few reported being distracted by visuals. For those who reported being distracted by visuals, most of them were under sequential static conditions (5 out of 7). Based on the results of the survey, the inclusion of visuals was considered to be beneficial for most participants. Some were reported to be distracted, particularly from sequential static imagery conditions.

The second question asked whether participants would focus on the speaker's face while listening. Most participants reported that they would pay attention to the speaker's face (91%). Among the six participants who reported "not paying attention" to speakers' faces, 5 out of 6 participants were from sequential static imagery conditions. The results indicated that participants' tend to devote their attention to the speaker's face while listening. In addition, for those who reported paying attention to speakers' faces were further required to specify the area they focused. In the video condition, the number of participants focusing on eyes and lips was nearly equal (16 versus 14), while in a static condition, more participants reported focusing on eyes than lips (18 versus 4). The report suggested that learners in dynamic video groups had more engagement in lip movement than those in static visual groups, but they exhibited similar patterns in viewing speakers' eyes.

Most participants who reported focusing on lips were under video conditions, and the information they learned from watching the speaker's lip was further discussed in their responses. Those who reported focusing on lips were inclined to use them to comprehend the words being said by the speakers. As the following statements:

- HDV-8

  *"I can better understand the words being said [by the speaker]."*

- LDV-44

  *"I can know what the speaker intends to say by watching the lip movement."*

- LDV-50

  "*I can ensure which words are being said by watching the lip movement*."

- LDV-41

  "*I can know the lip movement when certain words are said.*"

The benefits could multiply with the simultaneous presentation of verbal (aural) and nonverbal (lip movement) inputs.

- LDV-49

  "*I could better understand the conversation with lip movement and aural inputs presented simultaneously*."

Participants who reported focusing on the speaker's eyes could learn about the speaker's emotion, mood, or characteristics during the conversation. As the following responses suggested:

- HSSI-19

  *"I can infer the speaker's emotion either positive or negative by focusing on their eyes*."

- LDV-45

  "*I can infer the speakers' emotion at the moment, and so do their attitude."*

- HDV-16

  "*I can know a person's mood and personality*."

The fourth question asked whether they would pay attention to speakers' gestures. In the video condition, about 75% (27 out of 35) reported that they would focus on speakers' gestures, and the proportion was approximately the same in the static condition (25 out of 34). Learners who reported "not paying attention" to gestures were mostly from high-proficient groups (9 out of 14).

Participants who reported focusing on the speaker's gesture could learn about the relationship between the speaker and the interlocutor.

- LSSI-61

  "*I can know who the speaker is talking to by their gesture*."

- LSSI-60

  "*Gestures help me to clarify who the speaker is*."

Gestures may sometimes draw participants' attention to the important information within the conversation.

- HDV-13

  "*Gestures might imply the position of certain objects*."

- LSSI-68

  "*Gesture may indicate numbers*."

Some gestures carried no meaning but could indicate the meaning units within a conversation.

- LDV-54

  "*I can know which part is being emphasized [by focusing on gestures*]."

- LSSI-67

  “*Gesture will be used when it comes to an important part [of a conversation]*.”

The seventh question asked whether participants would focus on the objects that occurred in the conversation. In video condition, 74% of them reported focusing on the objects. In static condition, 78% of them reported focusing on the objects.

Participants could learn about location, or place, and the time and order of the conversation by focusing on objects.

- HDV-3

  “*I can know the place of the conversation*.”

- HDV-16

  “*I can know the time and location.*”

- HDV-6

  “*I can know the order of the conversation.*”

In addition, they could also learn about the gist or the topic of the conversation by focusing on the objects.

- HDV-15

  “*I can know the topic.*”

- HSI-21

  “*I can contextualize the conversation and get the gist faster.*”

When there was an unknown word, they could infer the meaning from the object.

- LDV-42

  “I can guess the unknown word from the objects.”

The objects were usually relevant to the conversation.

- LSI-38

  "The object might be relevant to the talk."

Lastly, anticipation could also be made by focusing on the objects.

- LSI-57

  "I can anticipate what is going to happen in the talk."

- LSI-66

  "I guess the questions might be related to the object."

## CHAPTER FIVE:

## DISCUSSION

The present study utilized two visual formats (i.e., dynamic video vs. sequential static imagery) as the input modes incorporated in the multimodal listening assessment. The first research question probed into the effect of the two visual formats on L2 learners, irrespective of their proficiency. The second research question addressed the effect of learners' L2 proficiency on their listening performance, regardless of input modes. The third research question investigated the interaction between visual formats and learners' L2 proficiency. The fourth research question further examined different proficient learners' perceived cognitive load within the two visual formats. In the ensuing paragraphs, an in-depth discussion will be initiated by exploring the key findings or the present study.

### 5.1 The Effects of Different Visual Modalities

The first research question examined whether the presentation of different visual formats will affect L2 learners' listening comprehension, regardless of their L2 proficiency. Results from the quantitative data necessitated the consideration of the various impacts resulting from diverse arrays of visuals. The incorporation of visual inputs significantly affected L2 learners' listening comprehension. Observed mean differences also suggested dynamic video yielded more significant influence than sequential static imagery. In other words, dynamic video might be more effective than sequential static imagery in eliciting participants' multimodal listening skills, in which participants were required to process both linguistic and nonlinguistic inputs simultaneously. Although the benefit of dynamic video to learners' listening performance has been well-informed in previous studies, such outcomes have shed light on the development of multimodal listening assessment. The following paragraph will further analyze the result on the theoretical premise of Rost's listening assessment model and provide a possible explanation for the different performances under two visual formats.

In Rost's listening assessment model, general knowledge could be incorporated into the listening construct when examining learners' listening performance. General knowledge, as the model suggested, has several components, and the most relevant component to the current finding is extra-linguistic knowledge. Extra-linguistic knowledge refers to the visual content and other nonverbal cues, such as gestures and facial expressions. The findings of the present study suggested that dynamic video might be a relatively ideal visual format for assessing learners' listening performance since it demonstrated detailed nonverbal features made by the speakers and interlocutors. In addition, the wide variety of objects that appeared in the dynamic videos and the constantly changing contextual settings have provided listeners with rich information to contextualize the conversation and mitigate their limited linguistic knowledge.

The above-mentioned benefits could be corroborated by participants' overall perceptions and qualitative responses toward the multimodal listening assessment. In the dynamic video group, over 90 percent of the participants considered visual information helpful to their listening comprehension, and these groups of learners also paid close attention to speakers' faces while listening. Among different parts of speakers' faces, their eyes and lips were two major areas of focus. Such allocation of learners' attentional resources was in line with Batty's study (2020) in which participants exhibited great attention on speakers' faces. In the follow-up open-ended questions, participants reported that they could differentiate speakers' emotions by focusing on their eyes and better understand the words being said by focusing on speakers' lips. These qualitative responses have verified the findings in previous studies on the role of different nonverbal features in learners' increased listening comprehension (Batty, 2020; Sueyoshi & Hardison, 2005).

In sum, the present study indicated that visual modalities might significantly affect L2 learners' listening comprehension. More specifically, the dynamic video might be more effective than sequential static imagery in facilitating learners' listening comprehension since the salient nonverbal features might not only complement their understanding of the video but also ameliorate their awareness of different nonverbal information.

### 5.2 The Effects of L2 Learners' Proficiency on Listening Performance

The second research question elicited the effects of learners' L2 proficiency. Findings from the analysis reinforced the influence of L2 proficiency, indicating that learners' L2 proficiency significantly affected learners' listening comprehension within different visual formats (Pusey & Lenz, 2014). The quantitative results also bolstered that high-proficient learners consistently outperformed low-proficient learners, regardless of visual formats. The major findings for the second research question seem reasonable since learners' L2 proficiency has been ascertained to be a predictor for their L2 listening comprehension ability (Vandergrift, 2006). The present study reassured the phenomenon and expanded the assumption in multimodal listening context, suggesting that learners with richer linguistic repertoire and mature decoding skills might exhibit better listening performance (Liu & Aryadoust, 2023).

Compared to high-proficient learners, low-proficient learners' seemed to demonstrate a relatively lower listening performance. However, they generally reported a positive attitude toward the inclusion of visuals, and attempted to utilize nonverbal information to complement their listening comprehension (Vandergrift, 2007). According to their qualitative responses, approximately 90 percent of the low-proficient learners consider visuals helpful to their listening comprehension. They paid close attention to the speakers' face, particularly to the speakers' eyes. Apart from facial expressions, they also devoted their attentional resources to the gestures made by the speakers. For low-proficient learners, the purpose of visual attention varied across different nonverbal information. For example, they could discern the emphasized content by focusing on speakers' gestures and infer some unfamiliar words they heard by focusing on speakers' lip movements. Their overall perception and focuses on visual information have implied that low-proficient learners, though their performance were inferior to their high-proficient counterpart, were still benefited from the incorporation of visuals in different aspects.

The findings above, especially on low-proficient learners' positive perception of visuals, were contradicted by Ockey's study (2007). In his study, low-proficient learners reported being slightly distracted by still images when one image switched to the other. As for their perceived usefulness of video, learners reported being over-emphasized on video, so they might overlook the listening task and jot down insufficient information from the talk. In addition to the discrepancy, the findings of the present study were also aligned with other findings in Ockey's study, particularly on the utilization of visuals. For example, low-proficient learners could predict that the speaker was about to say an important concept when they raised their hands. The finding corroborated that low-proficient learners might be able to utilize nonverbal information to mitigate their listening comprehension.

Findings from the first two research questions have illuminated the significant main effects of the two variables (i.e., visual formats and L2 proficiency) on learners' listening comprehension. The significant main effect of visual formats on listening comprehension was evidenced, which suggested that dynamic video has consistently led to increased performance compared to sequential static imagery, irrespective of L2 proficiency. Likewise, the significant main effect of L2 proficiency was also confirmed, indicating that high-proficient learners have exhibited substantially better performance in both visual formats.

### 5.3 The Interaction Effect between Visual Formats and Learners' L2 Proficiency

Despite the two variables having significant main effects, their efficacy did not interact with each other. The quantitative results did not show the interaction effects of the two variables. Such findings indicated that different proficient learners' listening comprehension might not depend on the visual formats they received. To elucidate the potential factors underlying the insignificant finding, it is imperative to conduct a thorough examination of the listening performance exhibited by learners of varying proficiency levels when exposed to the two visual formats. As Figure 5 suggested, the disparity in listening comprehension between the two visual formats was more pronounced among low-proficiency learners than their high-proficiency counterparts. In other words, low-proficient learners demonstrated a more significant improvement in dynamic video conditions than high-proficient learners. Although no statistical interaction effect was observed in the present study, the lines, as indicated by Figure 5, associated with high-proficient and low-proficient learners were not parallel. Such differences can be viewed as the source of the significant interaction. If low-proficient learners receive more exposure to dynamic video or more comprehension questions are included, it might be possible to observe an interaction effect.

A similar phenomenon could be found in Ginther's study (2002), which reported an interaction between participants' language proficiency and visual stimulus types. In his study, the interaction between participants' language proficiency and visual stimulus types was statistically significant. However, the line chart could not directly observe the interaction effect. Ginther then interpreted the interaction effect by using the unparalleled lines associated with the two visual stimulus types and concluded that the extended line would reach interaction. In line with Ginther, Sueyoshi and Hardison (2005) also asserted that different visual stimulus conditions have an interaction effect with proficiency level.

While the two variables in the present study have no significant interaction effects, the prominent disparity among low-proficient learners' performance in two visual formats has suggested that low-proficient learners' improvement in dynamic video was more than their high-proficient counterparts. As low-proficient learners gain more experience and strategies in processing dynamic video, their listening performance might one day overtake high-proficient learners. In this way, the interaction effect might be observed.

### 5.4 Learners' Perceived Cognitive Load

The last research question probed into the qualitative data on participants' perceived cognitive load: *intrinsic*, *extraneous*, and *germane cognitive load* among the four listening conditions. In the ensuing paragraphs, the major findings in this section will be summarized and followed by the possible interpretation of the findings.

As Table 7 suggested, low-proficient learner demonstrated higher perceived intrinsic cognitive load than their high-proficient counterpart. High intrinsic cognitive load is derived from the inherent complexity of the concept (Sweller, 1999). In the present study, the complex concept was associated with the dialogue content, language use, and speech rate according to the statements listed in the intrinsic cognitive load sector. In other words, low-proficient learners struggled with these components. The higher rating in this section also asserted that low-proficient learners might not maintain much cognitive capacity for processing non-linguistic information.

As for the perceived extraneous cognitive load, learners in dynamic video groups yielded higher ratings than those in sequential static imagery groups. Extraneous cognitive load derived from the material design (Sweller, 1999). In the present study, the material design was the manipulation of different visual formats in accompaniment with the concurrent presentation of auditory information. The finding implied that dynamic video might be an optimal material design for multimodal listening tasks. In addition to the overall rating, another interesting picture was in statement 6, particularly in the sequential static imagery group. In this statement, learners were asked whether the nonverbal cues, such as facial expressions and gestures, were effective for their listening comprehension. In this section, low-proficient learners' in the sequential static imagery group rated higher than higher-proficient learners. This result suggested that even though sequential static imagery might not be the best option, the nonverbal information embedded in each static image was still helpful for low-proficient learners. In other words, if static images were the only visual format available, select images that contained a variety of nonverbal features could still provide learners, especially low-proficient learners with effective visual enhancement.

Lastly, germane cognitive load referred to the mental exertion and motivation that was induced by the material design (Mayer, 2005). In the present study, both visual formats induced high perceive germane cognitive load, with dynamic video eliciting slightly higher rating. High perceived germane cognitive load might stimulate deeper mental engagement in processing multimodal inputs. Accordingly, learners in video groups performed slightly better in inferential comprehension questions.

# CHAPTER SIX: CONCLUSIONS

## 6.1 Summary of the Major Findings

The present study investigated the effectiveness of different visual formats (i.e., dynamic video vs. sequential static imagery) on L2 learners with different language proficiencies (i.e., high-proficient vs. low-proficient). The effect of visual formats was determined by participants' listening performance, the cognitive load (i.e., *intrinsic*, *extraneous*, and *germane*) derived from the visual formats, and participants' overall perception of the visuals formats. The major findings of the current study are listed as follows:

- Visual formats and L2 proficiency both significantly affected participants' listening comprehension. In particular, L2 proficiency yielded a larger effect size than visual formats, indicating that participants' L2 proficiency might be more impactful than visual formats in their listening performance. For example, high-proficient learners consistently demonstrated better performance than low-proficient learners, regardless of visual formats.
- The two variables had no significant interaction effect, suggesting that different proficient learners' listening performance might not depend on the type of visual formats they received.
- Low-proficient learners tended to report higher *intrinsic cognitive load* than high-proficient learners due to the inherent complexity of the linguistic inputs. Participants in sequential static imagery rated higher *extraneous cognitive load* than those in dynamic video groups, suggesting that dynamic visuals might be more suitable for multimodal listening assessment. Lastly, all four experimental groups rated high *germane cognitive load*, indicating that the design could successfully induce their mental efforts while listening.

The findings above ensured that different proficient L2 learners' listening performances vary across visuals. High L2 proficiency might lead to better listening performance and report lower perceived *intrinsic cognitive load*. While low-proficient learners generally demonstrated inferior listening performance, they still reported positive attitudes and have paid a close attention to different nonverbal information. As for the role of visual formats, dynamic video yielded lower perceived *extraneous cognitive load, and* higher *germane cognitive load*. These findings could serve as a reference for instructors when developing multimodal listening assessments or applying multimodal materials in teaching L2 listening.

### 6.2 Pedagogical Implications

When teaching L2 listening ability, instructors tend to develop materials that are authentic and aligned with real-life communication settings. With the increased availability of multimodal resources, visuals might be one of the most common materials applied for pedagogical purposes. To incorporate visuals that may enhance rather than disrupt their listening comprehension, instructors should consider learners' L2 proficiency and their interaction with different arrays of visual formats. Additionally, instructors should consider whether learners can maintain cognitive resources while processing multimodal inputs. In this vein, the following paragraphs will provide implications on the premise of the findings of the current study.

#### 6.2.1 Using video with level-appropriate spoken text to facilitate listening comprehension

The implementation of multimodal listening tasks required the simultaneous presentation of auditory and visual information. For L2 instructors, selecting visuals that could effectively enhance learners' listening performance and reduce extraneous cognitive load might be important. In order to attain the research objective, the study identified dynamic video as the plausible option. As suggested in the present study, learners in dynamic video conditions, regardless of proficiency, obtained notably higher scores than those in sequential static imagery conditions. Significant gains in dynamic video conditions was accompanied by low perceived *extraneous cognitive load*, indicating that the incorporation of dynamic video could enhance listening performance without escalating their perceived e*xtraneous cognitive load*. Given the supportive evidence, L2 instructors could consider incorporating dynamic video in designing multimodal listening tasks.

For instructors to optimize the effect of dynamic video, particularly for low-proficient learners, incorporating level-appropriate spoken texts in accompaniment with dynamic video is important. Findings of the present study indicated that low-proficient learners reported higher intrinsic cognitive load than high-proficient learners. Since intrinsic cognitive load is derived from the inherent complexity of materials, low-proficient learners might find linguistic information difficult to understand. When learners grapple with linguistic information beyond their proficiency, they retain few cognitive resources for processing visual information. The lack of sufficient cognitive capacity for visual information might preclude their opportunity to utilize visuals to compensate for their limited linguistic knowledge. Therefore, to optimize the benefits of visuals, particularly for low-proficient learners, instructors should select level-appropriate spoken text when designing multimodal listening tasks. Such implication also aligned with Mayer's guideline (2010) for designing multimedia materials, indicating that *intrinsic cognitive load* could be managed by providing level-appropriate materials.

In language programs with homogenous learner backgrounds (e.g., L2 proficiency level), instructors were encouraged to incorporate linguistic information that was suitable for the particular group of learners. If learners could easily process linguistic inputs, they could be able to reserve their cognitive resources for processing non-linguistic inputs. However, in the majority of local contexts, a mixed level of learners would be arranged in the same classroom. In such a context, presenting tailored materials for learners at different levels during class time might be challenging. A feasible method was to manipulate the sequence of presenting materials. Instructors were advised to consider the following approaches.

First of all, to manage low-proficient learners' intrinsic cognitive load, instructors could give an overview of the listening passage topic before presenting the video. The brief introduction of the listening passage might activate learners' prior knowledge. With the activation of existing knowledge, learners might draw on their mental representation to comprehend the incoming inputs and utilize *top-down processing* to mitigate their insufficient *bottom-up* decoding strategy. While presenting the multimodal listening tasks, learners might be able to handle the multimodal inputs without being overwhelmed by the complex linguistic information. This approach might allow instructors to employ the same listening passage to the entire class but still ensure the comprehensibility of linguistic inputs and the efficacy of visual information.

Secondly, instructors could also present the multimodal listening tasks several times with different purposes. For the first time, instructors could play the dynamic video in silent mode and require learners to focus solely on the visual information in the video. The reduction of auditory mode might allow learners to concentrate on the details in visual information. After watching the video for the first time, learners might have an overall understanding of the content of the listening tasks. In the second time, when both visual and auditory inputs were presented simultaneously, low-proficient learners might be able to shift their focus to linguistic information. This approach could direct learners' attention to non-linguistic information to establish their overall comprehension first. When receiving auditory and visual information, they could spend most of their attentional resources on processing auditory inputs. This study explored manipulating the multimodal listening tasks (through simplification or modification) as a strategy for effective implementation within mixed-level instructional settings.

### 6.2.2 Guiding learners' attention to various nonverbal cues while listening

Real-life communication is a multimodal discourse. In the majority of communicative events, listeners were able to see the speakers. They constructed their understanding of the communication via both verbal (e.g., auditory inputs) and nonverbal inputs (e.g., gesture, facial expression). Although auditory information constituted an integral component of the entire communicative event, a comprehensive apprehension of the communication necessitated the integration of both verbal and nonverbal cues. In other words, instructors were encouraged to integrate multimodal listening tasks and guide learners in focusing on the underlying meanings of nonverbal information.

The findings of the present study indicated that participants devoted their attention to different nonverbal cues in the videos. They reported a major focus on speakers' faces, including their eyes and lips. Learners could interpret the speakers' emotions by gazing at the speakers' eyes, and mitigate their comprehension through lip reading. In addition to facial areas, participants also revealed their purposes of gazing at speakers' gestures and the objects that appeared in the contextual settings. Based on the findings of the present study, instructors were encouraged to integrate these nonverbal cues in designing multimodal listening tasks. When presenting the materials, instructors could guide learners to be consciously aware of the message carried by the nonverbal information. Since the meaning of nonverbal information varies across different communication contexts, it is necessary for instructors to highlight the interconnection between verbal and nonverbal messages. With conscious awareness of the relationship between verbal and nonverbal information, learners could yield better listening performance in a multimodal environment. Additionally, the development of multimodal decoding skills may facilitate the transfer of this ability to their communication practices.

### 6.2.3 Incorporate visual components in listening assessment

The present study and many existing studies have jointly indicated the positive impact of visuals, particularly dynamic video, in enhancing L2 learners' listening comprehension. Based on the aforementioned research findings, it is recommended that L2 instructors integrate visual elements, particularly dynamic video content, into their assessment of learners' listening comprehension skills. When selecting videos for listening assessment, instructors can consult the selection criteria in the present study to avoid integrating videos that may increase learners' cognitive burden. In the development of multimodal listening assessments, it is advisable for instructors to utilize Rost's listening assessment model (2016) as a framework for conceptualizing the listening construct. With a comprehensive understanding of this construct, instructors can then operationalize test items based on their expertise as assessment developers. When listening assessments are designed to be multimodal and align with authentic communication contexts, instructors can draw more valid inferences regarding learners' real-world listening performance beyond the assessment environment.

## 6.3 Limitations and Suggestions for Future Direction

The present study has shed light on the incorporation of visuals in multimodal listening, both theoretically and pedagogically. In the present study, the effect of L2 proficiency and visual formats on learners' listening performance was being examined. Notably, the contributing factors of learners' listening performance were not limited to these variables, and a single administration experiment might refrain from interpreting other long-term effects. Accordingly, the results of the present study should be interpreted with caution. Some limitations are identified in the ensuing paragraphs.

First, the present study solely compared the efficacy of two visual formats (i.e., dynamic video vs. sequential static imagery) and found that low-proficient learners consistently yielded lower listening performance than high-proficient learners, irrespective of visual formats. However, the absence of performance data from the audio-only listening conditions may refrain researchers from determining whether low-proficient learners' listening performance supported by the two visual formats surpassed their performance without visual aids. The additional comparison between visual-mediated and audio-only conditions might strengthen the rationale for the incorporation of visuals, in listening assessment, particularly for the inclusion visuals which could effectively enhance low-proficient learners' listening performance. In sum, future study is encouraged to add a control group, which is an audio-only listening condition, in the research design to provide a comprehensive picture of the effectiveness of visuals.

Secondly, the findings of the present study indicated only individual main effect, but no interaction effect between L2 proficiency and visual formats. Even though there was no interaction effects, learners' performance data revealed a tendency that low-proficient learners gained more significant improvement in dynamic video conditions.

The improvement was expected to increase as learners receive more exposure to dynamic video, and the disparity between high and low-proficient learners might be offset in the long run. Since the present study was a short-term research with only a single exposure of visual stimuli, future study is encouraged to conduct a long-term experimental design with multiple exposure to the visual stimuli. The long-term study might increase the possibility of an interaction effect between the two variables.

Third, the present study, limited by methodological shortcomings, could only infer participants' visual utilization from their responses to the questionnaire. Although most participants reported their visual utilization patterns to various extants, such an indirect interpretation might overgeneralize their actual usage of nonverbal information. To formulate valid conclusions regarding learners' visual attention patterns, future research can incorporate comprehensive retrospective verbal protocols. These in-depth interviews will facilitate participants' recollection of their viewing behaviors during multimodal listening assessments, thereby enhancing the robustness of data interpretation. Collecting and analyzing their listening performance data, perceived cognitive load rating, and verbal reports may provide a holistic view of participants' visual utilization.

# REFERENCES


Anderson, J. R. (2000). *Cognitive psychology and its implications*. (5th ed.). Worth Publishers.

Aryadoust, V., Foo, S., & Ng, L. Y. (2022). What can gaze behaviors, neuroimaging data, and test scores tell us about test method effects and cognitive load in listening assessments? *Language Testing*, *39*(1), 56-89. https://doi.org/10.1177/02655322211026876

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Batty, A. O. (2014). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing, 32*(1), 3–20. https://doi.org/10.1177/0265532214531254

Batty, A. O. (2020). An eye-tracking study of attention to visual cues in L2 listening tests. *Language Testing*, *38*(4), 511–535. https://doi.org/10.1177/0265532220951504

Baltova, I. (1994). The impact of video on the comprehension skills of core French students. *Canadian Modern Language Review, 50*(3), 507–531. https://doi.org/10.3138/cmlr.50.3.507

Brett, P. (1997). A comparative study of the effects of the use of multimedia on listening comprehension. *System*, *25*(1), 39–53. https://doi.org/10.1016/s0346-251x(96)00059-0

Buck, G. (2001). *Assessing listening*. Cambridge University Press. https://doi.org/10.1017/cbo9780511732959

Campoy-Cubillo, M. C., & Querol-Julián, M. (2015). Assessing multimodal listening. In Camiciottoli, B. C., & Fortanet-Gómez, I. (Eds.), *Multimodal Analysis in Academic Settings* (pp. 193-212). Routledge.

Campoy-Cubillo, M. C. (2019). Functional diversity and the multimodal listening construct.*European Journal of Special Needs Education*, *34*(2), 204-219. https://doi.org/10.1080/08856257.2019.1581402

Chi, C., Chen, H. J. H., Tseng, W. T., & Liu, Y. T. (2023). Efficacy of different presentation modes for L2 video comprehension: Full versus partial display of verbal and nonverbal input. *ReCALL, 35*(1), 105-121. https://doi.org/10.1017/s0958344022000088

Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: a case study. *System, 29*(1), 1–14. https://doi.org/10.1016/s0346-251x(00)00057-9

Cross, J. (2011). Comprehending news videotexts: The influence of the visual content. Language, *Learning and Technology, 15*(2), 44–68

Education Testing Service. (2022). *TOEIC Examinee handbook: Listening and reading test.*

Education Testing Service. (2023). *TOEIC Listening and reading test official test preparation guide* (Vol. 7).

Field, J. (2008). Revising segmentation hypotheses in first and second language listening. *System*, *36*(1), 35-51. https://doi.org/10.1016/j.system.2007.10.003

Fries, C. C. (1945). Teaching and Learning English as a foreign language. *University of Michigan Press*

Furness, E. L. (1952). Techniques for the teaching of listening. *The Modern Language Journal, 36(3)*, 124–128. https://doi.org/10.1111/j.1540-4781.1952.tb01412.x

Ginther, A. (2001). Effects of the presence and absence of visuals on performance on TOEFL® CBT listening-comprehension stimuli. *ETS Research Report Series*, 2001(2), i-43.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing, 19*(2), 133–167. https://doi.org/10.1191/0265532202lt225oa

Goh, C. C. (2000). A cognitive perspective on language learners' listening comprehension problems. *System, 28*(1), 55-75.

Goh, C. C., & Vandergrift, L. (2021). *Teaching and learning second language listening: Metacognition in action*. Routledge.

Gruba, P. (1993). A comparison study of audio and video in language testing. *JALT Journal, 15*(1), 85–88

Guichon, N., & McLornan, S. (2008). The effects of multimodality on L2 learners: Implications for CALL resource design. *System, 36*(1), 85-93. https://doi.org/10.1016/j.system.2007.11.005

Jewitt, C. (2013). Multimodal teaching and learning. In C.A. Chapelle (Ed.), *The encyclopedia of applied linguistics (pp.4109-4114).* Oxford: Blackwell Publishing Ltd.

Kang, T., Gutierrez Arvizu, M. N., Chaipuapae, P., & Lesnov, R. (2019). Reviews of academic English listening tests for non-native speakers. *International Journal of Listening*, *33*(1), 1–38. https://doi.org/10.1080/10904018.2016.1185210

Kamiya, N. (2022). The limited effects of visual and audio modalities on second language listening comprehension. *Language Teaching Research, 10*(2), 79-88. 13621688221096213. https://doi.org/10.1177/13621688221096213

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in psychology, 8*, 1-8. https://doi.org/10.3389/fpsyg.2017.01997

Lado, R. (1964). *Language testing: The construction and use of foreign language tests*. London, UK: Longman.

Lesnov, R. O. (2018). Content-rich versus content-deficient video-based visuals in L2 academic listening tests: pilot study. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT), 8*(1), 15-30.

Lesnov, R. O. (2022). Furthering the argument for visually inclusive L2 academic listening tests: The role of content-rich videos. *Studies In Educational Evaluation, 72,* 1-11. https://doi.org/10.1016/j.stueduc.2021.101087

Lee, P. J., Liu, Y. T., & Tseng, W. T. (2021). One size fits all? In search of the desirable caption display for second language learners with different caption reliance in listening comprehension. *Language Teaching Research, 25*(3), 400-430.

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T. & Van Merriënboer, J. J. G. (2013) Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods, 45*(4), 1058–1072

Liu, T., & Aryadoust, V. (2023). Does modality matter? A meta-analysis of the effect of video input in L2 listening assessment. *System, 120*, 1-18. https://doi.org/10.1016/j.system.2023.103191

Londe, Z. C. (2009). The effects of video media in English as a second language listening comprehension tests. *Issues in Applied Linguistics, 17*(1).

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research.*Journal of English for Academic Purposes, 10*(2), 79–88. https://doi.org/ 10.1016/j.jeap.2011.03.001

Mayer, R.E. (2001). *Multimedia learning.* Cambridge University Press.

Mayer, R. E. (2005). *Cognitive theory of multimedia learning*. In Mayer, R. E. (ed.), The Cambridge handbook of multimedia learning. Cambridge University Press, 31–48. https://doi.org/10.1017/CBO9780511816819.004

Mayer, R. E. (2009). *Multimedia learning*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511811678

Mayer, R. E. (2010). Techniques that reduce extraneous cognitive load and manage intrinsic cognitive load during multimedia learning. (R. Moreno, Ed.). In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 131-152). Cambridge University Press.

Messick, S. (1996). Validity and washback in language testing. *Language testing*, *13*(3), 241-256.

Newton, J. M., & Nation, I. S. (2020). *Teaching ESL/EFL listening and speaking.* Routledge.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*(4), 517–537. https://doi.org/10.1177/0265532207080771

Ockey, G. J., & Wagner, E. (2018). *Assessing L2 listening: Moving towards authenticity* (Vol. 50). John Benjamins Publishing Company.

Parry, T., & Meredith, R. A. (1984). Videotape vs. Audiotape for listening comprehension tests: *An experiment* (ED254107). ERIC. https://files.eric.ed.gov/fulltext/ED254107.pdf.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1-4. https://psycnet.apa.org/doi/10.1207/S15326985EP3801_1

Pusey, K., & Lenz, K. (2014). Investigating the interaction of visual input, working memory, and listening comprehension. *Language Education in Asia*, *5*(1), 66-80.

Rost, M. (2016). *Teaching and researching listening* (3rd ed.). Pearson Longman.

Shin, D. (1998). Using videotaped lectures for testing academic listening proficiency. *International Journal of Listening*, *12*(1), 57–80. https://doi.org/10.1080/ 10904018.1998.10499019

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661–699. https:// doi.org/10.1111/j.0023-8333.2005.00320.x

Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. In C. A. Chapelle, H. G. Jun, & I. Katz (Eds.), *Developing and evaluating language learning materials*, 53-68. Ames, IA: Iowa State University.

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing, 32*(4), 463–483. https://doi.org/10.1177/0265532214562099

Suvorov, R., & He, S. (2022). Visuals in the assessment and testing of second language listening: A methodological synthesis. *International Journal of Listening*, *36*(2), 80-99. https://doi.org/10.1080/10904018.2021.1941028

Sweller, J. (1999). *Instructional design in technical areas*. Camberwell. Victoria: ACER Press.

Terao, T. (2024). Computer-Based Listening Test with Full Video, Visual-Limited Video, and Audio: A Comparative Analysis Based on Difficulty, Discrimination Power, and Response Time. *Applied Measurement in Education, 37*(1), 29-42.

Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency? *The modern language journal, 90*(1), 6-18.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language teaching, 40*(3), 191-210. https://doi.org/10.1017/s0261444807004338

Wagner, E. (2007). Are they watching? Test-Taker viewing behavior during an L2 video listening test. Language, *Learning and Technology, 11*(1), 67–86.

Wagner, E. (2008). Video listening tests: What are they measuring?. *Language Assessment Quarterly, 5*(3), 218-243.

Wagner, E. (2010a). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing, 27*(4), 493–513. https://doi.org/10.1177/0265532209355668

Wagner, E. (2010b). Test-takers' interaction with an L2 video listening test. *System, 38*(2), 280–291. https://doi.org/10.1016/j.system.2010.01.003

Wagner, E. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Language Assessment Quarterly, 10*(2), 178-195.https://doi.org/10.1080/15434303.2013.769552

Weir, C.J. and Vidakovic, I (2013). The measurement of listening ability 1913-2012. In C. Weir, I. Vidakovic & E. Galaczi (Eds.) *Measured constructs: a history of the constructs underlying Cambridge English language (ESOL) examinations 1913-2012: Studies in Language Testing 37 (pp. 347-419)*. Cambridge: Cambridge University Press.

## Appendix A: Listening Comprehension Questions

1. What is the conversation mainly about?
   (A) The budget for the next meeting
   (B) The rumors about budget reduction
   (C) The scariest experience in one's life
   (D) The new member in the accounting sector

2. Which place do these people work in?
   (A) They work at a detective agency.
   (B) They work at an accounting office.
   (C) They work for the city government.
   (D) They work for a media company.

3. Which of the following statements is true about the talk?
   (A) Mark is the only person being fired.
   (B) Mark is a member of the video sector.
   (C) People are ready to work in other places.
   (D) People are gossiping about the criminals.

4. What is the purpose of the meeting?
   (A) To review every member's updated resume
   (B) To discuss about members' work experience
   (C) To announce that all members will get raises
   (D) To assign new tasks for members to complete

5. What might happen to the staff after the meeting?
   (A) They will get raises next month.
   (B) They will all leave the company.
   (C) They will start doing the new tasks.
   (D) They will join the accounting sector.

6. What is the conversation mainly about?
   (A) Writing a book at a cafe
   (B) Preparing for a job interview
   (C) Meeting new people in the city
   (D) Finding a new partner for works

7. According to Ms. Weaver, what is Anna good at?
   (A) Writing a book
   (B) Giving job advices
   (C) Working with others
   (D) Driving the helicopter

8. Why does Anna think that Pete is the right person?
   (A) Because Pete is busy with his work
   (B) Because Pete is very different from her
   (C) Because Pete is reading the job postings
   (D) Because Pete is good at playing puzzles

9. Which of the following statements is true about Pete?
   (A) He enjoys staying in the cafe.
   (B) He loves to stay away from people.
   (C) He loves to eat during the interview.
   (D) He enjoys listening to other people's talk.

10. After knowing that Pete gets the job, what are Pete and Anna going to do next ?
   (A) They will book a restaurant to celebrate.
   (B) They will meet Ms. Weaver at her office.
   (C) They will discuss their new assignments.
   (D) They will find new jobs in the newspaper.

11. What is the conversation mainly about?
   (A) A short trip around the city
   (B) A fun fact about the tour bus
   (C) The history of the monument
   (D) The layout of the White House

12. How does Penelope decide the tour she wants to attend?
   (A) She refers to the travel books.
   (B) She chooses between two flyers.
   (C) She listens to Anna's suggestion.
   (D) She asks a young tour guide for help.

13. What was the DC duck originally used for?
   (A) To carry visitors in the city
   (B) To serve on the battlefield
   (C) To honor a former president
   (D) To maintain the city's security

14. Which of the following statements is true about the boy?
   (A) He is looking for a restroom.
   (B) He is also a tourist in the city.
   (C) He is good at telling funny things.
   (D) His father works in the White House.

15. After getting the tips, what is the boy most likely to do next?
   (A) He will ask for more tips.
   (B) He will stop sharing the fun facts.
   (C) He will return the tips to the women.
   (D) He will continue to share the fun facts.

16. What is the conversation mainly about?
   (A) A solution on how to protect bees
   (B) A lesson on how to interact with bees
   (C) The reasons why Anna is afraid of bees
   (D) The introduction on how to get fresh honey

17. Which of the following statements is true about Caroline?
(A) She is a friend of Anna.
(B) Her kids are also afraid of bees.
(C) She used to be stung by the bee.
(D) She raises the bees in her backyard.

18. Why does Anna get stung by the bees?
(A) Because she climbs up a tree
(B) Because she wears a colorful jacket
(C) Because she attacks the bees' habitat
(D) Because she gets too much honey out

19. Which of the following items does not belong to protective gear?
(A) A veil
(B) A suit
(C) A pair of boot
(D) A pair of gloves

20. After knowing that Anna is ready, what will Caroline most likely do next?
(A) She will ask Anna to open up the frame.
(B) She will ask Anna to spray more smoke.
(C) She will ask Anna to get close to the bees.
(D) She will ask Anna to put on safety clothing.

21. What is the conversation mainly about?
(A) The vacation plan of the singer
(B) The release of a singer's new album
(C) The interview with a famous rock singer
(D) The introduction about the singer's next concert

22. How does Emma G react to Anna's video scrapbook?
(A) She agrees to sign her signature on it.
(B) She questions the right to use the photos.
(C) She refuses to share it with the audiences.
(D) She appreciates Anna's efforts in making it.

23. Which of the following statements is true about Emma G's schedule?
(A) She will have a concert in July.
(B) She will drive six hours to meet Anna.
(C) She will have lunch with Anna's sister.
(D) She will teach at the music camp Chicago.

24. What happened to Emma G. while she was singing a song?
(A) She noticed that the song was not the one she wrote.
(B) She thought that the background music was too loud.
(C) She found that Anna was holding the wrong instrument.
(D) She insisted that Anna should not sing together with her.

25. What can we infer from the conversation?
    (A) Anna is not a fan of Emma G.
    (B) Anna’s sister lives near Chicago.
    (C) Anna is going to join Emma's concert.
    (D) Anna has spent six hours practicing singing.

26. What is the conversation mainly about?
    (A) An interview with Anna’s colleagues.
    (B) An introduction on disease prevention.
    (C) A report about Anna’s health examination.
    (D) A talk about everyone’s pursuit of a dream.

27. Why did Anna think that she might be dying?
    (A) She was infected by a fatal disease.
    (B) She was afraid of visiting the doctor.
    (C) She failed to check her blood test result.
    (D) She learned the message from the doctor.

28. Which of the following statements is true about Penelope’s dream?
    (A) She would like to be a nurse.
    (B) She would like to be a dancer.
    (C) She would like to be a dog trainer.
    (D) She would like to be a book writer.

29. Why did Pete get a call from his agent?
    (A) Because he was late for an audition.
    (B) Because he did not sleep well last night.
    (C) Because he needed to go on a business trip.
    (D) Because he booked a spaceship in Hollywood.

30. What can we infer from the talk?
    (A) Penelope is going to join a dance audition.
    (B) Ashley is going to bring her dog to a contest.
    (C) Anna is going to host the rest of the radio show.
    (D) Jonathan is going to post his bucket list on Twitter.

## Appendix B: Questionnaires

**基本資料**

性　別 □ 1.男 □ 2.女 □ 3.其他__________

學習英文的時間 □1.小於 3 年 □2. 3 到 6 年 □3. 6 年以上

學測英文聽力級分 □1. A 級 □2. B 級 □3. C 級或其他

英文檢定最佳成績 ___________(檢定名稱)__________(級數/成績)

**第一部分：**

請您就認同程度與自身感受，在 1 到 10 圈選出一個適當的數字，數字越大表示同意程度越高。(1=非常不同意; 10=非常同意)

1. 我覺得影片中對話的主旨難以理解

1 2 3 4 5 6 7 8 9 10

2. 我覺得影片中使用的英文字詞過於困難

1 2 3 4 5 6 7 8 9 10

3. 我覺得影片中說話者的語速太快難以跟上

1 2 3 4 5 6 7 8 9 10

4. 我覺得影像呈現有助於我理解對話發生的情境

1 2 3 4 5 6 7 8 9 10

5. 我覺得影像與音檔同時呈現有助於我理解對話內容與作答

1 2 3 4 5 6 7 8 9 10

6. 我覺得表情、手勢等非語言資訊有助於理解對話與作答

1 2 3 4 5 6 7 8 9 10

7. 我覺得我有努力去理解測驗中的影像以及聽力音檔

1 2 3 4 5 6 7 8 9 10

8. 我觀看影片的目的是想正確的理解所有內容並作答正確

1　2　3　4　5　6　7　8　9　10

9. 我覺得影片有涵蓋助於我理解對話內容與作答的資訊

1　2　3　4　5　6　7　8　9　10

**第二部份：**

1.　在測驗過後，你對於影像融入聽力測驗的看法是?

□1. 有幫助　□2. 使我分心　　□3. 其他__________

2. 在測驗過程中，請問你是否會關注說話者的臉部?

□1. 會關注　□2. 不會關注　□3. 其他___________

3. (承上題) 在眾多臉部特徵中，你最關注下列哪一項?

□1. 眼睛　　□2. 嘴唇　　□3. 其他___________

4. (承上題) 請問你認為透過觀察上述臉部特徵可以獲得哪些關於對話的資訊?

______________________________________________________

5. 除了臉部特徵，你是否會關注說話者所使用的手勢?

□1. 會關注　□2. 不會關注　□3. 其他_______

6. (承上題) 請問你認為透過手勢可以獲得哪些關於對話的資訊?

______________________________________________________

7. 在測驗過程中，請問你是否會關注對話場景中所出現的物品?

□1. 會關注　□2. 不會關注　□3. 其他_______

8. (承上題) 請問你覺得透過觀察物品可以獲得哪些關於對話的資訊?

______________________________________________________