

Abstract

Previous studies concluded that first-time film viewers (Schwan & Ildirar, 2010; Ildirar & Schwan, 2015) often had difficulty integrating shots into a coherent representation of the depicted events in the absence of a familiar action through the film cuts or a salient eye-gazing of a character in the film. In this study we investigated whether diegetic sound (i.e. sound that seems to originate from the depicted cinematic space) could effectively bridge shots for first-time viewers. Across a range of films, both dialog, and salient environmental sound (e.g., barking dogs) helped first-time viewers connect shots. However, sound was not always successful in supporting first-time viewers' interpretations. Whilst experienced viewers were able to understand less familiar linking sounds and environments, first-time viewers found this difficult. Overall, a range of diegetic sounds helped first-time viewers understand spatio-temporal relations between shots but these viewers still had difficulty integrating views of unfamiliar environments. (word count:148)

Keywords: continuity perception, film literacy, first-time viewers, diegetic sound, role of audio

Audio Facilitates the Perception of Cinematic Continuity by First-time Viewers

One of the most important questions organizing research on media is the degree to which understanding mediated communication depends on medium-specific skills vs. everyday perceptual and cognitive skills. The former hypothesis is often associated with research in the humanities, but it is also prevalent in the field of communications and, to a degree, within cognitive science. According to this hypothesis, media such as cinema present meaningful events using formal structures that differ qualitatively from everyday perception. Therefore, the relationship between specific formal structures in cinema and the events they represent must be learned instance-by-instance, much as the individual words making up a language must be mapped to their references (for review see Frith & Robinson, 1975; Lowe & Durkin, 1999; Smith, Anderson, & Fischer, 1985). Clearly, such a process would require extensive learning. The more naturalistic view of cinema is that it relies on existing perceptual skills. On this view, structural differences between cinema and everyday perception rarely interfere with understanding because the natural context for perceptual processes forced them to be robust over the kinds of variations induced by cinema. Accordingly, cinema requires little specific learning to understand (see for example, Messaris, 1994; Anderson, 1996; Levin & Simons, 2000; Smith, Levin, & Cutting, 2012).

There are several means of assessing the role of learning in cinema. One is to determine whether young children and infants, who presumably have had relatively few opportunities to learn cinema-specific codes, can understand films. Research documenting this understanding is organized by the broad hypothesis that younger infants' processing of films is primarily stimulus-driven (e.g. attention is drawn by movement and sudden changes) and that development is associated with more top-down control as the child matures cognitively and acquires general world knowledge as well as specific knowledge about formal features of film (i.e. gains film literacy; Levin & Anderson, 1976; Anderson, Lorch,

Field, & Sanders, 1981; Crawley, Anderson, Wilder, Williams, & Santomero, 1999; Lemish, 1987; Richards & Gibson, 1997). Recent research has documented several of these late-developing understandings. For example, children younger than 18 months are insensitive to sequential and linguistic comprehensibility in edited sequences (Richards & Cronise, 2000; Pempek, Kirkorian, Richards, Anderson, Lund, Stevens, 2010). Other research has documented a "video deficit" whereby infants and toddlers sometimes learn better from real-life experiences than from video (Troseth & Deloache, 1998; Anderson & Pempek, 2005; Hayne, Herbert & Simcock, 2003). Barr, Zack, Garcia, & Muentener (2008) argue that developmental achievements such as these are associated with four factors: age, prior exposure to content, parent-infant interaction style and formal features (see Barr, Zack, Garcia, & Muentener, 2008 for review). However, research testing whether children understand medium-specific formal features is rare, and primarily limited to the auditory domain (such as music, vocalizations, and sound effects; Somanader, Garcia, Miller, Barr, 2005) or to non-structural visual techniques (such as animation, and visual special effects, action, and pacing, Calvert, Huston, Watkins, & Wright, 1982). Furthermore, most infant studies (e.g. Richards & Cronise, 2000; Pempek et. al., 2010) use existing program material in which the use of such techniques is confounded with familiarity of narrative contexts. Systematic research on older children's (4+ years) understanding of cinematic techniques suggests that complex formal features (e.g. flashbacks, crosscutting) are first understood later, in comparison to more simple features (e.g. shot reverse shots, POV Shots; Abelman, 1989; Lowe & Durkin, 1999; Munk, Rey, Diergarten, Nieding, Schneider, & Ohler, 2012). However these studies do not clarify whether increases in comprehension of formal features are due to age-related increases in experience with the medium, or to general cognitive development.

Therefore, cross-cultural research documenting cinema understanding in the few populations of adults who lack experience is particularly interesting. Initial studies in these populations have demonstrated that inexperienced adult viewers can effectively understand *familiar* edited narratives as effectively as narratives that lack edits (Hobbs, Frost, Davis & Stauffer, 1988). However, other, more recent work has documented that these naïve viewers are sometimes limited in their understanding of edited films. Schwan and Ildirar (2010) recently identified a group of adults in the mountains near Isparta Turkey who, due to their isolation and the relative recency of electrification in their community, had seen no television or films. These first-time viewers watched a series of short films and exhibited only a limited understanding of a substantial number of simple techniques, including shot-reverse-shots, establishing shots, and point-of-view shots. For example, a shot-reverse-shot sequence of a man looking right, followed by a shot of a man looking left, with both actors shown against the same scenic background (Figure 1) was not interpreted as ‘Two men looking at each other’ but, instead, as two completely independent scenes: ‘First, there was a man, then he was gone, and then, another man appeared.’ On the other hand, first-time viewers easily understood films that depicted familiar activities using complex formal features such as ellipses (skipping of time segments), and cross-cutting between two simultaneous events taking place in different locations. One major conclusion drawn from the data was that first-time viewers can sometimes integrate shots using familiar narratives but cannot draw upon the full range of cognitive and perceptual view-linking skills that more experienced viewers possess. The findings of Schwan and Ildirar (2010) draw attention to the mechanisms of connecting adjacent shots in a semantically meaningful manner, which in turn can be considered an important prerequisite for establishing a coherent mental representation of the film’s content.

[Figure 1 here]

In addition to their difficulties with shot reverse shots, most of the first-time viewers failed to understand conceptual relations between shots such as switches from outside to inside views or from long shots to close-ups. Although this finding implies that medium-specific experience helps viewers to draw the fundamental conclusion that disparate views show a single scene, it is important to note that the failure of first-time viewers to understand some of the films used by Schwan and Ildirar (2010) may have occurred because of the subtle nature of the view-linking cues in these films. In the film depicted in figure 1, the actors did not speak or move, eliminating the possibility that the views could be linked by a common dialog or continuing events. Indeed, follow-up work demonstrated that first-time viewers could integrate views when, for example, one actor handed an object to the other across a cut or when an actor looks up to a tree top across a cut (Ildirar & Schwan, 2015). Combined, these results suggest that first-time viewers may have difficulty integrating views when few actions support the integration, but can do so when this support is provided.

However, the nature of the necessary support remains incompletely understood. It is interesting to note that some salient visual and spatial between-view commonalities are, by themselves, inadequate. Clearly, the two depicted in Figure 1 share very similar backgrounds, and the actors were nominally looking off-screen in complementary directions (although given their lack of movement, it may not have been clear that they were actually targeting their gaze at any particular off-screen object). Also, adding establishing shots showing both actors in a single shot before the two individual shots did not facilitate integration (Ildirar & Schwan, 2015). The view-linking cues that have previously facilitated understanding in this population seem to rely less on a common spatial perceptual or spatial framework than on linking actions or more abstract conceptual links. This implies that the visual integration

necessary to understand cinema can be supported by nonvisual information, there is good precedence for arguing that audio is central in cinematic experience.

Historically, cinema for the most part lacked synchronized sound for the first 35 years of its development, but even the earliest inventors of the motion picture believed that the multimodal recreation of picture and sound was necessary to effectively reproduce real events (Coe, 1992). Showings of silent films almost always featured live music, starting from the first public projection of movies by the Lumière Brothers in 1895. From the beginning, music was recognized as essential in the United States as well. This emphasis on sound accompaniment was quite widespread. The early cinema of Brazil featured *fitas cantatas*: filmed operettas with singers performing behind the screen. (Parkinson, 1996). In Japan, films not only included live music but also the *benshi*, a live narrator who provided commentary and character voices (Standish, 2006). It is interesting to note that Georges Demeny, along with his close associate E.J. Marey developed one of the first practical applications of cinema by creating brief films showing actors producing individual phonemes with the aim using them to help teach deaf individuals to speak. In an important sense, then, the foundations of cinema sometimes went beyond the use sound as an adjunct by marrying picture and audio to create an "image of sound" (Gunning, 2001, p14).

Film scholars have also argued that that addition of sound to cinema facilitates the temporal guidance of visual attention, allowing audiences to more effectively perceive fleeting visual events (Chion, 1994). In addition, film scholars have argued that sound can help integrate events. For example, Chion (1994) argues that the most important function of film sound consists of binding the flow of images, both by bridging visual breaks spatially and temporally, and by establishing atmosphere (p.47). According to Chion, by adding sound, two shots seem "magically to fall into a linear time continuum". More concretely, silent cinema clearly depicted a world of sound where people's lips moved, and characters could

clearly hear one another (Raynaud, 2001). In this historical and theoretical context, current film practice is often predicated on the assumption that visual and auditory stimuli are mutually supporting. The well-known film editor Walter Murch goes so far as to state that "the mutual influence of sound and picture is inextricable".

This is also true for the everyday perception. Many real-world social interactions require integrating visual and linguistic information. Nonverbal cues such as body movements, head nods, hand-arm gestures, facial expressions, eye gaze, posture, and interpersonal distance as well as the lip movements that accompany speech sounds are very helpful for communication. As Goldin-Meadow (1999) suggests "gesture serves as both a tool for communication for listeners, and a tool for thinking for speakers" (p. 419). Research in communications demonstrates that memory is improved by converging audio and visual information (so long as it doesn't conflict; for review see Lang, 1995). Although the present experiment is more focused on view integration than memory, Lang's review does support the hypothesis that encoding (as measured by recognition memory) is facilitated by multiple audio and visual channels (see Meyerhoff & Huff, 2016, for a recent demonstration). Other well-known work in perception demonstrates strong multimodal effects. For example, in the McGurk effect, auditorily presented phonemes combine perceptually with visually presented mouth movements (McGurk, & McDonald, 1976). Perceptual combinations such as these can extend to simple events as in Shams, Kamitani, & Shimojo (2000) where the number of perceived visual onsets and offsets is influenced by the number apparently co-occurring tones. Findings such as these have led researchers to argue that intermodal interaction is common, both behaviorally and neurally (Shimojo & Shams, 2001). Recent studies in cognitive neuroscience provide evidence for links between language and action in the brain: Motor areas activated in speech production are also activated when listening to speech sounds (for a review, see Williams & Hagoort, 2007).

Especially relevant for present purposes, other recent research suggests that audio narrative and sound effects can, in theory, support visual view integration on various levels of processing. First, sound can quickly guide viewers' attention to certain elements in a scene (Iordanescu, Grabowecky, Franconeri, Theeuwes, & Suzuki, 2010; Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008), which may help viewers to integrate differing views of a scene presented in adjacent shots. Second, appropriate combinations of audio and visual information have been shown to foster local causal bridging inferences, which may compensate for coherence breaks in a film's stream of events (Tibus, Heier, & Schwan, 2013). Synchronised audio has also been shown to decrease awareness of cuts, suggesting an increase in the perception of continuity especially in combination with matched action across a cut (Smith & Santacreu, 2016). In support of these inferences, audio narrative and sound effects can invoke strong mental imagery (Rodero, 2012) that also may support visual view integration. Finally, recent research has demonstrated that infants can use the spatial and temporal information in sound effects to help track an object that momentarily disappears behind an occluder (Bremner, Slater, Johnson, Mason, & Spring, 2012). Basic findings such as these have been incorporated into more general theoretical frameworks that attempt to describe how the psychological system processes multimedia narratives (e.g. Cohn, 2016 and Bateman & Wildfeuer, 2014). For example, Cohn (2013) argues that integration of meaning in multimedia is similar to the synthesis of meaning from speech and gesture.

Applied to the films presented by Schwan and Ildirar (2010) and Ildirar and Schwan (2015), the above review suggests that there are several means by which sound could aid in establishing cinematic continuity. In their original shot-reverse-shot films first-time viewers understood that the films depicted a common space, but adding sound in the form of dialog could establish a rich perceptual and conceptual link between shots that would reinforce the conclusion that the characters occupied the space at the same time. Not only does language

link shots because the meaning of utterances establishes a correspondence between initial phrases and initial shots and subsequent phrases and shots, but it also establishes simpler correspondences by increasing pre- and post-shot similarity and allowing predictions and postdictions over time. For example, when one shot shows a character looking off screen and saying “Hello, how are you?” and this is followed by a shot showing a second character saying “Fine”, the two shots can be linked by knowledge about the typical greeting. The viewer can use the greeting in the first shot to predict the response in the second shot, and then once that prediction is fulfilled assume that the two shots show the same event. This kind of integration draws upon a relatively rich conceptual linkages (understanding of language, pragmatics, and typical rote conversation) and perceptual similarities (in that the added dialog increases the perceptual similarity of the clips). Thus, it seems plausible that naïve viewers will be able to draw upon this rich context to conclude that individual shots go together. Although this might not seem surprising, it would at a minimum be evidence against the hypothesis that the sudden view transitions characteristic of motion picture edits is a strong barrier to view integration in the absence of extensive learning.

However, not all of the elements characteristic of dialog need be present in an audio track, and it is possible that first-time viewers will be able to link shots using sparser audio support. This is particularly relevant for the outdoor-to-indoor films previously presented to this population in which a house was shown from the outside, followed by a view of someone inside the house. In such a situation a dialog would be atypical, but simpler audio support for integration could be effective. For example, someone outdoors could be seen calling out, followed by a shot of someone indoors while the call is repeated at a lower volume. So, in this study we presented viewers with brief sequences that varied in the amount of conceptual and perceptual support that audio sources provided. These varying levels of support were spread across three groups of films. The first group depicted familiar settings and included

sound in the form of language, either in the form of dialog, or a single speaker. The second group of films depicted familiar settings and included non-language sound effects such as animal sounds. A third group of films depicted unfamiliar environments and included non-language sound effects such as continuous water sound or a non-language human sound such as whistling. Films in the first group (familiar setting with spoken language), varied across three levels of support (see Figure 2). High support films included dialog in which each speaker could be seen in turn when greeting each other. Medium support films included human voices but not dialog. These films depicted two persons in two different locations, as one of them called the other. The called-to person did not react to the call. Low support films in the first group included only offscreen verbalizations, whose source was invisible. In contrast to the medium support films of this category, neither the calling person nor called-to was not shown. Finally, one potential limit to audio-derived support might be difficulty that first-time viewers have in representing the source of off-screen sounds. Therefore, we added two single-shot scenes in which a character can be heard speaking from off-screen.

The second group of films (familiar setting with non-language sound; see Figure 3) included medium- and low-support films. We reserved the classification of high-support for films including language because language includes rich grammatical, semantic, and pragmatic sources of information, much of which is missing in non-language sound effects. Films in the second group tested whether support can derive from audio that affords continuity in the absence of familiar conversational patterns. Critically, successful use of audio to integrate these sequences requires viewers to identify the source of the sound in one shot and then to link the continuing sound with a representation of the now-offscreen source in a second shot. It should be noted that some environmental sound was present in the earlier Schwan and Ildirar (2010) and Ildirar and Schwan (2015) films but it was not particularly salient. Medium support films in this group included salient environmental sound continuing

through the cut. Low support films in this group either did not present continuing sound or did not depict the source of the sound. In some of these films, the sound was continuous across views but its source was never shown, and in some, the source was depicted in one shot, but the sound associated with the source was played in the other shot. We classified such films as “low support” because viewers would need to rely on relatively abstract linking predictions (or postdictions) to combine views.

Finally, we tested integration in films that depicted unfamiliar environments (see Figure 4). Whereas most of the films described above depicted familiar environments and individuals from the Isparta Turkey area, we created two additional low-support films depicting less familiar environments in the US. These relied upon sound effects to create continuity and therefore tested the limits of viewers’ ability to integrate views.

In the present experiment, these three groups of films were shown to viewers by the first author on a visit to the Isparta area during the summer of 2013. A set of naïve viewers with little or no film-viewing experience was compared with a similar set of viewers from the same area who did have film viewing experience. We tested whether the addition of high, medium, and low levels audio support would help the first-time viewers integrate views as effectively as the more experienced viewers.

Method

Participants

Forty participants (23 female, 53–85 years old, $M = 66.35$ years) took part in the study. The experimental group (20 first-time viewers, 13 female, 53–85 years, $M = 71.65$ years) knew of the existence of television and had some abstract ideas about it, but had no prior direct experience with the medium. This group lived in small isolated houses in the mountains south of Isparta, Turkey that had only recently been connected to the electrical grid. None of the experimental group however had come into direct contact with any film

screening and lacked even abstract ideas about it. However, 4 of these participants had been in the Schwan and Ildirar (2010) study and three had been in Ildirar & Schwan (2015). Although these participants recalled serving in these experiments when asked, they could only remember that they had seen some pictures in the experiment and could not recall any of the specific events they had seen. As reported below, results did not change substantively when these viewers' data were excluded.

All first-time viewers had some photos (mostly head shots of their children or grandchildren) and four had radios that receive signals from a very limited area. Many assumed that television is a “visual radio” with programs that showed pictures of the people who speak or sing on the radio. Seven of the group was illiterate and the average years of schooling was 1.96 years.

The control group (10 female, 55–72 years, $M = 61.05$ years) were from a similar geographic and cultural background as the experimental group. Critically, however, these participants all had some experience with television. They spoke the same dialect and had a similar lifestyle as the experimental group, but with a little more accessibility to luxuries. Three of them were illiterate and average education level was 3.30 years. This control group was significantly younger than the experimental group, $t(38)=4.451$, $p<.001$, and significantly more educated than the first-time group, $t(38)=2.200$, $p=.034$.

Stimuli

We showed participants 20 film clips. Eighteen of the films depicted familiar environments, people, sounds, and/or dialog, while two depicted unfamiliar settings. The familiar-setting films were created in the region of Turkey where participants resided, while the unfamiliar environment films were created in Nashville, TN.

Of the eighteen familiar-setting films, eleven included human voice/voices (see Figure 2). Four of the voice films were classified as high-support dialog films. These depicted

brief dialogs between two persons, in which each said “hello” to the other in turn (in Turkish). The dialog films depicted two persons standing across from each other in a typical shot-reverse-shot sequence. Two of these films were edited to be consistent with the 180 degree rule (and therefore showed actors looking off-screen in opposing directions; films 1 and 2), and two of the dialog films (films 3 and 4) violated this rule (and therefore violated the eyeline, showing actors looking off-screen in the same direction; for review see Baker & Levin, 2015). Two of the voice films (films 5 and 6) provided medium levels of support: In initial shot, one speaker called out to another who could be seen in the next shot (and who did not respond). In these films, the person calling out was outside, and the person being addressed could in one case be seen in a non-overlapping view outside, and in another case indoors (in this case, the audio level was lowered during the indoor shot to simulate how the call would sound from that location). In both of these films, the call could be heard in both shots, although it was attenuated in the second shot, as is typical when depicting an off-screen sound coming from a non-proximal location. Five of the voice films (films, 7, 8, 9, 10 and 11) provided low levels of support. First three films did not depict the speaker at all but instead showed two shots, each depicting an exterior of a building followed by an interior including a person. In two of the films (films 7 and 8) the person was the specific person being addressed, and in one (film 9) it was a man who was presumably hearing a general call to prayer. The last two films in this category (films 10 and 11) were one-shot films, in which viewers see just one person saying “hello” to someone not visible in the scene.

[Figure 2 here]

In the second set of film clips no human voice was used (figure 3). Instead, sound effects such as animal sounds, environmental sounds or sounds produced by actions linked

views. In medium support films (films 12 through 15), these sounds continued across the cut and could be heard during each shot. In one of these films (film 12), there was a clear semantic relationship between the shots (the first depicted a man cutting wood, and the second depicted a person inside a house sitting in front of a fire). In the other three films (13, 14, and 15) the relationship was less specific but thematically consistent. In film 13, a wooded area was shown in the first shot, and a bird in a tree in the second shot, and in film 14 a farmhouse with a garden (long shot) was shown in the first shot and then a rooster in the second (close-up shot). In film 15, hens nominally making the sound were shown in front of a village house (outdoor shot), and the second shot showed the inside of the house (indoor shot). In films 13-15, the sound of the animal was audible in both shots. Two of the sound effect films (films 16 and 17) were low-support films in which the sound was present for only one of the two shots. For example, the first shot of film 16 depicted a house while a donkey could be heard, and the second shot showed the donkey making no sound. In the last sound effect film (film 18), sound was present for both views but nonspecific with regard to source and referent. This film depicted the outside of a house while unseen dogs barked, then cut to a scene of the inside of the house with the dogs still barking (and still unseen).

[Figure 3 here]

Finally, two of the films were intended to test the role of familiarity in naïve viewers' view integration. These depicted unfamiliar settings (figure 4) with low support. Both were filmed in the US. The first of these films (film 19) showed the outside of an American house and then cut to a view of a woman washing dishes at a sink inside the house. In this film, the sound of running water could be heard at the end of the first shot and throughout the second shot. The other film (film 20) depicted the outside of a large building with columns, then cut to a man walking inside the building whistling. The whistling could be heard during both shots.

[Figure 4 here]

Procedure

Participants were tested in individual sessions at their homes. The experimental session that lasted 90 – 120 min was started only after the interviewer, who was one of the authors (S.I.) became familiar with the participant. First, in order to check for possible auditory, visual, or cognitive deficits, participants were asked to describe their present situation (e.g., what they saw outside the window). Next, they were interviewed about their experience with and their knowledge about television and films. This interview included indirect questions about such matters as whether participants had ever been outside of their village, whether they regularly visited their children, whether their children had a television set, whether they knew about the political agenda of their country and knew popular television stars. Then, a laptop with a 17.3 inch display, which had been set up at the beginning of the session, was introduced. The laptop was placed at a distance of ~ 60 cm.

Participants were told that they would see something on the display and were asked to describe it as they had previously described their present (real-life) situation. They were asked additional questions to clarify their interpretations of the film clips. For example, when they said that they had seen a man shouting to a woman (films 5 and 6, illustrated in Figure 2), they were then asked where the man and woman were. Answers such as "he was outside of her house", "One was outside and one was inside the house" were taken as standard interpretations because they demonstrated that participants understood that the two views showed the same scene. Similarly, when participants said that they had seen a hee-hawing donkey (film 16, illustrated in Figure 3), they were asked where the donkey was. When they said that the donkey was next to a house, then they were asked to describe the house to verify that they were referring to the house shown in the first shot.

All participants were presented all video clips in randomized order. All experimental sessions were video recorded and transcribed later. For each clip, a 'standard interpretation' was defined that was based on an appropriate understanding of editing cuts. We assumed the validity of the standard interpretation if it was given by the control group consists of experienced viewers. The verbal responses were coded from transcripts in considerable detail with the qualitative analysis program Atlas-ti. Then, for each participant and each clip, the correspondence of his or her interpretation with the standard interpretation was determined independently by two coders, one of whom was one of the authors (S. I.), with an intercoder reliability (Kappa) of .94.

Results

As summarized in Table 1, first-time viewers often successfully arrived at standard interpretations of the sequences, and in all cases 100% of experienced viewers agreed upon the standard interpretations. Also, the difference in the prevalence of standard interpretations between the first-time viewers who had been in previous experiments (84% standard interpretation) and those who had not (82% standard interpretations; $t(18)=.914$, $p=.373$) was very small. Also, for no individual film did the difference in proportion of standard interpretations between participants who had and had not been in the previous experiment approach significance (all $p's > .353$, Fischer's Exact tests).

Familiar setting voice films

All first-time viewers reported that they saw two men standing opposite each other greeting one another (films 1 and 2). This was also the case when interpreting the similar eyeline violation film clips (e.g. clips violating the 180 degree rule; films 3 and 4). Typically, first time viewers responded, "two men greeted each other" and when asked, viewers reported that the men appeared to be standing across for each other.

In previous studies, first-time viewers had difficulty interpreting transitions from an exterior view to an inside view (Schwan & Ildirar, 2010), even when the outside was indicative of an object or an activity that was performed inside (Ildirar & Schwan, 2015). However, with the addition of sound, first-time viewers had little difficulty interpreting these sequences. For example, 100% of first-time viewers correctly understood the relationship between an initial shot in which a man can be seen outdoors calling out a woman's name and a subsequent shot that depicted a woman inside a house even though she did not react to the call (film 5 and 6), and when the person doing the calling could not be seen (films 7, 8, 9). For example, film 8 shows an establishing shot of a house (the camera is tilted up slightly at the house) while an off-screen voice calls out "Sister Fatma!". The second shot shows a woman sitting inside, who does not respond to a repetition of the call (presented at a lower volume). One first-time viewer responded, "A woman is drinking tea and another is calling to her", and when asked where the other woman was, the viewer reported "Outside, underneath the house, and the woman inside is hearing but does not responding".

The first-time viewers also correctly interpreted the two one-shot films (films 10 and 11) as showing one person talking with another off-screen actor. All first-time viewers correctly interpreted these films, demonstrating that they had no difficulty understanding that an on-screen actor could be looking off screen at another unseen actor. For example, one first-time viewer described the film as follows: "Someone said hello to a woman but she/he didn't appear". When asked where the speaker was, the participant said, "She didn't appear but she was across [from] her I guess."

Familiar setting sound effect films

Similar to voices, salient ongoing environmental sounds often helped first-time viewers combine views. Three films (films 12, 13, and 14) depicted continuous sounds of

people's actions (cutting wood), or animals that could be seen in one shot and heard in both. First-time viewers almost always produced standard interpretations (in 85, 70, and 90% of cases) of these films. In the case of the wood-chopping film (film 12), first-time viewers did provide relatively more diverse interpretations about the specific relative locations of the man chopping wood and the woman in front of the fire. 90% of experienced viewers described the man as being outside a house, and the woman as being inside the house, while only 15% of first-time viewers described this relationship. Instead, many first-time viewers stated that the two were next to each other (40%), or gave spatially indeterminate responses (30%). First-time viewers were also successful in understanding the film clips in which sound connected a long shot of scenery and a medium close up of an object in that scene. However, sound was not always successful in supporting first-time viewers' interpretations. In one film (film 14) chickens could be seen and heard outside a house preceding a cut to the inside of the house where two women could be seen sharing a meal. Only 25% of first-time viewers integrated these views, while all experienced viewers did so.

Finally, no first-time viewers integrated the views in film 15. This film featured an animal sound that turned out to be ambiguous. It depicted a long shot of a pond with lilly pads and the buzzing of a bee, followed by an extreme close-up of the bee. Most first-time viewers (along with experienced viewers) misinterpreted the buzzing sound as a motorbike when they first heard it in association with the pond. Interestingly, the experienced viewers saw the bee and then reinterpreted the sound as buzzing bee when they saw the second shot, and on this basis integrated the views. For example, one experienced viewer commented while watching, "there must be a motorbike passing by the lake... no, no, it is a bee". In contrast, first-time viewers often did not reinterpret the sound, and in some cases misinterpreted the close-up as depicting a very large bee. For example, one first-time viewer reported, "There appeared a big bee. Before that I heard a noise".

First-time viewers gave however many standard interpretations even when the sound was not continuous across views. For example, in film 16 a donkey could be heard while the exterior of a house was shown, and then a second shot showed the donkey but included no sound. All first-time viewers described this sequence as depicting a donkey near a house (and it is important to note that for other films, first-time viewers felt free to explicitly report that the spatial relationship between shots was unknowable).

[Table 1 here]

Unfamiliar setting films

In contrast to many of the other films, first-time viewers had considerable difficulty interpreting the two unfamiliar-environment films (films 19 and 20). They often indicated that it was not possible to know the relationship between the views, and sometimes misinterpreted the depicted locations. For example, the first shot of film 20 depicted a man walking inside the lobby of a large building, and many first-time viewers thought he was actually outside of a building. Thus, when these viewers saw the very different outside view of the same building, it would have been natural to assume that this was a second building. First-time viewers showed a similar tendency to misinterpret the house shown in film 19 because it had several connected rooms and a back porch, which led some first-time viewers to indicate that the shot depicted several houses.

Discussion

This experiment demonstrates that first-time viewers successfully integrated views for film clips including audio tracks. These viewers were successful not only with films that included dialog, but also with films that relied upon animal and environmental sounds. Not only did these sounds support view integration, but they did so even when they were not

continuous across both shots. For example, viewers successfully integrated views when they heard a donkey sound cut under a long shot of a house, then saw the now-silent donkey in the next view. In addition, view integration was successful when the both the source and referent for sounds were not visible. First-time viewers successfully integrated views when an unseen person called out to someone inside a house, and even when a continuous sound effect such as dogs barking was audible in both views. In this latter case, the dogs were not visible, and there was no particular referent for the sound as the object of the dogs barking was nominally unseen. The ability to rely on sounds from unseen sources may be strongly supported by everyday experience in which an approximately 120 degree visual field is a subset of a 360 degree auditory field. These results both reveal first time viewers' abilities to integrate views and demonstrate the importance of using rich multimedia materials in settings where more sparse materials (such as silent films) might lack cues that support recruitment of everyday perceptual skills for understanding media. However, first-time viewers were not uniformly successful. They did not integrate views when the films depicted unfamiliar environments and they seem to have had difficulty when sound effects could be misinterpreted.

So, how might one characterize first-time viewer's capabilities and, more generally, the mix of basic perceptual skills and medium-specific skills necessary to understand cinema? One way of framing the question is to ask about the relative roles of formal meaning-independent perceptual cues and the meaning of events in integrating views in cinema (Levin and Baker, 2017). The data presented here suggests that first-time viewers rely heavily on meaning to integrate views. This recruitment of meaning when understanding visual events might be understood with reference to theories of text comprehension such as Kintsch's (1988) construction integration model. This model assumes that story comprehension involves constructing propositions from a text, and then associatively retrieving related propositions from long-term memory (i.e. background knowledge). In the absence of relevant

background knowledge this activation cannot occur, leaving a relatively sparse knowledge structure to guide ongoing event perception (Zacks, Speer, Swallow, Braver, & Reynolds, 2007). It is also interesting to note that a lack of background knowledge may induce misleading interpretations that could interfere with view integration. For example, after viewing a film of a man walking through an American university building with columns inside, one first-time viewer responded: “This must be a mosque” (the only building they can imagine with such a high ceiling and columns) “but he is whistling” (it is not acceptable to whistle in a mosque).

However, it is not possible to be certain about the degree to which the formal structure of cinema itself served as a form of cognitive/perceptual “glue” that supports view integration independent of meaning. The one direct test of this hypothesis in the experiment assessed the degree to which first-time viewers integrated views that violated the 180 degree rule. They did so just as successfully as rule-consistent films, but it is important to note that 100% of viewers successfully integrated both kinds of films, so any difference that might have been observed could have been obscured by a ceiling effect. That said, the most salient difference between first-time viewers successes and failures in this experiment and others seems to be the presence of familiar events.

If first-time viewers rely strongly on meaning to integrate views, why do they sometimes fail when more experienced viewers from the same culture succeed? One reasonable hypothesis is that the more experienced viewers better understand the pragmatic goals of cinema. That is, they understand that views are arranged in the service of telling a story, and that in cases where the story is not initially evident, some effort at problem-solving may be rewarded. This problem-solving likely involves assumptions about something akin to directorial intent. For example, if one assumes that views are arranged with the intent not of portraying some specific real environment, but rather to specify a coherent story, it becomes

clear that few salient elements of the film are accidental. This process is evident in a number of experienced viewers' responses. For example, when viewing the bee video, some of these viewers corrected their misinterpretation of the ambiguous bee sound when they saw the bee in the second shot. This reinterpretation would be difficult to activate if it were not clear that all elements in the film are intended as parts of a meaningful whole, as opposed to literal reproductions of one or more scenes.

One particularly interesting question for future research is whether non-diegetic sound (e.g. sound that does not appear to come from the depicted space and events) such as music or narration would be similarly effective in integrating views. The fact that the dogs-barking effect seemed to help first-time viewers suggests that visible diegetic sources are not necessary to support view integration. Although non-diegetic sound represents an added level of remove for naïve viewers because it is much less similar to everyday events than diegetic sound, this lack of similarity might be more apparent than real if simple non-diegetic sounds can be structured to mirror everyday story-telling or imagination. For example, it is likely that individuals often narrate events both when they teach others and when they recount recent experiences. Especially in the latter case, an individual who must understand another person's recounting of a previous event needs to integrate the speaker's current description with an imagined series of events, a task similar to forms of absent reference that are learned by young children, and are assumed to serve as an important basis for language learning.

On the other hand, recent research has demonstrated that 12-month old infants' absent reference skills can be confounded because infants are overly concrete in including irrelevant location information in their representation of referred-to objects. Osina, Saylor, & Ganea (2013) found that infants had difficulty locating a hidden referred-to object if the infants had learned about it in one room and then were tested in another room. This implies that linking the current discourse with an arbitrary non-present location is a secondary skill that relies

upon learning that may be domain-specific. If this is broadly true, then naïve viewer's flexibility in linking a narration with events portrayed in a different location from the narrator may be limited.

The possibility that other forms of nondiegetic audio such as music may facilitate view integration is also potentially interesting, although in this case, there is less evidence and theory that might explain how experienced viewers rely on this source of information. One possibility is that music may reinforce thematic relationships among dissimilar shots that can be integrated into a story. Given cross-cultural research demonstrating flexible access to both thematic and taxonomic relationships among a broad range of cultures, it is possible that naïve viewers can rely on this skill to infer that events associated with a song are unified. Again, however, this may also be difficult for them if media-specific learning is required to activate the idea that a music sound track should be interpreted in relation to the visual events portrayed in the sequence of shots.

A particularly interesting question that the pattern of successes and failures exhibited by naïve viewers raises is the amount of learning that would be necessary for naïve viewers to overcome difficulties in interpreting edited sequences. Previous research reveals a few instances where only a very small prompt is necessary to successfully interpret pictures. Classic research on picture perception demonstrated that picture-naïve viewers who initially focused on the glossy surface of photographic prints needed only a quick pointer to focus their attention on interpreting the patterns on the paper to achieve success (Messaris, 1994). In other situations, a simple attentional instruction may be insufficient, but a more effortful working through of an example may produce a readily generalizable application of existing skills. In either case, understanding the new medium would rely mostly on existing skills that could be activated and broadly applied after a quick lesson. However, such quick lessons may

be insufficient, or may fail to generalize, in cases where media literacy requires more substantive new skills that must be practiced or applied in multiple situations.

References

- Abelman, R. (1989). From here to eternity: Children's acquisition of understanding of projective size on television. *Human Communication Research, 15*, 463–481.
- Anderson, J. D. (1996). *The reality of illusion: An ecological approach to cognitive film theory*. Carbondale, IL: Southern Illinois University Press.
- Anderson, D. R., Lorch, E. P., Field, D. E., & Sanders, J. (1981). The effects of TV program comprehensibility on preschool children's visual attention to television. *Child Development, 52*, 151-157.
- Anderson, D. R., & Pempek, T. A. (2005). Television and very young children. *American Behavioral Scientist, 48*(5), 505–522.
- Baker, L.J., & Levin, D.T. (2015). The role of spatial triggers in event updating. *Cognition, 136*, 14-29.
- Barr, R., Zack, E., Garcia, A., & Muentener, P. (2008). Infants' attention and responsiveness to television increases with prior exposure and parental interaction. *Infancy, 13*(1), 30-56.
- Bateman, J. A., & Wildfeuer, J. (2014). A multimodal discourse theory of visual narrative. *Journal of Pragmatics, 74*, 180-208.
- Bremner, J.G., Slater, A.M., Johnson, S.P., Mason, U.C., & Spring, J. (2012). The effects of auditory information on 4-month-old infants' perception of trajectory continuity. *Child Development, 83*(3), 954-964.
- Calvert, S. L., Huston, A. C., Watkins, B. A., & Wright, J. C. (1982). The relation between selective attention to television forms and children's comprehension of content. *Child Development, 53*, 601-610.
- Chion, M. (1994). *Audio-vision: sound on screen*. Columbia University Press.

Coe, B. (1992). *Muybridge and chronophotographers. Museum of the Moving Image*: London.

Cohn, N. (2013). *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images*. A&C Black

Cohn, N. (Ed.). (2016). *The visual narrative reader*. Bloomsbury Publishing

Crawley, A. M., Anderson, D. R., Wilder, A., Williams, M., & Santomero, A. (1999). Effects of repeated exposures to a single episode of the television program Blue's Clues on the viewing behaviors and comprehension of preschool children. *Journal of Educational Psychology, 91*(4), 630.

Frith, U., & Robson, J. E. (1975). Perceiving the language of films. *Perception, 4*(1), 97-103.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences, 3*(11), 419-429.

Gunning, T. (2001). Doing for the eye what phonograph does for the ear. In Abel, Richard, and Altman, Rick, eds. *The sounds of early cinema*. Indiana University Press: Bloomington, IA.

Hayne, H., Herbert, J., & Simcock, G. (2003). Imitation from television by 24- and 30-month-olds. *Developmental Science, 6*(3), 254–261.

Hobbs, R., Frost, R., Davis, A., & Stauffer, J. (1988). How first time viewers comprehend editing conventions. *Journal of Communication, 38*(4), 50-60.

Ildirar, S., & Schwan, S. (2015). First-time viewers' comprehension of films: Bridging shot transitions. *British Journal of Psychology, 106*(1), 133-151.

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., & Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics*, 72, 1736-1741.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), 163.

Kovács Á.M., Téglás E., & Endress A.D. (2010). The social sense: susceptibility to others' beliefs in human infants and adults. *Science*, 330, 1830–1834.

Lang, A. (1995). Defining audio/video redundancy from a limited-capacity information processing perspective. *Communication Research*, 22(1), 86-115.

Lemish, D. (1987). Viewers in diapers: The early development of television viewing. Natural audiences: *Qualitative Research Of Media Uses And Effects*, 33-57.

Levin, S. R., & Anderson, D. R. (1976). The development of attention. *Journal of Communication*, 26(2), 126-135.

Levin, D. T., & Baker, L. J. (2017). Bridging views in cinema: a review of the art and science of view integration. *Wiley Interdisciplinary Reviews: Cognitive Science*

Levin, D.T., & Simons, D.J. (2000). Fragmentation and continuity in motion pictures and the real world. *Media Psychology*, 2, 357-380.

Lowe, P. J. & Durkin, K. (1999). The effect of flashback on children's understanding of television crime content. *Journal of Broadcasting & Electronic Media*, 43(1), 83-97.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.

Messaris, P. (1994). *Visual "literacy": Image, mind, and reality*. Westview Press.

Meyerhoff, H.S., & Huff, M. (2016). Semantic congruency but not temporal synchrony enhances long-term memory performance for audio-visual scenes. *Memory & Cognition*, 44, 390-402.

Munk, C., Rey, G. D., Diergarten, A. K., Nieding, G., Schneider, W., Ohler, P. (2012). Cognitive processing of film cuts among 4-to 8-year-old children: An eye tracker experiment. *European Psychologist*, 17(4), 257.

Murch, W. (2003). Touch of silence. In Sider, L., Sider, J., & Freeman, D. (Eds.) *Soundscape: The School of Sound Lectures 1998-2001*. Wallflower Press: New York.

Osina, M. A., Saylor, M. M., & Ganea, P. A. (2013). When familiar is not better: 12-month-old infants respond to talk about absent objects. *Developmental Psychology*, 49(1), 138.

Parkinson, D. (1996). *History of Film*. New York: Thames and Hudson

Pempek, T. A., Kirkorian, H. L., Richards, J. E., Anderson, D. R., Lund, A. F., Stevens, M. (2010). Video comprehensibility and attention in very young children. *Developmental Psychology*, 46(5), 1283.

Raynaud, I. (2001). Dialogues in early silent screenplays: What actors really said. In Abel, R., & Altman, R. R. (Eds.). *The sounds of early cinema*. Indiana University Press: Bloomington, IA.

Richards, J. E., Cronise, K. (2000). Extended visual fixation in the early preschool years: Look duration, heart rate changes, and attentional inertia. *Child Development, 71(3)*, 602-620.

Richards, J. E., & Gibson, T. L. (1997). Extended visual fixation in young infants: Look distributions, heart rate changes, and attention. *Child Development, 68(6)*, 1041-1056.

Rodero, E. (2010). See it on a radio story: Sound effects and shots to evoked imagery and attention on audio fiction. *Communication Research, 39*, 458-479.

Schwan, S., & Ildirar, S. (2010). Watching Film for the First Time How Adult Viewers Interpret Perceptual Discontinuities in Film. *Psychological Science, 21*, 107-113.

Shams, L., Kamitani, Y, & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature, 408*, 788.

Shimojo, S., & Shams, L. (2001). Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion In Neurobiology, 11(4)*, 505-509.

Smith, T. J., Levin, D. T., & Cutting, J. E. (2012). A window on reality: Perceiving edited moving images. *Current Directions in Psychological Science, 21(2)*, 107–113.

Smith, R., Anderson, D., & Fischer, C. (1985). Young Children's Comprehension of Montage. *Child Development, 56*, 962-971.

Tim J. and Martin-Portugues Santacreu, J.Y. (2016) Match-action: the role of motion and audio in creating global change blindness in film. *Media Psychology* , ISSN 1521-3269.

Somanader, M., Garcia, A., Miller, N., & Barr, R. (2005). Effects of sound effects and music on imitation from television during infancy. Paper presented at the Society for Research on Child Development, Atlanta, GA

Standish, I. (2006). *A New History of Japanese Cinema: A Century of Narrative Film*. New York: Continuum.

Tibus, M., Heier, A., & Schwan, S. (2013). Do films make you learn? Inference processes in expository film comprehension. *Journal of Educational Psychology, 105*, 329-340.

Troseth, G. L., & DeLoache, J. S. (1998). The medium can obscure the message: Young children's understanding of video. *Child Development, 69*(4), 950-965.

Van der Burg, E., Olivers, C.N.L., Bronkhorst, A.W., & Theeuwes, J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 1053-1065.

Willems, R. M., & Hagoort, P. (2007). Neural evidence for the interplay between language, gesture, and action: A review. *Brain and Language, 101*(3), 278-289.

Zacks JM, Speer NK, Swallow KM, Braver TS, Reynolds JR. (2007). Event perception: a mind-brain perspective. *Psychological Bulletin, 133*, 273-293.

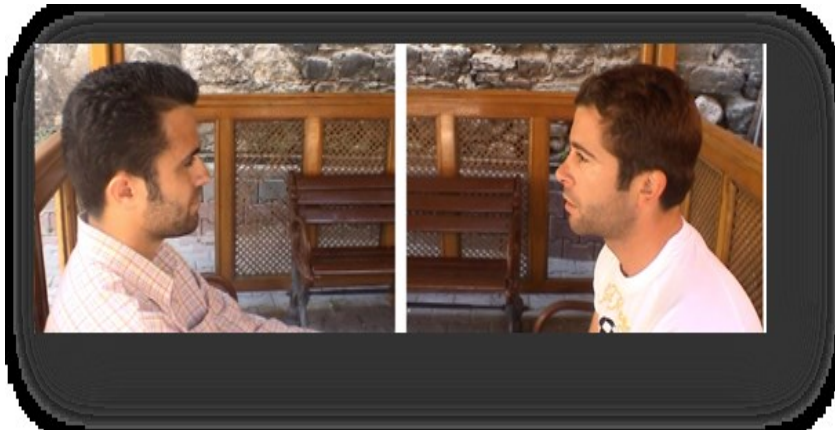
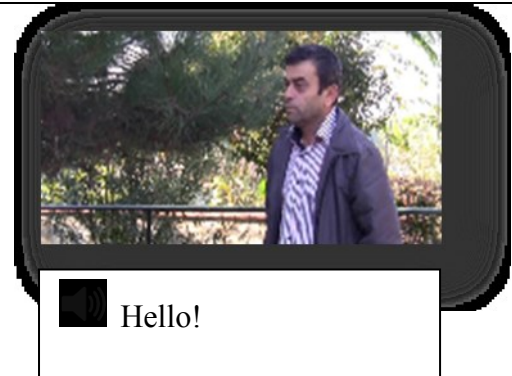
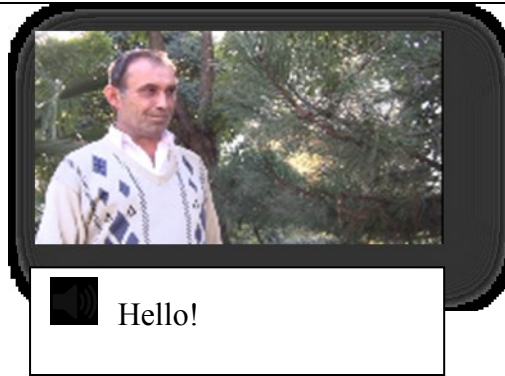


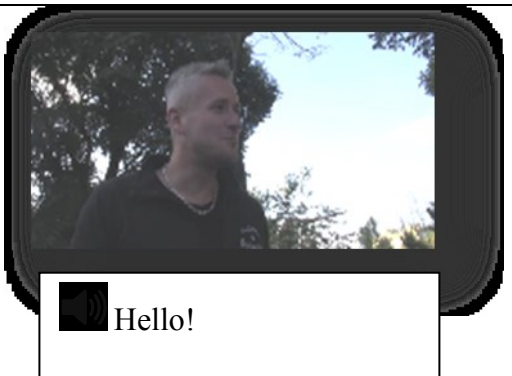
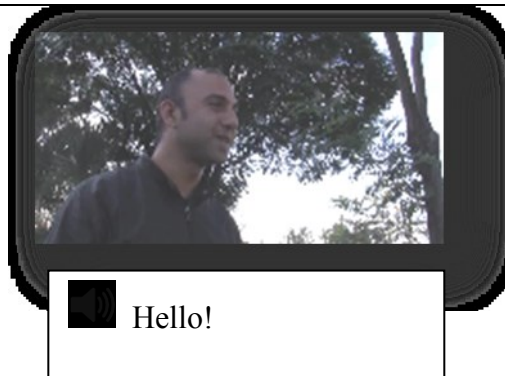
Figure 1. Example of sequence from Schwann and Ildirar (2010).

Familiar Settings, Voice Films

Films 1,2.
High Support
Dialog



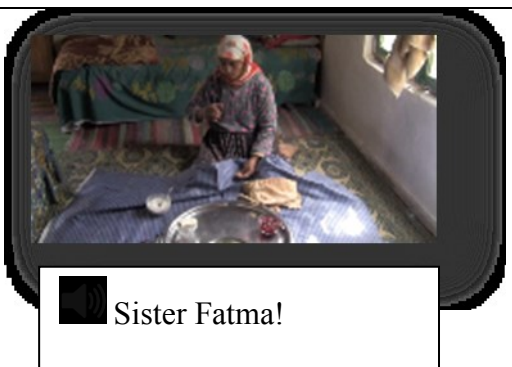
Films 3,4.
High Support
Dialog + 180
degree rule
violation



Films 5, 6.
Med. Support
One Speaker,
Visible
Silent
Partner



Films 7, 8, 9.
Low Support
No speaker,
visible
partner



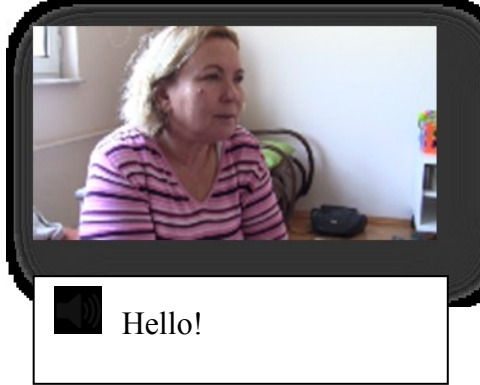
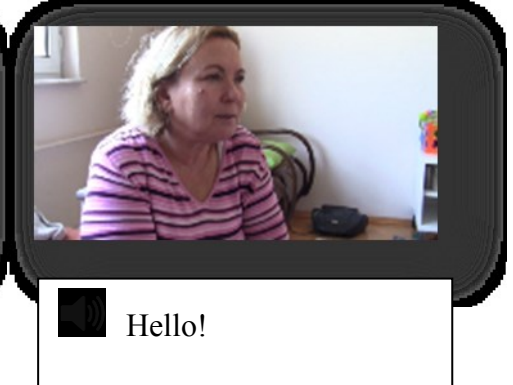
<p>Films 10, 11 Low Support Unseen partner</p>		
--	---	--

Figure 2: Film clips with human-voice

<i>Familiar Setting, Sound Effect Films</i>		
<p>Film 12. Medium Support. Continuous Sound</p>		
<p>Films 13, 14, 15. Medium Support Continuous animal sound</p>		
<p>Films 16, 17 Low Support Noncontinuous animal sound</p>		

<p>Film 18 Low Support continuous sound no visible sound</p>	 <p data-bbox="427 465 842 568">  Dogs Barking </p>	 <p data-bbox="943 465 1358 568">  Dogs Barking </p>
--	---	--

Figure 3: Film clips with environmental sound


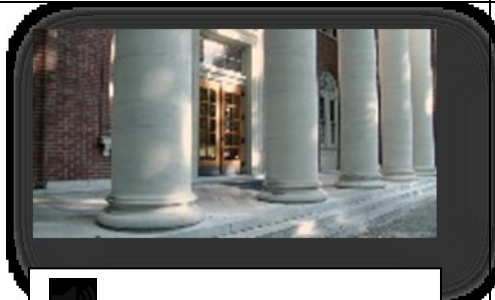

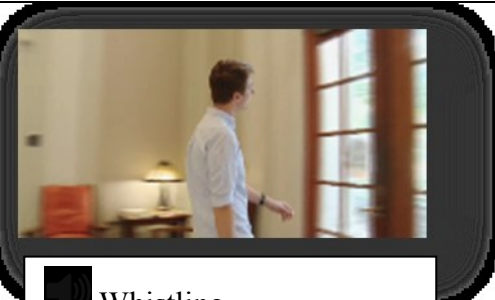

<p><i>Unfamiliar Setting</i></p>		
<p>Film 19 Low Support Continuous Sound</p>	 <p data-bbox="427 1182 842 1285">  Water Sink </p>	 <p data-bbox="943 1182 1358 1285">  Water Sink </p>
<p>Film 20 Low Support Continous Sound</p>	 <p data-bbox="427 1565 842 1668">  Whistling </p>	 <p data-bbox="943 1565 1358 1668">  Whistling </p>

Figure 4. Unfamiliar-setting films

Film	First-Time Viewers' Interpretation (% standard)
<i>Familiar Setting Voice Films</i>	
1 Dialog	100 (n.s)
2 Dialog	100 (n.s)
3 Dialog, eyeline violation	100 (n.s)
4 Dialog, eyeline violation	100 (n.s)
5 One speaker, visible partner	90 (n.s)
6 One speaker, partner not seen	100 (n.s)
7 No speaker, visible partner	100 (n.s)
8 No speaker, visible partner	100 (n.s)
9 No speaker, visible partner	100 (n.s)
10 One shot, unseen speaker	100 (n.s)
11 One shot, unseen speaker	100 (n.s)
<i>Familiar Setting Sound Effect Films</i>	
12 Continuous sound of action	85 (n.s)
13 Continuous animal sound	70 **
14 Continuous animal sound	90 (n.s)
15 Continuous animal sound	25 ***
16 Noncontinuous animal sound	100 (n.s)
17 Noncontinuous animal sound	80 (n.s)
18 Continuous sound no source	85 (n.s)
<i>Unfamiliar Setting Films</i>	
19 Unfam Env, continuous sound	20 ***
20 Unfam Env, continuous sound	0 ***

Table 1. First-time viewers' interpretations of films.