The Role of Cinematic Visual Context in Supporting Viewers' Language Processing

Dylan Kistler

Dr. Levin

PSY-PC 4999: Honors Seminar

March 2022

Abstract

Previous research has demonstrated that basic forms of visual context such as object identification and gaze support language comprehension. However, complex forms of narrative context may structure visual supports for language in ways research has yet to reflect. I investigated how cinematic cues such as edit timing and shot coverages (such as depicting actors in close-ups as opposed to wider views) support language comprehension. Participants were shown scenes that either maintain or disrupt the timing of cuts, shot coverage, and other elements of visual context. Participants were tested for their memory for conversation and theory of mind accuracy, as well as reported their perception of continuity from each scene. The experimental conditions had a significant effect on memory for conversation performance but not on theory of mind inference. Memory performance was significantly decreased in the slideshow and reordered conditions, and perceived continuity was significantly decreased for all three conditions compromising the original scene's visual context.

The Role of Cinematic Visual Context in Supporting Viewers' Language Processing

Language is almost always understood alongside accompanying visual information, rather than being interpreted as distinctly auditory information apart from any visual context. For instance, observing lip movements and gestures, leveraging basic visual information about one's environment, and detecting a speaker's eye gaze have all been shown to impact listener's understanding of language early during processing. These instances of visual supports have been robustly confirmed to aid in language processing in real-time, but they are certainly not descriptive of the multitudinous other ways in which visual information is used to inform how language is understood across a wide range of possible contexts.

Language abounds in the videos of the 21st century, from television, to YouTube, to various kinds of social media. As in most other naturalistic contexts, language in video is accompanied by visual context that likely aid in listener's understanding of what they hear. A history of filmmaking practice hints at the fact that professional creators of video have long understood the ties between effective visuals and the desired delivery of a line in a movie (Levin & Baker, 2017). Visual context in cinematic conversation may support language comprehension for viewers of cinema in ways suggested by various filmmaking best practices but not yet proven by empirical research. Likewise, the supports leveraged by filmmakers may suggest new ways in which visual context informs language processing not previously revealed by experiments focusing on lip movements, gaze, details about one's physical environment, or eye gaze.

Visual Context Outside of Film

Visual context in various instances has been well established as a real-time support for language processing. These findings have been confirmed for several unique visual supports, oftentimes alongside listener's interpreting potentially ambiguous language since those are times

at which it can be made evident at what point during the phrase being heard the listener's gaze or actions reveal that they have ascertained the language's meaning. One type of visual context which has been studied for its role in real-time language processing is a listener's detection of a speaker's lip movements to resolve ambiguous syllables in the speech. Another support that has been studied is the situational visual information in one's physical surrounding which contribute to more quickly understanding the meaning of ambiguous language. As an example, asking "is the food hot" is likely to be understood differently from the earliest moments of processing by someone looking at a spicy chili pepper as opposed to someone looking at a bowl of hot soup. Gestures are visual cues speakers offered with their hands to potentially add information to their speech, and adult gesturing has been shown to frequently offer added information not redundant with words being spoken when the subject of speech is increasingly far away.

Information movements are a subtle visual cue which people use to help identify differences between similarly sounding syllables they hear. Research from McGurk & Macdonald (1976) found that lip movements which were synchronized to a woman speaking the syllables "ba" or "ga" were understood as such by most normal adults, while dubbing her pronunciation of "ba" on top of video showing her saying "ga" led to most adults incorrectly believing the woman had said "da". Adults who listened to the voice without any visual input were also able to correctly understand the syllables, suggesting that witnessing incorrect visual support was more detrimental than the correct visual supports were helpful in this example of syllabic language processing.

Gesturing has been widely studied among infants as a crucial way in which infants strive to establish join-attention with adults to better their learning outcomes and relationships (Capone & McGregor, 2004). Later research suggested that infants are not the only ones who

point to establish joint-attention and word learning, as well as that pointing is consistently used by adults to coordinate attention. Bangerter (2004) demonstrated that as the distance away from a third object visible to each of two adults increased, participants used more pointing and less speaking to describe the object. The proportional tradeoff between the use of pointing and the use of language to communicate supports the view that pointing gestures are not typically redundant with linguistic information. Rather, gesturing is exercised strategically to build attention and modify language in situations where language is increasingly vague and needs visual support to efficiently convey the speaker's intentions.

For decades, disambiguation of language using visual context has been a popular domain of research for psychologists. Early experiments suggested that people use visual context to differentiate between ambiguous options presented in spoken language (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). For example, "candy" and "candle" are identical through the first half of their pronunciation. When only one of these two objects were placed in front of participants, the mean time to gaze at the correct object was 145 ms from the end of the keyword, whereas that time grew to 230 ms when both were physically present in front of the participant. This demonstrates that the visual context was used in on-line language processing early on as the participant sought to disambiguate the linguistic input. Tanenhaus et al. also evaluated the timing of incorporating visual context to disambiguate the phrase "put the apple on the towel in the box," a phrase that has possible meanings which are only discernible based on whether the apple is already on a towel, or whether another towel has been placed adjacent to the apple and the box with nothing on it. Eye-gaze tracking found that from the earliest moments of linguistic processing as "put the apple on the towel in the box" was being read aloud, participants sought visual references in accordance with their behavioral goals. In other words, from the earliest

measurable moments that participants were successfully processing the language of the instructions, they were already gazing at the various visual references available according to each part of the sentence read aloud, such that there was quick and continuous use of the physical items to disambiguate the sentence during every moment of on-line language processing. This supports theories of language comprehension which characterize linguistic and nonlinguistic information as being rapidly integrated into the earliest stages of language processing.

Trueswell, Sekerina, Hill, and Logrip (1999) built on the above findings with their famous study adapting the above experiment to compare adult and children performance disambiguating the phrase "put the frog on the napkin in the box." A selection of different stuffed animals and objects were used in further trials, and in some trials, participants contended with a physical context in which two different or identical animals were placed on and off of the napkin, in order to generate more possible interpretations during on-line processing of the prompts. Alternatively, a control featured the tracking of eye-gaze in the presence of the unambiguous modification "put the frog that's on the napkin in the box." The children performed worse than adults due to a systematic tendency to commit to an interpretation of the language early on when listening to the sentence, as evidenced by gazing at an incorrect object and subsequently refusing to edit their first guess when new linguistic information was added. Nevertheless, children showed a similar process of progressively integrating visual information during on-line language processing, in particular by employing the use of gaze to make predictions when presented with the unambiguous versions of prompts. Additionally, the significantly superior performance of adults over children supported the theory that applying pragmatic principles - which are common principles applied to make inferences about language - is a sophisticated skill used in on-line language processing until later in life.

Eye-gaze from fellow humans is another cue demonstrated to affect on-line linguistic processing early (Hanna & Brennan, 2007). In their experiment, naïve pairs of participants were separated by a barrier such that they could see one another's faces but not one another's display of identical objects in a row. One was assigned to be a director and the other a matcher. The director told the matcher to move a target object using a command following a sentence structure, for example "the blue circle with five dots on it." Depending on the other shapes in the display, there might be no competitor, meaning no alternative with the same color, or a competitor with the same color but a different number of dots. The competitors couldn't be discerned as incorrect until later in the sentence due to the color preceding the number of dots inside the shape in each director's description. Participants found the correct shape much more quickly when they could see the shapes in the same order as the director (and thereby take advantage of the director's gaze as a cue). Using eye gaze as a cue in this way proved to aid in finding the correct shape early during ambiguous sentences, even when a near-competitor shape threatened to distract from the correct shape due to having many similarities that might have required further language to distinguish. Therefore, eye-gaze produced by the speaker was used to resolve linguistic ambiguity early in the processing of the director's sentences and was used to discern sentence meaning earlier rather than merely modifying understanding if on-line sentence processing proved incorrect.

Visual Supports from Filmmaking

Filmmakers contribute to the list of visual supports that enhance understanding for conversation through the on-screen information they cut, curate, and present in cinema. Analyzing how these choices create meaningful support for conversational understanding in

viewers is likely to necessitate a complex, cognitive approach to studying visual context's effect on language processing.

Filmmaking has a long history of leveraging the cognitive strengths of viewers in their creative techniques; well-known early filmmakers like D.W. Griffith employed pseudo-empirical approaches to observing audiences and modifying films accordingly (Slide, 2012). The segmenting, ordering, and editing of visual context alongside language developed into filmmaking principles latent with insights about cognition, in effect placing the pioneers of early filmmaking as "informal cognitive scientists" (Levin, Brown-Schmidt, & Watson, 2020).

Still, little work has been done to determine the effect of the visual supports filmmakers utilize on viewers' processing of language in cinema. As a result, the intuitive practice of filmmaking has a storied history of gradually incorporating but not empirically backing cognitive principles through its approach. For instance, filmmakers from as early as the mid 20th century eventually embraced the notion that they did not have to always show a person who was talking (Levin & Simons, 2000), harkening to the findings about pointing at a third object during established communication settings as an effective way to opportunistically add details and maintain joint focus (Bangerter, 2004). While the filmmaker does not literally point in front of the audience member to direct attention, they do often cut to show a third object that is the focus of speech. By cutting during the sentence, filmmakers demonstrate reliance on the cognitive abilities of viewers to keep track of who is speaking while visually refocusing on a third object that the speech references. For instance, cutting from a shot of a car salesman pointing offscreen to a view of the car they are selling is likely to be understood by viewers because of the sequence's narrative logic. Additionally, editing techniques have increasingly encouraged the practice of varying shot distances and angles. The filmmaker must decide

between a wide angle shot showing both the speaker and listener, or to focus solely on the speaker's face (Bordwell & Thompson, 1993). These choices are both fundamental to the art of filmmaking, and directly demonstrate a cinematic analogy for the kinds of visual supports cognitive scientists commonly investigate in naturalistic conversation.

Before any more nuanced analysis of filmmaking techniques can be approached through a cognitive lens, it must first be determined to what degree the core visual inputs of speakers' expressions, timing of edits, and choices of shots verifiably support any understanding of language. Likewise, this study will determine whether compromising core visual inputs in cinematic conversation impedes viewers' understanding of on-screen language.

Understanding for Conversation

A person's understanding for a given conversation is a broad concept to measure, and one which is to a degree modified by the goals of the filmmaker or cognitive scientist in question. For the purpose of this study, we will focus the parameters of understanding to include memory for conversation, theory of mind inferences arising from the dialogue and social situations inherent to most cinema, as well as how these may be mediated by participants' perceived continuity of the scene. The study will seek to investigate to what degree the typical visual supports filmmakers include in their cinematic conversations are necessary to perform well on each of these three measures of understanding, as well as what interactions might exist between relative success or failure on the three measures when considered as independent components of effective language processing.

The first of these measures for understanding, memory for conversation is both a highly limited and crucial cognitive function. As far back as 1979, Ross and Sicoly performed the foundational memory for conversation study which found that after a period of several days,

adults remembered only 6% of the ideas they had communicated in a conversation and only 3% of ideas others had communicated. The study allowed for paraphrasing to qualify as correct to a broad degree, investigating only what information was correctly stored from the conversation.

This study will focus on item memory, defined as the content that is spoken in a scene, as opposed to source memory, which entails the knowledge of who said a given line. We hypothesize that overall memory for conversation will decrease when the visual context is compromised (Levin, Brown-Schmidt, & Watson, 2020).

Theory of mind, the second measure of interest, evinces an individual's ability to infer mental states such as beliefs, desires, and intentions, as well as to predict the behavior of others based on those inferences (Apperly, 2012). Research on theory of mind has historically focused on a narrow range of children, but recently is being expanded to include broader age demographics. These studies on children have also focused heavily on false-belief tasks, in which children's aptitude of theory of mind is measured by seeing how well they can identify someone whose who's beliefs are both false and different from form their own (Wimmer & Perner, 1983).

Theory of mind is often utilized by filmmakers to create suspense or impending drama. For instance, scenes where the audience is privy to knowing who will backstab who, or where the cop is bewildered but the audience is aware of the culprit, and permit audiences to observe rich developments in mental states among characters on screen. In a 1993 essay on Jane Austen's pivotal role in literary history, Zunshine argues that cognitive frameworks of theory of mind are needed to organize Austen's intricate webs of intersubjectivity, especially in social situations where the layers of inferences about character beliefs made by the author can encourage the reader to (or dupe them into believing they can) participate in tantalizingly complex social

analysis. In her article, Zunshine notes the similarities between deep intersubjectivity in film studies and literary studies, observing how "words and gestures" combine to paint a picture of social complexity in Austen novels, and that seeing multiple bodies at once allows for a more rapid and dynamic presentation of intersubjectivity than the one-by-one unfolding of details which is generally inherent to textual narratives. Through each of these points, Zunshine builds a convincing case not only for the merit of cognitive approaches to sophisticated literary art, but also that her claims are bolstered when applied to narratives accompanied by a visual context.

Third, this study will measure continuity in film as a possible mediator for understanding between memory for conversation and theory of mind. In the introduction to a recent Levin and Kai study, a range of useful definitions for cinematic continuity are given, including "creating the illusion of continuous action" and "preserving graphic, space, time, logical, and narrative connections between shots" (2020). Conceptual integration is defined by Levin and Baker as the arrangement of visual and conceptual cues by an audience to process an understandable sequence (2017). Notably, this conceptual integration can occur in the midst of heavily disrupted perceptual continuity (Levin & Kai, 2020). Levin and Kai also posit that perceptual continuity in cinema has been put forth by expert filmmakers as chiefly important for directing attention, which can likewise aid the audience in confidently subscribing meaning to the scenes they view. In this experiment, experiences of perceptual continuity were effectively manifested by stronger continuity cues in the editing, as well as that increased individual reports of perceptual continuity in the study coincided with an increase in one's ability to comprehend when a character was looking off-screen at another character. Because of the efficacy demonstrated by the Levin and Kai self-report measure of perceptual continuity, that measure

will be utilized in this study as the third and final component of audience understanding that is to be measured at the completion of viewing each scene.

Method

Participants

The study included 78 adults who signed up using either the Vanderbilt University SONA system (50 of the participants), or the Mechanical Turk online labor recruitment pool created by Amazon (the remaining 28 participants). Those who enlisted for the study from Mechanical Turk were vetted for certification and received financial compensation from Dr. Levin's lab. Participants from Mechanical Turk were sourced only from the pool of Master Workers, who have shown prior success across a variety of previous tasks. Students who enrolled through Vanderbilt University's SONA system could receive a modest hourly rate, but most opted to instead complete the study for credit needed in their introductory psychology or research courses. Additionally, participants were screened to identify whether they had seen the films from which the four scenes are sourced, either in full or in part.

Materials

Participants completed the experiment on either their own computer (Mechanical Turk participants vetted through Cloud Research) or a university computer (SONA participants).

Source movies were chosen such that no film is deemed popular enough that a large portion of possible participants would have to be eliminated for having seen it. Potential films released in the past 12 years could only be chosen if they had fewer than 40,000 total ratings on IMBD, and films released more than 12 years ago had to have fewer than 100,000 total ratings to be considered. Additionally, any chosen film had to have an IMDB popularity score of 500 or less. The films chosen were *The Science of Sleep* (2006), *The Ice Storm* (1997), *The Freshman*

(1990), and *North by Northwest* (1959). Based upon the main topic of each, the scene chosen from *The Science of Sleep* is later referred to as "Art Showing", the scene chosen from *The Ice Storm* is referred to as "Band Money", the scene chosen from *The Freshman* is referred to as "Job Offer", and the scene chosen from *North by Northwest* is referred to as "Train".

Only one scene from each film could be chosen. Scenes featured dialogue between 3 or fewer primary characters, with the focus on background characters kept to a minimum. Scenes chosen had to be long enough to fulfill the 60 to 120 second interval, and preference was given to scenes with more subtext for testing theory of mind and more simplistic edits that focus on the characters' faces.

The three modified conditions of each scene were created using FinalCut Pro, according to the editing parameters as established in the Design section. Special care was given such that for the slideshow-based manipulations, which included all three of the total manipulations, the transitions fell exactly on the frame where the filmmaker had originally placed a cut to a new speaker's face. The scenes were edited to a length of between 60 seconds to two minutes. The control group scene was not manipulated from its original video, and no scenes had their audio edited.

Memory for conversation questions were modeled after the methods used in Keenan, MacWhinney, & Mayhew (1977). Each question was followed by the correct answer, a near-miss option that paraphrases the correct answer, an answer which differs in its propositional content from the original (while remaining plausible within the context of the scene), and a paraphrase of the answer differing in propositional content. The four answers were presented in random order.

Finally, the continuity measure will be based on the Likert-scale self-report featured in Levin & Keliikuli (2020). The first three of the questions borrowed from the study measure conceptual integration, while the fourth represents perceptual flow as perceived by the participant. The questions asked participants to what degree (1) it was easy to understand how the shots fit together into events, (2) it was sometimes difficult to understand interactions between one person and the other, (3) they were sometimes confused because it appeared as though the shots were showing different things in different places, and (4) sometimes it looked as though objects suddenly changed location or were suddenly further or closer to the camera.

Design

The study had a repeated measures design using one independent variable, which was the manipulation of the scene's visual context. The visual context had three levels of manipulation beyond the control group, where scenes were presented in their original form. The first of three manipulations included changing the scene's video into slides showing the speaker's face according to the timing of who was shown in each of the filmmakers' shots. This was labeled the *slideshow manipulation*. The second was a slideshow like the first manipulation except that the timing of transitions between slides was adjusted by up to one second so that the introduction of the speakers' faces did not match the start of their lines in the scene and was referred to as the *miscut slideshow*. Finally, the third manipulation used the same basic slideshow model, except that its slides were reordered throughout the scene such that no slide ended up adjacent to a slide to which it was initially adjacent. This was known as the *reordered slides* manipulation. Each participant saw a control scene and one scene presented as one of each of the three manipulations, for a total of four scenes. Participants were randomly assigned to which scenes they saw under the original (control) condition and under each manipulation. Therefore,

differences in the scenes did not systematically alter how performance on the manipulations and control groups compared.

After viewing each of the four scenes, participants answered Qualtrics questions specific to that scene. Of these questions, six tested for memory in the conversation, six asked the participant to make a theory of mind inference, and four were self-report questions identifying the participant's perception of continuity in the scene.

Procedure

The experiment occurred in a single-sitting Qualtrics experiment and was estimated to last for one hour or less. On average, participants took 24.5 minutes to complete the experiment. Participants saw four scenes, randomly being assigned to see one control scene and one test scene of each of the three manipulations. After every scene, the participant answered questions, beginning with memory for conversation questions, then theory of mind questions, and finally the continuity self-report. The time to answer any given question was not limited. Participants were not able to go back to previous questions or scenes after choosing an answer and subsequently selecting "next" to proceed.

Results

Data was entered into a within-participants ANOVA for the overall effect of condition on memory, theory of mind, and continuity performance. There was a significant overall effect of condition on memory performance, $F(1,73) = 2.849$, $p = .038$, $n^2 = .036$, suggesting some overall impact of the visual context for performance on memory questions (See Figure 1). In contrasting specific conditions, there was an 8.5 percent decrease ($p = .035$) in memory performance between the original condition and the slideshow conditions, and a 9.8 percent decrease ($p = .015$) in memory performance between the original and reordered conditions.

However, there was not a significant overall effect of condition on theory of mind performance in the within-participants ANOVA, $F(1,73) = 1.348$, $p = 0.260$, $n^2 = .017$, suggesting there was not a significant impact of the visual context for performance on theory of mind questions (See Figure 2). In contrasting specific conditions, there was an 7.8% percent decrease ($p = 0.064$) in memory performance between the original condition and the reordered conditions, suggesting the most extreme obstruction of the intended visual context still was not significant.

Participants experienced a strong effect of condition on their reported perception of continuity, $F(3,288) = 57.482$, $p < .001$ $n^2 = .431$, between the original scenes and three modified conditions (See Figure 3). A significant difference in perceived continuity was also observed between the slideshow and reordered condition, $t(288) = 2.124$, $p = .035$, as well as when comparing the miscut condition with the reordered condition, $t(288) = 2.460$, $p = .015$.

Data was also entered into eights between-participants ANOVAs, including an analysis of memory performance and theory of mind performance for each of the four scenes (see Figure 4). There was no significant effect of condition on memory for the Job Offer scene, $F(3,74) = 2.023$, $p = .118$, $n^2 = .076$, nor was there a significant effect of condition on theory of mind, $F(3,74) = 2.073$, $p = .111$ $n^2 = .078$. Similarly, there was no significant effect of condition on memory for the Band Money scene, $F(3,74) = 2.478$, $p = .068$ $n^2 = .091$, nor was there a significant effect of condition on theory of mind, $F(3,74) = 1.316$, $p = .276$ $n^2 = .051$. While the analysis of the Art Showing scene did not demonstrate a significant effect of condition on memory, $F(3,74) = 1.364$, $p = .260$ $n^2 = .052$, there was a significant effect of condition on theory of mind for Art Showing, $F(3,74) = 3.822$, $p = .013$ $n^2 = .134$. Additionally, there was a significant effect of condition on memory for the Train scene, $F(3,73) = 7.142$, $p < .0001$ n^2

=.227, as well as a significant effect of condition on theory of mind, $F(3,73) = 4.449$, $p = .006$ n^2 =.155.

There were also notable contrast analyses between specific conditions in the individual scenes. The analysis of the Job Offer scene showed a significant difference between memory performance between the miscut and reordered conditions, $t(74) = 2.170$, $p = 0.033$. For the Train scene, individuals unexpectedly performed significantly better for memory in the reordered condition as compared with the slideshow condition in their memory scores, $t(73) = -3.855$, $p <$ .001. Moreover, for the Train scene, the reordered condition was surprisingly associated with significantly stronger theory of mind performance, $t(73) = -3.461$, $p = <.001$. In the analysis of the Art Showing scene, $t(74) = 3.228$, $p = .002$. For the Band Money scene, there was a significant effect on memory when comparing the original condition and the slideshow condition, $t(74) = 2.001$, $p = .049$, as well as a significant effect on memory when comparing the original condition and the miscut condition, $t(74) = 2.265$, $p = .026$, and finally a significant effect on memory when comparing the original condition and the reordered condition, $t(74) = 2.307$, $p = .024$.

Discussion

The findings from this experiment provides some support for the hypothesis that compromising the visual context provided by filmmakers can disrupt a viewer's ability to understand cinematic dialogue. While the original condition of the scenes produced better memory performance than the slideshow and reordered conditions, the miscut condition was not significantly worse. Replicating the study to ensure that original versions of films do produce better memory for conversation may be needed to confirm this finding due to the outlier comparison between the original and miscut conditions.

The effects on theory of mind performance were more uncertain, with performance trending in the direction hypothesized (each next condition of compromised visual context had decreasing performance for theory of mind) although the result was still not significant. It is an intriguing result that the reordered condition of the original scenes (showing only still shots in a fully random order) did not produce significantly worse performance for viewers' theory of mind inferences, suggesting that either the visual context was not meaningful, or perhaps that the visual supports available for understanding conversation were mitigated by the effect of viewers tuning out the visual context and focusing on only the scene's audio.

The continuity scores suggest the first manipulation (between the original condition and the slideshow condition) had by far the greatest effect on perceived continuity of the scenes shown, and the reordering manipulation was a smaller but still significant effect noticed by viewers. The measurement of memory, theory of mind, and continuity are perhaps each not sensitive enough or not accurately constructed in order to discern an effect of the miscut condition, otherwise suggesting that the specific timing of edits within a margin of one second is not ultimately meaningful for viewers understanding of cinematic conversation.

One of the reasons for not finding more conclusive effects on understanding for conversation from compromising visual context may be the quality of questions used for theory of mind performance. While the memory questions were based on a highly structured framework (Keenan, MacWhinney, & Mayhew, 1977) the theory of mind questions was based on principles of theory of mind inference but did not have a structured for writing incorrect multiple-choice answers. In a couple of instances, especially in the Job Offer scene, participants answered a particular wrong far answer more often than the correct answer, with this pattern being most consistent in the viewing of the original condition (suggesting that viewers with the

hypothetically best visual context were most likely to select this incorrect answer). This analysis implies that there may be disruption in the theory of mind findings due to some of the questions producing lower accuracy for participants with better visual context if an alternate answer was in fact superior in the eyes of most participants. An exploratory analysis of eliminating these problematic questions suing item response theory is needed for further investigation.

The pattern of results across the four scenes was strikingly different (see figure 4), suggesting that the variation in filmmaking was impactful on the experimental results of the conditions across each scene. For instance, the significantly better theory of mind performance in the reordered condition is only present for the Train scene, suggesting that when viewer's quickly note the visual information in that scene is not helpful and may instead dedicate more attention to the scene's audio information. It could be that the traditional style of shot-reverse shot - whereby shots follow the speakers face and cut only when a new speaker begins to talk - does not offer substantive visual context for understanding conversation. Thus, participants who quickly tuned out the visual context in favor of focusing on audio performed markedly better in the Train scene due to that visual context being unhelpful in informing theory of mind inferences.

Furthermore, the significantly worse memory performance in the reordered condition of the Job Money condition suggests that the ordering of key visual context was important to understand the dialogue in the scene. Since this scene features the most notable movement to new spaces during its progression, it's possible that the ordering of events is more crucial in scenes such as Job Offer due to the importance of sequencing for understanding narrative that takes place in several distinct, significant places within a single scene.

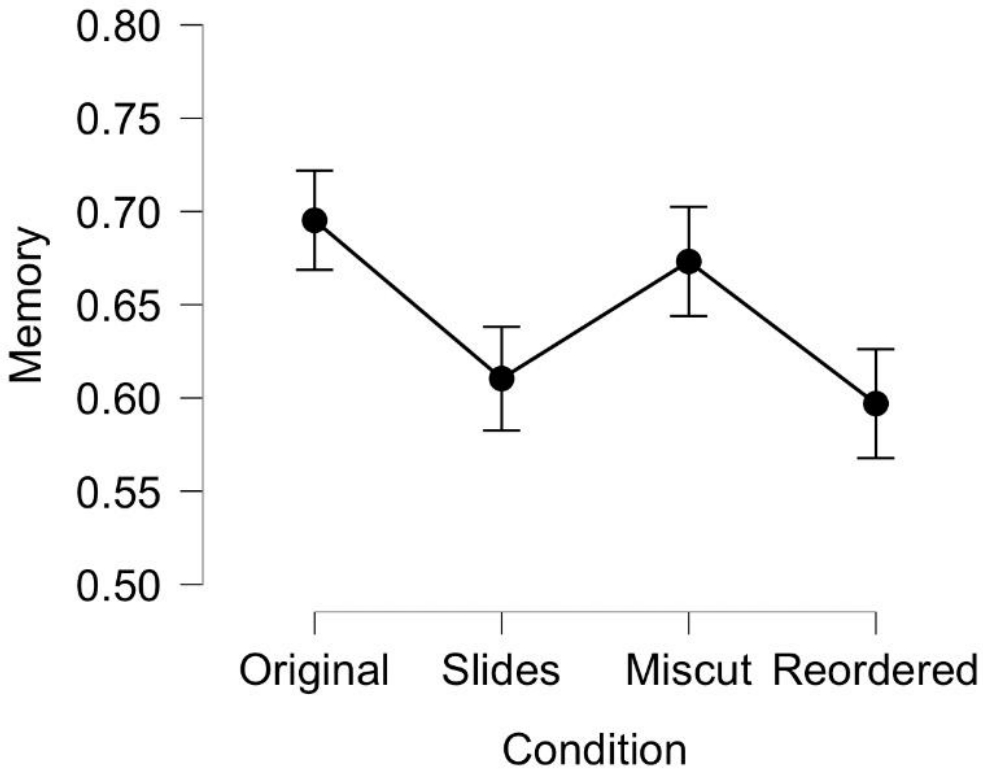
Conversely, memory performance was significantly better in the original condition of the Band Money scene, but not significantly worse in the reordered condition, suggesting that the facial expressions and other fine elements of acting only perceivable in the fluid (non-slideshow) original rendition of the visual context may be most critical to understand its dialogue. The subtle cues of the daughter's face when she is genuinely worried about her mother as opposed to when she is lying to her mother may be embedded in these finer elements of acting lost when the scene is shown in the slideshow condition (and the miscut and reordered conditions which are also in a slideshow format).

To better model the construct in question of disrupting visual context without dramatically changing the viewing experience, a future study might use extensive raw film footage to create stimulus with various edited conditions without having to rely on the dramatic and disruptive modification of creating slideshow conditions of the scenes. This was a dramatic enough change that the findings as to how much visual context impacted viewers' understanding of the scenes might be confused with the effect of viewing scenes in a mode (a slideshow) highly unexpected for the average viewer of modern cinema. Another more practical approach would be to select scenes which are more similar in filmmaking style, such as four scenes from a long film or a TV series (in which one would assume the style of filmmaking would remain relatively constant). While the scenes for this experiment were chosen for their richness of ground for writing theory of mind questions, with memory for conversation questions being able to be written for almost any scene of conversation, one could select a single series from a genre such as true crime or mystery where consistent gaps in character information create frequent theory of mind inferences. This approach should be sufficient for recreating this experiment using more consistent filmmaking practices to address the significant variation across scenes.

In conclusion, a future study can strengthen the significant findings for the impact of visual context on memory, as well as bring the nearly significant effect on theory of mind inference into a significant range by using scenes with more consistent filmmaking practices. In addition, multiple studies done on sets of scenes from different filmmaking styles can verify and compare the exploratory analysis done here on individual scenes suggesting that shot-reverse shot techniques, subtle moments of acting in moments of deception, and use of multiple physical spaces may each have distinct implications for what kinds of visual context impact viewers' understanding of cinematic conversation.

Figure 1

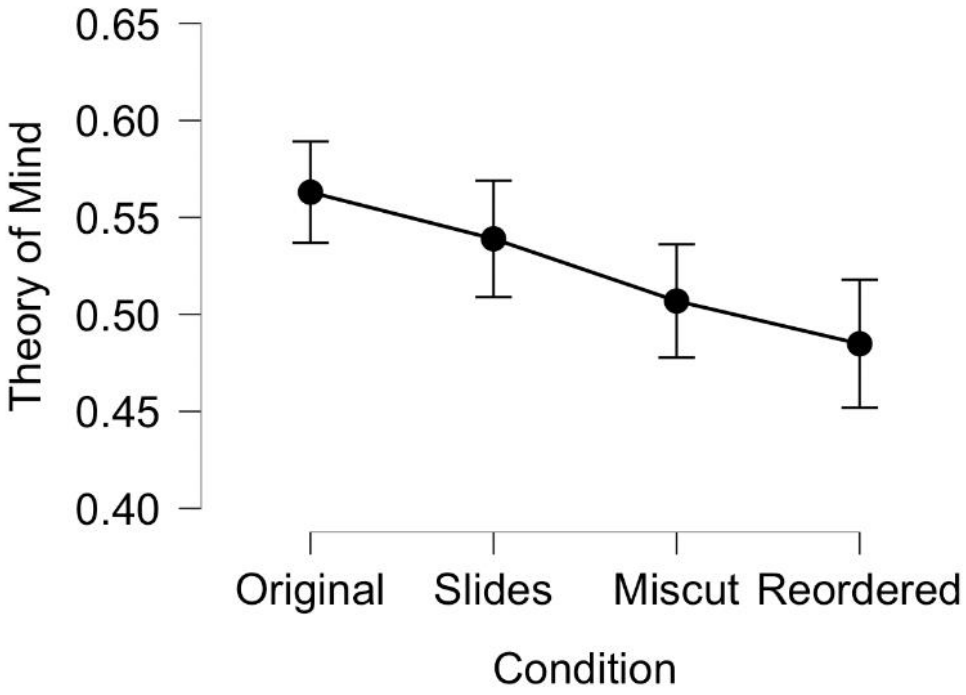
Effect of Condition on Memory for Conversation Performance Across All Scenes



Note. Error bars represent standard errors.

Figure 2

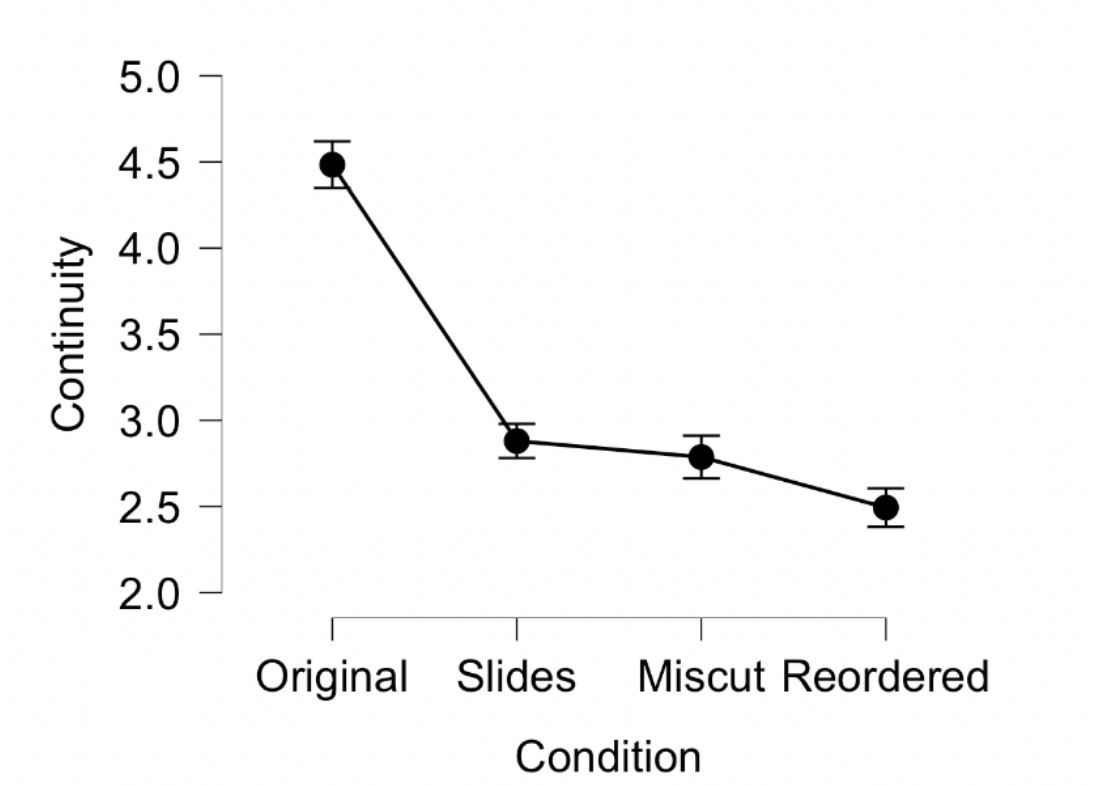
Effect of Condition on Theory of Mind for Conversation Performance Across All Scenes



Note. Error bars represent standard errors.

Figure 3

Effect of Condition on Continuity for Conversation Performance, Across all Scenes

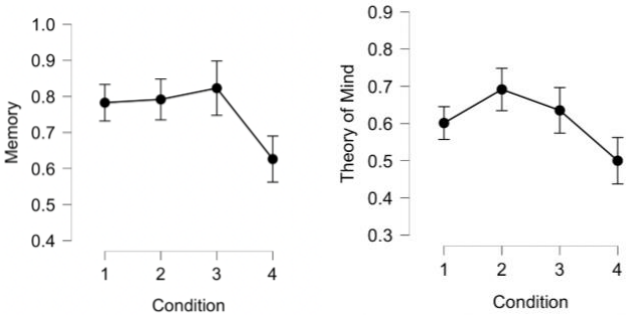


Note. Error bars are standard errors.

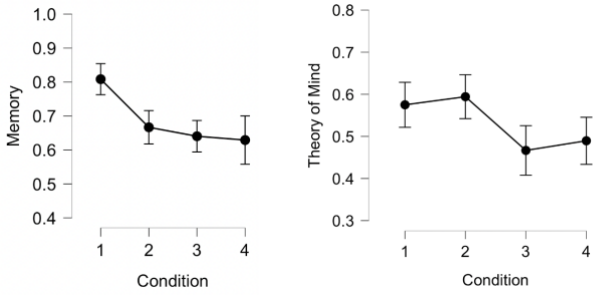
Figure 4

Scene-Specific Effects of Condition on Memory and Theory of Mind Performance

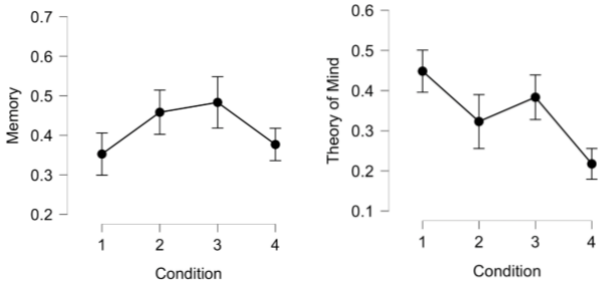
Job Offer



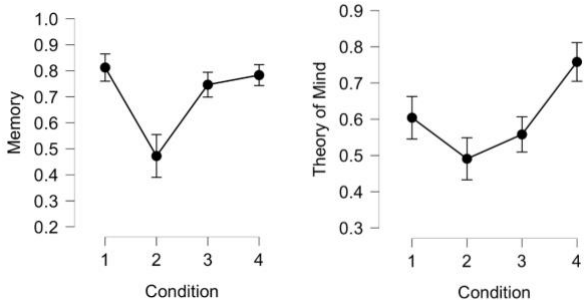
Band Money



Art Showing



Train



Note. Condition 1 is the original condition, Condition 2 is the slideshow condition, Condition 3 is the miscut condition, and Condition 4 is the reordered condition. Error bars are standard errors.

References

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of memory and language*, *57*(4), 502-518.

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, *65*(5), 825-839.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological science*, *15*(6), 415-419.

Bordwell, D., Thompson, K., & Smith, J. (1993). *Film art: An introduction* (Vol. 7). New York: McGraw-Hill.

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic bulletin & review*, *16*(5), 893-900.

Capone, N. C., & McGregor, K. K. (2004). Gesture development.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*(4), 596-615.

Keenan, J., MacWhinney, B., & Mayhew, D. (1977). Pragmatics in memory: A study of natural conversation. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 549–560. https://doi.org/10.1016/S0022-5371(77)80018-2

Knoeferle, P., & Kreysa, H. (2012). Can speaker gaze modulate syntactic structuring and thematic role assignment during spoken sentence comprehension? *Frontiers in Psychology*, *3*, 538.

Knutsen, D., & Le Bigot, L. (2012). Managing dialogue: How information availability affects

collaborative reference production. *Journal of Memory and Language, 67*(3), 326-341.

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*, *8*(5), e1436.

Levin, D., Brown-Schmidt, S., & Watson, D. (2019). *McDonnell Film Grant* [Grant Proposal].

Levin, D. T., Hymel, A. M., & Baker, L. (2013). Belief, desire, action, and other stuff: Theory of mind in movies. *Psychocinematics,* 244-266. doi:10.1093/acprof:oso/9780199862139.003.0013

Levin, Daniel T., and Kai Keliikuli. (2020). "An empirical assessment of cinematic continuity." *Psychology of Aesthetics, Creativity, and the Arts*.

McGurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748. doi:10.1038/264746a0

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103-128.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of personality and social psychology*, *37*(3), 322.

Scholl, B. J., Simons, D. J., & Levin, D. T. (2000). Implicit beliefs about change detection and change blindness. *PsycEXTRA Dataset*. doi:10.1037/e501882009-193

Slide, A. (2012). DW Griffith: Interviews. University Press of Mississippi: Jackson, MS.

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Faux pas recognition Test (Adult Version). *PsycTESTS Dataset*. doi:10.1037/t73330-000

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995).

Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632-1634.

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition, 73*(2), 89-134.

Zunshine, L. (2007). Why Jane Austen was different, and why we may need cognitive science to see it. *Style, 41*(3), 275-298.