Effect of sequential video shot comprehensibility on attentional synchrony: A comparison of children and adults

Heather L. Kirkorian[a,1] and Daniel R. Anderson[b]

[a]Human Development and Family Studies, University of Wisconsin–Madison, Madison, WI 53706; and [b]Psychological and Brain Sciences, University of Massachusetts Amherst, Amherst, MA 01003

Edited by David E. Meyer, University of Michigan, Ann Arbor, MI, and approved September 5, 2017 (received for review August 9, 2016)

To comprehend edited video, viewers must infer the meaning conveyed by successive video shots (i.e., continuous video segments separated by edit points, such as camera cuts). The central question here was whether comprehension-related top-down cognitive processes drive eye movements during sequential processing of video montage. Eye movements were recorded as 4 year olds and adults (n = 62) watched a video with the same constituent shots in either normal or random sequence. The key analyses compared eye movements to constituent shots when presented in normal order with those to the same shots presented in random order. The dependent variable was attentional synchrony or the extent to which viewers looked at the same location at the same time, indicating commonality of processing the video. This was calculated as the bivariate contour ellipse area within which points of gaze fell during each video frame. Results indicated that children were more scattered in their gaze locations than adults. Viewers became more similar to each other as normal vignettes unfolded over time; this was especially true in adults and possibly reflects a growing and shared understanding of the content. Conversely, adult attentional synchrony was reduced when watching random shot sequences. Thus, attentional synchrony during normal video viewing is driven not only by salient visual features, such as movement and areas of high contrast, but also, by the unfolding sequential comprehension of video montage, especially in adults. Differences between children and adults indicate that this top-down control of eye movements while watching video changes systematically over development.

development | cognition | television | eye movements | sequential comprehension

Despite the growing popularity of interactive screen technologies, the vast majority of children's media time is spent with television or other noninteractive video, which is often viewed on mobile digital devices (1). Film and television programs are ordinarily constructed as sequences of video shots (i.e., continuous video segments that are separated by edit points, such as camera cuts), from which overarching meaning emerges. As an example, a shot of the Eiffel Tower followed by a shot of a busy outdoor café would typically lead to the viewer's inference that the café scene takes place in Paris. A scripted television program consists of sequences individually conveying concepts, such as location, spatial layout, time transitions, implied events, character point of view, and many others. Such relations between shots are collectively called filmic montage. It has long been recognized that viewers have to learn how to comprehend filmic montage insofar as many shot sequences deviate from real-world perceptual experiences (2). By adulthood, most viewers readily extract the meaning from edited video sequences and are often not aware of the transitions between shots, such as cuts, fades, and wipes (3, 4). Research with adults who are inexperienced with audiovisual media indicates that comprehension of shot transitions often fails, suggesting that maturation and real-world experience alone are not enough for efficient processing of video (5).

Ordinarily, video producers avoid presenting images that are visually cluttered and complex. A television image is typically designed to have only one or two visual foci that represent the most informative portions of the image with respect to illustrating the ongoing scripted content. This may be particularly true of children's television programs (6). In addition, videos are edited to be coherent in the attempt to sustain attention as well as to encourage cognitive comprehension and emotional engagement. A successful video production directs visual attention to the most informative portions of the screen during each successive shot. It is highly desirable, therefore, that the audience looks at the same places on the screen at the same time. In line with this assertion, adults are more similar to each other in their visual fixations when viewing professionally edited Hollywood movie trailers than when viewing static scenes or videos of naturalistic scenes that have not been professionally edited (7). To the degree to which the audience is fixating the screen in common, it is likely that it is arriving at a common understanding of the video content. A corollary of this is that, to the degree to which the audience does arrive at a common understanding of the content, then fixation patterns should become progressively tighter with each new shot into the unfolding content.

Although transparently obvious to experienced adults, audiovisual media comprehension is much more challenging for young children. Only beginning at 18 mo, for example, do toddlers begin to look more at meaningful shot sequences compared with randomized sequences of the same shots, indicating that they have begun to perceive relationships between adjacent shots; by 24 mo, a preference for meaningful shot sequences is well-established (8, 9). Nonetheless, comprehension of edited video is likely a cognitively taxing activity for young viewers, with research indicating substantial growth in comprehension of simple edited video stories through middle childhood (10–12).

There has been relatively little research on how viewers process edited video in real time. The approach taken here is to examine eye movements during video viewing. With respect to edited video, the central question here is how eye movements to shots that are part of a meaningful video montage differ from eye movements to the same shots that occur in random order. One possibility is that eye movements are driven by bottom-up perceptual mechanisms

Aspects of this research were presented at the biennial meeting of the International Congress of Infant Studies, July 3–5, 2014, Berlin, Germany.

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Digital Media and Developing Minds," held, October 14–16, 2015, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Digital_Media_and_Developing_Minds.

Author contributions: H.L.K. and D.R.A. designed research; H.L.K. performed research; H.L.K. contributed new reagents/analytic tools; H.L.K. analyzed data; and H.L.K. and D.R.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: kirkorian@wisc.edu.

Published online October 1, 2018.

(also referred to as exogenous stimulus control) related to perceptual salience, such as movement or image regions of high contrast (13, 14). An additional possibility, explored here, is that, as a viewer develops a schematic understanding of the unfolding video montage, top-down cognitive processes (or endogenous control) increasingly drive eye movements. Top-down cognitive processes related to cognitive understanding of the content should be manifest in normal video and to a much lesser extent (if at all), in random shot sequences. If so, because adults are vastly more experienced in processing filmic montage compared with young children, adult eye movements should be relatively more influenced by disruptions in sequential shot comprehensibility.

When adults watch video, they exhibit attentional synchrony (15) [that is, a strong tendency to look at the same place on the screen at the same time as each other (7, 13, 16–21)]. This coherence between adult viewers is at least partly explained by common attentional capture by perceptual features, such as movement (13). In addition to the influence of exogenous features, attentional synchrony may reflect strong coherence in patterns of neural activation during film viewing (22) or in other words, a common perception and understanding of the unfolding content. In adult film viewing, individuals' memory and comprehension of the content are correlated with the degree to which strong synchronies in neural activation occur across individuals (23). When encountering new content that is unconnected with prior shots (as in the transition from television program to commercial), adults have a strong tendency to fixate the center of the screen, subsequently moving fixations to a point of central interest. In contrast, in response to shot changes within continuing content, they are much more likely to fixate other regions of the screen (13, 19, 20, 24). The latter finding likely reflects anticipation of the most informative region of the screen in the current shot as related to the immediately prior shot.

One study reported that adult attentional synchrony during film viewing could be successfully modeled without invoking top-down mechanisms involving anticipation based on prior shots (21). Nevertheless, shot anticipation in adult processing of edited video was verified using experimentally produced animations (25). In these animations, a series of shots showed a character or object moving continuously across a landscape. Following standard cinematic editing practices, the character or object disappeared from one edge of the screen only to reappear on the opposite edge. Adult viewers showed clear anticipation of this reappearance by moving their fixations to the opposite side of the screen; 4-y-old children, in contrast, reacted to the shot change more slowly and tended to center their gaze at the beginning of the next shot. This study shows that adults efficiently integrated spatial and action information across shots. This is consistent with other research showing that adults who receive contextual information (i.e., view preceding shots) exhibit greater attentional synchrony and are more likely to make critical inferences when viewing video clips (26). Unlike adults, 4 y olds tend to wait for a character or object to reappear on the opposite edge of the screen before they shift their gaze to track it (25).

Compared with adults, young children who are old enough to comprehend elementary filmic montage exhibit less attentional synchrony insofar as they are less likely to look at the same place on the screen at the same time as other children. They are also much more likely to center their gaze after each shot, regardless of whether the content is ongoing or new. This stands in contrast to infants, who not only show relatively little attentional synchrony but also, show no such tendency to center gaze after a shot boundary, indicating that they are probably not sensitive to shot transitions as informative features of video (19). Thus, on the one hand, preschool-aged children exhibit gaze patterns that suggest less systematic top-down processing than what is characteristic of adults' eye movements during video viewing. On the other hand, 4 y olds have substantial comprehension of age-directed, edited television programs, such as *Sesame Street* (11). Perhaps reflecting this comprehension, as shots within a content vignette progress, attentional synchrony among 4 y olds increases (19). These findings are suggestive that some top-down control of eye movements occurs in these young children but that, relative to adults, they are slow and less efficient.

In summary, the extant literature suggests that attentional synchrony may be the result of both bottom-up and top-down cognitive processes. Whether the age-related increase in attentional synchrony is primarily because of greater influence of top-down control remains unknown. Some eye-tracking evidence suggests that the importance of meaningful features, such as faces (relative to perceptually salient features), increases with age (16, 17). However, perceptual salience may be correlated with the locations of meaningful objects in a scene (27), making it difficult to separate effects of bottom-up exogenous control and top-down endogenous control using naturalistic stimuli. This study experimentally disentangles many of the factors that control eye movements within shots compared with between shots by using normal shot sequences compared with the same shots that occur in random order. In other words, we compared the same audiovisual content (largely controlling for bottom-up features) presented in normal sequences vs. random sequences to more directly assess the impact of top-down comprehension processes on attentional synchrony. To the extent that viewers efficiently guide eye movements to successive shots by their emerging comprehension of the content, they are able to exert top-down control over eye movements to normal video, and they should show greater disruption of attentional synchrony when the shots are in random order (21).

Overview of the Study

Adults and 4-y-old children were shown a version of the children's television program *Sesame Street* while their eye movements were recorded. This television program, while primarily intended as educational content for preschool children, also has a layer of humor directed at adult viewers to encourage parent coviewing. The program incorporates sophisticated editing that is characteristic of professionally produced television and incorporates multiple content segments of several minutes in duration that we refer to as vignettes. The story arc within each vignette is conveyed across multiple shots. There are thus three time courses over which eye movements unfold: across frames within the same video shot, across conceptually related shots within the same vignette, and across distinct vignettes that have no semantic relation. There were two experimental conditions: the normal, comprehensible sequence of shots in each vignette as originally broadcast and the same shots presented in random order within each vignette. Four year olds were used as the child comparison group, because they are at a prime age for viewing *Sesame Street*, and prior research has shown that they have a basic—albeit imperfect—comprehension of filmic montage (12). Sensitivity to canonical filmic montage first appears at about 18 mo and is well-established by about 24 mo (8, 9); however, comprehension of filmic montage continues to develop throughout early and middle childhood (10, 12).

Analyses consider attentional synchrony among viewers in the same age group and condition. That is, the analyses examine the extent to which the viewers looked at the same place on the screen at the same time as peers who watched the same video sequence. We predicted that eye movements to the normal version of *Sesame Street* would be more synchronous for adults than for children, consistent with greater strategic, comprehension-related information processing and greater experience viewing filmic montage. Such a finding would replicate earlier research (19).

Furthermore, to the extent that attentional synchrony among adults reflects ongoing comprehension, we expected synchrony to increase across successive shots within a vignette when adults viewed normal video sequences. In the random shot sequence

version, however, there is much less opportunity for top-down endogenous guidance of eye movements. Consistent with this prediction, comparisons with normal video have shown that random shot sequences engage different neural systems than normal shot sequences; in particular, the default mode network is not activated while watching random shot sequences (28, 29). Activation of the default mode network, in turn, is associated with sequential comprehension (30). Therefore, we predicted that, while watching random shot sequences, adult eye movements would be more variable and thus, more similar to child eye movements to the same shots. Given that 4-y-old children are able to comprehend *Sesame Street* but have much less experience than adults viewing edited video, it was an open question as to how much their eye movement patterns would change between normal and random shot sequences.

Results

The dependent variable was attentional synchrony across individuals as indicated by bivariate contour ellipse area (BVCEA). *Methods* has specific detail regarding BVCEA calculation. The unit of analysis was an individual frame, each of which was viewed by children and adults viewing either normal or random shot sequences. Thus, the general analytical design was a four-level hierarchical linear model (HLM) including age and condition (level 1), time into a shot (level 2), order of shots in a vignette (level 3), and order of vignettes in the video (level 4). To compare identical frames when viewed in normal vs. random sequence, one model included data from both conditions, with age, condition, and frame order (in seconds) as predictors. Given that shot order differed in the normal and random conditions, separate models examined the impact of shot order for each sequence. As an example of raw data, Fig. 1 presents scatterplots of gaze coordinates for 15 viewers in each of the four groups during one video frame.

Attentional Synchrony Within the Same Shot. Model 1 compared attentional synchrony among children and adults when watching the exact same shot as it was presented in either a normal or random sequence. This model also examined change in attentional synchrony over the course of a single shot. Results from model 1 are presented in Table 1 and plotted in Fig. 2.

The first set of effects in Table 1 (marked as constant) represents BVCEA for adults in the normal condition (i.e., when both age and condition = 0) and how attentional synchrony for this group changed over time into a single shot. There was a significant effect of frame order, such that adults' synchrony decreased over time into a shot (evidenced by greater BVCEA): $\delta_{0100} = 33.77$, SE = 3.37, $t(501) = 10.01$, $P < 0.001$.

The second set of effects in Table 1 shows the effect of age in the normal condition and the extent to which this age effect changed across individual shots. There was a significant main effect of age, such that children were less synchronous (larger BVCEA) than adults: $\delta_{1000} = 33.96$, SE = 2.81, $t(1,618) = 12.10$, $P < 0.001$. While this main effect was evident throughout shots, it decreased somewhat over time: $\delta_{1100} = -0.76$, SE = 0.28, $t(1,618) = -2.74$, $P = 0.007$.

The third set of effects in Table 1 depicts condition effects for adults (i.e., whether random video differs from normal video) and the extent to which this condition effect changed over time into individual shots. Adults' fixations were less synchronous (higher BVCEA) when watching random shot sequences than when watching shots in a coherent order: $\delta_{2100} = 8.21$, SE = 2.81, $t(1,618) = 2.93$, $P = 0.004$. This difference is consistent with the hypothesis that attentional synchrony among adults is at least partly the result of top-down endogenous control given that the audiovisual features that would drive bottom-up exogenous control were identical across the two conditions. The effect of frame time on condition was not significant ($P > 0.250$); thus, the effect of condition did not vary across a single shot.

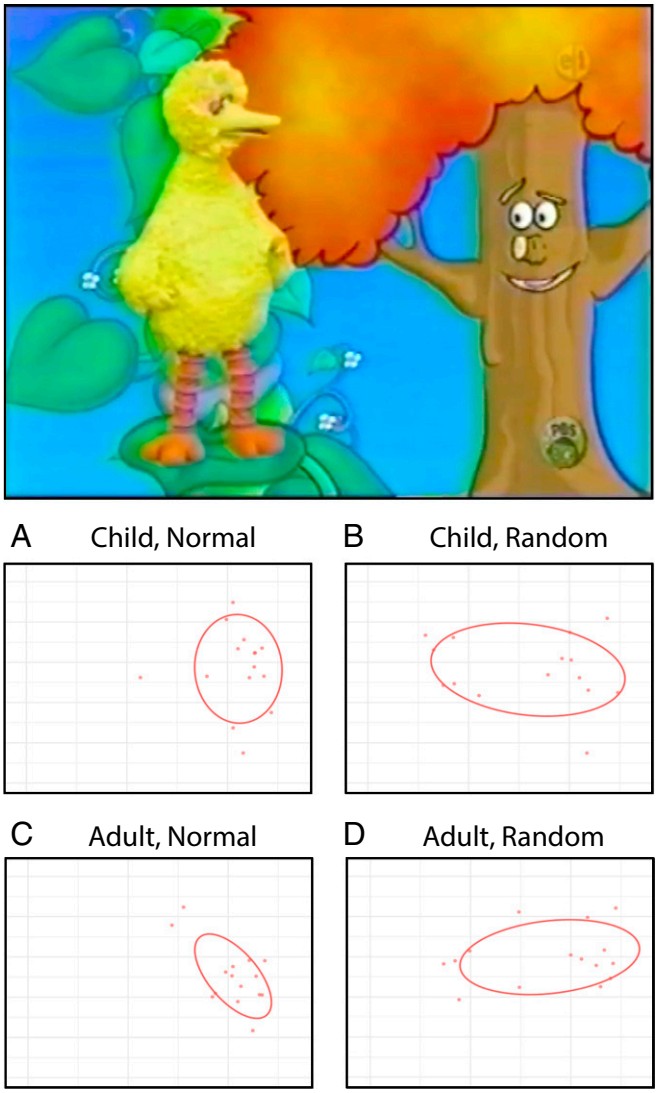

Fig. 1. Examples of contour ellipses around gaze coordinates. (*Upper*) A single video frame from the video. (*Lower*) Fifteen randomly selected gaze coordinates for children watching this shot in the normal sequence (*A*), children watching the random sequence (*B*), adults watching the normal sequence (*C*), and adults watching the random sequence (*D*).

The fourth set of effects in Table 1 depicts fixed effects for the condition by age interaction (i.e., the extent to which the condition effect differed between children and adults). The level 1 interaction term was significant: $\delta_{3000} = -8.35$, SE = 3.97, $t(1,618) = -2.10$, $P = 0.035$. Although adults were less synchronous when watching random sequences than when watching normal sequences, this was not the case for children. As with adults, the condition effect for children did not change significantly over time into a single shot ($P > 0.900$).

To summarize, adults' gaze locations were more similar to those of other adults when watching normal shot sequences than when watching the same shots in random sequences. This difference is consistent with the hypothesis that top-down control over eye movements in adults likely results in high attentional synchrony among viewers. Because these vignettes are easy for an adult to comprehend, there should be little difference between adults in their level of comprehension, and therefore, BVCEA should be small. In the random shot condition, in

Table 1. Fixed effects for model 1 examining attentional synchrony within the same shot

Predictor	Coefficient (SE)	*t* Ratio
Constant (δ_{0000})	33.77 (3.37)	10.01***
Frame within shot (δ_{0100})	1.04 (0.28)	3.76***
Age (δ_{1000})	33.96 (2.81)	12.10***
Age × frame (δ_{1100})	−0.76 (0.28)	−2.74**
Condition (δ_{2000})	8.21 (2.81)	2.93**
Condition × frame (δ_{2100})	−0.31 (0.28)	−1.12
Condition × age (δ_{3000})	−8.35 (3.97)	−2.10*
Condition × age × frame (δ_{3100})	0.03 (0.39)	0.09

The dependent variable was BVCEA (10° visual angle^2) for each group for each frame. Age was entered at level 1 as a dichotomous variable, with zero representing adults and one representing children. Condition was also a level 1 dichotomous variable, with zero representing the normal condition and one representing the random condition. Frame order (seconds since the start of the shot) was entered at level 2: df = 6 for the constant (δ_{0000}), df = 501 for frame order within a shot (δ_{0100}), and df = 1,618 for all other predictors. *P* values are $^*P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$.

contrast, adults likely cannot apply a common real-time comprehension scheme to a vignette. Consequently, cognitive interpretations of the shot sequence likely vary across individuals, explaining an increase in values of BVCEA. In comparison with adults, children were relatively less synchronous, and there appears to be relatively little impact of random shot sequences on eye movement patterns of young children. These findings are consistent with the hypothesis that younger viewers are less able to engage top-down comprehension processes in real time to guide anticipatory eye movements when watching normal shot sequences, resulting in relatively high variability across individuals, regardless of whether the shots are presented in normal or random order.

Attentional Synchrony Across Multiple Shots. To the extent that attentional synchrony across individual viewers is caused by ongoing comprehension processes, we would expect comprehension and thus, synchrony to increase over time in a coherent video sequence but not in a random sequence. To test this hypothesis, separate HLM4 models examined changes in BVCEA across coherent vs. random shot orders. Table 2 presents fixed effects for model 2 (normal sequences) and model 3 (random sequences). Fitted lines from these two separate models are overlaid in Fig. 3 for ease of visual comparison.

When viewing shots in normal order, adults' gaze locations became more synchronous across successive shots: $\delta_{0010} = -0.77$, SE = 0.26, $t(44) = -2.92$, $P = 0.006$. Also consistent with the hypothesis that increasing attentional synchrony reflects increasing comprehension of events within a vignette, there was no increase in synchrony across shots in a random sequence ($P > 0.500$).

As in model 1, children's eye movements were more scattered than those of adults in both the normal and random conditions: $\delta_{1000} = 20.74$, SE = 3.82, $t(500) = 5.43$, $P < 0.001$ vs. $\delta_{1000} = 20.46$, SE = 3.65, $t(500) = 5.61$, $P < 0.001$, respectively. Of greater interest here was the extent to which attentional synchrony varied across a sequence of shots. In the normal condition, children's gaze became more synchronous across shots in the same vignette but to a marginally lesser extent than that of adults: $\delta_{1010} = 0.51$, SE = 0.22, $t(500) = 2.33$, $P = 0.020$. In the random condition, children (like adults) did not exhibit any change in BVCEA across shots ($P > 0.900$).

Together, these findings are consistent with the hypothesis that adults (and to a lesser extent, young children) are able to use top-down strategies to guide visual attention when viewing normal shot sequences. When the same shots are presented in a random sequence, these strategies are disrupted, thus limiting viewers' ability to engage in common comprehension.

Discussion

There have been two broad (not necessarily mutually exclusive) points of view of the factors that control eye movements to video. One is that fixations are controlled by visual factors individually associated with each image. That is, the eyes are attracted to regions of high visual contrast, movement, and other visual features that are known to produce visual orienting, even in infants (13, 17). Subsequently, viewers may home in on a portion of the image that is most informative (21). The other broad point of view is that attention to the screen in general and fixations to specific portions of the images are driven by comprehension of the content as it unfolds across filmic montage. Experience gained viewing video, therefore, provides the viewer with efficient comprehension-related strategies of eye movements that can guide fixations across successive shots (19, 25).

While there is prior evidence for both bottom-up (13, 17, 21) and top-down (14, 25) factors controlling eye movements to video, this experiment attempts to developmentally disentangle them using identical video content that is presented in either coherent or incoherent sequences. By removing the logical or pragmatic connections between shots (that is, by placing them in random order), the primary factors that control eye movements are likely those that are found within each individual shot. In contrast to normal videos, where comprehension has the potential to grow over successive shots, there should also be a growing realization by viewers that a random sequence of shots provides little relevant information useful in guiding eye movements.

The findings for adults were quite consistent with hypotheses that both top-down and bottom-up cognitive and perceptual processes drive eye movements to video. Normal shot sequences led to eye positions clustered in relatively small regions of the screen. As shots within vignettes progressed, these regions became even smaller, consistent with the hypothesis that adults acquire a growing understanding of the content and greater expectations as to the nature of succeeding shots. Random shot sequences, however, greatly increased the variability between adult viewers as to where they looked in each shot. There was little decline in this variability between viewers as additional

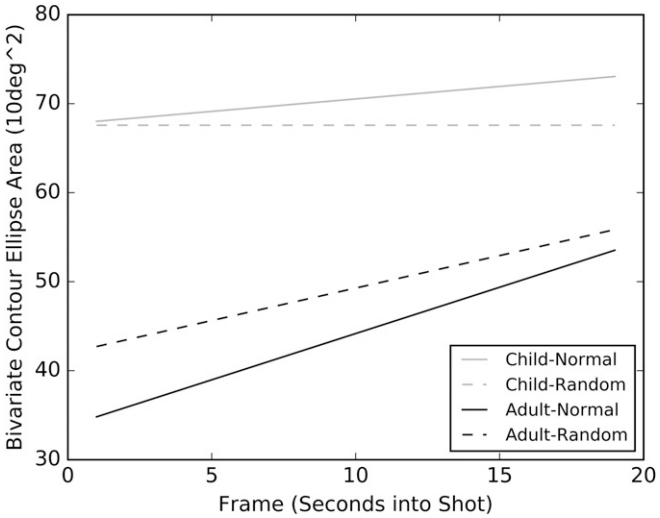

Fig. 2. Fitted lines for model 1 examining attentional synchrony as a function of age, condition, and frame (seconds into the shot). Table 1 shows fixed effects. The dependent variable was BVCEA (10° visual angle^2) for a group's gaze coordinates to a single frame ($n = 561$).

Table 2. Fixed effects for models 2 (normal) and 3 (random) examining age and shot order effects on attentional synchrony in each condition

	Model 2: normal		Model 3: random	
Predictor	Coefficient (SE)	*t* Ratio	Coefficient (SE)	*t* Ratio
Constant (δ_{0000})	53.00 (4.79)	11.07***	49.14 (5.56)	8.84***
Shot within vignette (δ_{0010})	−0.77 (0.26)	−2.92**	−0.20 (0.36)	−0.55
Age (δ_{1000})	20.74 (3.82)	5.43***	20.46 (3.65)	5.61***
Age × shot (δ_{1010})	0.51 (0.22)	2.33*	0.002 (0.28)	0.007

The dependent variable was BVCEA (10° visual angle^2). Age was entered at level 1 as a dichotomous variable, with zero representing adults and one representing children. Shot order (centered at the first shot in each vignette) was entered at level 3: df = 6 for the constant (δ_{0000}), df = 44 for shot order (δ_{0010}), and df = 500 for all other predictors. *P* values are $*P < 0.05$, $**P < 0.01$, and $***P < 0.001$.

shots from the vignette were presented in random order. This evidence is consistent with the position that adults anticipate content in succeeding shots during normal video and that they use this information to guide eye movements. Moreover, the fact that adults showed greater attentional synchrony than children to individual shots in the random sequences is suggestive that they have, with maturation and viewing experience, developed efficient strategies for finding informative visual features even when sequential comprehensibility is absent (21).

Compared with the findings for adults, the findings for children were quite different. Consistent with earlier research (19), this study found much greater variability in gaze locations among 4 y olds relative to adults. This indicates that children are less likely than adults to look at the same part of the screen at the same time as their peers. The differences in synchrony between normal and random shot sequences for children were relatively small, indicating little effect of sequentially meaningful shot relationships on eye movement control. In fact, if random shot sequences had any impact on the eye movements of children, it would be in the opposite direction than would be expected by comprehension-related top-down control of eye movements across shots. That is, children had slightly (albeit nonsignificantly) greater attentional synchrony during random sequences than normal ones. One possible explanation is that, during random shot sequences, children tended to center their gaze in the middle of the screen for each and every shot, thus reducing individual differences in fixation point. However, analyses of gaze distance from the center did not support this interpretation. We suggest instead that greater variability in gaze location among children during normal shot sequences likely reflects individual differences in filmic montage comprehension and perhaps, in overall content understanding. Some children may be in the early stages of anticipating the next shot, but others simply center their gaze with each new shot, regardless of shot coherence. Alternatively, less gaze variability during random sequences may represent a default mode of visual attention, wherein gaze is simply directed at the most salient portion of the image.

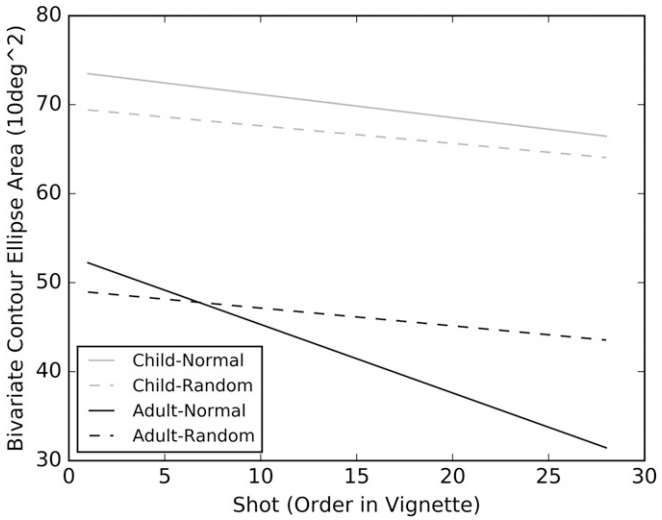

Fig. 3. Modeled attentional synchrony as a function of age, condition, and shot (order within vignettes). The dependent variable was BVCEA (10° visual angle^2) for a group's gaze coordinates to a single frame ($n = 561$). Plots for model 2 (normal condition; solid lines) and model 3 (random condition; dashed lines) were calculated separately (Table 2 shows fixed effects) but are overlaid here for ease of comparison.

We created experimental stimuli using *Sesame Street* to observe visual attention in a seminaturalistic viewing situation. While this approach provides some degree of external validity, it leaves some questions unanswered. Future research with carefully controlled audio and visual stimuli is needed to more directly examine the interplay between bottom-up and top-down processes in video viewing. Such research would also be useful in determining the specific features of random video sequences that decreased attentional synchrony in adults in this study (e.g., less perceptual overlap between adjacent shots at shot boundaries, inability to use audio information to predict what will happen in the next shot). It is also critical to complement these findings with direct assessments of content comprehension in both children and adults, particularly as it relates to attentional synchrony and ongoing narrative comprehension. Nonetheless, this study represents a critical step in bridging the research on filmic montage comprehension by young children and attentional synchrony among adults during video viewing.

Television, video, and television-like digital applications have become extraordinarily important aspects of modern children's lives (1). Video screens on a wide variety of large-screen and mobile devices have become principal sources of formal and informal education as well as entertainment. Especially with respect to education, it is essential to provide productions that are informed by evidence of how children process video and how that processing changes with development. This study shows that preschool children's patterns of eye movements are quite different from those of adults and that they process video montage differently. These differences likely arise from slower processing speed and above all, more difficulty in processing transitions between shots. An implication is that educational media producers should not rely on adult perceptions alone to produce effective content for young children (31, 32). Instead, for video and video-like interactive productions, judicious shot pacing and image composition should be high priorities for producers. Educational video programs (and video-like interactive applications) can and should make use of evidence-informed principles,

such as these, to produce engaging and comprehensible educational content.

Methods

Participants and Design. Participants were 30 4-y-old children (11 female, 19 male; mean age of 4.52 y old; range = 4.36–4.74 y old) and 32 adults (26 female, six male; mean age of 20.80 y old; range = 18.17–30.49 y old). An additional three children were excluded from analyses because of inability to calibrate visual fixation points; 12 adults were excluded for the same reason (e.g., because they wore glasses). Equal numbers of viewers in each age group were randomly assigned to watch normal, comprehensible video or random edit, sequentially incomprehensible video. Adults were studied, because they have high comprehension of editing techniques and exhibit relatively synchronous gaze patterns compared with younger viewers (16, 17, 19). Additionally, adults exhibit greater attentional synchrony when watching comprehensible shot sequences than when watching randomly edited video segments (21). Four year olds were studied, because prior research shows that young children have only a limited understanding of editing conventions relative to older children (12) and have fewer synchronous gaze patterns relative to adults when watching normal shot sequences (19).

The study was approved by the Institutional Review Board at the University of Massachusetts Amherst. Adult participants were recruited from undergraduate psychology courses. Names and addresses for potential child participants came from a database of birth records for children in western Massachusetts. Each family received a letter describing the project and a follow-up phone call requesting participation. This database typically results in a response rate of 10–20%. When participants arrived at the laboratory, an experimenter explained the study and offered to answer questions. Then, adult participants and parents signed an informed consent document.

Parents of child participants also completed a survey to report demographics and home media use. The majority (90%) of parents identified their child as white or Caucasian; the remaining 10% identified their child as biracial or mixed race. As a proxy for socioeconomic status, parents reported the number of years of education completed by each parent, with 12 y typically indicating a high school diploma, 16 y typically indicating a 4-y college degree, and so on. The average number of years per parent was 17.03 y (SD = 3.54 y, range = 10–28 y).

Parents reported their child's use of screen media at home by completing a retrospective viewing diary for a typical week. Each day (Monday through Sunday) was divided into 30-min intervals from 6:00 AM to 10:00 PM. Parents indicated times during which the television was typically on while the child was in the room. Of those times, parents distinguished between foreground exposure (when the child would typically be watching a program) and background exposure (when the child would be in the room with the television on but not watching a child-directed program). Parents reported daily averages of 1.94 h of foreground television (SD = 1.44 h, range = 0–5.93 h) and 0.65 h of background television (SD = 1.18 h, range = 0–4.57 h). Parent-reported television exposure did not predict the dependent variable of interest (attentional synchrony), and it did not differ significantly by condition; therefore, it was not considered further.

Video Stimulus. The video stimulus was a compilation of 10 vignettes from the children's program *Sesame Street*. The entire video lasted ∼19 min. Vignettes represented a mix of live action and computer-generated imagery as well as real-life actors, puppets, and animated characters. The narrative vignettes contained an average of 19 distinct shots (range = 3–47). Individual shots averaged 6 s in length but ranged from less than 1 to almost 40 s. The only difference between the two experimental conditions was the order of shots within each vignette. In the normal condition, the shots were played in their original order, creating a cohesive story for each vignette. In the random condition, the shots (including the audio) within each vignette were played in a random sequence. Questions regarding the videos can be directed to H.L.K.

In the random condition, the shots were separated either at the exact moment of an abrupt jump cut or at the midpoint of a wipe across the screen. Thus, to the extent possible, the formal features (visual and auditory characteristics) of each vignette remained largely unchanged in the random sequence, but the sequence of events within each vignette was disrupted, rendering each narrative less comprehensible. While it is possible that adjacent video shots are more perceptually similar than those that occur farther away from each other in a sequence and that randomly ordering shots within each vignette incidentally increased the perceptual differences between adjacent shots, such differences are unlikely to explain the condition effects observed in this study. To rule out this possibility, we calculated the Visual Activity Index (VAI) (33) immediately before and after each cut. This index describes the similarity in luminance in two separate images (in our case, each pixel of the video frame that appeared immediately before vs. after each cut). The average VAI across all frame pairs was identical for the normal and the random edit video (normal video: mean = 0.70, SD = 0.14, range = 0.34–1.11; random video: mean = 0.70, SD = 0.19, range = 0.11–1.16). Therefore, any condition effects observed in this study are unlikely to be caused by systematic differences in the perceptual similarity of adjacent shots, giving more confidence to the interpretation that condition effects are caused by differences in viewers' ongoing comprehension of each vignette.

Another video was used to calibrate the eye-tracker cameras. This video consisted of small animated images alternating between the top left and bottom right corners of the screen. Each image was on the screen for 4 s. In other research, this procedure has resulted in equally accurate calibration across age groups insofar as there were no significant age differences in horizontal and vertical variability of the samples that were averaged to calculate gaze location for the two calibration points. Because the same calibration procedure was used for all participants before they viewed the *Sesame Street* video, any condition effects that are reported here are not likely to be explained by systematic differences in the quality of the calibration.

Setting and Apparatus. The study took place in a laboratory space on a university campus. A curtain was hung from the ceiling to separate the control area from the viewing area where participants watched the video (1.93 × 2.54 m). The stimulus display screen, a 66-cm television set (standard 330-line resolution screen,

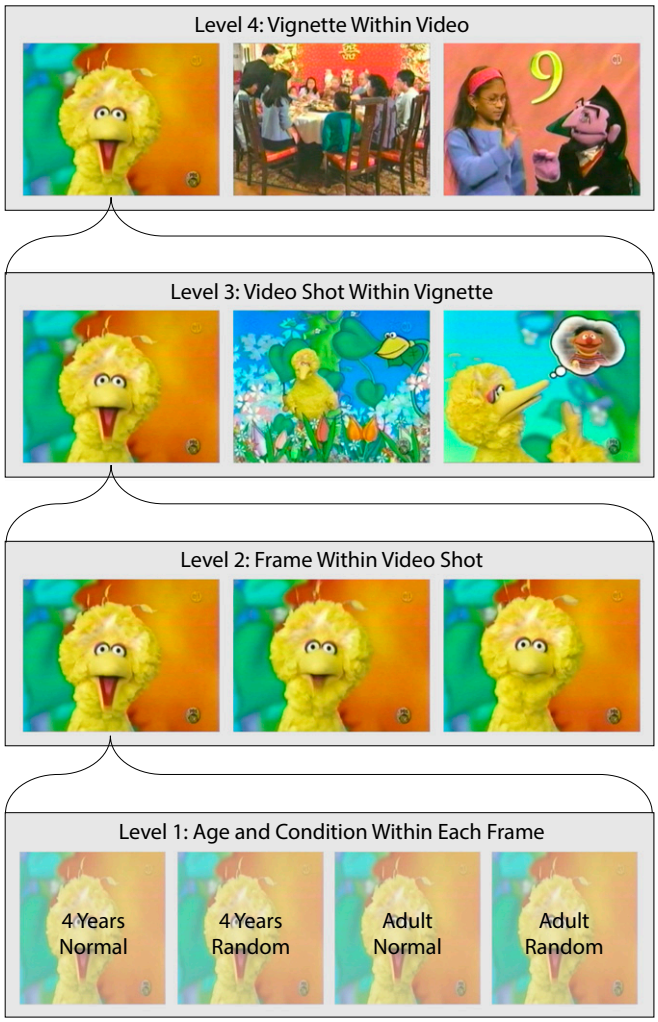

Fig. 4. Graphical representation of the analytical design. Age and condition (level 1) were nested within frames (level 2). Frames were nested within shots (level 3). Shots were nested within vignettes (level 4). © 2017 Sesame Workshop.® Sesame Street® and associated characters, trademarks, and design elements are owned and licensed by Sesame Workshop. All rights reserved.

4:3 aspect ratio), was centered along one short wall. The visible image on the display screen was 40.60-cm high and 54.60-cm wide. Given the dimensions of the screen and typical distance from the viewer (100 cm), the video image subtended ~23° visual angle vertically and ~30.5° visual angle horizontally.

A chair was positioned ~65 cm from the eye-tracking apparatus for optimum focus and ~100 cm from the television screen; children sat on a booster seat to approximate the height and viewing angle of adults. Parents sat in a chair to the right of the child participants. Dark curtains hung along the walls on all sides of the viewing area. In the control area on the other side of the curtain divider, a switcher allowed the experimenter to toggle between the calibration and video stimuli as needed.

The eye-tracking cameras were mounted on a tripod and located between the participant and the stimulus display screen (30.5 cm from the television). The eye camera was the Applied Science Laboratories (ASL) Eye-Trac 6000 Series, a near-IR corneal reflection system with remote pan-tilt optics. Effective accuracy is claimed to be 0.25° visual angle (about 0.25 cm^2 at a distance of 50 cm), and temporal resolution is 60 Hz (60 data points per second). The head-tracking camera was the ASL VH2 model that uses face recognition software to locate the viewer's head in space. The pan-tilt eye camera used information from the video head tracker to follow the participant's head movements and minimize data loss.

A second computer was used to keep the participant's eye on the camera, calibrate the eye tracker, and save gaze data. A recording deck was used to capture digital video of the scene with overlaid eye cursor. An ASL Digital Frame Overlay was also used, which overlaid the digital frame from the ASL Control Unit onto the scene video, allowing the experimenter to sync the data file from the user interface computer with the digital video record.

Procedure. On entering the study room, the participant was seated in front of the video display screen. Parents remained in the room with child participants, and they were asked to refrain from directing their child's attention to any particular area on the screen after the stimulus video began. The two-point (top left and bottom right) "quick calibration" procedure was used for all participants. Adult participants were asked to look at each image, whereas children were asked to "play a guessing game" by identifying the images (e.g., mouse, robot) as they appeared. After calibration, the experimenter started the stimulus video and began recording the gaze file and the digital video. Throughout the video, the experimenter ensured that the head tracker was enabled and that the eye camera was focused on the eye.

Data Reduction. We adapted the analytical procedure described by Kirkorian et al. (19). In particular, attentional synchrony was measured using BVCEA, where larger values represented larger ellipses that fit around more widely scattered gaze coordinates, thus reflecting less synchrony among viewers. The unit of analysis was a single video frame. Fig. 1 shows examples. We wrote scripts using Python 3.4 to reduce the data and to calculate BVCEA for each video frame. Data files and code can be found at https://uwmadison.box.com/s/raohe6rxz7e9i1anue2wlldynfmx17c0.

The first step in data reduction was to identify usable data points and to smooth the data. We analyzed raw gaze coordinates (rate of 60 data points per second) instead of clustering the data into fixations to include more data from the children and minimize any impact of systematic data loss. Raw gaze coordinates were considered usable if they (*i*) had valid eye data (e.g., pupil size > 0), (*ii*) had a gaze coordinate that landed within the bounds of the video screen, and (*iii*) occurred during the stimulus video. These criteria were met for just under one-half of all possible data points for the children (48% in the normal condition, 45% in the random condition) and the majority of data points for adults (83% in the normal condition, 84% in the random condition). Because we used raw gaze coordinates rather than aggregated fixations, the raw gaze coordinates were smoothed by averaging each *x* and *y* coordinate with the previous three data points.

Smoothed gaze coordinates were then matched to a corresponding frame in the stimulus video based on the exact moment when each gaze coordinate occurred. Time stamps in the normal and random videos were synced to compare the same frames when they were presented in their original order and when they were presented in a random sequence. If an individual participant had two valid data points for a given video frame, then only the first data point was considered for analysis.

After each participant's gaze coordinates were synced with the corresponding stimulus video, data were further reduced to ensure that synchrony estimates were based on a majority of participants in both age groups and conditions and that the same video frames were analyzed for all groups. For this step, frames were sampled at a rate of 4 Hz (i.e., one frame every 250 ms). As in the work by Kirkorian et al. (19), an individual video frame was included in analyses only if it had usable gaze data from more than one-half of the individuals in each category (minimum of nine participants per group). This criterion was met for the majority of frames for adults (91% in the normal condition, 92% in the random condition). However, this criterion was met for only 34% of frames for children watching normal video and 23% of frames for children watching randomly edited video. When matching frames across both age groups and both conditions, 12% of video frames had usable gaze data from at least nine individuals within each of the four groups. We also analyzed the data using less stringent criteria to include more video frames (e.g., lowering the minimum number of participants to eight increased the number of usable frames to 24%). The general pattern of results was the same. Because our outcome measure is sensitive to the number of samples included in calculating BVCEA, we adopted more stringent criteria for the analyses presented here.

After eligible frames were identified, we calculated attentional synchrony for each group and for each frame using the same formula for BVCEA as that used in prior studies of between-subject differences in eye movements to video (18, 19). In brief, this formula produced the spatial area of the best fit bivariate contour ellipse surrounding at least 61.33% of the data points for each group for each frame. The equation was

$$\mathrm{BVCEA} = 2k\pi\delta_H\delta_V\left(1-p^2\right),$$

where δ_H was the SD of the horizontal gaze coordinates, δ_V was the SD of the vertical gaze coordinates, p was the product–moment correlation between horizontal and vertical values, and k was the enclosure; k was calculated as

$$p = 1 - e^{-k},$$

where p was the proportion of points included in the ellipse. In this study, k = 0.95, such that 61.33% of data points were required to be included in the best fit ellipse. This is similar to the threshold used in previous research (18, 19).

Analysis Plan and Preliminary Analyses. The unit of analysis was an individual video frame. However, each frame was viewed by four groups and therefore, had four BVCEA values (children vs. adults; normal vs. random). Moreover, each frame was nested within a video shot, which was then nested within a vignette. Therefore, we utilized HLM to analyze BVCEA as a function of age, condition, and time. We created four-level nested models using the HLM 7 software package. Fig. 4 shows a graphical representation. At level 1, we compared the BVCEA for the exact same video frame when it was presented to children vs. adults and in a normal vs. random sequence. We also considered time into a single video shot at level 2, the order of shots within a single vignette at level 3, and the order of vignettes in the video at level 4. Thus, we were able to consider the effect of age and condition over time into a single shot, over one story arc, and across multiple story arcs.

The frequency distributions of frames and shots were positively skewed, such that most frames occurred within the first 20 s of an individual shot and within the first 30 shots of an individual vignette. Preliminary analyses examined the overall pattern of results when including only those frames that occurred during this early period of individual shots and earlier shots in the vignettes. Trimming the dataset in this way did not change the general findings. In addition, BVCEA can be affected by the number of data points included in the analysis. Thus, preliminary analyses also examined the overall pattern of results when the data file included exactly nine data points per group per frame (randomly selected from all possible data points for that frame). Again, the general pattern of findings did not differ when trimming the data file in this way. The results presented here are based on data for frames that occurred at any point within a shot or vignette. Finally, preliminary analyses indicated that adding vignette order did not improve any of the HLM4 models; therefore, this predictor was not considered in the analyses presented here.

ACKNOWLEDGMENTS. This research was supported by National Science Foundation Grant BCS-0623888. Findings and opinions expressed in this manuscript do not reflect endorsement by the National Science Foundation.

1. Rideout V (2017) The Common Sense Census: Media use by kids age zero to eight. Available at www.commonsensemedia.org. Accessed November 9, 2017.
2. Münsterberg H (1970) *The Film: A Psychological Study: The Silent Photoplay in 1916* (Dover Publications, Mineola, NY).
3. Kraft RN (1986) The role of cutting in the evaluation and retention of film. *J Exp Psychol Learn Mem Cogn* 12:155–163.
4. Smith TJ, Henderson JM (2008) Edit blindness: The relationship between attention and global change blindness in dynamic scenes. *J Eye Mov Res* 2:1–17.

5. Ildirar S, Schwan S (2015) First-time viewers' comprehension of films: Bridging shot transitions. *Br J Psychol* 106:133–151.
6. Wass SV, Smith TJ (2015) Visual motherese? Signal-to-noise ratios in toddler-directed television. *Dev Sci* 18:24–37.
7. Dorr M, Martinetz T, Gegenfurtner KR, Barth E (2010) Variability of eye movements when viewing dynamic natural scenes. *J Vis* 10:28.
8. Anderson DR, Lorch EP, Field DE, Sanders J (1981) The effects of TV program comprehensibility on preschool children's visual attention to television. *Child Dev* 52: 151–157.
9. Pempek TA, et al. (2010) Video comprehensibility and attention in very young children. *Dev Psychol* 46:1283–1293.
10. Calvert SL, Scott MC (1988) Television production feature effects on children's comprehension of time. *J Appl Dev Psychol* 9:263–273.
11. Lorch EP, Bellack DR, Augsbach LH (1987) Young children's memory for televised stories: Effects of importance. *Child Dev* 58:453–463.
12. Smith R, Anderson DR, Fischer C (1985) Young children's comprehension of montage. *Child Dev* 56:962–971.
13. Mital PK, Smith TJ, Hill RL, Henderson JM (2010) Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognit Comput* 3:5–24.
14. Smith TJ, Mital PK (2013) Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *J Vis* 13:16.
15. Smith TJ, Henderson JM (2008) Attentional synchrony in static and dynamic scenes. *J Vis* 8:773.
16. Franchak JM, Heeger DJ, Hasson U, Adolph KE (2016) Free viewing gaze behavior in infants and adults. *Infancy* 21:262–287.
17. Frank MC, Vul E, Johnson SP (2009) Development of infants' attention to faces during the first year. *Cognition* 110:160–170.
18. Goldstein RB, Woods RL, Peli E (2007) Where people look when watching movies: Do all viewers look at the same place? *Comput Biol Med* 37:957–964.
19. Kirkorian HL, Anderson DR, Keen R (2012) Age differences in online processing of video: An eye movement study. *Child Dev* 83:497–507.
20. Tosi V, Mecacci L, Pasquali E (1997) Scanning eye movements made when viewing film: Preliminary observations. *Int J Neurosci* 92:47–52.
21. Wang HX, Freeman J, Merriam EP, Hasson U, Heeger DJ (2012) Temporal eye movement strategies during naturalistic viewing. *J Vis* 12:16.
22. Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R (2004) Intersubject synchronization of cortical activity during natural vision. *Science* 303:1634–1640.
23. Hasson U, Furman O, Clark D, Dudai Y, Davachi L (2008) Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. *Neuron* 57:452–462.
24. Le Meur O, Le Callet P, Barba D (2007) Predicting visual fixations on video based on low-level visual features. *Vision Res* 47:2483–2498.
25. Kirkorian HL, Anderson DR (2017) Anticipatory eye movements while watching continuous action across shots in video sequences: A developmental study. *Child Dev* 88: 1284–1301.
26. Loschky LC, Larson AM, Magliano JP, Smith TJ (2015) What would Jaws do? The tyranny of film and the relationship between gaze and higher-level narrative film comprehension. *PLoS One* 10:e0142474.
27. Einhäuser W, Spain M, Perona P (2008) Objects predict fixations better than early saliency. *J Vis* 8:18.1–18.26.
28. Anderson DR, Fite KV, Petrovich N, Hirsch J (2006) Cortical activation while watching video montage: An fMRI study. *Media Psychol* 8:7–24.
29. Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N (2008) A hierarchy of temporal receptive windows in human cortex. *J Neurosci* 28:2539–2550.
30. Nakano T, Kato M, Morito Y, Itoi S, Kitazawa S (2013) Blink-related momentary activation of the default mode network while viewing videos. *Proc Natl Acad Sci USA* 110:702–706.
31. Anderson DR (2004) Watching children watch television and the creation of Blue's Clues. *Nickelodeon Nation: The History, Politics, and Economics of America's Only TV Channel for Kids*, ed Hendershot H (New York Univ Press, New York), pp 241–268.
32. Hirsh-Pasek K, et al. (2015) Putting education in "educational" apps: Lessons from the science of learning. *Psychol Sci Public Interest* 16:3–34.
33. Cutting JE, DeLong JE, Brunick KL (2011) Visual activity in Hollywood film: 1935 to 2005 and beyond. *Psychol Aesthetics Creativity Arts* 5:115–125.