Behavior Research Methods
https://doi.org/10.3758/s13428-022-01832-5

Measuring event segmentation: An investigation into the stability of event boundary agreement across groups

Karen Sasmita[1] · Khena M. Swallow[1]

Accepted: 11 March 2022

Abstract

People spontaneously divide everyday experience into smaller units (*event segmentation*). To measure event segmentation, studies typically ask participants to explicitly mark the boundaries between events as they watch a movie *(segmentation task*). Their data may then be used to infer how others are likely to segment the same movie. However, significant variability in performance across individuals could undermine the ability to generalize across groups, especially as more research moves online. To address this concern, we used several widely employed and novel measures to quantify segmentation agreement across different sized groups ($n = 2–32$) using data collected on different platforms and movie types (in-lab & commercial film vs. online & everyday activities). All measures captured nonrandom and video-specific boundaries, but with notable between-sample variability. Samples of 6–18 participants were required to reliably detect video-driven segmentation behavior within a single sample. As sample size increased, agreement values improved and eventually stabilized at comparable sample sizes for in-lab & commercial film data and online & everyday activities data. Stabilization occurred at smaller sample sizes when measures reflected (1) agreement between two groups versus agreement between an individual and group, and (2) boundary identification between small (*fine-grained*) rather than large (*coarse-grained*) events. These analyses inform the tailoring of sample sizes based on the comparison of interest, materials, and data collection platform. In addition to demonstrating the reliability of online and in-lab segmentation performance at moderate sample sizes, this study supports the use of segmentation data to infer when events are likely to be segmented.

Keywords Event segmentation · Segmentation agreement · Naturalistic perception · Event cognition · Online data collection

Introduction

Everyday perception involves segmenting experience into distinct units (*events;* Zacks et al., 2001a, b; Zacks et al., 2007). This process, *event segmentation*, is typically studied by asking participants to watch a movie and simultaneously mark the *boundaries* between events with button presses (*segmentation task;* Newtson, 1973; Newtson & Engquist, 1976). In healthy adults, segmentation task performance tracks measurable changes in brain activity (Speer et al., 2007; Zacks, Braver, et al., 2001a), cortical representational states (Baldassano et al., 2017), and cognitive function (Faber et al., 2018; Swallow et al., 2009) during task-free video watching. The event segmentation task thus appears to reflect a cognitive process that helps shape everyday cognitive function (Richmond et al., 2017). However, because the segmentation task is relatively unstructured, quantifying and interpreting performance in this task poses unique analytical challenges. In this paper, we examine the stability and sensitivity of measures of segmentation task performance, focusing on measures that capture the degree to which groups and individuals agree with each other on when event boundaries occur in naturalistic stimuli (*segmentation agreement*).

In a typical segmentation task, participants are asked to press a button whenever they believe that one natural and meaningful unit of activity has ended and another has begun (Newtson, 1973; Newtson & Engquist, 1976; Zacks et al., 2001a, b). By design, participants are given limited directions and are told to rely on their own judgment when marking event boundaries. Despite the task's ambiguity, participants tend to press the button at similar

✉ Khena M. Swallow
kms424@cornell.edu

[1] Department of Psychology, Cornell University, 211 Uris Hall, Ithaca, NY 14850, USA

Published online: 19 April 2022

times during the videos, resulting in moments that most people are likely to report an event boundary (*normative boundaries*), and other moments that people are unlikely to report an event boundary. Further, when instructed to identify event boundaries at multiple levels of granularity (Kurby & Zacks, 2011; Newtson, 1973), performance on the segmentation task reflects the hierarchical structure of goal-directed activities: smaller *fine-grained* units (e.g., grinding coffee beans, pouring hot water) are grouped within larger *coarse-grained* units (e.g., making coffee; Hard et al., 2006; Hard et al., 2011; Zacks et al., 2001a, b). In addition, the consistency of boundary identification within and across individuals (Newtson, 1973; Speer et al., 2003) as well as its relationship to identifiable features of the videos (Hard et al., 2006; Magliano et al., 2001; Newtson et al., 1977; Swallow et al., 2018) supports the growing practice of using segmentation data from one group to examine the effects of event boundaries on cognitive and neural activity in another group (e.g., Ben-Yakov & Henson, 2018; Faber et al., 2018; Swallow et al., 2009).

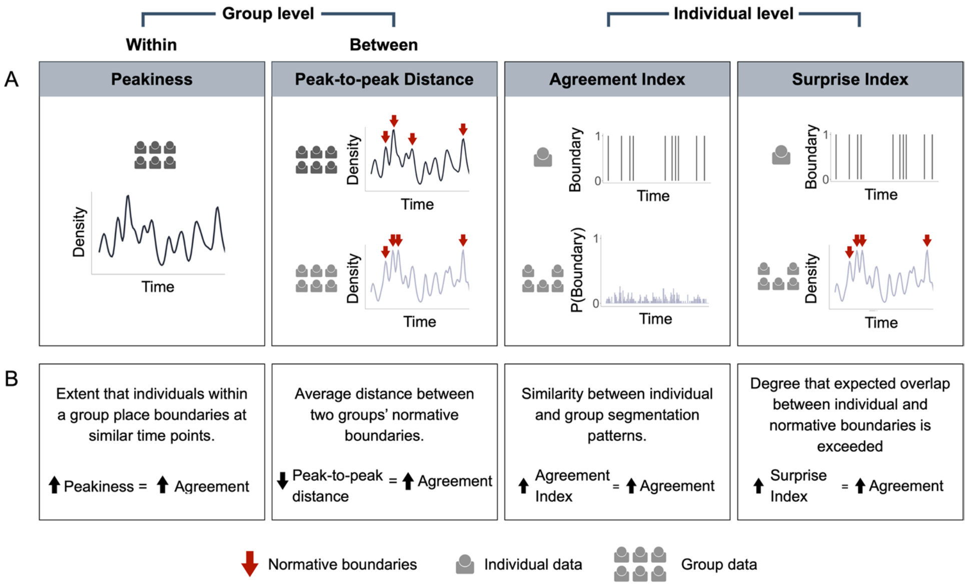
However, performance on the segmentation task can differ markedly from one person to the next. Variability in task performance across individuals could reflect real differences in how individuals segment events (Bläsing, 2015; Kurby & Zacks, 2011; Levine et al., 2017; Newberry et al., 2021; Papenmeier et al., 2019; Sargent et al., 2013; Swallow & Wang, 2020). It could also reflect a variety of other factors, including differing interpretations of ambiguous task instructions, attentional lapses, erroneous button presses, or accidental misses. Some of these factors may be exacerbated by online data collection, a practice that is increasing and which offers little control over either the context in which data collection occurs or the hardware that is being used. Factors influencing performance error may also play a larger role in the segmentation of realistic depictions of everyday experiences (e.g., lab-produced videos of everyday activities) than of videos that have been structured by their creators (e.g., directors and editors of commercial film) to increase engagement and similarity in viewing patterns across individuals (Dorr et al., 2010; Hutson et al., 2017; Loschky et al., 2015).

Individual differences and errors in segmentation task performance should reduce the degree to which normative boundaries identified by small groups are predictive of individual task performance. Yet, studies that use one group's segmentation behavior to infer event segmentation during task-free viewing have used groups that vary widely in sample size ($N = 1$–41) and expertise (Ben-Yakov & Henson, 2018; Chen et al., 2017; Faber et al., 2018; Kosie & Baldwin, 2019; Levine et al., 2017; Swallow et al., 2009). These studies have provided a wealth of data pointing to the segmentation task's effectiveness in capturing meaningful behavior. However, the assumption that segmentation data from a small number of individuals is sensitive to the commonly perceived structure of naturalistic stimuli has received little formal attention.

Estimates of segmentation task performance may also be influenced by how it is quantified, and a variety of approaches appear in the literature. Most simply, group performance can be quantified as the rate of boundary identification (e.g., Bläsing, 2015; Jeunehomme & D'Argembeau, 2018). However, because researchers are often interested in when those button presses occur, they typically generate a *group time series* by calculating the proportion of participants that identified a boundary within discrete time windows, or *bins* (often 1–5 s in duration) that span the entirety of the video (e.g., Massad et al., 1979; Zacks et al., 2001a, b). The binned group time series can then be correlated with that from another group (as in, e.g., Hard et al., 2006) or with the button presses of individual participants (Kurby & Zacks, 2011; Newberry & Bailey, 2019; Swallow & Wang, 2020; Zacks et al., 2006). Alternatively, the group time series has been used to identify intervals during which most participants press a button (normative boundaries; as in, e.g., Massad et al., 1979). One group's normative boundaries can then be compared to another group's (Massad et al., 1979) or to the boundaries identified by individuals from a separate group (Zalla et al., 2013).

Confidence in the methods used to quantify segmentation task agreement is justified by the repeated replication and validation of results that relied on them. However, there are several concerns that result from their use. Metrics that rely on correlations or signal detection theory may have requirements that may not necessarily be true of segmentation data, including that individual and group time series values are normally distributed, are linearly related, and have homoscedastic variance, that outliers are rare or not present, and that the range of values is not restricted. For example, quantifying individual performance using d' (e.g., Zalla et al., 2013) assumes that normative boundaries capture true signal (i.e., the presence of a boundary in specific bins) and that individuals generate button presses based on signals embedded in normally distributed noise. In this paper we therefore introduce the *surprise index* (Katori et al., 2018) as an alternative way to measure segmentation agreement. This metric captures the degree to which the observed overlap between individual and group button presses exceeds expectations (based on the Poisson distribution) without imposing assumptions about how individuals generate button presses.

Currently employed metrics also only assess relative agreement—whether two groups agree with each other or whether individuals agree with groups about when boundaries occur—and therefore do not quantify segmentation agreement within a single group of participants. To address this gap, we introduce a new metric, *peakiness*, that measures within-group agreement by quantifying variability in the group time series over time.



Extent that individuals within a group place boundaries at similar time points.

↑ Peakiness = ↑ Agreement

Average distance between two groups' normative boundaries.

↓ Peak-to-peak distance = ↑ Agreement

Similarity between individual and group segmentation patterns.

↑ Agreement Index = ↑ Agreement

Degree that expected overlap between individual and normative boundaries is exceeded

↑ Surprise Index = ↑ Agreement

Fig. 1 Illustration (**A**) and description (**B**) of group- and individual-level agreement measures. Group time series are illustrated as the density of button presses over time (peakiness, peak-to-peak distance, and surprise index) or as the proportion of participants that pressed a button within a 1-s-long time bin (agreement index). Individual time series are represented as vertical lines marking button presses at every 1-s time bin (for agreement index) or continuously over time (for surprise index). Normative boundaries are defined as the times of the highest n-peaks, where n = mean number of button presses

The standard practice of generating group time series by calculating the proportion of participants that pressed the button within time bins poses additional challenges. One major concern is that the selection of bin size can vary markedly across studies (from 1 to 15 s; e.g., Swallow et al., 2018; Boggia & Ristic, 2015) and is not always explicitly motivated in the literature. Bin size could have outsized effects on the results of an analysis: whereas small bin sizes run the risk of preserving variance from sources that do not operate at the group level (e.g., lapses in attention and motor speed), large bin sizes may obscure the effects of brief, but relevant sources of variation (e.g., changes in the movie content). The choice of bin size thus creates another, often overlooked source of variation in results across studies. A second major concern is that binning the data in this way distorts the continuous nature of events. It treats a boundary as a binary event that is either present or absent during a discrete interval of time, rather than as an entity that could be probabilistically distributed over time. Even if boundaries are punctate, binning them potentially treats two button presses that are far from each other as the same and two button presses that are near each other as different, depending on whether or not they fall within the same bin.

Recognition that binning segmentation data may not adequately capture segmentation task performance is evident in a growing number of studies that have characterized boundary identification as probabilistic over time (Huff, Maurer, et al., 2017a; Huff, Papenmeier, et al., 2017b; Newberry et al., 2021; Smith et al., 2020). In these studies, group time series are created by aggregating the density of participants' button presses over time. The density of button presses is estimated by centering a Gaussian kernel on each button press. These approaches preserve the continuity of performance in a segmentation task, but, similar to approaches that bin the data, do not describe methods for determining the temporal resolution of the group time series (in this case set by the bandwidth of the Gaussian kernel). Like these more recent studies, we characterize group time series probabilistically, but estimate group button press density over time from pooled individual data. We also describe a method for systematically determining the bandwidth of the kernel used to generate the group time series, thereby avoiding the need for potentially arbitrary decisions about the unit of time in which boundaries can occur.

To address concerns surrounding the quantification of segmentation task performance, this paper systematically examines the efficacy of both new and existing measures of segmentation agreement (illustrated in Fig. 1). The measures we examined quantified agreement in boundary placement within a group of individuals (*peakiness*), between two groups of individuals (*peak-to-peak distance*), and between an individual and a separate group (*agreement index,* Zacks et al., 2006; and *surprise index,* adapted from Katori et al., 2018). In all but one case (agreement index), the measures are based on probabilistic, continuous group time series, with a clear basis for determining the bandwidth of the kernel used to generate them. We evaluated these measures according to three criteria: (1) the ability to distinguish segmentation behavior from noise, (2) the ability to distinguish segmentation of one movie from another, and (3) the sample size needed for the measure to stabilize and show little improvement with the addition of more participants. The impact of sample size on these metrics is of particular interest because it influences the ability to measure segmentation task performance in two ways: first, by contributing to the stability of the group time series across samples, and second, by influencing the power of statistical tests. Therefore, to examine the influence of sample size on metrics of segmentation agreement, we first bootstrapped estimates of each agreement measure for samples of different sizes. This allowed us to characterize measures of segmentation agreement across samples. Characterizing agreement measures across samples is important because all measures of agreement reflect sample (group)-level behavior, not just the behavior of an individual (Fig. 1). For example, in the agreement index, individual performance is evaluated by referencing it to the group, making stable estimates of group segmentation behavior a central component of this metric. We therefore fit a linear mixed-effects model to the sample estimates to characterize how each measure differs when used on samples of real data, random data, and data with conflicting signals, as well as how the measure changes as sample size increases. Finally, we estimated the number of participants needed to distinguish segmentation data from noise or conflicting signals within a single sample.

We assessed the utility of segmentation agreement metrics across different modes of data collection and types of videos using two independent data sets collected for separate projects. The first consisted of segmentation data collected in the lab with commercially made movies (Sasmita & Swallow, in prep). The second consisted of segmentation data collected online with videos depicting everyday activities (Swallow & Wang, 2020). While lab-based experiments offer a more controlled environment, online experiments offer benefits such as the capacity to recruit from a more diverse population and efficiency in collecting data from a large number of participants (e.g., Birnbaum, 2004). Additionally, videos of everyday activities can be more realistic, but lack the structure and richness of commercial films (Cutting et al., 2011) that may influence segmentation by systematically guiding viewers' attention over time (Dorr et al., 2010; Hutson et al., 2017; Loschky et al., 2015). Notably, the mode of data collection and the materials are confounded in our data such that those conditions expected to increase error in task performance (i.e., online segmentation of unedited videos) are combined. Rather than comparing these data sets, we limit our investigation to whether segmentation agreement metrics can be reliably applied to segmentation data collected online using videos depicting everyday activities as well as to data collected in-lab with commercial movies.

Materials and methods

This paper evaluates the sensitivity of several segmentation agreement metrics to the presence of structure in group and individual data. Sampling distributions for each metric were created using a bootstrapping procedure. This procedure created samples of different sizes by randomly drawing individual data from data sets collected for two other projects (Sasmita & Swallow, in prep; Swallow & Wang, 2020). The projects differed in the type of video stimuli used, how the data were collected, and the sampled population. The first project (*commercial-lab*) examined segmentation of excerpts from commercially produced movies. These data were collected in the lab with undergraduate participants. The second project (*everyday-online*) utilized unedited videos of actors performing everyday activities produced by the lab. Data were collected online with participants recruited through Amazon's Mechanical Turk and CloudResearch (Swallow & Wang, 2020). Data reported in this paper and the code used for data processing and analyses are available on GitHub: https://github.com/ksasmita/esMethods.

Participants

All participants provided informed consent and all procedures were approved by the Cornell Institutional Review Board.

Commercial-lab

Participants were recruited from the Cornell University community. Analyses reported in this paper focused on data from 64 participants (27 male, 37 female) between the ages of 18 and 41 ($M = 21.04$, $SD = 4.27$). Participants were compensated with course credit or $15 for their time.

Everyday-online

Participants were recruited through Amazon's Mechanical Turk using CloudResearch. Data were collected from participants in India and in the United States, but analyses in this paper only utilized data from participants in the United States ($N = 72$; 41 male, 31 female; 19–58 years old; $M = 33.79$, $SD = 8.26$). Participants were compensated $5 or $7.50 for their time.

Experimental design

Commercial-lab

Video stimuli were constructed using excerpts from two commercial movies: *3 Iron* (Kim, 2004) and *Corn Island*

(Ovashvili, 2015). These movies were selected because they depict distinct activities (e.g., cooking in the kitchen, sawing and hammering wood), have little to no dialogue, and are set in naturalistic settings.

For each movie, excerpts were created by extracting scenes defined by natural breakpoints created by an editor's cut, scene or location changes, or breaks in the narrative. Each scene depicted one or more activity (e.g., the actor picking up dirty clothes from the floor, then washing the clothes and hanging the clothes on a drying rack) and consisted of multiple shots. The extracted scenes were then joined together to form one continuous video (*3 Iron* = 9.82 min; *Corn Island* = 9.43 min). Scenes in the final video were not necessarily contiguous in the original movie, but their order was preserved. The audio tracks for both videos were removed.

The videos were divided into ten 1-min-long *clips* (the last clip was shorter than 1 min). For half of the participants, the 10 clips from each movie were presented without interruption and in order (*uninterrupted*). For the other half of the participants, the clips were presented in order but with 3 s of white noise interrupting the clips (*interrupted*). Data from a third group were collected for the original project but were not included in the analyses reported in this paper.

Everyday-online

Stimuli were videos of actors performing everyday activities set in the United States: doing laundry (5.41 min), making coffee (5.73 min), doing the dishes (6.03 min), and making a bed (4.55 min). The original project also used recordings of the same four activities set in India; however, data from this condition were not included in the following analyses.

Videos were recorded using a GoPro HERO4 silver edition (1920 × 1080 pixels, 29.98 frames per second [fps], narrow field of view). The activities were performed by different actors and were filmed from several feet away. Therefore, in each video, an actor can be seen performing one activity within one room, with no changes in camera angle or scene cuts. Videos were presented with the audio tracks removed.

Segmentation task

All participants segmented every video that they watched. They were instructed to press the spacebar every time they believed one natural and meaningful unit of activity had ended and another had begun. Participants were instructed to mark the smallest (*fine* grain) and largest (*coarse* grain) units of activity change. They were told to press the spacebar as many times as necessary to mark all of the units in the video.

All experiment sessions began with a practice segmentation task. For *commercial-lab*, a short practice video was constructed by extracting a 1.5-min clip from the movie *3 Backyards* (Mendelsohn, 2010). For *everyday-online,* a short practice video (1.65 min) was recorded depicting an actor sitting outside, eating snacks, looking through books, and using his phone. To increase similarity across participants and reduce overlap between fine and coarse conditions, practice was repeated until participants reached a performance criterion. The criterion was based on mean unit durations reported in earlier literature (e.g., Zacks et al., 2001a, b) or on pilot segmentation data with the practice video (as described in Swallow & Wang, 2020): *commercial-lab,* mean duration of coarse = 11.25–30 s (3–8 button presses), fine = 2.5–6 s (15–36 button presses); *everyday-online,* mean duration of coarse = 16.5–49.5 s (2–6 button presses), fine = 3.3–8.25 s (12–30 button presses).

Participants in *commercial-lab* performed coarse and fine segmentation on one movie before moving on to segmenting the next movie. In contrast, participants in *everyday-online* segmented all videos in one grain first before moving on to segmenting the videos again in the other grain. Video and grain orders were fully counterbalanced in *commercial-lab*. Although most participants in *everyday-online* started with coarse grain segmentation due to a programming error, segmentation patterns between participants who started with fine segmentation and those who started with coarse segmentation were similar (Swallow & Wang, 2020).

Following segmentation of each video, participants in *commercial-lab* performed a free recall task, typing everything they remember happening in the video into a document, while participants in *everyday-online* completed some questionnaires. The following analyses examined only segmentation task performance.

Data processing

Our analyses focused on four methods of quantifying group- and individual-level segmentation agreement. We examined a summary of group-level agreement (*peakiness*), a comparison of segmentation performance between two groups (*peak-to-peak distance*), and two measures that evaluate individual segmentation performance by comparing it to the group (*agreement index* and *surprise index*). Measures were selected based on their use in prior research or their potential to provide additional insight into the consistency of segmentation patterns across individuals.

For each participant, button presses that occurred within 500 ms of the previous button press were excluded from the analysis. This removed button presses that were likely to reflect recording artifacts (e.g., holding down the keyboard button) rather than genuine button presses. Segmentation data also were restricted to the length of the shortest videos to allow for comparison of segmentation performance across different movies (commercial-lab = 566 s, everyday-online = 273 s).

Creating data sets of different sample sizes

We used a bootstrapping procedure to estimate the distribution of each agreement measure when *subsamples* of different sizes were used (Singh & Xie, 2010). For each sample, we randomly selected with replacement *n* individuals from the group of interest 1000 times, where *n* ranged from 2 to 32, incremented by 2. We then obtained a sample-level estimate of agreement for each movie for use in subsequent analyses. This process was performed separately for each agreement measure. Resampling with replacement makes it likely that one or more individuals' data will be replicated several times within a single subsample, particularly when the *n* is large. This could raise the concern that sampling with replacement could inflate measures of agreement[1]. However, at $n=32$, the means of the bootstrapped agreement measures were not systematically greater than the values for our full samples (*actual values*; Appendix A), which fell within the 95% confidence interval of bootstrapped estimates in almost all cases. The bootstrapping approach employed in our analysis thus provides reasonable estimates of agreement for samples of different sizes.

Calculating measures of segmentation agreement

Group-level agreement For measures of group-level agreement, we characterized group segmentation performance as the observed probability density of a button press over time (group density). Using the density function in base R (stats package; R Core Team, 2018), the subsample of participants' button press times were combined and every time point was smoothed using a Gaussian kernel. The kernel's bandwidth was determined using a smoothing function that accounts for variations in the data and normalizes the density distribution so that it has an area under the curve equal to 1 (option "SJ" in the density function; Sheather & Jones, 1991). For each grain, the computed bandwidth was multiplied by a value that generated a group time series with visually distinctive peaks and valleys for all sample sizes ("adj" parameter: coarse = 0.1, fine = 0.05; Fig. 2a). Density was estimated for the duration of the movie, padded by a pre- and post-movie window to capture the probability function around button presses at the beginning and end of the movie. This also allowed the density estimate to approach zero at those times. The size of the window added to the beginning and end of the movie was twice the final bandwidth of the smoothing function, rounded to the nearest whole number. *Normative boundaries* were defined as the time points of the highest *j* peaks in the group density distribution, where *j* is the mean number of button presses generated by participants in the group. Whenever the number of observed peaks was lower than the average number of group button presses, *j* was defined as the maximum number of peaks identified in the group density estimate. This method provides clear criteria for generating probabilistic group time series and defining normative boundaries. It therefore avoids the need for what may be arbitrary decisions about the bandwidth of the smoothing kernel or unit of time at which the data should be binned.

Peakiness To quantify agreement among individuals in a single group, we introduce a new measure that we refer to as *peakiness*. This measure reflects the expectation that greater amounts of within-group agreement will lead to group density functions with higher "peaks" and lower "valleys" (i.e., the group density time series are less flat). For every subsample, we calculated the moment-to-moment change in amplitude of the group density function over the duration of the movie (rugosity) using the rugo function (seewave package; Sueur et al., 2008). This value was then scaled by dividing the observed rugosity by the minimum possible rugosity value given the number of button presses and the density algorithm. The minimum was defined as rugosity when the combined participants' button presses were uniformly distributed over time. A high peakiness value, therefore, suggests high within-group agreement on boundary placement.

Peak-to-peak distance To measure the consistency of normative boundaries identified across groups, we calculated the peak-to-peak distance. This measure is defined as the mean distance between the normative boundaries defined by two groups, with a value of 0 indicating perfect agreement.

To calculate peak-to-peak distance, we first set the number of normative boundaries to examine as the minimum of the following three values: the number of peaks in the group density time series for group 1, the number of peaks in the group density time series for group 2, and the mean of the number of normative boundaries in group 1 and group 2 (rounded to the nearest whole number). We then obtained a set of normative boundaries for each time series and calculated the distance (in seconds) between the normative boundaries in one group to the normative boundaries in the other group (and vice versa). These distances were then averaged to form the peak-to-peak distance metric.

Individual-level agreement The next two measures quantify agreement between boundaries identified by an individual and those identified by an independent group of participants. The first measure, agreement index, has been commonly used in prior research (Kurby & Zacks, 2011). We applied the second

[1] This is unlikely: Repeated occurrences of an individual with idiosyncratic patterns of segmentation may increase their agreement with the group, but it would decrease agreement for other group members. Repeated occurrences of an individual with typical segmentation patterns should result in representative group time series.

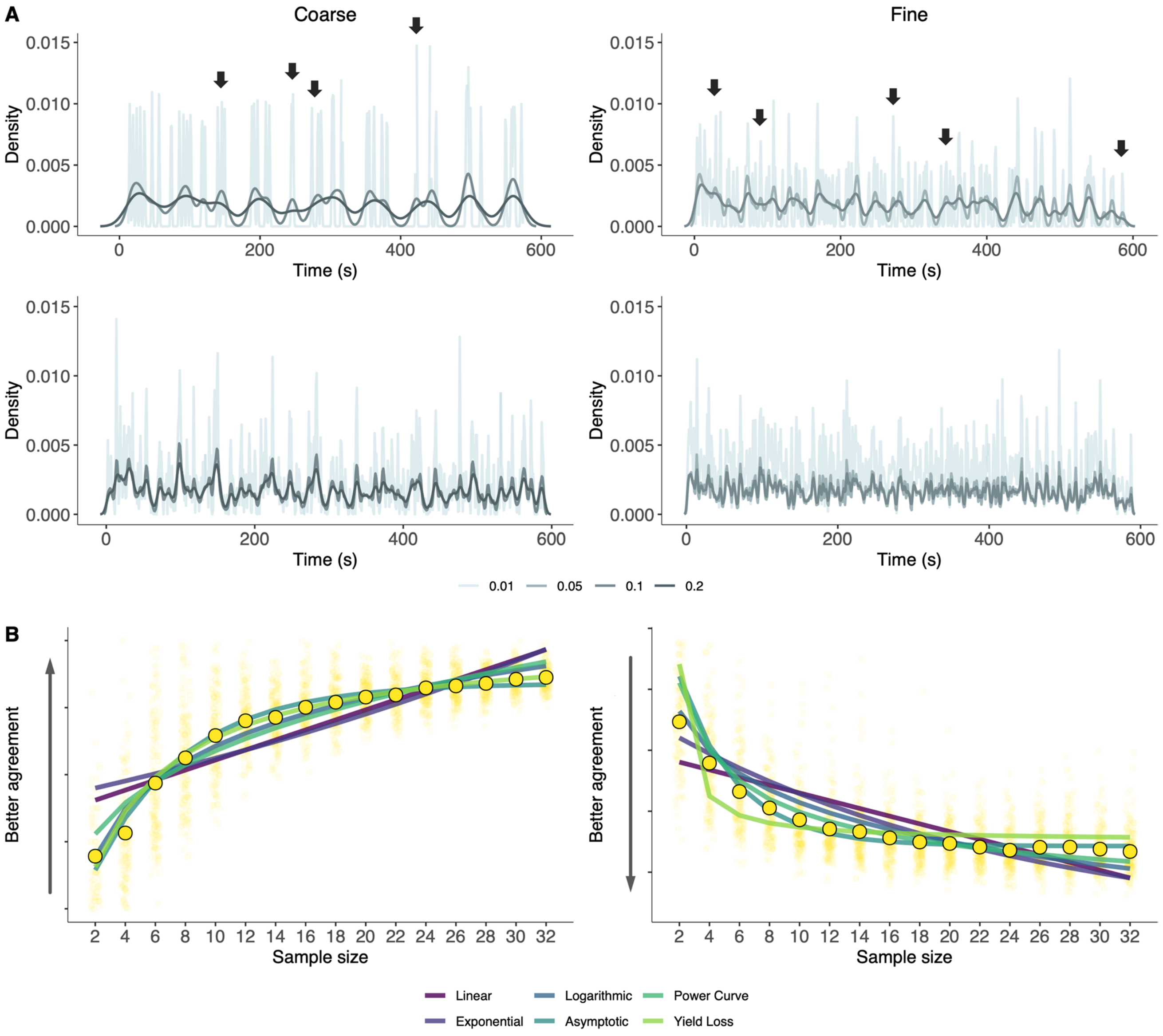


Fig. 2 (**A**) Example of density estimates with different bandwidth adjustments for small ($n=2$; upper panel) and large ($n=32$; lower panel) sample sizes. In all cases, the lower adjustment value (0.01) seems to capture individual button presses rather than the group's consensus button presses. For the large sample size (lower panel), the middle and higher adjustment values do not strongly influence the shape of the peaks and valleys of the density estimate. However, for small sample size (upper panel), distinctive peaks and valleys are formed in the density estimate using the middle adjustment value (0.1 for coarse, 0.05 for fine). The highest adjustment value (0.2 for coarse and 0.1 for fine) reduces the difference between the peaks and valleys, and even eliminates several peaks (arrows). Therefore, we chose the middle bandwidth adjustment (0.1 for coarse and 0.05 for fine) for our density estimation for all sample sizes. (**B**) Examples of growth (left) and decay (right) function fits. Small dots represent the average agreement estimate for individual bootstrap iteration. Large dots represent the average agreement estimate across all bootstrap iterations with each sample size. Functions with the lowest BIC value were selected as the best-fitting curve

measure, the surprise index (Katori et al., 2018), to segmentation data for the first time to evaluate whether this measure of overlap between two time series (which makes fewer assumptions than the correlation coefficient) provides a more sensitive or more stable index of individual-group agreement.

Individual segmentation performance has also been quantified as a measure of individual accuracy in identifying normative boundaries (e.g., Zalla et al., 2013). We briefly explored this approach and present our findings in Appendix B.

Agreement index This measure captures the similarity of segmentation patterns between an individual and an independent group of participants (Zacks et al., 2006). To calculate the agreement index, we transformed individual button press data into a time series indicating the presence or absence of a button press in 1-s-long bins (*individual time series*), as is standard for these measures. We correlated every individual time series in a subsample with the group time series created by calculating the proportion of the remaining participants who pressed a button within each 1-s bin (obs. *r*). The correlation was then scaled with the minimum and maximum possible correlations (min *r* and max *r*) for that individual using the following formula (Kurby & Zacks, 2011): ai = (obs. r − min r)/ (max r − min r). The agreement index therefore ranged from 0 to 1, with higher values indicating greater agreement. Individual agreement index values were then averaged to obtain the subsample's mean agreement index value.

Surprise index This measure reflects the extent to which the expected overlap between individuals' button presses and an independent group's normative boundaries is exceeded. It is adapted from work that estimates the probability of spike co-occurrence across two neurons using the Poisson distribution (Katori et al., 2018). We used the surprise index to capture the co-occurrence of an individual's button presses with a group's normative boundaries. For every participant, we first extracted the time points in the individual time series that contained button presses (individual boundaries). Next, we identified normative boundaries for the rest of the group using the group density time series. We then calculated *k* as the number of times the individual's boundaries overlapped with a group normative boundary. For this analysis, overlap was defined as occurring when the individual boundary fell within a 1-s window centered on the normative boundary. This provided a resolution that was comparable to that of the agreement index (which used 1-s bins). The *rarity* of overlap was calculated as the summed probability of overlap occurring *k* to *K* number of times, where *K* is total number of possible overlaps for a given video. We define *K* as the total duration of each video in seconds, to match the resolution used for identifying overlap. Surprise index was then calculated by $-\log_2$ transforming rarity (Katori et al., 2018). Individual surprise index values were then averaged to obtain the mean for the subsample.

Evaluating the sensitivity and specificity of segmentation agreement measures

To evaluate each measure's sensitivity to the event boundaries participants are likely to identify in a movie, we calculated agreement values for (1) segmentation data from one movie (*same;* as described above), (2) randomly generated segmentation data (*random*), and (3) segmentation data from different movies (*cross-movie*). Random data were generated for each participant by randomly sampling (without replacement) *l* times from a continuous uniform distribution with a minimum value of 0 and a maximum value equal to the video duration (in seconds), where *l* was the number of times the participant pressed the button. Random agreement was then calculated using a group of random data (in the case of peakiness) or by comparing two random data sets (in the case of peak-to-peak distance, agreement index, and surprise index). Random agreement measures therefore represent agreement of completely random segmentation data. If a measure reliably captures information about how individuals or groups segment a video, then it should show better agreement for same movie comparisons than for random data or cross-movie comparisons.

Statistical analyses

Our bootstrap approach estimates the sampling distribution of sample-level descriptive statistics (e.g., the sampling distribution of the mean agreement index) of segmentation of different movies at different granularities for samples of different sizes. To quantitatively and efficiently characterize how sample-level agreement estimates change with sample size and to better account for the effects of movies on these measures, we fit linear mixed-effects models to the bootstrapped agreement measures. Sample size (2 to 32, incremented by 2), segmentation grain (fine vs. coarse), and measurement condition (same, random data, or cross-movie) were included as fixed effects and movies were included as random effects using the lmer function from the lme4 package in R (Bates et al., 2015). Separate models were fit for the commercial-lab and everyday-online data sets. In addition to evaluating the fixed effects in these models, we characterize the models with planned contrasts that compared sample-level estimates of same agreement values to sample-level estimates of agreement in random data and cross-movie agreement values. We refer to analyses based on the linear mixed-effects models as examining agreement *across subsamples.* Because they are performed on metrics that characterize agreement for a subsample of data (i.e., the peakiness of a group time series, or the mean agreement index for those within a subsample), these comparisons described whether, on average, studies that utilize a particular agreement measure capture meaningful segmentation behavior.

We used 1000 bootstrap iterations to ensure that we adequately characterized the sampling distributions of the agreement measures when different sample sizes were used. Though this results in high statistical power to detect small effects, not all comparisons reached statistical significance in our analyses. This suggests that there was not sufficient power to classify any trivially small difference as significant. However, the large number of iterations that went into the models makes it even more important to characterize the distributions and the size of the effects at different sample sizes (Lakens, 2013). We therefore computed the standardized difference between the means of same, random, and cross-movie subsample statistics:

$$t_{est} = \frac{\mu_{\overline{X}s} - \mu_{\overline{X}c}}{\sqrt{\frac{(n_s-1)\ \sigma^2_{\overline{X}s} + (n_c-1)\sigma^2_{\overline{X}c}}{n_s+n_c\ \ -2}}} \tag{1}$$

Subscripts "s" or "c" refer to values calculated using same data or the comparison data (random or cross-movie) respectively. To preserve the maximum amount of variance, we performed this calculation without aggregating the bootstrapped values. Thus, for every sample size and segmentation grain, n_s and n_c are defined as the number of bootstrapped values calculated on every movie for those data (i.e., $n_s = 2$ for commercial-online and 4 for everyday-online * 1000 bootstrap iterations). This approach provides us with the most conservative t_{est} values. We list the t_{est} values in Appendix A along with descriptive statistics for the same, random, and cross-movie sampling distributions.

These standardized differences provide a basis for estimating the sample size needed to detect a difference in agreement for same versus random data or cross-movie comparisons within a single study (*within study*). Because the standardized difference calculations are performed on the bootstrapped sampling distributions[2], they approximate the t-statistic (t_{est}) for two independent samples, which are each of size n (e.g., a single same movie sample and a single random data sample). Therefore, the smallest sample size at which t_{est} exceeds the critical t-value of a one-sided test with $\alpha = .05$ and degrees of freedom $= n_1 + n_2 - 2$, should be sufficient for detecting differences between individual samples of same movie, random, and cross-movie agreement values in traditional null hypothesis testing. This additional analysis provides a practical basis for estimating the minimum sample size needed for a single study to capture population-level nonrandom and video-specific segmentation data. Alternatively, one could evaluate whether the means of the random and cross-movie data fall outside the 95% confidence interval around the mean of the same data. However, this does not take variance of both sampling distributions into account, and it typically (but not always) produces results that are identical or close to those based on t_{est} (see Appendix A).

Fitting linear mixed-effects models to the subsample statistics also allowed us to investigate how segmentation agreement changes with increases in sample size. To do this, we examined the polynomial trends of the linear mixed-effects model estimates for the effect of sample size using emmeans (Lenth, 2021). Segmentation agreement should, on average, improve with the increase in sample size, since adding more participants to the group should improve estimates of when individuals or groups are likely to segment the movie. Therefore, we expected a significant positive linear contrast for peakiness, agreement index, and surprise index, and a significant negative linear contrast for peak-to-peak distance.

If segmentation agreement estimates stabilize with increasing sample size, there will be a point at which the rate of change in agreement values slows as sample size increases (the relationship will be nonlinear). Therefore, we expected significant negative quadratic effects for peakiness, agreement index, and surprise index (for which larger values are better) and a positive quadratic effect for peak-to-peak distance (for which smaller values are better). Whenever a quadratic effect in the linear mixed-effects model was significant, we estimated the sample size at which agreement stabilized. To do this, we first fit each bootstrapped agreement measure to several growth and decay functions (Fig. 2b)—linear, exponential, logarithmic, asymptotic regression, power curve, and yield loss—using lm, nls (R Core Team, 2018), drm (drc package, Ritz et al., 2015), and aomisc (Onofri, 2020) in R. These functions were selected for their potential ability to fit the shape of the data and were not theoretically motivated. Next, we determined the best fitting curve as the fit with the lowest Akaike information criterion (AIC) and Bayesian information criterion (BIC) values. In all but one case (see Appendix C), the best-fitting curve was nonlinear. Finally, the elbow was defined as the point on the fitted curve that was farthest from a straight line connecting the start and end points of the curve. The elbow was then taken as the sample size at which the measure was stabilizing and where the inclusion of more participants would, on average, provide diminishing returns in estimating agreement.

Results

In this section, we report analyses for each agreement measure separately. For each measure, we first characterize its sensitivity to segmentation data that came from the same movie rather than from random noise (same vs. random comparison) or from a different movie (same vs. cross-movie comparison). We report the z-statistics and 95% confidence intervals of model-based estimates of the effects of measurement condition across subsamples (obtained with emmeans). We also report estimates of the minimum sample size needed to detect differences between

[2] The independent-sample t-statistic is the difference between the mean of one group (e.g., same data) minus the mean of the other group (e.g., random data) divided by the estimated standard error of the mean difference. The numerator in (1) is an estimate of the mean difference between two groups. The denominator in (1) is the pooled standard deviation of the sampling distributions for the two groups, and therefore approximates the standard error of the mean difference (Hays, 1994).

Table 1 Practical characteristics of all agreement measures

Level	Measure	Characteristic	Commercial-lab		Everyday-online	
			Coarse	Fine	Coarse	Fine
Within group	Peakiness	Detect	14	10	14	14
		Stabilize	-	16	-	-
		Asymptote	-	20.21	-	-
Between group	Peak-to-peak distance	Detect	8	6	18	14
		Differentiate	6	6	8	10
		Stabilize	10	10	10	10
		Asymptote	2.61	0.93	4.35	0.70
Individual-group	Agreement index	Detect	6	6	6	8
		Differentiate	6	6	6	6
		Stabilize	14	14	18	18
		Asymptote	0.47	0.62	0.53	0.69
	Surprise index	Detect	8	6	12	8
		Differentiate	8	6	14	8
		Stabilize	16	14	16	14
		Asymptote	5.24	10.53	4.74	7.52

Note: Detect = estimated sample size at which same movie comparisons may be distinguished from random data within a single study. Differentiate = estimated sample size at which same movie comparisons may be distinguished from cross-movie comparisons within a single study (cannot be calculated for peakiness). Stabilize = estimated sample size at which the mean across subsamples began to stabilize. Asymptote = asymptotic value of the mean across subsamples. For the asymptote, peak-to-peak distance is reported in seconds

same data and random and cross-movie data within a single study. Next, we describe how sample-level agreement values change as sample size increases. Should the measure of agreement stabilize, we report the sample size at which stabilization begins. Lastly, we report the presence of any differences between random and cross-movie agreement values. For better visualization, we plotted a random selection of 10% of the bootstrapped estimates and the log-transformed values for peakiness and peak-to-peak distance. However, all analyses were performed on the untransformed agreement values. For all measures, practical evaluations of the sample sizes required to detect nonrandom and video-specific segmentation patterns within one study, as well as the sample size needed to achieve stable segmentation agreement across subsamples, are summarized in Table 1.

Peakiness

Peakiness was sensitive to participants' agreement about the timing of event boundaries (Fig. 3). The model indicated that same peakiness values were overall higher than random peakiness values in the commercial-lab data set, $z = 251.78$, $p < .001$, 95% CI [2.79, 2.83] for coarse, $z = 405.43$, $p < .001$, 95% CI [4.51, 4.55] for fine and in the everyday-online data set, $z = 49.43$, $p < .001$, 95% CI [14.20, 15.38] for coarse, $z = 18.40$, $p < .001$, 95% CI [4.92, 6.09] for fine. Pairwise comparisons revealed that, across subsamples, peakiness began to differentiate real data from noise at a sample size of 6 for coarse and fine commercial-lab segmentation, zs > 2.11, ps < .001, smallest 95% CI [0.0068, 0.18] and 4 for coarse and fine everyday-online segmentation, zs > 2.14, ps < .001, smallest 95% CIs [0.22, 4.91]. This suggests that the average peakiness value of studies utilizing samples of 4–6 participants should reflect real agreement in the data, rather than only random behavior. However, to detect nonrandom peakiness within a single sample, a larger number of participants is required. The standardized differences (t_{est}) indicated that, for commercial-lab segmentation, the required within-study sample sizes were 14, $t_{est}(26) = 2.05$, 95% CI = [1.96, 2.14] for coarse and 10, $t_{est}(18) = 2.18$, 95% CI = [2.09, 2.27] for fine segmentation. For everyday-online segmentation, they were 14, $t_{est}(26) = 1.72$, 95% CI = [1.66, 1.78] for coarse and 14, $t_{est}(26) = 1.89$, 95% CI = [1.83, 1.95] for fine segmentation. Because peakiness provides a summary of segmentation agreement within one group and not a comparison between different groups, cross-movie peakiness values could not be calculated.

Our analysis showed that increasing sample size increased peakiness in most cases (Fig. 3). Significant linear contrasts indicated that across subsamples, same peakiness values increased with sample size in the commercial-lab data, $z = 195.27$, $p < .001$, 95% CI [225.34, 229.91] for coarse, $z = 243.68$ $p < .001$, 95% CI [281.77, 286.34] for fine and in the everyday-online data, $z = 5.92$, $p < .001$, 95% CI [123.55, 245.87] for coarse and $z = 4.09$,

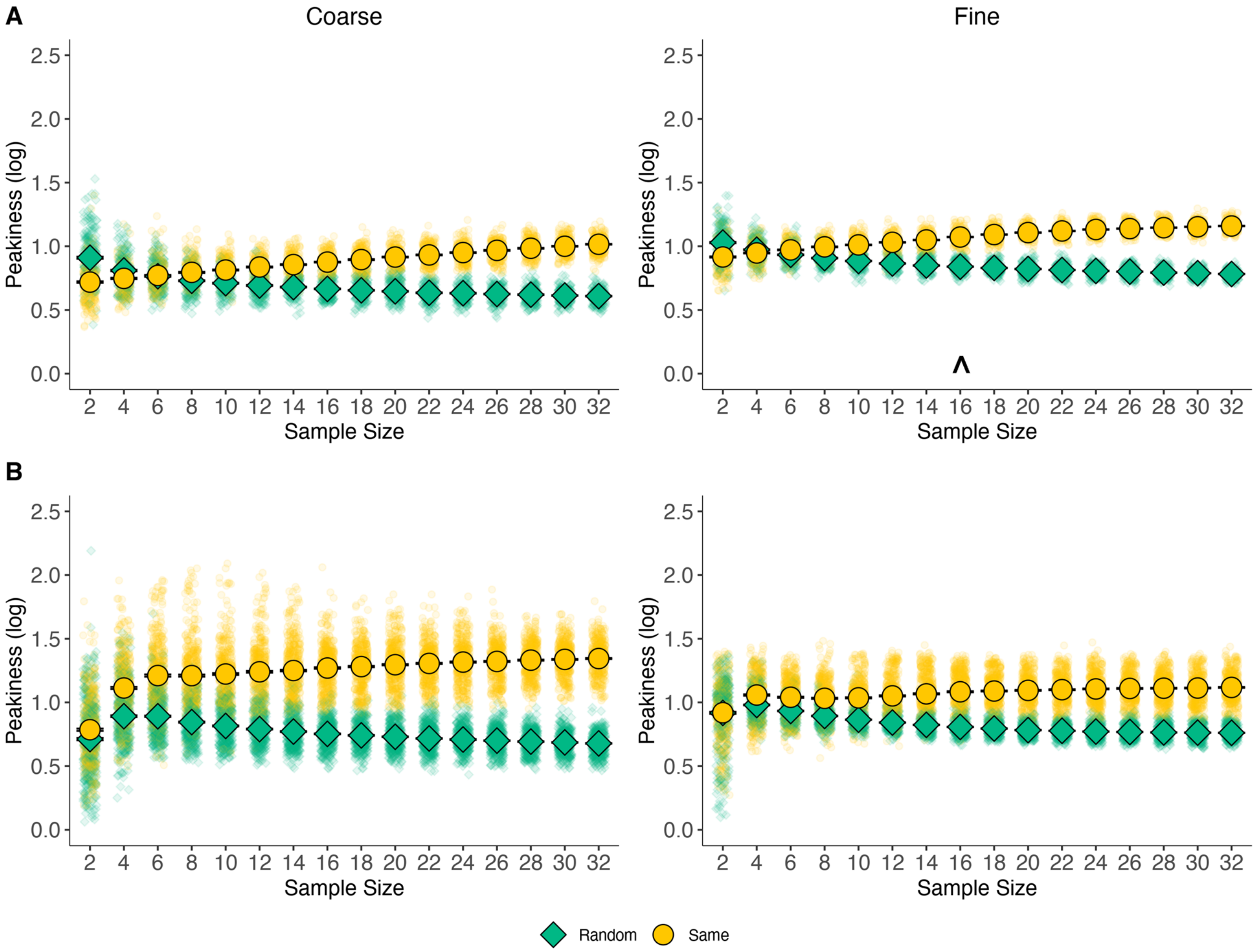


Fig. 3 Log$_{10}$-transformed peakiness values over increasing sample sizes for: (**A**) commercial-lab and (**B**) everyday-online data sets. Small shapes depict the values calculated from a single bootstrap iteration (subsample; only a randomly selected 10% of the bootstrapped values are plotted). Larger shapes depict the average value across all bootstrapping iterations. One low peakiness value and seven high peakiness values for coarse everyday activity segmentation were excluded from the plot due to the *y*-axis limit. Error bars represent 95% confidence interval and carets (^) represent the elbows

p < .001, 95% CI [66.71, 189.03] for fine. There was minimal evidence that peakiness stabilized at larger sample sizes: the negative quadratic effect of sample size on same peakiness values was significant only for fine segmentation in commercial-lab, $z = 27.20$, *p* < .001, 95% CI [−69.66, −60.30], stabilizing at sample size 16. However, we found no evidence for stabilization for coarse segmentation in commercial-lab, as peakiness continued to increase with the increase in sample size, resulting in a significant positive quadratic contrast, $z = 7.90$, *p* < .001, 95% CI [14.19, 23.56]. There was also no evidence for stabilization for segmentation in everyday-online data, as the quadratic effects of sample size on same peakiness were not significant $z = 0.87$, $p = .39$, 95% CI [−69.88, 180.81] for coarse and $z = -0.99$, $p = .32$, 95% CI [−188.65, 62.03] for fine.

Overall, these findings suggest that peakiness can be used to quantify within-group agreement about boundary placement. Across subsamples, this information may be, on average, captured with sample sizes as small as 4–6 participants. However, to capture meaningful segmentation patterns within a study, 10–14 participants are needed. Further, peakiness did not appear to consistently stabilize as sample size increased. When it did, stabilization occurred with 16 participants.

Peak-to-peak distance

Peak-to-peak distance was sensitive to boundaries identified in specific movies, distinguishing group boundaries from the same movie from group boundaries identified using random data and other movies (Fig. 4). Same peak-to-peak distance

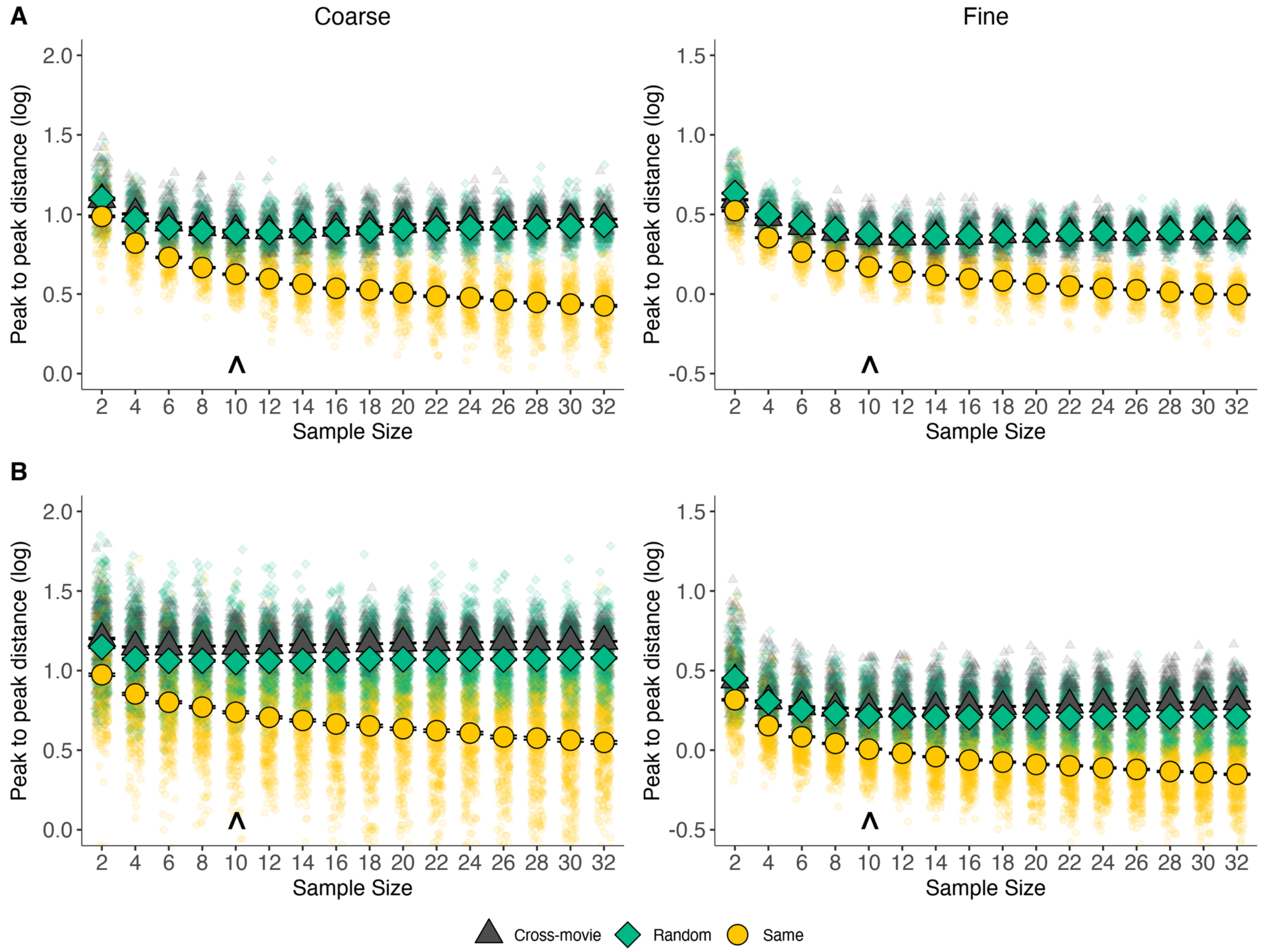


Fig. 4 Log$_{10}$-transformed peak-to-peak distance over increasing sample sizes for segmentation of: (**A**) commercial-lab and (**B**) everyday-online. Small shapes depict values calculated from a single bootstrap iteration (subsample; only a randomly selected 10% of the bootstrapped values are plotted). Larger shapes depict the average value across all bootstrapping iterations. The minimum and maximum values of the *y*-axis for each plot are adjusted between grains to better capture the degree of change in peak-to-peak distance values, but the ranges are kept consistent. Twelve high peak-to-peak distance values for coarse commercial-lab and five high peak-to-peak distance values for coarse everyday-online were excluded from the plot due to the limits set for the *y*-axes. Error bars represent 95% confidence intervals and carets (^) represent the elbows

across subsamples was overall lower than random peak-to-peak distance in the commercial-lab data set, $z = 413.31$ $p < .001$, 95% CI [4.39, 4.43] for coarse, $z = 107.70$, $p < .001$, 95% CI [1.13, 1.17] for fine, and in the everyday-online data set, $z = 375.82$, $p < .001$, 95% CI [6.87, 6.94] for coarse and $z = 41.37$, $p < .001$, 95% CI [0.72, 0.80] for fine. Across subsamples, same peak-to-peak distance was also overall lower than cross-movie comparisons in the commercial-lab data set, $z = 458.33$, $p < .001$, 95% CI [4.87, 4.92] for coarse, $z = 103.42$, $p < .001$, 95% CI [1.08, 1.13] for fine and in the everyday-online data set, $z = 502.10$, p < .001, 95% CI [9.19, 9.26] for coarse, $z = 54.12$, $p < .001$, 95% CI [0.96, 1.03] for fine.

Model-based pairwise comparisons revealed that across subsamples, peak-to-peak distance differentiated same data from random and cross-movie data starting at sample size 2 for all segmentation grains and experiment conditions, *zs* > 12.82, *ps* < .001, smallest 95% CIs [0.46, 0.63], for $n = 2$ to 32 with coarse and fine commercial-lab segmentation, and *zs* > 68.00, *ps* < .001, smallest 95% CIs [0.42, 0.71], for $n = 2$ to 32 with coarse and fine everyday-online segmentation. This suggests that the average of peak-to-peak distance values across studies utilizing samples of 2 participants should reflect video-driven segmentation behavior. However, a larger number of participants is required to detect such differences within a single study. Peak-to-peak distance was lower for same movie comparisons than for comparisons of random data starting at sample sizes of 8 for coarse, $t_{est}(14) = 2.13$, 95% CI = [2.04, 2.22] and 6 for fine, $t_{est}(10) = 2.41$, 95% CI = [2.31, 2.51] commercial-lab segmentation, and 18 for coarse, $t_{est}(34) = 1.74$, 95% CI = [1.68, 1.80], and 14 for fine, $t_{est}(26) = 1.78$, 95% CI = [1.72, 1.84] everyday-online segmentation. Same peak-to-peak distance was also lower than cross-movie peak-to-peak distance starting at sample sizes 6 for coarse, $t_{est}(10) = 1.92$, 95% CI = [1.83, 2.01] and 6 for fine, $t_{est}(10) = 2.02$, 95% CI = [1.93, 2.11] commercial-lab segmentation, and 8 for coarse, $t_{est}(14) = 2.11$, 95% CI = [2.05, 2.18] and 10 for fine, $t_{est}(18) = 1.77$, 95% CI = [1.71, 1.83] everyday-online segmentation.

Across subsamples, the difference between cross-movie and random peak-to-peak distance depended on segmentation grain and data set. In the commercial-lab data set, cross-movie peak-to-peak distance was higher than random peak-to-peak distance for coarse segmentation, $z = 45.02$, $p < .001$, 95% CI [0.46, 0.50]. The reverse was true for fine segmentation, in which cross-movie distances were lower than random distances, $z = 4.28$, $p < .001$, 95% CI [−0.067, −0.025]. In the everyday-online data set, cross-movie peak-to-peak distance was consistently higher than random distances for coarse, $z = 126.27$, $p < .001$, 95% CI [2.28, 2.36] and fine, $z = 12.75$, $p < .001$, 95% CI [0.20, 0.27] segmentation.

Peak-to-peak distance decreased as sample size increased and eventually stabilized (Fig. 4). Significant linear contrasts indicated that peak-to-peak distance decreased with increasing sample size in the commercial-lab data set: $z = 198.98$, $p < .001$, 95% CI [−223.82, −219.45] for coarse, $z = 64.65$, $p < .001$, 95% CI [−74.20, −69.83] for fine, and in the everyday-online data set: $z = 112.32$, $p < .001$ 95% CI [−219.01, −211.49] for coarse, $z = 22.86$, $p < .001$, 95% CI [−47.56, −40.05] for fine. Significant positive quadratic fits in commercial-lab coarse, $z = 117.42$, $p < .001$ 95% CI [263.57, 272.52] and fine $z = 38.37$, $p <.001$, 95% CI [83.10, 92.05] segmentation and in everyday-online coarse, $z = 55.72$ $p < .001$, 95% CI [211.14, 226.53] and fine, $z = 13.36$, $p < .001$, 95% CI [44.79, 60.18] segmentation were consistent with stabilization in this measure. Subsequent analyses suggested that peak-to-peak distance started to stabilize at sample size 10 for coarse and fine segmentation in the commercial-lab data set and in the everyday-online data set. Asymptotic values for peak-to-peak distance differed across grains but were consistent across data sets (Table 1 and Fig. 3).

These findings suggest that peak-to-peak distance can be used to quantify two groups' agreement about boundary placement within a specific movie. Across subsamples, information about normative group boundaries captured by peak-to-peak distance may be, on average, present in samples as small as 2 for all segmentation conditions. However, a minimum of 6–18 participants are needed to capture meaningful segmentation behavior within a single study, depending on the mode of data collection. Further, as sample size increased, peak-to-peak distance stabilized with 10 participants.

Agreement index

As a measure of how well an individual agrees with a separate group of observers, agreement index was sensitive to the specific movie being segmented (Fig. 5). Across subsamples, model-based comparisons indicated that it distinguished segmentation of the same movie from random button presses and segmentation of a different movie in both the commercial-lab data set and the everyday-online data set. The model indicated that same agreement index values were overall higher than those for random data in the commercial-lab data set, $z = 344.57$, $p < .001$, 95% CI [.116, .118] for coarse segmentation, $z = 302.14$, $p < .001$, 95% CI [.102, .103] for fine segmentation, and in the everyday-online data set, $z = 581.79$, $p < .001$, 95% CI [.193, .195] for coarse segmentation, $z = 482.46$, $p < .001$, 95% CI [.160, .162] for fine segmentation. Same agreement index values were also higher than cross-movie agreement index values in the commercial-lab data set, $z = 551.79$, $p < .001$, 95% CI [.187, .188] for coarse segmentation, $z = 481.64$, $p < .001$, 95% CI [.163, .164] for fine segmentation, and in the everyday-online data set, $z = 674.93$, $p < .001$, 95% CI [.224, .226] for coarse segmentation, $z = 564.68$, $p < .001$, 95% CI [.188, .189] for fine segmentation.

Model-based pairwise comparisons also revealed that across subsamples, the agreement index began to differentiate same data from random and cross-movie data starting at sample size 2 for all segmentation grains and experiment conditions, $zs > 14.96$ $ps < .001$, smallest 95% CIs [.018, .023] for coarse and fine commercial-lab segmentation and $zs > 45.72$, $ps < .001$, smallest 95% CIs [.058, .064] for coarse and fine everyday-online segmentation. This suggests that the average of agreement index values across studies that utilize samples with as few as 2 participants should reflect nonrandom and video-specific segmentation behavior. However, a larger number of participants is required to detect differences within a single sample. The standardized differences indicated that the required sample sizes to detect differences between same and random agreement index within a single study were 6 for coarse, $t_{est}(10) = 1.97$, 95% CI = [1.88, 2.06] and fine, $t_{est}(10) = 1.88$, 95% CI = [1.79, 1.96] commercial lab segmentation, and 6 for coarse, $t_{est}(10) = 2.10$, 95% CI = [2.04, 2.16] and 8 for fine, $t_{est}(14) = 2.27$, 95% CI = [2.20, 2.34] everyday online segmentation. Sample sizes of 6 for coarse, t_{est} $(10) = 2.25$, 95% CI = [2.16, 2.34] and fine, $t_{est}(10) = 2.25$, 95% CI = [2.16, 2.34] segmentation of commercial lab and 6 for coarse, $t_{est}(10) = 2.11$, 95% CI = [2.05, 2.17] and fine, $t_{est}(10) = 2.30$, 95% CI = [2.23, 2.37] segmentation of everyday-online were required to detect differences between same and cross-movie agreement index within a single study.

Further, model-based comparisons revealed that, unlike the other segmentation agreement measures, agreement

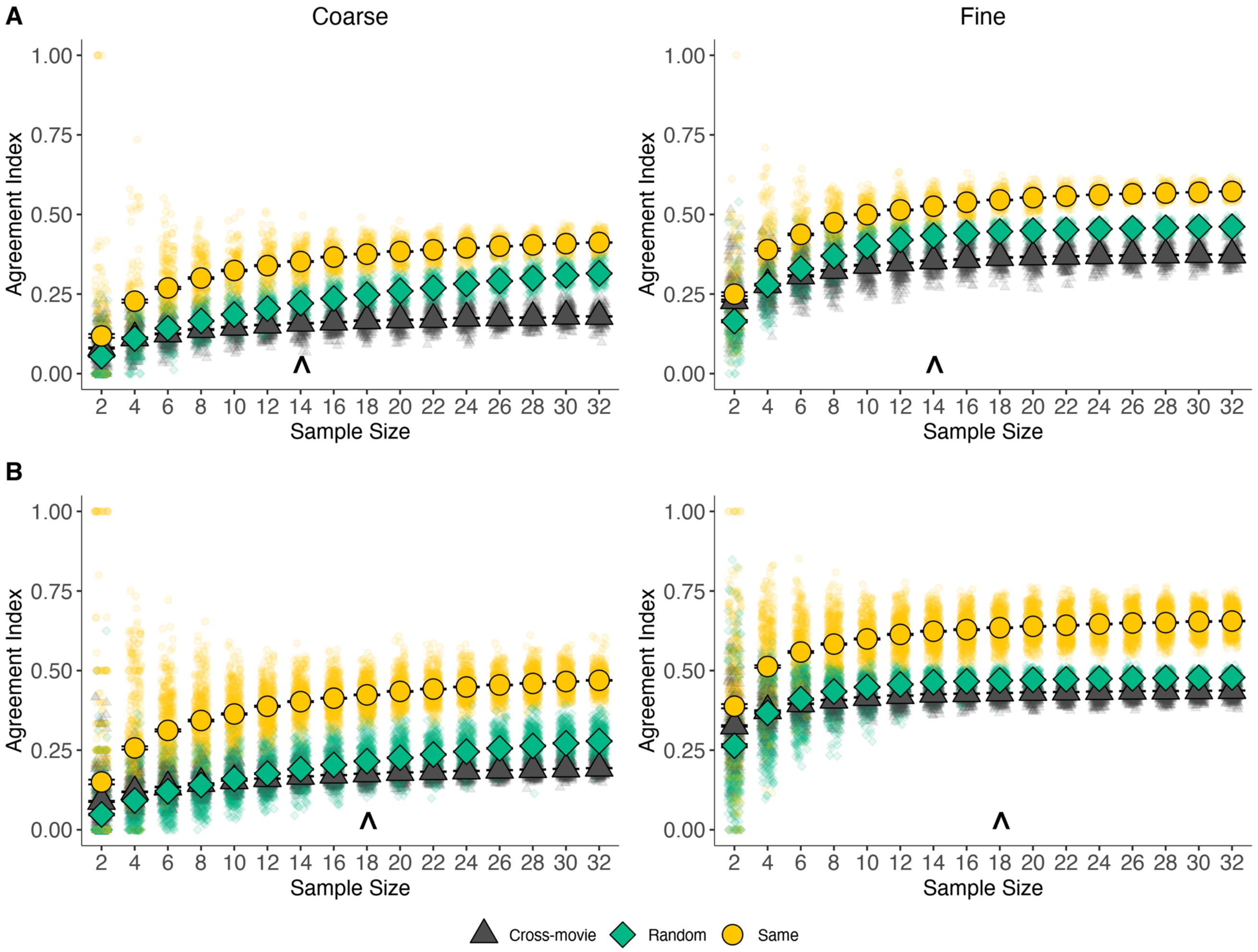


Fig. 5 Agreement index over increasing sample size for segmentation in: (**A**) commercial-lab and (**B**) everyday-online. Small shapes depict values calculated from a single bootstrap iteration (subsample; only a randomly selected 10% of the bootstrapped values are plotted). Larger shapes depict the average value across all bootstrapping iterations. Error bars represent 95% confidence interval and carets (^) represent the elbows

index can differentiate cross-movie comparisons from random data. Across subsamples, cross-movie agreement index was overall lower than the random agreement index for all segmentation grains in the commercial-lab data set, $z = 207.22$, $p < .001$, 95% CI [0.070, 0.071] for coarse segmentation, $z = 179.50$, $p < .001$, 95% CI [0.060, 0.062] for fine segmentation and in the everyday-online data set, $z = 93.14$ $p < .001$, 95% CI [0.030, 0.032] for coarse segmentation, $z = 82.12$, $p < .001$, 95% CI [0.027, 0.028] for fine segmentation. Lower cross-movie agreement values could reflect incompatible event structures in different movies. Consistent with this possibility, the group time series from different movies on the original sample were weakly correlated (Appendix D).

Agreement index improved and eventually stabilized with increasing sample size. Positive, significant linear contrasts indicated that same agreement index increased with increasing sample size in the commercial-lab data set, $z = 281.99$, $p < .001$, 95% CI [9.92, 10.06] for coarse segmentation, $z = 286.97$, $p < .001$, 95% CI [10.10, 10.24] for fine segmentation and the everyday-online data set, $z = 319.11$, $p < .001$, 95% CI [11.03, 11.17] for coarse segmentation, $z = 229.63$, $p < .001$, 95% CI [7.92, 8.06] for fine segmentation. This effect of increasing sample size on the agreement index decreased as sample size increased, as evidenced by the significant negative quadratic fit in the commercial-lab data set, $z = 133.09$, $p < .001$, 95% CI [−9.81, −9.52] for coarse segmentation, $z = 164.91$, $p < .001$, 95% CI [−12.12, −11.83] for fine segmentation, and in the everyday-online data set, $z = 143.56$, $p < .001$, 95% CI [−10.37, −10.09] for coarse segmentation, $z = 138.77$, $p < .001$, 95% CI [−10.03, −9.75] for fine segmentation. Subsequent analyses indicated that the agreement index started to stabilize at sample size 14 for coarse and fine segmentation in the commercial-lab data set and 18 for coarse and fine segmentation in the everyday-online data set. Asymptotic values for the agreement index differed across grains but were comparable across data sets (Table 1 and Fig. 5). These values were also comparable to those reported in other work (e.g., Kurby & Zacks, 2011).

These findings suggest that the agreement index can be used to quantify how well the boundaries identified by an individual agree with those identified by a separate group. Across subsamples as small as two participants, the mean agreement index captured nonrandom, video-driven segmentation behavior on average for all segmentation conditions. However, for the agreement index to capture segmentation agreement that reliably differs from chance within a single study, a minimum of 6–8 participants are needed. Further, as sample size increased, the agreement index stabilized with 18 participants or fewer. Lastly, the agreement index also differentiated random segmentation data from segmentation of different movies. This property was not observed in the other agreement measures tested in this study.

Surprise index

Overall, surprise index was sensitive to whether individual segmentation was compared to normative boundaries from groups that were segmenting the same movie (Fig. 6). The linear mixed-effects model indicated that across subsamples, the surprise index distinguished segmentation of the same movie from random button presses and segmentation of a different movie in both the commercial-lab data set and the everyday-online data set. According to the model, same surprise index was higher than random surprise index in the commercial-lab data sets, z =265.42, p < .001 95% CI [2.30,

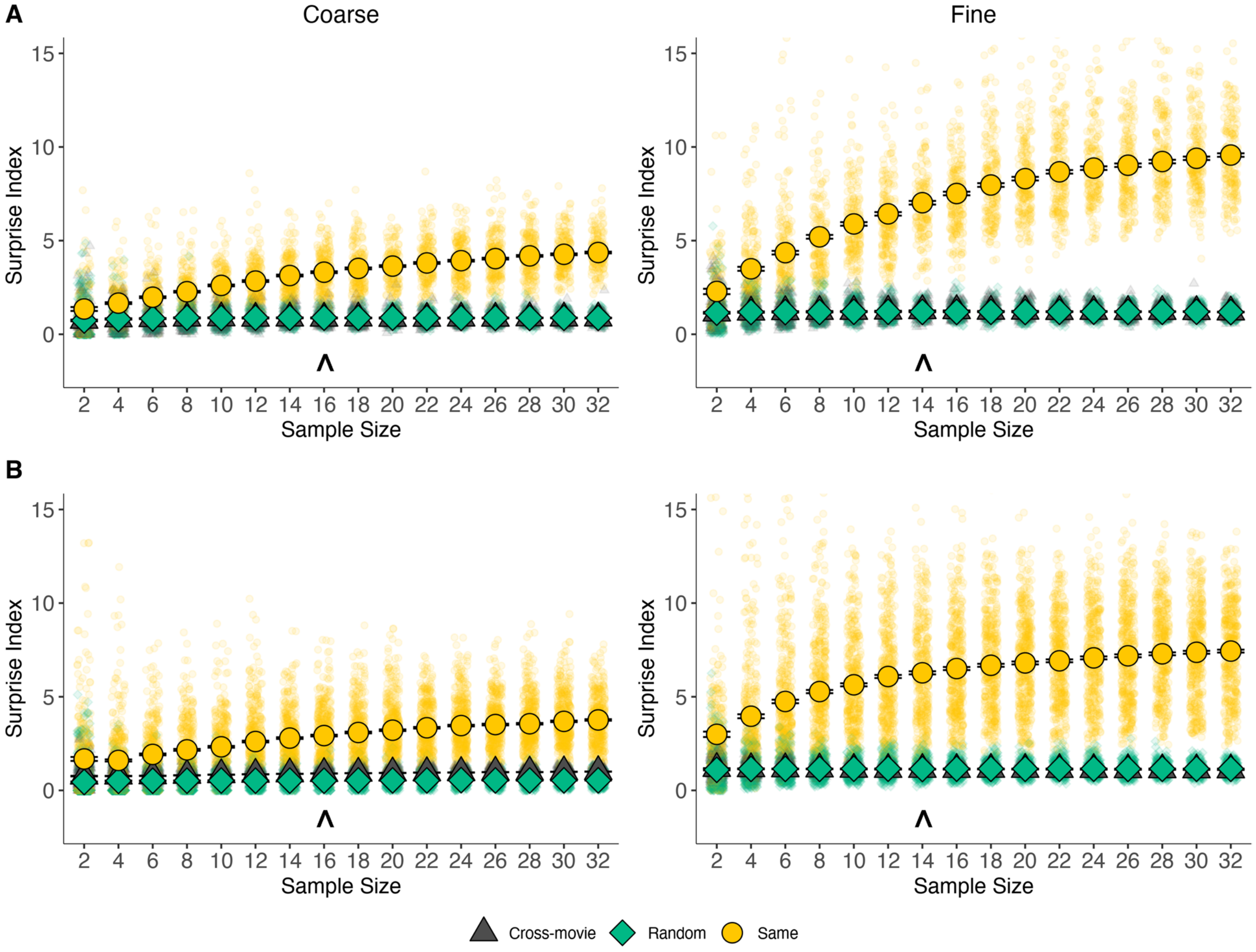


Fig. 6 Surprise index over increasing sample size for segmentation in: (**A**) commercial-lab and (**B**) everyday-online. Small shapes depict values calculated from a single bootstrap iteration (subsample; only a randomly selected 10% of the bootstrapped values are plotted). Larger shapes depict the average value across all bootstrapping iterations. Error bars represent 95% confidence intervals and carets (^) represent the elbow.

2.34] for coarse segmentation, $z = 673.31$, $p < .001$, 95% CI [5.87, 5.90] for fine segmentation, and in the everyday-online data sets, $z = 311.05$, $p < .001$, 95% CI [2.32, 2.35] for coarse segmentation, $z = 665.78$, $p < .001$, 95% CI [4.98, 5.01] for fine segmentation. Surprise index was also overall higher for same than for cross-movie group comparisons in the commercial-lab data set, $z = 266.51$, $p < .001$, 95% CI [2.31, 2.35] for coarse segmentation, $z = 670.52$, $p < .001$, 95% CI [5.84, 5.88] for fine segmentation, and in the everyday-online data set, $z = 261.41$, $p < .001$, 95% CI [1.94, 1.97] for coarse segmentation, $z = 669.09$, $p < .001$, 95% CI [5.00, 5.03] for fine segmentation.

Model-based pairwise comparisons revealed that the average surprise index across subsamples differentiated same movie agreement from agreement in random noise and cross-movie agreement, starting at sample size 2 for all conditions, all zs > 17.25, ps < .001, smallest 95% CI [0.53, 0.67] for coarse and fine segmentation in commercial-lab, and all zs > 27.78, ps < .001, smallest 95% CI [0.77, 0.89] for coarse and fine segmentation in everyday-online. However, for the surprise index to reflect nonrandom and video-specific agreement within a single study, a larger number of participants is required. The standardized differences indicated that to detect a difference between same movie and random data, the surprise index requires a sample size of at least 8 for coarse, $t_{\text{est}}(14) = 1.78$, 95% CI = [1.70, 1.86] and 6 for fine, $t_{\text{est}}(10) = 1.96$, 95% CI = [1.87, 2.05] commercial-lab segmentation, and 12 for coarse, $t_{\text{est}}(22) = 1.91$, 95% CI = [1.85, 1.97] and 8 for fine, $t_{\text{est}}(14) = 2.03$, 95% CI = [1.97, 2.09] everyday-online segmentation. Sample size needed to be at least 8 for coarse, $t_{\text{est}}(14) = 1.80$, 95% CI = [1.72, 1.88] and 6 for fine, $t_{\text{est}}(10) = 1.96$, 95% CI = [1.87, 2.05] segmentation in commercial lab, and 14 for coarse, $t_{\text{est}}(26) = 1.83$, 95% CI = [1.77, 1.89] and 8 for fine, $t_{\text{est}}(14) = 2.06$, 95% CI = [2.00, 2.12] segmentation in everyday-online for surprise index to reflect differences between same and cross-movie agreement.

Across subsamples, surprise index did not consistently differ between cross-movie and random data. In commercial-lab segmentation, the surprise index did not differ significantly across cross-movie and random data, $z = 1.09$, $p = .27$, 95 % CI [−0.0076, 0.027]. However, it did differ for fine segmentation, $z = -2.79$, $p = .0052$, 95% CI = [0.0073, 0.042]. For everyday-online segmentation, the surprise index was greater for cross-movie data than random data for coarse, $z = 49.64$, $p < .001$, 95% CI = [0.36, 0.39], but not fine, $z = 3.31$, $p = .00090$, 95% CI = [0.010, 0.039] segmentation.

Surprise index improved and eventually stabilized with increasing sample size, as indicated by significant positive linear contrasts for the commercial-lab data set, $z = 148.20$ $p < .001$, 95% CI [133.33, 136.90] for coarse, $z = 340.33$, $p < .001$, 95% CI [308.50, 312.07] for fine and the everyday-online data set, $z = 129.95$, $p < .001$, 95% CI [100.08, 103.14] for coarse, $z = 220.28$, $p < .001$, 95% CI [170.72, 173.78] for fine. The positive effect of sample size on surprise index grew smaller with larger sample sizes, as evidenced by the significant negative quadratic fit in the commercial-lab data set, $z = 30.52$, $p < .001$, 95% CI [−60.69, −53.37] for coarse, $z = 97.58$, $p < .001$, 95% CI [−185.99, −178.67] for fine, as well as in the everyday-online data set, $z = 22.95$, $p < .001$, 95% CI [−39.92, −33.64] for coarse, $z = 84.83$, $p < .001$, 95% CI [−139.09, −132.80] for fine. Additional analyses indicated that surprise index began to stabilize at sample size 16 for coarse and 14 for fine segmentation in commercial-lab and 16 for coarse and 14 for fine segmentation in everyday-online. Asymptotic surprise index values differed markedly across conditions (Table 1; Fig. 6).

These findings suggest that surprise index can be used to quantify agreement between an individual's boundary placement and normative group boundaries within a specific movie. The results suggest that, across studies, the surprise index will capture nonrandom and video-specific segmentation agreement with sample sizes as small as 2 participants, on average. However, a minimum of 6–14 participants are needed, depending on the mode of data collection, to capture meaningful segmentation agreement within a single study. Further, the surprise index stabilized with 16 or fewer participants.

Discussion

This study systematically explored various segmentation agreement measures, each quantifying performance at the group (*peakiness* and *peak-to-peak distance*) or individual (*agreement index* and *surprise index)* level. The results confirmed previous findings on the segmentation task's utility in measuring meaningful behavior. They demonstrated that multiple agreement measures can capture group and individual segmentation patterns with relatively few participants. This observation is true for data collected using structured commercial films in the lab and unedited videos of everyday activities online. The results also validated the use of group segmentation data to infer when other people are likely to segment, especially when large sample sizes are used. These results provide insight into potential limits to the amount of agreement that can be observed at the group level and inform the selection of sample sizes and measures for future research.

One advantage of using the segmentation task to study event perception is its use of naturalistic stimuli, such as movies, to approximate the complex and continuous nature of everyday experience. However, unlike presenting simple stimuli such as images or tones in a discrete and temporally controlled manner, the mapping between stimuli and responses in a segmentation task is undefined. Although boundary identification is associated with changes in the features of a movie (e.g., action or location changes; Hard et al., 2006; Magliano et al., 2001; Newtson et al., 1977; Swallow et al., 2018), events are fundamentally a manifestation of the mind. Therefore, establishing a ground truth to compare segmentation performance with is not only challenging but potentially misleading. This study showed that, despite no "correct" behavior in a segmentation task, segmentation performance is far from random and is not generalizable across different movies. In all cases, agreement calculated based on segmentation on the same movie (*same*) was better than agreement calculated based on randomly generated data (*random*) or segmentation on different movies (*cross-movie*). When evaluated across subsamples, nonrandom and video-specific segmentation behavior appeared to be present with sample sizes as small as 2. Although larger samples are needed to detect signal-driven segmentation behavior within a single sample, this finding highlights the efficacy of using the segmentation task to capture genuine and meaningful segmentation of naturalistic events. Because most studies use samples larger than this, it also increases confidence in the literature using naturalistic stimuli to study event perception, its cognitive implications, and its neural correlates (Sonkusare et al., 2019).

Further, we found that the reliability of segmentation task measures extends beyond controlled laboratory settings. Despite differences in the type of videos used, how the data were collected, and the sampled population, on average, group- and individual-level agreement measures were comparable across the commercial-lab and everyday-online data sets. This finding demonstrates that the segmentation task is a reliable method for assessing segmentation behavior with different types of materials, in diverse populations, and on a platform where there is minimal experimental control, and especially when moderate to large samples are used. This information is valuable considering the increased interest in online data collection methods for studying human cognition (Stewart et al., 2017) and the heightened reliance on web-based alternatives due to global pressures (e.g., the COVID-19 pandemic).

Although, on average, sample-level measures of segmentation agreement were sensitive to structured segmentation data, large amounts of variability across samples poses clear limits on the use of small sample sizes in future studies. For small sample sizes, normative boundaries could shift by several seconds (on average) from one sample to another (Table 1, Appendix A, and Fig. 4). Further, the minimum sample size needed to detect differences between same and random or cross-movie data in a single comparison was always higher than the sample size needed to detect differences across sample-level statistics. The minimum sample size was overall higher for online segmentation of everyday activities (up to 18 participants). These findings call for caution when using data from a small group of participants as a basis for inferring both whether normative boundaries are movie-specific and when individuals are likely to segment an event.

In addition to using large enough samples to capture meaningful segmentation patterns, future studies should consider the sample size needed to obtain stable segmentation agreement. In this study, we found that agreement started to stabilize with moderately large sample sizes (18 participants or fewer). However, although adding participants beyond this stabilization point resulted in diminishing gains in agreement, studies investigating potentially small differences in segmentation agreement may need to aim for larger sample sizes. This is particularly true in those cases where between-sample variability necessitated larger sample sizes to ensure the ability to detect video-driven segmentation data within a single study (e.g., peak-to-peak distance and surprise index for everyday-online segmentation). The sample sizes at which segmentation agreement started to stabilize differed between individual- and group-level agreement measures, with smaller sample sizes needed for measuring between-group agreement (Table 1). Thus, studies interested in estimating the similarity between individual and group performance would require larger sample sizes than studies interested in identifying normative boundaries (see Appendix A for values useful for power analyses). Of note, peakiness was the one measure that did not consistently stabilize with increasing sample size, highlighting the impact that individual variability in segmentation task performance can have on overall group agreement. Performance error and individual differences in segmentation task performance also limit estimates of group and individual levels of agreement, though the precise limits are likely to depend on a variety of factors (e.g., materials, population, and instructions). Future studies should be cautious of this limit.

This study introduced new methods to quantify segmentation task agreement. First, we describe a systematic and straightforward method for defining normative group boundaries that preserves the continuous nature of segmentation behavior and accounts for subtle variations in when participants report an event boundary. Though more recent studies also treat segmentation data probabilistically (Huff, Maurer, et al., 2017a; Huff, Papenmeier, et al., 2017b; Newberry et al., 2021; Smith et al., 2020), this study includes methods for avoiding arbitrary decisions about kernel bandwidth, basing it on the data. We applied our probabilistic quantification of group segmentation to calculate between-group agreement (*peak-to-peak distance*) and introduced new ways to quantify individual and within-group agreement: the *surprise index* and *peakiness*. Notably, the surprise index provides a probabilistic method to quantify whether the observed overlap between individual button presses and normative boundaries is greater than expected based on their frequency. As a result, it avoids limitations that arise from correlating individual and group time series, as in the agreement index (e.g., assumptions regarding normality and restriction of range).

In addition to the type of agreement one needs to measure (within-group, between-group, or individual to group), several factors should be considered when deciding which agreement metric to use in future research. These include whether the entire group time series should be quantified (peakiness, agreement index) or just the normative boundaries (peak-to-peak distance, surprise index), whether button presses should be treated discretely (agreement index) or probabilistically (peakiness, peak-to-peak distance, and surprise index), and whether assumptions underlying the metric are met. We also note that the agreement index was the only measure for which cross-movie agreement was significantly worse than random agreement in all segmentation grains and conditions. This finding could imply that the agreement index is sensitive to uncorrelated signals from different movies (correlations of the time series for each movie were all below .2, see Appendix D).

This study has several limitations. Although the overall patterns of results between commercial-lab and everyday-online data sets were comparable, there were notable differences. More variability was observed across the online data subsamples. These differences are difficult to interpret. Greater variability in the everyday-online data could reflect common concerns with online task performance, such as differences in levels of attentional engagement, equipment used to perform the task, or in the sample population. It could also reflect the use of lab-produced videos of everyday activities in the everyday-online data set. Unlike commercially produced movies, everyday activity videos do not include editing features (e.g., cuts and scene changes) that promote synchrony in gaze direction (Dorr et al., 2010; Hutson et al., 2017; Loschky et al., 2015) and that may systematically facilitate boundary identification (Magliano et al., 2020; Magliano & Zacks, 2011). However, because everyday activity videos were segmented online, we could not disentangle online performance effects from video type. Future work is necessary to address this question.

Additionally, although our data sets were relatively large for this type of research ($N=64$, and $N=72$), our analyses and the conclusions that can be drawn from them are limited by sample size, the population that they were drawn from, and the degree to which participants in our data were representative of those populations (an assumption of bootstrapping procedures for estimating sampling distributions). Finally, we have restricted this investigation to a narrow set of questions aimed at addressing whether segmentation task performance is stable across individuals and groups of different sizes. We have largely ignored questions about other aspects of event segmentation, including measures of hierarchical alignment of event boundaries for different grains, the relationship between boundary identification and video features or other aspects of behavior (e.g., eye movements), the influence of individuals with atypical segmentation patterns on these measures, and the various methods for creating group time series. These and other questions will need to be addressed in future research.

Conclusion

This study demonstrates that group-to-group or individual-to-group segmentation agreement measures reflect behavior that is meaningful and sensitive to the structure of the movie, and that this is true for segmentation data collected online as well as in the lab. When estimating segmentation agreement, large samples may not show large advantages over moderately sized samples, and even small samples may be sufficient to detect nonrandom task performance with some metrics. However, the inherent variability in segmentation performance across participants may limit the certainty with which normative boundaries can be identified. Thus, we propose that studies interested in investigating segmentation agreement should consider (1) the type of agreement under consideration (group or individual level), (2) the mode and medium in which data are collected (commercial vs. everyday movie and in-lab vs. online), and (3) the minimum sample size needed to reliably capture video-driven segmentation patterns within a single sample (Table 1). Given the extensive influence of event segmentation on other cognitive processes such as attention and memory, this research can inform and supplement other studies on event segmentation and its cognitive implications.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-022-01832-5.

Acknowledgements The authors would like to thank Melissa Yap, Ariella Ratner, Sofie Prieto, and Joel Cruz for assistance in data collection for the in-lab data set, Taylor Beauvais, Clara Lee, Isabella Poulous, and Hyung Sop Shin for assistance in conducting the online study, and Stephen Parry from the Cornell Statistical Consulting Unit for advice on data analysis. The authors declare no conflicts of interest. This research was financially supported by the Cornell Center for Social Sciences (CCSS 297) and Cornell College of Arts and Sciences.

Author contributions K. Sasmita – study concept and design, data collection, data analysis design, code development, writing first draft of manuscript and manuscript editing.

K. Swallow – study concept and design, data collection, data analysis design, manuscript revisions and editing.

Funding This research was financially supported by the Cornell Center for Social Sciences (CCSS 297) and Cornell College of Arts and Sciences.

Declarations

Competing interests The authors have no conflicts of interest to declare.

Ethics approval All procedures involved in collecting and analyzing data from human subjects were conducted in accordance with the ethical standards of the 1964 Declaration of Helsinki and its subsequent amendments. Procedures were reviewed and approved by the Cornell Institutional Review Board.

Consent to participate Informed consent was obtained from all individual participants who contributed data to this study.

References

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron, 95*(3), 709–721.e5. https://doi.org/10.1016/j.neuron.2017.06.041

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Ben-Yakov, A., & Henson, R. N. (2018). The Hippocampal Film Editor: Sensitivity and Specificity to Event Boundaries in Continuous Experience. *Journal of Neuroscience, 38*(47), 10057–10068. https://doi.org/10.1523/JNEUROSCI.0524-18.2018

Birnbaum, M. H. (2004). Human Research and Data Collection via the Internet. *Annual Review of Psychology, 55*(1), 803–832. https://doi.org/10.1146/annurev.psych.55.090902.141601

Bläsing, B. E. (2015). Segmentation of dance movement: Effects of expertise, visual familiarity, motor experience and music. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.01500

Boggia, J., & Ristic, J. (2015). Social event segmentation. *The Quarterly Journal of Experimental Psychology, 68*(4), 731–744. https://doi.org/10.1080/17470218.2014.964738

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience, 20*(1), 115–125. https://doi.org/10.1038/nn.4450

Cutting, J. E., DeLong, J. E., & Brunick, K. L. (2011). Visual activity in Hollywood film: 1935 to 2005 and beyond. *Psychology of Aesthetics, Creativity, and the Arts, 5*(2), 115–125. https://doi.org/10.1037/a0020995

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision, 10*(10), 28. https://doi.org/10.1167/10.10.28

Faber, M., Radvansky, G. A., & D'Mello, S. K. (2018). Driven to distraction: A lack of change gives rise to mind wandering. *Cognition, 173*, 133–137. https://doi.org/10.1016/j.cognition.2018.01.007

Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory & Cognition, 34*(6), 1221–1235. https://doi.org/10.3758/BF03193267

Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General, 140*(4), 586–604. https://doi.org/10.1037/a0024310

Hays, W. L. (1994). *Statistics* (5th ed.). Hardcourt Brace College Publishers.

Huff, M., Maurer, A., Brich, I., Pagenkopf, A., Wickelmaier, F., & Papenmeier, F. (2017a). Construction and Updating of Event Models in Auditory Event Processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*. https://doi.org/10.1037/xlm0000482

Huff, M., Papenmeier, F., Maurer, A. E., Meitz, T. G. K., Garsoffky, B., & Schwan, S. (2017b). Fandom Biases Retrospective Judgments Not Perception. *Scientific Reports, 7*(1), 43083. https://doi.org/10.1038/srep43083

Hutson, J. P., Smith, T. J., Magliano, J. P., & Loschky, L. C. (2017). What is the role of the film viewer? The effects of narrative comprehension and viewing task on gaze control in film. *Cognitive Research: Principles and Implications, 2*(1), 46. https://doi.org/10.1186/s41235-017-0080-5

Jeunehomme, O., & D'Argembeau, A. (2018). Event segmentation and the temporal compression of experience in episodic memory. *Psychological Research*. https://doi.org/10.1007/s00426-018-1047-y

Katori, K., Manabe, H., Nakashima, A., Dunfu, E., Sasaki, T., Ikegaya, Y., & Takeuchi, H. (2018). Sharp wave-associated activity patterns of cortical neurons in the mouse piriform cortex. *European Journal of Neuroscience, 48*(10), 3246–3254. https://doi.org/10.1111/ejn.14099

Kim, K-D. (Director). (2004). *3 Iron* [Film]. Kim Ki-Duk Film & Cineclick Asia.

Kosie, J. E., & Baldwin, D. (2019). Attention rapidly reorganizes to naturally occurring structure in a novel activity sequence. *Cognition, 182*, 31–44. https://doi.org/10.1016/j.cognition.2018.09.004

Kurby, C. A., & Zacks, J. M. (2011). Age differences in the perception of hierarchical structure in events. *Memory & Cognition, 39*(1), 75–91. https://doi.org/10.3758/s13421-010-0027-2

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-test and ANOVAs. *Frontiers in Psychology, 4*, 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Lenth, R. V. (2021). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.6.2-1. https://CRAN.R-project.org/package=emmeans

Levine, D., Hirsh-Pasek, K., Pace, A., & Michnick Golinkoff, R. (2017). A goal bias in action: The boundaries adults perceive in events align with sites of actor intent. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(6), 916–927. https://doi.org/10.1037/xlm0000364

Loschky, L. C., Larson, A. M., Magliano, J. P., & Smith, T. J. (2015). What Would Jaws Do? The Tyranny of Film and the Relationship between Gaze and Higher-Level Narrative Film Comprehension. *PLOS ONE, 10*(11), e0142474. https://doi.org/10.1371/journal.pone.0142474

Magliano, J. P., & Zacks, J. M. (2011). The Impact of Continuity Editing in Narrative Film on Event Segmentation. *Cognitive Science, 35*(8), 1489–1517. https://doi.org/10.1111/j.1551-6709.2011.01202.x

Magliano, J. P., Miller, J., & Zwaan, R. A. (2001). Indexing space and time in film understanding. *Applied Cognitive Psychology, 15*(5), 533–545. https://doi.org/10.1002/acp.724

Magliano, J. P., Kurby, C. A., Ackerman, T., Garlitch, S. M., & Stewart, J. M. (2020). Lights, camera, action: The role of editing and framing on the processing of filmed events. *Journal of Cognitive Psychology*, *32*(5–6), 506–525. https://doi.org/10.1080/20445911.2020.1796685

Massad, C. M., Hubbard, M., & Newtson, D. (1979). Selective perception of events. *Journal of Experimental Social Psychology, 15*(6), 513–532. https://doi.org/10.1016/0022-1031(79)90049-0

Mendelsohn, E. (Director). (2010). *3 Backyards* [Film]. Caruso/ Mendelsohn Productions.

Newberry, K. M., & Bailey, H. R. (2019). Does semantic knowledge influence event segmentation and recall of text? *Memory & Cognition*. https://doi.org/10.3758/s13421-019-00926-4

Newberry, K. M., Feller, D. P., & Bailey, H. R. (2021). Influences of domain knowledge on segmentation and memory. *Memory & Cognition.* https://doi.org/10.3758/s13421-020-01118-1

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology, 28*(1), 28–38. https://doi.org/10.1037/h0035584

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology, 12*(5), 436–450. https://doi.org/10.1016/0022-1031(76)90076-7

Newtson, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology, 35*(12), 847–862.

Onofri, A. (2020). *The broken bridge between biologists and statisticians: A blog and R package*. Statforbiology, IT, web: https://www.statforbiology.com

Ovashvili, G. (Director). (2015). *Corn Island* [Film]. Alamdary Film, George Ovashvili Production, 42film GmbH, Arizona Films Productions, Axman Production, FocusFox Studio, Kino Company (KZ).

Papenmeier, F., Maurer, A. E., & Huff, M. (2019). Linguistic Information in Auditory Dynamic Events Contributes to the Detection of Fine, Not Coarse Event Boundaries. *Advances in Cognitive Psychology, 15*(1), 30–40. https://doi.org/10.5709/acp-0254-9

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Richmond, L. L., Gold, D. A., & Zacks, J. M. (2017). Event Perception: Translations and Applications. *Journal of Applied Research in Memory and Cognition, 6*(2), 111–120. https://doi.org/10.1016/j.jarmac.2016.11.002

Ritz, C., Baty, F., Streibig, J. C., & Gerhard, D. (2015). Dose-response analysis using R. *PLOS ONE, 10*(12), e0146021. https://doi.org/10.1371/journal.pone.0146021

Sargent, J. Q., Zacks, J. M., Hambrick, D. Z., Zacks, R. T., Kurby, C. A., Bailey, H. R., Eisenberg, M. L., & Beck, T. M. (2013). Event segmentation ability uniquely predicts event memory. *Cognition, 129*(2), 241–255. https://doi.org/10.1016/j.cognition.2013.07.002

Sheather, S. J., & Jones, M. C. (1991). A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society. Series B (Methodological), 53*(3), 683–690.

Singh, K., & Xie, M. (2010). Bootstrap: A statistical method. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., pp. 46–51). Elsevier Science.

Smith, M. E., Newberry, K. M., & Bailey, H. R. (2020). Differential effects of knowledge and aging on the encoding and retrieval of everyday activities. *Cognition, 196*, 104159. https://doi.org/10.1016/j.cognition.2019.104159

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic stimuli in neuroscience: critically acclaimed. *Trends in Cognitive Science, 23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Speer, N. K., Swallow, K. M., & Zacks, J. M. (2003). Activation of human motion processing areas during event perception. *Cognitive, Affective, & Behavioral Neuroscience, 3*(4), 335–345. https://doi.org/10.3758/CABN.3.4.335

Speer, N. K., Zacks, J. M., & Reynolds, J. R. (2007). Human Brain Activity Time-Locked to Narrative Event Boundaries. *Psychological Science, 18*(5), 449–455. https://doi.org/10.1111/j.1467-9280.2007.01920.x

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences, 21*(10), 736–748. https://doi.org/10.1016/j.tics.2017.06.007

Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave, a Free Modular Tool for Sound Analysis and Synthesis. *Bioacoustics, 18*(2), 213–226. https://doi.org/10.1080/09524622.2008.9753600

Swallow, K. M., & Wang, Q. (2020). Culture influences how people divide continuous sensory experience into events. *Cognition, 205*, 104450. https://doi.org/10.1016/j.cognition.2020.104450

Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General, 138*(2), 236–257. https://doi.org/10.1037/a0015631

Swallow, K. M., Kemp, J. T., & Candan Simsek, A. (2018). The role of perspective in event segmentation. *Cognition, 177*, 249–262. https://doi.org/10.1016/j.cognition.2018.04.019

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., & Raichle, M. E. (2001a). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience, 4*(6), 651–655. https://doi.org/10.1038/88486

Zacks, J. M., Tversky, B., & Iyer, G. (2001b). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General, 130*(1), 29–58. https://doi.org/10.1037/0096-3445.130.1.29

Zacks, J. M., Speer, N. K., Vettel, J. M., & Jacoby, L. L. (2006). Event understanding and memory in healthy aging and dementia of the Alzheimer type. *Psychology and Aging, 21*(3), 466–482. https://doi.org/10.1037/0882-7974.21.3.466

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin, 133*(2), 273–293. https://doi.org/10.1037/0033-2909.133.2.273

Zalla, T., Labruyère, N., & Georgieff, N. (2013). Perceiving Goals and Actions in Individuals with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders, 43*(10), 2353–2365. https://doi.org/10.1007/s10803-013-1784-0

Open practices statement Data and analytical code used for this study are publicly available in the GitHub repository at the following link: https://github.com/ksasmita/esMethods. None of the experiments was preregistered.