# What Makes Content Engaging? How Emotional Dynamics Shape Success

JONAH BERGER
YOON DUK KIM
ROBERT MEYER

Some cultural products (e.g., books and movies) catch on and become popular, while others fail. Why? While some have argued that success is unpredictable, we suggest that period-to-period shifts in sentiment—what we term sentiment volatility—enhance engagement. Automated sentiment analysis of over 4,000 movies demonstrates that more volatile movies are evaluated more positively. Consistent with the notion that sentiment volatility makes experiences more stimulating, the effect is stronger in genres where evaluations are more likely to be driven stimulation (i.e., thrillers rather than romance). Further, analysis of over 30,000 online articles demonstrate that people are more likely to continue reading more volatile articles. By manipulating sentiment volatility in follow-up experiments, we underscore its causal impact on evaluations, and provide evidence for the role of stimulation in these effects. Taken together, the results shed light on what drives engagement, the time dynamics of sentiment, and cultural analytics or why some cultural items are more successful.

*Keywords*: narratives, natural language processing, experiences, movies, automated textual analysis

Why do some cultural products succeed while others fail? Cultural propagation, artistic change, and the diffusion of innovations have been examined across disciplines with the goal of understanding why things catch on (Bass 1969; Boyd and Richerson 1985; Cavalli-Sforza and Feldman 1981; Kashima 2014; Rogers 1995; Salganik et al. 2006; Simonton 1980). Some movies become blockbusters, while others languish. Some news gets lots of attention while other stories flop. What leads certain items to win out in the marketplace of ideas?

One possibility is that success is random. Even domain experts are notoriously bad at recognizing hits and failures in advance (Bielby and Bielby 1994; Hirsch 1972) and Hollywood often spends millions of dollars promoting movies that end up being duds. This has led some to argue that success is driven more by patterns of social influence than anything about the cultural items themselves (Adler 1985; Salganik et al. 2006).

Another possibility is that individual-level psychological processes play an important role. Research on cross-cultural psychology demonstrates how culture can impact psychological processes (Markus and Kitayama 1991), but the reverse is also true: psychological processes influence what people remember, like, and share, which in turn shapes collective culture (Berger et al. 2012; Berger and Packard 2018; Kashima 2008; Schaller and Crandall 2004). Arousal can increase social transmission (Berger 2011), for example, leading stories that evoke more high arousal emotions to be more likely to go

Jonah Berger (jberger@wharton.upenn.edu) is an Associate Professor of Marketing, Yoon Duk Kim (yoonduk@wharton.upenn.edu) is a doctoral student in Marketing, and Robert Meyer (meyerr@wharton.upenn.edu) is the Frederick H. Ecker/MetLife Insurance Professor of Marketing, all at the Wharton School at the University of Pennsylvania, 3730 Walnut St., Philadelphia, PA 19104. Please address correspondence to Jonah Berger. The authors thank Jaehwan Song, Bowen Liu, Russell Richie, and Reihane Boghrati for assistance with various aspects of the project and participants at the Ohio State University Psychology seminar, Behavioral Insights from Text Conference, and Marketing Science conference for helpful feedback. Supplementary materials are included in the web appendix accompanying the online version of this article.

*Editor: J. Jeffrey Inman*

*Associate Editor: Andrew T. Stephen*

*Advance Access publication February 23, 2021*

• Vol. 48 • 2021
DOI: 10.1093/jcr/ucab010

viral (Berger and Milkman 2012). Thus, psychological processes may act as a selection mechanism, determining which things prosper and which fall flat (Berger and Heath 2005; Norenzayan et al. 2006).

Along these lines, we suggest that period-to-period shifts in sentiment—what we term "sentiment volatility"—can enhance engagement, and thus shape narrative success. We test this possibility in both the lab and the field. First, automated sentiment analysis on over 4,000 movies examines whether more volatile movies are received more positively. Second, analysis of over 30,000 online articles examines whether people are more likely to continue reading more volatile articles. Third, experiments directly manipulate sentiment volatility, testing its causal impact, and ruling out alternative explanations.

This work makes three main contributions. First, while research has examined how emotions influence evaluations, and specific moments that are weighted more heavily (Fredrickson and Kahneman 1993) there has been less attention to emotional dynamics (i.e., period-to-period changes). We demonstrate how one dynamic feature, sentiment volatility, shapes responses, and outline others that might deserve further attention.

Second, while researchers have long speculated about narrative structure, there have been few empirical tests. Kurt Vonnegut argued that "stories have shapes which can be drawn on graph paper," and suggested that that there were eight main trajectories. Others have argued that for seven (Booker 2004), twenty (Tobias 1993), or thirty-six (Polti 1921) basic plots or theorized about emotion and narratives in film (Tan 1996). But while these suggestions are intriguing, little work has actually empirically tested them (Reagan et al. 2016). Further, while classifying plot types is worthwhile, it does not address whether and why certain features of narratives might make them more engaging. We begin to address this question, quantifying one feature of stories, suggesting why it might be valuable, and demonstrating its impact.

Third, we illustrate how automated text analysis can be used for cultural analytics and to study cultural success. While there has been great interest in why things catch, measurement has proved difficult. Recent work has highlighted the value automated textual analysis for consumer research (Berger et al. 2021b; Humphreys and Wang 2018; Moore and McFerran 2017; Netzer, Lemaire, and Herzenstein 2018; Rocklage and Fazio 2015) and begun to apply this approach to narratives (Eliashberg, Hui, and Zhang 2014). We demonstrate how automated text analysis can measure key features at scale, opening new avenues for future research.

## EVALUATION OF HEDONIC EXPERIENCES

Decades of research have examined how aspects of experiences shape evaluations. Most simply, experiences can be described by valence. Some experiences, like eating tasty food, are positive, others, like getting fired, are negative. Not surprisingly, people generally prefer positive experiences to negative ones, and a great deal of evidence supporting the hedonistic principle (i.e., approach pleasure and avoid pain) is consistent with this perspective.

A key question, though, is how multiple periods of hedonic experience are integrated into summary evaluations. Movies, for example, have multiple moments, some more positive and some less so. Articles have positive paragraphs and less positive ones. How might these multiple moments of differing valence, or sentiment, be combined into an overall response?

Early work used the mean response over the course of a stimulus (Aaker, Stayman, and Hagerty 1986), but subsequent work (Fredrickson and Kahneman 1993) suggests that some moments of experiences (e.g., peaks and ends) are weighted more heavily. Making a painful colonoscopy longer should make it more unpleasant, but consistent with the notion that endpoints matter, adding an additional less negative period to the end actually made it less unpleasant (Redelmeier, Katz, and Kahneman 2003). Similarly, ads that end more positively are liked more (Baumgartner, Sujan, and Padgett 1997).

Other research focused on rates of change (Baumgartner et al. 1997; Hsee and Abelson 1991). Carver and Scheier (1990), for example, theorize that in addition to the absolute level, hedonics are also driven by the slope of affective trajectories. Similarly, Plantinga (2009) argues that rising action in film may evoke the strongest response.

Most prior work, however, has focused on the integration of affective stimuli of a single valence (Pracejus and Olsen 2004). How making a positive experience longer (Fredrickson and Kahneman 1993) or adding a less negative end to an already negative experience (Kahneman et al. 1993) shapes overall evaluations. There has been less attention, however, to how positive and negative components, or different sentiments are combined into an overall affective response.

More generally, while research has examined peaks and ends, as well as rates of change, these are not the only features of affective dynamics (Kuppens and Verduyn 2017). Might period-to-period shift in sentiment affect evaluations, and if so, how?

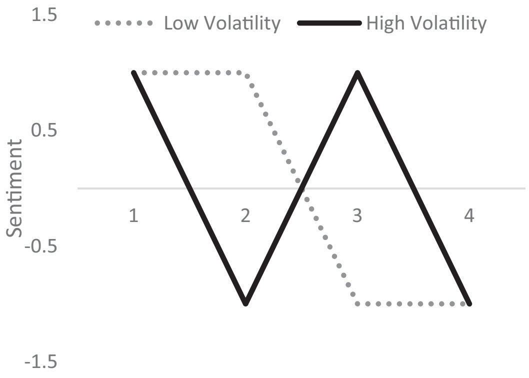
## VOLATILITY

We suggest that volatility plays an important role. The term volatility is often used to describe variation or dispersion. In the case of the stock market, for example, volatile stocks are those that frequently fluctuate up and down.

Volatility can also be used to describe the emotional nature of an experience. Consider, for example, two different four-part sequences, such as chunks of a movie

**FIGURE 1**

EXAMPLE OF VOLATILITY



(figure 1).The low volatility sequence has two positive periods followed by two negative ones (i.e., $P_1P_2N_1N_2$), while the high volatility sequence has the same four periods but alternates positive and negative, positive and negative (i.e., $P_1N_1P_2N_2$). Both experiences have the same average valence, the same peak and end, and even the same average distance from the mean, but the second is more volatile. It is characterized by greater period-to-period shifts in valence.

We suggest such volatility in valence, what we term “sentiment volatility,” can increase evaluations. While no work has empirically examined this possibility, some have theorized that rapid changes, changes in direction, or stringing positive and negative scenes together might make stories more engaging (Gergen and Gergen 1988).

One reason sentiment volatility may be beneficial is because it increases stimulation. People get bored of eating the same foods (Rolls et al. 1981) or listening to the same music (Ratner et al. 1999). Consequently, variety can have a positive impact (Redden 2014). Variation can be stimulating (McAlister and Pessemier 1982; Pessemier and Handelsman 1984), and this stimulation, in turn, can have beneficial effects. Doing more varied activities over the course of an hour can make that time feel more stimulating and exciting, for example, which increases how happy people were with that time (Etkin and Mogilner 2016). While we are unaware of work demonstrating that similar effects occur with *hedonic* variation, research on thrill rides shows something similar. Emotional ups and downs can serve both to induce positive stress (Rietveld and van Beest 2007) and elevate dopamine levels (Norbury and Husain 2015)—both of which can boost evaluations.

Taken together, we hypothesize that sentiment volatility can enhance engagement by boosting stimulation. Like the physical ups and downs of roller coasters, people may also enjoy narratives that provide emotional thrill rides, changing sentiment from one portion to the next.

Building on volatility measures used in finance (Markowitz 1952), we define sentiment volatility as the standard deviation (SD) of differences in sentiment between adjoining chunks of an experience. Specifically, let

$N$ = Total number of chunks
$s_i$ = Average sentiment of chunk $i$
$d_i = s_{i+1} - s_i$

$$\bar{d} = \frac{\sum_{i=1}^{N-1} d_i}{N-1}$$

We define the sentiment volatility of a hedonic sequence as:

$$\text{Sentiment Volatility} = \text{stdev.}(d) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(d_i - \bar{d})^2} \quad \textbf{(1)}$$

There are, of course, alternative ways that temporal variance could be represented, such as the number of directional changes in sentiment (Pham et al. 2001), or the mean absolute deviation. We focus on the SD of differences because it has two attractive psychological properties relative to these alternatives: it allows perceived volatility to be affected by the magnitude (rather than just existence) of directional changes, and gives more weight to extreme changes sentiment—changes that are more likely to be noticed by viewers.[1]

## THE CURRENT RESEARCH

Testing the relationship between sentiment volatility and engagement requires quantifying volatility at scale. Prior work often has people turn dials or move their mouse to rate on-line experiences (Baumgartner et al. 1997; Tan and van den Boom 1992), but while possible for a few items, manually doing this for thousands of cultural items is challenging.

To solve this issue, we use automated textual analysis. Study 1 uses sentiment analysis to calculate sentiment volatility of over 4,000 movies and examine whether volatile movies are evaluated more favorably. We also examine sequels (to provide a stricter test of volatility’s impact),

[1] Our measure also differs from overall standard deviation of sentiment, or variance about the mean (Pham et al. 2001). Overall standard deviation is indifferent to *when* changes in sentiment occur, but for volatility, timing is particularly important. Supporting this idea, a pilot study found that compared to grouping positive and negative chunks together, alternating between them made content seem more volatile (see study 3 for more detail). Nevertheless, we report results with more than one approach to show robustness.

and how the effect varies across genres (to test the underlying role of stimulation).

To test whether these effects generalize to a different domain, and moment-to-moment engagement (rather than after an experience has concluded), study 2 examines over 30,000 pieces of online content. We measure sentiment volatility, and examine whether people are more likely to continue reading more volatile articles.

Finally, to directly test volatility's causal impact, we conduct simple experiments. We take the same four chunks of a movie, manipulate their order, and measure evaluations. We also test and provide evidence for at least one potential mechanism underlying volatility's effects, showing stimulation through both mediation (study 3 and replicates) and moderation (study 4).

Note that we do not mean to suggest that sentiment volatility is the only feature of narratives. Various writers have theorized about narratives arcs or dramatic structure (Freytag 1900; Frijda 1986; Plantinga 2009; Tan 1996) and one paper (Reagan et al. 2016) examines whether the sentiment arc across the entire narrative of books can be clustered into basic shapes. Importantly, however, all this work is about the aggregate, overall arc of a story or narrative, rather than period-to-period emotional changes. Movies may have the same overall aggregate "shape" but with quite different period-to-period progressions along the way. To use an analogy, Reagan et al. (2016) as well as theoretical work on overarching dramatic arcs or narrative structure, focuses on how many peaks different mountain ranges have, while the current paper focuses on whether the walk up and down the sides of those mountains is smooth or more jumpy and jagged (i.e., volatile). We discuss these differences, as well as how other features might shape success, in the general discussion.

## STUDY 1: EMPIRICAL ANALYSIS OF 4,000 MOVIES

Study 1 uses automated textual analysis to measure the sentiment volatility of thousands of movies. We predict that more volatile movies will be more successful.

### Method

First, we collected data on movies. We analyzed English subtitles in the OpenSubtitles2013 corpus, a collection of subtitles gathered from http://www.opensubtitles.org/ (Tiedemann 2012). Most movies were released between 1981 and 2013, and include everything from small indie films (The Marsh) to blockbusters (Star Wars). They span all genres, but the most frequent are dramas, comedies, romances, and thrillers. To ensure similar movies are being compared, we ignored shorts (e.g., <30 minutes) or those with very few words (i.e., <2000), and any files that were mislabeled or inaccurate, leaving 4118 movies. To focus on text that appeared as spoken dialogue, parenthetical indicators (e.g., [music], (laughter), and [gunshot]) were filtered out.

Second, we measure the sentiment of each word in the script. We rely on prior work (Dodds et al. 2011) which scored over 10,000 words based on how positive or negative they made people feel on a nine-point scale.[2] Words like laughter, happiness, and love are rated as highly positive while words like terrorist, suicide, and murder are rated as highly negative and ancillary analyses show that ratings of chunks of test are well correlated with human perceptions.[3] Following prior work (Reagan et al. 2016), we focus on words with clear emotional content (i.e., $\geq 6$ or $\leq 4$). Mean sentiment was 5.93 (SD = .22)

Third, we calculate sentiment volatility, or the SD of differences in sentiment between adjoining chunks. We focus on volatility between sizable chunks of a movie, like scenes, or portions of them. While volatility may also occur at a more granular level (e.g., second-to-second), this is less likely to leave an enduring impression and more likely to be measured with error. Further, if a movie repeatedly oscillated back and forth between highly negative and positive in a matter of seconds, it might exhaust the viewer. Consequently, we examine volatility on a larger scale, looking at how emotional variations across chunks of dialogue (e.g., 1% of the movie) relates to success.

One challenge in constructing emotional trajectories is determining chunk length (i.e., how to break up the text). Unfortunately, there is no clear answer (see Vanden Abeele and Maclachlan 1994 for a discussion of similar challenges in segmenting commercials). While one could imagine chunking movies by scene, scenes length varies greatly. Some movies have shorter scenes and others have longer ones. Even within movies, some scenes are longer than others. Thus, using scene boundaries would involve comparing apples and oranges as scenes would be of different lengths, which could influence sentiment volatility displayed. Further, in many movies, emotional variation

[2] While Linguistic Inquiry and Word Count (Pennebaker et al. 2015) is also well known, it is less ideal for a number of reasons. First, rather than giving words a continuous score, LIWC only classifies them as positive or negative, making it much less fine grained. Second, LIWC covers 85% fewer words (less than 1400) so it provides less coverage. Third, the Dodds et al. (2011) measure has been shown to be more accurate in correctly identifying positivity and negativity in passages of text (Reagan et al. 2016). Indeed, our own preliminary investigation on a small set of movies found that Dodds et al. (2011) measure more accurately reflected actual respondents' ratings of different chunks of movies.

[3] To validate the dictionary, three research assistants (blind to the hypotheses) were given 200 movie chunks and asked to rate how positive or negative they were (−3 to 3). There was reasonably high reliability across raters ($\alpha = .84$) and their ratings were averaged. Their ratings were reasonably well correlated ($r = .66$, $p < .001$) with the Dodds et al. suggesting the automated dictionary does a reasonable job of capturing viewers emotional reaction.

occurs *within* scenes, suggesting that scenes may not be the ideal unit of analysis. Finally, at a practical level, the subtitles data does not demarcate where scenes end and begin and while dividing content into segments by hand can work for a small set of shorter stimuli (Pham et al. 2001) it is challenging to scale.

Consequently, our main analysis relies on the chunking strategy used in prior work (Reagan et al. 2016), applying the same overlapping window to all parts of all movies. As in Reagan et al. (2016), we divide each movie into the same number of segments (i.e., 100) of consistent lengths (i.e., 500 words each) and average sentiment across the words within that segment. That said, as shown below, results are robust to various other chunking approaches: Using the same segment size but different number of segments (e.g., 500 words and 50 segments), different segment sizes (e.g., 1000 words and 50 segments), fixing segment size as well as overlap between segments (e.g., 500-word segments with 100-word overlap), or fixing segment size but having no overlap between segments. This robustness casts doubt on the possibility that the results are driven by the type of segmenting used.

Fourth, we measure cultural success. We recorded user ratings of each movie from IMDB.com (1–10 scale, mean number of ratings per movie = 50,547). We focused on this measure of cultural success, rather than say, critics' reviews, because it is more likely to be driven by individual preferences rather than a small number of institutionalized actors. Finally, an OLS regression examines the relationship between volatility ($M = .74$, SD = .02) and movie success.

## Results

As predicted, more volatile movies receive higher ratings ($b = 9.31$, s.e. = .78, $p < .001$, partial $\eta^2$ = .034, table 1, model 1 and figure 2).[4]

*Robustness Checks.* We included numerous covariates to assess the stability of the main result and test alternative explanations. First, one might wonder whether longer movies or movies from certain genres are more volatile and positively evaluated, and length or genre, rather than volatility, is driving the observed relationship. To test this possibility, we control for movie length (i.e., number of word or running time) and genre fixed effects.

Second, rather than volatility, one might wonder whether the mere presence of emotion (i.e., some movies are more emotional) or valence (i.e., some movies are more positive or negative) is driving the effect. To test this possibility, following prior work (Berger and Milkman 2012) we control for emotionality by the proportion of affect-laden words in the script using Linguistic Inquiry and Word Count (Pennebaker et al. 2015) and control for valence using average sentiment across all segments.[5]

**TABLE 1**

SENTIMENT VOLATILITY AND MOVIE EVALUATIONS

| | (1) | Controls (2) | Budget (3) |
|---|---|---|---|
| Sentiment volatility | 9.31*** | 3.15 *** | 2.81 ** |
| | (.78) | (.90) | (1.02) |
| Movie length | | .01 *** | .01 *** |
| | | (.00) | (.00) |
| Emotionality | | −.01 | −.01 |
| | | (.02) | (.02) |
| Avg. valence | | −.46*** | −.67*** |
| | | (.13) | (.15) |
| Peak | | .27 | .20 |
| | | (.23) | (.26) |
| End | | .03 | .08 |
| | | (.07) | (.08) |
| Complexity | | .01 | .01 |
| | | (.01) | (.01) |
| Stand. Dev of sent | | −.60 | −.62 |
| | | (.55) | (.61) |
| Budget (in 100M) | | | .14 * |
| | | | (.06) |
| Genre dummies | No | Yes | Yes |
| Time dummies | No | Yes | Yes |
| Intercept | 5.58*** | 8.09 *** | 8.98 *** |
| | (.06) | (.96) | (1.13) |
| Adjusted $R^2$ | .033 | .243 | .232 |
| *N* | 4118 | 4118 | 3350 |

NOTES.– *** $p < .001$, ** $p < .01$, * $p < .05$.

Third, one could wonder whether volatile movies have higher peaks or ends, and those factors, rather than volatility, is driving the effect. To address this possibility, we control for each movie's peak emotion (using the difference between the global maximum and mean of the overall emotional trajectory) and end (using the film's last segment minus the mean). Given the significant debate about whether peak and end effects always occur (Miron-Shatz 2009; Tully and Meyvis 2016), we do not necessarily expect to find them, and merely to include them as controls. Results are robust to different approaches to measuring peak (e.g., maximum minus minimum) and end (e.g., the last segment, last few segments, or last minus minimum).

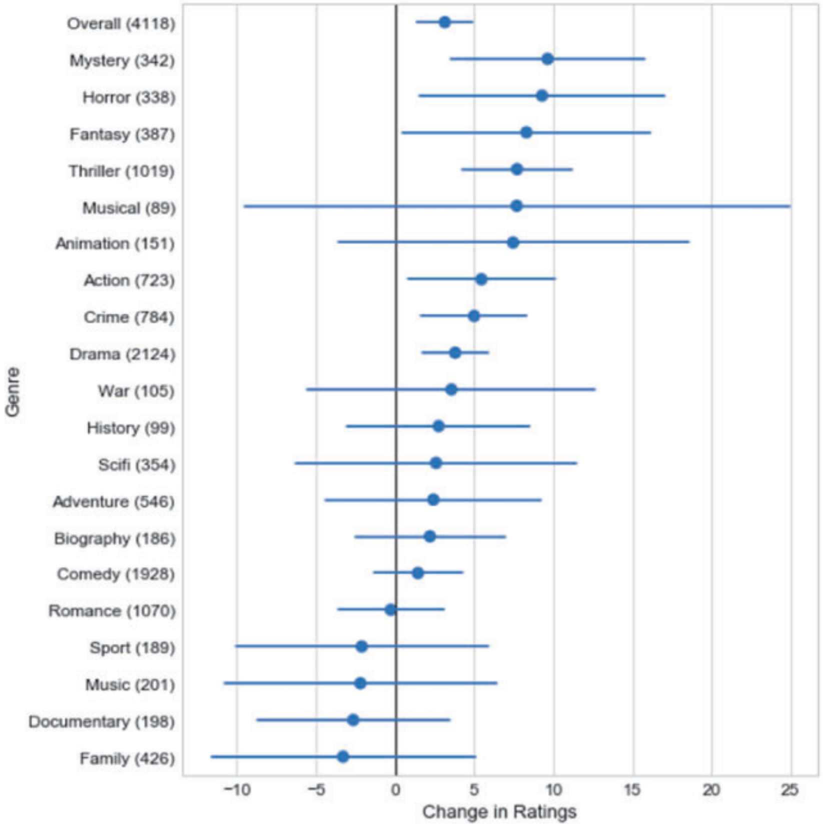
Fourth, maybe volatile movies are also more complex, and complexity is driving success. To address this, we control for complexity using Flesch-Kincaid Grade Level (Kincaid et al. 1975) which measures the number of years of education required to understand a text.

[4] There is no significant quadratic effect of volatility, casting doubt on the notion that the relationship between sentiment volatility and movie ratings is nonlinear.

[5] While one might suggest using hedonometer (Dodds et al. 2011) for this analysis, it is less ideal because it would require picking cut points. Hedonometer provides a more continuous measure of sentiment, but to determine whether certain movies are more or less emotional, we need a way to determine whether words are emotional or not rather than how positive or negative they are. Thus, LIWC seemed more appropriate.

**FIGURE 2**

IMPACT OF SENTIMENT VOLATILITY ACROSS GENRES



Note: Coefficient estimates and 95% confidence intervals for all genres (first row) and for each individual genre (subsequent rows). The effect is significant for a genre if the confidence interval does not intersect with zero. Numbers in parentheses are the number of films in each genre. Most films are tagged with multiple genres.

Fifth, rather than short-term sentiment volatility, maybe our measure is picking up extremeness, or how much a movie's emotional trajectory diverges from the mean, and that is driving the effect. To address this possibility, we control for how much a movie's emotional trajectory diverges from the mean using the SD of sentiment.

Sixth, one might wonder whether recent movies are more volatile and rated more positively, and thus time explains the observed relationship. To address this possibility, we control for time using release year dummies. Results are the same using a continuous measure of time.

Even controlling for all these factors, however, the link between sentiment volatility and success persists (table 1, model 2). While the coefficient's size is reduced by two-thirds, it remains highly significant. A 31.75% increase in sentiment volatility is linked to a one-point increase in ratings.

One could also wonder whether the results were somehow driven by the movies' budget. Budget information was not available for all the movies, but even among the reduced set, the result still persists (with controls: $b = 2.81$, s.e. $= 1.02$, $p = .006$, partial $\eta^2 = .002$, table 1, model 3). There is almost no relationship between budget and ratings ($r = .04$).

*Chunking Strategy*. Results are also robust to different ways of dividing movies into chunks. Whether using a different number of segments (e.g., 500 words, 50 segments: $b = 5.54$, s.e. $= .48$, $p < .001$, partial $\eta^2 = .032$) or different segment sizes (e.g., 1000 words and 50 segments: $b = 10.79$, s.e. $= .92$, $p < .001$, partial $\eta^2 = .032$, or 100 segments: $b = 17.42$, s.e. $= 1.51$, $p < .001$, partial $\eta^2 = .031$), results remain the same.

We also examined alternate ways of chunking, such as fixing the segment size as well as the overlap between segments (e.g., 500-word segments with 100-word overlap) or fixing the segment size but having no overlap between segments. The latter allows us to avoid any concerns about forcing movies of different lengths into the same number of segments. We control for movie length in these approaches, as some movies have more segments. In all cases, results remain the same.

*Identifying the Effect.* To best identify volatility's effect, one would ideally keep all other movie aspects the same, vary volatility, and measure its influence on success. To approximate this, we examine movies with sequels (e.g., Harry Potter). While many focal actors and production members remain the same, sentiment volatility varies across different movies in the series, providing a stricter test of volatility's impact. If among movies in a particular series, more volatile ones are more successful, this would further the notion that volatility, rather than some other factor, is driving success. We analyze the 175 films that are part of a series (either the original or a sequel) using an analysis approach similar to difference-in-differences, testing whether difference in sentiment volatility between movies in a series is linked to the difference in ratings between those movies.

Underscoring the prior findings, even within a series, more volatile films are evaluated more favorably ($b = 8.86$, s.e. $= 4.32$, $p = .04$, partial $\eta^2 = .025$). While one could argue that this is driven by the original films, this is not the case. Ignoring originals and just examining later films in a series (i.e., the second vs. third) shows the same relationship; even looking among sequels in the same series, those sequels that are more volatile are more successful ($b = 19.37$, s.e. $= 6.19$, $p = .003$, partial $\eta^2 = .144$).

*Variation by Genre.* Further evidence that sentiment volatility increases ratings comes from looking across genres. What make a good thriller, for example, is different than what makes for a successful romance. While thrillers should be more appealing the more stimulation they elicit, stimulation should have less of an impact in genres like romance. Consequently, if sentiment volatility is truly shaping evaluations, as we suggest, it should have a stronger effect in genres where stimulation matters more (e.g., thrillers rather than romances).

Consistent with this notion (figure 2), the relationship was strongest in genres like thrillers ($b = 12.50$, s.e. $=$ 1.52, $p < .001$, partial $\eta^2 = .006$) and mysteries ($b = 12.93$, s.e. $= 2.41$, $p < .001$, partial $\eta^2 = .008$) and weakest in genres like music, documentaries, and romance ($b = 2.55$, s.e. $= 3.48$, $p = .46$, partial $\eta^2 = .003$; $b = 3.40$, s.e. $=$ 2.23, $p = .13$, partial $\eta^2 = .012$; and $b = 2.94$, s.e. $= 1.51$, $p = .05$, partial $\eta^2 = .004$, respectively).

## DISCUSSION

Textual analysis of thousands of movies provides preliminary evidence of the value of sentiment volatility. More volatile movies were evaluated more positively.

Robustness checks cast doubt on a number of alternative explanations. While the relationship between volatility and success relationship is reduced when controlling for other features (i.e., average valence, peaks, and ends) and factors (i.e., genre, movie length, emotionality, complexity, extremity, and release year), it persists, and is robust to a range of other chunking approaches. Further, even within movies that are part of the same series, those that are more volatile are liked more.

In addition, consistent with the underlying role of stimulation, rather than being equally beneficial across genres, volatility has a larger impact on evaluations in genres where stimulation is more desirable (e.g., thrillers rather than romance).

*Ancillary Analyses.* We also conducted a number of other robustness checks (see web appendix for more details). We focused on ratings rather than sales because they are more likely to be driven by individual preferences rather than institutionalized actors (i.e., studio and theater executives). Further, sales depend on things like how many theaters show the movie, which not only are unobserved, but could also be endogenous. Consequently, ratings are cleaner.

That said, ticket sales show the same effect (web appendix table A1). We collected sales data, as well as controls such as budget and studio size, for all movies for which it was available. More volatile movies have higher box office sales ($b = 162.54$ M, s.e. $= 40.64$ M, $p < .001$, partial $\eta^2 =$ .004). This persists when including the control variables ($b = 142.77$ M, s.e. $= 47.30$ M, $p = .003$, partial $\eta^2 = .003$) or taking the log of ticket sales to reduce the impact of outliers ($b = 11.35$ M, s.e. $= 2.12$, $p < .001$, partial $\eta^2 =$ .007). It is also robust to different block sizes (e.g., 500 or 1000 words), numbers of blocks (e.g., 50 or 100), overlap between blocks (e.g., 500 words with 100 words of overlap), or allowing number of blocks to vary across movies.

Results are also the same using other measures of identifying sentiment (e.g., Evaluative Lexicon's valence measure, Rocklage, Rucker, and Nordgren 2018; see web appendix table A3) or using an alternate measure of complexity (i.e., number of named entities, or noun phrases that refer to specific individuals, places, or organizations, see web appendix table A4).

We also looked at several alternate measures of sentiment volatility. First, results are the same when standardizing the sequences (i.e., scaling between 0 and 1, $b = 1.73$, s.e. $= .81$, $p = .03$, partial $\eta^2 = .001$, see web appendix table A2).[6] This

[6] Let $\boldsymbol{S}'$ denote the transformation of sentiment sequence $\boldsymbol{S}$, such that all sentiment scores are standardized to be between 0 and 1.

$$\boldsymbol{S}' = \boldsymbol{S} - \frac{\boldsymbol{S} - \min(\boldsymbol{S})}{\max(\boldsymbol{S}) - \min(\boldsymbol{S})}$$

Let $\boldsymbol{D}$ denote the sequence of sentiment changes in $\boldsymbol{S}$ of length $\boldsymbol{N}$. $\boldsymbol{D}$ is comprised of $\boldsymbol{d_i} = \mathrm{s_{i+1}} - \mathrm{s_i}$, where $\mathrm{s_i}$represents the sentiment score of $\boldsymbol{S}$ at position $\boldsymbol{i}$, $\boldsymbol{i} = 1, 2, 3, \ldots, \boldsymbol{N} - 1$. Then, the volatility for the standardized sequence $\boldsymbol{S}'$ may be calculated as follows:

$$\boldsymbol{d}_i' = \boldsymbol{s}_{i+1}' - \boldsymbol{s}_i' = \frac{\boldsymbol{s}_{i+1} - \min(\boldsymbol{S})}{\max(\boldsymbol{S}) - \min(\boldsymbol{S})} - \boldsymbol{s}_i - \min(\boldsymbol{S}) \quad \max(\boldsymbol{S}) - \min(\boldsymbol{S}) = \frac{\boldsymbol{d}_i}{\max(\boldsymbol{S}) - \min(\boldsymbol{S})} \quad \text{Volatility}_{\boldsymbol{S}'} = \sigma(\boldsymbol{D}') = \sigma\left(\frac{\boldsymbol{D}}{\max(\boldsymbol{S}) - \min(\boldsymbol{S})}\right) = \sigma(\boldsymbol{D})$$

casts doubt on the notion that the results are somehow driven by how extreme the emotions are overall. Second, while one could count the number of directional changes, this does not seem ideal (i.e., ignores the sizes of changes) and indeed, using this instead of sentiment volatility shows no effect (and reduces the variance explained). Third, overall SD does not really capture volatility because it ignores where different sentiment occurs. Indeed, this measure is highly correlated with peak ($r$ =.72) and including it instead of volatility reduces the amount of variance explained. Fourth, the mean of the absolute differences between consecutive segments would be a sensible alternative measure of volatility. It is highly correlated with the measure of sentiment volatility used ($r$ = .99) and leads to identical results when it is included in the model.

## STUDY 2: EMPIRICAL ANALYSIS OF 30,000 ARTICLES

Results of study 1 are consistent with our theorizing, but one could be concerned that the data does not include auditory and visual features. Unfortunately, scoring sentiment from these aspects is much more challenging. Further, while these aspects shape audience experiences, their incidence is likely correlated with the dialogue, and thus it is unclear how not being able to examine them would lead to incorrect results. Indeed, film theorists have argued that it is the narrative itself, rather than other features, that drives the viewer experience (Plantinga 2009).

That said, to avoid this concern, study 2 examines a context where auditory and visual features play less of a role: written content. People often read content online, but while some content garners sustained attention (i.e., people read the whole thing), other articles are barely glanced at before people move on to something else. We analyze over 600,000 page read events from over 30,000 articles (all without pictures or videos) and examine whether people are more likely to continue consuming more volatile articles.[7]

### Method

We analyze data from Berger, Moe, and Schweidel (2021), including 647,171 page-read events from 32,085 online articles (see web appendix for more detail). This dataset contains a random sample of page-read events (i.e., when a potential reader loads an article) from nine online content sites over a two-week period. The specific sites cover a wide range of topics (e.g., global news, sports, and lifestyle) and were selected because they use fixed layouts (i.e., content is organized the same way across articles), do not include ads, and are not responsive, meaning the page shows up the same way across devices.[8]

*Dependent Variable.* The dataset includes information on how far down the page a user scrolled, as well as textual features of each paragraph. This allows us to examine how the text of prior paragraphs relates to whether a user continues into the next paragraph. In particular, whether users are more likely to continue consuming articles with higher sentiment volatility. Said another way, if a reader is on paragraph 10, are they more likely to continue to paragraph 11 if the prior paragraphs have been more volatile.

*Independent Variable.* We used the same approach as study 1 to measure sentiment volatility. Sentiment ($M$ = 6.00, SD = .64) was measured at the paragraph level, and since paragraphs vary slightly in length, this provides robustness to the chunking methods of study 1. For each paragraph of each article, we calculate sentiment volatility ($M$ = .67, SD = .38), or the SD of differences in sentiment between the adjoining prior paragraphs. If a given reader is on paragraph 10, for example, but has yet to move to paragraph 11, volatility is calculated using the changes in sentiment from paragraphs 1 through 10. When the reader moves on to paragraph 12, volatility is calculated from paragraphs 1 to 12, and so on. Since volatility requires at least three chunks of sentiment, it cannot be calculated for the first and the second paragraph of an article.

*Model.* We use the model for measuring reading depth specified in Berger et al. (2021a). Each reading session $i$ is a sequence where the reader either continues reading or stops at the end of the paragraph. $Y_{ij}$ denotes the action made after paragraph $j$ of reading session $i$, in which $Y_{ij}$=1 if the reader continues and $Y_{ij}$ =0 if they do not. The probability of continuing past paragraph $j$ of reading session $i$ is assumed to be a function of the paragraph-level content variables and control variables. Formally, we estimate the following logistic regression:

$$Y_{ij} \sim \text{Bernoulli}\left(p_{ij}\right)\text{where}$$

$$\text{logit}(p_{it}) = \beta_0 + \sum_k \beta_k \cdot X_{ijk} + \sum_c \gamma_c \cdot Z_{ijc}$$

where $X_{ijk}$ denotes the $k$th independent variable that

$\cdot \frac{1}{\max(\boldsymbol{S})-\min(\boldsymbol{S})}$

[7] Note, we do not mean to suggest that study 2 is a conceptual replication of study 1. Study 1 examined evaluations after a movie has been consumed and study 2 examines a moment-to-moment measure of whether people continue to read online articles. Further, some work (Tan 1996) suggests a distinction between appreciation of a movie after it has been viewed and moment-by-moment interest. That said, some content features contribute to both. Good acting, for example, should increase moment-to-moment interest, and overall evaluations afterwards. Similarly, while the two studies look at different domains, and dependent variables, these differences help test whether the effects generalize to different ways of tapping engagement.

[8] This means that regardless of whether an article was viewed on a phone, desktop, or other device, the content was not reformatted based on viewport size and the line breaks are the same.

represents the content characteristics of paragraph $j$ of reading event $i$ and $Z_{ijc}$ denotes the $c$th control variable.

## Results

Consistent with study 1, volatile content was more successful. People were more likely to continue consuming more volatile articles ($b = 0.1768$, s.e. $= 0.0039$, $p < .001$, OR $= 1.193$, table 2, model 1). Other ways of considering prior paragraphs show the same results, including calculating volatility from the five ($b = 0.2649$, s.e. $= 0.0045$, $p < .001$, OR $= 1.303$) or 10 most recent paragraphs ($b = 0.3862$, s.e. $= 0.0074$, $p < .001$, OR $= 1.471$).

*Robustness Checks.* To test alternative explanations, we include various ancillary measures from study 1 (i.e., emotionality, average valence, end, and complexity).[9] We also included control variables from Berger et al. (2021a) to address dataset specific factors. We controlled for the *publisher*, as different publishers may attract different types of readers, attract readers when they have more or less time, or publish types of articles that encourage longer or shorter reads. We controlled for the *device* (i.e., desktop, mobile, or tablet) as different types of people may use different devices, people may use different devices at different times, and different devices may impact behavior (Ransbotham, Lurie, and Liu 2019). We controlled *article topic* using latent Dirichlet allocation and a 25-topic solution. Rather than assigning each article to one topic, we include the posterior topic probabilities as control variables, allowing us to control for the mix of different topics that may appear in a given article. We controlled for *paragraph length*, given that longer paragraphs might encourage (or discourage) people from reading. We controlled for *percentage read* through article length in words up to that point, using both a linear and quadratic term, as people may be more or less likely to continue depending on how long they have read already. Finally, we control for article level features examined in Berger et al. (2021a), such as the concreteness and familiarity.

Even including over 40 controls, however, the relationship between sentiment volatility and consumption persists. People are more likely to continue consuming content when articles are more volatile ($b = 0.0414$, s.e. $= 0.0065$, $p < .001$, OR $= 1.042$, table 2, model 2).[10] A one-unit increase in sentiment volatility was associated with a 4.23% increase in odds of continued consumption. Results also remain the same if volatility is standardized between 0 and 1 as in study 1 ($b = 1.1252$, s.e. $= 0.0092$, $p < .001$, OR $= 3.081$, see web appendix table A8).[11]

**TABLE 2**

SENTIMENT VOLATILITY AND READING

| | (1) | (2) |
|---|---|---|
| Sentiment volatility | .1768*** | .0414 *** |
| | (.0039) | (.0065) |
| Paragraph length | | −.0063*** |
| | | (.0000) |
| Emotionality | | −.0037*** |
| | | (.0005) |
| Avg. valence | | −.0529*** |
| | | (.0092) |
| Peak | | .1603*** |
| | | (.0072) |
| End | | −.0038* |
| | | (.0029) |
| Complexity | | −.0244*** |
| | | (.0004) |
| Publisher | No | Yes |
| Reading device | No | Yes |
| Article topic | No | Yes |
| Position in article | No | Yes |
| Article level features | No | Yes |
| Intercept | 1.9572*** | 1.0733*** |
| | (.0030) | (.1546) |
| −2LL | 3,396,808 | 2,874,485 |
| *N* | 4,864,005 | 4,864,005 |

NOTES.– *** $p < .001$, ** $p < .01$, * $p < .05$.

## Discussion

Study 2 provides further evidence that sentiment volatility shapes success. People were more likely to continue consuming more volatile content. While the magnitude is reduced when controls are included, the effect remains significant. Demonstrating the effect in an alternate domain, where only textual features play a role, speaks to its generalizability.[12]

Rather than stimulation, one could wonder whether the results of study 2 could be driven by uncertainty reduction. One could argue that emotional volatility makes people feel uncertain, so they want to keep reading to find out how things resolve. That said, note that effect on sentiment

[9] Given not all readers made it to an article's end, end refers to the last paragraph read rather than the last paragraph of the whole article. Standard deviation of sentiment was highly correlated with volatility ($r = .89$), and given collinearity concerns, could not be included. This may have occurred because articles have many fewer segments ($M = 9.55$) than movies (100 in study 1).

[10] One could wonder whether individual random-effects could be used, but unfortunately given the amount of data and number of individuals in the dataset, it was just too computationally intensive. Using latent classes, however, does something similar (though coarser) and finds the same effects as the main analysis.

[11] A decision to continue reading cannot be impacted by things readers have not yet seen, and most readers are exposed to only a portion of the text, not the full thing, so we only use the paragraphs they saw to compute emotional volatility (and thus to standardize this measure).

[12] Given that people tend to think of news as pretty dry and unemotional, one could wonder whether sentiment volatility should matter in this context. That said, one could also argue the exact opposite. Given the domain is somewhat dry, volatility could be even more important in driving continued attention. Regardless, the results here suggest that at least in the context of the 30,000+ articles examined sentiment volatility increases engagement.

volatility persists ($b = 0.0386$, s.e. $= 0.0065$, $p < .001$, OR $= 1.039$) even controlling for a measure of linguistic certainty (i.e., LIWC's certainty measure).

As noted previously, we do not mean to suggest that what makes a movie enjoyable is identical to what encourages continued reading. Consumers may watch movies to be transported into different worlds and read content to inform themselves, so there are certainly factors that shape engagement in one domain and not the other. That said, sentiment volatility may be one factor, among others, that plays a role in both. Moment-to-moment experiences like sentiment volatility may even be less important for things like movies (where once viewers get engaged, they tend to stay until the end) and more important for things like online content (where people may opt-out at any point if interest wanes).

## STUDY 3: EXPERIMENTALLY MANIPULATING VOLATILITY

To directly test sentiment volatility's causal impact, study 3 manipulates it. We take the same four chunks of a movie, two positives and two negatives, and by manipulating their order, manipulate sentiment volatility and measure its impact. We predict that alternating positive and negative scenes will increase evaluations, and that this will be driven by sentiment volatility.

In addition, we measure stimulation, and test whether it plays a role in driving the effect.

### Method

Participants ($N = 142$) completed a film study as part of a larger group of studies (see web appendix for more detail on all experiments). They were told the experimenters were interested in how people react to film scripts, and that they would read some scenes from a movie script and answer some questions.

All participants read an overview of the movie. They were told about some college students who loved a singer that was coming to perform and volunteered to help put on the show in exchange for tickets. At the last minute, the company won't give the students the tickets, so the students hatch a four-part plan to steal them. Each part of the plan is independent, occurs simultaneously, and involves different pairs of the students, so could occur in any order.

Participants were then exposed to each of the four different scenes. Two of the scenes ($P_1$ and $P_2$) were pretested to be positive (e.g., they successfully sneak into the box office) and two ($N_1$ and $N_2$) were pretested to be negative (e.g., try to hack into the surveillance system but leave a trail that makes them easy to find).[13]

The only difference between conditions was the order in which the scenes occurred. The scenes were designed so that they could be considered in any order and still make sense. In the low volatility condition, same valence scenes were grouped together. Participants were either exposed to two positive scenes followed by two negative ones, or two negative ones followed by two positive ones. In the high volatility condition, however, positive and negative scenes were interspersed, to generate a more volatile experience (i.e., negative, positive, negative, positive, or positive, negative, positive, negative). The appearance of scenes was fully randomized, so among participants who saw a positive scene first, for example, some saw $P_1$ first while others saw $P_2$. Pretest data confirmed that the movie was seen as more volatile when it alternated between positive and negative scenes.[14]

Next, we measured the hypothesized underlying process. Participants were asked how stimulating the plot was (1 = not at all, 7 = extremely).

Then, after reading a brief summary of the end of the movie, participants completed the main dependent variable. They reported how much they liked the movie (1 = didn't like at all, 7 = liked a great deal).

Finally, we tested alternative explanations. First, to examine whether mood could explain the results, participants rated the overall mood of the movie (1 = very positive, 7 = very negative). Second, to test whether the volatile script was simply more natural or the plot made more sense, participants rated "how natural was the order of the four scenes you read" (1 = not at all, 7 = extremely) and "how much sense did the plot make" (1 = very little, 7 = a great deal).

### Results

First, as predicted, a one-way ANOVA found that sentiment volatility influenced evaluations. Boosting volatility made participants like the movie more ($M = 4.01$ vs. 3.41; $F(1, 141) = 4.39$, $p = .038$, $\eta^2 = .018$).

Second, as expected, volatility also influenced stimulation. Boosting volatility made participants think the movie was more stimulating ($M = 4.76$ vs. 4.12; $F(1, 141) = 6.79$, $p = .01$, $\eta^2 = .046$).

Third, as predicted, increased stimulation mediated the effect. A bias-corrected bootstrapping mediation analysis

[13] Participants read one scene and rated how they felt ($-3$ = very negative, 3 = very positive). The two positive scenes evoked a positive mood ($M = 0.54$ and 1.00, compared to the scale midpoint $ts > 49$, $ps < .001$) and the two negative scenes evoked a negative mood ($M = -0.78$ and -1.01, compared to the scale midpoint $ts > 29$, $ps < .001$)

[14] After reading the script, participants ($N = 258$) were asked "Thinking back to the different scenes, how emotionally volatile was the movie? That is, how much did the emotion valence change between scenes?" As expected, alternating between positive and negative scenes made the movie seem more volatile ($M = 4.73$ vs. 4.32; $F(1, 257) = 11.26$, $p = .01$), confirming the effectiveness of the manipulation.

generated a 95% confidence interval around the indirect effect of stimulation that excluded zero (Hayes 2018) and found that stimulation drove the impact of volatility on evaluations [$ab = .48$, 95% CI .11 to .92]. Boosting volatility made the movie seem more stimulating ($b = .64$, $t = 2.48$, $p = .011$) which made people like it more ($b = .75$, $t = 7.96$, $p < .001$).

## Discussion

Study 3 provides experimental support for the results observed in the field. Reorganizing a movie to intersperse positive and negative chunks made people like it more. Further, this was mediated by stimulation.

*Alternative explanations.* One could wonder whether the effects were driven by mood, but this did not differ between conditions ($M = 5.04$ vs. 4.95; $F(1, 141) = 0.28$, $p > .6$). There was no difference between conditions on either how natural the movie seemed ($M = 4.85$ vs. 4.66; $F(1, 141) = .75$, $p = .39$) or how much sense the plot made ($M = 5.22$ vs. 5.20; $F(1, 141) = 0.01$, $p = .94$).

# REPLICATES—STUDY 3A, 3B, AND 3C

To test robustness, we also ran three replicates of study 3, using the same general set-up but different stimuli. Participants read "scripts from a movie," and we rearranged scene order to manipulate emotional volatility. Further, we tested whether uncertainty reduction could explain the effects. More detail appears in the web appendix, but below are the main results.

*Study 3A*. Increasing sentiment volatility made participant like the movie more ($M = 3.78$ vs. 2.00); $F(1, 105) = 6.19$, $p = .014$, $\eta^2 = .056$, and think it was more stimulating ($M = 4.00$ vs. 3.27); $F(1, 105) = 5.72$, $p = .019$, $\eta^2 = .052$, and a bias-corrected bootstrapping mediation analysis found that stimulation drove the impact of scene order on evaluations [$ab = .50$, 95% CI .11 to .98]. There was no impact of condition on how uncertain participants felt in either study 3A ($F < .01$, $p > .9$), study 3B ($F < 0.9$, $p > .35$), or study 3 C ($F < 0.75$, $p > .4$).

*Study 3B*. Boosting volatility increased movie evaluations ($M = 4.32$ vs. 3.60); $F(1, 111) = 4.72$, $p = .032$, $\eta^2 = .041$, and stimulation ($M = 4.61$ vs. 3.64); $F(1, 111) = 9.45$, $p < .001$, $\eta^2 = .079$, and stimulation mediated the effect of volatility on evaluations [ab = .76, 95% CI .29 to 1.26].

*Study 3C*. Boosting volatility increased evaluations ($M = 4.54$ vs. 3.72; $F(1, 108) = 5.84$, $p = .017$, $\eta^2 = .052$) and stimulation ($M = 4.58$ vs. 3.74); $F(1, 108) = 6.12$, $p = .015$, $\eta^2 = .054$, and stimulation mediated the effect of volatility on evaluations [ab = .63, 95% CI .13 to 1.15].

# STUDY 4: PROCESS BY MODERATION

Study 4 further tests the role of stimulation through moderation. If sentiment volatility boost evaluations through increasing stimulation, as we suggest, then its effects should be larger among people who prefer stimulation. Study 4 tests this possibility.

## Method

Participants ($N = 168$) completed a similar study to study 3. We manipulated sentiment volatility and measured how much participants liked the movie. In addition, we measured individual differences in optimal stimulation (i.e., seven items adapted from Steenkamp and Baumgartner 1996 such as "When things get boring, I like to find some new and unfamiliar experience," averaged to form an index). A regression examined movie evaluations as a function of volatility condition, individual differences in optimal stimulation level (mean-centered), and their interaction.

## Results

In addition to an effect of volatility ($B = .44$, SE = .25, $t(164) = 1.78$, $p = .078$), results revealed the predicted volatility × optimal stimulation interaction ($B = .60$, SE = .30, $t(164) = 2.02$, $p = .045$). Spotlight analyses at one SD above and below the mean of optimal stimulation level provide deeper insight into the pattern of results. Among people who prefer higher levels of stimulation (+1SD), making movie more volatile increased evaluations ($B = .95$, SE = .35, $t(164) = 2.72$, $p = .007$). Among people who do not desire higher levels of stimulation (−1SD), however, there was no such effect ($B = -.07$, SE = .36, $t(164) = -.20$, $p = .84$).

## Discussion

Study 4 provides further support for the underlying role of stimulation in these effects. Boosting sentiment volatility made people like a movie more, but this effect was stronger among people who prefer more stimulation.

Note that attention has difficulty explaining the results. Increased attention can be seen as a signal of liking (Norton, Mochon, and Dan 2011), so if sentiment volatility encourages people to pay more attention to the content, one could argue that increased attention, rather than stimulation is increasing evaluations. But this does not seem to be the case. There was no effect of sentiment volatility on time spent on any of the individual pages ($F$s $< 2.05$, $p$s $> .15$), time spent was not correlated with movie evaluations ($|r|$s $< .06$, $p$s $> .47$), and there was no interactive effect of sentiment volatility and optimal stimulation on time spent. Taken together, that casts doubt on the possibility that increased attention is driving the effects.

# GENERAL DISCUSSION

Why are some cultural items more successful? A combination of field data and experiments demonstrates that sentiment volatility may play a role. Automated textual analysis of thousands of movies demonstrates that more volatile content is evaluated more positively (study 1). Similar analysis of over 30,000 pieces of online content demonstrate that people are more likely to continue consuming more volatile content (study 2). Experimental evidence underscores volatility's causal effect (study 3, study 4, and replicates). Taken together, these results demonstrate that sentiment volatility can impact evaluations and shows how these effects may play out in the field.

Additional results provide insight into the underlying process through both mediation and moderation. One reason volatility increases evaluations is because it makes content more stimulating (study 3 and replicates). Consistent with this notion, volatility has a more positive effect on evaluations among people who prefer stimulation (study 4) and in genres where evaluations are more likely to be driven by how stimulating content is (e.g., thrillers rather than romances, study 1). While sentiment volatility's effect of is likely multiply determined, these results demonstrate that stimulation plays a role.

## Contributions

This work makes several contributions. First, these findings contribute to research on the psychological foundations of culture (Berger et al. 2012; Kashima 2008; Schaller and Crandall 2004). When shared across individuals, psychological processes can shape the processing, evaluation, and sharing of cultural items, which in turn, shapes their success (Berger and Packard 2018; Norenzayan et al. 2006). In this case, sentiment volatility boosts the evaluation and consumption of content.

Second, the results shed light on emotional dynamics. While prior work illustrates the importance of specific moments (e.g., peak and end), or average valence, this research demonstrates that how sentiment evolves over the course of an experience also plays a role. Future work might examine other aspects of emotional trajectories and how they shape evaluations.

Third, the findings contribute to research on variety and hedonic adaptation. While decades of variety research has examined varied consumption experiences (e.g., eating different foods), there has been less attention to variation in emotional experiences. Similarly, while inserting breaks (Nelson and Meyvis 2008) or other experiences (Nelson, Meyvis, and Galak 2009) between chunks of an experience can stem hedonic adaptation, our research suggests that variation within an experience itself may also provide benefits. Ads, for example, may make an enjoyable television show more enjoyable, but even within the show itself, sentiment volatility should shape evaluations.

Fourth, the work contributes to a burgeoning stream of literature extracting insight from text. Researchers have long been interested in why some things succeed and fail, but measurement has been a key challenge. Natural language processing, however, provides a reliable method of extracting features, and doing so at scale (Berger et al. 2021b; Humphreys and Wang 2018). This emerging toolkit can hopefully shed light not only on cultural analytics and cultural success, but a range of other interesting questions.

Fifth, the findings have obvious implications for cultural producers. Boosting sentiment volatility (at least within a reasonable range) should increase evaluations. Writers might consider incorporating more sentiment volatility into books, movies, and other content, and as shown in the experiments, this may be as simple as reordering narrative chunks.

## Limitations

First, as noted, movies are much more than just the words actors speak. Automatically measuring visual and audio information, however, is not trivial. Software packages like Praat (Boersma 2002) can be used to extract paralinguistic features (e.g., pitch and tone) from interactions, but there is less literature linking these (or image) features to psychological processes. That said, the fact that a measure constructed from words alone helps explain outcomes suggests that the effect of sentiment volatility could be even larger once other features are taken into account.

Second, one might wonder whether volatility's effects are driven solely by changes in direction. In our measure, volatility can come from changes in direction (i.e., action becoming more positive and then more negative) or changes that do not involve changes in direction (i.e., action accelerating from good, to slightly better, to much, much better). To attempt to separate these aspects empirically, we decomposed the study 1 data into chunk-to-chunk changes that involved changes in direction and those that did not. Then, we took the average size of each type of change, as well as the proportion of each type, across each movie and related these aspects to ratings. Results suggest that both changes in direction, as well as changes in the same direction separately contribute to evaluations. These analyses are far from perfect, but they suggest that the effect of sentiment volatility is driven by swings in emotion, whether of whether they involved changes in direction.

Third, sentiment volatility's effect is likely multiply determined. While stimulation seems to play a role, so may hedonic contrasts. The first bite of a tasty sandwich is delicious, but people soon adapt, and the tenth bite is not as hedonically positive. Volatility, however, may stem this reduction in subjective intensity. While five positive scenes in a row may make the last scene feel less positive, alternating between more and less positive periods, should intensify the experience.

Fourth, we do not mean to suggest that sentiment volatility always has a positive effect. In study 2, for example, there is a negative quadratic effect such that the benefit of sentiment volatility increases at a decreasing rate. That said, we do not see the same pattern in study 1, where if anything the quadratic effect is positive. This difference could be driven by the fact that volatility is measured over a much shorter timescale in study 2, or because the range of volatility is much higher. One could even imagine that too much volatility might eventually become overwhelming, and thus volatility might be additive over the course of an experience.

Fifth, cultural success is clearly driven by a host of other factors. The evaluation of movies, for example, may also depend on judgments after the film has been processed (Tan 1996), such as whether it made the viewer think, changed their perspective, or seemed relevant to their own lives. Whether readers continue to read online articles may depend on whether the topic is of interest or the content evokes uncertainty.

Sixth, Studies 3 and 4 are quite simplistic. They were designed to test the causal impact of sentiment volatility, and whether stimulation could contribute to this effect, but future work could design more complex investigations.

## Other Features of Narratives

Volatility focuses on a micro level, or period-to-period change, but more macro shifts, or changes across several periods, should also be important. In narratives, for example, characters often have to overcome various intermediate barriers to success before achieving a happy ending. Consequently, the broader trajectory seems to follow a long wave-like pattern: Starting low and then slowly building up before going down and up again.[15] Overcoming lows may make the highs more impactful. Said another way, reaching the top of a mountain is more exciting and satisfying if you hiked up from a valley rather than got dropped off by a helicopter. This may relate to ideas of balance being disturbed and then restored (Frijda 1986; Tan 1996) or that the essence of drama is conflict (Field 2005). Consequently, enjoyable narratives may blend volatile moment-to-moment changes with larger aggregate waves of ups and downs.

The distance between such peaks might also influence evaluations. Vertical change, from negative to positive, may be beneficial, but the duration over which that change occurs is likely also important. Put differently, part of the reason hiking up the mountain all the way from the valley is so impactful is that it took a while to get there. Wins are savored more if they took a while to develop. Consequently, beyond period-to-period volatility, too much aggregate change, too quickly, is likely not as positive.

The way peaks develop might also play a role. Levels of a video game often build on one another. In the first level, the character must overcome a small challenge. The next level, a slightly larger challenge, and so on. This engages a range of players, and lets them practice and build their skills, but there may also be a narrative benefit. Saving the biggest challenge for last may lead to the most emotional payoff. Slowly increasing extremity may increase evaluations. At the beginning, both the lows and highs are small. But as the narrative develops, they get larger and larger. That said, one could also make an argument for an alternate structure. The largest valley near the beginning to set the stage, and smaller ones thereafter. It would be interesting to empirically examine this more deeply.

Other features beyond sentiment also deserve attention. Some stories move relatively quickly, for example, while others spend longer on related ideas before moving on to different ones. Consequently, one could imagine measuring the speed of semantic progression, or pace (Laurino Dos Santos and Berger 2021). Moving too quickly may lose an audience, but moving too slowly may be boring. Consequently, the effect of semantic progression may depend on genre. Faster semantic progression may be good for thrillers but bad for romantic comedies. Similarly, one could imagine capturing how much ground the narrative covers, or the circuitousness of its path (Toubia, Berger, and Eliashberg 2021). Covering a lot of ground may make narratives more interesting and increase the impact of other forms of discourse like academic papers.

Plot development would also be interesting to measure. Many authors have theorized about narrative arcs (Freytag 1900; Tan 1996) but quantification is challenging. Ely, Frankel, and Kamenica (2015) apply a theory of surprise and suspense to things like tennis matches and elections, for example, but while this works when there is only one main question (e.g., who will win) it's harder to apply to complex narratives where there are multiple outstanding questions evolving simultaneously (e.g., will the hero defeat the villain, will they reconcile with their parents, and will they end up reuniting with their lost love). Identifying how sections of content relate to these questions, let alone how they move the plot forward on those dimensions, is not trivial. This area provides a rich set of questions for further work, and one that natural language processing tools may be helpful in addressing.

# CONCLUSION

In conclusion, academics and practitioners alike have long been interested in cultural analytics and why some cultural products succeed while others fail. While there is

[15] Pham et al. (2001) looked at the number if runs, or directional changes in valence across commercials as a dependent variable. In addition to number, examining the length of such runs for movies may be useful as an independent variable predicting success.

certainly an art to writing an engaging narrative, this work suggests that there may be some underlying science as well. Natural language processing provides an exciting method for extracting features of narratives and doing so at scale. Hopefully, this method will help unlock the mystery of why some content is so impactful.

## DATA COLLECTION INFORMATION

The first and second author collected the data for the first study online in 2015 by using existing sources. All three authors jointly analyzed the data. The first author collected the data for the second study in 2014 from a company and the second author analyzed it. Data for studies three and four and all replicates were collected via Mturk and analyzed by a research assistant. The data are currently stored in a project directory on the Open Science Framework.

## REFERENCES

Aaker, David. A., Doug M. Stayman, and M. R. Hagerty (1986), "Warmth in Advertising: Measurement, Impact, and Sequence Effects," *Journal of Consumer Research*, 12 (4), 365–81.

Adler, Moshe (1985), "Stardom and Talent," *The American Economic Review*, 75 (1), 208–12.

Bass, Frank (1969), "A New Product Growth for Model Consumer Durables," *Management Science*, 15 (5), 215–27.

Baumgartner, Hans, Mita Sujan, and Dan Padgett (1997), "Patterns of Affective Reactions to Advertisements: The Integration of Moment-to-Moment Responses into Overall Judgement," *Journal of Marketing Research*, 34 (2), 219–32.

Berger, Jonah and Chip Heath (2005), "Idea Habitats: How the Prevalence of Environmental Cues Influences the Success of Ideas," *Cognitive Science*, 29 (2), 195–221.

Berger, Jonah (2011), "Arousal Increases Social Transmission of Information," *Psychological Science*, 22 (7), 891–3.

Berger, Jonah, Eric Bradlow, Alex Braunstein, and Yao Zhang (2012), "From Karen to Katie: Using Baby Names to Study Cultural Evolution," *Psychological Science*, 23 (10), 1067–73.

Berger, Jonah and Katherine L. Milkman (2012), "What Makes Online Content Viral?" *Journal of Marketing Research*, 49 (2), 192–205.

Berger, Jonah and Grant Packard (2018), "Are Atypical Things More Popular?" *Psychological Science*, 29 (7), 1178–84.

Berger, Jonah, Wendy Moe, and David Schweidel (2021a), "What Leads to Longer Reads? Reading Depth in Online Content," Working Paper.

—— (2021b), "Linguistic Drivers of Content Consumption," Working Paper.

Bielby, William T. and Denise D. Bielby (1994), "All Hits Are Flukes: Institutionalized Decision Making and the Rhetoric of Network Prime-Time Program Development," *American Journal of Sociology*, 99 (5), 1287–313.

Booker, Christopher (2004), *The Seven Basic Plots: Why We Tell Stories*, London: A&C Black.

Boyd, Robert and Peter J. Richerson (1985), *Culture and the Evolutionary Process*, Chicago, IL: University of Chicago Press.

Boersma, Paul (2002), "Praat, a System for Doing Phonetics by Computer," *Glot International*, 5, 341–5.

Carver, Charles S. and Michael F. Scheier (1990), "Origins and Functions of Positive and Negative Affect: A Control-Process View," *Psychological Review*, 97 (1), 19–35.

Cavalli-Sforza, Luigi Luca and Marcus W. Feldman (1981), *Cultural Transmission and Evolution*, Vol. 16. Princeton, NJ: Princeton University Press.

Dodds, Peter Sheridan, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth (2011), "Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter," *PLoS ONE*, 6 (12), e26752.

Eliashberg, Joshua, Sam K. Hui, and John J. Zhang (2014), "Assessing Box Office Performance Using Movie Scripts: A Kernel-Based Approach," *IEEE Transactions on Knowledge and Data Engineering*, 26 (11), 2639–48.

Etkin, Jordan and Cassie Mogilner (2016), "Does Variety among Activities Increase Happiness?" *Journal of Consumer Research*, 43 (2), 210–29.

Field, Syd (2005), *Screenplay: The Foundations of Screenwriting*. New York City, NY: Delta Trade Paperbacks.

Fredrickson, Barbara L. and Daniel Kahneman (1993), "Duration Neglect in Retrospective Evaluations of Affective Episodes," *Journal of Personality and Social Psychology*, 65 (1), 45–55.

Freytag, G. (1900), *"Freytag's Technique of the Drama, an Exposition of Dramatic Composition and Art,"* by Dr. Gustav Freytag: An Authorized Translation from the Sixth German Edition by Elias J. MacEwan, M.A., Chicago: Scott, Foresman and Company.

Frijda, N. H. (1986), Studies in Emotion and Social Interaction, The emotions. New York, NY, US: Cambridge University Press; Paris, France: Editions de la Maison des Sciences de l'Homme.

Gergen, K. J. and M. M. Gergen. (1988), "Narrative and the Self as Relationship," *Advances in Experimental Social Psychology, Vol. 21. Social Psychological Studies of the Self: Perspectives and Programs*. San Diego, CA: Academic Press, 17–56.

Hirsch, Paul M. (1972), "Processing Fads and Fashions: An Organization-Set Analysis of Cultural Industry Systems," *American Journal of Sociology*, 77 (4), 639–59.

Hsee, C. K. and R. P. Abelson (1991), "Velocity Relation: Satisfaction as a Function of the First Derivative of Outcome over Time," *Journal of Personality and Social Psychology*, 60 (3), 341–7.

Humphreys, Ashlee and Rebecca Jen-Hui Wang (2018), "Automated Text Analysis for Consumer Research," *Journal of Consumer Research*, 44 (6), 1274–306.

Kahneman, Daniel, Barbara L. Fredrickson, Charles A. Schreiber, and Donald A. Redelmeier (1993), "When More Pain is Preferred to Less: Adding a Better End," *Psychological Science*, 4 (6), 401–5.

Kashima, Yoshihisa (2008), "A Social Psychology of Cultural Dynamics: Examining How Cultures Are Formed, Maintained, and Transformed," *Social and Personality Psychology Compass*, 2 (1), 107–20.

—— (2014), "How Can You Capture Cultural Dynamics?" *Frontiers in Psychology*, 5, 995.

Kincaid, J. Peter, Robert P.. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom (1975), "Derivation of New Readability

Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," *Institute for Simulation and Training*, 56.

Kuppens, Peter and Philippe Verduyn (2017), "Emotion Dynamics," *Current Opinion in Psychology*, 17, 22–6.

Laurino Dos Santos, Henrique and Jonah Berger (2021), "Speed of Semantic Progression," Working Paper.

Markowitz, H.M. (1952), "Portfolio Selection," *The Journal of Finance*, 7 (1), 77–91.

Markus, Hazel R. and Shinobu Kitayama (1991), "Culture and the Self: Implications for Cognition, Emotion, and Motivation," *Psychological Review*, 98 (2), 224–53.

McAlister, Leigh and Edgar Pessemier (1982), "Variety Seeking Behavior: An Interdisciplinary Review," *Journal of Consumer Research*, 9 (3), 311–22.

Miron-Shatz, (2009), "*Evaluating Multi-Episode* Events: A Boundary Condition for the Peak-End Rule," *Emotion*, 9 (2), 206–21.

Moore, Sarah G. and Brent McFerran (2017), "She Said, She Said: Differential Interpersonal Similarities Predict Unique Linguistic Mimicry in Online Word of Mouth," *Journal of the Association for Consumer Research*, 2 (2), 229–45.

Nelson, Leif D. and Tom Meyvis (2008), "Interrupted Consumption: Disrupting Adaptation to Hedonic Experiences," *Journal of Marketing Research*, 45 (6), 654–64.

Nelson, Leif D., Tom Meyvis, and Jeff Galak (2009), "Enhancing the Television-Viewing Experience through Commercial Interruptions," *Journal of Consumer Research*, 36 (2), 160–72.

Netzer, Oded, Alain Lemaire, and Michal Herzenstein (2018), "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications," Working Paper.

Norbury, Agnes and Masud Husain (2015), "Sensation-Seeking: Dopaminergic Modulation and Risk for Psychopathology," *Behavioural Brain Research*, 288, 79–93.

Norenzayan, Ara, Scott Atran, Jason Faulkner, and Mark Schaller (2006), "Memory and Mystery: The Cultural Selection of Minimally Counterintuitive Narratives," *Cognitive Science*, 30 (3), 531–53.

Norton, Michael, Daniel Mochon, and Dan Ariely (2011), "The IKEA Effect: When Labor Leads to Love," *Journal of Consumer Psychology*, 22 (3), 453–60.

Packard, Grant, Sarah Moore, and Brent McFerran (2018), "(I'm) Happy to Help (You): the Impact of Personal Pronoun Use in Customer-Firm Interactions," *Journal of Marketing Research*, 55 (4), 541–55.

Pessemier, Edgar and Moshe Handelsman (1984), "Temporal Variety in Consumer Behavior," *Journal of Marketing Research*, 21 (4), 435–44.

Pennebaker, J. W., R. J. Booth, R. L. Boyd, and M. E. Francis (2015), "*Linguistic Inquiry and Word Count: LIWC2015*", Austin, TX: Pennebaker Conglomerates (www.LIWC.net).

Pennebaker, J.W., R.L., Boyd, K., Jordan, and K. Blackburn. (2015), "*The Development and Psychometric Properties of LIWC2015*", Austin, TX: University of Texas at Austin.

Pham, Michel Tuan, Joel B. Cohen, John W. Pracejus, and G. David Hughes (2001), "Affect Monitoring and the Primacy of Feelings in Judgment," *Journal of Consumer Research*, 28 (September), 167–88.

Plantinga, Carl (2009), *Moving Viewers: American Film and the Spectator's Experience*, Berkeley: University of California Press.

Polti, Georges (1921), *The Thirty-Six Dramatic Situations*, OH: J.K. Reeve.

Pracejus, J. W. and G. Olsen (2004), "The Role of Brand/Cause Fit in the Effectiveness of Cause-Related Marketing Campaigns," *Journal of Business Research*, 57 (6), 635–40.

Ransbotham, Sam, Nicholas H. Lurie, and Hongju Liu (2019), "Creation and Consumption of Mobile Word of Mouth: How Are Mobile Reviews Different?" *Marketing Science*, 38 (5), 773–92.

Ratner, Rebecca K., Barbara E. Kahn, and Daniel Kahneman (1999), "Choosing Less-Preferred Experiences for the Sake of Variety," *Journal of Consumer Research*, 26 (June), 1–15.

Reagan, Andrew J., Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds (2016), "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes," *EPJ Data Science*, 5 (1), 31.

Redden, Joseph P. (2014), "Desire over Time: The Multi-Faceted Nature of Satiation," in *The Psychology of Desire*, ed. Wilhelm Hofmann and Loran Nordgren. New York City, NY: Guilford Press, 82–103.

Redelmeier, Donald A., Joel Katz, and Daniel Kahneman (2003), "Memories of Colonoscopy: A Randomized Trial," *Pain*, 104 (1 – 2), 187–94.

Rietveld, Simon and Ilja van Beest (2007), "Rollercoaster Asthma: When Positive Emotional Stress Interferes with Dyspnea Perception," *Behaviour Research and Therapy*, 45 (5), 977–87.

Rocklage, Matthew D., and Russell H. Fazio (2015), "The Evaluative Lexicon: Adjective Use as a Means of Assessing and Distinguishing Attitude Valence, Extremity, and Emotionality," *Journal of Experimental Social Psychology*, 56, 214–27.

Rocklage, Matthew D., Derek D. Rucker, and Loran F. Nordgren (2018), "Persuasion, Emotion, and Language: The Intent to Persuade Transforms Language via Emotionality," *Psychological Science*, 29 (5), 749–60.

—— (2018), "The Evaluative Lexicon 2.0: The Measurement of Emotionality, Extremity, and Valence in Language," *Behavior Research Methods*, 50 (4), 1327–44.

Rogers, Everett M. (1995), *Diffusion of Innovations*. New York: The Free Press.

Rolls, Barbara J., Edmund T. Rolls, Edward A. Rowe, and Kevin Sweeney (1981), "Sensory Specific Satiety in Man," *Physiology & Behavior*, 27 (July), 137–42.

Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts (2006), "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science*, 311 (5762), 854–6.

Schaller, Mark and Christian S. Crandall (2004), *The Psychological Foundations of Culture*. Mahwah, NJ: Psychology Press.

Simonton, Dean K. (1980), "Thematic Fame, Melodic Originality, and Musical Zeitgeist: A Biographical and Transhistorical Content Analysis," *Journal of Personality and Social Psychology*, 38 (6), 972–83.

Steenkamp, Jan-Benedict and Hans Baumgartner (1996), "Exploratory Consumer Buying Behavior: Conceptualization and Measurement," *International Journal of Research in Marketing*, 13 (2), 121–37.

Tan, Edward S. (1996), *Emotion and the Structure of Narrative Film: Film as an Emotion Machine (B. Fasting, Trans.)*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Tan, Edward and J.M.I. van den Boom (1992), "Explorations in the Psychological Affect Structure of Narrative Film," in *Reader Response to Literature*, ed. E .F. Nardocchio. Berlin and New York: de Gruyter, 57–94.

Tiedemann, Jörg (2012), “Parallel Data, Tools and Interfaces in OPUS,” *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Turkey (May 21–27).

Tobias, Ronald (1993), *20 Master Plots and How to Build Them*. OH: Writers Digest Books.

Toubia, Olivier, Berger Jonah, and Eliashberg Josh (2021), “Quantifying the Semantic Progression of Texts,” Working Paper.

Tully, Stephanie, and Tom Meyvis (2016), “Questioning the End Effect: Endings Are Not Inherently Over-Weighted in Retrospective Evaluations of Experiences,” *Journal of Experimental Psychology: General*, 145 (5), 630–42.

Vanden Abeele, P. and D. I. Maclachlan (1994), “Process Tracing of Emotional Responses to TV Ads: Revisiting the Warmth Monitor,” *Journal of Consumer Research*, 20 (4), 586–600.