International Journal of Communication 17(2023), 2621–2642 1932–8036/20230005

# Predicting Romantic Comedy Success From Content

MELISSA M. MOORE
YOTAM OPHIR[1]
University at Buffalo, State University of New York, USA

Predicting film success has proven challenging, with prior research examining factors including budget, production studios, and stars, to varied degrees of accuracy. Missing was the impact of film scripts and latent linguistic features, examined here through textual analysis. Recent computational study identified the latent thematic content of nearly 200 romantic comedy films, revealing an increasing focus on romantic relationships and tumultuous courtship. We harness the same linguistic model to test whether changes in thematic content are associated with success in reviews, awards, and financial earnings. We find relationship-centered content is positively associated with earnings, mediated by the number of theaters.

*Keywords: cinema/film, cultural studies, Hollywood, media economics, media industries*

Romantic comedy films (romcoms) reflect and shape our cultural understanding of romance, courtship, and gender relations (Segrin & Nabi, 2002). Nevertheless, like other forms of entertainment media, the genre has been considered low-brow and superficial, and therefore largely unworthy of critical review or scientific study (McDonald, 2007). Despite being dismissed as "chick flicks," romcoms are widely favored among 76% of audience members across varied demographics (Morning Consult & The Hollywood Reporter, 2018). They also have an economic foothold; the 10 top-grossing romcoms alone earned a combined $1.7 billion dollars (BoxOfficeMojo [BOM], n.d.). Ultimately, romantic comedy is a popular, successful, and impactful genre worthy of more detailed examination.

Although the genre as a whole is successful and popular, very little is known about what makes a romcom more or less successful over other films in the genre. Those in the film industry often express the sentiment that "nobody knows anything," as success is nearly impossible to predict even for experienced professionals (Walls, 2005, p. 1). The problem of predicting the success or failure of a cultural object is not unique to films. For example, some in the world of books have argued that success is mostly a matter of luck, almost a random occurrence (Archer & Jockers, 2016). Likewise, artist Nia King has argued that the

Melissa M. Moore: mmoore6@buffalo.edu
Yotam Ophir: yotamoph@buffalo.edu
Date submitted: 2022-03-02

[1] We greatly appreciate the contributions from our research assistant, Irena Cao, who collected much of the metadata that facilitated our analyses.

value of art is largely arbitrary (Clements, 2016). In fact, among artists, writers, and publishers, there seems to be a common truism that "success is all about an established name, marketing dollars, or expensive publicity campaigns" (Archer & Jockers, 2016, p. 6).

However, recent studies, and especially those relying on computational methods for the analysis of big data (Archer & Jockers, 2016; Joseph, 2019), have managed to identify latent features that are predictive of success. Importantly, some research demonstrates that success could be predicted not merely from factors external to the artistic object itself (marketing, star power, and so on) but also from objective features internal to the object—or, in the context of books, the text and linguistic components (Archer & Jockers, 2016). Despite important advancements in methodological tools and theory, little is known about the predictors of success in films, or, in particular, film genres. The current study aims to fill that scientific lacuna by estimating the impact of linguistic features in romcom scripts on their financial and critical success. Since a film's success is often determined in box offices, we examine the mechanism of affect in addition to the impact of latent content. Specifically, we examine whether the effect of content on success is mediated by the number of theaters screening the films.

Previous study of the romantic comedy genre, while limited, has focused almost exclusively on how these depictions of courtship and romance affect viewers' perceptions of relationships (e.g., Hefner & Wilson, 2013). As such, very little is known about what makes these films successful. Other studies have constructed predictive models for film success more broadly based on information available before release—most commonly the film's budget, production studio, runtime, release date, number of theaters screening the film, genre, and MPAA ratings (ratings ranging from G to R from the Motion Picture Association of America; e.g., Joseph, 2019; Walls, 2005). For example, Ericson and Grodman (2013) considered data including budget, box office earnings, critic and audience scores, previous awards that actors and directors had won, and top words appearing in plot synopses, among many other variables. Utilizing machine learning, their results explored success as defined by critic and audience scores, gross earnings, and opening weekend earnings. Ultimately, the strongest predictors of earnings were studio and genre, with genre being most important for critic and audience ratings as well. For plot synopses, the words "life" and "story" were positively correlated with higher ratings, while "find" and "must" were correlated with higher box office gross (Ericson & Grodman, 2013). Notably, these data were skewed toward post-1980 films because of limited availability. More importantly, the analysis did not rely on actual language in films but rather on plot summaries that are limited in scope and potentially biased by the person writing the summary.

Some studies focused on the actors and production teams. Although Ericson and Grodman (2013) considered "star power" as far as previous Academy Awards and Golden Globes, Sharda and Delen (2006) averaged recent movie-making prices to define the degree to which actors and actresses contributed to the sale of the movie. They also considered common variables from prior studies, including genres and MPAA ratings, adding whether the film was a sequel or had technical effects (i.e., animation and science fiction content). The researchers used artificial neural networks, a type of machine-learning algorithm, to forecast box office earnings. Their results revealed that the strongest factors were the number of theaters, high use of technical effects, and star power and that neither MPAA rating nor genre was a significant contributor. De Vany and Walls (1999) also included star power, here defined by listings of influential people from other

sources. They found that the most important factors for predicting financial successes were budget, time in theaters, and number of theaters. Ultimately, with time in theaters having such a significant impact, the authors concluded that the audience was the biggest driver of success.

Absent in those analyses and predictive models is the script. A few studies have incorporated text as big data for the analysis of films' successes, but those were limited to either film discussion forums (Krauss, Nann, Simon, Fischbach, & Gloor, 2008) or plot synopses (Lash & Zhao, 2016). In another example, Joshi, Das, Gimpel, and Smith (2010) harnessed sentiments from critic reviews to predict box office revenue with relative success. Krauss et al. (2008) used social network analysis and sentiment analysis of audience forums on the Internet Movie Database (IMDb) and successfully predicted Oscar nominations. Finally, Lash and Zhao (2016) conducted topic modeling on plot synopses to determine financial success based on return on investment (revenue less budget).

Yet, to date, none of these analyses have examined film content directly. Perhaps methodological barriers have prevented large-scale content analyses, particularly the ability to analyze massive corpora of film scripts. As a result, little is known about the effects of discourse and language, especially as directly measured through text. In the lack of empirical evidence, some scholars have theorized and hypothesized about how content would impact success. For example, a core argument of the institutional approach to media analysis (Turow, 1997) is that the predictability of content in different genres, such as the romcom, stems from attempts to cater to audiences' expectations (McDonald, 2007) with the aim to maximize profits (Mortimer, 2010). However, little research has tested whether commercial and critical film success is influenced by linguistic and thematic content. This data would be of interest for investors assessing the value of their contributions toward a film and theaters deciding which films to screen (as suggested by Lash & Zhao, 2016), as well as writers choosing a distributor, production studios choosing details related to release date, genre, and MPAA rating (as suggested by Joseph, 2019), or any of these parties looking to predict revenue.

A recent study (Moore & Ophir, 2022) presented the first systematic, computational analysis of thematic development of romcoms over recent decades. Analyzing the scripts of nearly 200 films using the Analysis of Topic Model Networks approach (ANTMN; Walter & Ophir, 2019), we argued that romcom scripts could be clustered into two broad themes (Moore & Ophir, 2022). The first, *relationships*, involves the plot's central courtship and all relevant elements of that romantic relationship. The second, *life*, contains the films' premises and the nonromantic life events of the characters. The authors identified 40 topics within these two themes, the largest being breakups and screwups (depicting relationships ending and regretful mistakes), introductions (characters meeting), and conflicting feelings (characters experiencing mixed emotions, such as simultaneous joy and embarrassment). Importantly, the results pointed to a consistent increase in *relationships* content, including an increase in depictions of romance that other scholars (e.g., McDonald, 2007) have identified as containing misleading messages about relationships, gender, and romance.

Although analyses from Moore and Ophir (2022) revealed important patterns and trends in the consistent use of *relationships* content in romcoms, it did not examine the real-world impact of those changes. Other scholars have argued that predictable plot lines and tropes contribute to the genre's success (McDonald, 2007; Mortimer, 2010). Yet, no prior study has demonstrated a relationship between financial

and/or critical success and scripts themselves. Here, we fill that scientific gap by examining the relationship between movie content and success using the data and model from Moore and Ophir (2022). More specifically, we examine the relationship between films' language and three indicators of success: lifetime earnings at the box office (defined by BOM, now owned by the IMDb), critical reception by audiences and film critics, and industry awards. Thus, our first question is:

*RQ1: Is there a relationship between a romcom's relationships content and financial earnings?*

Even if a relationship between content and financial success exists, we do not believe it is a direct one. Previous studies have attributed film success, in part, to the number of theaters showing each film (De Vany & Walls, 1999; Joshi et al., 2010; Sharda & Delen, 2006). We argue that the number of theaters screening a movie could mediate the impact of films' content on success. Because audiences cannot anticipate a movie's full content before purchasing tickets, we believe it is more likely that theaters are the decision makers, and it is their decision to screen a film that influences success. We thus ask and hypothesize:

*RQ2: Is there a relationship between a film's relationships content and the number of theaters who will screen a film?*

*H1: The number of theaters screening a film will mediate the relationship between relationships content and financial earnings.*

Previous research has investigated the relationship between studios, critics, and audiences to determine how critics influence box office earnings (i.e., audience demand; Elberse & Eliashberg, 2003; Hsu, 2006; Terry, Butler, & De'Armond, 2005). Although there is debate as to whether professional critics' reviews and scores act as forecasters or influencers, it is commonly assumed that they are a uniquely important audience that studios appeal to directly (Hsu, 2006) and who can significantly impact earnings (Elberse & Eliashberg, 2003; Terry et al., 2005). For example, Elberse and Eliashberg (2003) demonstrated how review scores directly impact the number of theaters screening a film. Interestingly, more positively reviewed films tend to open with fewer screens, and more poorly reviewed films have wider openings. The authors suggest that distributors with low-rated films attempt to recoup potential losses with wide openings, while highly rated films may build momentum over a longer run (and avoid high advertising costs that come with a wide release). Thus, critical acclaim is positively related to opening-week revenue, but negatively related to opening-week screens.

Hsu (2006) explored how films' box office earnings, IMDb reviews, and critic reviews were affected by differences in genre appeal. Hsu's (2006) results show that as films appeal to more genres (creating a "wider niche width"), the audience size increases, but ratings decrease. That is, more critics and audiences are likely to see a "wide niche" film and review it, but both groups are less likely to rate such a film positively. Critics are more likely to review films that they see as relevant to audiences and are therefore more likely to review releases from major distributors, films with high star power (previously successful actors and directors), and films appealing to a wider array of genres. However, critic reviews and audience appeal were both higher for films with narrower "niche widths," likely because there is a stronger consensus for what

those films should look like. Although Hsu (2006) did not include romantic comedy specifically, as it is a boundary-spanning genre, we might expect both audiences and critics to react more positively to films that adhere more closely to the formula of the genre (i.e., including a stronger focus on courtship and *relationships* content rather than deviating from those depictions).

In our consideration of audience and critical reviews, we referred to the same database that provided the list of top-grossing romcoms: IMDb. This is a source that both lay audiences and scholars use for film information (e.g., Ericson & Grodman, 2013; Krauss et al., 2008). IMDb is a comprehensive source containing everything from cast and crew billings to box office earnings, reviews, and awards. Thus, we sourced our audience-review data from the user ratings on IMDb and our critic reviews from the Metascore, which IMDb sources from Metacritic.com. Scholars have established that audiences of romantic comedies expect, desire, and appreciate the genre's focus on romantic themes (Caperello & Migliaccio, 2011). We thus expected that:

*H2: There will be a positive relationship between a romcom's relationships content and its critical reception by (a) audiences and (b) professional critics.*

One of the largest indications of a film's accomplishments is industry awards. In the United States, the two most celebrated ceremonies are the Academy Awards (the Oscars) and the Golden Globe Awards (New York Film Academy, 2018). Some research has looked at predictive factors to see which films are most likely to earn these awards and nominations. As Simonton (2004) illustrates, there is a strong consensus of nominated and awarded films between various awarding bodies. Across the film industry, from the Oscars and Golden Globes to the British Academy of Film and Television Arts, there is a high rate of agreement resulting from a limited number of films receiving honors across the industry.

Given this relationship, it is perhaps unsurprising that the strongest predictor of awards and nominations is the receipt of previous awards and nominations. For instance, Pardoe and Simonton (2008) show that Golden Globes are strong predictors of Oscar success. Nominations for certain awards are more likely for those who have a history of success in related categories. The authors suggest that if this trend continues, it will be possible to predict winners with greater than 70% accuracy depending on the award category (and up to 96% for directors). Similar findings have appeared for other researchers, including Kaplan (2006), who also found that receiving the most nominations in a season and winning the Golden Globe for Best Picture are strong predictors for Oscar Best Picture winners. Further, Kaplan saw that the strongest contenders by far for Best Picture are films that are both epic and biographical and those that have directors honored by the Director's Guild of America.

Although these predictive models shine an important spotlight on an awarding industry that is difficult to understand, we are interested in exploring factors beyond previous successes. We thus step outside the usual analyses in this area to see how the content itself relates to awards. Acknowledging that romcoms rarely win industry awards, we ask:

*RQ3: Is there a relationship between a romcom's relationships content and nomination or winning of awards?*

## *Method*

We began with a topic model of romcom films' thematic content, from Moore and Ophir (2022), built from 188 top-grossing romcoms in the United States as defined by BOM (n.d.). A series of metatextual variables were collected for each film to serve as covariates in models for each of the areas described above. These variables and the complete list of films are displayed in Appendix A in our Open Science Framework repository (https://bit.ly/AppendixAB).

Building the topic model involved multiple steps to identify thematic content and ultimately demonstrate the changes in relationships content discussed throughout this study (summarized in Appendix B; https://bit.ly/AppendixAB). The process began with collecting complete scripts for the 200 top-grossing romantic comedies in the United States from 1980 to 2019, as classified by BOM (n.d.). The scripts were preprocessed for text analysis, and ANTMN (Walter & Ophir, 2019) calculated a topic model based on words cooccurring across scenes. These words become a network of topics (distribution lists of words), with nodes as topics and edges based on words' cooccurrence in scenes. Broader themes were identified with a community detection algorithm (Fastgreedy; Clauset, Newman, & Moore, 2004) based on topics' frequency of coappearing across scenes. Topics and themes were labeled after qualitative examination of highly representative scenes and unique top words for each topic. Finally, Analysis of Variance (ANOVA) and multiple regression examined how themes and topics change over time and now investigates which factors are associated with those changes.

### *Thematic Content*

The key independent variable, the theme of "*relationships*," was based on a quantitative estimation of the percentage of language in a corpus of 188 scripts (for films released between 1980 and 2019) that is linguistically associated with relationships done by Moore and Ophir (2022). In their work, Moore and Ophir used a novel computational method based on unsupervised machine learning to estimate the thematic structure of each film's full script. Their model identified two themes, showing that scripts of romantic comedies are linguistically composed as a balance between topics related to "*relationships*" (e.g., breakups and screwups; conflicting feelings about the relationship; sex and attraction; and more) and topics related to "*life*" (e.g., financial transactions, religious ceremonies, work, and business). The films thus offered a balance between characters' personal lives and their romantic relationships. A major advantage of this approach is its ability to estimate this balance for each film. Notice that the variables are continuous and not binary. In other words, the measurement did not indicate whether a film had a breakup scene or not, but rather how much of the film's script was associated with the breakup topic. That allowed for measuring the prevalence of specific topics (like the breakups and screwups topic) but also an estimation of the cumulative use of the two themes. For example, the film *Don Jon* (Bergman & Gordon-Levitt, 2013) used the *relationships* theme the most, with 80.5% of the language used in the script being associated with relationships. Similarly, *The Five-Year Engagement* (Apatow, Rothman, & Stoller, 2012) used *relationships* language in 76.39% of its content. However, films like *The American President* (Reiner, 1995) relied heavily on *life* topics (66.79% of the script was dedicated to *life* topics), as did *Legal Eagles* (61%; Reitman, 1986), *Pretty Woman* (59.24%; Milchan, Reuther, & Marshall, 1990) and others. These estimations allow us to use thematic content as a continuous variable in our estimations in the current study.

*Financial Earnings*

Financial data were collected from the aforementioned BOM list to indicate opening weekend and lifetime box office earnings (in millions of dollars, adjusted for inflation; Federal Reserve Bank of Minneapolis, n.d.).

*Audience and Critic Reviews*

Audience and critic reviews were collected from IMDb, with audience ratings on a 1–10 scale and critic ratings on a 1–100 scale, sourced from MetaCritic (n.d.). The Metascore is "distilled" into a weighted average from respected critic reviews based on their quality and stature (Metacritic, n.d.). The details of these calculations are available on Metacritic's website.

*Industry Awards*

Additional data about major industry awards and nominations were also collected from IMDb, sourced from the Academy of Motion Picture Arts and Sciences (Academy Awards, a.k.a. Oscars) and from the Hollywood Foreign Press Association (Golden Globe Awards).

*Theater Screens*

The number of theaters screening films during their releases was considered both for opening weekend and in total (referred to as "opening screens" and "lifetime screens," respectively; BOM, n.d.).

*Production Data*

The year of release, production studio, MPAA rating (i.e., G, PG, PG-13, R), and budget, all collected from BOM, were used as control variables. Note that year of release was recoded to indicate the age of each film.

## Results

Table 1 shows the descriptive statistics and correlations between all continuous variables.

***Table 1. Descriptive Statistics and Correlations for Study Variables.***

| Variable | *M* | *SD* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. *Relationships* content | 0.66 | 0.10 | | | | | | | | | | |
| 2. Budget[a] | 44.09 | 27.04 | −.13 | | | | | | | | | |
| 3. Ranking | 94.56 | 57.71 | −.09 | −.42** | | | | | | | | |
| 4. Age (year) | 18.55 | 8.77 | −.34** | .02 | −.02 | | | | | | | |
| 5. IMDb rating | 6.26 | 0.72 | −.08 | −.17* | −.20** | .20** | | | | | | |
| 6. Metascore | 54.24 | 16.93 | −.09 | −.18* | −.15* | .22** | .76** | | | | | |
| 7. Opening screens | 2156.51 | 994.64 | .37** | .38** | −.22** | −.58** | −.44** | −.51** | | | | |
| 8. Lifetime screens | 2366.61 | 766.77 | .36** | .38** | −.35** | −.71** | −.33** | −.38** | .89** | | | |
| 9. Awards and nominations | 1.03 | 0.25 | −.05 | .02 | −.30** | .31** | .45** | .45** | −.45** | −.35** | | |
| 10. Opening weekend earnings[a] | 18.45 | 13.10 | .20** | .46** | −.65** | −.13 | −.16* | −.25** | .61** | .56** | −.17* | |
| 11. Lifetime earnings[a] | 62.28 | 41.47 | .10 | .34** | −.88** | .01 | .21** | .15 | .18* | .33** | .31** | .65** |

[a] In millions, adjusted for inflation. *$p < .05$. **$p < .01$.

Our first question (RQ1) asked whether there is a relationship between a romcom's *relationships* content and financial earnings. We conducted a linear regression comparing opening weekend earnings with the production data indicated above as control variables. This model was significant, $F(11, 164) = 6.35$, $p < .001$, adj. $R^2 = .25$. Budget ($t = 6.56$, $p < .001$) and *relationships* content ($t = 2.72$, $p < .007$) were significantly associated with higher financial earnings. Full details are shown in Table 2. We removed the nonsignificant covariates to present the most parsimonious model, $F(2, 173) = 29.54$, $p < .001$, adj. $R^2 = .25$. Budget ($t = 7.46$, $p < .001$) and *relationships* content ($t = 3.52$, $p < .001$) remained strongly associated with financial earnings during opening weekend.

To explore this connection in greater detail, we conducted a second linear regression exploring lifetime financial earnings based on *relationships* content, controlling for production data. This initial model was significant, $F(11, 164) = 2.95$, $p < .001$, adj. $R^2 = .11$ Both budget ($t = 4.72$, $p < .001$) and *relationships* content ($t = 2.81$, $p < .006$) were significant variables, such that higher budget and increased *relationships* content were associated with higher earnings (see Table 2). We then removed the nonsignificant factors individually to obtain the most parsimonious model, controlling only for budget, $F(2, 173) = 14.44$, $p < .001$ adj. $R^2 = .13$. Budget remained a significant variable ($t = 5.14$, $p < .001$), as did *relationships* content ($t = 2.71$, $p < .007$). Thus, it is clear that both *relationships* content and budget play a significant role in earnings over the lifetime of a theatrical run.

***Table 2. Linear Regression of Thematic Content and Production Attributes on Theatrical Revenue.***

| | Opening weekend | | | | Lifetime | | | |
|---|---|---|---|---|---|---|---|---|
| Effect | Estimate | *SE* | *t* | *p* | Estimate | *SE* | *t* | *p* |
| Intercept | −18.442 | 10.300 | −1.790 | .08 | −51.445 | 36.134 | −1.424 | .16 |
| *Relationships* content | 31.048 | 11.412 | 2.721 | .007** | 112.513 | 40.036 | 2.810 | .006** |
| Year | −18.442 | 10.300 | −1.790 | .08 | 0.302 | 0.414 | 0.729 | .47 |
| Studio[a] | | | | | | | | |
| NBCUniversal | 8.571 | 5.031 | 1.704 | .09 | 14.708 | 17.649 | 0.833 | .41 |
| Other | 2.805 | 5.547 | 0.506 | .61 | 11.276 | 19.459 | 0.579 | .56 |
| Sony Pictures | 9.043 | 4.848 | 1.865 | .06 | 19.078 | 17.006 | 1.122 | .26 |
| ViacomCBS | 9.994 | 5.003 | 1.998 | .05* | 14.501 | 17.550 | 0.826 | .41 |
| Walt Disney Studios | 8.133 | 4.730 | 1.720 | .09 | 16.024 | 16.593 | 0.966 | .34 |
| WarnerMedia | 7.367 | 5.015 | 1.469 | .14 | 18.461 | 17.595 | 1.049 | .30 |
| Budget[b] | 0.222 | 0.034 | 6.555 | <.001*** | 0.559 | 0.119 | 4.717 | <.001*** |
| MPAA rating[c] | | | | | | | | |
| PG-13 | 1.407 | 2.501 | 0.563 | .57 | −11.524 | 8.773 | −1.314 | .19 |
| R | −0.979 | 2.977 | −0.329 | .74 | −6.502 | 10.443 | −0.623 | .53 |

[a] The reference was MGM Holdings. [b] In millions, adjusted for inflation. [c] The reference was PG.

Our second question (RQ2) asked whether there is a relationship between a romcom's *relationships* content and the number of theaters screening the film. We repeated the linear regression with the same production data, examining opening screens first. The result was a significant model $F(11, 164) = 19.73$, $p < .001$, adj. $R^2 = .54$ (see Table 3). The most significant variables were year ($t = -8.74$, $p < .001$), budget ($t = 7.04$, $p < .001$), and *relationships* content ($t = 2.49$, $p < .01$). Removing nonsignificant variables revealed a model with only three significant variables: year ($t = -9.87$, $p < .001$), budget ($t = 7.70$, $p < .001$), and *relationships* content ($t = 2.19$, $p < .03$), $F(3, 172) = 62.37$, $p < .001$, adj. $R^2 = .51$.

We continued this exploration with lifetime screens, again reaching a significant model, $F(11, 164) = 38.59$, $p < .001$, adj. $R^2 = .70$. Again, the significant variables were year ($t = -13.00$, $p < .001$), budget ($t = 9.04$, $p < .001$), and *relationships* content ($t = 3.45$, $p < .001$; see Table 3). This held in the most parsimonious model, $F(3, 172) = 128.8$, $p < .001$, adj. $R^2 = .69$, maintaining only the three aforementioned variables as significant (year, $t = -14.85$, $p < .001$; budget, $t = 9.64$, $p < .001$; *relationships* content, $t = 3.32$, $p < .001$). Thus, whether examining theaters hosting the film during opening weekend or across the film's theatrical run, newer films with higher budgets and higher percentages of *relationships* thematic content are more likely to appear and stay in theaters.

***Table 3. Linear Regression of Thematic Content and Production Attributes on Theater Screens.***

| | Opening weekend | | | | Lifetime | | | |
|---|---|---|---|---|---|---|---|---|
| Effect | Estimate | *SE* | *t* | *p* | Estimate | *SE* | *t* | *p* |
| Intercept | 1241.492 | 594.104 | 2.090 | .04* | 1787.566 | 379.974 | 4.704 | <.001*** |
| *Relationships* content | 1639.490 | 658.265 | 2.491 | .01** | 1453.386 | 421.009 | 3.452 | <.001*** |
| Year | −59.450 | 6.800 | −8.743 | <.001*** | −56.516 | 4.349 | −12.995 | <.001*** |
| Studio[a] | | | | | | | | |
| NBCUniversal | 268.213 | 290.183 | 0.924 | .36 | 79.833 | 185.593 | 0.430 | .67 |
| Other | 158.730 | 319.938 | 0.496 | .62 | 97.339 | 204.624 | 0.476 | .63 |
| Sony Pictures | 224.301 | 279.613 | 0.802 | .42 | 13.795 | 178.833 | 0.077 | .94 |
| ViacomCBS | 564.448 | 288.546 | 1.956 | .05* | 287.021 | 184.547 | 1.555 | .12 |
| Walt Disney Studios | 350.169 | 272.810 | 1.284 | .20 | 140.716 | 174.482 | 0.806 | .42 |
| WarnerMedia | 525.620 | 289.284 | 1.817 | .07 | 318.041 | 185.019 | 1.719 | .09 |
| Budget[b] | 13.719 | 1.949 | 7.038 | <.001*** | 11.270 | 1.247 | 9.040 | <.001*** |
| MPAA rating[c] | | | | | | | | |
| PG-13 | 106.401 | 144.249 | 0.738 | .46 | 65.249 | 92.258 | 0.707 | .48 |
| R | −236.710 | 171.700 | −1.379 | .17 | −81.005 | 109.815 | −0.738 | .46 |

[a] The reference was MGM Holdings. [b] In millions, adjusted for inflation. [c] The reference was PG.

Because of our expectation that theaters would act as a mediator between studios and audiences, we conducted analyses to examine this relationship in greater detail (H1). Since *relationships* content influences theaters' selection of films and time in theaters influences financial success, we conducted a mediation analysis using the mediation package in R (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014) with opening weekend earnings as the outcome, opening screens as mediator, and *relationships* content as the predictor (while controlling for production data). The unstandardized indirect effects were computed for each of 1,000 bootstrapped samples. The mediated effect of *relationships* content on opening weekend earnings was statistically significant (Effect = 14.85, 95% C.I. [1.70, 30.7], $p < .03$). We repeated the analysis with lifetime earnings and lifetime screens, with similar results. The mediated effect of *relationships* content on lifetime earnings was statistically significant (Effect = 52.00, 95% C.I. [17.36, 97.12], $p < .001$). These results are illustrated in Figure 1. Thus, in support of H1, we find that the number of theatrical screenings is a significant mediator in the relationship between *relationships* content and financial earnings, both during opening weekend and throughout the theatrical release of films in this genre.

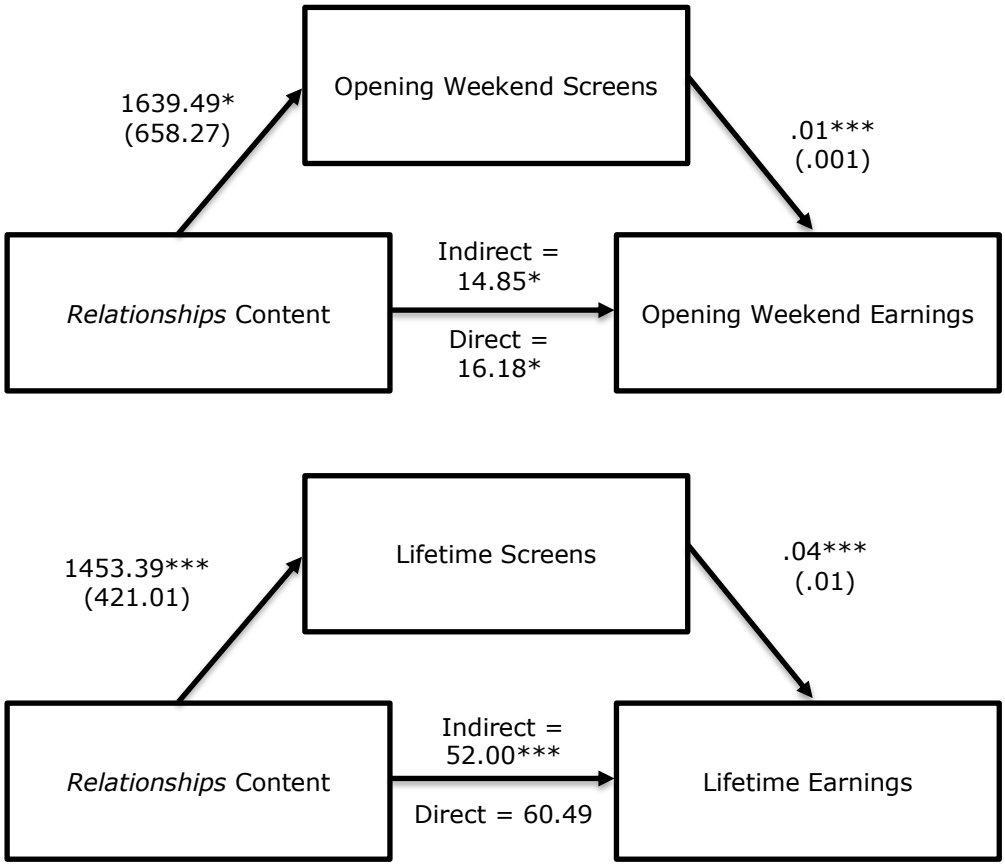


***Figure 1. The mediating effect of theater screenings on earnings during opening weekend and in a total theatrical run. ***p < .001, **p < .01.***

*Note*. Production data were included as control variables and the results are not standardized.

Our second hypothesis (H2) expected that there would be a positive relationship between a romcom's *relationships* content and its critical reception by (a) audiences and (b) professional critics. We first ran a correlation between these two ratings and found that IMDb and Metacritic scores are highly correlated, $r(179) = .74$, $p < .001$. That is, lay audiences and professional critics typically agree on film quality in this genre. To test H2, we conducted a linear regression comparing first IMDb and then Metacritic scores based on *relationship* content, controlling for production data.

The first model with IMDb scores (H2a) was significant, $F(11, 164) = 2.51$, $p < .006$, adj. $R^2 = .09$, and indicated budget ($t = -1.97$, $p < .05$) and year ($t = 2.31$, $p < .02$) as significant variables (see Table 4). We removed the nonsignificant factors to present the most parsimonious model, leaving only year, budget, and *relationships* content, $F(3, 172) = 4.66$, $p < .004$, adj. $R^2 = .06$. However, while budget ($t = -1.99$, $p < .05$) and year ($t = 2.91$, $p < .004$) remained significant variables, romantic content was nonsignificant ($t = 1.50$, $p = .14$). Thus, romantic content does not significantly affect audience ratings on IMDb. Interestingly however, audiences rated older movies and lower budget productions more positively.

We repeated the linear regression above, this time for Metacritic rather than IMDb (H2b). The initial model was significant, $F(11, 159) = 3.21$, $p < .001$, adj. $R^2 = .12$. Only the year was a significant variable ($t = 2.42$, $p < .02$), with marginal differences among studios (see Table 4). We removed nonsignificant factors, leaving budget, year, studio, and *relationships* content $F(9, 161) = 2.81$, $p < .004$, adj. $R^2 = .09$. Budget was marginally significant ($t = -1.86$, $p < .07$) with year as the strongest variable ($t = 2.57$, $p < .01$). *Relationships* content failed to reach significance ($t = 1.52$, $p = .13$). Several studios were significant factors, but an ANCOVA (analysis of covariance) revealed that the studio itself was not a significant variable ($F = 1.28$, $p = .27$). Post hoc Tukey revealed no significant differences between any studios. The initial significance was because of the higher mean Metacritic ratings for MGM Holdings ($M = 63.4$, $SE = 6.86$), though the large standard error explains the nonsignificant differences between other groups ($M_{range} = 47.8–54.4$, $SD_{range} = 2.67–4.13$). Thus, neither studio nor *relationships* content significantly affect Metacritic scores. However, like audiences, critics rated older movies and lower-budget films more positively.

***Table 4. Linear Regression of Thematic Content and Production Attributes on Reviews.***

| | IMDb | | | | MetaCritic | | | |
|---|---|---|---|---|---|---|---|---|
| Effect | Estimate | *SE* | *t* | *p* | Estimate | *SE* | *t* | *p* |
| Intercept | 5.987 | 0.625 | 9.580 | < .001*** | 49.226 | 15.056 | 3.269 | .001*** |
| *Relationships* content | 0.616 | 0.692 | 0.890 | .37 | 16.659 | 16.199 | 1.028 | .31 |
| Year | 0.017 | 0.007 | 2.312 | .02* | 0.401 | 0.166 | 2.421 | .02* |
| Studio[a] | | | | | | | | |
| NBCUniversal | −0.194 | 0.305 | −0.636 | .53 | −10.980 | 7.511 | −1.462 | .15 |
| Other | −0.555 | 0.337 | −1.649 | .10 | −16.411 | 8.149 | −2.014 | .05* |
| Sony Pictures | −0.266 | 0.294 | −0.904 | .37 | −11.383 | 7.255 | −1.569 | .12 |
| ViacomCBS | −0.552 | 0.304 | −1.820 | .07 | −17.027 | 7.389 | −2.304 | .02* |
| Walt Disney Studios | −0.462 | 0.287 | −1.612 | .11 | −15.348 | 7.066 | −2.172 | .03* |
| WarnerMedia | −0.302 | 0.304 | −0.991 | .32 | −13.900 | 7.444 | −1.867 | .06 |
| Budget[b] | −0.004 | 0.002 | −1.968 | .05* | −0.082 | 0.047 | −1.732 | .09 |
| MPAA rating[c] | | | | | | | | |
| PG-13 | −0.007 | 0.152 | −0.047 | .96 | −2.744 | 3.579 | −0.767 | .44 |
| R | 0.275 | 0.181 | 1.522 | .13 | 6.093 | 4.221 | 1.444 | .15 |

[a] The reference was MGM Holdings. [b] In millions, adjusted for inflation. [c] The reference was PG.

Our final question (RQ3) explored the potential relationship between a romcom's *relationships* content and nomination or winning of awards. We conducted a linear regression comparing all award information (Oscar and Golden Globe wins and nominations), based on *relationships* content, while controlling for production data. The initial model was significant, $F(11, 164) = 2.93$, $p < .001$, adj. $R^2 = .11$. Full details are shown in Table 5. Removing nonsignificant factors resulted in a parsimonious model, controlling for only year, $F(2, 185) = 7.87$, $p < .001$, adj. $R^2 = .07$. Although year ($t = 3.77$, $p < .001$) was a significant variable, *relationships* content was not ($t = 0.31$, $p = .76$). Thus, the increasingly rare nominations and wins from awarding bodies in the romcom genre are unrelated to the *relationships* content in each film.

***Table 5. Linear Regression of Thematic Content and Production Attributes on Industry Awards.***

| Effect | Estimate | *SE* | *t* | *p* |
|---|---|---|---|---|
| Intercept | 1.614 | 2.163 | 0.746 | .46 |
| *Relationships* content | −0.113 | 2.397 | −0.047 | .96 |
| Year | 0.076 | 0.025 | 3.085 | .002** |
| Studio[a] | | | | |
| NBCUniversal | −2.335 | 1.057 | −2.210 | .03* |
| Other | −2.316 | 1.165 | −1.988 | .05* |
| Sony Pictures | −1.532 | 1.018 | −1.505 | .13 |
| ViacomCBS | −2.737 | 1.051 | −2.605 | .01** |
| Walt Disney Studios | −2.338 | 0.993 | −2.354 | .02* |
| WarnerMedia | −1.912 | 1.053 | −1.815 | .07 |
| Budget[b] | 0.002 | 0.007 | 0.271 | .79 |
| MPAA rating[c] | | | | |
| PG-13 | −0.200 | 0.525 | −0.381 | .70 |
| R | 0.796 | 0.625 | 1.273 | .20 |

[a] The reference was MGM Holdings. [b] In millions, adjusted for inflation. [c] The reference was PG.

## Discussion

Our approach used and extended previous computational analyses of the scripts of romantic comedies to demonstrate an impact of latent linguistic features on film success. Prior work (Moore & Ophir, 2022) revealed major themes and topics in romantic comedy, and that illustrated a consistent increase in *relationships* content at the expense of characters' lives outside of their romantic plots. In the current study, we take an institutional approach to media analysis (Turow, 1997) to explore whether this trend in romcom content is attributable to film executives catering to critic and audience expectations (McDonald, 2007) and with the aim to maximize profits (Mortimer, 2010).

Our results reveal that romantic content is indeed a significant factor in opening weekend and lifetime box office earnings, as are higher film budgets. Similarly, theaters are more likely to screen films with more *relationships* content, both during opening weekend and through the lifetime of the theatrical run, which in turn increases earnings (adjusted for inflation). However, *relationships* content was not associated with IMDb or Metacritic scores. Nor was there an association between industry awards and *relationships* content, though we

note that older films and lower-budget productions had higher ratings. Similarly, Oscar and Golden Globe wins and nominations were more prevalent among older films and those with female producers at the helm. These analyses also revealed that higher *relationships* content is associated with newer films, lower-budget films, and stronger MPAA ratings. Thus, while revenue increases as scripts focus on *relationships* content (thanks in part to more exposure from theaters), audiences and awarding bodies prefer productions that take a different perspective and include more depictions of protagonists' lives beyond romance.

Because of the methodological limitations of the pre–big data era, and perhaps because scholars are reluctant to quantify art, not much is known about the systematic antecedents of commercial and critical success of specific pieces (Archer & Jockers, 2016). Most importantly, although studies of film have examined factors such as budget, studio, year of release, number of screening theaters, and MPAA rating (all of which we have controlled in our models), none have accounted for the role of language. Our unique contribution to this area of study is the direct and systematic analysis of text. By including full scripts in our analysis, we present an improved means of predicting financial success in a popular genre and a method that could conceivably be applied to or expanded to include additional genres in future work. As argued by Archer and Jockers (2016), this information has tremendous importance to film industry experts, ranging from investors determining their contributions to theaters selecting films, writers choosing a distributor, and production studios determining which attributes to assign to a film (e.g., release date, genre, MPAA rating).

Importantly, beyond their practical usefulness, our findings also contribute to the development of a theory of film language. An experiment conducted by the author (Moore, Green, Ophir, & Wang, forthcoming) found that increased focus on the language of *relationships* could lead to stronger idealized romantic beliefs. Such beliefs were associated with dissatisfaction in one's own romantic relationships (e.g., Stafford & Merolla, 2007; Tomlinson, Aron, Carmichael, Reis, & Holmes, 2014) and could thus have detrimental effects on audiences. Our findings suggest that the same content audiences often romanticize is financially profitable and is thus likely to continue to dominate the genre. Indeed, prior work (Moore & Ophir, 2022) has demonstrated a consistent increase in the use of such language in romantic comedies. Our findings in this study show that a focus on *relationships* content does not necessarily increase positive reactions to films as manifested in ratings. This may suggest that a studio's decision to increase the focus on *relationships* content is financially motivated, prioritizing audiences' actual behavior over their cognitive and emotional needs. Future studies may further examine these potential discrepancies.

We acknowledge that future studies employing our method would need to consider the idiosyncratic nature of specific genres. It may be that different genres have different factors contributing to their success. One factor pointing to such a discrepancy is popularity among audiences and critics. In their analysis of top box office films, Elberse and Eliashberg (2003) found that critical acclaim was positively related to opening-week revenue but negatively related to opening-week screens. In our study of romcoms (which, as noted, are less likely to receive acclaim), IMDb rating and Metascore were both negatively related to opening weekend revenue *and* to opening screens. Both were also negatively related to lifetime screens, though IMDb score was positively related to lifetime earnings. Metascore had no significant relationship with lifetime earnings. Whether this is an indication that top box office films

are different than others or that the romcom genre is unique in its relationship to reviews could be determined with additional analyses of box office flops and other genres.

In that vein, we also note prior research examining how adherence to genre formulae has been associated with more positive reviews from critics and audiences (Hsu, 2006). Were it the case that audiences preferred romcoms that stay closer to the expectations of the genre, we might expect increased *relationships* content to be associated with more positive ratings. However, neither IMDb score nor Metascore was significantly related to thematic content. This presents another potential area for additional investigation. Although *relationships* content has been noted to contain more tropes and romantic content typical of the genre (Moore & Ophir, 2022), that does not necessarily mean scripts with higher amounts of this thematic content are meeting the expectations for the genre's plotlines, character typologies, or other qualitative content like music and visuals. Future research might explore ways to include these components in a quantitative analysis.

It is important to note two limitations to these findings. First, textual analysis of scripts does not capture the full viewing experience. Additional variables, such as soundtrack, actors, and locations are not captured here or in the original textual analysis (Moore & Ophir, 2022). Second, additional research and experimentation to examine how audiences respond to these thematic shifts, not only in the ratings examined here but in their attitudes and beliefs related to romance. Notably, star power, which we did not include, has previously been shown to increase the total screen number and earnings (based on prior awards, Ericson & Grodman, 2013; previous film revenue, Sharda & Delen, 2006; or influence noted from popular media, De Vany & Walls, 1999). We hope that the analyses provided here will lead to future research that examines the romantic comedy genre from additional angles, as well as the relationships between text and real-world impact of films in general.

A final limitation to note is our consideration of theatrical revenue as a measure of financial earnings in a dynamic media environment that over the years has moved toward relying on alternative revenue options, including streaming platforms and direct-to-streaming releases. We acknowledge many of these movies came out in different entertainment environments before there were more opportunities to watch quickly at home. Still, to account for the wide range of decades, theaters are the most accurate comparison. Notably, the date range examined ends in 2019, before theaters were closing because of the coronavirus pandemic. Importantly, our analysis controlled for year of release in the regressions. This covariate should be seen as accounting, at least in part, for technological changes (VCR, TiVo, streaming, etc.).

Despite these limitations, this is the first study to demonstrate an empirical relationship between films' scripts and real-world success. Specifically, we examined whether the film industry's adherence to predictable plotlines in romantic comedies can be explained by associations with commercial and critical success. These detailed findings may be of interest to investors, studios, and stakeholders such as theaters who are interested in predicting revenue and optimize their marketing strategies. It may also be useful for screenplay writers looking for distributors, as they can further illustrate the value of their scripts. We look forward to continued work in this area to better understand what makes romantic comedies successful.

**References**

Apatow, J., Rothman, R. (Producers), & Stoller, N. (Director). (2012). *The five-year engagement* [Motion picture]. Universal City, CA: Universal Pictures.

Archer, J., & Jockers, M. L. (2016). *The bestseller code: Anatomy of the blockbuster novel*. New York, NY: St. Martin's Press.

Bergman, R. (Producer), & Gordon-Levitt, J. (Director). (2013). *Don Jon* [Motion picture]. Los Angeles, CA: Voltage Pictures.

BoxOfficeMojo. (n.d.). *Top grossing romantic comedy movies at the box office*. Retrieved from https://www.boxofficemojo.com/genres/chart/?view=main&sort=gross&order=DESC&pagenum=5&id=romanticcomedy.htm

Caperello, N., & Migliaccio, T. (2011). Women's interactions with romantic comedies and the impact on their relationships happily ever after. In N. K. Denzin & T. Faust (Eds.), *Studies in symbolic interaction* (Vol. 37, pp. 195–219). Bingly, UK: Emerald Group Publishing.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*(6), 066111. doi:10.1103/PhysRevE.70.066111

Clements, A. (2016, November 16). *What are the chances? Success in the arts in the 21st century*. Los Angeles Review of Books. Retrieved from https://lareviewofbooks.org/article/chances-success-arts-21st-century/

De Vany, A., & Walls, W. D. (1999). Uncertainty in the movie industry: Does star power reduce the terror of the box office? *Journal of Cultural Economics, 23*(4), 285–318. doi:10.1023/A:1007608125988

Elberse, A., & Eliashberg, J. (2003). Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures. *Marketing Science, 22*(3), 329–354. doi:10.1287/mksc.22.3.329.17740

Ericson, J., & Grodman, J. (2013). *A predictor for movie success*. CS229: Machine learning. Stanford University. Retrieved from http://cs229.stanford.edu/proj2013/EricsonGrodman-APredictorForMovieSuccess.pdf

Federal Reserve Bank of Minneapolis. (n.d.). *Consumer price index, 1913*–. Retrieved from https://www.minneapolisfed.org/about-us/monetary-policy/inflation-calculator/consumer-price-index-1913-

Hefner, V., & Wilson, B. J. (2013). From love at first sight to soul mate: The influence of romantic ideals in popular films on young people's beliefs about relationships. *Communication Monographs, 80*(2), 150–175. doi:10.1080/03637751.2013.776697

Hsu, G. (2006). Jacks of all trades and masters of none: Audiences' reactions to spanning genres in feature film production. *Administrative Science Quarterly, 51*(3), 420–450. doi:10.2189/asqu.51.3.420

Joseph, S. E. (2019). *What makes a movie successful: Using analytics to study box office hits*. Chancellor's Honors Program Projects. Retrieved from https://trace.tennessee.edu/utk_chanhonoproj/2252

Joshi, M., Das, D., Gimpel, K., & Smith, N. A. (2010, June). Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 293–296). Los Angeles, CA: Association for Computational Linguistics.

Kaplan, D. (2006). And the Oscar goes to . . . A logistic regression model for predicting Academy Award results. *Journal of Applied Economics & Policy, 25*(1), 23–41.

Krauss, J., Nann, S., Simon, D., Fischbach, K., & Gloor, P. (2008). Predicting movie success and academy awards through sentiment and social network analysis. *16th European Conference on Information Systems, 2008*(1), 2026–2037. Retrieved from http://aisel.aisnet.org/ecis2008/116

Lash, M. T., & Zhao, K. (2016). Early predictions of movie success: The who, what, and when of profitability. *Journal of Management Information Systems, 33*(3), 874–903. doi:10.1080/07421222.2016.1243969

McDonald, T. J. (2007). *Romantic comedy: Boy meets girl meets genre*. New York, NY: Columbia University Press.

Metacritic. (n.d.). *How we create the metascore magic*. Retrieved from https://www.metacritic.com/about-metascores

Milchan, A., Reuther, S. (Producer), & Marshall, G. (Director). (1990). *Pretty woman* [Motion picture]. Burbank, CA: Touchstone Pictures.

Moore, M. M., Green, M. C., Ophir, Y., & Wang, H. (forthcoming). The effects of corrective strategies on romantic belief endorsement. *Communication Research*. doi:10.1177/00936502221138428

Moore, M. M., & Ophir, Y. (2022). Big data, actually: Examining systematic messaging in 188 romantic comedies using unsupervised machine learning. *Psychology of Popular Media, 11*(4), 355–366. doi:10.1037/ppm0000349

Morning Consult & The Hollywood Reporter. (2018). *National tracking poll #181150, November 29–December 02, 2018*. Morning Consult. Retrieved from https://morningconsult.com/wp-content/uploads/2018/12/181150_crosstabs_HOLLYWOOD_REPORTER_FINAL_120418.pdf

Mortimer, C. (2010). *Romantic comedy*. London, UK: Routledge.

New York Film Academy. (2018, July 6). *A guide to the most important film award shows*. Retrieved from https://www.nyfa.edu/student-resources/guide-important-film-award-shows/

Pardoe, I., & Simonton, D. K. (2008). Applying discrete choice models to predict Academy Award winners. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 171*(2), 375–394. doi:10.1111/j.1467-985X.2007.00518.x

Reiner, R. (Producer & Director). (1995). *The American president* [Motion picture]. Beverly Hills, CA: Castle Rock Entertainment.

Reitman, I. (Producer & Director). (1986). *Legal eagles* [Motion picture]. Universal City, CA: Universal Pictures.

Segrin, C., & Nabi, R. L. (2002). Does television viewing cultivate unrealistic expectations about marriage? *Journal of Communication, 52*(2), 247–263. doi:10.1111/j.1460-2466.2002.tb02543.x

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems With Applications, 30*(2), 243–254. doi:10.1016/j.eswa.2005.07.018

Simonton, D. K. (2004). Film awards as indicators of cinematic creativity and achievement: A quantitative comparison of the Oscars and six alternatives. *Creativity Research Journal, 16*(2–3), 163–172. doi:10.1080/10400419.2004.9651450

Stafford, L., & Merolla, A. J. (2007). Idealization, reunions, and stability in long-distance dating relationships. *Journal of Social and Personal Relationships, 24*(1), 37–54. doi:10.1177/0265407507072578

Terry, N., Butler, M., & De'Armond, D. A. (2005). The determinants of domestic box office performance in the motion picture industry. *Southwestern Economic Review, 32*(1), 137–148.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software, 59*(5), 1–38. doi:10.18637/jss.v059.i05

Tomlinson, J., Aron, A., Carmichael, C. L., Reis, H. T., & Holmes, J. G. (2014). The costs of being put on a pedestal: Effects of feeling over-idealized. *Journal of Social and Personal Relationships, 31*(3), 384–409. doi:10.1177/0265407513498656

Turow, J. (1997). *Media systems in society: Understanding industries, strategies, and power*. New York, NY: Longman Publication Group.

Walls, W. D. (2005). Modeling movie success when "nobody knows anything": Conditional stable-distribution analysis of film returns. *Journal of Cultural Economics, 29*(3), 177–190. doi:10.1007/s10824-005-1156-5

Walter, D., & Ophir, Y. (2019). News frame analysis: An inductive mixed-method computational approach. *Communication Methods and Measures, 13*(4), 248–266. doi:10.1080/19312458.2019.1639145