Atlantic Marketing Journal

Volume 5 | Number 3 Article 9

February 2017

The Cross-Platform Consistency of Online User Movie Ratings

Dan Baugher
Pace University, Lubin School of Business, dmbaugher@aol.com

Chris Ramos
Pace University, Lubin School of Business, cdcramos@aol.com

The Cross-Platform Consistency of Online User Movie Ratings

Cover Page Footnote
A prior version of this paper was presented at the 2016 meeting of the American Society of Business and Behavioral Sciences (ASBBS) and was also published in the 2016 ASBBS Conference Proceedings.

The Cross-Platform Consistency of Online User Movie Ratings

Dan Baugher, Pace University

dbaugher@pace.edu

Chris Ramos, Pace University

cramos@pace.edu

Abstract - Online movie ratings are often used in studies of Internet word-of-mouth, but the cross-platform consistency of such ratings has not been well established. Mean user ratings were obtained for 500 movies using the rating systems in place on Netflix and IMDb. Mean volume was 1,480,931 (SD = 2,636,388) and 104,746 (SD = 193,532) on the Netflix and IMDb sites, respectively, suggesting the movies were visible, but also variable in visibility. Volume was moderately correlated (r = .724, $p \leq .001$). Mean user ratings on the Netflix and IMDB sites averaged 3.57 (SD = .37) and 3.84 (SD = .34). Mean ratings were moderately consistent (r = .600, $p \leq .001$). Mean DVD ratings on the Amazon platform were also considered and showed only modest correlations of .153 ($p \leq 001$) and .167 ($p \leq 001$) with ratings on Netflix and IMDB, respectively.

Keywords - Word-of-Mouth Advertising, eWOM, Online Product Ratings, Rating Valence, Rating Volume, Movie Ratings, Validity, Reliability

Relevance to Marketing Educators, Researchers and/or Practitioners – The paper shows the validity and cross-platform consistency of online mean product ratings (valence) should not be assumed. Mean movie ratings across Netflix and IMDb appear to be valid measures of movie quality, but showed only moderate reliability suggesting some attenuation of their true correlation with criteria due to unreliability is likely. Given its correlation across sites, rating volume may serve as a proxy for movie visibility. However, Amazon ratings did not correlate well with Netflix and IMDB ratings and may not be a valid measure of movie quality.

Introduction

The substantial growth of the Internet has had a tremendous impact on consumer buying behavior and competition. Products that once were sold in a local competitive

ISSN: 2165-3879 (print), 2165-3887 (electronic)

landscape now must compete globally. This can readily be seen for used and rare books (Raugust, April 12, 1999). Rare books are no longer solely sold in small Antiquarian book stores. Buyers can now purchase rare books from such global sites as Alibris (http://www.alibris.com/), Biblio (http://www.biblio.com), and AbeBooks (https://www.abebooks.com). DVDs are another product that has moved almost completely from local distribution to online sale or rental. Few products have escaped the shift to sales online. As shipping improves and firms adopt favorable return policies, it is likely availability of products online will continue to increase.

With the move to online availability, sources of information on products have also changed. While salespeople had their biases, they were a common source of information in local stores. Word-of-mouth largely took the form of friends or family who purchased a product. With the Internet, word-of-mouth advertising has often replaced friends and family. Online product reviews are available for a wide range of products including hotels, restaurants, books, games, and computers (Zhang et al., 2010). Ratings and written reviews are often considered electronic word-of-mouth or eWOM (Hennig-Thurau et al., 2004) and have been found to be a major source of information for making purchases by online consumers (Godes and Mayzlin, 2004; Liu, 2006).

As a result, users no longer need to wait for an expert review or an issue of *Consumer Reports* to determine the quality of products they wish to purchase. The recommendations of online peer consumers are readily available (Dean and Biswas, 2001; Floyd et al., 2014; Riegner, 2007). In fact, such feedback by peers is a constant on websites in the digital marketplace (Wang, 2005; Ziegle and Weber, 2015). Some research suggests that peer feedback may even have a greater impact on purchase decisions than experts (Gilly et al., 1998).

When it comes to word-of-mouth advertising, movies are one product where a shortage of opinions has never been a problem. Critics have become well known for their views (e.g., Siskel and Ebert). Friends and family have also always been ready with an opinion. The main difference is peer opinions are far easier to access than in the past. The five-star rating system used for many products (Koehn, 2003) is readily available on many Internet sites involved with movies, though some sites, such as IMDb, use a ten-point system.

Availability of this huge reservoir of ratings has not been lost on researchers. The number of studies making use of viewer ratings has increased considerably over the past decade. One common measure has been the mean viewer rating for movies, often referred to as valence. It has been related to such criteria as box office revenue, though the results have not been entirely consistent. Some have shown valence to predict sales (Baugher, Noh & Ramos, in press, 2017; Dellarocas and Zhang, 2007; Gruhl et al., 2005) and others have found no relationship between valence and sales (Duan et al., 2008; Chintagunta, Gopinath & Venkataraman, 2010).

This inconsistency in results may be due to differences in the movies considered. Purnawirawan et al. (2015) found the impact of valence is moderated by type of

product and brand. However, it is also possible there are problems with the measures of quality used. Some have expressed concerns about the trustworthiness of online data (Miller, 2001). With the data so easily accessible, it is easy to ignore its limitations. Many of the concerns of traditional survey sampling remain in the use of online ratings including problems with rating scales, response accuracy, and missing ratings from those who refuse to rate or have no access to a rating site.

Also, online user ratings are not collected in controlled settings and the rating scales used can sometimes be weak. None have been developed for the purpose of academic research. These weaknesses can foster random measurement error and result in rating inconsistency between database sources.

While there are many possible sources of error for inter-rater reliability across online platforms including differing rater standards and different rating scales, it is beneficial to begin by thinking about what correlation might be useful. Typically, measures are considered to have an acceptable level of reliability when they reach a .80 correlation (Nunnally, 1978). To the extent that reliability is much lower than .80, it may attenuate the correlation between such ratings and criteria resulting in lost power to discover relationships (Ghiselli, Campbell & Zedeck, 1981; Nunnally, 1978; Traub, 1994). Lower reliability can also call into question the validity of ratings as reliability is a precursor to achieving strong validity (Kline, 1993; Rozeboom, 1966).

Literature Review

Concerns over the reliability of movie ratings has resulted in laboratory research on reliability. Cosley et al. (2003) used a rate-rerate procedure for 40 randomly selected movies which received middle ratings (2, 3, 4) on a 5-point rating scale. The movies were rated by 212 study participants where the initial rating of participants was made months or even years before the study. They reported a test-retest reliability of .70. Typically, a .90 reliability coefficient is considered good for test-retest reliability (Murphy and Davidshofer, 1996) though the purpose of a measure can result in different levels of acceptable reliability (Van Ness, Towle & Juthani-Mehta, 2008).

Amatriain, Pujol & Oliver (2009) conducted a study over three time periods to better assess the overall reliability and stability of movie ratings. In their study, 118 users rated 100 movies from the Netflix Prize database using a 5-point rating scale. They found a reliability of .90 for the shorter timeframe of one day and .88 for the longer timeframe of at least 15 days. They also found extreme ratings showed greater consistency over time than mild opinions. At the same time, they did not find users to be especially consistent in rating whether they had seen a movie or not, despite consistent ratings once they decided to make a rating, bringing into question the validity of their ratings.

While these studies suggest the test-retest reliability of viewer movie ratings is acceptable, differences can exist across rating site databases. It is one thing for a user

to consistently give high or low ratings to the same movies over time and quite another for two users on different sites to agree the same movies deserved comparable high and low ratings. Also, it is the reliability of the predictor that matters (Tett, Jackson & Rothstein, 1991). If mean ratings are used, it is their reliability that is of the greater concern. The reliability of individual ratings does not assure the consistency of mean ratings across sites.

Boor (1990) addressed the issue of rater consistency for movies by examining critic ratings from the *Video Movie Guide* by Leonard Maltin. Boor (1990) found a correlation of .67 for critic ratings of 3,144 movies and considered the ratings to be of sufficient rater agreement for viewers to make good use of them though they did not reach the high reliability for test-retest found for users by Amatriain, Pujol & Oliver (2009). However, Agresti and Winner (1997) did not find strong agreement between the movie critics, Siskel and Ebert. Across 160 movies, they found Cohen's kappa for their agreement was .39, only moderate at best. When they compared other pairs of critics, agreement was worse with the next best pair yielding a kappa of .28.

Plucker et al. (2009) conducted an interesting study of the consistency of movie ratings, using 169 student raters, online raters (referred to as "self-defined" novice critics) and critics for 680 films though the number of films rated by students varied as a function of whether they viewed the film. Critic data came from metatcritic.com. Novice critic data came from boxofficemojo.com and IMDb.com. They found a low-moderate correlation of .43 between mean ratings from students and critics.

In contrast, mean ratings for novice critics from IMDb and boxofficemojo.com in their study showed quite high consistency with a correlation of .86. The correlation between students and mean ratings for novice critics was moderate at .65 across both novice sites as was the correlation between the mean ratings for novice critics and critics at .72 to .77 for the two sites. They also found the correlation between students, novice critics, and critics dropped considerably as a function of student experience. The mean ratings of students who saw far fewer movies showed a low correlations of .22 with mean ratings for critics and .30 or less with mean ratings for novice critics.

Hon (2014) conducted a study of critics and users from the Yahoo! and the metacritic.com sites for 243 popular movies released between 2008 and 2009 and found the correlation between mean critic and mean user ratings from Yahoo! was low at .413, but the correlation between mean critic ratings was quite high at .92. The correlation between Yahoo user and Metacritic user scores was not provided, though noted as significant at the 1% level. The number of online users making ratings in Hon's study from these sites was upward of 30,000 to 40,000 users though relatively low by comparison to the volume of ratings typical for the Netflix and IMDb platforms. This lower volume may be due to the relative newness of the movies considered in the study as they were all recent releases.

Thus, it appears movie ratings can vary considerably in their consistency across raters depending on rater and approach. In the Agresti and Winner (1997) study, pairs of critics did not show highly consistent ratings with a Kappa of .39, but Hon

(2014) found the correlation between mean critic ratings was quite high at .92 and Plucker et al. (2009) found it to be .86. Clearly, mean critic ratings are more consistent than those from individual pairs of critics. Similarly, Hon (2014) found the correlation between mean critic and mean user ratings to be low at .413, but the Plucker et al. (2009) study reported the correlation between mean user ratings across multiple platforms ranged between .72 and .77. However, Plucker et al. (2009), also found inter-rater reliability for mean ratings across platforms went as low as .30 and .22, depending on raters.

Test-retest reliability for individual raters seems less variable and less likely to go to lower levels than estimates of inter-rater reliability. Test-retest reliability was quite high at .88 and .90 for student movie ratings in the Amatriain, Pujol & Oliver (2009) study though possibly over-estimated by the very short time interval between ratings. The test-retest reliability of user movie ratings in the Cosley et al. (2003) study was lower at .70, though the long timeframe between the ratings may have resulted in real changes in views and an underestimate of reliability. In both cases, test-retest reliability did not go lower than .70.

Study Purpose

The primary goal of this study is to ascertain whether mean user ratings of movie quality are consistent across two different, but frequently used platforms in research. To this end, it examines the association between mean user ratings of movie quality (valence) on the IMDb and Netflix sites. It also examines the correlation between rating volume across the two sites. Secondarily, it assesses the consistency of such ratings with those made by buyers of the same movie on Amazon, who unlike raters on the two movie sites, made a monetary investment in the specific product.

It is hoped the study will shed light on the somewhat contradictory results found for consistency or inter-rater reliability of user movie ratings in other studies. The Netflix site was selected because it is one of the most popular benchmarks used in the Recommender Systems (RS) literature (Amatriain, Pujol & Oliver, 2009) and heavily used by researchers. The IMDb database is also used by researchers though its ratings tend to be the result of far less rating volume than those on Netflix. Of all sites, none provides the high rating volume of the Netflix site where it is often in the millions. Also, it is expected the study will show the extent to which researchers can assume consistency and validity for user movie ratings across different platforms.

Hypotheses

The research on reliability and rating consistency for movies covered in the literature review suggests user ratings across sites should be correlated. However, we do not expect the correlation to be any greater than moderate, and nothing near the .86 found by Plucker et al. (2009) across sites. While test-retest reliability is not the same as consistency, the range of .40 to .75 suggested by Fleiss (1986) as "fair to good"

reliability is instructive in this setting. It is expected that the correlation between mean ratings or valence across the two movie sites will fall in this range, not exceed it.

In part, our view that the correlation will be lower is a result of the Plucker et al. (2009) study. They found lower consistency between the mean ratings of students and users of movie rating sites than between users of the movie sites. We believe those using Netflix are more akin to students in their study who did not show as much consistency with IMDb and boxofficemojo.com users as site users, possibly because students were less experienced raters. The correlation they found between mean ratings by students and site users was .65, but it declined to .30 or less for students with less experience.

While it is expected there will be a correlation between mean user ratings of movie quality on Netflix and IMDb and those on Amazon, a lower correlation is not unexpected. All three sites are impacted by movie quality, but those on Amazon are also affected by the technical quality of the physical product as well as product delivery through the postal system or United Parcel Service (UPS).

Rating volume was considered across the movie sites because it often serves as an indicator of movie visibility and we wanted a wide range of visibility. For example, Dellarocas and Zhang (2007) found volume had a positive, significant impact on future national box office movie performance and increased visibility. Duan, Gu & Whinston (2008) suggest that volume builds as a result of prior sales success in movie attendance. We've not found any study showing the correlation between volume across such sites. Given that visibility should be comparable across users, we are comfortable in hypothesizing that volume will correlate across the two movie sites.

Three hypotheses were tested. In Hypothesis 1, we predicted that a positive correlation would exist between mean user ratings of movie quality (valence) across the movie sites. In Hypothesis 2, we predicted that a positive correlation would exist between user rating volume across the movie sites. In Hypothesis 3, we predicted that a positive correlation would exist between mean user ratings of product quality on the Amazon site (valence) and mean user ratings of movie quality (valence) on the Netflix and IMDb sites. However, we do not anticipate the correlations with the Amazon site will be as great as those for the movie sites because Amazon ratings involve more than movie quality.

The following formal hypotheses were tested:

H1: Mean user ratings of movie quality (valence) will be positively correlated across the two movie rating sites.

H2: Mean user volume will be positively correlated across the two movie rating sites.

H3: Mean user ratings of movie quality (valence) from the two movie rating sites will be positively correlated with mean user ratings of product quality

(valence) from the Amazon site.

Method

User rating were obtained from the Netflix, IMDb, and Amazon U.S. sites. Specifically, they were acquired from: http://movies.netflix.com/WiHome, http://www.imdb.com, and http://www.amazon.com.

IMDb ratings are made on a 10-point scale while those from Netflix and Amazon are made on a five-point scale. For ease of comparison, IMDb mean ratings were converted to a five-point range by re-anchoring to a range of one to five through the transformation of (.44*IMDb mean ratings) + .56) (IBM, 2010).

A sample of 500 movies was selected for rating comparison to assure a wide range of movie quality, movie type, and visibility. The process began with a search of DVD listings on amazon.com from its decade by decade listing. Blockbusters and other movies from all traditional movie genres (Mckenzie, 2010) were selected including drama, horror-sci-fi, comedy, mystery-thriller, action/adventure, and musical. To increase the diversity of the mix, foreign films from the Criterion Collection, gay/lesbian films, and plays, most of which were dramas, were added to the original mix. The Criterion Collection offers classic films from across the world.

The resulting mix included blockbusters in drama, horror-sci-fi, comedy, mystery-thriller, and action-adventure. It also included musicals and art films with far less box appeal as well as foreign films and plays.

Volume of user ratings was also assessed for each movie to ascertain its market visibility. It was obtained from each respective site.

Findings

Table 1 provides summary data. The movies covered eight genre. Drama dominated the list, followed by horror/sci-fi and foreign films. A one-way-ANOVA showed no significant difference by genre for Netflix ratings ($F = 1.42$, $df = 7.492$, ns). IMDb ratings showed a significant difference by genre ($F = 7.29$, $df = 7, 492$, $p \leq .001$). Post hoc analysis showed the source of this difference was lower ratings for horror-sci-fi and comedy and higher ratings for drama and foreign films, with four of the comparisons of these to the other seven genre showing significant differences ($p \leq .05$). As noted earlier, foreign films were mainly drama.

The average valence for Netflix ratings was 3.57 (SD = .37) with a range of 1.90 to 4.50. For IMDb, the average valence was 3.84 (SD = .34) with a range of 2.19 to 4.65. As expected, volume was considerably different across the Netflix and IMDb platforms. The average volume for Netflix was 1,480,931 while that for IMDb was 104,746. The mix of movies also resulted in considerable variability in rating volume for both sites with a standard deviation of 2,636,388 and 193,532 for Netflix and

IMDb, respectively. The average valence for Amazon ratings was 4.29 (SD = 1.83) ranging from 1.80 to 5.00. Not surprisingly, far fewer users rated movies on the Amazon site with an average volume of 306 (SD = 405).

Table 1: Summary Data for Movie List (n=500)

Variable	Summary Statistics			
DVD Genre	Drama (228-46%), Horror-Sci-Fi (76-15%), Foreign (51-10%), Comedy (47-9%), Mystery/Thriller (36-7%), Musical (26-5%), Action/Adventure (26-5%), Gay/Lesbian (10-2%)			
	Mean	SD	Minimum	Maximum
Valence Netflix	3.57	.37	1.90 *The Incredible Melting Man*	4.50 *The Shawshank Redemption*
Valence IMDb	3.84	.34	2.19 *The Incredible Melting Man*	4.65 *The Shawshank Redemption*
Valence Amazon	4.29	1.83	1.80 *Dragon Country* (T. Williams Play)	5.00 *Scenes from a Marriage* (Bergman)
Volume Netflix	1,480,931	2,636,388	100 *Ah Wilderness* (O'Neill Play)	19,478,084 *Titanic*
Volume IMDb	104,746	193,532	18 *The Merry Wives of Windsor* (Shakespeare Play)	1,517,052 *The Shawshank Redemption*
Volume Amazon	306	405	2 *Beyond the Horizon* (O'Neill Play)	4,466 *Star Trek: Into Darkness*

H1 was tested by finding the correlation between mean ratings on the Netflix and IMDb sites. The zero-order product moment correlation is the preferred statistic for studying reliability and validity (Tett, Jackson & Rothstein, 1991). The results show support for H1. The Pearson correlation between the two sets of ratings was .600 ($p \leq .001$). The results suggest a moderate degree of consistency falling as we predicted in the "fair to good" range of .40 to .75 proposed for reliability by Fleiss (1986).

H2 was tested by finding the correlation between mean volume on the Netflix and IMDb sites. The results support H2. The Pearson correlation between volume on the Netflix and IMDb sites was .724 ($p \leq 001$), though volume differed considerably across the two sites.

H3 was tested by finding the correlation between mean ratings on the Amazon

site with those on the Netflix and IMDb sites. The results support H3, though showing modest correlations as expected. The Pearson correlation between mean ratings of DVD quality on Amazon and mean ratings of movie quality on Netflix and IMDb were .153 ($p \leq 001$) and .167 ($p \leq 001$), respectively.

Discussion

This study focused on determining the consistency of mean user ratings for movies available on DVDs across two movie rating platforms: Netflix and IMDb. While studies have focused on the agreement of movie critics and the correlation between users and movie critics, we know of only one study that provided a measure of the consistency of ratings across different user platforms (Plucker et al., 2009) though the volume of ratings on which the ratings were based was unclear.

We know of no study that has examined the degree to which Netflix ratings, a heavily used source of movie ratings by investigators, are related to those from another movie site. While some studies have examined the reliability of individual raters (Amatriain, Pujol & Oliver, 2009; Cosley et al., 2003; Murphy and Davidshofer, 1996; Plucker et al., 2009), it is the reliability of the mean ratings that matter when they are used as the predictor.

Since reliability places an upper bound on the correlation of a measure with criterion variables, we believe examining the consistency of mean user ratings across platforms is one approach for determining their reliability as measures and thus, understanding the extent to which lower reliability might affect validity and possibly weaken the correlation of the measures with such criteria as sales. To this end, a very diverse range of movies was considered.

The mean ratings were based in some instances on millions of users though volume varied from site to site. Ratings on Netflix were often based on hundreds of thousands of ratings. By contrast, IMDb ratings were based on far fewer users, about one-tenth as many, on average. Netflix ratings are made by a wide range of users as a part of the Netflix system, often after viewing a movie. IMDb ratings require more initiative. However, volume across the two sites correlated moderately at .724, suggesting it might serve as a useful indicator of movie visibility.

Despite these large differences in volume, users, how users made ratings, and rating scales used, the results showed a moderate degree of consistency for mean ratings of movies or valence across the two movie sites with a correlation of .600. Since Netflix is heavily used in studies, it is good to see ratings on the two sites correlate moderately well with each other. The meaning of results from studies using Netflix or IMDb ratings would be far more difficult to interpret without such cross-platform consistency.

That said, the reliability uncovered is some distance from the minimum standard for reliability of .80 suggested by some (Nunnally, 1978) though in the middle of the

"fair to good" range for inter-rater reliability suggested by others (Fleiss, 1986). While it will depend on the reliability of the criteria, some attenuation of the true relationship of these measures of movie quality with criteria is likely given their reliability is well below .80 (Osborne, 2003). However, few studies correct for attenuation and such underestimation is not uncommon in the behavioral sciences where variables are often difficult to measure (Pedhazur, 1997).

The much weaker relationship found between Amazon ratings and those on the Netflix and IMDb platforms raises a caution about quickly generalizing the ratings of users on one platform to those on another platform who may appear to be rating the same thing. While we expected the correlation between ratings from the movie sites and the Amazon site to be lower, we didn't expect them to be as low as uncovered. At .153 and .167, they were modest at best. This result points to issues of validity that can easily cross platforms in any study. Differences in users and rating goals can change what is being measured. In the case of Amazon ratings, it doesn't appear they would be a valid measure of movie quality.

Also, Amazon ratings were made by far fewer users than on the Netflix and IMDb sites, a very distant third in volume from IMDb. The problem of insufficient raters for representation is known to occur for movie ratings (Zhou and Lange, 2009) and could play a role in the low consistency of ratings from Amazon with those on the two movie sites. It may be the Amazon ratings are not as reliable as necessary and its users don't reflect perceptions of movie quality in the broader marketplace. Of course, rating a physical product, like a DVD, is more complex than rating movies and could be another source of the inconsistency.

While a diverse mix of movies was selected, the mix was not a random selection of all movies and thus the results are limited by the movies considered. However, a key benefit of the mix is the inclusion of movies with staying power in the marketplace and a list of movies that would be easily recognized by anyone with even a passing interest in the genres and includes a full range of movies from blockbusters to art films.

Another area of concern is whether the ratings were made by consumers who intentionally manipulated their ratings for personal gain (Jindal and Liu, 2008; Lui, 2010) which might increase reliability, but reduce validity. While this could be an issue on the IMDb platform and even more so on the Amazon platform, the sheer volume of ratings on the Netflix platform makes it an unlikely problem for that site. Of course, there is always the possibility of self-selection in online reviews which can bias their representativeness (Li and Hitt, 2008).

It is also possible some of the consistency in ratings uncovered is a result of ratings being influenced by previously posted ratings, creating a herding effect. (Moe and Trusov, 2011). While this might occur within a site and cause individual ratings to be alike there, it is less likely that one site will cause the same herding effect on a different site.

The possibility of valence differences across genre was suggested in this study by the lower ratings found on the IMDb site for horror-sci-fi and comedy as compared to drama which showed higher ratings. This may suggest different standards or differences in movie quality by genre worthy of further investigation. An investigation of whether genre moderates cross-platform consistency would be an interesting study.

De Langhe, Fernback & Lichtenstein (2014) pose another potential source of rating consistency that can occur due to factors other than validity. They found consumer ratings of technical quality were reliable, but not very diagnostic of quality and thus, not especially valid. They suggest this can occur when consumer ratings are more the result of marketing actions on users' ratings than quality. When they controlled for quality ratings from *Consumer Reports*, user ratings were higher for more expensive products and brands with reputations linked to emotional benefits.

Such marketing influences on rater consistency are not as clear-cut when it comes to movies though movie stars might be thought of as providing a brand and there are a few visible distributors/producers such as Disney, New Lion Cinema, Lionsgate, and Sony. Price differences exist in the sale of DVDs where the Criterion brand offers higher quality at a higher price, but are more often associated with how the movie is packaged (e.g., DVD versus Blu-ray) than the movie itself. While there are no objective sources of movie quality such as *Consumer Reports*, it might be possible to make a similar assessment for movies by finding the partial correlation between mean ratings across platforms when brand or star power are controlled.

Conclusions

We conclude that mean user ratings of movie quality (or valence) on the Netflix and IMDb platforms are moderately reliable and valid measures of movie quality. Investigators can also expect volume to show a moderately strong correlation across movie sites suggesting it may be a useful measure of visibility.

Generally, the reliability of a measure will not be less than its correlation with a criterion, but as it lessens from 1.0, the observed correlation may be attenuated due to errors of measurement (Traub, 1994). With a correlation of .600 across the sites, well below the normally expected reliability of .80, some attenuation of the true correlation of these measures of movie quality with criteria due to unreliability is likely. Though this is not uncommon in behavioral research and there is no way for an investigator to improve the ratings, a higher level of reliability for these measures would give greater comfort in their ability to correlate with criteria.

As expected, the results demonstrate it is unwise to assume mean quality ratings across platforms are valid for the same purpose even if at first glance they appear to be measuring the same thing; one reason research of this sort is valuable. While it might seem ratings of movies on Amazon should reflect movie quality, this study

shows their relationship with ratings from well-established, movie platforms is weak suggesting they are not likely valid for this purpose.

References

Agresti, A., & Winner, L. (1997) Evaluating agreement and disagreement among movie reviewers. *Chance: New Directions for Statistics and Computing*, 10 (2), 10-14.

Amatriain, X., Pujol, J., & Oliver, N. (2009) I like it. I like it not: Evaluating user ratings noise in recommender systems. *UMAP '09 Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization: formerly UM and AH*, June, Trento, Italy, 247-258. https://xamat.github.io//pubs/xamatriain_umap09.pdf [Accessed 4th January 2016]

Baugher, D., Noh, S., & Ramos, C. (in press) (2017) The relationship of online Netflix user reviews to days to sale for new DVDs on Amazon. *Academy of Marketing Studies Journal.*

Boor, M. (1990) Reliability of ratings of movies by professional movie critics. *Psychological Reports,* 67 (1), 243-257.

Cosley, D., Lam, S.K., Albert, J., Konstan, J.A., & Riedl, J. (2003) Is seeing believing? How recommender system interfaces affect users' opinions. *The CHI 2003 New Horizons Conference Proceedings: Conference on Human Factors in Computing System,* April, Ft. Lauderdale, 585-592. http://files.grouplens.org/papers/conform-chi03.pdf [Accessed 4th January 2016].

Chintagunta, P.K., Gopinath, S., & Venkataraman, S. (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, 29 (5), 944-957.

De Langhe, B., Fernbach, P., & Lichtenstein, D. (2014) Navigating by the stars: What do online user ratings reveal about product quality? In: Cotte, J. and Wood, S. (eds.) *Advances in Consumer Research*, Volume 42, Duluth, MN, 453-453.

Dean, D.H., & Biswas, A. (2001) Third-party organization endorsement of products: An advertising cue affecting consumer pre-purchase evaluation of goods and services. *Journal of Advertising*, 30 (4), 41-57.

Dellarocas, C., & Zhang, X.M. (2007) Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, 21 (4), 23-45.

Duan, W., Gu, B., & Whinston, A. B. (2008) The dynamics of online word-of-mouth and product sales--an empirical investigation of the movie industry. *Journal of Retailing*, 84 (2), 233-242.

Fleiss, J.L. (1986) *The Design and Analysis of Clinical Experiments*. New York, NY: Wiley & Sons.

Floyd, K., Freling, R., Alhoqail, S., Cho, H., & Freling, T. (2014) How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing*, 90 (2), 217-232.

Ghiselli, E.E., Campbell, J., & Zedeck, S. (1981) *Measurement Theory for the Behavioral Sciences*. San Francisco, CA: Freeman.

Gilly, M.C., Graham, J.L., Wolfinbarger, M.F., & Yale, L.J. (1998) A dyadic study of interpersonal information search. *Journal of the Academy of Marketing Science*, 26 (2), 83-100.

Godes, D., & Mayzlin, D. (2004) Using online conversations to study word-of-mouth communication. *Marketing Science,* 23 (4), 545-560.

Gruhl, D., Guha R., Kumar, R., Novak, J., & Tomkins, A. (2005) The predictive power of online chatter. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, August, Chicago,78-87. http://dl.acm.org/citation.cfm?id=1081883 [Accessed 4th January 2016].

Hennig-Thurau, F., Gwinner, K., Walsh, G., & Gremler, D. (2004) Electronic word of mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? *Journal of Interactive Marketing*, 18 (1), 38-52.

Hon, L.Y. (2014) Expert versus audience's opinions at the movies: Evidence from the North-American box office. *Marketing Bulletin,* 25, article 1, 1-22. http://marketing-bulletin.massey.ac.nz/V25/MB_V25_A1_Hon_FINAL.pdf [Accessed 4th January 2016].

IBM, 2010, August 23 Transforming different Likert scales to a common scale. http://www-01.ibm.com/support/docview.wss?uid=swg21482329 [Accessed 4th January 2016].

Jindal, N., & Liu, B. (2008) Opinion spam and analysis. *WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining,* February, Palo Alto, 219-230. https://pdfs.semanticscholar.org/16f2/deb863ef6d3d6f432de12a2e81149ab03e5a.pdf [Accessed 4th January 2016].

Kline, P. (1993) *The Handbook of Psychological Testing*. New York, NY: Routledge.

Koehn, D. (2003) The nature and conditions for online trust. *Journal of Business Ethics,* 43 (1), 3-19.

Li, X., & Hitt, L. (2008) Self-selection and information role of online product reviews. *Information Systems Research,* 19 (4), 456-474.

Liu, Y. (2006) Word of mouth for movies: Its dynamics and impact on box office revenues. *Journal of Marketing,* 70 (3), 74-89.

Lui, B. (2010) Sentiment analysis and subjectivity. In: Indurkhya, N. and Damerau F. J. (eds.) *Handbook of Natural Language Processing*. Boca Raton, FL: Chapman & Hall/CRC Press.

Mckenzie, J. (2010) How do theatrical box office revenues affect DVD retail sales? Australian empirical evidence. *Journal of Cultural Economics*, 34 (3), 159-179.

Miller, T. W. (2001) Can we trust the data of online research? *Marketing Research*, 13 (2), 26-32.

Moe, W., & Trusov, M. (2011) The value of social dynamics in online product ratings forums. *Journal of Marketing Research,* 49 (June), 444-456.

Murphy, K.R., & Davidshofer, C.O. (1996) *Psychological Testing: Principles and Applications (4th edition).* Boston, MA: Addison-Wesley.

Nunnally, J.C. (1978) *Psychometric Theory (2nd edition).* New York, NY: McGraw-Hill.

Osborne, J. (2003) Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from Educational Psychology. *Practical Assessment, Research & Evaluation,* 8 (11), 1-7. http://pareonline.net/getvn.asp?v=8&n=11 [Accessed 2nd December 2016).

Pedhazur, E. J. (1997) *Multiple Regression in Behavioral Research (3rd edition).* Orlando, FL: Harcourt Brace.

Plucker, J.A., Kaufman, J.C., Temple, J.S., & Qian, M. (2009) Do experts and novices evaluate movies the same way? *Psychology & Marketing*, 26 (5), 470-478.

Purnawirawan, N., Eisend, M., De Pelsmacker, P., & Dens, N. (2015) A meta-analytic investigation of the role of valence in online reviews. *Journal of Interactive Marketing,* 31 (August), 17-27.

Raugust, K. (1999, April 12) Used and rare books go online. *Publishers Weekly*, 245 (15), 22-25. http://www.publishersweekly.com/pw/print/19990412/23257-used-and-rare-books-go-online.html [Accessed 4th January 2016].

Riegner, C. (2007) Word of mouth on the web: The impact of Web 2.0 on consumer purchase decisions. *Journal of Advertising Research,* 47 (4), 436-447.

Rozeboom, W. W. (1966) *Foundations of the Theory of Prediction*. Homewood, IL: Dorsey Press.

Tett, R.D., Jackson, D., & Rothstein, M. (1991) Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, 44 (4), 730-742.

Traub, R. (1994) *Reliability for the Social Sciences: Theory and Applications (Vol. 3).* Thousand Oaks, CA: Sage.

Van Ness, P., Towle, V., & Juthani-Mehta, M. (2008) Testing measurement reliability in older populations: Methods for informed discrimination in instrument selection and application. *Journal of Aging Health*, 20 (2), 183-197.

Wang, A. (2005) The effects of expert and consumer endorsements on audience response. *Journal of Advertising Research*, 45 (4), 402-412.

Zhang, Z., Quiang Y., Law, R., & Yijun, L. (2010) The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management*, 29 (4), 694-700.

Zhou, H. & Lange, K. (2009) Rating movies and rating the raters who rate them. *American Statistician*, 63 (4), 297-307.

Ziegle, M., & Weber, M. (2015) Example, please! Comparing the effects of single customer reviews and aggregate review scores on online shoppers' product evaluations. *Journal of Consumer Behaviour*, 14 (2), 103-114.

Author Information

Dan Baugher has a Ph.D. in psychology from Rutgers University and is Professor of Management and Associate Dean and Director, Graduate Programs in the Lubin School of Business at Pace University. He has published in numerous journals and proceedings including the Decision Sciences Journal of Innovative Education, Journal of Research in Personality, and Personnel Review.

Chris Ramos has a BFA from the University of Hawaii in dance and a MPA from Pace University in not-for-profit management. He is a Clinical Assistant Professor of Management and Executive Director of the Arts & Entertainment Management Program in the Lubin School of Business at Pace University. He has published in the Academy of Marketing Studies Journal and Personnel Review and for twenty years was artistic director for his own modern dance company in New York City.