Strikingly Low Agreement in the Appraisal of Motion Pictures

Pascal Wallisch and Jake Alden Whritner

Abstract: Neuroimaging research suggests that watching a movie synchronizes brain activity between observers. This is surprising in light of anecdotal reports that viewers construct their experience radically differently, consistent with contemporary cognitive media theory. This article empirically tests the degree of agreement in the appraisal of commercially produced major motion pictures. Ratings for more than two hundred carefully selected movies were solicited from a diverse pool of more than three thousand study participants. Doing so shows that intersubjective movie appraisal is strikingly low but significantly different from zero. The article also shows that these ratings correlate only weakly with the judgment of professional movie critics. Taken together, this study supports the notion that movies are an extremely rich, highly dimensional narrative stimulus with many degrees of freedom for viewers to construct their subjective experience in a highly idiosyncratic fashion.

Keywords: appraisal, cognitive media theory, consensus, expert judgment, major motion pictures, psychocinematics

Background

Over the past hundred years, filmmakers have developed a rich tool kit that allows them to tightly control viewers' attention. This toolbox comprises what is commonly referred to as the continuity editing style, which includes techniques such as matches on action, shot reverse shot, point-of-view editing, and the 180-degree rule (Bordwell and Thompson 2010). In addition, variable framing helps to create further hierarchies within the film image and provides salient cues for viewers to attend to certain aspects of the frame such as faces or narratively significant objects (Seeley and Carroll 2014). As Tim J. Smith (2012) argues in "The Attentional Theory of Cinematic Continuity," the conventions of continuity editing take advantage of natural attentional cues to guide viewer attention and maintain their interest. These techniques establish and sustain engagement of the viewer for long periods of time—typically for well over an hour. We know that these techniques work because filmmakers successfully guide the attention and eye movements (Smith et al. 2012) across

Volume 11, Issue 1, Summer 2017: 102–120 © Berghahn Books
ISSN 1934-9688 (Print), ISSN 1934-9696 (Online)

viewers, even in cases of long tracking shots (Wang et al. 2012). Importantly, this synchronization does depend on the kind of natural stimulation—a commercially produced movie achieves much higher synchronization between viewers than footage of unstructured scenes (Hasson et al. 2008). This is also reflected in the intersubjective synchronization of brain activity (Hasson et al. 2004, 2009), particularly in the back of the brain, which is mostly involved with perceptual processing (Wallisch and Movshon 2008).

Whereas these kinds of low-level phenomena have been explored extensively in terms of neural and attentional processing—arriving at a converging picture that they are in fact highly consistent between viewers—the degree of agreement in the appraisal of a given movie between viewers—whether or not they liked it—remains largely unexplored. We want to be clear that this more wholistic appraisal that we focus on here pertains to a different cognitive process, one that is likely more complex. There are several reasons to suspect that this kind of intersubjective agreement might be less than perfect. First, brain activity in large regions of frontal cortex does not appreciably synchronize between viewers, even though they—by virtue of synchronized eye movements and attention—received the same bottom-up signal and thus effectively saw the same movie. Second, there are ample anecdotal reports—for example, virtually any message board on the Internet Movie Database (IMDb) or any film forum—that people disagree vehemently as to how much they like or dislike any given movie. This might be due to the notion that people think they argue about objective facts when considering the quality of a movie (Kivy 2015). Finally, cognitive media theory suggests that movies are extremely rich audiovisual stimuli that allow viewers tremendous degrees of freedom to reconstruct the narrative experience subjectively and idiosyncratically (Bordwell 2013).

The degree of agreement in the appraisal of a given movie between viewers—whether or not they liked it—remains largely unexplored.

Previous research on appraisal of commercially produced major movie releases (hereinafter referred to as films) has focused on the agreement among critics, which was found to be quite high (Boor 1990, 1992; Wanderer 2011a, 2011b). Moreover, studies have investigated the question of whether commercial returns of films can be predicted (Eliashberg 1997) and whether the recommendations by movie critics contribute to the commercial success of said films (Boatwright 2007; Hadida 2008; Holbrook 1999; Kim et al. 2013; Zhang 2004). Both of these questions were answered in the affirmative. However, this research might have overestimated the consensus among nonexpert viewers and between viewers and critics. The emphasis on critics and the consensus among them is also typical of other fields, such as the cognitive study of music (Lundy 2010, 2013).

In contrast to studies showing high agreement between critics and convergent brain activity among noncritic viewers, the degree of agreement between

noncritic viewers and between critics and noncritics remains unexplored. There is reason to believe that the degree of agreement will be low. For instance, *Batman v Superman: Dawn of Justice* (Zack Snyder, 2016) was universally panned among critics (as indicated by its Rotten Tomatoes critic rating), whereas the movie received one of its highest ratings among noncritics (as indicated by its Rotten Tomatoes user rating; Mancini 2016; see also Child 2016 for a discussion of this effect). In addition, debates about the merits and shortcomings of any given movie can be observed on almost any online discussion forum. We want to know whether these anecdotal observations are representative or not and to quantify the degree of agreement in these populations, which has—to our knowledge—not been done. In general, much of the existing literature focuses on box office success and whether it can be predicted, not whether viewers liked a movie.

The purpose of this study is to establish the degree of agreement among nonexpert viewers of films empirically and to quantify the average agreement between the laity of moviegoers and critics in a high-powered (Wallisch 2015) fashion and with a large and representative sample of critics and movies.

Methods

To answer these questions—whether a strong degree of agreement exists between nonexpert viewers and expert critics, as well as within noncritic viewers—we employ the following procedure.

Stimulus Materials

As stimuli, we used a list of 209 major motion pictures released between 1985 and 2004. Given the nature of our study, it was critical to use a representative sample of popular films. Popularity matters because we wanted to ensure a fair chance that the movies we asked about had been seen by our participants and would have garnered several critic ratings. Participants were asked to only rate movies that they had seen. Within this constraint, we wanted to pick a sufficiently large set of representative films chosen mostly at random. Specifically, we selected movies in the following way:

(1) Financial success and popularity (fifty films): The top ten grossing movies for each year from 2000 to 2004.
(2) Critic rating diversity (forty-five films): Movies from Roger Ebert's full range of ratings, five movies from each rating (0 to 4 stars in steps of 0.5 stars), at random. If one of these movies was already including in (1), it was redrawn at random.
(3) Popular acclaim (fifty films): Movies from the IMDb Top 250, at random. If a movie was already contained in (1) or (2), it was redrawn at random.

(4) Diversity (sixty-four films): Drawn from the IMDb completely at random unless already contained in (1), (2), or (3).

Additional constraints: To ensure that our participants had a fair chance of having seen the movie, all films we picked had to have at least fifteen hundred user ratings on IMDb at the time of selection (if a movie picked according to the criteria described above fell below that threshold, it was redrawn at random). To ensure homogeneity of the movie materials, we did not include foreign language films. In addition, to avoid "glow" or nostalgia effects, we picked movies released after 1984. This had the additional benefit that every movie in our sample included a publicly accessible rating by Roger Ebert, the most popular movie critic of the late twentieth century. As far as we can tell, this selection process yielded an unbiased yet differentiated sample of films with a fair chance of our participants having seen them.

Survey

We asked people to appraise these movies. We operationalized appraisal as how much participants "liked" a movie, implemented by asking them to assign a "star rating," as a critic would do. Specifically, we instructed them to rate these movies on a nine-point scale from zero to four stars, including intermediate half-star steps, but only if they had seen the movies. We stressed not to rate movies they had not seen and in particular that they should go with their gut feeling. Afterward, we also asked participants to report demographic information such as age and gender and whether they consider recommendations from movie critics in their choice to see a movie. This survey was deployed online, on surveymonkey.com.

Participants

Participants included a diverse range of undergraduate and graduate students at the University of Chicago and New York University, who were recruited by flyers, as well as people solicited through online ads, specifically through Google AdWords ("Take a survey to test your movie taste"). Responses were collected from 2005 to 2015. We managed to gather data from 3,204 participants in this way. As far as we can tell, our results are qualitatively the same for all subpopulations of participants, which is why we report our results in a pooled fashion, regardless of how we recruited the participants. In addition, we gathered the corresponding ratings for these same movies from forty-two publicly accessible critics or rating sites such as Roger Ebert, IMDb, *Film Comment,* Allociné, and so on. Thus, we ended up with two distinct datasets—one from study participants and one from professional movie critics—that were analyzed separately. The respective University of Chicago and New York University Institutional Review Boards approved all procedures.

Data Analysis

In general, we calculated the Spearman (1904) rank correlation coefficient between any given participant and other participants or critics, as well as between critics and between individuals and averaged ratings. This is the appropriate coefficient, as our data corresponds to measurements on an ordinal level (Stevens 1946). However, we also performed all analyses with Pearson correlation coefficients (Pearson 1920), and all qualitative results remain entirely unchanged. Before calculating the correlation coefficients, we normalized (z-scored) the ratings of each participant to mitigate against biases in how individuals used the rating scale that would introduce non-normal rating distributions. However, this normalization did not affect any of our qualitative results, which is why we believe our findings are robust as to the details of the analysis.

Results

What Are Our Sample Characteristics?

We logged data from 3,204 participants and 42 critic sources. Of the critics, 29 were individual critics (e.g., Roger Ebert, James Berardinelli, Armond White), 4 were aggregated sources of critic information (e.g., Rotten Tomatoes, Metacritic), and 9 were aggregated noncritic users (e.g., IMDb user ratings, Yahoo Movie ratings). We only analyzed data from participants who had seen at least 5 percent of the movies we asked about. Given the nature of the corpus of movies in our survey, we think this is justified, as someone who is a member of the moviegoing public can be reasonably expected to have seen at least 5 (or 10) percent of these popular films. This criterion retained 87 percent of our sample in terms of number of participants and more than 99 percent of the actual ratings. In other words, we analyzed data from 2,784 participants, and those provided almost all of the ratings in our sample. These participants reported to have seen—on average—just less than half of our movies, with a large variation in the number of movies seen between individuals (mean = 99, median = 94, SD = 45). In this sample, 1,199 participants identified as male, 1,216 as female, and the remaining 369 did not provide this information. Given the demographics of our sample (mostly students and Internet users), it unsurprisingly skews toward younger people; however, this age group is also overrepresented among the moviegoing public. Specifically, the mean age in our sample is 25.14 years, with a median of 21 and a standard deviation of 9.86 years. The three highest-rated movies in our sample were *The Shawshank Redemption* (Frank Darabont, 1994), *Schindler's List* (Steven Spielberg, 1993), and *Goodfellas* (Martin Scorsese, 1990), with average z-scores of 0.93, 0.78, and 0.71, respectively. The lowest-rated movies were *Gigli* (Martin Brest, 2002), *Battlefield Earth* (Roger Christian, 2000), and *Crossroads* (Tamra Davis, 2002), with respective z-scores of -1.58, -1.36, and -1.31. The fact that both highest and lowest scores were achieved for "notoriously" good or bad movies raises the pos-

sibility that our measures confound reputation with personal liking, at least at the extreme end of the scale (for movies that have a reputation). However, we would like to note that when we tracked the stability of ratings on IMDb for a few movies released in 2004, we could not discern much of an effect of reputation; the average ratings were remarkably stable from release to several months later, despite sometimes an order of magnitude or more increase in the number of ratings. Also, few people in our sample did rate *Gigli* or *Battlefield Earth*, implying that those who rated them are also those who saw them. Interestingly, those were not dissuaded by reputation, but they still did not like the movie. Finally, our results suggest that there is not much of a reputation effect at all; we were surprised how little agreement there was even for obvious classics like *The Shawshank Redemption*. As you can see, extreme z-scores are rare; even the "best" (most highly rated) movies in our sample do not even achieve an average z-score of 1. This is a hint that there will be relatively little agreement about movies that everyone likes whereas there is a stronger consensus about movies that are disliked, as the lowest-rated movies in our sample exceeded z-scores of -1.

Given This Sample, What Is the Average Level of Agreement between Participants?

To answer this question, we correlated the vector of movie ratings of each participant with that of all other participants. To ensure the stability of the correlation coefficients, we only included relations between people if they jointly reported ratings for at least ten movies. Correlation coefficients below that number are too unstable, for example, correlations that are based on two data points are necessarily either -1 and 1. We plot the distribution of joint movie ratings reported in Figure 1.

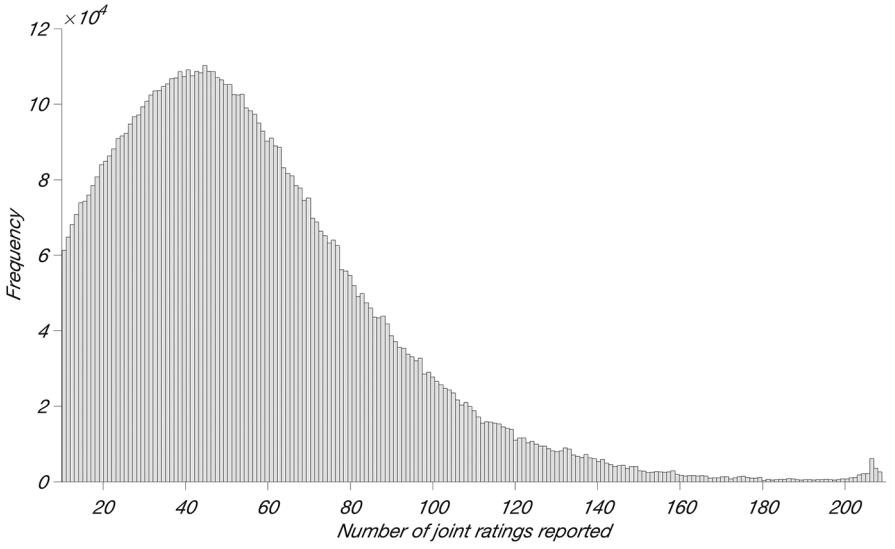

Figure 1. Distribution of joint ratings. The x-axis represents how many movies participants have seen jointly, whereas the y-axis represents the number of correlation coefficients in that bin.

This distribution is highly skewed with a mean of 55.4 and a standard deviation of 32.1. Interestingly, this distribution is bimodal, suggesting that there seems to have been a subpopulation of people on the right side of the distribution who report to have seen virtually every movie on our survey. These people probably represent either "movie buffs" or people who did not follow the instructions (and rated all movies instead of just movies they saw) or a combination of these possibilities. Note that we cut this distribution on the left side to only show the data that we actually use in the rest of the analysis: people who saw less than 5 percent of the movies in our sample are not representative of the movie-watching public as reflected by this distribution. In addition, as we made ten jointly seen movies to be the criterion for our correlation analysis, these participants would not be able to contribute any data to that analysis. However, as noted above, this cutoff retains 87 percent of study participants and more than 99 percent of the ratings. Doing the analysis on the full (noncut) dataset does not qualitatively change any of our results.

Based on this distribution of joint rankings, we now calculate a full cross of all individual rating vectors with all others, which we show in Figure 2.

As you can see, the average correlation coefficient is relatively low (mean = 0.2592, median = 0.27), with large variation (SD = 0.2) and distributed as a shifted normal distribution with a very long tail into the negative correlations. This average correlation is significantly different from zero; the 95 percent confidence interval is 0.2590 to 0.2593. If one averages the correlations first, on a per-person basis, the mean agreement between people is 0.258 (median = 0.273, SD = 0.09). A one-sample t-test establishes that this is different from

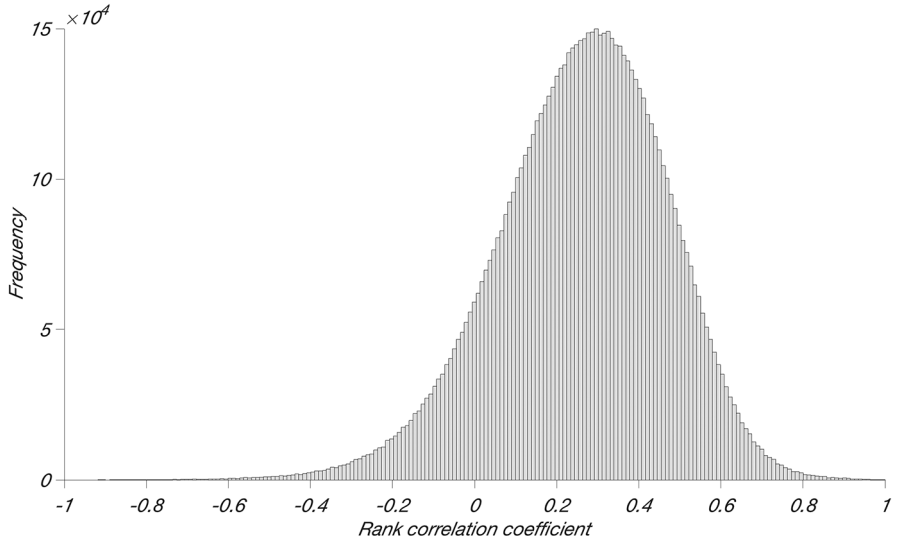

Figure 2. Average intercorrelation. The x-axis represents the magnitude of the Spearman rank correlation coefficient between any two ratings vectors and the y-axis the number of instances that achieve a given correlation.

zero ($p < 0.001$, $t = 153$, $df = 2781$, CI = [0.255,0.262]). Most individuals have moderate correlations with other individuals. There are only a few "superpredictors" with average correlations around 0.5 and only a thin tail extending into the negative correlations (see Figure 3).

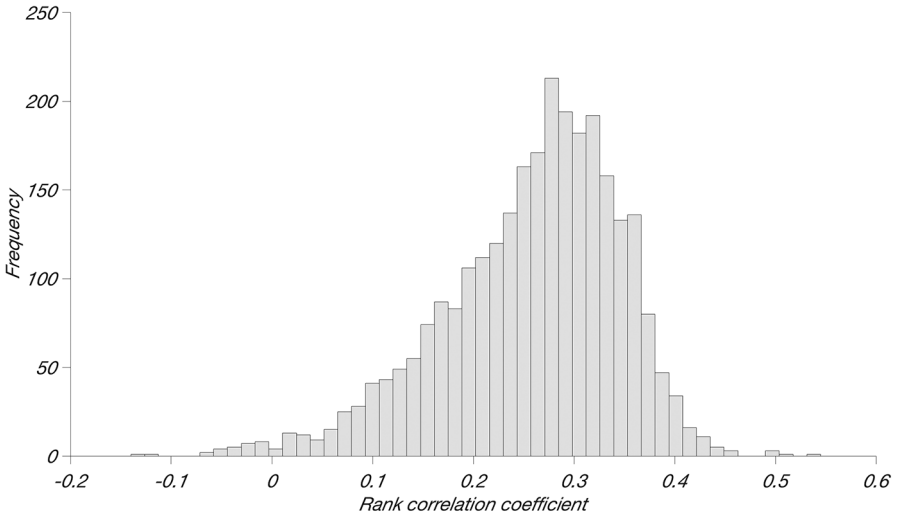

Figure 3. Intercorrelations per individual. The x-axis represents the magnitude of the average Spearman rank correlation coefficient between an individual and all other individuals in the sample and the y-axis the number of instances that achieve a given correlation.

This average correlation is a function of the number of movies rated, as can be seen in Figure 4.

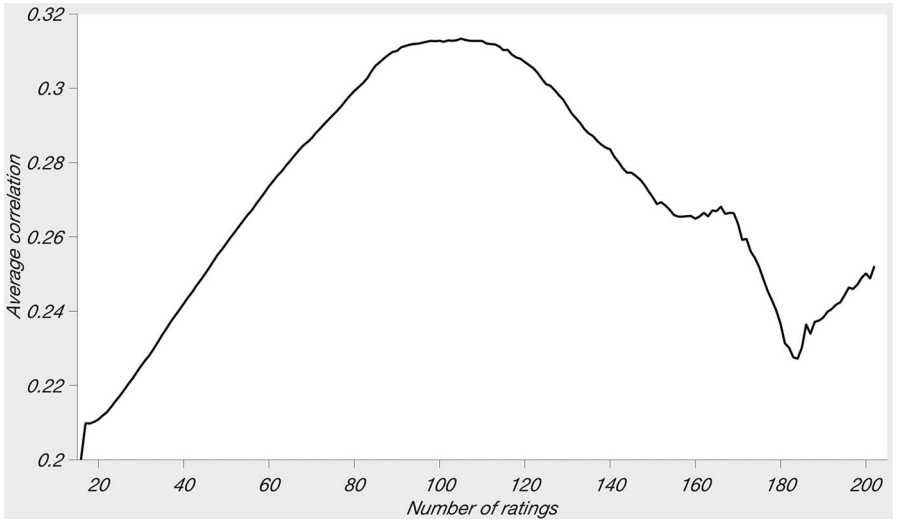

Figure 4. The fine structure of correlation magnitude as a function of joint number of movies seen. On the x-axis is the number of ratings, and on the y-axis is the average correlation based on that number of ratings in our sample.

We convolved the raw data with a kernel of bin width of fifteen before plotting it. As you can see, we can discern four clear groups, from left to right: if people have seen few movies, the average correlation is low. In other words, these people tend to have an uncalibrated movie taste. As the joint number of movies increases, the average correlation reaches a maximum around half of the movies in our sample. The distribution seems to form a stable plateau that might well represent the average movie taste of the casual moviegoer in the United States. Interestingly, the average intercorrelation then dips again as a function of increasing number of movies seen. However, this dip is hard to interpret. Movie taste possibly becomes more idiosyncratic as viewers become more discerning—as a function of the number of movies seen. As the nadir of this dip is close to the maximum possible—around 180 movies jointly seen—it could be that this is an artifact of improper survey completion by those who report to have seen all of the movies. Moreover, not many people underlie this part of the curve, so statistical reliability of the dip is low. Regardless of the reason for the dip, agreement rises again as people report to have seen more than 180 movies together. This could reflect the formation of a truly expert movie taste held by movie buffs, film students and the like, perhaps informed by exposure to similar instruction material such as film theory, classic movies, or familiarity with the history of medium. If true, this has two implications. First, if we had set a lower cutoff for joint movies rated in order to calculate correlations, the average agreement among participants would be even lower. Second, as the agreement among participants who have seen many movies was high, we might expect that a separate population—that of the critics—could exhibit a similarly high degree of agreement, simply by virtue of the fact that they—as a group—also have seen a lot of the movies in the sample. A more prosaic explanation is that those at the far end of the curve did not properly fill in the survey. We cannot distinguish between these possibilities at this point. One way to do so would be to administer a substantially longer instrument, for instance, one with several hundred more movies.

Do Demographic Factors Such as Age or Gender Modulate This Average Correlation?

It is often thought that demographic variables such as age or gender are important predictors of psychological phenomena in general. In the film industry specifically, there is the pervasive notion that certain kinds of movies are designed to appeal to specific demographics (e.g., "chick flicks" and melodramas for women or action and sci-fi movies for men). In addition, gender has been a topic of considerable interest within film theory (Doane 1987; Mulvey 1975; Williams 1991). Thus, we wondered whether these factors could modulate the intersubjective agreement between these groups: is the intersubjective agreement within a group (e.g., among female movie watchers) higher than the

agreement between groups (e.g., between male and female watchers)? If true that gender is an important predictor of movie taste—and modulates individual perception of movies along gender lines, we would expect high inter-rater agreement within a gender but low inter-rater agreement between genders.

Doing so, we noticed the following results: first, male movie taste seems to be more consistent than female movie taste. The average intersubjective correlation among males is 0.31 but 0.24 for females, and this difference is significant ($t = 21.40$, $df = 2411$, $p < 0.001$) and of an appreciable effect size: $d = 0.87$. This difference cannot be attributed to the fact that males saw fewer movies than females, which could artificially inflate correlations. In fact, males saw more movies (68, on average) than females (51), and this difference is also significant ($t = 20.1$, $df = 2411$, $p < 0.001$) and of a considerable effect size ($d = 0.82$). One possible explanation for this could be the notion that Hollywood is a male-dominated industry (Dargis 2014), the idea being that movies that appeal more strongly to one gender (as they are made by people of that gender) are more effective at doing so. However, there is no evidence that the movie taste between males and females (mean = 0.249) is notably different from the average movie taste in general, where everyone is correlated with everyone else (mean = 0.258). While this difference is significant ($t = 3.56$, $df = 4034$, $p < 0.01$), it is of questionable relevance; the effect size is very small ($d = 0.1$).

Similarly, when dividing our sample by median age, we found no differences in agreement within age groups (young vs. old = 0.265 and 0.271, respectively, $t = 1.7$, $df = 2548$, n.s.) or between age groups ($t = 0.57$, $df = 4048$, n.s.), despite the fact that the older participants had seen substantially more movies: 62 versus 54 ($t = 9.4$, $df = 2548$, $p < 0.001$); $d = 0.37$. This should not be surprising, as they had a lot more time to watch movies, but this had no apparent implications on taste. We conclude that age was not a factor in determining movie taste but point out that this study was not designed to elicit whether there is such a difference. In other words, our entire sample was rather youthful; a more diverse age distribution might have revealed stronger effects.

Finally, we wondered whether there is a difference in taste between people who report having seen more than the median number of movies in our sample (94) versus those who report having seen less than the median. Presumably, these are either different populations (i.e., "movie nerds" vs. casual moviegoers), or exposure to so many movies might have shaped their taste in itself. Empirically, we find that this does make a difference. The people who saw more movies than average have a more homogenous movie taste than those who saw less (0.31 vs. 0.23 in terms of intersubjective correlation, $t = 23.16$, $df = 2792$, $p < 0.001$), and this effect is strong ($d = 0.90$). Of course, there is also a substantial difference in the number of movies they have seen, because that is how these are defined. If this was not so, the observed difference in taste would be puzzling (95 vs. 28 movies seen, respectively, $t = 132.83$, $df =$

2792, d = 4.68). However, when comparing movie taste between these groups to that of everyone, the difference in intercorrelation among those who have seen a lot versus a few movies is extremely modest (0.253 vs. 0.258, t = 2.42, df = 4217, $p < 0.05$, d = 0.08).

What Is the Correlation between People and Critics as Well as between Different Kinds of Critics?

We wondered as to the correlation within different groups of critics versus correlations between them. This information is best represented in the form of a table (see Table 1).

Table 1. Average intercorrelation between groups

	Individuals	Critics	Aggregated critics	Aggregated individuals
Individuals	0.260	0.267	.0329	0.437
Critics	–	0.394	0.552	0.488
Aggregated critics	–	–	0.784	0.666
Aggregated individuals	–	–	–	0.807

Correlations increase as the number of individuals involved increases, or as the homogeneity of the group (e.g., critics) increases. Interestingly, critics seem to be best at predicting the responses of other critics—individually or together—but not the responses of people. The best predictor of individual noncritic movie taste is aggregated individual noncritic movie taste.

Discussion

We showed that the intersubjective agreement of movie taste is low but significantly different from zero. There is some agreement, but not much. We also showed that demographic qualifiers do not matter much in terms of moderating these numbers, indicating that movie taste is an idiosyncratic quality—of the individual, not of demographically defined groups like age or gender. This issue has recently been raised anecdotally and seems to be exploited commercially (Barrett 2016). Moreover, we have shown that the agreement between people and individual critics is low, whereas the agreement among critics themselves is high. This emphasizes and quantifies the fundamental disconnect between people's and critics' tastes that has been highlighted anecdotally. So it seems that the *Batman v Superman* example we introduced is perhaps a strong example of the critic/audience disconnect, but it is by no means an outlier. This divide has sparked many journalistic reflections on the importance of film criticism (Child 2016; Mancini 2016). Ironically, something about being a critic seems to make the recommendations of critics unsuitable—relatively speaking—for predicting

Intersubjective movie appraisal is low but significantly different from zero.

the movie taste of regular people. We want to emphasize that we cannot distinguish possible explanations for why this might be the case. It is possible that critics are fundamentally different people and unrepresentative of the population at large. Critics and audience members may also have different cognitive goals while watching a movie. Moreover, the correlated inputs of the critics—the movies they have seen, the film theory they have read, the conventions they are aware of—may also synchronize their judgments but put them at odds with the judgment of average viewers. If this is the case, pooled predictions based on critics should saturate early, whereas the decorrelated judgments of individuals should not, which is what we see empirically (Zohary et al. 1994). None of these explanations is mutually exclusive; they could all be true at the same time. Sorting them out will have to be the subject of future research.

Finally, we showed that the best predictor of individual movie ratings was aggregated individual movie ratings (e.g., IMDb rating, average Yahoo user rating), with correlations close to the theoretically highest possible value, given the inherent diversity of movie taste. Of course, even better predictions are possible if one tailors recommendations to the individual, but not to groups, as the lack of consensus about the appraisal of films in the population precludes giving ratings that appeal to everyone. That is, when the predicted rating for one movie becomes too high, it gets out of touch with the people who did not like that particular movie, which is why companies like Netflix use closeness to clusters in taste space to make predictions. The idea is that groups in a highly dimensional taste space share a particular taste but also have a taste that is different from those of other groups. If one has not seen a given movie, any given taste cluster will have seen it, and one's predicted rating corresponds to the rating of the nearest cluster. After launching in 130 countries in 2016, Netflix representatives revealed that the company relies on taste profiles in this taste space rather than demographic information or geographical location (Barrett 2016). In other words, what matters when predicting how much someone will like a given movie is how close that person is to someone in taste space, not physical or demographic proximity.

From a psychometric perspective, the extreme ratings of individual critics make them poor predictors of individual movie taste, whereas aggregate ratings (e.g., IMDb) tend to avoid extreme ratings. Moreover, it is possible that movie critics pursue different motivations when assessing a movie. Whereas audience responses to a film are perhaps dominated by their emotional experience, critics might also consider what the filmmakers were trying to achieve when making the movies and whether they succeeded doing so. These considerations put the very purpose of a movie critic into question. Of course, the cultural role of the critic has been discussed at length (Cameron 1995; Debenedetti 2006). We show empirically that if people are to take the ratings

We show empirically that if people are to take the ratings of movie critics as viewing recommendations, they are better off either consulting average ratings or figuring out which movie critic is most predictive of* their *taste.

of movie critics as viewing recommendations, they are better off either consulting average ratings or figuring out which movie critic is most predictive of *their* taste. This implies that professional movie critics are popular not because of the accuracy of their forecasts but because of their presentation (e.g., the quality of their writing or their insights about the movie). Movie critics would perhaps respond that their job is not to pander to the basest tastes of their audience but rather to educate them about the finer nuances of film evaluation or analysis. In other words, professional movie critics might see themselves as educators and tastemakers, not as bellwethers. This is in line with Pierre Bourdieu's (1984) reflections on the relationship between taste and cultural capital. Bourdieu points out that those with a large amount of cultural capital are able to determine what constitutes taste. Perhaps this is why the judgment of movie critics—the societally accepted cultural arbiters of taste—is accepted despite lacking predictive validity. However, this is not how many people use the information from movie critics. In our sample, about a third of our participants explicitly stated that they do use rating information from movie critics as recommendations whether to see a movie, more than any other potential information source—such as advertising or word of mouth. Most of these participants mentioned Roger Ebert as the most frequently consulted and the most accurate individual movie critic.

Our research touches on a deeper issue as well, namely the conflict in the perceived merits of experts versus collectives of individuals (Tetlock 2015). This issue plays out in many areas, going back to ancient times. Who makes more just judgments, expert judges or juries? Who can be expected to rule more wisely, a philosopher king or democracy? In terms of forecasting and which forecasts are most reliable, the notion of a "wisdom of the crowds" has taken hold (Surowiecki 2004). The aggregated judgment of many laypeople consistently outperforms the predictions by experts in a diverse range of fields, including searching for submarines but also political judgment. In the arena of political judgment, aggregating polling data with smart algorithms seems to far outperform the judgment of pundits (Tetlock 2005). Given our results, the same seems to be true for movie taste.

In terms of the prediction of movie taste, sites that aggregate individual judgments like IMDb are the clear winner. This might be counterintuitive, as sites that aggregate critic judgments like Rotten Tomatoes or Metacritic have more perceived authority, as they are curated by professionals. Unfortunately, these are only good at predicting the taste of other professionals, not of civilians.

Finally, we want to note that this research opens up an exciting prospect: psychology has explored mental responses to simple stimuli for well over a hundred years. The promise of this analytical approach was that it would allow us—over time—to understand how the brain and mind work, as evidence would accumulate in a linear fashion (Rust and Movshon 2005). For instance, if one can characterize the response of the brain to oriented lines, one can then predict and investigate its response to a combination of such lines. However, there are several concerns about this approach. People and their brains are complex—and the system inherently nonlinear—enough that progress has been slow and the response to complex stimuli can often not be predicted from that of the system to a linear combination of the components that make up the stimulus (Felsen and Dan 2005; Jazayeri et al. 2012).

More recently, a complementary approach to using simple and carefully designed stimuli has gained currency, namely using "natural scenes" or "natural images" (Geisler 2008). Movies are not the only stimulus material used in this research. For instance, investigations of neuroaesthetics have yielded interesting results, linking aesthetic experiences to the activation of specific networks in the brain (Vessel et al. 2012). However, movies have properties that provide the researcher with several critical affordances. We know, for example, that they activate the back of the brain and the systems that guide attention and gaze similarly and strongly (Hasson et al. 2004; Smith 2012; Wang et al. 2012). Interestingly, the appraisal of these highly engaging stimuli seems to be up to the individual. In other words, movies might allow us to present people with a compelling piece of (virtual) reality that have been suggested to be—for much of the brain—hard to distinguish from reality (Grodal 2006). Of course, there is some debate about whether this is the case. For instance, it is plausible that the large prefrontal cortex of human primates allows them to override the bottom-up signals from sensory cortices, telling them what they are seeing is "just a movie." However, there might be some a priori arguments in favor of Torben Grodal's thesis: for most of the evolutionary history of the brain, it was not challenged with the necessity to distinguish virtual from actual stimuli, so much of the brain might be prone to take perceptual inputs on face value (Anderson 1996; Zacks 2015). Of course, both of these things might be going on at the same time: the back of the brain processing virtual and real stimuli in an essentially indistinguishable fashion (Hasson et al. 2004), whereas the prefrontal cortex is engaged in reality monitoring and can suspend belief when watching unbelievable things (Grubb et al. 2010). Ultimately, it will be an empirical question of which of these processes dominates in regular observers; it might be interesting to see whether there are individual differences in the propensity to take virtual realities at face value. If true, this enables researchers to look at differential responses of individuals to the same reality, allowing

them to investigate the elusive top-down processes by which people bring about the subjective reality that they inhabit. As this—probing how people sample and interpret reality—is difficult to do in a laboratory setting, it remains understudied. Here, we show that movies are a controllable snippet of "virtual reality" that can be deployed in the lab and will yield idiosyncratic but systematic individual responses.

These findings raise several questions that can be investigated in future research. The first question is what mechanism accounts for the diversity in appraisal. We already know that people do broadly see the same movie, as eyes and attention are engaged similarly and trained on the same focal points (Smith 2012; Wang et al. 2012). However, it remains unclear whether people differ in the appraisal of what they see or already in what they think they perceive (i.e., how they characterize the stimulus). It is possible that someone perceives a comedy as a drama and they do not like the movie because they like comedy, but not drama, whereas someone else with the same preferences differs in their appraisal because they—correctly—perceive the movie as a comedy. If so, this would imply a multistep process where diversity of movie appreciation could enter: either at the point of perception (early) or at the point of valuation (late). Conversely, it is possible that both observers agree in terms of what they see, just not whether they like it (e.g., both perceive it as a comedy, but only one of them finds it funny). Which of these accounts underlies the difference in appraisal will have to be elucidated by future research.

In conclusion, we note that it remains to be investigated how even more immersive technology like genuine virtual reality (VR) systems are able to engage the brain. It might turn out that movies are already sufficient in terms of providing a compelling virtual experience that can be interpreted idiosyncratically. Still, the differences in interpretation of movies might be due to the degrees of freedom afforded by the medium (e.g., lack of haptic feedback, weaker depth and motion cues, lack of 360-degree immersion, lack of agency) and not due to the idiosyncratic reconstructive processes of individual brains. If this is the case, intersubjective agreement should be higher in experiences induced by VR.

Whether movies are sufficient to reveal irreducible and real differences in how individuals interpret reality or whether the apparent lack of agreement in appraisal is due to an impoverished environment relative to genuine VR remains to be decided empirically.

Acknowledgments

Parts of this article have previously been presented in abstract form (Wallisch 2005; Whritner and Wallisch 2015; Wallisch and Whritner 2016).

Pascal Wallisch, PhD, serves as a clinical assistant professor of psychology at New York University. His research interests concern motion perception, generally conceived, for example, its neural basis, its attentional impact, and its appraisal. His work was recognized with a variety of honors, including the Wayne C. Booth Prize and Golden Dozen Award for Excellence in Teaching, as well as the First Eagleman Prize. He also was a recipient of a German National Merit Scholarship.

Jake Alden Whritner is a laboratory/tech assistant at the Cognitive and Data Science (CoDaS) Lab at the Rutgers University-Newark, where he works on a visual perception project that uses augmented reality to study and train the visual system. His research interests include dynamic event cognition and the interaction between visual processing and attention.

References

Barrett, Brian. 2016. "Netflix's Grand, Daring, Maybe Crazy Plan to Conquer the World." *Wired,* 27 March. http://www.wired.com/2016/03/netflixs-grand-maybe-crazy-plan-conquer-world.

Boatwright, Peter, Suman Basuroy, and Wagner Kamakura. 2007. "Reviewing the Reviewers: The Impact of Individual Film Critics on Box Office Performance." *Quantitative Marketing and Economics* 5 (4): 401–425.

Boor, Myron. 1990. "Reliability of Ratings of Movies by Professional Movie Critics." *Psychological Reports* 67: 243–257.

Boor, Myron. 1992. "Relationships Among Ratings of Movie Pictures by Viewers and Six Professional Movie Critics." *Psychological Reports* 70: 1011–1021.

Bordwell, David. 2013. "The Viewer's Share: Models of Mind in Explaining Film." In *Psychocinematics: Exploring Cognition at the Movies,* ed. Arthur P. Shimamura, 29–53. New York: Oxford University Press.

Bordwell, David, and Kristin Thompson. 2010. *Film Art: An Introduction.* 9th ed. New York: McGraw-Hill.

Bourdieu, Pierre. 1984. *Distinction: A Social Critique of the Judgement of Taste.* Cambridge, MA: Harvard University Press.

Cameron, Sam. "On the Role of Critics in the Cultural Industry." *Journal of Cultural Economics* 19 (4): 321–331.

Child, Ben. 2016. "Is the Biggest Batman v Superman Smackdown between Fans and Critics?" *Guardian,* 28 March. http://www.theguardian.com/film/filmblog/2016/mar/28/is-the-biggest-batman-v-superman-smackdown-between-fans-and-critics.

Dargis, Manohla. 2014. "In Hollywood, It's a Men's, Men's, Men's World." *New York Times,* 24 December. http://www.nytimes.com/2014/12/28/movies/in-hollywood-its-a-mens-mens-mens-world.html.

Debenedetti, Stéphane. 2006. "The Role of Media Critics in the Cultural Industries." *International Journal of Arts Management* 8 (3): 30–42.

Dellarocas, Chysanthos, Neveen Farag Awad, and Xiaoquan (Michael) Zhang. 2004. "Exploring the Value of Online Reviews to Organizations: Implications for Revenue Forecasting and Planning." *ICIS 2004 Proceedings* 30. http://aisel.aisnet.org/icis2004/30.

Doane, Mary Ann. 1987. *The Desire to Desire: The Woman's Film of the 1940s*. Bloomington: Indiana University Press.

Eliashberg, Jehoshua, and Steven M. Shugan. 1997. "Film Critics: Influencers or Predictors?" *Journal of Marketing* 61: 68–78.

Felsen, Gidon, and Yang Dan. 2005. "A Natural Approach to Studying Vision." *Nature Neuroscience* 8: 1643–1646. doi:10.1038/nn1608.

Geisler, Wilson S. 2008. "Visual Perception and the Statistical Properties of Natural Scenes." *Annual Review of Psychology* 59: 167–192. doi:10.1146/annurev.psych.58.110405.085632.

Grodal, Torben. 2006. "The PECMA Flow: A General Model of Visual Aesthetics." *Film Studies* 8 (1): 1–11. doi:10.7227/FS.8.3.

Grubb, Michael, David J. Heeger, Uri Hasson, and Pascal Wallisch. 2010. "Subjective Preference and Its Effect on the Reliability of Cortical Activity during Movie Viewing." Poster presented at the 40th Annual Meeting of the Society for Neuroscience, San Diego, CA, 13–17 November.

Hadida, Allègre L. 2008. "Motion Picture Performance: A Review and Research Agenda." *International Journal of Management Reviews* 11 (3): 297–335. doi:10.1111/j.1468-2370.2008.00240.x.

Hasson, Uri, Ohad Landesman, Barbara Knappmeyer, Ignacio Vallines, Nava Rubin, and David J. Heeger. 2008. "Neurocinematics: The Neuroscience of Film." *Projections: The Journal for Movies and Mind* 2 (1): 1–26. doi:10.3167/proj.2008.020102.

Hasson, Uri, Rafael Malach, and David J. Heeger. 2009. "Reliability of Cortical Activity during Natural Stimulation." *Trends in Cognitive Sciences* 14 (1): 40–48.

Hasson, Uri, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. 2004. "Intersubject Synchronization of Cortical Activity during Natural Vision." *Science* 303 (5664): 1634–1640. doi:10.1126/science.1089506.

Holbrook, Morris B. 1999. "Popular Appeal versus Expert Judgments of Motion Pictures." *Journal of Consumer Research* 26: 144–155.

Jacobs, Ruud S., Ard Heuvelman, Somaya Ben Allouch, and Oscar Peters. 2015. "Everyone's a Critic: The Power of Expert and Consumer Reviews to Shape Readers' Post-Viewing Motion Picture Evaluations." *Poetics* 52: 91–103. doi:10.1016/j.poetic.2015.07.002.

Jazayeri, Mehrdad, Pascal Wallisch, and J. Anthony Movshon. 2012. "Dynamics of Macaque MT Cell Responses to Grating Triplets." *Journal of Neuroscience* 32 (24): 8242–8253. doi:10.1523/JNEUROSCI.5787-11.2012.

Kim, Sang Ho, Namkee Park, and Seung Hyun Park. 2013. "Exploring the Effects of Online Word of Mouth and Expert Reviews on Theatrical Movies' Box Office Success." *Journal of Media Economics* 26 (2): 98–114. doi:10.1080/08997764.2013.785551.

Kivy, Peter. 2015. *De Gustibus: Arguing about Taste and Why We Do It*. Oxford: Oxford University Press.

Ladhari, Riadh, and Miguel Morales. 2004. "The Influence of Individual Characteristics on Moviegoers' Satisfaction Ratings." Paper presented at the Atlantic Schools of Business Conference, Mount Saint Vincent University, Halifax, Nova Scotia, 4–6 November.

Lundy, Duane E. 2010. "A Test of Consensus in Aesthetic Evaluation among Professional Critics of Modern Music." *Empirical Studies of the Arts* 28 (2): 243–258. doi:10.2190/EM.28.2.h.

Lundy, Duane E. 2013. "Critiquing the Critics: Statistical Analysis of Music Critics' Rating Distributions as a Measure of Individual Refinement." *Empirical Studies of the Arts* 31 (1): 59–79. doi:10.2190/EM.31.1.d.

Mancini, Vince. 2016. "'Batman v Superman' and the Myth That Box Office Repudiates Critics." *Uproxx,* 28 March. http://uproxx.com/filmdrunk/batman-v-superman-audience-critic-divide.

Mulvey, Laura. 1975. "Visual Pleasure and Narrative Cinema." *Screen* 16 (3): 6–18. doi:10.1093/screen/16.3.6.

Pearson, Karl. 1920. "Notes on the History of Correlation." *Biometrika* 13 (1): 25–45.

Plucker, Jonathan A., James C. Kaufman, Jason S. Temple, and Meihua Qian. 2009. "Do Experts and Novices Evaluate Movies the Same Way?" *Psychology & Marketing* 26 (5): 470–478. doi:10.1002/mar.20283.

Rust, Nicole, C., and J. Anthony Movshon. 2005. "In Praise of Artifice." *Nature Neuroscience* 8: 1647–1650. doi:10.1038/nn1606.

Schrage, Scott. 2012. "The Impact of Movie Reviews vs. Word of Mouth on Post-Viewing Evaluations of Films." MA thesis, Iowa State University.

Seeley, William, and Noël Carroll. 2014. "Cognitive Theory and the Individual Film: The Case of *Rear Window*." In *Cognitive Media Theory,* ed. Ted Nannicelli and Paul Taberham, 235–253. New York: Routledge.

Simonton, Dean Keith. 2007. "Is Bad Art the Opposite of Good Art? Positive Versus Negative Cinematic Assessments of 877 Feature Films." *Empirical Studies of the Arts* 25 (2): 143–161. doi:10.2190/2447-30T2-6088-7752.

Simonton, Dean Keith. 2011. *Great Flicks: Scientific Studies of Cinematic Creativity and Aesthetics.* Oxford: Oxford University Press.

Smith, Tim J. 2012. "The Attentional Theory of Cinematic Continuity." *Projections: The Journal for Movies and Mind* 6 (1): 1–27.

Smith, Tim J., Daniel Levin, and James E. Cutting. 2012. "A Window on Reality: Perceiving Edited Moving Images." *Current Directions in Psychological Science* 21: 101–106. doi:1177/0963721412436809.

Spearman, C. 1904. "The Proof and Measurement of Association between Two Things." *The American Journal of Psychology* 15 (1): 72–101. doi:10.2307/1412159.

Stevens, S.S. 1946. "On the Theory of Scales of Measurement." *Science* 103 (2684): 677–680.

Surowiecki, James. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* New York: Anchor.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

Tetlock, Philip E., and Dan Gardner. 2015. *Superforecasting: The Art and Science of Prediction.* New York: Crown.

Vessel, Edward A., G. Gabrielle Starr, and Nava Rubin. 2012. "The Brain on Art: Intense Aesthetic Experience Activates the Default Mode Network." *Frontiers in Human Neuroscience* 6 (66): 1–17. doi:10.3389/fnhum.2012.00066.

Wallisch, Pascal. 2005. "Affective Responses to Moving Complex Audio-Visual Stimuli Are Better Predicted by 'Simple' Response Pools Than by Individual or Pooled Expert Predictors." Paper presented at the Annual Meeting of the Midwestern Psychological Association, Chicago., 6 May.

Wallisch, Pascal. 2015. "Brighter Than the Sun: Powerscape Visualizations Illustrate Power Needs in Neuroscience and Psychology." arXiv:1512.09368.

Wallisch, Pascal, and J. Anthony Movshon. 2008. "Structure and Function Come Unglued in the Visual Cortex." *Neuron* 60: 195–197. doi:10.1016/j.neuron.2008.10.008.

Wallisch, Pascal, and Jake Alden Whritner. 2016. "Surprisingly High Concordance Rates in the Categorization of Movies." Paper presented at the annual conference for the Society for Cognitive Studies of the Moving Image, Ithaca, NY, 1–4 June.

Wang, Helena, Jeremy Freeman, Elisha Merriam, Uri Hasson, and David Heeger. 2012. "Temporal Eye Movement Strategies during Naturalistic Viewing." *Journal of Vision* 12 (1): 1–27.

Wanderer, Jules J. 2011a. "Scaling Professional Critics: Men and Women Rate Films." *Empirical Studies of the Arts* 29 (2): 209–223. doi:10.2190/EM.29.2.e.

Wanderer, Jules J. 2011b. "When Film Critics Agree: Does Film Genre Matter?" *Empirical Studies of the Arts* 29 (1): 39–50. doi:10.2190/EM.29.1.c.

Whritner, Jake Alden, and Pascal Wallisch. 2015. "A Neurocinematic Approach to the Appraisal of Film." Paper presented at the annual conference for the Society for Cognitive Studies of the Moving Image, London, 17–20 June.

Williams, Linda. 1991. "Film Bodies: Gender, Genre, and Excess." *Film Quarterly* 44 (4): 2–13.

Zohary, Ehud, Michael N. Shadlen, and William T. Newsome. 1994. "Correlated Neuronal Discharge Tate and Its Implications for Psychophysical Performance." *Nature* 370: 140–143.