Do Experts and Novices Evaluate Movies the Same Way?

Jonathan A. Plucker
Indiana University

James C. Kaufman, and Jason S. Temple
California State University, San Bernardino

Meihua Qian
Indiana University

ABSTRACT

Do experts and novices evaluate creativity the same way? This question is particularly relevant to the study of critical and public response to movies. How do public opinions differ from movie critic opinions? This study assessed college student (i.e., novice) ratings on movies released from 2001 to 2005 and compared them to expert opinions and those of self-declared novices on major movie rating Web sites. Results suggest that the student ratings overlapped considerably—but not overwhelmingly—with the self-described novices, student ratings correlated at a lower magnitude with critic ratings, and the ratings of students who saw the most movies correlated more highly with both critics and self-described novices than those of students who saw the least movies. The results suggest a continuum of creative evaluation in which the distinctions between categories such as “novice,” “amateur,” and “expert” are blurry and often overlap—yet the categories of expertise are not without importance. © 2009 Wiley Periodicals, Inc.
Psychologists studying creativity, with a long history of applied research in the arts and literature, have recently turned their empirical attention to the study of film. For example, Simonton (2004a, 2004b, 2005a, 2005b, 2006, 2007a, 2007b, 2007c) has conducted a series of increasingly complex studies examining movie awards, critic ratings, box office, and their relationships, among many other phenomena. Delmestri, Montanari, and Usai (2005) have studied artistic merit and commercial success of Italian films, and Plucker, Holden, and Neustadter (2008) have investigated the psychometric properties of various film ratings and the ratings’ correlation with specific box office data. Zickar and Slaughter (1999) have used the careers of film directors to investigate patterns of longitudinal creative production.

The reasons for this sudden growth in interest include the large, readily available pool of extant data on movies, the relatively thin pool of studies on creativity in film relative to other popular art forms, and the traditional, popular fascination with movies. However, perhaps the major motivation for studying cinematic creativity is the ability to influence the creation and, especially, the marketing of movies. Much of the existing empirical research on film marketing comes, not surprisingly, from marketing researchers. But creativity researchers, with their appreciation for both the creative product and process, may have insights into how all aspects of film—writing, directing, acting, production, distribution, marketing—can be better understood (and perhaps improved) by the application of models and theories from the broader literature on creativity and innovation.

THE CRITERION PROBLEM

A problem encountered by researchers interested in cinematic creativity is a version of the traditional criterion problem that affects all creativity research. MacKinnon (1978) argued that “the starting point, indeed the bedrock of all studies of creativity, is an analysis of creative products, a determination of what it is that makes them different from more mundane products” (p. 187). Csikszentmihalyi (1999) wrote,

If creativity is to have a useful meaning, it must refer to a process that results in an idea or product that is recognized and adopted by others. Originality, freshness of perception, divergent-thinking ability are all well and good in their own right, as desirable personal traits. But without some sort of public recognition they do not constitute creativity. ... The underlying assumption [in all creativity research] is that an objective quality called “creativity” is revealed in the products, and that judges and raters can recognize it. (p. 314)

Many psychologists have shared MacKinnon’s and Csikszentmihalyi’s belief in the importance of the creative product. The importance of the creative product emerged in response to perceived needs for external criteria by which researchers could compare other methods of measuring creativity for the purposes of establishing evidence of validity. However, an absolute and indisputable criterion of creativity is not readily available, hence the criterion problem (Plucker & Renzulli, 1999).

Although approaches to the study of creative products vary greatly (see Plucker & Renzulli, 1999), the most common method for the measurement of creative products utilizes the ratings of external judges, especially the use of consensual assessment techniques (Amabile, 1983; Baer, Kaufman, & Gentile, 2004; Hennessey & Amabile, 1988; Kaufman, Baer, & Gentile, 2004). The Consensual Assessment Technique (CAT) is based on the belief that future creativity is best
Although we acknowledge that one use of critics’ opinions in film marketing is to get people to buy tickets on opening weekend, regardless of whether laypeople are likely to agree with the critics, this strategy is only effective if laypeople believe that their assessments will be similar to those of critics. If not, both the short-term and long-term impacts of using critics’ ratings in marketing efforts are questionable.

**EVALUATING FILM QUALITY**

Although no pure measures of a movie’s creativity have been created, a significant amount of research has focused on identifying existing, proxy measures. Simonton (2002, 2004c) has made progress in defining and refining select criteria such as awards and reviews, but a much larger range of quantitative data, such as critics’ and fan Web site users’ ratings, are available for use by researchers. A major implication of this work is the examination of critic and layperson ratings of movies. Marketing efforts are obviously most effective when they target characteristics of the creative product that consumers consider when they decide whether to purchase the product. If layperson ratings diverge sharply from those of critics, the use of professional film critic reviews in marketing efforts would appear to be less than ideal. And if data provide evidence that layperson ratings are similar to critics’ evaluations under certain conditions (e.g., romantic comedies without major stars released mid-winter), marketing for specific films released under those same conditions can be tailored accordingly.

The research on critic and layperson agreement on film quality is relatively thin. Boor (1990, 1992) found evidence that inter-rater agreement for critics was high (ranging from around 0.50 to 0.75), with a slightly lower range of correlations for critics and non-critics. Holbrook (1999) found a correlation of only 0.25 between critic and non-critic ratings. Yet somewhat surprisingly, Plucker, Holden, and Neustadter (in press) found evidence that professional critics’ ratings did not differ significantly from those of laypeople. But the possibility exists that the layperson ratings used in that study—i.e., ratings drawn from users of popular Web sites devoted to movies—may not truly be laypeople. They could include, for example, ratings from people who work in the film industry, ratings from people who are so devoted to movies that their opinions and aesthetic tastes have begun to mirror those of critics, ratings from non-critics who use the Internet a great deal, or perhaps all of the above.

A similar controversy exists in other areas of creativity research. In the fine arts, the opinions of non-professional artists about their own work were more accurately reflected by their artistic peers than by the judgments of professional artists (but not, it should be noted, art critics) (Runco, McCarthy, & Svenson, 1994). In creative writing, Kaufman, Gentile, and Baer (2005) found evidence of a high correlation between expert ratings and the ratings of gifted novices. However, the same research team, studying expert and non-gifted novice judgments of creativity in poetry, found only a moderate degree of agreement between expert and non-expert judges, with lower ratings by the expert judges (Kaufman et al., 2008). Indeed, when the number of raters was adjusted, the non-expert raters also showed a lower amount of agreement among themselves than the experts.

---

1 Although we acknowledge that one use of critics’ opinions in film marketing is to get people to buy tickets on opening weekend, regardless of whether laypeople are likely to agree with the critics, this strategy is only effective if laypeople believe that their assessments will be similar to those of critics. If not, both the short-term and long-term impacts of using critics’ ratings in marketing efforts are questionable.
No solid consensus exists about the relationship among expert and non-expert evaluators of creative products, whether in the fine arts, creative writing, or movies. As a result, within the context of film, the degree to which critics’ ratings are valid and influential measures of the popular opinion of cinematic creativity has not been determined. The purpose of this paper is to extend the research of Plucker, Holden, and Neustadter (2008) and compare the ratings of critics, self-proclaimed novices (who may have a greater interest in movies than most novices), and true novices (i.e., college students) who are major consumers of movies.

METHOD

Measures and Data Sources
The primary sample for this study included all films that played on at least 1,000 screens in the United States during the calendar years of 2001–2005. This resulted in a sample of 680 films.

Three types of ratings sources were utilized. First, ratings for professional critics were drawn from publicly available data on metacritic.com. Metacritic ratings are a normalized, weighted average of critics’ reviews from 42 major publications (newspapers, weekly and monthly magazines, and Web sites), reported on a 0–100 scale. The metacritic database includes ratings data on nearly every movie released in the United States since 1999.

Second, ratings for self-defined novices were collected from publicly available data on boxofficemojo.com and imdb.com. These sites were chosen based on the recommendation of Plucker, Holden, and Neustadter (2008) due primarily to the psychometric quality of the available data. We consider these ratings to come from “self-defined” novices because little is known about the users of each site that choose to submit their ratings: They could be a broad cross section of moviegoers, or they could be a highly self-selected sample, such as people working in the film industry itself.

Third, ratings for movies were collected from 169 students at a large public university in the southwestern United States. The university is mid-sized (approximately 14,000 students), and serves a student body that comprises a substantial portion of nontraditional students. The mean age of the student body is 25.9 years. The student population is ethnically diverse (32.75% Hispanic, 11.74, African American, 4% Asian-Pacific Islander), and the university is recognized by the Department of Education as a Hispanic-serving institution.

Data Analyses
The data were analyzed in three steps. First, means and standard deviations were calculated for all four groups of raters. Analysis of variance was used to examine differences among the mean ratings of the four groups. Second, a correlation matrix for the four sets of ratings was created and examined. Third, the student ratings were calculated separately for students who saw a low, average, or high numbers of movies. These separate student means were compared to those of the critics and self-defined novices.
RESULTS

Descriptive statistics from the four groups of raters appear in Table 1. Ratings were converted to a 0–10 scale. The data included in Table 1 suggest that significant differences existed among the four groups’ ratings, and a subsequent ANOVA \((F(2.04) = 352.98, p < 0.001, \text{partial } \eta^2 = 0.40 \text{ (Greenhouse-Geisser-adjusted)})\), provided evidence of this difference. The critics’ ratings were significantly lower than those for the three other groups, perhaps as large as a full-pooled standard deviation. However, the lack of mean differences among the student and two novice groups does not necessarily mean that students and novices rated movies similarly. For this reason, Pearson product–moment correlations were calculated for the complete set of ratings (Table 2). The correlation between critic and student ratings was moderate in magnitude but much lower than the correlations between critics and novices. Indeed, the magnitude of correlations between critics and novices was at the high end of the range noted in previous research for agreement among critics.

Student Ratings by Number of Movies Seen

To investigate whether a “familiarity effect” influences the correlations among ratings groups, we split the student ratings into three new groups: low, average, and high exposure to films. The number of movies seen by average student raters was between \(-1\) and \(+1\) standard deviations of the mean number of movies viewed by the sample. In the same vein, the number of films seen by the low group of student raters was more than one standard deviation below the mean; the converse was true for the high group of student raters.
Table 3. Mean Movie Ratings of Students by Number of Movies Rated.

<table>
<thead>
<tr>
<th>Rating Source</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students low</td>
<td>7.30</td>
<td>1.63</td>
</tr>
<tr>
<td>Students average</td>
<td>6.69</td>
<td>0.93</td>
</tr>
<tr>
<td>Students high</td>
<td>6.36</td>
<td>1.04</td>
</tr>
</tbody>
</table>

*Note: Films rated on a scale of 1 to 10, 10 being highest.*

Table 4. Correlations Between Movie Ratings of Critics, Novices, and Students by Number of Movies Rated.

<table>
<thead>
<tr>
<th>Rating Source</th>
<th>Critics</th>
<th>Novice Critics (IMDB)</th>
<th>Novice Critics (BOM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students low</td>
<td>0.22**</td>
<td>0.30**</td>
<td>0.29**</td>
</tr>
<tr>
<td>Students average</td>
<td>0.39**</td>
<td>0.63**</td>
<td>0.62**</td>
</tr>
<tr>
<td>Students high</td>
<td>0.44**</td>
<td>0.58**</td>
<td>0.59**</td>
</tr>
</tbody>
</table>

*Note: Students low = ratings from undergraduates who saw +1 SD fewer movies than average; Students average = ratings from undergraduates who saw between -1 SD and +1 SD movies; Students high = ratings from undergraduate who saw +1 SD more movies than average.

**p < 0.01.

The means and standard deviations for the three groups are presented in Table 3. An analysis of variance provided evidence that the scores were significantly different [$F(1.32) = 103.94$, $p < 0.001$, partial eta squared = 0.19 (Greenhouse-Geisser-adjusted)], with the low group of students having significantly higher ratings (and greater variability within those ratings). Ratings for the average and high student groups were roughly a standard deviation lower than the low student group but still higher than for novices and critics (see Table 1).

Table 4 includes correlations between the three student groups’ ratings and the critic and novice ratings used earlier. The low group of students had low correlations with all three non-student groups, yet the average and high groups had moderate correlations with the critics’ and novices’ ratings. Interestingly, the correlations for the average and high groups of students were very similar to those of the entire student group included in Table 2.

**DISCUSSION**

The results of this study indicate that the novice ratings used in studies such as Plucker, Holden, and Neustadter (2008) are indeed different from those of laypeople—but they are also different from the ratings of professional critics. In general, critics have lower ratings than non-critics (i.e., novices submitting ratings to film Web sites), who in turn have lower ratings than college students. College students who see relatively few movies appear to have especially high ratings and, predictably, greater variation in their ratings compared to the other groups.
We find it especially interesting that the film Web site users appear to be different from both critics and college students. In a sense, the Web site users are firmly in between the two other groups: Their mean ratings roughly split the difference between the mean ratings of the other two groups, and their ratings correlate with critics at a magnitude \( r > 0.70 \) that is on the high end of how critics tend to correlate with each other (see Boor, 1990, 1992). In this sense, they are truly “novice critics”—not quite laypeople, but not quite professional critics, either.

We would argue that the Web site novice critics are akin to the gifted novices tested by Kaufman, Baer, and Gentile (2004) as opposed to true novices. Indeed, most of the users of imdb.com (for example) have much more experience and interest in movies than a typical novice. Many aspiring actors become members of the site hoping to network with industry professionals. It is impossible to determine how many of the Web site novice ratings were actually given by experts.

Given these findings and the past research by Kaufman, Baer, Runco, and others, it seems that there truly is some type of difference between novice and expert ratings, but the key is determining who is genuinely an expert and who is genuinely a novice. “True” novices—college students who are very unlikely to work in the film industry—were not only significantly different from professional critics (a clear example of “expert” raters) but were also different from supposed novices from movie Web sites. We wonder whether it may be beneficial to approach the novice–expert Consensual Assessment Technique ratings question as a continuum. Perhaps instead of strict categories such as “novice” and “expert,” we need more labels that reflect the vast diversity of experience that can be found for nearly all domains. Future research on the marketing of movies should recognize the continuum of evaluation expertise that exists in the available film data.

One model of creativity that includes a developmental trajectory is Beghetto and Kaufman’s (2007) Four-C model. This model traces creativity from the personal discoveries of mini-c to the everyday innovations of little-c to the professional accomplishments of Pro-c to the eminent genius of Big-C. A similar arc of expertise categories may be needed. Just as Beghetto and Kaufman argue that creativity is too diffuse to be constrained into the little-c and Big-C labels, so too may expertise be unable to be pigeonholed into “novices” and “experts.”

Viewing expertise—or, perhaps more accurately, experience evaluating the quality of creative products—on a continuum has other advantages. For example, Csikszentmihalyi (1988), in his popular model of creativity, talks about the role of creative gatekeepers who help decide what products within a specific field are to be judged as being creative. He allows for contexts in which the stature of the gatekeepers may make it easier or harder for a creative product to be accepted, and the results of the present study suggest that the role of “gatekeeper” be viewed broadly, at least in the case of film.

For example, the continuum implies that a gatekeeper for one person may be a well-known critic; for another, novice critics on the most popular film Web sites; for yet another, their next-door neighbor or best friend. The continuum perspective also suggests that gatekeepers may occupy hybrid positions between categories, such as the case of a novice critic who starts her or his own movie blog, then adds a weekly review column on a film Web site, then begins to review occasionally for a local paper, et cetera. This person is clearly moving along the continuum, primarily in the gray areas that separate laypeople from novice critics.
from professional critics. We anticipate this sort of movement to occur increasingly with the explosion in avenues for expression provided by information technology. Indeed, with the new avenues for communication provided by technological advances, these transitory experts may be the “tastemakers” or, as Gladwell (2000) describes them, the Mavens, Connectors, and Salespeople that marketers seek to reach.

Limitations

This study has a number of important limitations. First and foremost, the sample chosen to represent laypeople consisted of a very specific student population that is probably more diverse than typical college students. At the same time, research with college students often involves samples lacking in diversity, so the present sample is not a major weakness—but it does reinforce the need to replicate these results with different student and layperson populations (i.e., non-students).

Second, although we collected novice critic ratings based on recommendations from previous research, other sites and data collection strategies may produce different patterns of results. Third, data were collected at one point in time, at least 1–2 years after each film was released in the theaters. For this reason, we could not control for when a movie was viewed (for all subsamples) or when a rating was published or released (for the professional and novice critics). Future research should consider the impact of time on the nature of movie ratings.

REFERENCES


The authors appreciate the assistance of Gary Zhou, Haiying Long, and Minjing Duan, all of whom helped with gathering extant data and constructing the final database for the study.

Correspondence regarding this article should be sent to: Jonathan A. Plucker, Center for Evaluation and Education Policy, 1900 E. 10th Street, Bloomington, IN 47406 (jplucker@indiana.edu).