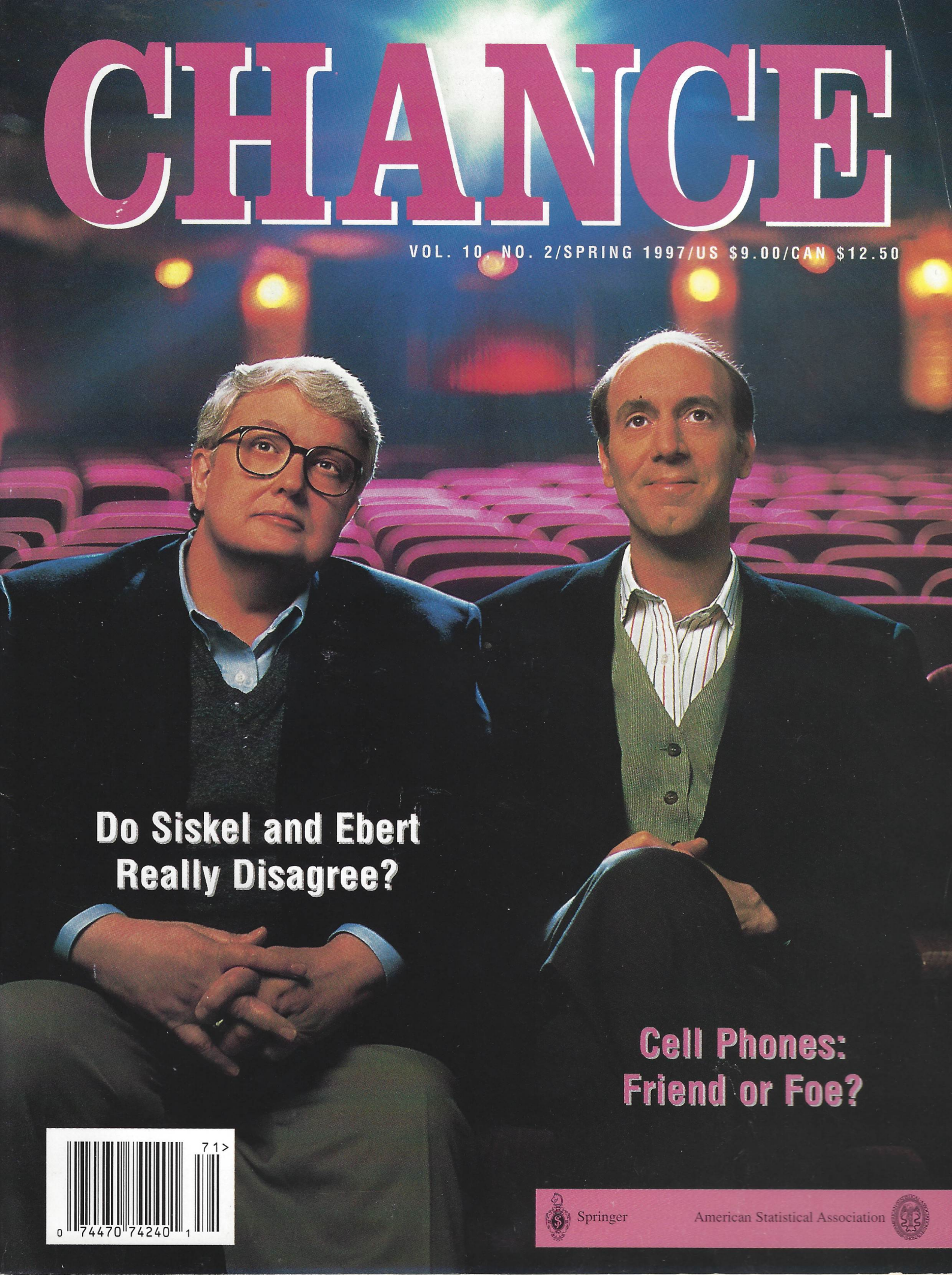
CHANCE

VOL. 10, NO. 2/SPRING 1997/US $9.00/CAN $12.50



Do Siskel and Ebert Really Disagree?

Cell Phones: Friend or Foe?

71>
0 74470 74240 1

Springer

American Statistical Association

Gene Siskel and Roger Ebert entertain us with their high-spirited debates, but how much do they—and other movie reviewers—really disagree?

Evaluating Agreement and Disagreement Among Movie Reviewers

Alan Agresti and Larry Winner

Thumbs up, or thumbs down? Two reviewers face each other across a theater aisle, arguing—sometimes forcefully—the merits and faults of the latest film releases.

This is the entertaining—and often imitated—format that Chicago's Gene Siskel and Roger Ebert originated some 20 years ago with a local television program in their home city. Siskel and Ebert's growing popularity led to their *Sneak Previews* program on PBS and later their syndicated show, currently distributed by Buena Vista Television, Inc.

In their day jobs, Siskel and Ebert are rival movie critics at the *Chicago Tribune* and the *Chicago Sun-Times*, respectively. They highlight this friendly rivalry in their on-camera face-offs, often creating the impression that they strongly disagree about which movies deserve your entertainment time and dollars.

But how strongly do they really disagree? In this article, we'll study this question. We'll also compare their patterns of agreement and disagreement to those of Michael Medved and Jeffrey Lyons, the reviewers on *Sneak Previews* between 1985 and fall 1996. We then look at whether the degree of disagreement between Siskel and Ebert and between Medved and Lyons is typical of movie reviewers by evaluating agreement and disagreement for the 28 pairs of eight popular movie reviewers.

The Database

Each week in an article titled "Crix' Picks," *Variety* magazine summarizes reviews of new movies by critics in New York, Los Angeles, Washington, DC, Chicago, and London. Each review is categorized as Pro, Con, or Mixed, according to whether the overall evaluation is positive, negative, or a mixture of the two.

We constructed a database using reviews of movies for the period April 1995 through September 1996. The database contains the *Variety* ratings for these critics as well as some explanatory variables for the movies, discussed later, that could influence the ratings.

Summarizing the Siskel and Ebert Ratings

Table 1 shows the ratings by Siskel and Ebert of the 160 movies they both reviewed during the study period. This square contingency table shows the counts of the nine possible combinations of ratings. For instance, for 24 of the 160 movies, both Siskel and Ebert gave a Con rating, "thumbs down." The 24 + 13 + 64 = 101 observations on the main diagonal of Table 1 are the movies for which they agreed, giving the same rating. Their agreement rate was 63% (i.e., 101/160). Fig. 1 portrays the counts in Table 1.

To achieve perfect agreement, all observations would need to fall on the main diagonal. Fig. 2 portrays corresponding ratings having perfect agreement. When there is perfect agreement, the number of observations in each category is the same for both reviewers. That is, the row marginal percentages are the same as the corre-

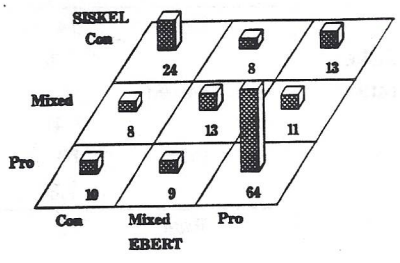


Figure 1. Movie ratings for Siskel and Ebert.

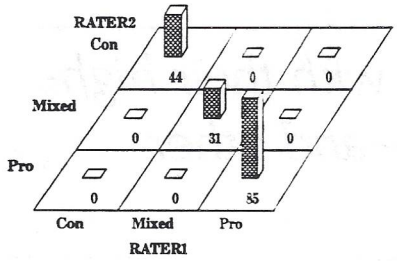


Figure 2. Movie ratings showing perfect agreement.

sponding column marginal percentages, and the table satisfies *marginal homogeneity*. In Table 1, the relative frequencies of the ratings (Pro, Mixed, Con) were (52%, 20%, 28%) for Siskel and (55%, 19%, 26%) for Ebert. Though they are not identical, the percentage of times that each of the three ratings occurred is similar for the two raters. There is not a tendency for Siskel or Ebert to be easier or tougher than the other in his ratings. If this were not true, it would be more difficult to achieve decent agreement. If one reviewer tends to give tougher reviews than the other, the agreement may be weak even if the statistical association is strong between the reviewers.

The agreement in Table 1 seems fairly good, better than we might have expected. In particular, the two largest counts occur in cells where both Siskel and Ebert gave Pro ratings or they both gave Con ratings. If the ratings had been statistically independent, however, a certain amount of agreement would have occurred simply "by chance." The cell frequencies expected under this condition are shown in parentheses in Table 1. These are the expected frequencies for the Pearson chi-squared test of independence for a contingency table. If Siskel's and Ebert's ratings had no association, we would still expect agreement in (11.8 + 6.0 + 45.6) = 63.4 of their evaluations (39.6% agreement rate). The observed counts are larger than the expected counts on the main diagonal and smaller off that diagonal, reflecting better than expected agreement and less than expected disagreement.

Of course, having agreement that is better than chance agreement is no great accomplishment, and the *strength* of that agreement is more relevant. Where does the Siskel and Ebert agreement fall on the spectrum ranging from statistical independence to perfect agreement?

A popular measure for summarizing agreement with categorical scales is *Cohen's kappa*. It equals the difference between the observed number of agreements and the number expected by chance (i.e., if the ratings were statistically independent), divided by the maximum possible value of that difference. For the 160 observations with 101 agreements and 63.4 expected agreements in Table 1, for instance, sample kappa compares the difference 101 − 63.4 = 37.6 to the maximum possible value of 160 − 63.4 = 96.6, equaling 37.6/96.6 = .389. The sample difference between the observed agreement and the agreement expected under independence is 39% of the maximum possible difference. Kappa equals 0 when the ratings are statistically independent and equals 1 when there is perfect agreement. According to this measure, the agreement between Siskel and Ebert is not impressive, being moderate at best.

Table 1—Ratings of 160 Movies by Gene Siskel and Roger Ebert, with Expected Frequencies in Parentheses for Statistical Independence

		Ebert rating			
		Con	Mixed	Pro	Total
Siskel rating	Con	24 (11.8)	8 (8.4)	13 (24.8)	45
	Mixed	8 (8.4)	13 (6.0)	11 (17.6)	32
	Pro	10 (21.8)	9 (15.6)	64 (45.6)	83
	Total	42	30	88	160

Source: Data taken from *Variety*, April 1995 through September 1996.

The rating scale (Pro, Mixed, Con) is ordinal, and Cohen's kappa does not take into account the severity of disagreement. A disagreement in which Siskel's rating is Pro and Ebert's is Con is treated no differently than one in which Siskel's rating is Pro and Ebert's is Mixed. A generalization of kappa, called *weighted kappa*, is designed for ordinal scales and places more weight on disagreements that are more severe. For Table 1, weighted kappa equals .427, which is also not especially strong.

Symmetric Disagreement Structure

Table 1 is consistent with an unusually simple disagreement structure. The counts are roughly symmetric about the main diagonal. For each of the three pairs of categories (x, y) for which the raters disagree, the number of times that Siskel's rating is x and Ebert's is y is about the same as the number of times that Siskel's rating is y and Ebert's is x.

The model of *symmetry* for square contingency tables states that the probability of each pair of ratings (x, y) for (Siskel, Ebert) is the same as the probability of the reversed pair of ratings (y, x). In fact, this model fits Table 1 well. The symmetry model has cell expected frequencies that average the pairs of counts that fall across the main diagonal from each other. For instance, the expected frequencies are (13 + 10)/2 = 11.5 for the two cells in which one rating is Pro and the other is Con. The Pearson chi-squared statistic for testing the fit of the symmetry model is the sum of (observed–expected)^2/expected for the six cells corresponding to the three disagreement pairs. It equals .59, based on df = 3, showing that the data are consistent with the hypothesis of symmetry ($P = .90$).

Whenever a square contingency table satisfies symmetry, it also satisfies marginal homogeneity. The counts in Table 1 are within the limits of sampling error both for the conditions of symmetry and marginal homogeneity. Nonetheless, both these conditions can occur even if the ratings show weak agreement or are statistically independent, so we next focus on the kappa measures for summarizing strength of agreement.

Agreement for Other Raters

Our summary of Table 1 using kappa seems to confirm what viewers see on television—that Siskel's and Ebert's level of agreement is not especially strong. But how does it compare to the agreement between other movie reviewers? We next study Table 2, which shows joint ratings of the most recent *Sneak Previews* reviewers, Michael Medved and Jeffrey Lyons. One difference we notice immediately is that the ratings are not symmetric. For instance, there are many more cases in which Lyons's rating is Pro and Medved's is Con than the reverse. The marginal row and column totals suggest that Medved is a much more critical reviewer, tending to give more negative reviews than Lyons.

As already mentioned, strong agreement is difficult to achieve when the overall distributions of ratings differ substantially between two reviewers. For Table 2, in fact, weighted kappa is only .204, about half as large as the value of .427 that Siskel and Ebert share. The values are not directly comparable, because the movies being rated are not identical for the two tables and because kappa has some dependence on the marginal counts, but these data certainly suggest stronger agreement between Siskel and Ebert than between Medved and Lyons. (If weighted kappa values of .427 and .204 were based on independent samples, the difference of .223 would have a standard error of .096; these samples are partially dependent, but the actual standard error is likely to be similar.)

Fig. 3 portrays the mediocre agreement between Medved and Lyons. Notice the large incidence of severe disagreements in which Medved's rating is Con while Lyons's rating is Pro. Except for the spike in the cell where both ratings are Con, Medved's rating is essentially independent of Lyons's. In fact, the model of independence applied only to the other eight cells in Table 2 (which is a special case of a *quasi-independence model*) fits well. Its chi-squared goodness-of-fit statistic equals .8 with df = 3.

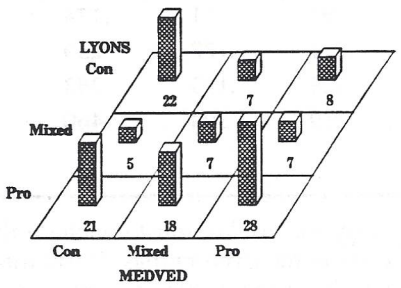


Figure 3. Movie ratings for Medved and Lyons.

We mention this statistic only as an informal index, because the figure motivated this model selection. Nonetheless, conditional on Medved's and Lyons's ratings falling in a cell other than (Con, Con), it seems plausible that their ratings are statistically independent. In these cases, their agreement is no better than if they were each randomly rating the movies without even watching them!

Besides exhibiting poor agreement between themselves, neither Lyons nor Medved show much agreement with Siskel or Ebert. The weighted kappa values are .229 between Medved and Siskel, .178 between Medved and Ebert, .267 between Lyons and Siskel, and .209 between Lyons and Ebert. The Siskel and Ebert agreement looks better all the time!

Variety reports ratings for several reviewers, so we next analyzed how the Siskel and Ebert agreement compares to agreement among other popular reviewers. Table 3 is a matrix of kappa and weighted kappa values for eight reviewers, the four already mentioned as well as Peter Travers of *Rolling Stone*, Rex Reed of *New York Observer*, Gene Shalit of *The Today Show* (NBC), and Joel Siegel of *Good Morning America* (ABC). Of the 28 pairs of reviewers, the Siskel and Ebert agreement is the strongest, according to either agreement measure. For instance, the next strongest kappa after the value of .389 between Siskel and Ebert is .283 between Shalit and Travers. Though the values are not entirely comparable, the overwhelming impression one gets from this table is that many reviewers simply don't agree much more strongly than they would if they were randomly and blindly making their ratings.

In Table 3, one reviewer stands out from the others. Michael Medved has very poor agreement with all the other raters. For two raters (Siegel and Shalit), in fact, his agreement with them is not even significantly different from chance agreement. This is not surprising to the film buff, because Medved is a maverick among reviewers who has been outspoken in criticizing much of what Hollywood does. See, for instance, Medved (1992). He recently argued that "Gratuitous violence, reckless sex, gutter language and a hatred for organized religion in movies and on television contribute to a general climate of violence, fear and self-indulgence" (*The Salt Lake Tribune*, June 28, 1996).

The presence of these characteristics in a film may cause Medved to give a negative rating when other reviewers give a positive rating because they consider the movie to be well-made and interesting. Examples may be *Leaving Las Vegas*, *Pulp Fiction*, *Seven*, and *Trainspotting*, for which Medved disagreed with nearly all movie reviewers. But it is also worth noting that Medved was the only reviewer in our study who gave the Con rating to Disney's *Pocahontas* and *The Hunchback of Notre Dame*—two animated films that do not so obviously display the qualities to which he reacts so negatively.

Table 2—Ratings of 123 Movies by Michael Medved and Jeffrey Lyons

		Medved rating			
		Con	Mixed	Pro	Total
Lyons rating	Con	22	7	8	37
	Mixed	5	7	7	19
	Pro	21	18	28	67
	Total	48	32	43	123

Source: Data taken from *Variety*, April 1995 through September 1996.

Table 3—Matrix of Agreement Indexes for Eight Movie Reviewers and a Consensus Rating, With Weighted Kappa Above Main Diagonal and Cohen's Kappa Below Main Diagonal

	Ebert	Lyons	Medved	Reed	Shalit	Siegel	Siskel	Travers	Consensus
Ebert	1.0	.209	.178	.200	.361	.224	.427	.210	.431
Lyons	.182	1.0	.204	.330	.183	.285	.267	.178	.404
Medved	.140	.177	1.0	.210	.081	.103	.229	.161	.356
Reed	.138	.259	.180	1.0	.295	.226	.250	.215	.403
Shalit	.232	.110	.033	.215	1.0	.217	.305	.374	.552
Siegel	.218	.224	.081	.211	.174	1.0	.227	.348	.410
Siskel	.389	.240	.211	.177	.239	.176	1.0	.246	.499
Travers	.123	.124	.129	.143	.283	.275	.170	1.0	.475
Consensus	.315	.284	.270	.304	.460	.303	.387	.383	1.0

Who's the Toughest, Who's the Easiest?

Table 4 shows the distribution of ratings across the three categories for each of the eight reviewers. This table reveals one reason why the agreement tends to be, at best, moderate. The reviewers varied considerably in their propensity to assign Pro and Con ratings. For instance, Travers gave a Pro rating only 28.0% of the time, whereas Siegel gave it 56.0% of the time.

Nonetheless, even given this variability in ratings' distributions, the agreement shown in Table 3 is quite unimpressive. For instance, raters with the margins displayed in Table 2 for Lyons and Medved have the potential for a weighted kappa as high as .707, much larger than the observed value of .204. This would occur for the counts portrayed in Fig. 4, which show the greatest agreement possible for the given margins. We must conclude that, even for the given ratings distributions, ratings among movie reviewers show weak agreement.

Table 4—Marginal Distributions of Ratings by the Eight Reviewers

Reviewer	Pro	Mixed	Con
Ebert	55.0%	18.8%	26.2%
Lyons	52.0%	16.3%	31.7%
Medved	35.5%	25.7%	38.8%
Reed	42.6%	17.9%	39.5%
Shalit	49.1%	26.3%	24.6%
Siegel	56.0%	15.7%	28.4%
Siskel	51.9%	20.0%	28.1%
Travers	28.0%	34.8%	37.1%
Consensus	27.2%	56.4%	16.4%

Agreement With a Consensus Rating

For each movie rated, we also noted the consensus rating. We define this to be the rating nearest the mean of the reviewers' ratings, based on equally spaced scores for the three response categories. As Table 4 shows, the consensus rating is much more likely to be Mixed than is the rating by any particular reviewer. How strong is the agreement between each reviewer and this consensus rating?

Table 3 also shows the kappa and weighted kappa values between each reviewer and the consensus. Not surprisingly, the reviewers tended to agree more strongly with the consensus than with other reviewers. For each reviewer, the kappa and weighted kappa values with the consensus exceed the corresponding values with every other reviewer, the only exception being kappa with Siskel and Ebert. The strength of agreement is still not exceptionally strong, however. The agreement tends to be slightly weaker yet if we form kappa between a reviewer and the consensus of the other reviewers (i.e., excluding that reviewer in determining the consensus).

What Causes Disagreement?

We next studied whether certain explanatory variables suggested reasons for the weak agreement. We added three indicator explanatory variables representing our judgment about whether the film contained large amounts of violence, contained large amounts of sex, or was strong in promoting family values. We also recorded a film's studio, genre (action/adventure, comedy, drama, etc.), and rating by the Motion Picture Association of America (MPAA). Perhaps surprisingly, none of these factors showed much association with observed disagreement between reviewers. Partly this may reflect the relatively small number of movies in this sample judged to have large amounts of sex or violence.

We also studied the effects of these factors on the ratings themselves by the various reviewers. The only noticeable effects concerned high sexual content and the ratings by Medved and Lyons. For instance, for movies with high sexual content, the percentage of Con ratings was 78% for Medved and 62% for Lyons, compared to Con percentages of 37% by Medved and 31% by Lyons for other movies. By contrast, the Con percentages for movies with high sexual content versus other movies were 33% and 30% for Siskel and 41% and 25% for Ebert.

So, what film qualities help determine a reviewer's or other movie-goer's rating? It's not possible to draw conclusions based on our limited sample, but some research has been conducted in the fields of marketing and management to quantify differences among individuals in movie enjoyment. Eliashberg and Sawhney (1994) attempted to predict enjoyment of hedonic experiences in general, and movie enjoyment in particular. Their variables used in prediction were

- Temporary moods of individuals
  - –Arousal (low, high)
  - –Pleasure (low, high)
- Individual sensation-seeking tendency (ordinal scale—four levels)
- Individual-specific mood transition variability (times spent in mood states)
- Movie-specific emotional content (varies from scene to scene)
  - –Arousal (low, high)
  - –Pleasure (low, high)

Although their model, at best, had limited success at predicting the enjoyments (on two scales) of 36 individuals watching a single made-for-HBO movie, their theoretical framework seems reasonable and provides possible reasons for differences in ratings among viewers of the same movie. For instance, violent and sex-themed

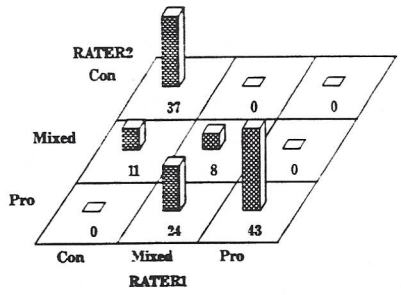


Figure 4. Movie ratings showing the strongest possible agreement, given the margins exhibited by Medved and Lyons.

movies might be more popular among viewers with higher "sensation-seeking tendencies," and family fare may be more popular among viewers with lower "sensation-seeking tendencies." Of course, initial mood state and mood transition rates could affect ratings as well. Unfortunately, there are no means of measuring these effects with our current database, and obtaining reliable and fair measurement would probably require replication, which would be impossible.

In our opinion, there are likely to be many very diverse reasons for the potential enjoyment of a movie, most of which have small effects. That's probably for the best, or else Hollywood would use a standard regression formula and make movies that are even more predictable than they currently are (anyone for *Rocky/Rambo n*?).

The effects of whatever factors do influence ratings probably differ between professional movie reviewers and the general movie-going public. One might well be more successful using such factors to predict movie enjoyment for the general public than for critics. For instance, individuals are likely to form specific preferences based on genre or degree of sex or violence in a movie, whereas most critics aim to assess the overall quality of the film, regardless of such characteristics. A result is that variation in levels of agreement among members of the movie-going population is probably at least as great as among movie reviewers.

Based on the weak agreements found in this article among "expert" reviewers, we should not be surprised to find ourselves disagreeing with the assessment of any particular reviewer. But over time, our individual experience may help us to determine with which reviewer we tend to agree most strongly. For these two authors, Agresti listens most closely to Gene Siskel (weighted kappa = .507) and Winner listens most closely to Roger Ebert (weighted kappa = .475). See you at the movies.

[The authors thank Jacki Levine, Assistant Managing Editor of *The Gainesville Sun*, for many helpful comments.]

Disagreements Among Movie Raters

Disagreement among a group of professional movie reviewers is not uncommon. For the movies in this database for which Siskel, Ebert, Lyons, and Medved all provided ratings, only 21.3% had the same rating by all four reviewers. Here are examples of movie ratings in which disagreements occurred, often between Medved and the others (P = Pro, M = Mixed, and C = Con):

MOVIE	Siskel	Ebert	Lyons	Medved
Nixon	P	P	P	C
Forget Paris	P	P	P	C
The Cable Guy	P	M	C	P
Pocahantas	P	P	P	C
Dangerous Minds	C	C	P	C
Seven	P	P	P	C
Leaving Las Vegas	P	P	C	C
Waterworld	C	M	C	C
Bridges of Madison County	P	P	P	C
Jumanji	C	C	P	P
Hunchback of Notre Dame	P	P	P	C
How to Make an American Quilt	M	C	P	C

References and Further Reading

Eliashberg, J., and Sawhney, M. (1994), "Modeling Goes to Hollywood: Predicting Individual Differences in Movie Enjoyment," *Management Science*, 40, 1151–1173.

Medved, M. (1992), *Hollywood versus America: Popular Culture and the War on Traditional Values*, New York: Harper Collins.

Spitzer, R. L., Cohen, J., Fleiss, J. L., and Endicott, J. (1967), "Quantification of Agreement in Psychiatric Diagnosis," *Archives of General Psychiatry*, 17, 83–87.