Article title: By Unanimous Decision? A Second Look At Consensus In The Film Industry.

Author names:

Erwin Dekker

Erasmus University Rotterdam

Erasmus School of History, Culture & Communication

Postbus 1738

3000 DR Rotterdam

The Netherlands

e.dekker@eshcc.eur.nl

010-4082460

Request for reprints should be addressed to Erwin Dekker

Zuzanna Popik

Erasmus University Rotterdam

Erasmus School of History, Culture & Communication

Running head: Consensus in the film industry reconsidered

Acknowledgements:

We are grateful for prof. Simonton for sharing his original dataset with us and the helpful comments of the participants of the Econ & Culture Seminar at the Erasmus School of History, Culture & Communication.

Submission date: 10th of February 2013

Abstract

This paper analyzes the verdicts of various film organizations that annually present awards to motion pictures and investigates whether they award/nominate the same movies in a given year. This research disputes previous findings which reported a high level of agreement between those juries, by the means of reliability analysis and the Cronbach's Alpha composite. Arguments were raised for why these earlier findings were flawed and why the use of Cronbach's Alpha is problematic. Different aspects of consensus are discussed after which a new measure (β) is introduced. This is followed by a detailed comparison between particular juries with regard to the percentage share of their decisions that award the most successful (chosen by multiple other juries as well) and the least successful (uniquely awarded) films. This measure shows how often a singular jury decides in line with the others and how much does it stray from the consensus.

The paper also broadens the theoretical discussions about the reasons for (not) expecting a consensus to arise between various expert juries. It argues that by adopting a cultural economic perspective we become aware of various reasons, most importantly competition between the award events and the juries tend towards a lower level of consensus.

Keywords: Consensus, Awards, Cronbach's Alpha, Movie industry, Oscars

By Unanimous Decision? A Second Look At Consensus In The Film Industry.

Recently there has been an interesting extension of cultural economic research to look at quality in a quantitative way. And while this sounds like an oxymoron, promising progress has been made in this way, especially by researchers who have been careful enough to realize that what they are really after is the measurement of the level of intersubjective consensus, rather than a conclusive settlement of what objectively constitutes quality.

The qualitative approach to quality in arts traditionally comprises descriptions of the art works with regards to some accepted scheme of values and characteristics. In motion pictures the qualities described by the experts pertain usually to the particular work of filmmakers involved: a film is judged by its script, acting, directing, costumes, or other artistic contributions. The quantitative approach to the quality of movies is based on counting success indicators like box office performance and other earnings (hence, consumers approval), critics' opinions (expressed in the amount of either stars awarded or publications that mention a title) and the peer or industry acclaim - expressed in cinematic awards. This is true also for other sectors like books or music. Numerous studies explore the relationship between the three indicators: critical acclaim to earnings (Gemser, Van Oostrum & Leenders, 2007), types of awards to earnings (Gemser, Leenders & Wijnberg, 2008), previous awards - "artist track record" to consecutive earnings (Hadida, 2010), annual awards to long-term recognition (Ginsburgh, 2003) and other combinations.

But an important step to be made before such relationships can be meaningfully interpreted is to establish to what extent experts and peers actually agree on quality; to what extent there is a consensus between various award juries and other experts in a particular cultural industry. In this paper the degree of consensus between film juries was examined. This paper is not the first attempt to do so, especially Simonton (2004, 2011) has done work on it before. But it will be argued that the methods he had used are inappropriate to study the consensus between film juries, mainly due to limitations of the data. Then various alternative methods which could be used to study the level of consensus between various film juries will be proposed, and their relative merits will be discussed. The paper will

therefore devote substantial attention to methodological issues, as well as present the results of these types of measurements for the movie industry. Since the award system in the movie industry, nominations or shortlists, from which the eventual winners are chosen is not unique, these methods can also be used to study consensus in different cultural industries.

The following sections of this paper will be first of all devoted to the discussion why we might (not) expect a consensus to form, after which a detailed examination of the data and methods will proceed, and finally alternative methods to measure this level of consensus, including a measure specifically developed for this type of inter-jury data will be introduced.

AWARDS AND JURIES IN MOTION PICTURES

The research on award-granting juries is interesting for various reasons. First of all, answering questions about the accordance of expert opinion allows one to establish whether experts generally agree on quality and contributes to the underlying debate on the convergence or divergence of taste of individuals, as discussed in Blaug's overview article (Blaug, 2001). Especially since these experts provide one of the most important signals about quality of products in the creative industries which are characterized by great uncertainty and risks (Caves, 2000; Wijnberg, 2003).

Secondly, this effort creates a clearer picture of the role of various experts and gatekeepers in the cultural sector (Ginsburgh, 2003). Their relations or interdependency can also be institutionally important as shown by a recent proposal by the Academy of Motion Picture Arts and Sciences. The Academy has announced, given the current overwhelming number of submissions, that for a documentary to be considered, it has at least to be reviewed in either the New York Times or the Los Angeles Times. This makes winning an award for a documentary dependent on recognition by other gatekeepers, and thus creating an institutional relationship between various critics.

The motion picture sector is particularly interesting when it comes to available expert opinions, awards and judgments. This is because of the number of organizations and bodies that independently assess the quality of movies. Before being released to the public, movies enter various

film festival competitions judged by professional juries. After their theatrical premieres they are considered by critics or journalists associations and award-granting professional organizations. Out of these the Oscars, Golden Globes and BAFTAs are just a few, but probably the most renowned awards series. This leads to some seemingly straightforward questions: do these juries use the same standard of quality, do they similarly define achievements in filmmaking - do they choose the same or different movies as winners? Are they in competition or do they corroborate on a success of a single title? What are then the consequences for audiences and the perceived quality of the films?

Many factors contribute to what can be hypothesized about the relations between the juries and their verdicts. It is possible to view quality in arts as a type of constant. It is conceivable that people, especially those professionally involved in motion pictures and in possession of expert knowledge on the subject, share a certain view on best filmmaking; regardless of the organization they happen to be members of. This is also sustained by the fact that the juries in question use the same or similar phrases to describe what they award. The terms "achievement", "excellence" or "merit" are repeatedly used and they all signify similar if not the same things. At the same time such terms remain quite vague: no emphasis is given to a movie's topic, origin or component, therefore it could be expected that similar verdicts will be given out.

Another reason to expect a high level of consensus between the experts is the fact that the award season is a cycle of events and galas that traditionally proceed in a fixed order from autumn to early spring the following year. This would allow for contagion effects to occur, as mentioned by Pardoe and Simonton (2008). The experts themselves might trust and follow each other when it comes to naming winners. Winning an award generates a lot of publicity for the honored movie, especially in the industry publications, and it is not impossible that the film will enjoy the benefits of increased visibility also during the voting process of consequent award series. On the other hand, this effect could just as well be reversed (anti-contagion). Members of organization that name their choices later in the award season might strive to give original verdicts. The two effects might even take place simultaneously given the democratic nature of voting in most expert juries (the jury panels consists of many individuals and everyone's vote counts equally). It is clear that not in all cases the contagion or

anti-contagion effects have a chance to develop - some ballots have to be handed in before previous verdicts are announced. But most of this must remain speculative, for very few organizations reveal the details of their voting procedures with specific dates and time-frames, which is why detailed research on the occurrence or direction of such effects is impossible without their permission and collaboration.

Then, there are reasons not to expect a high level of agreement between these juries. If they were all just confirming each other's choices, neither the public nor the industry would have the need for more than one (or a couple of) award series. After losing the first competition filmmakers would have no hope of winning a different award and the thrill and surrounding buzz inherent to the announcement of the verdicts would greatly diminish. Moreover, some of the organizations in the motion picture industry have been created especially to balance out the others. For instance, the awards of the New York Critics Film Circle were initially established to function as a counterweight to the dominance of the Oscars, which were perceived as biased (Simonton, 2011). Such claims have been recurring (Wiley & Bona, 1996, Holden, 1993). The Academy was accused of giving verdicts swayed by the local (Hollywood) tastes, which would please the big movie studios.

Another, more economic reason for divergence in opinions between panels is that all award ceremonies are important events for the movie industry. In fact the attendance of important actors and directors is clear evidence that the industry recognizes the importance of these events in order to attract additional attention to the movie industry as a whole, and award-winning movies in particular. That attention then can influence the box office earnings and the dvd sales both for the awarded movie and the acclaimed filmmakers' future projects (Gemser, Van Oostrum & Leenders, 2007; Gemser, Leenders & Wijnberg, 2008; Hadida, 2010).

METHOD

Although the objective of this research was to learn something about the expert juries rather than on the cultural products themselves, the data that needed to be gathered are movie titles - the winners and

nominees honored annually by different organizations. The dataset comprised not only a number of years or award seasons, but also spread across different categories. That way the results showed whether the juries agreed not only on which film was the best, but also who was considered to be the best actor/actress, director and so on in the other categories, which allowed to make comparisons and gain a complete picture of the juries' coexistence and decision-making.

As in any competition, the juries have their own specific set of rules for the eligibility of motion pictures that can be considered for awards. Those have been carefully considered, because it was crucial for the measurement of consensus that the groups of contending films are the same. Otherwise looking for an overlap in the verdicts would not make much sense. For this reason national and regional competitions that choose only between products produced in a given region have been excluded. Research into the official selections of festivals in Cannes, Venice and Berlin revealed no overlap between the competition films in a given year. There are, however, numerous Anglo-Saxon organizations (USA-based and the British Academy) that proved to be good subjects for the comparison of verdicts.

This group of Anglo-Saxon organizations is also the one group which Simonton (2004) analyzed. His dataset comprised 28 years of ceremonies and looked into the rulings in ten major award categories that included best picture, acting, directing, screenplay, cinematography and musical contributions to a single film. His research employed reliability analysis that measured the level of inter-panel agreement, but also had the capacity to point out organizations which deviated most from the aggregated decisions of the others, and measured every jury's contribution to the consensus - how close they were to it. In his results, no panel turned out to announce verdicts very different from the collective. Simonton concluded a considerable consensus between the seven mentioned organizations: The Academy of Motion Pictures Arts and Sciences (Oscars), The British Academy of Film and Television Arts (BAFTA), Hollywood Foreign Press Association (Golden Globe Awards), The National Board of Review, The National Society of Film Critics, The Los Angeles Film Critics Association and The New York Film Critics Circle. The Cronbach's alpha values reported in his article ranged from .86 to .59 across ten categories.

Insert table 1 about here

The table shows the composite alpha-scores, which were used as a measure of the degree of consensus for all the juries which give out an award in that category. The other columns show this degree when the particular jury is removed from consideration. Simonton's results suggested that the consensus is relatively high, the literature generally accepts .7 as the threshold, and that removing juries nearly consistently lowered the reported values. This can be interpreted as the finding that all juries contributed to the established consensus, even though some do more than others. Removing the Oscars usually led to the largest drop in the value of alpha, which was interpreted by Simonton as a sign of their exceptional expertise on the quality of motion pictures.

Firstly, it might be dangerous to draw this last conclusion, which is based on the assumption that all juries were measuring with the same standard. Not only is there no proof that the juries really agree on a definition of merit, excellence or achievement, but since it is consensus that is measured, not quality, a more cautious interpretation is more appropriate. An intersubjective, rather than an objective interpretation of Cronbach's alpha would be more appealing for this type of research. A low alpha in this intersubjective interpretation would indicate that the majority of the variance in the total composite score was really due to heterogeneous, rather than consensual valuations of movies.

Secondly, given what was outlined above, it is strange to find such a high level of consensus between seven different juries, with every single one adding to the overall consensus. To put these findings to the test and to look deeper into the relations between the award series, different methods were applied to these the exact same data, in terms of the choice of juries, categories and the time-frame.

As already mentioned, to establish the degree of consensus between seven film juries Simonton has used a popular psychometric measure called Cronbach's alpha (Simonton, 2004). The main argument in this paper is that the use of this measure in the way that Simonton employed it does not give a good indication of the level of consensus between film juries. This is mainly due to the restrictions of the data. The only data available were nominations for three out of the seven juries, and

the winners in each category for all seven juries. This means effectively that the data were mainly dichotomous (win or no win). Furthermore, the value of alpha crucially depended on the size of the dataset and the inclusion or exclusion of certain groups of movies. But more importantly Cronbach's alpha did not measure the consensus on quality, but was very strongly influenced by something which does not warrant as much attention, consensus on films which did not deserve awards. This paper will present the findings that when a more relevant dataset is used the consensus found, using Cronbach's alpha almost completely. More importantly there are fundamental problems with the use of Cronbach's alpha to measure this kind of inter-jury consensus.

THE PROBLEMS WITH CRONBACH'S ALPHA FOR MEASURING CONSENSUS

Cronbach's alpha was originally designed as a generalized measure of internal consistency, especially of test scores. It is most used to differentiate observations or scores into a true score and an error score. According to the test theory underlying it, different evaluators or juries can have identical standards but differ in the application of those standards owing to random errors. Increasing the number of evaluators or juries than allows one to eliminate these errors and arrive at the true score (for a full discussion see Cortina, 1993). Stemler discusses the use of Cronbach's alpha to measure consensus between juries or judges as was done above (Stemler, 2004). He argues that: “Cronbach's alpha coefficient is a measure of internal consistency reliability and is useful for understanding the extent to which the ratings from a group of judges hold together to measure a common dimension”. According to the standard and his interpretation a low alpha means that the variance in the scores (i.e. awards/nominations) is mainly caused by the measurement method (i.e. various film juries) and not by variance in true merit between films.

Such a measure would work generally well, if the movies scored on say a 1-10, or a 1-100 scale in each competition. Higher scores for film A and lower scores for film B by all juries, and so on, would lead to a high alpha, while mixed reviews for all films would lead to low alpha scores. This was unfortunately not the nature of the data analyzed here. Simonton's dataset contains only three scoring

categories: winning an award=2, nominated for an award=1, and no award or nomination=0. This means that alpha does not reflect whether most juries have given a high rating, but rather to what extent juries all award or not award the same movies. So far so good, one might say, but there is a major problem lurking. The high alpha scores in Simonton's original research were not primarily generated by a consensus on awards or nominations. Instead the high alpha scores were mainly generated by a majority of zero scores from the juries, with an occasional nomination or win mainly lowering alpha. The only situation in which this would not be the case is if there were large group of films in the dataset which score 5 or even more wins/nominations. Only in that situation not winning/being nominated, a zero score, would actually lower alpha. For this dataset that was not at all the case. On the contrary, in the best picture category for example only 2 movies out of 124 winning titles have won five awards from the seven juries. Add to this that four out of the seven juries do not name nominees and what is left is almost a dichotomous variable on merit.

In Simonton's original research this problem was even more serious. For instead of just considering the movies nominated or awarded in a particular category, he considered the entire dataset of over 1100 movies which were awarded or nominated, regardless of the category in which this happened. In every single category, there are only between 183 and 387 films nominated. To give an example, if Simonton calculates alpha for the best picture category, he did not only consider the 310 movies which are nominated or win an award in this category, but also an additional 822 unsuccessful titles. The additional 822 cases are in fact films that have received no honors at all in that particular category; all seven juries fully agreed *not* to nominate nor award them for best picture. This proportion of non-successful films is so large that the consensus found has been mainly caused by juries agreeing that movies should *not* win instead of the other way around. It might furthermore explain why there was so little difference between the alpha scores for various categories. Table 2 includes Simonton's original values and new results produced while using the more restricted population for each category: only films which were nominated or won in that particular category were considered.

Insert table 2 about here

The results are very different from those in table 1. First of all no single category reached the reliability threshold of 0.7, even though some of the major categories did come relatively close to it, most notably the best actor/actress category. What is also evident is that contrary to Simonton's results there is no longer any evidence that each jury adds to the consensus. The BAFTA juries deviated in 5 out of 10 categories, and in the other categories the alpha score with the BAFTA excluded was close to the score with all juries included.

To demonstrate this problem somewhat clearer table 3 shows two generated datasets which both score an alpha value of .74. To simplify matters only four juries were used, which all shortlisted four movies from which they picked a winner. A single movie, movie A, was let to win all the awards in both situations. So the only difference between the two tables was found in which films were nominated. On the left hand side you will find a situation which should score higher on a measure of consensus. Since on the left hand side in table 3 two juries have consensually nominated movies B-G for an award, while in on the right hand side in table 3 all juries have nominated unique movies, movies B-M.

Insert table 3 here

Alpha is equally affected by consensus on not nominating a movie (by zero's in this case), as it is by a consensus on nominating a movie. This undesirable effect is even stronger if one would expand the left hand side with zeros for movies H-M, including them in the sample regardless of their apparent lack of quality. This raised the alpha value to a very respectable 0.86, but of course without being in any meaningful way a reflection of a higher degree of consensus. This example clearly showed that alpha is inappropriate for this type of data and therefore alternative methods to measure consensus were considered and developed for the type of data that was available. But more importantly this highlighted and important problem with any measure of correlation which does not discriminate between consensus on awards or shortlists and that other type of consensus not shortlisted or awarded.

RESULTS

Consenus Between Juries

The choice for alternative methods was not an easy one, both because a very specific objective was set: measure the level of consensus between these juries, and because the data was of a rather specific character (almost dichotomous). The fact that all juries had to be considered at the same time made various correlation methods unsuitable. But the problem highlighted in the previous section is perhaps the most important problem. Any correlation measure will inevitably not discriminate between the two types of consensus identified: consensus on quality, and consensus on films which should not win. In other words it will always overestimate the level of consensus which is truly of interest to us, the consensus on quality. And additionally the choice of the films included in the dataset will always be subject to discussion. This forced the choice for methods which focused on winning or being shortlisted only, rather than looking at that part of the data which shows agreement on not-winning or not-being shortlisted.

One alternative method considered was latent-class analysis as a way to analyze the results. In this method the various movies are divided up in classes with different probabilities to win a particular award. But it was found that type of analysis was more suitable for a different kind of question: what increases the likelihood of winning a particular award. A question which was further explored in by Simonton and others (Simonton, 2011, Pardoe & Simonton, 2008 & Kaplan, 2006). While yet other methods which look at conditional probabilities (does winning an Oscar increase the likelihood of winning a BAFTA) might be helpful in examining the tricky issue of contagion. This issue remains tricky because even if there was such an increased likelihood it would still be very difficult to distinguish this from a consensus on quality.

And thus our approach has been to go back to basics, and more importantly to look only at that part of the consensus which is most important, consensus on awards and shortlists. The results below are not all statistically very sophisticated, nor is there one measure which tells the whole story. Instead various dimensions of the level of consensus were looked at. This section will propose alternative measures of consensus, which can be used to compare differences between categories. The next section includes alternative measures which can be used to compare different juries. The first and most obvious way of

measuring consensus is by setting a certain threshold for consensus. For example if at least four out of seven juries award a particular movie, we call this a consensus. A similar threshold could be set for movies which are nominated for an award by at least three juries. This information is summarized in table 4 for each considered category. The percentages were calculated by comparing the amount of individual movies winning three or more awards with the total number of movies winning an award in that category. Hence, they are cumulative percentages of the amount of movies winning 3/4 or more awards, and being nominated 2 or more, or 3 times. These lower thresholds for nominations are used, because information on nominations is only available for three juries (AMPAS, HFPA and BAFTA).

Insert table 4 here.

These data show that it is difficult to distinguish clearly between the levels of consensus on awards in different categories. What is readily observable in table 4 is that on average in all categories roughly one movie per year won three awards or more (remember that the database covers 28 years). The cinematography percentages are slightly higher, but they were based on the evaluations of five juries instead of seven and are therefore difficult to compare to the other categories. Notable is that the degree of consensus on nominations for the director category was quite a bit higher than for the other categories.

A downside to this method was that one does not have an absolute standard to compare against. And besides, as it was stressed in the previous section, it is important to control for the length of the dataset itself. In general, the shorter the dataset which contains all winners and nominees, the higher the degree of consensus. So we have constructed a measure, β, for this type of consensus.

$$\beta = (pw-aw)/(pw-mw)$$

With *pw* being potential winners, or in other words juries multiplied by the number of awards. With *aw* being the actual unique winners and *mw* the minimum potential amount of winners, or in other words the amount of movies awarded in the case of perfect consensus. Of course one could easily construct a similar formula for nominations, simply by substituting the relevant variables. Looking back to our fictional example above from table 4, we see that when winners are considered, both

results would score the value of 1, by way of (4-1)/(4-1). There, a single film won all awards from the four fictional juries. In the case of nominations, and include the winner as a nominee, the left hand side would score (16-7)/(16-4)= 0.75, while the right-hand side would score (16-13)/(16-4)= 0.25.

These values were calculated for each category, and similarly to the results in table 1, further results were calculated by omitting the individual juries, to see if this affected the consensus and how. There was one issue that was encountered which was that, the Golden Globe jury awards twice as many awards in three categories thus significantly increasing the *mw* value for our calculations in these categories (The Golden Globes are awarded to Best Motion Picture - Drama as well as to Best Motion Picture - Comedy or Musical, the lead actors' and actresses' performances are also split into these two (drama/comedy) categories, both containing their own disjoint sets of nominees). To overcome this minor issue the average values of movies awarded or nominated in each category was used instead of the maximum amount of awards by a particular jury. This did not affect the scores greatly.

Insert table 5 here

The results in table 5a show clearly that there are no large observable differences between various categories regarding the β-score. The overall scores for all categories are relatively similar. For nominations this is different, the consensus is higher for the director category, and quite a bit lower for the foreign language category. Overall consensus on nominations is also higher, but this is not very surprising given the fact that each jury that announces nominations can nominate five movies (on average), while they can only pick one winner. A priori one would expect a larger degree of consensus on a top five list, than the number one on that list. Within the nominations the British Academy added virtually nothing to the overall consensus, except in the foreign language category. This supported some of the findings presented in the next section.

The overall values for the β measure were not very high, for the awards they are around the .5 mark, indicating the middle ground between consensus and dissensus. Of course this absolute value should be interpreted with caution; there is not much to compare it against and the measure should be

calibrated using other datasets. Furthermore the distribution between 0 and 1 might be skewed. This will undoubtedly improve when it will be possible to extend this research to other industries such as the book or music industry. But given the fictional examples from above which scored alpha values of 0.25 and 0.75, it can be concluded that there is not a very high degree of consensus between the various juries. This was also indicated by the data in table 7 below which shows that each jury awards a considerable amount of movies which receive no other awards. For all juries this was true of at least 20% of their awards. This was no different for the nominations.

One could argue that our alternative β-measure is upwards biased, since if wins were randomly distributed over the dataset the score would probably not be very close to zero. This objection is not relevant however, since the dataset is not a random sample from all movies, but the population of all winning/nominated movies. The population of all movies released in the period is of course much larger, and hence we can eliminate the possibility that scores significantly different from zero are caused by chance. So while it can be argued that there is not a very high degree of consensus between various movie juries, there is considerable overlap which is not caused by chance, but by an intersubjective consensus in the respective categories.

Significant differences between juries

Next to the examination of consensus between juries, it is also of interest to examine deviations from that consensus by particular juries. This was done by comparing the top end of the spectrum and the bottom end of the spectrum. The top end consisted of those movies winning various awards/nominations in a particular category. While the bottom end of the spectrum consisted of movies winning only one award/nomination.

The top end was considered first, movies winning various awards, which were also considered before in table 4. This method examined whether particular juries were more often part of the group of juries which award/nominate these 'successful' movies. Successful here designated that they have won multiple awards/nominations.

Insert table 6 here

Table 6 shows how often various juries are part of a consensus, the threshold used is at least 3 awards, and 2 nominations (3 nominations would of course by definition include all juries). Table 6a shows the total amount of movies winning at least 3 awards per category, while the other columns show what fraction of these movies were awarded by particular juries. For example, 28 films won 3 awards or more in the best picture category and in 71% of these cases one of these awards was an Oscar. It was subsequently tested whether these fractions differed significantly from the average fraction of all juries, using a binomial test. These levels of significance can be found below the fractions for particular juries, while the < or > sign indicates a significantly lower/higher score than the average. Finally, the overall fraction has also been calculated, for all categories.

Some clear patterns emerged. The AMPAS (O) was significantly more often part of the consensus overall, and also in two individual categories. And in all other categories, except the foreign language films, they scored a higher fraction than the average. A similar pattern occurs for the LAFCA (L), which scored significantly higher in two individual categories and overall, while only scoring lower than average in the best picture category. The opposite was true for the BAFTA (B); they scored significantly lower in three individual categories, and overall. Again, this was the result of a clear pattern since their fraction was consistently lower than the average fraction. And while some of the other juries differed significantly overall, this was not the result of a clear pattern of scoring more or less in the individual categories.

The results were confirmed by the results in table 6b for nominations. Again the BAFTA scored significantly lower, while the AMPAS scored significantly higher. The results for the HFPA (G) in table 7b should be interpreted with caution. Since the HFPA nominates twice as many movies in the picture, male lead and female lead categories the higher fractions there are no surprise. In the screenplay category a similar caveat applies for the BAFTA and AMPAS jury who divide this category up in adapted and original screenplay.

After the top end of the consensus was considered, similar methods were applied to the bottom end of the spectrum. Which juries did most often nominate/award movies which no other jury shortlisted or awarded? A very similar method as above was used, again counting for each individual jury how often it awarded a unique movie and comparing this against the average percentage.

Insert table 7 here.

Many of the patterns observed for the top end were recurring. The BAFTA again stood out as showing the least overlap with the other juries, scoring twice as high a percentage of unique awards then the AMPAS. And in 8 out of 10 categories it scored the highest or second highest percentage of unique awards. The absolute percentages were also relatively high; 45% of the movies which received an award of the British Academy received no award from any of the other juries. Only the average percentage of the NBRMP (R) came close with 39%. As observed above, the LAFCA was most often part of the consensus, but when the bottom end of the spectrum was considered this pattern was not very clearly confirmed. Overall the LAFCA scored the least unique movies awarded alongside the NYCC, after the AMPAS, but there was not as clear a pattern in the individual categories.

These results are also not as clear for the nominations, but one has to be cautious with these figures. As indicated above the HFPA awards and nominates twice as many movies in the best picture, male and female lead category, and remember this is also the case for the screenplay category for the AMPAS and the BAFTA. This means that the percentages for these categories should be interpreted with caution. Even so, it was clear that the AMPAS nominated the fewest unique movies with the exception of the foreign language category, further supporting the pattern observed above. And in some of the categories which can be compared more easily such as male and female support, the earlier noted pattern emerged that the BAFTA deviates most from the other juries.

DISCUSSION

This analysis of consensus between prominent movie expert juries demonstrated that consensus is not easily captured with one statistical measure. Instead consensus is best considered, and measured, as a

multi-dimensional concept. In this study we have identified the following dimensions: consensus on particular high-quality movies within award categories, the level of consensus on short-lists using our measure β, being part of the dominant opinion, and awarding movies which are not awarded by other juries. The results found using these measures indicate a much lower level of a consensus between various expert juries than the result in a previous study (Simonton, 2004). These results however should be interpreted with caution, since they are best compared with level of consensus in other time periods or other industries, this is especially true of the β-results.

The main method-related argument was that the use of Cronbach's alpha for this type of data is highly problematic. Firstly, because the measure is very sensitive to the inclusion/exclusion of particular groups of movies. And secondly, because that even if we could agree on the relevant population or sample of movies, it gives equal weight to consensus on not awarding movies as it does to awarding them, which makes the method unfit to measure the degree of consensus on 'quality' or 'merit'.

The causes underlying these findings were not investigated, although it was shown which factors might cause a higher and a lower level of consensus. The results, which indicate a moderate level of consensus between expert movie juries, are in line with what was predicted from our theoretical arguments. Since all juries seek to award 'merit' in filmmaking it is expected that there is a tendency towards consensus. It was however argued that that for various other reasons one would not expect anything close to complete consensus. From the point of view of consumers of this expert information, and the accompanying award events, it would make little sense to have the same movies win over and over again. From the point of view of the movie industry, for which the awards generate attention, it also makes sense to avoid too much overlap, so that each award ceremony is able to attract sufficient attention. Finally, some juries have explicitly been established with the goal of being a correction to a perceived biased opinion of the other dominant juries, and hence one would expect them to deviate as well. Contagion and anti-contagion effects were also discussed. Further research could examine to strength of each of these effects.

REFERENCES

Blaug, M. (2001). Where Are We Now On Cultural Economics? *Journal of Economic Surveys*, 15(2), 123–143.

Caves, R. (2000). *Creative Industries: Contracts between Art and Commerce*. Cambridge: Harvard University Press.

Cortina, J. M. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, 78(1), 98–104.

Gemser, G., Leenders, M. A. A. M., & Wijnberg, N. M. (2008). Why Some Awards are More Effective Signals of Quality than Others: A Study of Moview Awards? *Journal of Management*, 34(1), 25–54.

Gemser, G., Van Oostrum, M., & Leenders, M. A. A. M. (2007). The Impact of Film Reviews on the Box Office Performance of Art House versus Mainstream Motion Pictures. *Journal of Cultural Economics*, 31(1), 43–63.

Ginsburgh, V. (2003). Awards, Success and Aesthetic Quality in the Arts. *The Journal of Economic Perspectives*, 17(2), 99–111.

Ginsburgh, V. A., & Van Ours, J. C. (2003). Expert Opinion and Compensation: Evidence from a Musical Competition. *The American Economic Review*, 93(1), 289–296.

Hadida, A. L. (2010). Commercial Success and Artistic Recognition of Motion Picture Projects. *Journal of Cultural Economics*, 34(1), 45–80.

Holden, A. (1993). *Behind the Oscar: The Secret History of the Academy Awards*. New York: Simon & Schuster.

Kaplan, D. (2006). And the Oscar Goes to...: A Logistic Regression Model for Predicting Academy Award Results. *Journal of Applied Economics and Policy*, 25(1), 23–41.

Pardoe, I., & Simonton, D. K. (2008). Applying Discrete Choice Models to predict Academy Award Winners. *Journal of the Royal Statistical Society*, 171(2), 375–394.

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

Simonton, Dean Keith. (2004). Film Awards as Indicators of Cinematic Creativity and Achievement : A Quantitative Comparison of the Oscars and Six Alternatives. *Creativity Research Journal*, 16(2-3), 163–172.

Stemler, S. E. (2004). A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment, Research & Evaluation*, 9(4).

Wijnberg, N. M. (2003). Awards. In R. Towse (Ed.), *A Handbook of Cultural Economics* (pp. 81–84). Cheltenham: Edward Elgar.

Wiley, M., & Bona, D. (1996). *Inside Oscar: The Unofficial History of the of the Academy Awards*. New York: Ballantine Books.

Table 1, Simonton's original results

Category	Alpha	O	G	B	N	R	S	L
Picture	.76	.67	.71	.71	.73	.74	**.76**	.75
Director	.78	.71	.72	.77	.77	.77	.77	.76
Male Lead	.73	.66	.67	.71	.70	-	.72	.70
Female Lead	.76	.67	.74	.73	.73	**.76**	.75	.72
Male Support	.77	.69	.75	.75	.74	.75	.76	.74
Female Support	.74	.66	.66	.73	.70	.73	.72	.72
Cinematography	.74	.65	.68	.73	.72	**.74**	.71	.69
Screenplay	.66	.58	-	.57	.64	-	**.67**	.57
Music Score	.59	.47	.41	.49	-	-	-	**.63**
Song	.86	.72	.75	**.91**	-	-	-	-

O= AMPAS (Oscar's), G=HFPA (Golden Globes), B=BAFTA, N=New York Film Critics Circle, R=National Board of Review, S=National Society of Film Critics, L=The LA Film Critics Association.

Higher (or equal) than overall scores in bold.

Table 2, Alpha scores considering only movies nominated in a specific category

Category	Alpha	O	G	B	N	R	S	L
Picture	0,60	0,46	**0,60**	0,55	0,55	0,55	**0,61**	0,57
Foreign	0,37	**0,49**	0,34	0,32	0,25	0,23	**0,38**	0,29
Director	0,52	0,43	0,49	0,50	0,50	0,50	0,47	0,45
Male Lead	0,63	0,51	0,64	**0,63**	0,57	**0,64**	0,62	0,55
Female Lead	0,67	0,58	**0,67**	**0,67**	0,61	0,63	0,66	0,62
Male Support	0,46	0,35	0,38	**0,54**	0,37	0,45	0,41	0,40
Female Support	0,46	0,35	0,44	**0,53**	0,40	**0,48**	0,37	0,34
Cinematography	0,40	**0,43**	-	**0,46**	0,34	-	0,37	0,04
Screenplay	0,50	0,46	0,45	**0,57**	0,38	**0,54**	0,43	0,38
Song*	0,39	-	-	-	-	-	-	-
Music score	0,03	**0,10**	-0,13	0,01	-	-	-	**0,09**

*=since the BAFTA in this category was only awarded four times, we exclude it. Since only two juries

are left,
other
values
cannot
be
calculate
d.

Higher (or equal) than overall scores in bold.

Table 3, Two fictional results of 4 juries.

	Jury 1	Jury 2	Jury 3	Jury 4	Jury 1	Jury 2	Jury 3	Jury 4
Movie A	2	2	2	2	2	2	2	2
Movie B	1	1	0	0	1	0	0	0
Movie C	1	1	0	0	1	0	0	0
Movie D	1	1	0	0	1	0	0	0
Movie E	0	0	1	1	0	1	0	0
Movie F	0	0	1	1	0	1	0	0
Movie G	0	0	1	1	0	1	0	0
Movie H					0	0	1	0
Movie I					0	0	1	0
Movie J					0	0	1	0
Movie K					0	0	0	1
Movie L					0	0	0	1
Movie M					0	0	0	1

Table 4 - Amount of movies which (almost) win a majority of awards/nominations

	3+ Aw.		4+ Aw.	% of Win.	2+ Nom.	% of Nom.	3 Nom.	% of Nom.
Picture	28	22,58%	10	8,06%	150	48,39%	64	19,94%
Director	25	24,51%	8	7,84%	113	57,07%	59	27,57%
Male Lead	26	20,63%	14	11,11%	133	40,92%	56	16,47%
Female Lead	22	18,64%	12	10,17%	132	42,17%	56	17,23%
Male support	27	22,69%	9	7,56%	111	48,26%	34	13,33%
Female Support	22	17,46%	11	8,73%	104	46,64%	36	13,90%
Cinematography*	28	37,84%	13	17,57%	56	31,82%	-	-
Screenplay	25	19,69%	6	4,72%	171	50,00%	74	20,96%
Foreign Language	22	20,18%	6	5,50%	144	48,98%	82	27,89%

* Since only 5 juries awarded films in this category, we have used a threshold of 2/3 for wins.

Table 5a - Our β-values, winners

	β	O	G	B	N	R	S	L
Picture	0,52	0,44	0,44	0,50	0,48	0,47	0,51	0,51
Director	0,53	0,47	0,47	0,51	0,49	0,51	0,48	0,49
Male Lead	0,54	0,48	0,51	0,53	0,50	**0,54**	0,53	0,48
Female Lead	0,55	0,48	0,50	0,54	0,52	0,52	0,54	0,50
Male support	0,52	0,47	0,46	**0,53**	0,47	0,50	0,47	0,49
Female Support	0,47	0,42	0,45	**0,47**	0,45	**0,47**	0,43	0,42
Cinematography	0,49	0,44	-	0,47	0,48	-	0,46	0,38
Screenplay	0,47	0,38	0,43	**0,51**	0,40	**0,49**	0,42	0,43
Foreign Language	0,49	0,48	0,43	0,47	0,43	0,45	0,47	0,44
Average per jury	0,51	0,45	0,46	0,50	0,47	0,49	0,48	0,46

Table 5b - Our β-values, nominations (including award winners)

	β	O	G	B
Picture	0,61	0,45	0,59	0,57
Director	0,72	0,62	0,67	**0,73**
Male Lead	0,55	0,36	0,51	**0,57**
Femal	0,56	0,36	0,52	**0,59**

e Lead				
Male support	0,58	0,31	0,41	**0,69**
Female Support	0,58	0,38	0,42	**0,65**
Cinematography*	0,36	-	-	-
Screenplay	0,63	0,48	0,57	**0,75**
Foreign Language	0,40	0,37	0,13	0,19
Average per jury	0,55	0,42	0,48	**0,59**

*=2 juries only.

Bold values indicate higher (or equal) than overall scores

Table 6a, Which juries are most often part of a consensus (3 awards or more)?

	3+ wins	Aver. f	O	G	B	N	R	S	L
Picture	28	0,49	**0,71**	0,64	0,39	0,46	0,54	**0,29**	0,43
			>.016*	0,084	0,186	0,446	0,404	**<.02***	0,304
Director	25	0,47	**0,68**	0,52	**0,28**	0,40	0,44	0,48	0,52
			>.031*	0,396	**<0.039***	0,296	0,446	0,554	0,396
Male Lead	26	0,57	0,69	0,69	0,50	0,58	0,42	0,42	0,69
			0,146	0,146	0,295	0,558	0,093	0,093	0,146
Female Lead	22	0,61	0,73	**0,91**	0,50	0,59	0,45	**0,36**	0,73
			0,183	**>.002****	0,2	0,507	0,102	**<.017***	0,183
Male support	27	0,50	0,59	0,67	**0,19**	0,59	0,41	0,52	0,56
			0,227	0,064	**<.001****	0,227	0,215	0,508	0,358
Female Support	22	0,55	0,68	0,45	**0,32**	0,55	0,59	0,55	0,68
			0,141	0,261	**<.027***	0,585	0,416	0,585	0,141
Cinematography *	28	0,54	0,57	-	0,50	0,39	-	0,50	**0,75**
			0,458	-	0,393	0,08	-	0,393	**>.02***
Screenplay	25	0,48	0,64	0,40	0,32	**0,72**	**0,04**	0,56	**0,68**
			0,08	0,275	0,08	**>0.013***	**<.000****	0,274	**>.035***
Foreign Languag e	22	0,49	0,41	0,55	0,36	0,64	0,50	0,32	0,68
			0,281	0,394	0,156	0,131	0,562	0,075	0,06
Overall	225	0,52	**0,64**	**0,60**	**0,37**	0,54	**0,42**	**0,44**	**0,63**
			>.000**	**>.012***	**<.000****	0,295	**<.003****	**<.012***	**>.001****

Table 6b, Which juries are most often part of a consensus (2 nom. exactly)?

	2 Nom.	Aver. f	O	G	B

Picture	86	0,67	0,72 0,188	**0,90** **>0,000**** *	**0,38** **<0,000**** *
Director	54	0,67	**0,81** **>0,014***	**0,81** **>0,014***	**0,37** **<0,000**** *
Male Lead	77	0,67	**0,83** **>0,001**** *	**0,95** **>0,000**** *	**0,22** **<0,000**** *
Female Lead	76	0,67	**0,87** **>0,000**** *	**0,95** **>0,000**** *	**0,18** **<0,000**** *
Male support	77	0,67	**0,96** **>0,000**** *	**0,83** **>0,001**** *	**0,21** **<0,000**** *
Female Support	68	0,67	**0,88** **>0,000**** *	**0,88** **>0,000**** *	**0,24** **<0,000**** *
Screenplay	97	0,65	**0,90** **>0,000**** *	**0,45** **<0,000**** *	0,61 0,224
Foreign Language	63	0,67	**0,41** **<0,000**** *	**0,92** **>0,000**** *	0,67 0,525
Overall	598	0,66	**0,81** **>0,000**** *	**0,82** **>0,000**** *	**0,36** **<0,000**** *

*=significant at the 0.05 level, **=significant at the 0.01 level, bold values indicate

significant deviations, higher (>), or lower (<).

Table 7a - Which juries do most often nominate unique movies

	Perc.	O	G	B	N	R	S	L
Picture	30,6%	**3,6%**	32,1%	40,7%	25,0%	25,8%	40,0%	43,3%
		<0.000**	0,46	0,15	0,34	0,36	0,21	0,10
Director	27,4%	18,5%	18,5%	36,4%	29,6%	39,3%	22,2%	28,6%
		0,21	0,21	0,24	0,47	0,51	0,12	0,52
Male Lead	28,8%	**3,7%**	**40,4%**	33,3%	22,6%	43,8%	32,1%	**14,7%**
		<0.001**	**>0.041***	0,37	0,29	0,05	0,42	**>0.046***
Female Lead	28,8%	**10,3%**	38,3%	37,0%	20,8%	25,0%	38,5%	21,9%
		<0.017*	0,07	0,23	0,27	0,44	0,19	0,26
Male support	29,8%	21,4%	**13,8%**	**54,2%**	25,0%	38,2%	25,7%	33,3%
		0,23	**<0.040***	**>0.011***	0,35	0,19	0,37	0,39
Female Support	37,7%	**18,5%**	39,3%	51,9%	40,6%	**52,8%**	31,0%	27,3%
		<0.027*	0,50	0,10	0,43	**>0.047***	0,30	0,15
Cinematography	37,7%	38,5%	-	46,2%	45,0%	-	41,7%	**19,2%**
		0,54	-	0,24	0,32	-	0,42	**<0.037***
Screenplay	34,6%	32,7%	25,0%	**66,7%**	**11,1%**	71,4%	22,2%	25,0%
		0,45	0,19	**>0.000****	**<0.006****	0,05	0,12	0,19
Foreign Language	33,7%	46,4%	25,0%	40,7%	23,1%	34,5%	38,1%	28,6%
		0,11	0,22	0,28	0,18	0,53	0,41	0,36
Overall	31,8%	**22,5%**	31,3%	**46,6%**	**26,7%**	**38,9%**	31,8%	**26,8%**
		<0.000**	0,451	**>0.000****	**<0.044***	**>0.015***	0,522	**<0.048***

*=significant at the 0.05 level, **=significant at the 0.01 level

Table 7b - Significant difference

s for nominations				
	Perc.	**O**	**G**	**B**
Picture	30,5%	**7,4%**	**49,3%**	**11,8%**
		<0.000**	**>0.000****	**<0.000****
Director	22,4%	18,9%	**29,0%**	19,4%
		0,20	**>0.039***	0,28
Male Lead	37,4%	**8,4%**	**53,4%**	31,1%
		<0.000**	**>0.000****	0,11
Female Lead	36,1%	**6,9%**	**52,6%**	30,0%
		<0.000**	**>0.000****	0,12
Male support	32,4%	**20,1%**	31,7%	**50,0%**
		<0.001**	0,47	**>0.000****
Female Support	33,6%	**25,6%**	34,0%	**44,1%**
		<0.031*	0,49	**>0.023***
Cinematography	51,7%	56,6%	-	45,6%
		0,15	-	0,13
Screenplay	35,4%	35,5%	**18,1%**	**48,7%**
		0,50	**<0.000****	**>0.000****
Foreign Language	42,0%	**56,8%**	35,9%	**32,2%**
		>0.000**	0,08	**<.018***
Overall	35,0%	**27,2%**	**41,5%**	35,6%
		<0.000**	**>0.000****	0,348

*=significant at the 0.05 level, **=significant at the 0.01 level, bold values indicate significant deviations,

higher (>),
or lower
(<).