Journal of Cultural Economics (2005) 29: 59–78

Expert Judgment Versus Public Opinion – Evidence from the Eurovision Song Contest

MARCO A. HAAN, S. GERHARD DIJKSTRA and PETER T. DIJKSTRA

Department of Economics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands
E-mail: m.a.haan@eco.rug.nl; sgdijkstra@gmx.net; ptdijkstra@gmx.net

Abstract. For centuries, there have been discussions as to whether only experts can judge the quality of cultural output, or whether the taste of the public also has merit. This paper tries to answer that question empirically, using national finals of the Eurovision Song Contest. We show that experts are better judges of quality in the sense that the outcome of finals judged by experts is less sensitive to factors unrelated to quality than the outcome of finals judged by public opinion. Yet, experts are not perfect; their judgment does still depend on such factors. This is also the case in the European finals of the contest.

Key words: Eurovision Song Contest, expert judgment, public opinion

1. Introduction

Ancient wisdom has it that there is no arguing about tastes. Yet, for many centuries artists, critics, philosophers and economists, amongst others, have done exactly that. In particular, they have argued about whether only specialists can assess the quality of art, or whether the taste of the general public also has some merit.[1]

This discussion has important implications for the question as to whether there is a market failure in the provision of the arts, and whether government should intervene. If the general public is a bad judge of artistic quality, then market provision of the arts, which effectively boils down to judgment by the general public, would not be the ideal institution to foster and promote the quality of the arts. In that case, government would have a role in supporting artists who are judged by experts (but not by the public) as being worthwhile. This is the classic merit good argument, introduced by Musgrave (1959).

Indeed, there are those that argue that "producers of popular culture tend to aim their offerings at the lowest common denominator thereby degrading cultural products by catering to the relatively uncultivated tastes of ordinary consumers"[2] (see Holbrook, 1999 and the references therein). This concern dates back at least to Plato, who argued in *The Republic* that attempts to please the audience would decrease the quality of theatrical productions. Adherents of this view thus argue that judgments of the artistic merits of cultural production should be left to experts

who are familiar with the particular art form, and who can put the offerings into their proper perspective.

On the other end of the spectrum, there are those that argue that market competition "augments rather than undermines the quality and quantity of cultural creations".[3] Economic incentives encourage artists to address the needs and interests of audiences. Economists and even critics and philosophers, the argument goes, cannot judge objectively the quality of art, just as a central planner will not be able to decide on the proper production and allocation of goods and services. Such jobs can only be done by the market, i.e., by the general public. One of the most outspoken proponents of this view is Tyler Cowen (1998) who argues that "aesthetic judgments that divide 'high' culture from 'low' culture fail to appreciate adequately the vitality of commercial culture and the efficacy of market forces in stimulating and sustaining creativity in all areas of artistic expression." Cowen dismisses people having such judgments as cultural pessimists, who "wish to supercede the workings of the market with their own moral and aesthetic judgments."

It seems impossible to judge which of these views is correct, and whether experts or the general public are best able to judge the quality of cultural output.[4] Any attempt to do so, it seems, inevitably implies the need for making judgments about quality to start with. Obviously, such an approach can never yield an objective evaluation of the judgment of quality. It seems that one cannot evaluate judgments of cultural merit without making such judgments oneself. Yet, in this paper, we do exactly that.

We show that the judgment of quality by a team of experts is inefficient in the sense that the random order of appearance of participants in a contest has a systematic effect on the final ranking of those participants, as decided by a jury of experts. Glejser and Heyndels (2001) argue that this is an inefficiency in the jury process. If jurors really evaluate contestants purely on their merit, then their order of appearance should have no influence on the final ranking.[5] When there is a correlation between order of appearance and final ranking, this then indicates that the jury is influenced by exogenous factors that should not influence their judgment.

Arguably, worse judges are more inefficient in this sense, as they are more strongly influenced by such exogenous factors. If judgments by the general public are more inefficient than judgments by a jury of experts, then we may argue that expert judgment is superior in the sense that it aggregates information in a way that is unambiguously better than that of the general public. In this paper, we show that, at least in the dataset we use, this is indeed the case. Thus, we are not, and never will be, able to judge whether the evaluation criteria that are used by the general public to judge cultural quality are "better" or "worse" than the criteria used by experts. But we are able to show that, however appropriate or inappropriate those evaluation criteria may be, experts at least do a better job than the general public in using them to evaluate the quality of cultural output.

Our research is inspired by Glejser and Heyndels (2001; henceforth GH), who study the Queen Elisabeth International Music Competition, a prestigious classical music contest held annually in Brussels, and judged by a panel of jurors who are leading experts in their field. Finalists perform on six consecutive nights, with two finalists per night. The order in which contestants perform is drawn by lot. Yet, GH show that this order has a systematic influence on the final ranking. Finalists that perform later in the week, do significantly better on average. The second finalist on a given night does better than the first one. As noted, the authors interpret this as evidence for the inefficiency of the jury process.

In this paper, we use data from the Eurovision Song Contest (ESC), an annual festival organized by the European Broadcasting Union (EBU), in which several countries participate, each with one song. Juries from all participating countries decide who is the winner, by awarding points to their favorite songs. The ESC is an annual event with a long history. It is shown live on television throughout Europe, and attracts roughly one hundred million viewers each year. We give further background on this contest in Section 2. Up to 1998, national juries consisted of experts that evaluated the songs and awarded points to the different contestants. In Section 3, we study these festivals. We find that songs that are performed later during the contest do significantly better, even though the order of appearance is determined randomly. This finding is consistent with GH.

Usually, the song representing a country in the ESC is chosen in a National Final or National Song Contest (NSC) that is broadcast on national television. The EBU does not issue any strict rules as to how to select a song, but most countries choose a format that is very similar to that of the ESC itself, often involving separate regional juries. The number of entries in an NSC is usually around 10. In Section 4, we extend the analysis by looking at NSCs. Interestingly, jury procedures used to elect the national winner differ across countries and through time. Originally, expert juries were used. Yet, increasingly, countries use a system of televoting, where each viewer can decide which song (s)he likes best, and then make a call to a phone number that is assigned to that particular song. In many countries, hundreds of thousands of viewers make such a call. If it is true that experts are a better judge of quality than the general public, then we would expect that the inefficiency noted by GH and in Section 4 of this paper, is much stronger in contests in which the public decides, than it is in contests with an expert jury. Section 4 shows that this is exactly the case.

One could argue that our results are merely driven by the fact that in the contests we consider, expert juries may be able to hear songs more often than the general public, and are therefore able to make a better judgment. This concern is addressed in Section 5, where we restrict our attention to a set of contests for which we are certain that the public was able to hear the participating songs more than once. We show that this does not affect our results.

Admittedly, few people would argue that the ESC represents high-brow culture. Many commentators claim that the participating songs are of dismal quality. Yet,

as argued above, it is not up to us to judge the quality of the contestants of the ESC, or the overall quality of the festival. We are only interested in the extent to which experts and the public are able to evaluate the participating songs, and pick the best. Regardless of the extent to which we feel that they indeed pick the song with the highest quality, at least their choice should be based purely on the perceived merits of the song itself, and not on any exogenous factors that have nothing to do with the quality of the songs. If these factors, such as the order in which songs are performed, have a stronger influence, then we can safely argue that the judgment of quality is more flawed.

In related work, Ginsburgh (2003) takes the extent to which works of art are still appreciated in the longer run as a proxy for their true aesthetic quality. He shows that, in the cases of movies and books, prizes awarded shortly after the production of an artwork or rankings that result from competitions are correlated with economic success and may even influence or predict it, but are often poor predictors of survival of the work. Ginsburgh and van Ours (2003) find strong support for the hypothesis that doing well in the Queen Elisabeth competition does help musicians in their career. This also implies that inefficiencies in the jury process have a lasting effect on a musician's career.

2. The Eurovision Song Contest: Background and Details

In the early 1950s, television networks were formed throughout Europe. In an effort to improve the quality of programs and to try to achieve economies of scale, networks in 10 countries[6] decided to join forces and establish the European Broadcasting Union (EBU). Under the *Eurovision* banner, the EBU started to distribute pan-European TV programs. In 1955 Marcel Benençon, Director General of Swiss Television, proposed to also organize and broadcast a song contest, initially modelled after the San Remo Festival, established in 1951. The purpose of the contest is to "promote high-quality original songs in the field of popular music, by encouraging competition among artists, songwriters and composers through the international comparison of their songs." (EBU/EUR, 2001). On May 24, 1956 the first edition of the Eurovision Song Contest took place in Lugano, Switzerland. Seven countries participated, each with two songs. Since 1957 each country can participate with only one song.

Each contest follows a by now standard format. First, after an initial introduction, the songs are performed in an order predetermined by lot.[7] Second, there is a break of about 5 minutes, in which national juries can decide on their vote. Third, the votes of the national juries are revealed, following the same order as that of the actual contest. This stage takes almost 1 hour. Fourth, the winner is announced, and the winning song is performed once more. Nowadays the entire show takes roughly 3 hours.

Contestants are often relatively unknown at the time of the festival, although there are exceptions. In general, the song they perform is not written by themselves,

but rather by some professional composer and songwriter. Songs have to be new, in the sense that they have not been recorded earlier. The original idea was that jurors would hear the songs for the very first time during the contest. This, however, is not always feasible. Since 1960, jury members are allowed to hear songs before the actual contest, but not to see them being performed. During the contest, songs are performed live.[8] Until the 1999 contest, all contestants in the ESC had musical backing from a symphony orchestra, provided by the host country. Sometimes, songs winning the ESC become huge hits, and winning artists manage to pursue a major national or international career.[9] In other cases, both the songs and their performers are soon forgotten.

Surprisingly, there are no restrictions on the nationality or citizenship of the performing artists or the composer of a song. Indeed, in the past it has often happened that winners were representing countries different from their own.[10] There have been restrictions, however, on the number of performers of a song. Starting in 1957, only 2 singers could be on stage, without any further vocal accompaniment. This rule was modified only in 1971, when the maximum was set to six performers. Also, since 1989 there has been an age limit of 16. Since 1962, the time limit for a song has been 3 minutes.

A widely discussed issue is the freedom of language. In early contests there were no rules with regard to the language in which songs were performed. Yet, each contestant still chose to use her own language. This changed in 1965, when the Swedish contestant sang in English. This led to a restriction in place since 1966 that performers could only sing in (one of) the official language(s) of their country. It is often argued that this restriction gives a huge advantage to Ireland and the U.K. – the only countries allowed to perform in English. In 1973 freedom of language was reinstated, but it was re-abolished in 1977. Since 1999, there has again been freedom of language.

The way in which the contest is judged has differed throughout the years. The exact details of the voting procedure for all the ESCs we consider is given in appendix A. In most cases, each participating country has a national jury that consists of a fixed number of members. Each member awards points to her favorite songs. For each national jury, these points are aggregated to yield a ranking of the songs for that particular jury. Each country then awards points based on that ranking. Since 1975, a national jury's favorite song receives 12 points, their second favorite 10 points, while 8, 7, 6, 5, 4, 3, 2, and 1 point are awarded to the third through tenth favorite. Juries cannot award points to the song representing their own country. Only the points of a national jury are revealed – not those of its individual members. Points given by each national jury are aggregated and determine the final ranking. All jurors have to cast their votes before the start of the voting stage in the television show, in an attempt to prevent strategic voting.

In principle, every country that wants to can participate in the ESC. The only restrictions are that the network broadcasting the contest has to be a member of EBU, and that the contest has to be shown live on television in that country in the

year of participation, as well as the year before. Membership is not restricted to European countries. In the past, for example, Morocco and Israel have been contestants in the ESC. Yet, since 1993, the number of countries wanting to participate has increased sharply. To prevent the contest from running too long, different qualification mechanisms have been introduced.[11] Hence, since 1993, a country that wants to participate is no longer guaranteed a spot in the contest. The exceptions to this rule are Germany, France, Spain and the United Kingdom, the so-called Big Four. These countries contribute a large amount of the EBU's budget and are therefore guaranteed participation.[12]

In the ESC, the *televoting* system was introduced in 1998. Every citizen in a participating country can make a call to a phone number corresponding to her favorite song. Each household can vote only three times. Calls can be placed during a period of 5 minutes, after all contestants have performed. A country lacking the necessary infrastructure for televoting uses the old system with 16 jury members. In either case, votes are translated per country to the now usual format of 12, 10, 8, 7, . . . , 1 point. Countries with televoting have a back-up jury of 8 members, in case problems with televoting occur.[13] In some national finals, televoting has already been in use for a much longer time.

Clearly the system with televoting is fundamentally different from the system with juries. Rather than a small number of carefully selected jurors, anyone with a phone can now be a part of the voting process. And many people choose to do so: in many countries, the number of people calling in to register their vote is in the hundreds of thousands. Therefore one can argue that, with televoting, public opinion determines the winner. With a jury system, the result is determined by experts, or at least by people that have been carefully selected and are committed to an honest and fair contest and moreover realize that their vote can be of crucial importance in determining the winner of the contest. In our study of the ESC, we therefore restrict attention to those contests that have been judged by a jury. For the NSCs, we test for the difference between a jury of experts and public opinion, by taking advantage of the fact that some NSCs are judged by a jury of experts, while others are decided by televoting.

3. Inefficiency in the ESC

In this section, we test for efficiency in the ESC. Note that the order of appearance in a contest is randomly drawn. Therefore, the final ranking of the songs should not be influenced by this order. A jury is supposed to determine the final ranking of the contestants purely on the basis of quality. Any evidence of systematic influence of a factor that is exogenous to that quality therefore implies inefficiency in the jury's decision making process (for an extensive discussion, see GH). The order of appearance clearly is such a factor.[14] In this section, we therefore test to what extent the order of appearance influences the final ranking in the ESC, while also taking other potentially important factors into account.

We collected data for all ESCs in the period 1957–1997. Data from order of appearance and final ranking are taken from Eeuwes (2002) and Walraven and Willems (2000), but are also available from many other websites. Starting in 1998, an increasing number of countries started to use televoting to determine their votes in the ESC. Therefore, these contests are not included.[15] This gives us a total of 41 festivals. Some summary statistics are given in Table I. For each contestant, we observe its order of appearance in the festival in which it participated. Since not all festivals have the same number of contestants, we normalize these values to the interval [0,1]. In a contest with n participants, the contestant that performs as number i of the evening, has a value for APPEARANCE that equals $(i-1)/(n-1)$. Hence the first contestant gets a value of zero, and the last contestant a value of one. We use the same normalization for the variable RANK, which gives the place of a contestant in the final ranking. Note therefore that a lower value of RANK implies a *better* performance.

As an example, suppose a festival has 21 contestants. A certain contestant performs as the 6th of the evening, and is number 15 in the final ranking. The observation for APPEARANCE for this contestant then equals 0.25 and the observation for RANK equals 0.70. In the case of ties, each of the tying contestants is awarded a ranking that is equal to the average of all rankings in that tie.

It is often observed that some countries almost always do particularly well at the ESC, while others perform particularly badly. The United Kingdom, for example, managed to secure second place in no fewer than 15 festivals, and won another 6. To allow for systematic quality differences in the contributions of the different countries, we have included country dummies. A total of 35 countries have participated in the ESC in our sample period.[16] Obviously, this implies that only 34 country dummies can be used. For ease of interpretation, we use as a benchmark the country with the most "average" performance over all contests in our sample. This turns out to be Denmark: the average value of RANK for this country is 0.49.

Table I. Summary statistics

Year	ESC 1957–1997	NSC expert 1993–2001	NSC televoting 1988–2001
Male	0.306		
Female	0.447		
Duo	0.117		
Group	0.129		
Observations	758	492	256
Contests	41	44	26
Participants per contest	18.49	11.18	9.85

The coefficients of the country dummies should thus be interpreted as how well particular countries are doing, *ceteris paribus*, relative to the average participating country – which happens to be Denmark.

Also, the country hosting an ESC (as a rule, the winner of the previous year) always seems to do particularly well. This can be due to the fact that the host country puts particular effort in selecting a fitting song. Alternatively, the other countries may be willing to judge the contribution of the host country more sympathetically.[17] We therefore include a dummy for the host country as well. Finally, we allow for the possibility that one type of contestant performs better than another type. We have therefore divided the 758 contestants in our data set into four categories: male singers, female singers, duos, and groups. For the first three categories, we include a dummy. Data on the types of contestants were found using Walraven and Willems (2000) and numerous websites, in particular Eilers (2002).

In column I of Table II, we explain the final ranking of a contestant at an ESC from its order of appearance plus a dummy for the host country, a dummy for the country the contestant is representing,[18] and dummies for a solo male performer,

Table II. Estimation results ESC (dependent variable: RANK. *t*-values in parentheses)

	I	II	III
Constant	0.551*** (9.11)	0.576*** (9.44)	0.562*** (9.09)
Appearance	−0.086** (−2.59)	−0.124*** (−3.44)	−0.105** (−2.67)
Host	−0.110* (−2.43)	−0.110* (−2.42)	−0.110* (−2.44)
United Kingdom	−0.293*** (−4.33)	−0.286*** (−4.24)	−0.286*** (−6.29)
Ireland	−0.213** (−2.99)	−0.202** (−2.85)	−0.202** (−2.85)
France	−0.206** (−3.03)	−0.198** (−2.93)	−0.197** (−2.91)
Belgium		0.134* (1.99)	0.133* (1.98)
Finland	0.199** (2.86)	0.210** (4.19)	0.210** (3.02)
Norway	0.136* (1.99)	0.137* (2.02)	0.140* (2.05)
Portugal	0.170* (2.44)	0.183** (2.62)	0.185*** (2.66)
Turkey	0.217** (2.66)	0.214** (2.63)	0.209* (2.57)
Male	0.043 (1.26)	0.038 (1.11)	0.038 (1.10)
Female	−0.029 (−0.89)	−0.034 (−1.06)	−0.036 (−1.10)
Duo	−0.049 (−1.19)	−0.054 (−1.31)	−0.056 (−1.36)
Opening		−0.127** (−2.63)	−0.114* (−2.29)
Second			0.062 (1.29)

All country dummies (except Denmark) are included in all regressions. Only country dummies that are significant at the 5% level are reported in the table. Full estimation results are available from the authors upon request. Column II is the preferred specification.
*Significant at 5%-level.
**Significant at 1%-level.
***Significant at 0.1%-level.

solo female performer, and duo. The equation is estimated using ordinary least squares. We have only reported the country dummies that are significant at the 5%-level. These fall into two groups: those for countries that do systematically better, and those for countries that do systematically worse than the average country.[19] The first group consists of the United Kingdom, Ireland, and France. Countries in the second group are Finland, Norway, Portugal and Turkey. The performance of the United Kingdom is especially impressive. On average, in a contest with 20 participants, the artist representing this country has a final ranking that is 5.5 places better than the average country. The artist representing Turkey, however, has a final ranking that is more than 4 places worse than the average country. The HOST dummy is also significant. None of the type-of-performer dummies are.

It is sometimes argued by avid followers of the contest that the song that is performed as the very first one in a contest has a better chance of winning. To test for this, we included the dummy OPENING, which equals 1 if and only if the particular song was performed as the first one of a contest. Column II in Table II shows that the coefficient for OPENING is indeed negative and significant. This could be due to some non-linearity in the true relationship between RANK and APPEARANCE, which is not explicitly modelled in our specification, but is picked up by OPENING. As a robustness check, we therefore also included a dummy SECOND, which equals 1 if and only if a song was performed as the second of a contest. If the negative coefficient of OPENING is indeed due to non-linearity, then the coefficient of SECOND should also be negative and significant. Column III of Table II shows that this is not the case: the effect of SECOND is insignificant. Also, including this dummy has little influence on the estimated effect of OPENING. As a further test for non-linearity, we used Ramsey's RESET-test (Ramsey, 1969). This also provided no evidence for non-linearity. Hence, there is truly an effect of being the first performer during an ESC.[20] Therefore, column II in Table II is our preferred specification.

When a jury bases its decision purely on the merits of the song under consideration, the final ranking should be independent of the order of appearance. This is evidently not the case in our data: the coefficient of APPEARANCE is negative and significant at the 0.1%-level. Hence, a song that is performed later during the contest stands a much better chance of obtaining a low value for RANK, and therefore does better on average. This is in line with Glejser and Heyndels (2001). The coefficient of APPEARANCE that we find, 0.124, implies that *ceteris paribus* a song that is performed last has a final ranking that is 12% better than a song that is performed near the beginning of the contest. For a contest with 20 participants, this boils down to roughly 2.3 places in the final ranking. Yet, we also see that the very first performer has a clear advantage relative to this effect: the coefficient of OPENING is significant and equal to 0.127. Interestingly, this is virtually the same value as the one we find for APPEARANCE. This implies that on average the very first and the very last song perform equally well. Apart from this, there is a negative relation between appearance and final ranking.

We thus find new evidence that juries are influenced by factors that should have no influence on their opinion: in this case the order of appearance of the contestants. So far we have looked only at contests that have an expert jury. In the next section we look at a different data set, that of national finals. Here, final rankings are sometimes determined by a jury of experts, and sometimes by the general public. We take advantage of this heterogeneity to test whether the inefficiency that we identified in this section is stronger for public opinion or for expert juries.

4. National Finals: Expert Judgment Versus Public Opinion

To test the main hypothesis of this paper, we used Stoddart (2002) to obtain data for a total of 70 national finals[21] (for a full list, see appendix B). Most national finals are judged exclusively by an expert jury. But in recent years, the number of countries that exclusively use televoting in their national final has increased. We use data from finals that are as recent as possible.[22] Summary statistics are given in Table I. Note that 9 out of the 26 televoting contests were held in the U.K., which used televoting as early as 1988. For the purposes of this study, we could pool all our data, including both ESCs and NSCs, but the ESCs have a different format and an element of international competition that is lacking in NSCs. Also, the number of contestants in an ESC is often much higher than that in an NSC. To avoid any possible influence these factors may have, we therefore restrict attention to the NSCs.

To test for a possible difference in efficiency between experts and the public, it would be ideal to have a situation in which the *same* contests are judged by both experts and the public. Alas, this is not possible. A number of countries do use some combination of both systems, but unfortunately, in most cases the results of the experts and the public are not reported separately. Therefore, we have to rely on different contests, some of which are judged by the general public, and some by experts.

There is a difficulty in testing for the difference between expert jury and public opinion. In vector notation, the equation we want to estimate in each subsample is

$$\text{RANK} = \alpha + \beta\ \text{APPEARANCE} + \varepsilon, \tag{1}$$

with ε a vector of iid error terms. Yet, this specification imposes too much structure. When the jury process is efficient, we will find a value for β that equals 0, and a value for α that equals 0.5. Inefficiency implies both that β differs from 0, and that α differs from 0.5. When we do not use any additional dummies, the intercept α fully determines the slope β since, by construction, the regression line passes through (0.5, 0.5). However, to compare levels of efficiency, and to do so in a statistically meaningful manner, we need to have one single coefficient that fully captures efficiency.

Formally, this can be done as follows. Note that we have normalized both RANK and APPEARANCE to lie in the interval [0,1]. By construction, the average value for

both RANK and APPEARANCE equals 0.5. By virtue of ordinary least squares, the regression line given by (1) therefore necessarily passes through (0.5, 0.5). This implies that we must have $0.5 = \alpha + \beta \times 0.5$ or $\beta = 1 - 2\alpha$. Plugging this back into (1) yields

$$\text{RANK} = \alpha + (1 - 2\alpha)\,\text{APPEARANCE} + \varepsilon,$$

or

$$\text{RANK} - \text{APPEARANCE} = \alpha\,(1 - 2 \times \text{APPEARANCE}) + \varepsilon.$$

By defining the transformed variables TRANSRANK $\equiv$ RANK $-$ APPEARANCE and TRANSAPPEAR $\equiv 1 - 2 \times$ APPEARANCE, we can thus find an unbiased estimate for α (and, by implication, β) by regressing TRANSRANK on TRANSAPPEAR, since we now have

$$\text{TRANSRANK} = \alpha\,\text{TRANSAPPEAR} + \varepsilon. \tag{2}$$

Define the dummy EXPERT to equal 1 if and only if the observation is from an NSC with an expert jury. Note that we have defined the inefficiency of the jury process as the extent to which β differs from 0, which is equivalent to the extent to which α differs from 0.5. Thus, televoting is less efficient than an expert jury when the value of α is significantly higher for observations with televoting. We can test for this by interacting EXPERT with TRANSAPPEAR and adding that expression to the equation above:

$$\text{TRANSRANK} = \alpha\,\text{TRANSAPPEAR} + \gamma\,(\text{TRANSAPPEAR} \times \text{EXPERT}) + \varepsilon. \tag{3}$$

When public opinion is inefficient in the sense that contestants that perform later have an advantage (as was the case in the ESCs), then the value of α should be significantly higher than 0.5. Moreover, when expert juries are indeed more efficient than the general public, then the value of γ should be significantly negative.

Table III gives the result of this regression. Note that the coefficient of TRANSAPPEAR is highly significant. The t-statistic reported here is that for the hypothesis that

Table III. Estimation results NSC (dependent variable: TRANSRANK. t-values in parentheses)

Transappear	0.623** (4.02)
(Transappear) × expert	−0.091* (−2.39)

See main text for definitions of variables.
Note: The t-value reported for TRANSAPPEAR is for the null hypothesis that TRANSAPPEAR = 0.5.
*Significant at 5%-level.
**Significant at 1%-level.

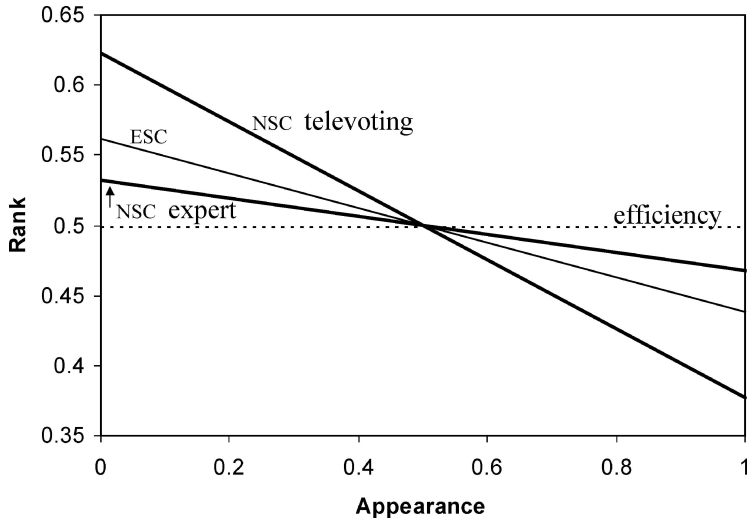

Figure 1. Estimated relationships between RANK and APPEARANCE.

the coefficient equals 0.5, the case in which the jury process is efficient. The second thing to note is that the coefficient of the interacted variable is indeed negative and significant, with a p-value of 0.0171. This establishes that public opinion leads to a decision that is arguably inferior to that of a team of experts. Hence, at least in our case, experts are a better judge of quality than the general public.[23]

Figure 1 summarizes the results of our regression analyses. When the jury process is efficient, the order of appearance should have no systematic effect on the final ranking of a contestant. In that case, any value of APPEARANCE should on average lead to a final RANK that equals 0.5. This is given by the dotted line. The two heavy lines give the estimated relationships between APPEARANCE and RANK for the two different NSC samples: the one with televoting and the one with expert juries. For reference, we also give the line that is implied by the coefficient estimate of 0.124 that we found for the variable APPEARANCE in the ESC sample. We calibrated this line such that it also passes through the point (0.5, 0.5). Note that the inefficiency in the televoting sample is remarkably large. The regression result implies that, *ceteris paribus*, the song that is performed first has a rank that is on average 0.245 lower than the song that is performed last. In a contest with 11 contestants, this boils down to roughly 2.5 places in the final ranking.

5. A Robustness Check

We have argued that experts are a more efficient judge of quality than the general public, in the sense that experts are less influenced by factors that are exogenous to the quality of the contestants. As we explained in Section 2, however, since 1960

expert juries, at least in the ESC, have been allowed to hear participating songs more than once. One could therefore argue that the results we find are purely driven by the fact that expert jurors have better information, in that they have heard the songs more often, and for that reason are able to make a more balanced judgment.

Especially in the last few years, a number of countries that use televoting have also used semi-finals to help determine which performer should represent their country at the ESC. In such a system, many contestants participate in a number of shows, all of which are shown on national television. From each show, the best contestant(s), to be determined by televoting, go on to the national final. In that final, the ultimate winner is determined, again through televoting. Hence, when contestants perform in the final, their songs are already known to the public. Therefore, if our results are solely driven by the fact that experts have heard the songs already at the time of the contest, then in televoting finals that were preceded by a semi-final, we would expect the same level of inefficiency as in NSCs judged by experts.

We collected data on 17 NSCs that were judged by the public and that were preceded by semi-finals in which the public was already able to see the songs being performed. Not all of these NSCs used televoting in a strict sense. In some, postcard voting was used, in which the public could send postcards indicating their favorite. In other instances, a random sample of viewers was used. In all cases the public ultimately decided upon the final ranking of contestants.

Table IV reports on a regression in which we use the final ranking in this sample of shows as the dependent variable, and the order of appearance plus an intercept as the independent variables. From the discussion in the previous section, the estimated value of the intercept can be directly compared to the coefficient of TRANSAPPEAR in Table III. For the regression in Table IV, we find an intercept of 0.641, where we found one of 0.623 in our original dataset. Hence, if anything, the inefficiency in the public's judgment in these shows is even larger than that for the shows in the original dataset. This strongly suggests that the greater inefficiency that we found in public judgment relative to the judgment of experts is *not* due to the fact that experts were able to hear the songs more often.

Table IV. Estimation results NSC with preceding semi-finals (dependent variable: RANK. *t*-values in parentheses)

Constant	0.641** (3.30)
Appearance	−0.282** (−3.91)

See main text for definitions of variables.
Note: The *t*-value reported for APPEARANCE is for the null hypothesis that APPEARANCE = 0.5.
*Significant at 5%-level.
**Significant at 1%-level.

6. Conclusion

The contribution of this paper is two-fold. First, following Glejser and Heyndels (2001), we provided additional evidence that there are ordering effects when judging music contests. We did so using two new data sets: one for international finals of the Eurovision Song Contest, and one for national finals. Moreover, we showed that an ordering effect exists not only for contests judged by experts, but also for those judged by the general public. In the contests we consider, participants that perform later do better on average, regardless of the fact that the order of appearance is determined randomly. In addition, we found evidence that the very first contestant also does substantially better on average. Why these order effects exist remains a mystery. One may argue that jurors are better able to remember the performance of later contestants, while the performance of the very first contestant also sticks in their mind. But *a priori* it is not clear why jurors would judge contestants they remember well, more favorably than contestants of whom their memories have faded.

Order effects can be a major source of economic inefficiency, not only in cultural contests, but also in other contexts where the quality of several candidates needs to be compared. Examples include job interviews and the grading of exams. Our results suggest that job candidates that either are the very first, or among the last to be interviewed, stand a better chance of being hired. Indeed, other fields also have started to address the relevance of ordering effects. For example, Stewart et al. (2002) use survey data to assess the public's willingness to pay for three different health care programs. They find that the order in which the three programs are presented to respondents has an influence on their willingness to pay.

The second and more innovative contribution of this paper is that we shed new light on the age-old question as to whether experts or the general public are better able to assess the quality of cultural output. To do so, we developed a method that enables us to address this question without having to resort to subjective quality judgments. We showed that, in our data, experts are unambiguously better judges of quality, at least in the sense that the outcome of contests judged by experts are less sensitive to exogenous factors that clearly do not influence the quality of output. Nevertheless, we showed that experts are not perfect, in the sense that their judgment does depend on such factors.

Of course, our results only shed light on part of the debate on the merit of expert judgment versus public opinion. It could very well be that the current public is a better predictor of the views of future experts than are current experts. A stronger ordering effect from the public does not rule out greater prescience at the same time. The standards that experts apply may still be inferior, in whatever sense, to the standards the common man applies. But we do show that, at least, experts apply these standards more consistently.

Admittedly, the data we used in this paper, those for Eurovision Song Contests, are a bit unusual, and not the first that spring to mind when one considers studying the judgment of quality of cultural output. Yet, the character of the data, with

contests that are very similar and only differ in that some are judged by experts and some by the general public, provides a unique opportunity to test for differences between the two. We believe that our results also generalize to other cases where the quality judgment of cultural output is an issue.

Acknowledgments

The authors thank Victor Ginsburgh, Joyce Jacobsen, Peter Kooreman, Lambert Schoonbeek, Adriaan Soetevent, Jan-Egbert Sturm, Linda Toolsema, and two anonymous referees for useful comments and discussion.

Appendix A: May I Have Your Votes, Please?

Throughout the years, the voting procedure in the ESC has often been changed. In this appendix we give the details for all contests. We only document the changes in the procedure, from year to year. If a year is not listed, the voting procedure has not been changed.[24]

In the first contest, in 1956, every country sent 2 jurors. Each juror voted for his/her favorite song, including that of her own country, on a scale from 1 to 10. For this contest, however, only the winner was announced, not the ranking of the other participants. Therefore, we do not include it in our data. In 1957, each national jury consisted of 10 jurors. Each juror voted for her favorite song. The number of votes determined the final ranking. In 1959 professional composers and publishers were banned from being a juror.

In 1962 the voting system changed again. Each juror could now choose three songs, awarding 3 points to her favorite, 2 to the second best and 1 to the third best. These points were aggregated to determine a ranking for each national jury. Each national jury then gave 3, 2, and 1 point to its three highest-ranked songs. In 1963 national juries were expanded to 20 members. Jurors now awarded 5, 4, 3, 2, and 1 point to their 5 favorite songs. This was aggregated to a vote of 5, 4, 3, 2, and 1 point for each national jury. In 1964 juries were scaled back to 10 members. Jurors could now divide 9 points freely over all (other) countries. Points were aggregated over national juries and translated to 5, 3 and 1 point corresponding to the national jury's three favorite songs.[25] In 1966 the EBU decided that members of every national jury should consist of representative members of the public. Juries were allowed to have light music and pop music experts but no professional composers, record manufacturers or publishers. The voting system of 1957, in which each juror only voted for her favorite song, was reintroduced in 1967.

Another change took place in 1971. Each country now had only 2 jurors, one under and one over 25 years of age. Each juror rated all songs on a scale from 1 to 5. All the individual scores were added to determine the final ranking in the contest. In 1974 national juries again consisted of 10 members, 5 under and 5 over 25, and preferably 5 men and 5 women. Minimum age was 16, maximum 60, with at least a 10 year age difference between the youngest and the oldest member. Each juror voted for her favorite song. In 1975, national juries had 11 members. Every member rated all songs on a scale from 1 to 5. Based on the total scores of its national jury, each country then awarded 12 points to its favorite song, 10 points to its second-favorite, and 8, 7, 6, 5, 4, 3, 2, and 1 point to its third through tenth favorite. This system was still in use in 1997, the end of our sample.

Since 1988 each national jury in the ESC has 16 members, with 4 between 15 and 25 years old, 4 between 26 and 35, 4 between 36 and 45, and 4 between 46 and 60. People with an interest in the music industry were barred from being a juror. Every jury member now rates songs on a scale from 1 to 10. The final vote system did not change. Nowadays, the tie-breaking rule is that the song that has received the highest number of maximum scores (i.e. 12 points) wins. In case that number is also equal, the number of second-highest scores is decisive, etc. In 1991 such a tie did actually occur.[26]

Appendix B: Sample of National Finals

Starting from 2001 and working backwards, we used data from all available national finals that either exclusively used televoting, or exclusively used expert juries, and had at least 7 contestants. We used all such national finals for which data are available, going back to 1988 for televoting, and to 1993 for expert juries.

The national finals that are included in our data are summarized in Table BI.

Table BI. Type of judgment used in the finals included in our data set

Televoting		Expert juries	
Belgium	1998 (10), 2000 (10)	Austria	1994 (8)
Denmark	1997 (10)	Bosnia-Herzegovina	1999 (17), 2001 (19)
Finland	1994 (10), 1996 (10)	Croatia	1993 (15), 1994 (21), 1997 (20)
Germany	1997 (9), 1999 (11), 2000 (11), 2001 (12)	Cyprus	1993 (8), 1994 (8), 1997 (8), 1998 (8), 1999 (9), 2000 (11)
Great Britain	1988 (8), 1989 (8), 1990 (8), 1991 (8), 1992 (8), 1993 (8), 1994 (8), 1995 (8), 2000 (8)	Estonia	1993 (8), 1994 (10), 1996 (13), 1997 (8), 1998 (10), 1999 (10), 2000 (10), 2001 (8)
Iceland	2001 (8)	Finland	1993 (8)
Ireland	1999 (8), 2000 (8), 2001 (7)	Hungary	1994 (15), 1997 (19)
FYR Macedonia	1998 (20)	Iceland	1993 (10)
Romania	2000 (13)	Ireland	1993 (8), 1994 (8), 1997 (8), 1998 (8)
Slovenia	1997 (13), 1998 (14)	Israel	1993 (12)
		Malta	1997 (16), 1998 (20), 1999 (16), 2000 (16)
		The Netherlands	1993 (8), 1994 (8)
		Norway	1993 (8), 1996 (8)
		Portugal	1996 (10), 1998 (8), 1999 (8)
		Slovenia	1993 (12)
		Sweden	1997 (12)
		Switzerland	1993 (7)
		Turkey	1998 (10)

Number of contestants in parentheses.

Note: In 2000 Great Britain held a semi-final with 8 contestants. The top 4 competed again the next evening to decide the winner. We use the 8-contestant semi-final.

Table BII. 'Public judgment' NSC finals with preceding semi-finals

Croatia	2003 (12), 2004 (12)
Finland	1984 (11)[a]
France	1977 (6), 1978 (6), 1980 (6), 1981 (6)[b]
Germany[c]	1983 (12), 1984 (12), 1985 (12), 1987 (12), 1988 (12)
Greece	1998 (9)
Lithuania	2004 (18)[d]
The Netherlands	2003 (8), 2004 (10)[d]
Slovenia	2004 (16)[d]

These festivals used regular televoting unless otherwise noted.
[a]Postcard voting rather than televoting was used.
[b]A representative sample of viewers was used. Its size is unknown.
[c]In all German festivals listed, a representative sample of 1,000 viewers was used.
[d]This final used a combination of televoting and an expert jury. However, the result of the televoting is separately known. We use this is in our data.

The national finals used for our analysis in Section 5 (and only in that analysis) are summarized in Table BII.

Notes

1. Wijnberg (1995) distinguishes three basic types of selection system for such cases: market selection, peer selection, and expert selection. In the case of market selection, the producers are the selected and the consumers are the selectors. In peer selection, on the other hand, the selectors and the selected are part of the same group. In the case of expert selection, the selectors are neither producers nor consumers, but have the power to shape selection by virtue of specialized knowledge and distinctive abilities. See also Wijnberg and Gemser (2000).
2. Holbrook (1999), p. 144.
3. The quotes in this paragraph are from Lipsitz (1999), in a review of Cowen (1998).
4. In economics, there is a small literature that looks at how experts and the general public assess the quality of movies. Holbrook (1999) tries to assess which movie characteristics have a positive influence on either popular appeal or critical acclaim. Ginsburgh and Weyers (1999) claim that the general public is more time consistent in their evaluation of movies. This is based on the following observations. Box office receipts are strongly and positively correlated with the number of times a movie appears on television after having been produced. There is a much larger discrepancy, however, between movies that win awards and those that make it to critics' best movie lists many years later.
5. Unless, of course, if the order of appearance influences the quality of the performance. There is little reason however to assume that this is the case.
6. These countries were Austria, Belgium, Denmark, France, Italy, Luxembourg, the Netherlands, Switzerland, the United Kingdom, and West-Germany.
7. For example, the draw for the 2002 contest took place on November 9, 2001 (see EBU/UER, 2001).
8. At least, the vocals are. Originally, it was allowed to have some instrumental backing on tape. Effective beginning in 1997, all instrumental backing can be on tape and only vocals have to be performed live.

9. The most notable examples are the Swedish band ABBA, who won the 1974 contest, and Canadian singer Celine Dion, who won in 1988 representing Switzerland.
10. This is particularly true for Luxembourg, which has won the contest 5 times – but never while being represented by a singer with Luxembourg nationality.
11. In 1993 there was a pre-contest in which 7 countries competed for 3 places in the ESC. In 1994 the 7 worst-performing countries were not allowed to participate the following year. In 1996 there was a pre-selection, in which all countries that wanted to participate had to send their song on tape to the EBU. A system of national juries, different from the juries in the finals, then decided which countries could participate in the actual contest. In 1997 the qualifying stage changed again. Each country could now participate at least once every two years. The other contestants were the winner of the previous year, plus the countries with the highest average score during the previous 5 years. Since 2001, the 13 highest-scoring countries in a given year automatically qualify for the next year. The numbers 14 and 15 may also qualify – dependent on the exact number of members of the Big Four that qualifies among the highest-scoring countries.
12. Italy is also entitled to participate in every single contest, as a tribute to the fact that the contest was modelled on the San Remo festival. However, Italy no longer chooses to exercise that right; since 1994, it has only participated once.
13. Such a case occurred, for example, in the Netherlands in 2000, when transmission of the ESC was interrupted to allow for news coverage on a major accident that took place in the town of Enschede.
14. Clearly, there are other such exogenous factors as well, which could also be taken into account in an analysis. In our study, we focus solely on the order of appearance, as this is the factor that is by far the most easy to observe. Note that failing to take other exogenous factors into account does not affect the analysis, as long as these factors are uncorrelated with the order of appearance, which they necessarily are, as the order of appearance is random.
15. As one referee argued, the co-existence of two parallel voting systems at the ESC level should allow testing our main hypothesis in addition to our present analysis that only uses NSE data. A major advantage would be that the analysis is not complicated by organisational differences between countries. Unfortunately however, the number of countries using expert jurors after 1998 is very small, so there is an insufficient number of observations to do this analysis.
16. These countries are Austria, Belgium, Bosnia-Herzegovina, Croatia, Cyprus, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Lithuania, Luxembourg, Malta, Morocco, Monaco, the Netherlands, Norway, Poland, Portugal, Russia, Roumania, Slovenia, Slovakia, Spain, Sweden, Switzerland, Turkey, the United Kingdom, and Yugoslavia. Most former communist countries have only participated since the 1990s.
17. Note that the issue of "home advantage" is well established in sports economics (see e.g., Vergin and Sosik, 1999, who test whether betting markets take this advantage into account in a manner that is consistent with efficient markets). In sports, this advantage is attributed to learning factors (e.g., familiarity with the stadium and its playing surface), travel factors (visiting teams experience physical and mental fatigue and disruption of routine), and crowd factors (crowds may provide social support) (see Schwartz and Barsky, 1977). Courneya and Carron (1991) suggest that referee bias in favor of the home team contributes toward home field advantage. Learning factors can hardly be a factor in the ESC, as the contest is only held once in each venue. Travel factors also seem of minor importance, as contestants already travel to the venue one week in advance. However, crowd factors and referee bias (or rather, jury bias) may be an issue.
18. Except when this is Denmark. See above.
19. Again, note that a *lower* rank reflects a *better* performance. Hence the countries that do systematically better, have a negative value for their dummy.
20. We also tested whether there is an additional effect of being the last performer, by including a dummy CLOSING. This dummy, however, was not significant.

21. Honesty (and an anonymous referee) obliges us to admit that we cannot guarantee that the order of appearance in all NSCs under consideration are determined in a random fashion as well. Nevertheless, we are reasonably confident that this is the case. Most probably, contestants would object if the order of appearance were determined in any other manner. And even if in some contests the order of appearance is not determined by a random draw, we would be in trouble only if this would be the case systematically more often in contests with televoting than in contests with expert judgment – or the other way round.
22. At the time of our initial data collection.
23. Note that (3), the equation we estimate, only follows from (1), the original specification, under the assumption that RANK(0.5) = 0.5. When we add dummies to the original specification, this condition is no longer satisfied. Therefore, it does not make sense to add dummies for the opening act or the type of performer to (3).
24. Source: Schwarm-Bronson (2001).
25. In theory, we could have a case in which all jurors of a national jury awarded all of their points to just one song. Should that occur, all 9 points of a national jury would have been awarded to that song. When only two songs were to receive points from a national jury, this would have been translated to 6 points for the highest-scoring song, and 3 points for the other. These contingencies, however, did not occur.
26. The 1969 contest had no fewer than 4 winners, but in that year there was no tie-breaking rule yet in place.

References

Courneya, K.S. and Carron, A.V. (1991) "Effects of Travel and Length of Home Stand/Road Trip on the Home Advantage." *Journal of Sport and Exercise Psychology* 13(1): 42–49.

Cowen, T. (1998) *In Praise of Commercial Culture*, Harvard University Press, Cambridge and London.

EBU/UER. (2001) *Rules of the Eurovision Song Contest 2002*, EBU/UER, Geneva, http://www.ebu.ch/tv-cec_2002_rules.pdf

Eilers, A. (2002) *Eurovision Song Contest Record Covers*, http://members.fortunecity.com/mcdeil69/

Eeuwes, G. (2002) *www.songcontest.nl*, http://www.songcontest.nl

Glejser, H. and Heyndels, B. (2001) "Efficiency and Inefficiency in the Ranking in Competitions: The Case of the Queen Elisabeth Music Contest." *Journal of Cultural Economics* 25: 109–129.

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

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

Ginsburgh, V. and Weyers, S. (1999) "On the Perceived Quality of Movies." *Journal of Cultural Economics* 23: 269–283.

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

Lipsitz, G. (1999) "T. Cowen, In Praise of Commercial Culture, Book Review." *Journal of Economic Literature* 37: 1741–1742.

Musgrave, R.A. (1959) *The Theory of Public Finance*. McGraw Hill, New York.

Ramsey. (1969) "Tests for Specification Errors in Classical Linear Least Squares Analysis." *Journal of the Royal Statistical Society, Series B* 31: 350–371.

Schwarm-Bronson, N. (2001) *Eurovision Song Contest: the Story*, EBU/UER, Geneva, http://www.ebu.ch/tv-cec_story.pdf

Schwartz, B. and Barsky, S.F. (1977) "The Home Advantage." *Social Forces* 55: 641–661.

Stewart, J.M., O'Shea, E., Donaldson, C. and Shackley, P. (2002) "Do ordering effects matter in willingness-to-pay studies of health care?." *Journal of Health Economics* 21: 585–599.

Stoddart, C. (2002), *Eurovision Song Contest National Finals Homepage*, http://www.geocities.com/national_finals/

Vergin, R.C. and Sosik, J.J. (1999) "No Place Like Home: An Examination of the Home Field Advantage in Gambling Strategies in NFL Football." *Journal of Economics and Business* 51: 21–31.

Walraven, H. and Willems, G. (2000) *Dinge-dong – Het Eurovisie Songfestival in de twinstigste eeuw*, Forum, Amsterdam.

Wijnberg, N.M. (1995) "Selection Processes and Appropriability in Art, Science, and Technology." *Journal of Cultural Economics* 19(3): 221–235.

Wijnberg, N.M. and Gemser, G. (2000), "Adding Value to Innovation: Impressionism and the transformation of the Selection System in Visual Arts." *Organization Science* 11(3): 323–329.