Interdisciplinary Studies in Musicology 10, 2011

MARIA MANTURZEWSKA (Warszawa)

The reliability of evaluation musical performance by music experts

ABSTRACT: To test the reliability of the evaluation of musical performances by musical experts, the protocols of the jury of an international music contest have been subjected to statistical analysis.
The object of the analysis were 2156 jurors' points and rank ratings, given by 28 members of the jury, assessing 77 different performances of one of Fryderyk Chopin's polonaises, evaluated during the first stage of an international music competition.
The analysis revealed the following: 1. very large interpersonal (inter-rater) differences of the jurors' ratings, 2. despite these differences, there was a very high level of statistical significance of the inter-rater agreement ($p<.001$) of the jurors' evaluation, 3. despite the high level of statistical significance of inter-rater agreement of the evaluation of musical performance, this accounts for only 1/3 of the general variance of ratings.
Conclusion: individual ratings of musical performance are not a reliable measure of musical achievement, even when given by music experts of the highest level.

KEYWORDS: evaluation, musical performance, music contest, reliability, music experts

Evaluation of performances occurs
in the everyday activity of music critics,
music teachers and musicians.
However, there are hardly any agreed
criteria either for what should be judged,
or for how the judgments should be made.
Judges may be unaware of what criteria
they actually use in their assessments[1]

[1] Alf Gabrielsson, 'Music performance research at the millennium', *Psychology of Music* 31/3 (2003), 255.

Introduction

Evaluation of musical performance by music experts is one of the basic criteria in assessing the achievement both of students and professional music performers. They often determine the educational and artistic careers of musicians. Expert ratings are also among the most frequently used external criteria in the validation of tests of musical ability and achievement. It is therefore important and useful, both from the point of view of the psychology of music and its application in psychological counseling and promotion of musical talent, as well as from the point of view of musicology, music education, musical praxis and musical life, to assess the reliability of expert ratings.

A study of the assessment of piano performances by music experts was one of the topics in a large research project carried out in 1965-1967 in collaboration by the Psychometric Laboratory, Polish Academy of Sciences in Warsaw, and the Chair of Sound Engineering at the Warsaw Music Academy.

The problem emerged during my research on the psychological determinants of the pianists' success, where the basic group (subjects) consisted of participants in the VI International Fryderyk Chopin Piano Competition.

In that research I used the jurors' ratings as the main criterion of musical achievement of the investigated pianists.[2] Naturally, this meant that I was interested in the reliability of this criterion.

I wanted to know whether the members of the jury of the International Piano Competition who assessed the participants of the contests were consistent with each other as to their ratings. I also wanted to know if their ratings were comparable with the evaluation of the same piano performances by other music experts who were not on the jury. I was also interested in finding out if the evaluations by music experts of musical performance were stable, i.e., whether music experts, assessing the same performance twice, evaluate it in the same way.

I also wanted to know if the ratings given by music experts differed from the ratings of the same musical performances by music novices.

To receive the answers to all these questions, four sets of ratings were analyzed:

(1) 2156 ratings, given by the 28 members of the jury of the sixth International F. Chopin Piano Competition in Warsaw, evaluating 77 performances of one of Chopin's Polonaises at the audiovisual rehearsals of the first stage of this competition in the concert hall;

[2] Maria Manturzewska, *Psychologiczne warunki osiągnięć pianistycznych* [Psychological determinants of pianist achievement] (Wroclaw, 1969); Maria Manturzewska, *Badania nad rolą psychicznych, fizycznych i biograficznych wyznaczników powodzenia w zawodzie pianisty* [Research into the role of psychological, physical and biographical indicators of success of career pianists], unpublished doctoral thesis, (Kraków, 1963).

(2) 70 ratings and free verbal assessments of 7 blind performances of the Polonaise-Fantaisie op. 61 (5 different and 2 repeated) selected from the above-mentioned 77 contest performances and evaluated by 10 carefully selected music experts in an individual laboratory setting;
(3) 420 ratings and verbal characterisation of the same 7 performances of Chopin's Polonaise-Fantaisie Op.61), given by 60 professional pianists, professors at four different music academies in Poland under the same conditions as group (2), but in a group setting;
(4) 210 quantitative ratings and verbal descriptions (characteristics) of the same tape recorded 7 performances of the Chopin Polonaise as in (3), given by 30 musical novices.

The ratings were analyzed in order to obtain quantified answers to the following questions:
(1) to what extent are the ratings of the different members of the international jury consistent?
(2 to what extent are the ratings of the members of the jury comparable to the ratings of music experts not on the jury?
(3) to what extent are the music experts' ratings stable, i.e., do they give the same ratings when rating the same performance twice?
(4) do music experts' ratings differ from the ratings of music novices?

The research was designed to be exploratory. I was at that time primarily interested in the reliability of the evaluation of musical performances by musical experts. This issue seems to me to be still worth testing today, because, despite very heated discussions and much controversy over the idea of music contests, they often determine the future artistic careers of young musicians. Yet we still don't know enough about either reliability of the ratings in such contests, or about the criteria for assessing musical performances and the psychological processes involved in it.

For this reason I decided to submit for public discussion the results of my earlier research on the problem of the evaluation of musical performances by music experts, in the hope that some young researchers interested in this problems, might, after reading my paper, be willing to repeat this research, basing its data on other international music contests, and be able to comment on my results in the light of modern theories of the psychology of evaluation.

I was fortunate in that the jury of the contest which served as the subject of my research included music experts of the highest international reputation; on the other hand, that contest took place in 1960, when the political situation in the world and in Europe had some special characteristics, which probably may have affected some of the members of the international jury of that competition.

Today, for my contribution to the volume honouring Professor Andrzej Rakowski, I have decided to use only the first part of the report on my research. This part concerns the analysis of the protocols of the jury. It primarily, but not only, addresses the problem of concordance of the jurors's verdicts on musical performances.

This is the first time that my research is being published in English.

Other parts of this research were published in Polish in a number of journals and books.[3]

The concordance of the evaluation of musical performances by music experts. (Study of the jurors of the music competition)

Our first research question may be formulated as follows: Which ratings from a 26-point scale (0-25) are most frequently chosen by the jurors, and what is the distribution of the ratings given during the contest? We predicted that the average ratings (12-15) would be given most frequently. However, we wanted to know the ratios of average to extreme ratings.

An analysis of the distribution of ratings given during Stage One of the contest by 28 members of the jury rating 77 pianists revealed that the most frequent ratings were 15 and 16. This represents as much as 19 percent of the total pool of ratings, i.e., 2.5 times more that could be expected from random distribution. Extremely high (24 and 25) and extremely low (0, 1, and 2) ratings constitute only 5 percent of all ratings.

[3] Maria Manturzewska, 'Zgodność ocen wykonawstwa muzycznego wydawanych przez ekspertów muzycznych' [Inter-rater consistency of music experts rating musical performance], *Biuletyn Psychometryczny* 1 (1966); Maria Manturzewska, 'Z badań nad ocenami wykonawstwa muzycznego wydawanymi przez ekspertów muzycznych' [Studies of ratings of musical performance by music experts], *Zeszyty Naukowe* 4 (Warszawa, 1968); Maria Manturzewska, 'O trudnej sztuce oceniania wykonań i wykonawców muzyki Chopina' [About the difficulties of the evaluation F. Chopin performances and performers], in *Muzyka w kontekście kultury* eds. Małgorzata Janicka-Słysz, Teresa Malecka, Krzysztof Szwajger [Music in the context of culture] (Kraków, 2001).

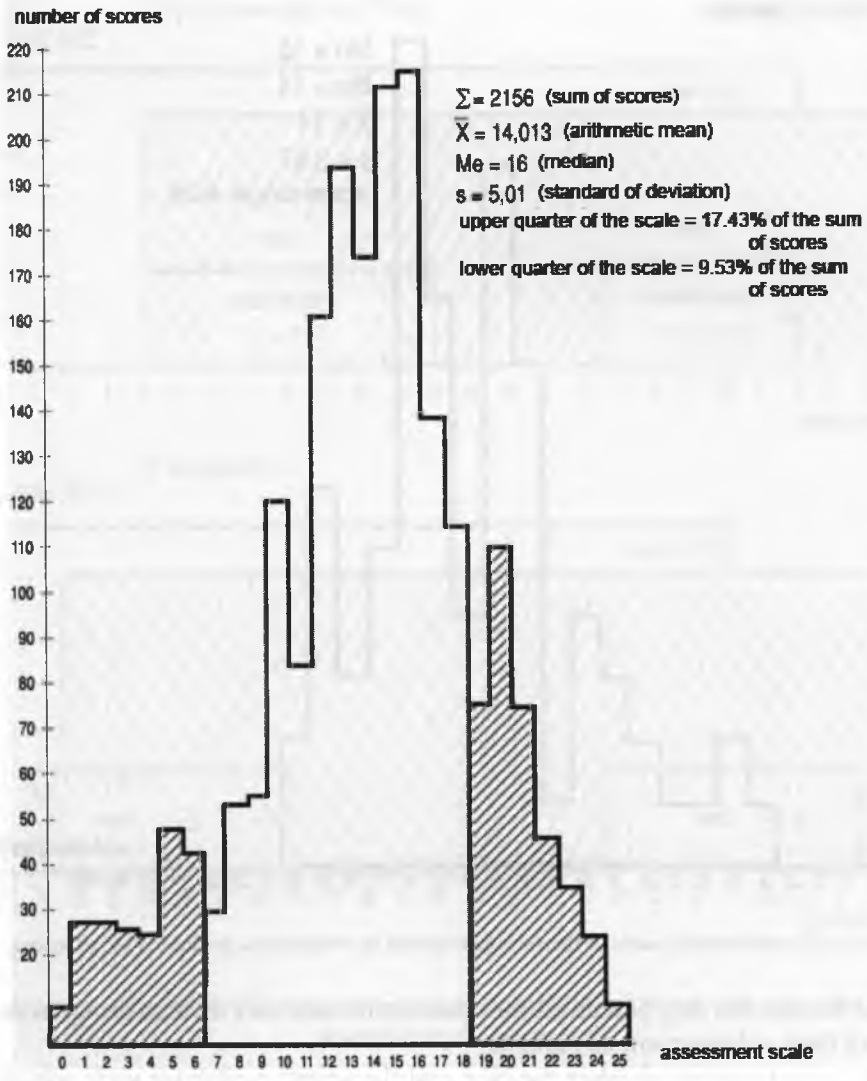

Figure 1. Ratings in scale 0-25, given by 28 jurors, of the International F. Chopin Piano-competition, evelicating 77 participants of this contest.

Before computing the coefficient of inter-rater consistency we analyzed the distribution of mean ratings obtained in Stage One of the contest by each of the 77 candidates.

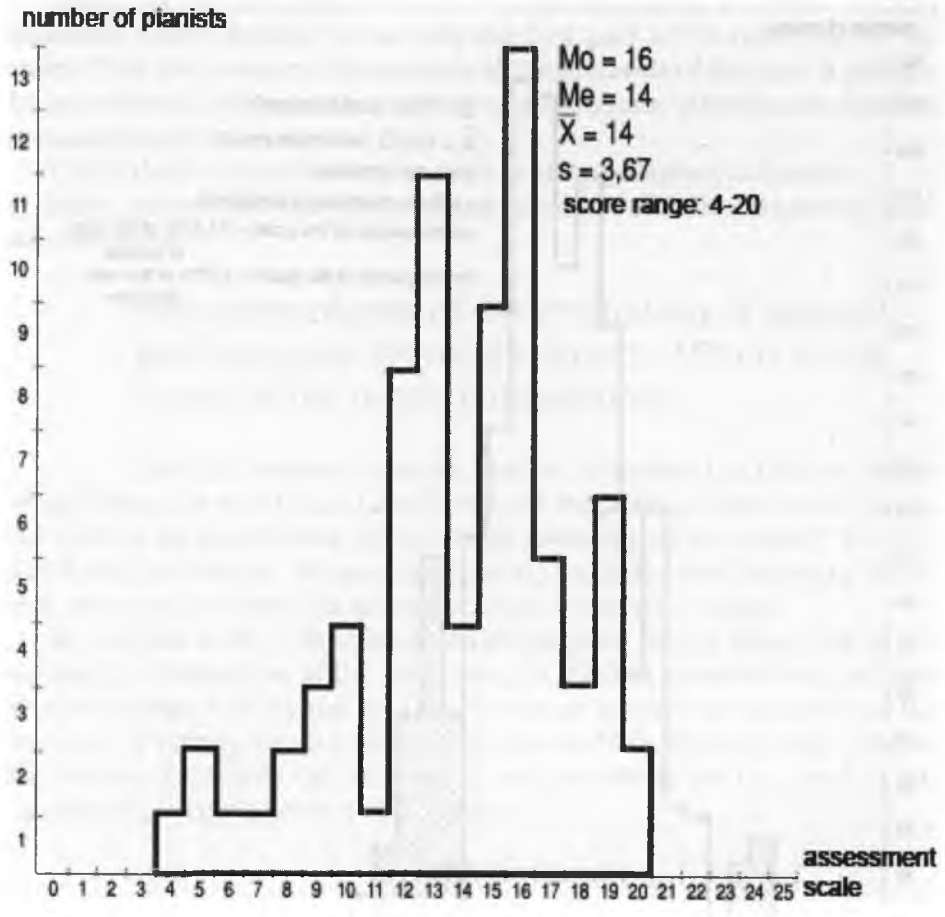

Figure 2. The arithmetic mens of the ratings received by 77 pianists, evaluated by 28 jurors.

As we see, the jury perceived the contestants as a very diverse group in regard of their achievement as pianists.

In order to check whether the best and the worst candidates were consistently rated by the majority of the jury, we compared the dispersions of individual ratings and the arithmetical means of raw scores within two extreme quartiles of pianists (25 percent of the best and 25 percent of the worst contestants). The results are shown in Fig. 3. The dispersion of individual scores is very large and the ranges of scores of the best and worst pianists overlap very considerably. The ratings of the best pianists range from 4 to 25 points, $\bar{x}$ = 18.96, s = 3.11, whereas those of the worst pianists range from 0 to 22 points, with $\bar{x}$ = 8,00, s=3,5

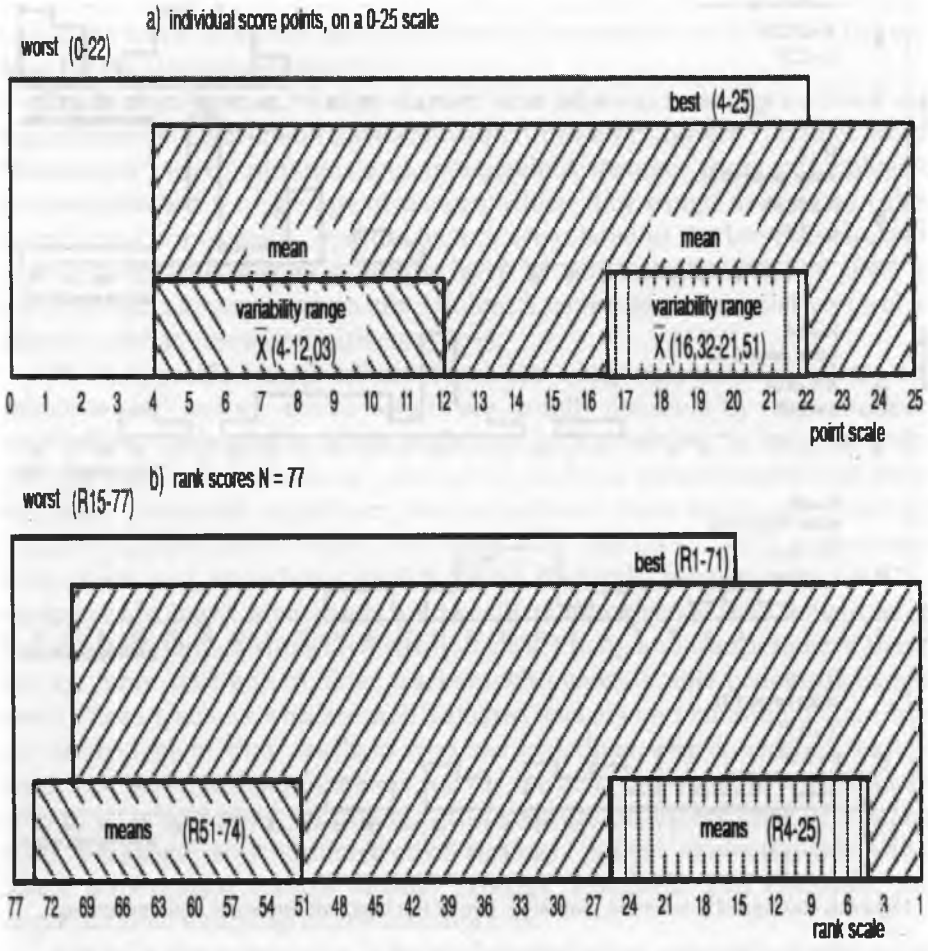

Figure 3. Range of the individual differences in ratings 25% of the best and 25% of the worst pianists from the group of 77.

Mean scores of the two groups differ with significance level of $p < 0.001$ but individual scores can be quite bewildering.

In order to analyze the dispersion of ratings given to individual pianists by the 28 members of the jury, we compared the distributions of ratings given to five different pianists for their performance of the Polonaise-Fantaisie op. 61. The performances selected were: two best, two worst, and one average. Figure 4 shows the distributions of ratings for each pianist. As we see, the range of ratings given to the same pianist by different judges is very large, from seven to fifteen points. Differences amount to two fifths of the entire scale.

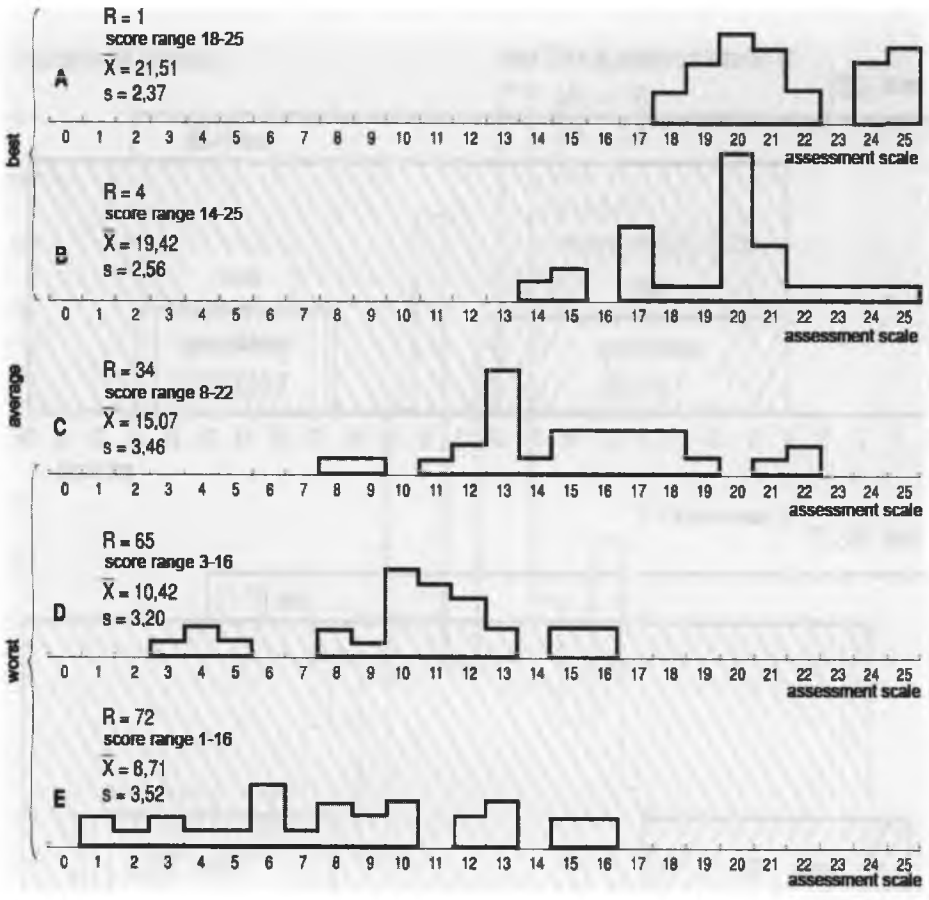

Figure 4. Ratings of 5 sellected pianists – two of the best two the worst and one average.

Here we may ask which judge was "right": the one who gave performance B 14 points, i.e., fair, or the one who gave 25 points, the highest possible rating. And who is right for performance E? The judge who gave 1 point (almost the worst rating possible) or the one who rated the performance at 16, i.e., good? We might say the truth lies in the middle, but where is the middle? Depending on who is on the jury, the "middle" will be at different points of the scale.

At all international contests, the arithmetic mean of rating is customarily accepted as the measure of achievement. This mean indicates the central tendency of the distribution of ratings obtained by a given pianist from all the judges. However, the psychologist may be intrigued as to why members of the jury, who have the highest possible competence as regards piano performance, differ to such an extent in their ratings of different pianists in a maximally simple situation. All they have to do is to assess the performance of the

same, well-known piece, by a well-known composer, at a contest with old traditions and a relatively stable standard of requirements. What are the reasons for these inconsistencies?

Studies by Fishman,[4] Kelly,[5] Carter,[6] and others concerning teachers' ratings of school achievement have shown that these ratings are variables which "summarize" many different and independent factors. These psychologists have established four groups of factors, which they categorized as so-called specific and non-specific. Specific factors are related to the level of scholastic achievement and include absolute level of achievement, relative level of achievement as compared to the student's capacities, and relative level of achievement as compared with class level.

The nonspecific factors fall into three sub-categories: pupil-, teacher-, and school-related factors. Better results are usually obtained by those students who behave like model students as defined by the teacher. It was also found that the pupil's sex, manners, personality, physical attractiveness and socio-economic status are significant determinants of their achievement ratings. Teacher characteristics which determine their ratings of pupil achievement include sex, age, experience, qualifications, demands, temperament, and "assessment ideology". Male raters are usually more objective than female raters but they also tend to show favoritism to girls. Young and inexperienced teachers are more severe than older teachers with considerable pedagogic experience. Those teachers who are well adjusted socially and emotionally are usually more lenient than teachers who are maladjusted and schizoid. Assessment can serve various functions for the teacher. For some it is a means of reward or punishment, for others it is a measure of achievement or a part of their pedagogic and administrative strategy. School characteristics which affect achievement ratings include prestige, standards and demands, and organizational and programme assumptions.

Although the assessment of musical performance, especially during an international contest, may not seem comparable with assessment of school achievement, we may assume that at least some of the determinants of school-achievement-rating may interfere with performance rating, causing inter-rater discrepancies. In order to test the significance of only one of the subjective determinants of assessment, i.e., between-judge differences in requirement standards, we calculated the means of ratings given by each of the 28 judges to the same 77 pianists in the same contest conditions. The distri-

---

[4] Joshua A. Fishman, 'Unsolved Criterion Problems in the Selection of College Students', *Harvard Educational Review* 28 (1958), 340-349.

[5] Eldon G. Kelly, 'A Study of Consistent Discrepancies between Instructor Grades and Term-End Examination Grades', *Journal of Educational Psychology* 49 (1958), 328-334.

[6] Robert E. Carter, 'Non-Intellectual Variables involved in Teachers Marks', *Journal of Educational Research* 47 (1953), 81-95.

bution of these mean ratings is shown in Figure 5a. As can be seen from the figure, the judges differed considerably as to their requirement standards. The set of performances of just one piece was given a mean rating of 10 by one judge and 21 by another. Therefore, according to one judge the candidates were very poor whereas according to the second one they were very good. Which judge was right and what do the various points on the rating scale mean for each of them? Figure 5b shows the individual differences between two jurors, evaluating the same group of pianists.

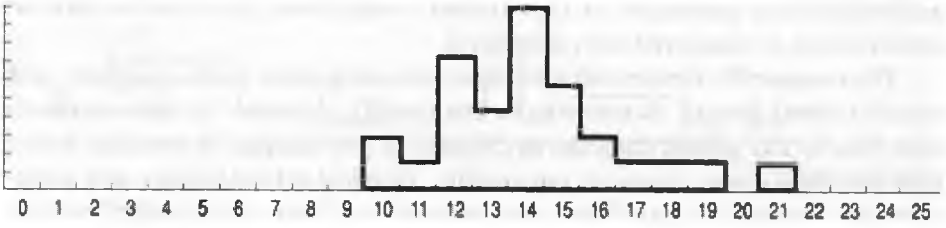

Figure 5. a The means of ratings given by each of 28 jurors, evaluating 77 pianists.

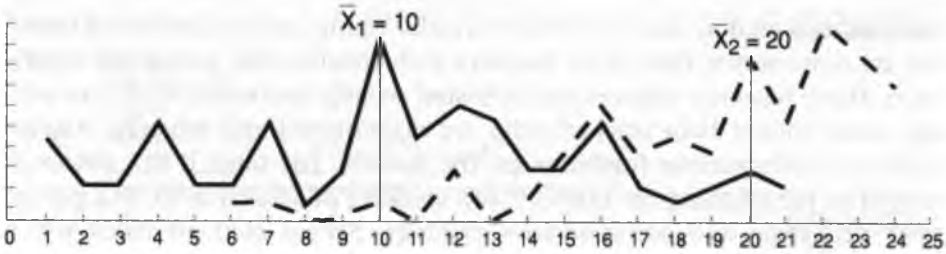

Figure 5. b. Ratings given by two different jurors, evaluating the same group of pianists.

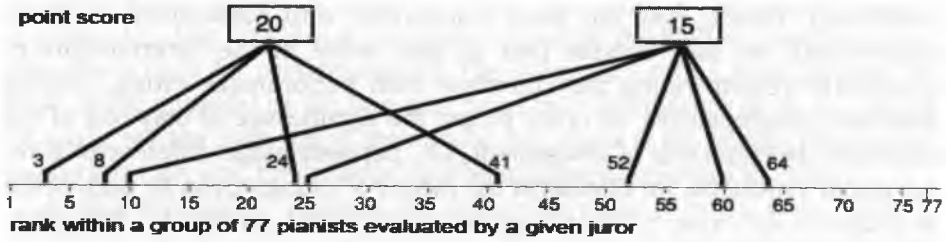

Figure 5. c. The relative value of the same points ratings, given by different jurors.

For each individual judge the rating scale is at least an ordinal scale, i.e., the pianist rated 25 is better than the pianist rated 20 by the same judge and this pianist, in turn, is better than the one rated 15, etc. However, the points

ratings 20, 25, or 10 given by one judge are not equivalent to the same performance score as the same points ratings given by a different judge. To investigate the relativity of the points ratings I decided to transform the point ratings into rank-order ratings and order the pianists according to the rank-order individually for each of the judges. The point scores are transformed into rank-order scores by giving the first rank to the pianist who obtained the highest point score, and the 77^{th} rank to the pianist who obtained the lowest point score from that judge, irrespective of the values of the point scores given by this expert.

When all the point scores were transformed into rank-order scores, we found that the same point scores were equivalent to different rank-order scores (Fig. 5c). For instance, for one judge, point 15 was equivalent to rank-order 8, whereas for another judge point 15 was equivalent to rank-order 24, and for a third judge - rank order 65, i.e., one of the last. Score 20 given by one judge placed the pianist in position 5 in the rank ordering, whereas the same score given by a second judge placed the pianist in position 21, and by a third judge in position 41 in the group of 77 pianists.

As we see, therefore, only one factor, i.e., individual differences in requirement standards, can considerably affect individual assessment of performances. The conclusion from this finding is that we must be very cautions in accepting point scores for performance even when they are given by people with the highest level of competence in instrumental performance. In order to check whether transformation of scores into rank-order would improve between-judge consistency, we once again compared the distribution of individual scores of the top and bottom quartiles of pianists, this time basing our analysis on rank-order positions. The results of this analysis are shown in Fig. 3. As we see from the figure, when individual differences in requirement standards are eliminated by transforming point scores into rank-order scores, the between-judge discrepancies drop only slightly, i.e., from 80 to 75 percent of the scale for individual scores and from 16 to 12.5 percent for mean scores. In order to obtain quantitative information about the degree of between-judge consistency in our international jury, we applied two statistics, i.e., (Fig. 6a) matrix inter-correlations for the 28 judges were calculated and Kendall's W test was used to assess the coefficient of consistency based on analysis of variance of rank-order scores (Fig. 6b). Both statistics revealed that the contest jury, despite the discrepancies discussed earlier, are consistent in their assessment of musical performance (the coefficients are significant) $\bar{r} = .596$, $W = .61$, both significant statistically at the level $p < .001$, i.e., contests jurors ratings are not random but are based on some sort of common criteria. However, these common criteria account for only about one third of the common variance. The remaining two thirds of the variance are attributable to differences in criteria in either individual judges or judge subgroups.

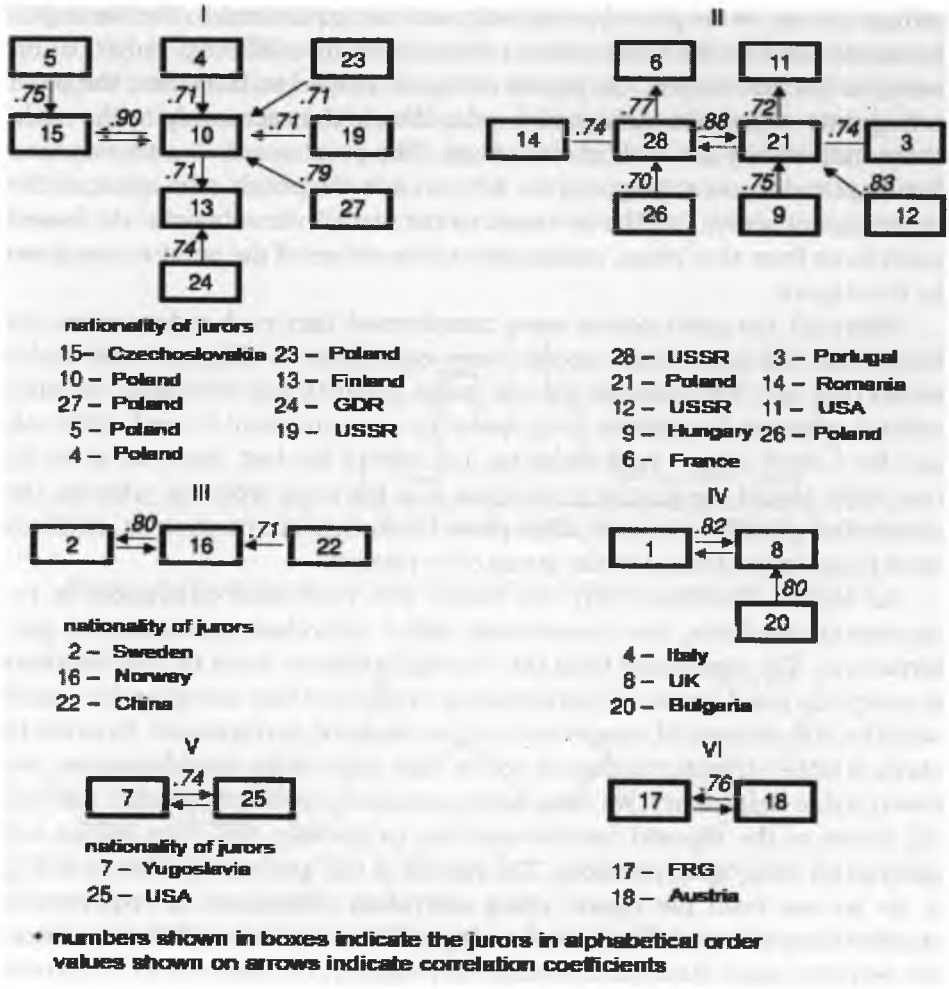

Figure 6. Groups of jurors after McQuitty cluster analysis.

In order to test whether it is possible to distinguish different subgroups of judges, we used the McQuitty cluster analysis which consists in grouping judges whose scores correlated highest. This enabled us to distinguish six subgroups of judges.

The information we had about the judges implied that one of the clustering criteria was nationality. Separate clusters were found for German and Austrian, Scandinavian, American, Italian, and English judges.

However, nationality was not the only variable responsible for differences between judges. The two largest clusters both contained judges from Poland, the Soviet Union, and Czechoslovakia.

Analysis of personal data suggested that the responsible factors were personality traits and different schools of pianist interpretation.

Conclusion

As we can see, reliable evaluation of musical achievement is not easy even for the highest level, very competent music experts, and even in such a typical situation as the evaluation of performances of a very well known, traditional, quite brief piece of music.

All of us: music psychologists, music teachers and musicologists, should be conscious of this fact and this truth when we use evaluation by musical experts as the criterion of musical achievement or musical talent.

My old research only touched on this very important and very difficult problem. It is my view that we should continue this research, to try to explain the difficulties of music evaluation, to analyze them, in order to understand the multidimensional process of music reception and evaluation, their structure, course and their determinants, both external and internal, specific and non–specific. We should be using the new paradigm of psychology of evaluation, and different methods of research, both quantitative and qualitative, psychometric and humanistic, interpretative approaches.

Translated by Helena Grzegułowska-Klarkowska