RICHARD PARNCUTT

THE RELIABILITY/ VALIDITY OF COGNITIVE/EMOTIONAL APPROACHES TO THE EVALUATION OF MUSICAL PERFORMANCE: IMPLICATIONS FOR COMPETITION JURIES

Introduction

Psychological research into music competitions can help organisers refine their procedures to increase the probability of identifying and rewarding the best musicians or the best performances. The aim of this contribution is to survey some relevant empirical literature, looking for findings that could be especially useful for competition organisers. On that basis, I will recommend changes to procedures for identifying winners in music competitions such as the Fryderyk Chopin International Piano Competition. Given the complexity of the performance evaluation situation, only part of which can be considered in a typical psychological study, my recommendations will necessarily be tentative and controversial.

The relevant psychological research is scientific. It belongs to the natural, social and formal sciences rather than the humanities. Today's psychologists tend to avoid qualitative, intuitive, speculative, theoretic and hermeneutic approaches, preferring approaches based on quantitative measurement and modelling. Research of this kind tends to be positivist, in the sense that the researcher is looking for 'truth' that is independent of time, place and context, whether that context be social, historical, political or cultural. When considering the challenge of identifying the best performer at a music competition, a scientist will tend to assume that a clear answer to this question exists.

Humanities scholars tend to question or even reject much of this scientific logic. Interesting 'truths' in such matters can rarely be considered absolute. Ideas can only be evaluated for truth-content relative to context. There is no such thing as the 'best' musician at a music competition; we can only speak of the musician whom the judges consider to be the best. The choice of winner depends on the judges' personal criteria and other aspects of the social, historical, political and cultural context. This limitation in turn restricts the conclusions that can reasonably be drawn from the relevant scientific research. Given the contrasting approaches of scholars in the humanities and sciences to these and other questions, I aim in this contribution to interpret the results of scientific studies from a humanities viewpoint, emphasising the dependence of results on contexts of different kinds.

At this point, I should mention relevant aspects of my personal background, because they could bias my approach and conclusions. I have a music degree from the University of Melbourne, which included a strong performance component. My instrument is

the piano, and I have considerable experience as both soloist and accompanist in concert performance of the 'classical' piano repertoire, but most of these activities stopped about two decades ago. I have published research into expressive timing and dynamics in piano music that is relevant to evaluations of performances of nineteenth-century piano repertoire.[1] I have also published research into piano timbre,[2] emphasising that although a pianist is unable to change the physical properties of individual tones independent of their loudness, there is considerable scope for changing the timbre of combinations of tones or musical passages, and the audience's impression of timbre also depends on the pianist's visible movements. I sympathise with the argument that a musical performance includes body movements (gestures) and interaction with the audience, but if asked to evaluate a performance, I tend to focus on timing and dynamics – the sound alone. Finally, from a political viewpoint, I am anxious to treat all entrants in music competitions equally, in spite of differences in appearance, cultural background and socio-economic status.

Reliability versus validity

A music competition is a test of musical performance ability, which in turn is an amalgam of partly independent skills.[3] As such, a music competition can be compared with a psychological test to measure specific abilities.

Psychologists place great emphasis on evaluating and standardising such tests before applying them. Two aspects that are often considered are reliability and validity. A reliable test is one that produces essentially the same result on different occasions, whereas a valid test is one that measures what it is purported to measure and not something else. The reliability of a music competition, when considered as a kind of psychological test, is its ability to identify the same winner(s) on different occasions, assuming that the same musicians enter the same competition at different times or in different places. Reliability becomes problematic in high-level competitions in which the best competitors are all excellent, making it difficult to choose between them (hence the so-called ceiling effect).

Validity is also an important issue, given the inherent difficulty of defining the quality of musical performance. Is a good performance one that accurately reproduces the musical score, conveys the composer's apparent intentions, gives a work a surprising new interpretation based on the personal voice of the musician, or merely excites the audience? If it can be all of those things, how should such aspects be evaluated relative to each other? How important are audible mistakes by comparison to the quality of the interpretation? Experts will always disagree on such questions,

1
Richard Parncutt, 'Accents and expression in piano performance', in Klaus Wolfgang Niemöller and Bram Gätjen (eds), *Perspektiven und Methoden einer Systemischen Musikwissenschaft* (Frankfurt am Main: Peter Lang, 2003), 163–185.

2
Richard Parncutt, 'Piano touch, timbre, ecological psychology, and cross-modal interference', in Aaron Williamon and Werner Goebl (eds), *Proceedings of the International Symposium on Performance Science* (Brussels: European Association of Conservatoires, 2013), 763–768.

3
Richard Parncutt and Gary E. McPherson (eds), *The Science and Psychology of Music Performance: Creative Strategies for Teaching and Learning* (Oxford: Oxford University Press, 2002).

because musical culture is constantly changing and developing. So music competitions can never be perfectly valid in the psychological sense – just as they can never be perfectly reliable.

There is a trade-off between reliability and validity: procedures that are more reliable tend to be less valid and vice versa.[4] A reliable procedure may be one in which different aspects of a performance are separately evaluated and an overall score is calculated on the basis of those different aspects. This can be done only at the expense of validity, since different experts will disagree on which points should be included in such a list and how important they are relative to each other. A competition should be open to new developments that could not have been predicted in advance. An inflexible set of factors is hardly likely to enable jurors to respond adequately to excellent performers that exceed their highest expectations. For that reason, good competitions tend not to constrain jurors by asking them to evaluate different aspects of a performance. Instead, jurors are given the freedom to evaluate along any dimensions they consider appropriate, taking advantage of their special expertise. The result is an increase in validity at the expense of reliability. In the end, validity is more important.

Of central interest to competition organisers and participants are the many sources of bias that could affect a jury's decisions. Situational biases include the performance hall, the order of pianists on the programme, the jury's prior knowledge about performers and the reaction of the audience as perceived by the jury. Musical biases include the juror's preference for, or familiarity with, the music being played. Biases to do with the performer include visible aspects such as gender, skin colour, attractiveness, age, height and weight. Finally, perceptual modality can also be a source of bias – whether the performance is only heard, or is both heard and seen. The latter involves body movement and the shapes and coordination of gestures.

Before turning to specific sources of bias, let us consider for a moment what the jury members are actually doing and experiencing as they attempt to make judgements that are both valid and reliable. One way to look at this process from a psychological viewpoint is to divide it into cognitive (analytical, systematic) and emotional (holistic) aspects. Cognitive processes tend to involve the systematic consideration of different features (technique versus interpretation, and further subcategories), which, as we have seen, makes judgements more reliable but less valid. Cognitive processes have useful teaching applications: teachers who can clearly articulate what is going on in different categories such as technique and interpretation can more effectively communicate useful ideas to their students, which their students can then implement in an analytical fashion. Emotional 'overall gut-reactions' have a different set of advantages and disadvantages. They tend to be more valid,[5] because (one might argue) emotional

4
Paul T. von Hippel, 'Achievement, learning, and seasonal impact as measures of school effectiveness: It's better to be valid than reliable', *School Effectiveness and School Improvement*, 20/2 (2009), 187–213.

5
Janet Mills, 'Assessing musical performance musically', *Educational Studies*, 17/2, 173–181 (1991), cited by Gary E. McPherson and William F. Thompson, 'Assessing music performance: Issues and influences', *Research Studies in Music Education*, 10/1 (1998), 12–24.

reactions are ultimately what music performance is all about. But they are less reliable, because they depend more on the context in which a performance is experienced.

Swanwick implemented a relatively holistic approach to performance evaluation for competition juries. [6] He drew up a list of level descriptors that did not separate technique from interpretation, but instead grounded evaluations in clear statements, in an attempt to simultaneously optimise reliability and validity, and balance cognitive and emotional aspects.

A more cognitive, systematic approach to evaluation was taken by Abeles, who asked experts to rate clarinet performances on a large number of scales, which were reduced to six factors by factor analysis: interpretation, intonation, rhythmic continuity, tempo, articulation and tone (timbre). [7] The use of such a scale makes evaluations reliable in the sense that one evaluator gives similar evaluations on different occasions (intra-judge reliability) and different evaluators give similar evaluations (inter-judge). But one may question the validity of such an approach insofar as it constrains the responses of the evaluators and downplays the importance of their intuitive spontaneous reactions, which are arguably the most important from the point of view of concert audiences.

A study by Thompson et al. showed that it is possible to discover which analytical aspects of evaluation are having the greatest effect on holistic evaluations. Six pianists performed Chopin's Etude, Op. 25 No. 6, and the recorded performances were evaluated by five other musicians.[8] The evaluations included both overall ratings and ratings of specific features. Correlation analysis suggested that right-hand expression and phrasing were the most important features determining the evaluation of these performances. Surprisingly, tempo was not important: overall evaluations of relatively slow performances were no different from those of relatively fast performances. This example demonstrates another disadvantage of analytical procedures: the relative importance of specific features may depend on the piece or style, and the nature of this dependency cannot (or cannot easily) be predicted in advance.

Gesture and movement

Should competition jurors be influenced by gesture and movement, and if so how much? Or should they judge the performance purely on the basis of the sound? If a juror is distracted by the flamboyant gestures of a performer, should she try to ignore them and focus on the sound? Is that possible?

Wapnick et al. obtained video recordings of performances at international piano competitions.[9] Observers rated the recordings in two different versions: audiovisual and audio only. On average,

6
Keith Swanwick, 'Music education liberated from new praxis', *International Journal of Music Education*, 28 (1996), 16–24, cited by McPherson & Thompson, 1998.

7
Harold F. Abeles, 'Development and validation of a clarinet performance adjudication scale', *Journal of Research in Music Education*, 21/3 (1973), 246–255, cited by McPherson & Thompson, 1998.

8
William F. Thompson, C. T. Patrick Diamond and Laura-Lee Balkwill, 'The adjudication of six performances of a Chopin Etude: A study of expert knowledge', *Psychology of Music*, 26/2 (1998), 154–174.

9
Joel Wapnick, Charlene Ryan, Nathalie Lacaille and Alice-Ann Darrow, 'Effects of selected variables on musicians' ratings of high-level piano performances', *International Journal of Music Education*, 22/1 (2004), 7–20.

audiovisual presentations were given higher ratings than audio, suggesting that the visual aspect makes a positive contribution to the perceived quality of a performance.

In a study of the effect of gesture and movement on music audiences, Dahl and Friberg asked musicians (marimba, bassoon, saxophone) to perform the same piece of music with different basic emotions: happiness, sadness, anger and fear.[10] Observers watched videos in three versions: audiovisual, audio only and visual only. In the visual only conditions, three of the four emotions – happiness, sadness and anger (not fear) – were well communicated by gesture alone. The result confirmed that the experience and evaluation of music performance depends strongly on vision.

Juchniewicz instructed pianists to play the same pieces in three different conditions: no movement (trying not to move their body while performing), head and facial movement, and full body movement.[11] Performances were recorded on video and presented to listeners. The main result was not surprising: performance ratings improved with more movement. It is unclear whether the improvement in ratings was due to the sound alone or to the movements as perceived by the audience; probably both contributed positively. In any case, body movements may help a performer to express sound structures, and from the performer's viewpoint, restricting movement obviously inhibits the ability to express. 'Full body movement' is the normal way of playing and that with which pianists are most familiar; so of course they can express themselves better in those conditions. Observers may also generally prefer performers who can be seen to move a lot while they are performing, regardless of the sound, perhaps because they have learned in the past to associate visible body movements with good performances.

Serial order effects

An important bias that can affect the results of a music competition is the serial order effect. When a human observer is asked to rate a series of stimuli on a subjective rating scale, the rating for a given stimulus generally depends on what happened previously. It depends most strongly on the immediately preceding stimulus, because observers tend to compare stimuli with each other. If for example a wine taster is evaluating a series of wines and a below-average wine is followed by an average wine, the average wine may be rated above average because it is better than the previous wine. Even the most experienced and reputable wine tasters exhibit serial order effects such as a preference for the first (primacy) or last (recency) wine in a sequence.[12]

Flôres and Ginsburgh investigated whether the results of the Queen Elisabeth Competition in Belgium depended on the order in which musicians performed.[13] They considered ten violin

10
Sofia Dahl and Anders Friberg, 'Visual perception of expressiveness in musicians' body movements', *Music Perception*, 24/5 (2007), 433–454.

11
Jay Juchniewicz, 'The influence of physical movement on the perception of musical performance', *Psychology of Music*, 36/4 (2008), 417–427.

12
Antonia Mantonakis, Pauline Rodero, Isabelle Lesschaeve and Reid Hastie, 'Order in choice: Effects of serial position on preferences', *Psychological Science*, 20/11 (2009), 1309–1312.

13
Renato G. Flôres Jr and Victor A. Ginsburgh, 'The Queen Elisabeth musical competition: How fair is the final ranking?', *Journal of the Royal Statistical Society*, 45 (1996), 97–104.

competitions (120 performers) and eleven piano competitions (132 performers). The evaluation indeed depended on position on the programme: musicians who performed near the start of a competition (e.g. first day) were less likely to win, while those performing near the end (final day) had a better chance. Possible psychological explanations include familiarity and liking (the jurors liked the pieces more at the end of the competition than at the start, having heard them several times, or they enjoyed the rating task more for a similar reason); recency (more recent performances are more vivid in the memory or tend to overwrite the experience of previous performances); and changes in evaluation strategy (jurors may start the competition with high expectations and gradually become more realistic as the competition proceeds). Effects of this kind can be magnified if they are combined, as when a competitor plays a relatively unfamiliar piece on the first day of a competition.

The effect of prior knowledge about the performers

Chmurzyńska asked whether prior knowledge about performers affects assessments of their performances – even when juries are instructed to consider only the current performance.[14] In her study, music school students rated different performances of a Chopin waltz. In one situation, they received prior knowledge about the pianists; in the other, no prior knowledge. The results were consistent with the intuition that juries either deliberately allow prior knowledge to affect their judgements (even if they agree not to do so) or are incapable of eliminating the influence of prior knowledge on their judgements. The problem could be solved by presenting jurors with sound recordings only (blind ratings), on the assumption that this prevents them from identifying performers. But sound is only one part of musical performance, and many would argue that the jurors should in fact be influenced by other or all aspects of the performance situation, including prior knowledge about the performers.

Effects of appearance: engagement, gender and 'race'

Behne and Wöllner enquired how jurors' ratings of a piano performance depend on the appearance of the pianist.[15] Videos were manipulated such that the same sound track was presented together with different images – as if two different pianists, with different appearance and movements, had produced exactly the same sounding performance. The jurors were unaware of this manipulation and rated the performances as a whole. The result was that ratings were strongly affected by appearance. There was also an

14
Małgorzata Chmurzyńska, 'Influence of a priori information on music performance assessment', paper presented at the triennial conference of the European Society for the Cognitive Sciences of Music, Manchester, UK, 2015.

15
Klaus Ernst Behne and Clemens Wöllner, 'Seeing or hearing the pianists? A synopsis of an early audiovisual perception experiment and a replication', *Musicae Scientiae*, 15/3 (2011), 324–342.

interesting effect of gender: male interpreters were considered more 'precise', whereas female interpreters were more 'dramatic'.

Ryan and Wapnick studied the effect of performers' attractiveness on performance evaluations, using recordings from the Van Cliburn International Piano Competition as experimental stimuli.[16] Raters had different levels of expertise, and stimulus conditions were audio, visual or audiovisual. The result was comforting for competition jurors: there was no clear or consistent attractiveness bias for raters with high levels of expertise, although some of the data suggested that attractiveness had a bigger effect for female than male performers. Another study found that attractiveness bias fell with increasing performance duration.[17]

Elliott systematically investigated effects of 'race' and gender.[18] Four flautists and four trumpeters were recorded on video playing the same piece. Half of the performers were white and the other half were black. Half of the black performers were female and the other half were male; the same applied to the white performers. As in the studies by Behne, raters saw manipulated video recordings in which the sound was the same but the image was different. On average, black and female performers received lower ratings than white and male performers, corresponding to racist and sexist stereotypes, although these stereotypes were often weak and absent, depending on the judges. Davidson and Edgar used videos of piano performance for a similar investigation. They then went a step further, having musicians rate either videos or point-light displays in which only the movement of certain points on the body could be seen.[19] The results revealed no significant effect of skin colour. Regarding gender, there was no main effect, but female raters tended to favour female performers – perhaps intuitively counteracting everyday sexism that female musicians have traditionally encountered.

Ideas for future research

Jurors' ratings evidently depend on how they feel emotionally about a performance, and perhaps also how they feel generally – and why not? We are all human, and some would argue that a good performance should touch our soul. That being the case, it would be surprising if this central issue had never been systematically investigated.

The emotional state of jurors could be monitored by videoing their facial expressions and later having independent observers evaluate the emotional content of the photos. The relationship between facial expressions and universal basic emotions is well known and surprisingly robust,[20] so this method may be more valid than one based on physiological measures such as heartbeat, breathing and skin conductivity. Compared with facial expressions, physiological measures tend to give more information about

16
Charlene Ryan, Joel Wapnick, Nathalie Lacaille and Alice-Ann Darrow, 'The effects of various physical characteristics of high-level performers on adjudicators' performance ratings', *Psychology of Music*, 34/4 (2006), 559–572.

17
Wapnick et al., 'Effects of selected variables'.

18
Charles A. Elliott, 'Race and gender as factors in judgments of musical performance', *Bulletin of the Council for Research in Music Education*, 127 (1995), 50–56, cited by McPherson & Thompson, 1998.

19
Jane W. Davidson and Richard Edgar, 'Gender and race bias in the judgement of western art music performance', *Music Education Research*, 5/2 (2003), 169–181.

20
Paul Ekman and Harriet Oster, 'Facial expressions of emotion', *Annual Review of Psychology*, 30/1 (1979), 527–554.

arousal, but less about valence (positive versus negative emotion) and categories of emotion such as happiness, sadness, anger, fear, surprise or disgust.

Another promising idea for future research is to allow performers to improvise, and to include improvisation among the aspects of a performance that are to be holistically evaluated. Improvisation plays an important role in almost all music, at most times and in most places, up to and including the nineteenth century in the European classical tradition.[21] In the twentieth century, the art of improvisation had shifted towards jazz. The positivistic focus on the Urtext within the classical tradition and the subsequent suppression of improvisation in score- and interpretation-based performance practice is regrettable – but not inevitable.

If performers in the Chopin Competition in Warsaw, for example, felt able to include improvisations in their performances, as Chopin himself would have done, those improvisations could be judged according to style, taste, appropriateness, and so on. The competition would at once become more interesting and more authentic, for both experts and the general public. It would not be necessary for the jurors themselves to be experts in improvisation to evaluate improvisations by competitors in the context of their overall performances. An innovation of this kind could reform the performance tradition of Chopin and classical music generally, bringing it closer to the world in which the composers themselves lived and composed.

Tentative recommendations

Referring to a series of empirical studies, Manturzewska proposed that 'We should be using the new paradigm of psychology of evaluation, and different methods of research, both quantitative and qualitative, psychometric and humanistic, interpretive approaches.'[22] Scientific research results suggest that competition evaluation procedures could be improved in a number of ways. Yet given the absence of an absolute standard by which to measure the best performer or performance – and even if there was such a standard, the limited ability of at least some jurors to evaluate according to that standard – it may be appropriate to run several different kinds of evaluation in parallel, and to combine them to obtain a final result.

Perhaps the most important problem, and the easiest to solve, is the serial order effect. The evaluation of a musical performance usually depends on the quality of the previous performance in a series. If a performance is disappointing and receives a low grade, jurors may overestimate the quality of the following performance – a kind of contrast effect. The effect may ultimately depend on the emotional state of the jurors, who feel relief and optimism when

21
See Jean-Jacques Eigeldinger, *Chopin: Pianist and Teacher as Seen by his Pupils*, tr. Naomi Shohet with Krysia Osostowicz and Roy Howat, ed. Roy Howat (Cambridge: Cambridge University Press, 1986); Martin Gellrich, 'Instrumental practice in the eighteenth and nineteenth centuries', *Bulletin of the Council for Research in Music Education*, 119 (1993), 137–145.

22
Maria Manturzewska, 'The reliability of evaluation [of] musical performance by music experts', *Interdisciplinary Studies in Musicology*, 10 (2011), 97–109.

the current performance is better than the previous one. More generally, happiness is often considered to be relative, depending on the difference between one's current situation and a previous situation, although this idea has also been challenged.[23] I am not saying that jury members should stop allowing their feelings to influence their evaluations. On the contrary, feelings are essential for validity. Instead, performances should be presented in a random order that is different for every jury member – just as trials in psychology experiments are presented in a random order that is different for every experimental participant, and different again when a participant repeats the experiment (to check his or her reliability).

In music competitions, serial order effects could be reduced in two specific ways. First, when remote jurors are judging a series of performances, computer software for psychological experiments could be used to present those performances in different random orders. The members of such a jury, who could work at home in different countries, would agree to isolate themselves from any news about the competition until their task was complete. They would listen to different performances of the same piece in a unique random order that was determined by the software. As a control, they would try to guess the identity of the performers, and their data would be discarded if their guesses were correct more often than by chance. The results of such an evaluation would then be compared with results obtained by more traditional methods. A jury at a higher level ('superjury') would consider these different sources of information and decide on the winner.

A second way to address the order problem is to ensure that the order of the live programme is random. The procedure for deciding the order should be transparent and public – similar to procedures for deciding who wins a lottery draw. A procedure of this kind can be compared with the selection of experimental and control groups in medical studies, in which completely random allocation to experimental groups is essential if results are to be taken seriously.[24] If each performer competes in several different stages of a music competition, different random orders at each stage may be determined by Latin squares – a statistical technique that can further reduce possible biases for or against given performances.

Another issue is the use of quantitative scales to rate performances. Results of psychological experiments in which participants use rating scales (e.g. from 1 = very poor performance to 7 = exceptional performance) can depend on how the scale is labelled: whether numbers are used, if so which numbers, and which points are verbally labelled.[25] Experience suggests that results of rating studies are most reliable and valid when at least three points on the scale (left, right and middle) are marked with qualitative descriptors, or if the middle point has an obvious meaning. For example, if colours are rated on a scale from yellow

23
Ruut Veenhoven, 'Is happiness relative?', *Social Indicators Research*, 24/1 (1991), 1–34.

24
Ben Goldacre, *Bad Science* (London: Harper Collins, 2008).

25
Bert Weijters, Elke Cabooter and Niels Schillewaert, 'The effect of rating scale format on response styles: The number of response categories and response category labels', *International Journal of Research in Marketing*, 27/3 (2010), 236–247.

to red, one of the intermediate points should be labelled 'orange'. Without an anchor of this kind, points on the scale can lose their meaning. In the case of competitions that start with a large number of competitors and become smaller as the competition proceeds, heading towards a final competition between a small number of semi-final winners, it should be clear to jurors at each stage of the competition whether their grades are high enough for the pianist to continue to the next stage.

On the assumption that (ecological) validity is (even) more important than reliability, psychological research supports the use of single holistic overall ratings rather than combinations of separate ratings for technique (virtuosity), interpretation, authenticity and so on. While there is no harm in allowing different aspects to be separately rated – indeed, this process may help jurors to offer clear justifications for their final decisions – the psychological concept of validity suggests that those final decisions should ultimately be based on overall ratings or gut reactions.

A test is valid only if it is clear what question is being asked. What exactly is being assessed? What is valuable to us, as listeners in a given cultural context, about performances in general, and a given performance in particular? Why do we like it, exactly? What musical values are implicitly being examined in a music competition? What, for example, is meant by 'authenticity', and how exactly can a performer in a competition be 'authentic'? To what extent is 'authenticity' already present in the score, and to what extent can it be added by the performance? If jurors are claiming to recognise the realisation of a composer's intentions, as many do, we may reasonably ask as scientific observers how they can possibly do that. We may similarly ask what they may mean by Polish, Russian, French or Chinese performance styles, or how well jurors can judge such differences independent of the performers' appearance or prior knowledge about them. In any case, performance traditions have changed enormously since the music played at most competitions was composed, and this variation should itself be considered part of the evaluation, because it is an important and valuable aspect of the cultural context.

While there are no clear answers to such questions, jury members nevertheless benefit from articulating and discussing them, trying to see the different sides of each issue, and avoiding extreme positions. From a practical perspective, juries might be invited to discussions of this kind before, during or after the competition. Psychologists are interested in how people interact in a group when given a particular task, and qualitative research can benefit from focus group interviews in which group members bounce ideas off each other.[26] Jurors' evaluations might become more reliable and valid if they have a chance to discuss their general approaches in a group, noting the contrasting opinions of their peers, and if they later discuss specific musicians and performances before coming to final

26
Sharon Vaughn, Jeanne Shay Schumm and Jane M. Sinagub, *Focus Group Interviews in Education and Psychology* (Thousand Oaks, CA: Sage, 1996).

decisions. After such a discussion, final grades should be submitted confidentially, to minimise biases due to power relationships within the group.

Combining methods

It bears repeating that all methods for trying to identify the best performer or to predict the future development of performers are problematic. For that reason, it may be wise to combine different methods, exploiting the contrasting benefits of tradition, modern research and modern technology. Existing methods can be complemented by adding new ones inspired by psychological research, while at the same time maintaining existing methods to get the best of both worlds. This strategy may improve both the reliability and the validity of juries' decisions.

The traditional jury could be replaced by a series of subjuries. Subjury 1 might be the traditional jury, comprising mainly professional musicians and including some previous winners of the same competition. They would attend the competition performances in the usual way. Subjury 2 might include expert audiences, critics and teachers who are sitting at different places in the concert hall; they could use new technologies (e.g. mobile phones) to rate pianists during or soon after performances. Subjury 3 might comprise international experts at home in different countries, who would be asked to participate in a quasi-psychological experiment shortly after each section of the competition, with performances presented in a random order that is different for each evaluator, as described above. They might first rate the sound alone, and later rate audiovisual presentations to give contrasting data.

Finally, one might envisage a kind of superjury that would consider the similarities and differences of the quantitative results and qualitative findings of the subjuries before coming up with a final result. The superjury would mix expertise from different disciplines, including music performance, music analysis, composition, music history, music psychology, music sociology and the music industry. Superjury members might be experienced colleagues who have served repeatedly on different subjuries in the past.

For procedures of this kind to work well, one would also need a strategy to manage conflicts of interest. Procedures might be similar to those of peer-reviewed academic journals. Jurors would be asked to evaluate their personal relationship with each performer and his or her teachers.

To avoid sexism, juries should have a balance between female and male judges. That may be difficult given the continuing male dominance in many areas of music performance; for example, almost all past winners of the Chopin Competition have been male.

However, I know of no evidence that men are fundamentally better at performing music of any kind. In the absence of such evidence, and given that sexism is often unconscious and unintended,[27] we should apply 'reverse discrimination' until it is clear that young female musicians enjoy the same level of opportunity as young male musicians, as measured by the relative numbers of women and men winning competitions. Sexism can take many forms; it includes sexist ideas in the families in which girls grow up, and the apparently innocent idea that performing music at a high level is okay for women as well as for men, but at some point women are expected to give up their career and have a family instead. In fact, both sexes can give up, interrupt or deprioritise their career to make time for children. And if even leading feminist and anti-sexist activists can be shown in a psychological test to have sexist biases, it is clear that long-term affirmative action is necessary.

27
Sam Cameron, 'The political economy of gender disparity in musical markets', *Cambridge Journal of Economics*, 27/6 (2003), 905–917.

ABSTRACT
Procedures to evaluate the quality of musical performance, like psychological tests generally, vary in reliability and validity. How can both be optimised? The subjective world of jurors comprises input (sensations) and output (thoughts and emotions). Cognitive approaches to performance evaluation, in which different aspects are analytically considered and the results combined, are more reliable than emotional approaches, which are ultimately based on holistic 'gut reactions'. However, emotional approaches may be more valid. Both depend on the serial order in which performances are presented to evaluators, suggesting a need for independent, computer-controlled procedures in which jurors evaluate performances in different random orders. Jurors can be influenced by performers' appearance and movements, as well as knowledge about past performances, whether or not they believe they should be; additional blind evaluations (sound only) could help. Evaluations may depend primarily on specific features such as right-hand melodic phrasing in romantic piano music. It would be interesting to systematically track the emotional state of jurors during a competition in order to better understand the interaction between their thoughts and emotions. In general, traditional approaches to performance evaluation might be supplemented (not replaced) by psychologically inspired, computer-based procedures. A 'superjury' could compare evaluations from different methods or 'subjuries'.

KEYWORDS
music performance, evaluation, psychological test, bias, reliability, validity, serial order

RICHARD PARNCUTT is a music psychologist and Professor of Systematic Musicology at the University of Graz, Austria, a position he has held since 1998. He has been director of the University's Centre for Systematic Musicology since 2009 and president of the European Society for the Cognitive Sciences of Music (ESCOM) since 2015. He is founding academic editor of the *Journal of Interdisciplinary Music Studies* (JIMS), and co-founder of the series Conference on Interdisciplinary Musicology (CIM) and International Conference of Students of Systematic Musicology (SysMus). His publications address musical structure (pitch, consonance, harmony, tonality, tension, rhythm, metre, accent), music performance (psychology, piano, applications), the origins of tonality and of music, and musicological interdisciplinarity.