# Measuring quality in the cultural sector

## The Manchester Metrics pilot: findings and lessons learned

**Catherine Bunting and John Knell**

**May 2014**


# Contents

## Foreword

We are delighted to present this report on the work of the Manchester Metrics group, with John Knell and Catherine Bunting, on the piloting of a potential metrics framework for assessing the quality of arts and cultural productions.

We arrived at this project from a sense that the arts and cultural sector struggles to provide arguments about the overall quality of its work in a way that both has credibility with funders and other stakeholders, and has the support of the arts sector. This is because relying purely on occasional peer review runs the risk of seeming like a conspiracy of self-interest, with arts people being reviewed by other arts people, and other measures (audience numbers for example) are not accurate measures of artistic quality.

We were interested to hear about the work in Western Australia, where the arts sector itself developed a metrics framework involving the triangulation of assessments of self, peer and public, and have trialled a similar approach in Manchester.

As the report sets out, we have found that it is possible to gain consensus on a core group of quality dimensions, and the pilot has shown that the idea of undertaking self, peer and public assessment can work in practice.

The pilot has also raised a large number of issues about refining the questions, expanding the range of artforms, dealing with large scale data, and the practicalities of implementation. With that in mind, we are delighted that we have been awarded a grant from the Digital R&D Fund to develop the metrics framework further over a 12 month period. This will allow it to be tested with a much wider range of artforms, scales of organisation and geographic locations, as well as providing time-series data and assessing the feasibility of this system at a large scale.

We are very grateful to Arts Council England and the Audience Agency for supporting this pilot phase of the project.

The Manchester Metrics Group
May 2014

# 1. Introduction

The Manchester Metrics pilot takes its inspiration from a project initiated in 2010 by the Department of Culture and the Arts (DCA) in Western Australia, who commissioned Michael Chappell of Pracsys and John Knell of Intelligence Agency to help them understand and measure the public value they create through their investments in arts and culture.

The key aims of that work were:

- to engage with the arts and cultural sector to try and produce a measurement framework that sensitively reflects their understanding of how best to foster and measure artistic quality, engagement and innovation
- to work with the arts and cultural sector to create a standardised and aggregatable metric system that measures what the cultural sector believes are the key dimensions of quality

As the metrics system developed in Western Australia it became clear to both the funder (DCA), and the funded arts organisations and artists taking part, that the arts and cultural sector had reached something of an impasse with regard to the measurement of the value they create, and that there was strong support for a more sophisticated approach to the measurement of cultural value.

The funded arts organisations and artists involved in the Western Australia project welcomed the opportunity to actively shape the measurement framework. This emphasis on co-producing the metrics with the arts and cultural sector was one of a number of key innovations trialled in Western Australia. The project quickly confirmed that the active involvement of the arts and cultural sector is fundamental to the creation of a credible and robust metric framework. The arts and cultural sector were being offered the chance to:

- shape the measurement of what they do in a way that reflects their artistic ambitions and intentions
- create better alignment between the data they need to inform their own creative programming and audience development activity and the data they need for public accountability purposes
- create a measurement and evaluation system that could diminish the reporting demands on them as funded clients, whilst increasing the quality and consistency of the evidence base
- build a stronger shared understanding of the sector's collective aims and ambitions

The resulting metric framework is now in its final implementation phase with the DCA planning to roll out the metrics, and supporting app and web based platforms for data collection, over the course of 2014.

The work in Western Australia has proved of interest to other cultural organisations and public funders. The potential of this work led a group of 13 Manchester cultural organisations (listed in Appendix A) to come together in spring 2013, with Arts Council England support, to work with John Knell to undertake the same process exploring whether, in principle, it would be possible to develop a commonly-agreed set of quality indicators.

The cultural organisations were excited by the possibility of working together to develop a new and robust approach to assessing the quality of the work that could be applied consistently across a number of different events and settings. They were also motivated by a belief that the outcomes and metrics that funders currently ask them to report against are not adequately capturing the quality and reach of the work they undertake.

Arts Council England were interested in supporting this work as they have been undertaking a review of their outcome and success measures for the sector, and were keen to support sector led initiatives which could contribute to their thinking on how best to measure the quality and reach of cultural activity.

The pilot began in April 2013, and has had two distinct phases:

- stage one (the proof of concept and metric identification phase)
- stage two (the testing phase) – on which this report focuses

***Stage one – the proof of concept and metric identification phase***

In generating quality metrics it was agreed that the group should start with a blank sheet of paper. In other words, the group would not review either existing metrics used by UK funders, or the emerging metric set produced by the DCA work. Rather the Manchester Metrics group set themselves the challenge of working from first principles, addressing the following questions:

- What do we mean by quality?
- What do we mean by reach?
- What outcome areas should we be measuring to capture quality and reach?
- What are the metric statements that best capture the essence of those outcome areas?
- Is it possible to develop a dashboard of measures capturing quality and reach that we would be happy to collectively endorse and use?

The aim was to develop a clear ‘outcome’ set for the key dimensions of quality and reach, and then to begin to identify metric statements to capture the essence of these outcomes.

Importantly the group had representatives from different artforms, and from the museum sector, and therefore were curious as to whether they would be able to agree a core set of metrics that could work well for the diversity of cultural forms and settings represented in the group.

The outcome set was developed between April and July 2013, with the Manchester Metrics group proving able to forge a strong degree of consensus about the outcomes they felt were most important to evaluate. They grouped these outcomes under three categories:

- quality – including: quality of product; quality of experience and engagement for audience members; and quality of creative collaboration with artists, curators, producers and so on
- reach – including an outcome for value adding partnerships
- organisational health – including outcomes for the quality of cultural leadership

These outcomes and findings were then shared with the Arts Council to determine whether further development of the outcomes and tentative metric statements was viewed as a worthwhile progression of stage one of the pilot.

***Stage two – the testing phase***

The Manchester Metrics group and the Arts Council decided that sufficient progress had been made in stage one, in terms of the coherence and focus of the 'outcome' set and metrics, to fund a stage two testing phase, on which this report focuses. It was decided that the testing phase would focus on quality rather than organisational health or reach, and in particular would seek to develop and test metrics to capture the quality of work and quality of experience for audience members and visitors. This testing phase has involved the following key elements:

- the Manchester Metrics Group developing standardised metric statements for some of the key quality outcome areas identified in stage one
- the use of a system called Culture Counts (first developed as part of the Western Australia project) to test self, peer and public responses to those metric statements at eight cultural events that took place between November 2013 and January 2014
- the analysis of the resulting data, and a reflection process on the merits of the approach and key implications for the future development and use of the metrics framework and Culture Counts system

This report provides a detailed account of the testing phase. Chapter two details the essential features of the Culture Counts system, the metrics used during this testing phase, and the event selection and data collection processes (including the identification of peers).

Chapter three analyses the data, presenting the scores awarded by self, peer and public assessors for all the different quality dimensions for each of the eight events in the Manchester pilot. We combine this assessment with corresponding artistic intention statements from each of the eight test organisations, in which they explain what they were trying to achieve with the work in a creative sense and their expectations of where it would score well on the quality dimensions.

Chapter four explores the individual quality metrics in more detail, comparing the scores received across all eight events for each metric in turn. We are primarily concerned here with the meaning and usefulness of the questions being asked, and with identifying which dimensions seem to be generating the most insightful data on the quality of the cultural events.

Chapter five offers some reflections on the process and findings, and we make some recommendations on the further development of the metrics and the Culture Counts system. We also sketch the wider implications of the pilot for other researchers involved in collecting data on experiences and perceptions of arts and cultural events.

Chapter six suggests some overarching conclusions and reflections on the current and potential value of this sort of approach to quality assessment and the data it generates – for cultural organisations, for audiences, for funders and for researchers more generally.

# 2. Methodology

## 2.1 Overview of the Culture Counts system

Culture Counts is a tool that captures that captures artist, peer and public feedback on the quality of arts and cultural events. It was developed by consulting firm Pracsys and John Knell as part of DCA's project in Western Australia to develop an overall framework for measuring the public value of its investment in culture. Culture Counts aims to provide value to: cultural practitioners and organisations by connecting them to peer and audience feedback; funders by measuring the quality of work supported by their funding; and the public by providing a structured forum for sharing views and opinions on arts and cultural experiences.

Culture Counts captures feedback on the quality of a work or event from three different groups:

- the artists, curators and/or cultural organisation that created the work or produced the event (self-assessment)
- expert peers such as other artists, people working in cultural organisations in the same field and academics; if appropriate peers can also include funders and representatives of business and political communities (peer assessment)
- audience members and visitors (public assessment)

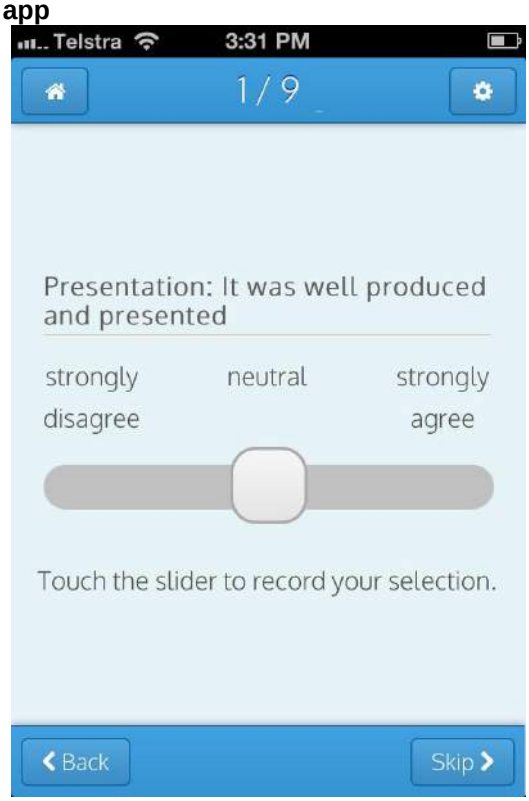
Quality is assessed by asking respondents to rate the work or event against a number of quality dimensions. Respondents complete a short survey in which each quality dimension is presented as a statement or 'metric' and respondents record the extent to which they agree or disagree with the metric using a sliding scale. Respondents indicate agreement by moving the slider to the right, disagreement by moving the slider to the left and a neutral response by clicking on the slider once to leave it at the mid-point of the scale. Respondents record a 'don't know' response by not moving the slider at all. As well as rating the event against the quality dimensions, respondents are asked to provide their gender, age and postcode.

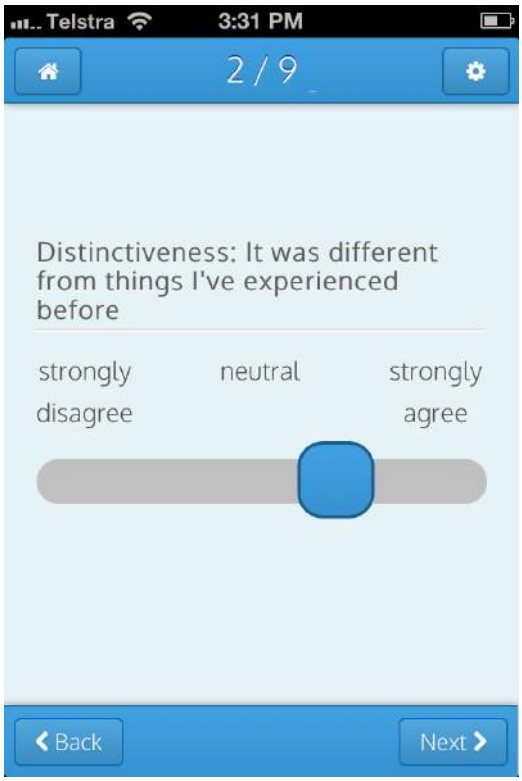
Self and peer assessment is carried out both before and after an event to explore how perceptions shift and the extent to which the event matches up to expectations. Self and peer assessment takes place via an online portal, with each assessor given unique login details and emailed instructions on how to complete both 'before' and 'after' surveys.

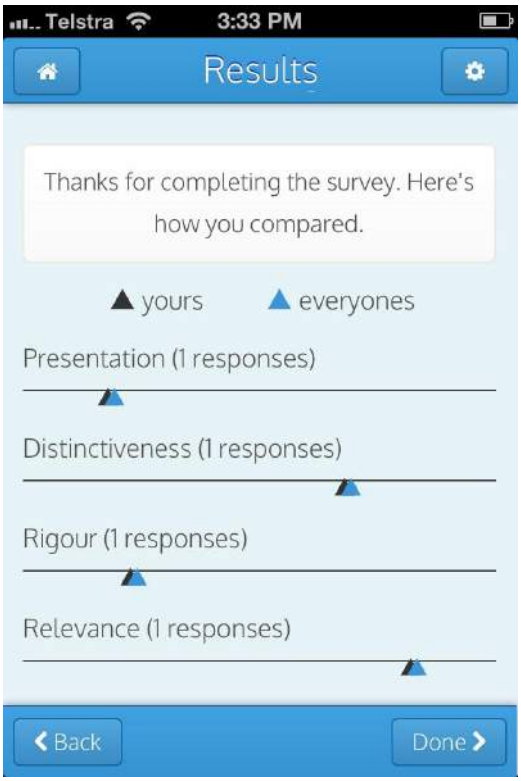
Public assessment takes place during or just after the event itself and captures 'real-time' feedback on how the audience is responding to the work. Audience members record their ratings using an app downloaded to a smartphone or tablet computer. A selection of screenshots from the app is included in Figure 1 to show how the survey appeared to members of the public who took part in the Manchester pilot.

Data from all respondents for every event is stored in a single database and exported to Excel files for analysis.

**Figure 1: Screenshots of the Culture Counts app**

## 2.2 Metric development

As explained in chapter one, at the start of the pilot the Manchester Metrics group had identified a number of dimensions of quality of cultural experience that they felt captured what they were trying to achieve with their work in terms of public and peer response.

The challenge for the group now was to translate these quality dimensions into a set of tightly defined metrics for inclusion in the Culture Counts system, with each metric composed of an essential term (eg 'distinctiveness') with an accompanying definition statement ('It was different from things I've experienced before'). The metric set had to meet some stringent criteria:

- the metric set could not be too large, as we wanted the survey process to be quick and straightforward for respondents, particularly members of the public. We set a limit of 10 metrics for inclusion in the public survey and 15 for the survey to be completed by self and peer assessors

- the metrics for public assessment had to be defined in a way that a diversity of audience members and visitors could understand and respond to

- the metrics had to be applicable to a number of different types of cultural experience, from a museum exhibition to a pantomime

As the project researchers, we took the outputs from stage 1 of the project (defined outcome areas with rough cut metric statements for the different dimensions of quality) and presented these back to the group with some suggested first stage refinements. Through a series of workshops the group debated, rejected, amended and added elements until they were in collective agreement on the metric statements. At the end of this process the group had produced a draft metric set which we took away and sense checked with colleagues in Australia. A final workshop was then held for the group to work on the revised metric set and further refinements were made and final agreement reached via email.

The final metric set comprised nine core metrics to be rated by self, peer and public assessors:

**Presentation:** it was well produced and presented
**Distinctiveness:** it was different from things I've experienced before
**Rigour:** it was well thought through and put together
**Relevance:** it had something to say about the world in which we live
**Challenge:** it was thought-provoking
**Captivation:** it was absorbing and held my attention
**Meaning:** it meant something to me personally
**Enthusiasm:** I would come to something like this again
**Local impact:** it is important that it's happening here

A further five metrics were included for self and peer assessment only:

**Concept:** it was an interesting idea/programme
**Risk:** the artists/curators really challenged themselves with this work
**Originality:** it was ground-breaking
**Excellence (national):** it is amongst the best of its type in the UK

**Excellence (global):** it is amongst the best of its type in the world

## 2.3 Event selection

At the same time as finalising the metric set, the Manchester Metrics group identified the arts and cultural events to be evaluated in the pilot. The selection was informed partly by a desire to include a number of different types of cultural experience in different settings in and around Manchester, but also by the practicalities of programming and what was happening at different venues during the agreed fieldwork window from November 2013 to January 2014. The following eight events were selected:

**Verdi bicentenary concert – Hallé Orchestra**: an event in Bridgewater Hall to celebrate the bicentenary year of Verdi's birth, this concert focused on the relationship between the composer and Arrigo Boito and included performances of the operas *Simon Boccanegra*, *Falstaff* and *Otello*.

***That Day We Sang* – Royal Exchange:** Victoria Wood's play about Manchester and the enduring power of music, first seen as part of the Manchester International Festival in 2011 and reinvented and re-orchestrated for the Royal Exchange.

***Jack and the Beanstalk* – Oldham Coliseum:** a Christmas family show in the traditional northern England 'panto' style.

***Robin Hood* – Octagon Bolton:** a new play based on the Robin Hood legend that combined serious drama with humour and music and included a cast of local children.

***Contact Young Actors Company (CYAC): Advent Avenue* – Contact:** a Christmas play aimed at teenage and up audiences that was devised, created and performed by a group of non-professional local actors between the ages of 16 and 24.

***Jeremy Deller: All That Is Solid Melts Into Air* – Manchester Art Gallery**: artist Jeremy Deller's personal look at the impact of the Industrial Revolution on British popular culture, originally curated at the Southbank Centre in London and brought to Manchester through a collaboration between Manchester Art Gallery and Hayward Touring.

**Vivarium – Manchester Museum:** the recently reopened 'Live Animals' gallery which enables visitors to experience a first-hand encounter with frogs, reptiles and lizards and explores the museum's role in helping to protect endangered species.

***The Radev Collection: Bloomsbury and Beyond* – Abbot Hall Art Gallery:** exhibition of works by British and international artists collected by Mattei Radev shown at Abbot Hall, a venue in Kendal, Cumbria run by Lakeland Arts Trust.

## 2.4 Self-assessors

Each organisation that participated in the pilot was asked to nominate a number of people to provide a self-assessment for the event being evaluated. There was no limit to the

number of self-assessors that could be nominated but we recommended that each organisation invite up to five people to participate. In total 32 people completed a ‘before’ self-assessment survey, an ‘after’ survey or both and this group included curators, artistic directors, staff working in marketing, education and administration roles and freelance associates, as well as members of the Manchester Metrics group.

## 2.5 Identifying peers

A key question for the group at the start of the process was who should be invited to participate as peer assessors. One option was for cultural organisations to nominate their own peers, which would enable organisations to select people with specialist expertise in the relevant artistic or cultural form and knowledge of the organisation and its previous work, and whose views they felt would be particularly meaningful and useful. However the group was concerned that if organisations nominated their own peers this might introduce some bias to the process and debated whether a more ‘objective’ assessment would be achieved if peers were allocated to events from a centrally appointed panel.

In the end the group decided to experiment with both approaches. First, the participating organisations were asked to recruit up to five peers from within their own networks and each organisation took a slightly different approach to this task. Most organisations nominated both Manchester-based peers and people working in similar organisations in other parts of the country. The Hallé Orchestra wanted to achieve a response that was as informed and unbiased as possible and decided to approach a number of classical music reviewers from the national press. Contact felt that it was particularly important to include artist representation while Manchester Art Gallery included a gallery chief executive as well as visual arts experts to provide a view on the exhibition as a visitor encounter as well as an artistic work.

In addition to organisations nominating their own peers, the group asked the Arts Council to invite a number of its regular artistic assessors to take part in the pilot. The Arts Council identified a group of assessors who were reasonably local to Manchester and had expertise in the types of work being evaluated. We matched these people to events based on their availability and area of specialist interest and were able to allocate Arts Council assessors to every event except the Verdi bicentenary concert and the Vivarium exhibition. The Arts Council assessors may or may not have been known to the organisation whose event they were evaluating.

In total 29 peer assessors took part in the pilot and completed a ‘before’ survey, an ‘after’ survey or both.

## 2.6 Public assessment

For most events, surveys with audience members and visitors were carried out by interviewers recruited from pools of market researchers who work for either the Manchester Museum or The Audience Agency in Manchester on a casual basis. There were two exceptions: interviewing for the Radev Collection exhibition at Abbot Hall Art Gallery was carried out by gallery staff and volunteers, and the professional interviewer team at the Verdi bicentenary concert was supplemented by staff and volunteers of the Hallé Orchestra. All those involved in interviewing attended an initial two-hour training session.

Interviewers were instructed to approach audience members and visitors at each event and ask them to give their feedback on the quality of the event by completing a short survey downloaded to a tablet computer. Interviewers were not given demographic quotas to meet but tried to ensure that their samples of respondents were reasonably representative of the audience as a whole in terms of age and gender. We aimed to achieve at least 50 public responses per event.

In the end a total of 637 surveys were completed by members of the public across the eight events. The total number of self, peer and public surveys completed for each event is shown in Table 1.

**Table 1: Number of self, peer and public survey responses received for eight events in the Manchester pilot**

| Event | Self | | Peer | | Public | Total |
|---|---|---|---|---|---|---|
| | Before | After | Before | After | | |
| **Verdi bicentenary concert** | 4 | 3 | 3 | 2 | 48 | 60 |
| ***That Day We Sang*** | 4 | 3 | 4 | 3 | 77 | 91 |
| ***Jack and the Beanstalk*** | 7 | 7 | 4 | 4 | 52 | 74 |
| ***Robin Hood*** | 4 | 3 | 3 | 3 | 57 | 70 |
| ***CYAC: Advent Avenue*** | 4 | 3 | 4 | 4 | 90 | 105 |
| ***Jeremy Deller: All That Is Solid Melts Into Air*** | NA | 5 | 3 | 3 | 133 | 144 |
| **Vivarium** | NA | 3 | 5 | 5 | 66 | 79 |
| ***The Radev Collection: Bloomsbury and Beyond*** | 2 | 2 | 4 | 4 | 114 | 126 |
| **All** | **54** | | **58** | | **637** | **749** |

## 2.7 Capturing feedback

This project was designed as a pilot and we were keen to ensure that as well as collecting data on the quality of different cultural events we learned as much as possible about the process of defining and measuring quality and the practical application of the Culture Counts system. We therefore included a number of mechanisms to capture formal and informal feedback from people involved in delivering the project.

All interviewers were provided with feedback sheets to complete at the end of every shift to tell us about their experiences of using the tablet computers, how people responded to the survey questions and the overall interview process. In addition, we hired two of Contact's Creative Experts (a pool of young and emerging arts consultants) to carry out an evaluation session with interviewers to explore their opinions about the Culture Counts app and the survey questions in more detail. We refer to feedback from interviewers throughout this report and the Creative Experts wrote up a separate paper on their evaluation session which is attached as Appendix B.

A number of people who participated as self and peer assessors gave informal feedback via email, and a draft of this report was sent to all self and peer assessors for comment. Finally, the Manchester Metrics group met at the end of the fieldwork phase to reflect on the overall process, and provided helpful feedback on earlier versions of this report.

# 3. Headline findings by event

## 3.1 Introducing the findings

This chapter explores the quality ratings awarded by audience members and visitors, peers and the cultural organisations themselves for each of the eight events in the Manchester pilot. For each event we start by presenting the average scores awarded by members of the public, with a note on the composition of the sample in terms of age and gender. We then explore peer ratings, identifying any interesting differences between the expectations of peers in advance of the event and their actual experience of the work. The final chart presented for each event compares self, peer and public ratings. Self-assessment scores are examined in more detail where there were notable differences in the ratings awarded before and after the event.

The charts in this chapter show mean scores awarded by particular groups (public, peers and self) for each quality dimension included in the survey on a scale of 0 to 1, where a score of 1 represents the strongest agreement, a score of 0 the strongest disagreement and a score of 0.5 is neutral. The public survey contained questions on nine dimensions and the survey for self and peer assessors included an additional five dimensions giving a total of 14. Comparisons across groups are only made for the nine dimensions common to all reviewers.

Where ‘average score’ is shown for a particular group, this is calculated by taking the mean score awarded to the event across all dimensions by each respondent in the group, then calculating the mean of these mean scores. In some cases sample standard deviations are reported to give an indication of the level of variation in the views of audience members and visitors at particular events. The sample standard deviation reported for a given event is the standard deviation of the individual ratings across all dimensions awarded by members of the public surveyed at that event.

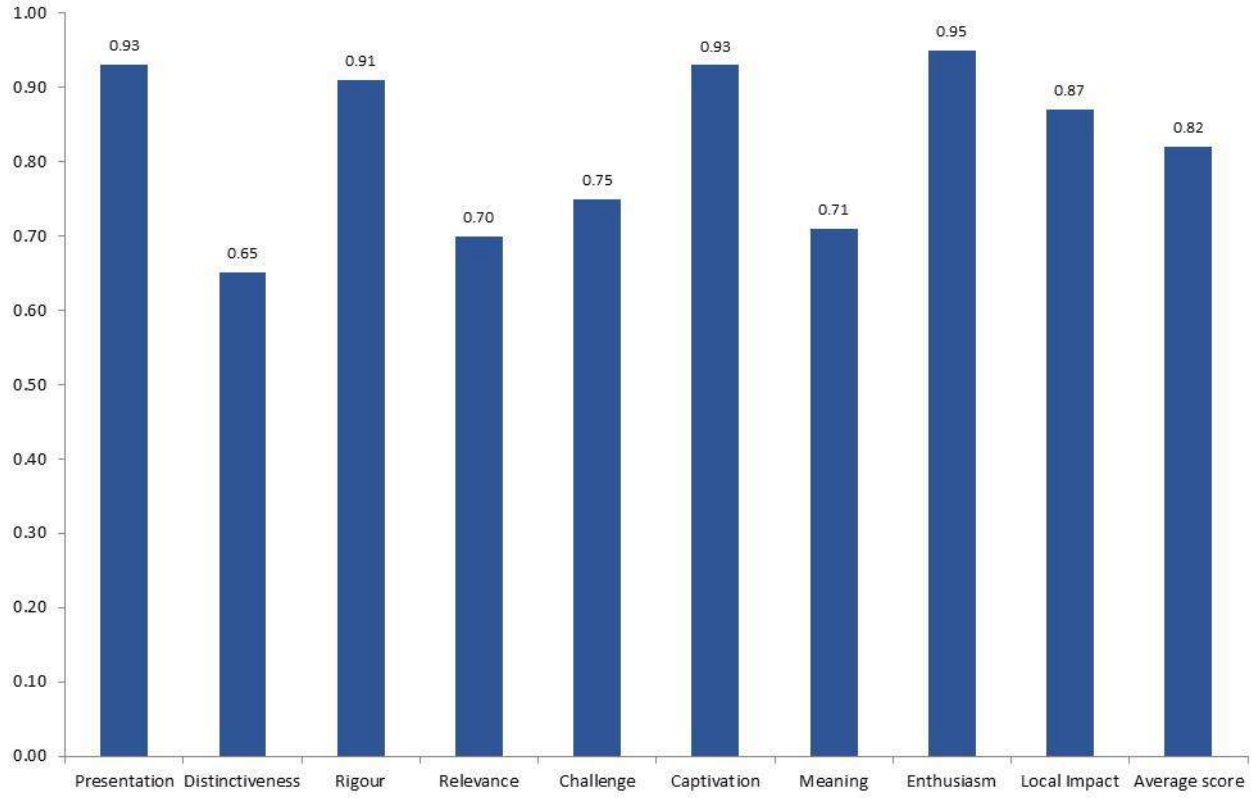
In addition to presenting the results for each event, we also provide some interpretive context in which to understand the data in the form of a creative intention statement from the cultural organisation that produced the work. Each statement explains: what the organisation was trying to achieve; which quality dimensions were most relevant and important to the work and where the organisation hoped to achieve its highest scores; and how the organisation understood its results and the aspects of the analysis that were most useful and insightful.

## 3.2 Verdi bicentenary concert – Hallé Orchestra

The Verdi bicentenary concert was performed by the Hallé Orchestra at Bridgewater Hall on 24 November 2013. A total of 48 audience members completed the Culture Counts survey and among the 44 who recorded their gender there were slightly more men (24 or 55 per cent) than women (20 or 45 per cent). The median sample age was 64, suggesting a relatively old audience that is fairly typical of classical music concerts.

Figure 2 shows the average scores awarded by audience members for the Verdi event. Ratings were particularly high for the quality dimensions ‘enthusiasm’ (average score 0.95), ‘captivation’ (0.93), ‘presentation’ (0.93) and ‘rigour’ (0.91), suggesting that audience members greatly enjoyed their experience and appreciated both the way in which the concert was planned and put together and the quality of the performance on the night. However the audience did not feel that the concert was particularly different from things they’d experienced before, awarding an average score of 0.65 for ‘distinctiveness’ which was relatively low compared to the ratings received for this dimension by other events in the pilot.

**Figure 2: Average public scores for Verdi bicentenary concert**

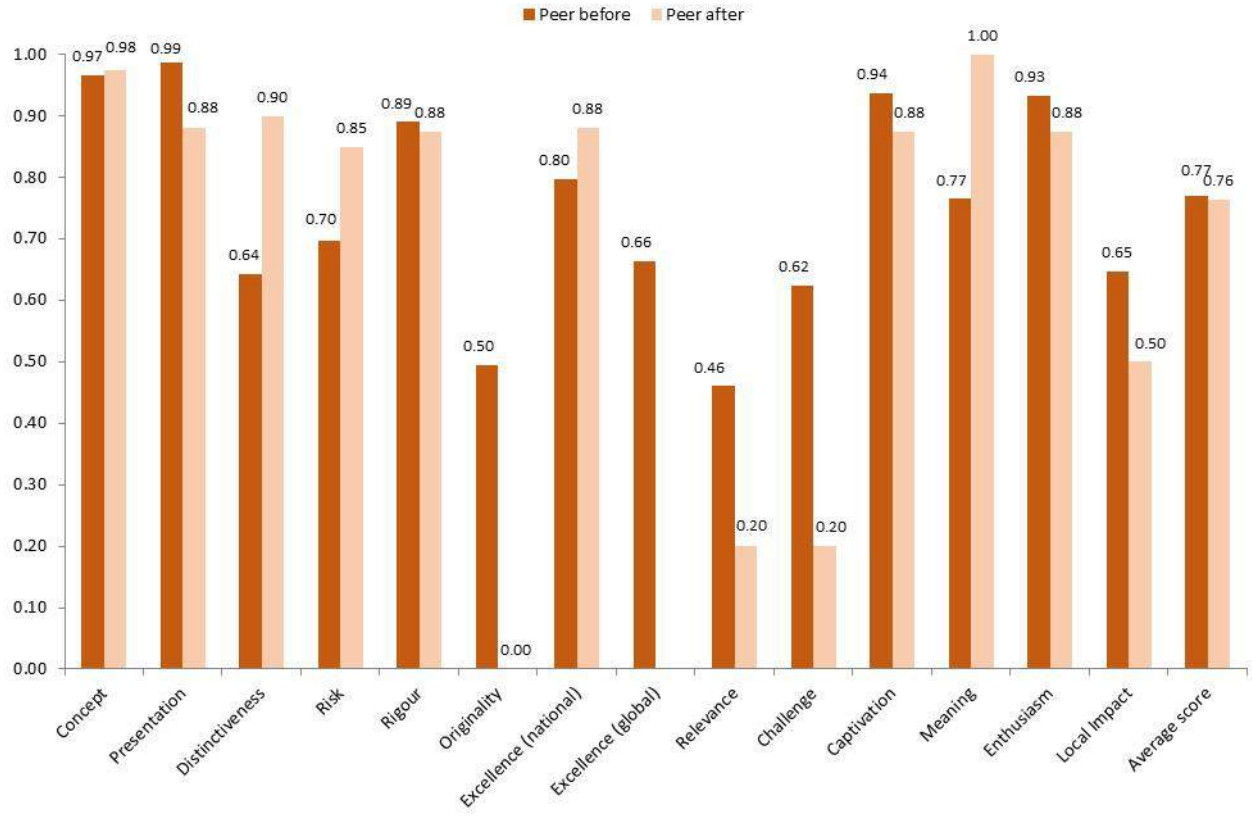



*n=48*

The peer group for the Verdi concert was slightly different to those assessing other events in the pilot in that the Hallé asked a number of critics from the national press to complete the ‘before’ and ‘after’ surveys as well as writing their usual published reviews. In addition, none of the peers nominated by the Arts Council were available to review this event. The Verdi concert received the highest average post-event score from peers (0.76) of all the events in the pilot. Figure 3 shows that peers had high expectations of the concert and were particularly impressed in relation to some of the more technical quality dimensions such as ‘concept’ (giving an average post-event rating of 0.98), ‘presentation’ (0.88), ‘rigour’ (0.88) and ‘risk’ (0.85). Peers felt that the concert was among the best of its type in the UK, awarding a notably high average score of 0.88 for ‘excellence (national)’, and clearly felt a strong personal connection to the work with high post-event scores for ‘captivation’, ‘meaning’ and ‘enthusiasm’. However peers did not feel that the concert was all that relevant to today’s world or thought-provoking, awarding an average post-event score of 0.2 for both ‘relevance’ and ‘challenge’, which was somewhat lower than their pre-event expectations.

Of the two peers who completed the post-event survey, only one provided ratings for the dimensions ‘distinctiveness’ and ‘originality’. Interestingly, that peer felt that the concert was different from things that she had personally experienced before, giving a score of 0.90 for ‘distinctiveness’, but didn’t break new artistic ground, giving a zero rating for ‘originality’. Neither peer was able to provide a post-event score for ‘excellence (global)’, with one commenting that it was ‘impossible to compare without travelling the world!’

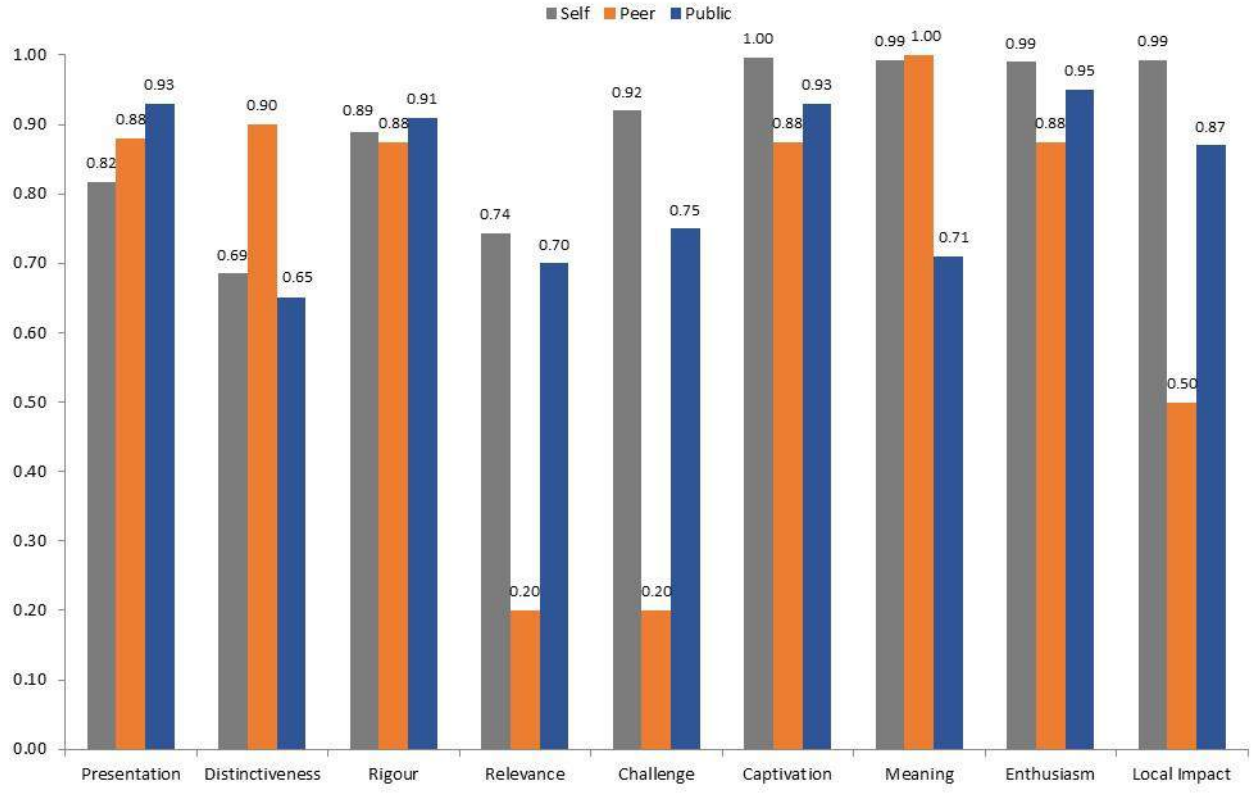
**Figure 3: Average ‘before’ and ‘after’ peer scores for Verdi bicentenary concert**




*Peer before n=3*
*Peer after n=2*

Figure 4 compares the ratings given to the concert by audience members, peers and the Hallé Orchestra staff who acted as self-assessors and shows that for several dimensions the three groups were in broad agreement. Interestingly the peer who rated the concert for its ‘distinctiveness’ was much more positive than both the self-assessors and the audience in response to this question, but peers were much less convinced that the work was relevant or challenging and less certain about the importance of its impact on the local area. Of all three groups audience members were least likely to feel that the concert meant something to them personally, possibly because they did not have the same depth of relationship with the music as orchestra staff and critics.

**Figure 4: Average self, peer and public scores for Verdi bicentenary concert (awarded after the event)**




*Self n=3*
*Peer n=2*
*Public n=48*

**Hallé Orchestra says:**

In a city where there is no opera company, and with a music director who has an international reputation as conductor of the works of Wagner and Verdi in particular, the Hallé decided to celebrate the Verdi bicentenary in an unusual way.

Using Verdi's relationship with Boito as a start-point, we focused on acts of three of the late great operas, using international singers of the highest calibre and vast orchestral and choral forces. Boito's collaboration with Verdi followed his early criticism of the older composer – and was the catalyst for an astonishing late outpouring of creative output by Verdi. A pre-concert event – involving speech and music – told this story, and the concert celebrated it.

Of the nine criteria, we thought it should have scored highly against 'presentation', 'captivation' and 'rigour'. We felt that our peers and public would judge us well on doing something which was world class in execution, imaginatively thought through and fit for the city and concert hall which is our home. Whilst, after 150 years, this music itself doesn't break new ground (it built it) we decided that doing something less obvious than, say, a whole opera or the Requiem, was an original contribution to the celebratory year. As far as we are aware no one else did anything at all similar – at a time when celebrations of Verdi's birth were going on all over the world.

We decided to select as our peers national reviewers – as we felt this would give an informed and unbiased response to what we were trying to achieve. This was a risk worth

taking as far as we were concerned, and we chose not to select people who were known to us.

What the ratings show us is that the peers, in particular, didn't quite 'get' the challenge in what we presented, but we were pleased to receive a high rating for 'distinctiveness'.

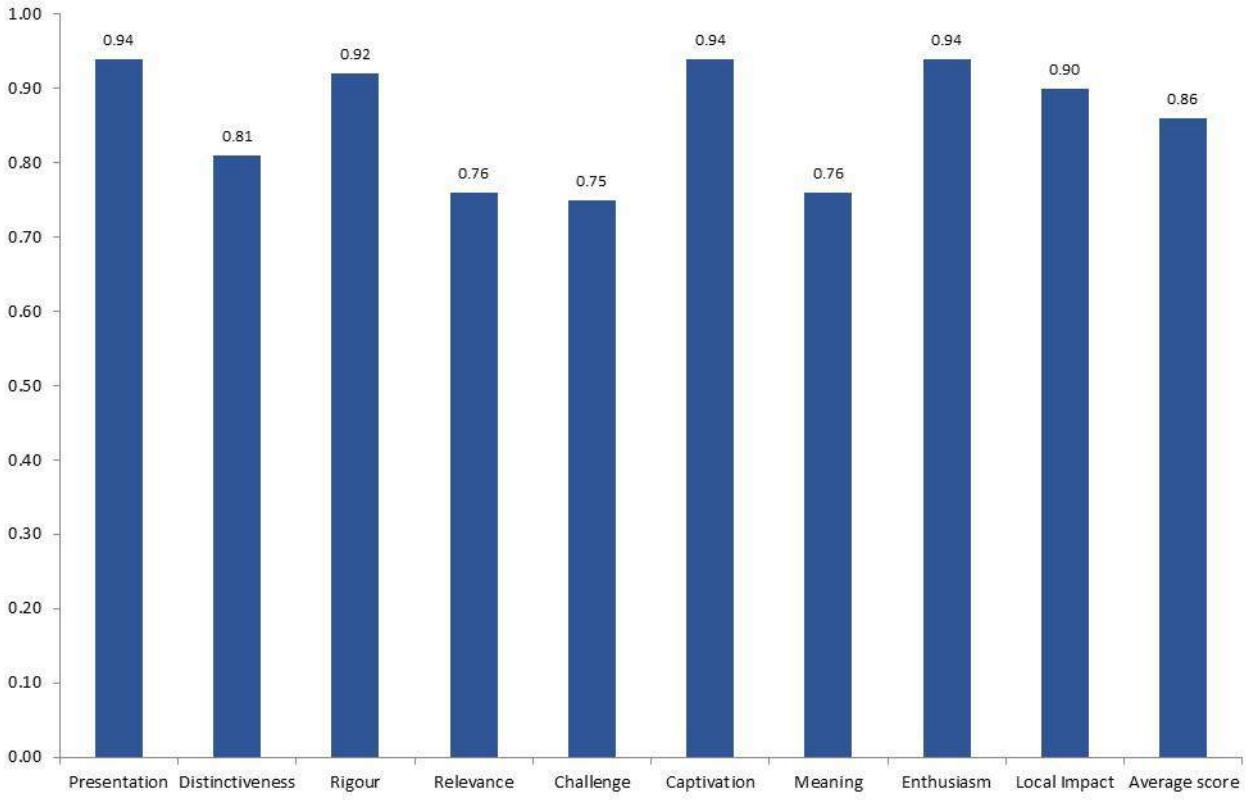
The most value for us in the process was outcome against expectation – as the approach develops we think this should lead to better and more targeted planning by the Hallé.

## 3.3 *That Day We Sang* – Royal Exchange

A total of 77 audience members were surveyed at one matinee and three evening performances of *That Day We Sang* at the Royal Exchange Theatre between 9 and 11 January 2014. The majority (59 per cent) of the sample were female and the median sample age was 58.

Of all the events in the Manchester pilot *That Day We Sang* was the most well received by members of the public who awarded an overall average score of 0.86. Figure 5 shows that the play scored well against all quality dimensions, with slightly lower scores for the more personal and subjective measures of 'relevance', 'challenge' and 'meaning'. The standard deviation of the public ratings for *That Day We Sang* was 0.20, which was fairly low compared to the amount of variation in responses for other events in the pilot, suggesting that audience members were relatively consistent in their reactions to the play.

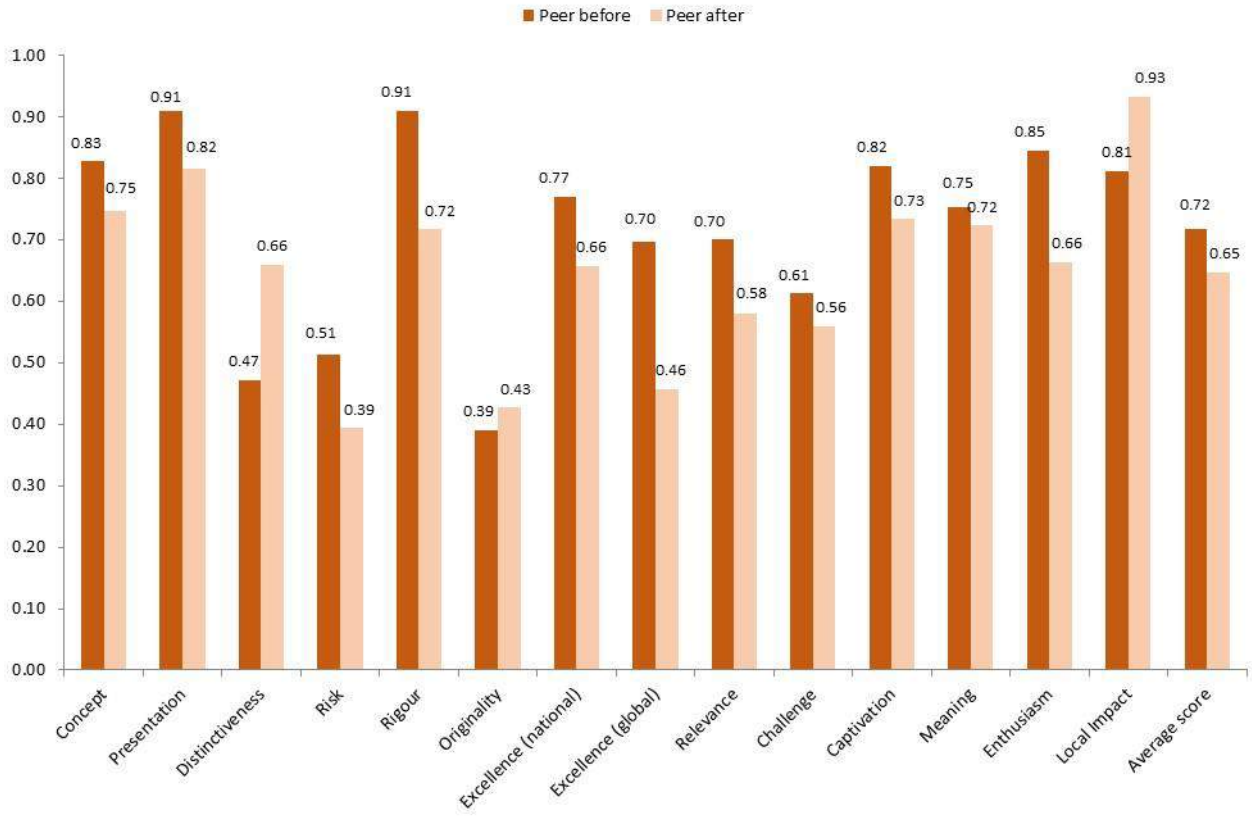
**Figure 5: Average public scores for *That Day We Sang***




*n=77*

Peers had high expectations of *That Day We Sang* and in particular expected a high level of technical accomplishment from the Royal Exchange, giving an average pre-event score of 0.91 for both 'presentation' and 'rigour'. Figure 6 shows that for most dimensions peer ratings afterwards were not quite as high as anticipated. One peer in particular felt that the show did not compare all that well to his experience of other work of this type in the UK and particularly internationally, resulting in a drop in the average peer score for 'excellence (global)' from 0.70 before the event to 0.46 afterwards. Peers were more inclined to agree afterwards that *That Day We Sang* was different from things they'd experienced before (awarding an average post-event score of 0.66 for 'distinctiveness') but as with the Verdi concert they didn't feel that the work was particularly ground-breaking, giving an average score of 0.43 for 'originality'. The lowest score awarded by peers after the event was 0.39 for 'risk', suggesting that the production wasn't seen as particularly challenging for the Royal Exchange.

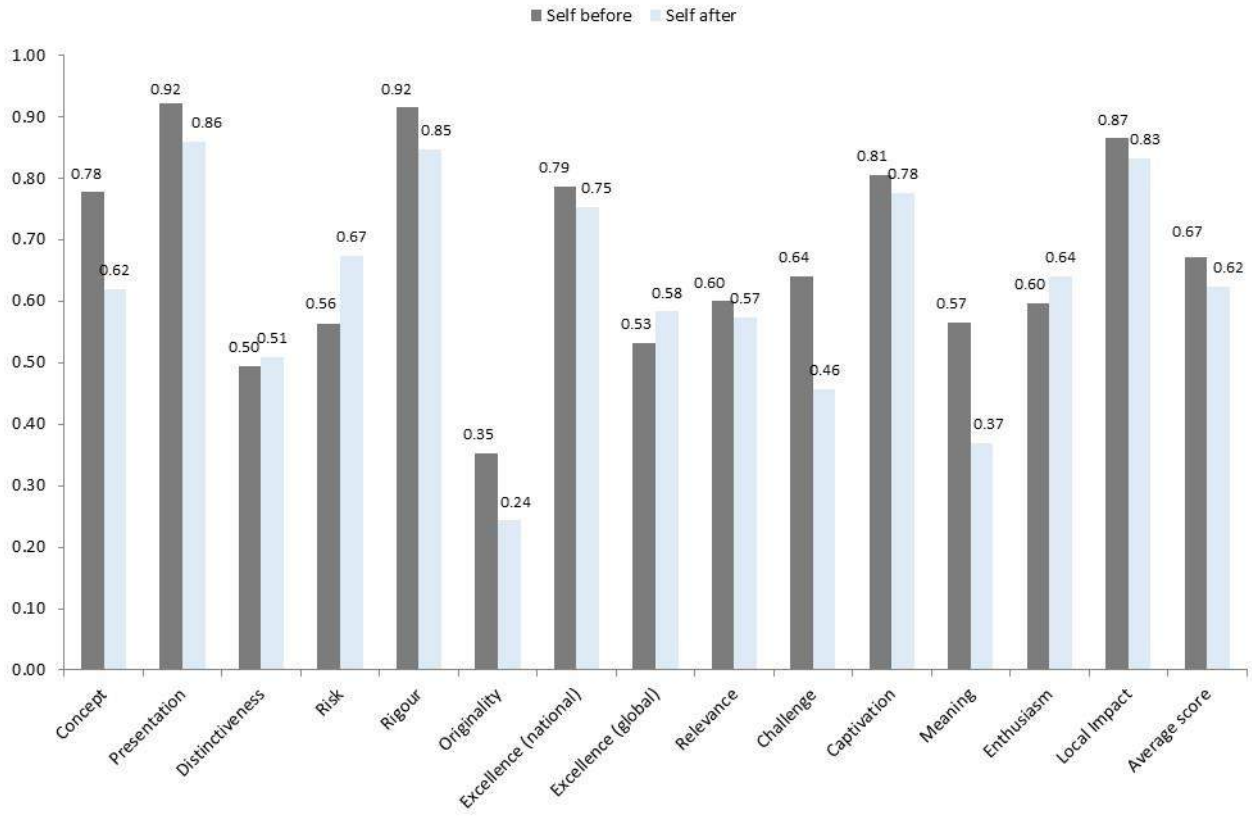
**Figure 6: Average 'before' and 'after' peer scores for *That Day We Sang***




*Peer before n=4*
*Peer after n=3*

Figure 7 shows that staff at the Royal Exchange also tended to be slightly less convinced about the quality of *That Day We Sang* after the event. Self-assessors were confident about the show's 'presentation' and 'rigour' but found that their experience of the work was less thought-provoking and meant less to them personally than they had anticipated, giving an average post-event score of 0.46 for 'challenge' and 0.37 for 'meaning'. The team at the Royal Exchange did not expect or perhaps intend *That Day We Sang* to be particularly 'original', but seemed to think that they had taken more of a 'risk' than peers appreciated.

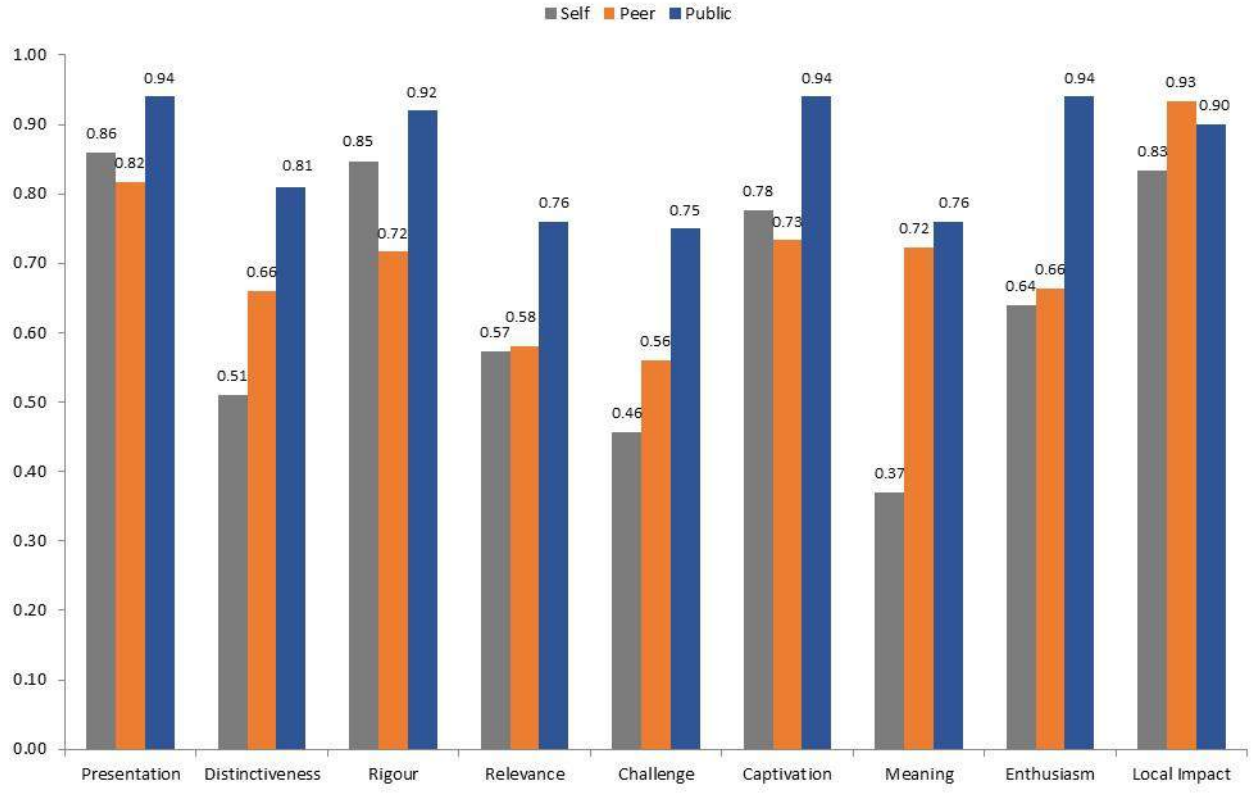
**Figure 7: Average 'before' and 'after' self scores for *That Day We Sang***




*Self before n=4*
*Self after n=3*

As shown in Figure 8, *That Day We Sang* is a good example of a show that is extremely popular with its audience despite experts both within the producing organisation itself and the wider peer community being slightly more restrained in their praise. Practically everyone in the audience at *That Day We Sang* who completed the survey was keen to go to something like this again: the public score for 'enthusiasm' was 0.94, compared with a peer score of 0.66 and a self-assessment score of 0.64.

**Figure 8: Average self, peer and public scores for *That Day We Sang* (awarded after the event)**




*Self n=3*
*Peer n=3*
*Public n=77*

**The Royal Exchange says:**

We programmed *That Day We Sang* by Victoria Wood in our Christmas slot. The Lowry was presenting War Horse this year so we wanted to find something that would not put us in direct competition to this and would complement the offer in the city in what has become a competitive market. *That Day We Sang* was first produced by Manchester International Festival 2011 in the Opera House. Alex Poots suggested that this might be something we would want to re-create for our theatre in the round space. We thought it was a perfect fit: it was a warm big hearted piece of work that would attract a broad range of audiences; it had the name of Victoria Wood attached to it; Sarah Frankcom our Artistic Director would direct; it was a Manchester love story that would suit our ambition to tell great stories in our intimate space; it was funny; it had a reasonable size cast with a chorus of 90 children who could take part in an intensive rehearsal period with an expert choir master Jeff Borrowdale; and it would stretch us and our production capability, as musicals were new to us this season.

The quality dimensions that were important to us were ’presentation’, ’rigour’, ‘captivation’, ‘excellence (national)’ and ‘local impact’. We expected that audiences would respond well to it – given it had been very successful at Manchester International Festival, had the name Victoria Wood attached to it and was being produced by us. The challenge for us was making it work in the round and that it was a musical play and therefore had huge requirements around sound and sound design. Musicals are a new departure for us and part of this season has been about testing our ability in this area with this and *Sweeney Todd*. We didn’t really expect our peers or audiences to recognise this as a challenge. We

also think that it was to be expected that both we and our peers would be harsher around dimensions to do with ‘risk’, ’originality’, ‘excellence (global)’ and ‘meaning’ because this was a popular Christmas show and in many ways a safe bet.

We are delighted that the audience responded so well across all dimensions and it is interesting to note that they did mark lower for ‘relevance’, ‘challenge’ and ‘meaning’ as expected. We think peer and self-assessors would perhaps always see the ‘Christmas show’ as being a more popular, less challenging piece with less national and international impact and less risk, and the survey confirmed this.

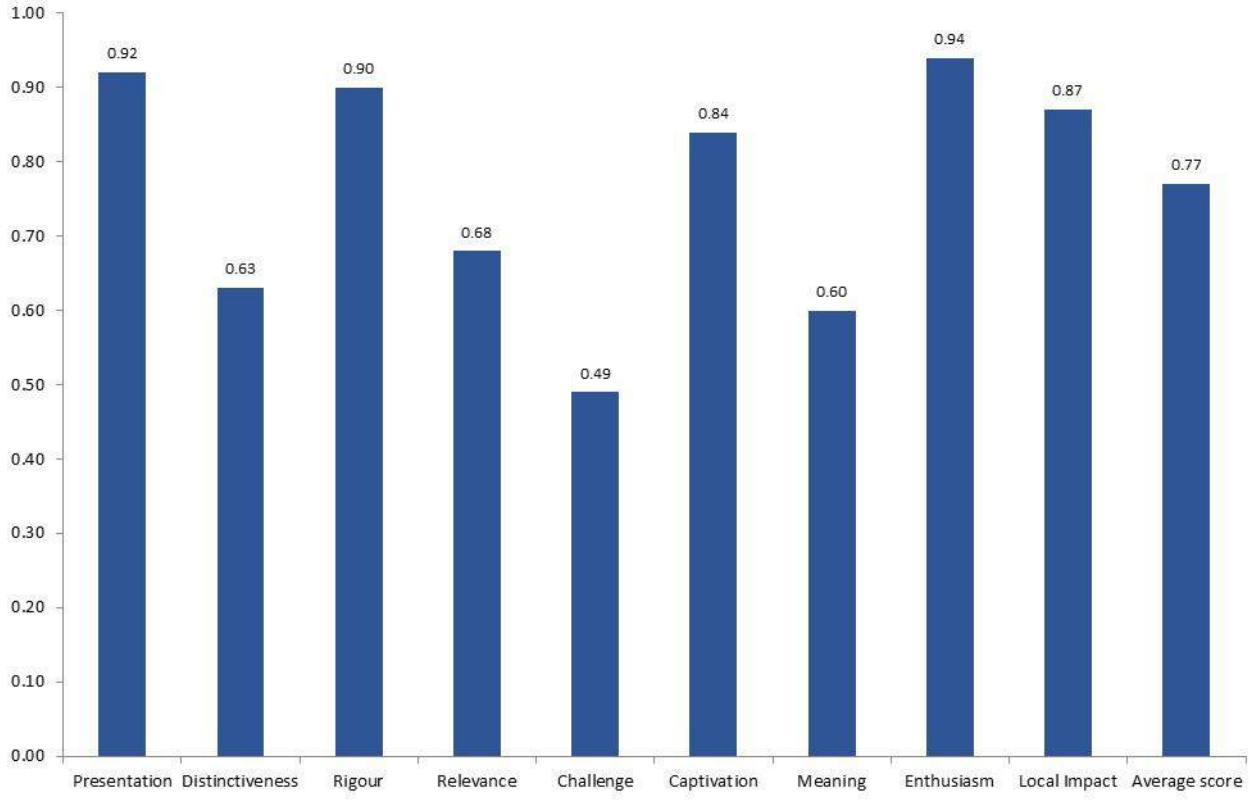
It would be fascinating to run the surveys again for a piece of work that we thought was more challenging and original.

## 3.4 *Jack and the Beanstalk* – Oldham Coliseum

A total of 52 audience members took part in the Culture Counts survey at *Jack and the Beanstalk* at Oldham Coliseum over one matinee and two evening performances on 6 and 7 December 2013. The sample was fairly evenly split between men and women and, as might be expected for a pantomime, the audience sample was relatively young, with a median age of 43.5.

Figure 9 shows that audience members enjoyed *Jack and the Beanstalk* and felt that it was well planned, produced and presented. Unsurprisingly for a pantomime, respondents didn’t seem to feel that it offered more than a good night out and the show received the lowest scores in the pilot for ‘challenge’ (0.49) and ‘meaning’ (0.60). *Jack and the Beanstalk* had the highest level of variation in audience response of all eight events in the pilot, with a sample standard deviation of 0.28.

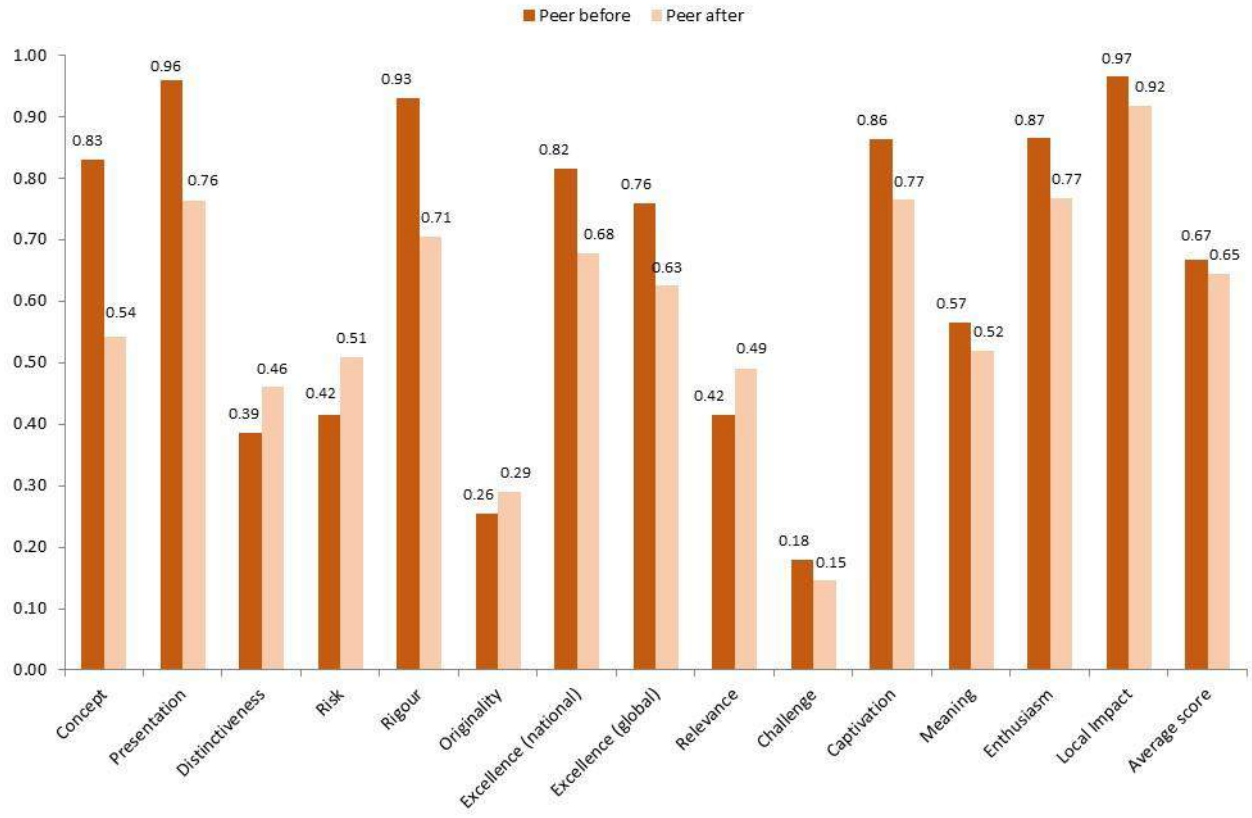
**Figure 9: Average public scores for *Jack and the Beanstalk***




*n=52*

Peers were clearly expecting *Jack and the Beanstalk* to be an extremely high quality example of its genre, giving very high pre-event scores for 'presentation' (0.96), 'rigour' (0.93), 'excellence (national)' (0.82) and 'excellence (global)' (0.76). Figure 10 suggests that the pantomime was not produced to quite the standard they were expecting, and in particular that peers weren't particularly inspired by the underlying idea, giving an average post-event score of 0.54 for 'concept' compared with 0.83 beforehand. Nonetheless the peer scores for the challenging criteria of 'excellence (national)' and 'excellence (global)' were still high after the event at 0.68 and 0.63 respectively, and, like members of the audience, peers clearly enjoyed their experience, awarding a post-event score of 0.77 for both 'captivation' and 'enthusiasm'.

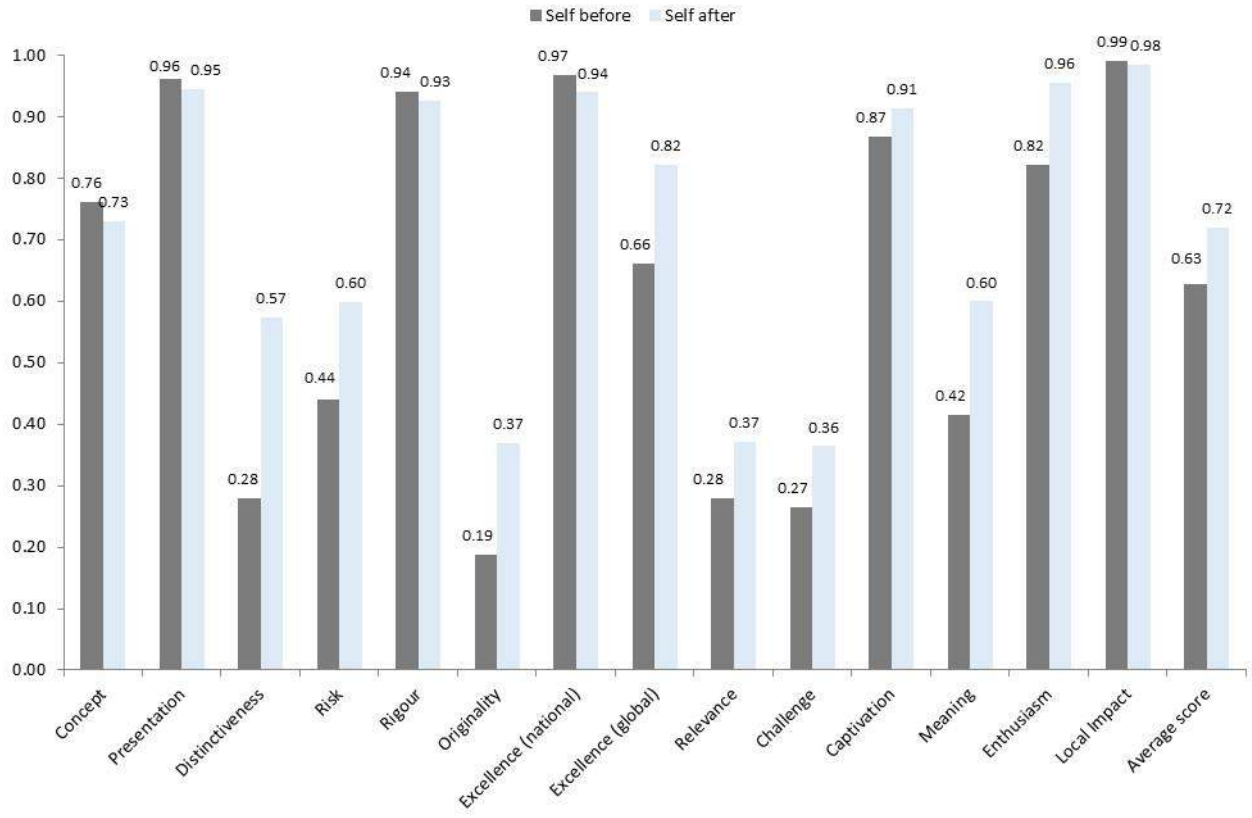
**Figure 10: Average 'before' and 'after' peer scores for *Jack and the Beanstalk***




*Peer before n=4*
*Peer after n=4*

Figure 11 indicates that the team responsible for *Jack and the Beanstalk* at Oldham Coliseum did not experience the same slight disappointment with the technical qualities of the pantomime as peers. Self-assessment scores were high both before and after the event for 'presentation' and 'rigour' and the team felt confident about the excellence of the show relative to others of its type both nationally and globally. Both peer and self-assessor groups thought that the pantomime was more distinctive, risky and original than expected.

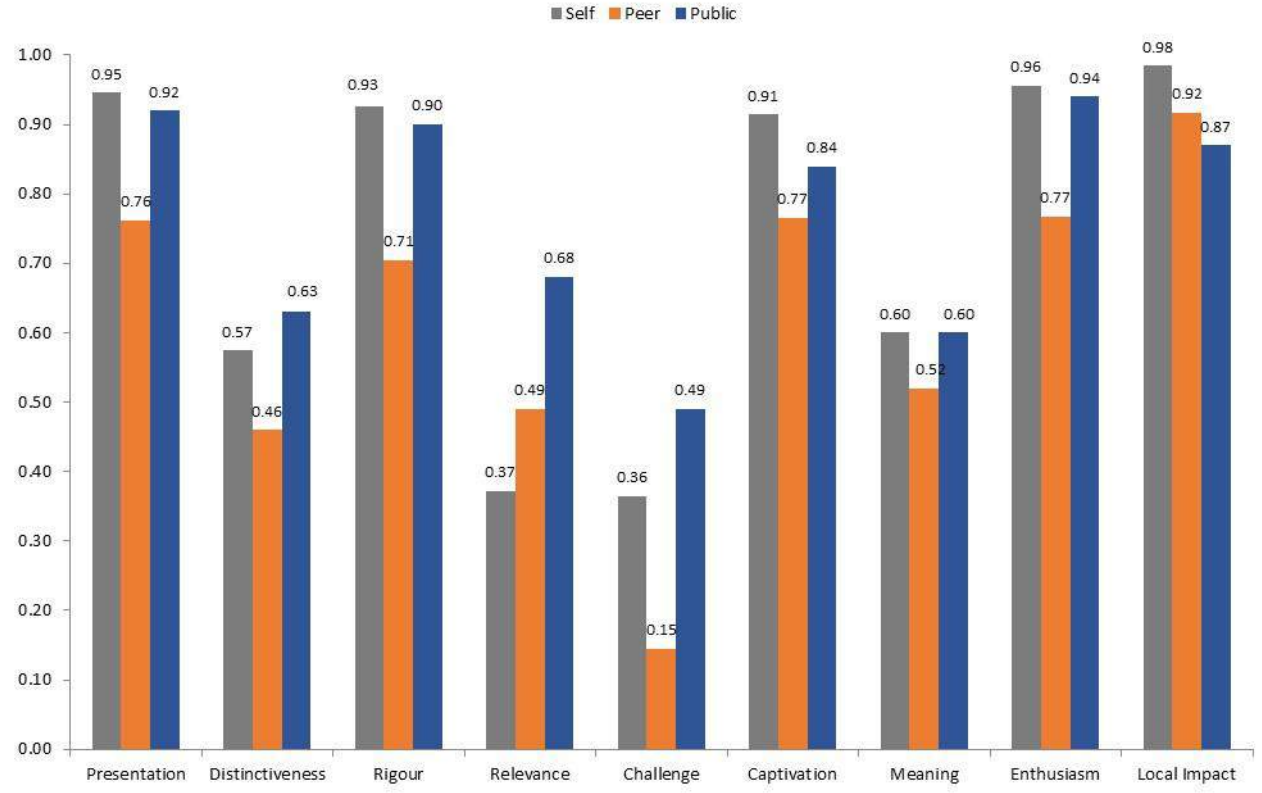
**Figure 11: Average 'before' and 'after' self scores for *Jack and the Beanstalk***




*Self before n=7*
*Self after n=7*

The comparison of self, peer and public scores in Figure 12 shows that the theatre and its audience were in broad agreement about the quality of the pantomime, while peers were generally slightly more critical. Audience members were more likely to feel that the show had something to say about the world today, awarding a 0.68 for 'relevance' compared to 0.49 from peers and just 0.37 from the Oldham Coliseum team.

**Figure 12: Average self, peer and public scores for *Jack and the Beanstalk* (awarded after the event)**




*Self n=7*
*Peer n=4*
*Public n=52*

**Oldham Coliseum says:**

The Coliseum's annual pantomime was included in the initial application of the metrics as a deliberate challenge to the pilot. Our intent was to commission and produce a well-crafted, entertaining family show that built on the long-established tradition of 'panto' in the north of England. Consequently we expected assessment from peers, self and public to score highly in 'presentation', 'rigour', 'excellence (national)' and 'captivation'. We also expected a high score for 'Local impact', considering the absence of competing activity within the borough and our perception that the local audience was not well-disposed to travel.

We neither intended nor expected to score highly in areas of 'risk', 'relevance' or 'challenge'; indeed we were aware of an element of awkwardness in answering specific questions associated with these elements in relation to this particular product. Reflection on the metric of global excellence of a pantomime was a particularly strange exercise, as this piece was the expression of a geographically localised and specific popular culture. Indeed it could be argued that as this form of cultural production is not seen elsewhere, it must be globally excellent as well being of a high quality nationally.

Our expectations were broadly met by peer and audience response to our work, although the public assessment of our work scored highly in these latter areas, perhaps reflecting a general loyalty towards the Coliseum brand. In those areas where we expected the show to excel, the analysis of our staff and the public were broadly similar, although here the perception of peers suggested a lower quality.

An interesting trend which we don't really understand and would like to explore further is that peer reviewers of the work exhibited slight disappointment with the realisation of the work, whereas self-assessors believed the realisation was slightly better than expected. We suspect that the differences in these data are probably not statistically significant, but we would like to explore whether there is an underlying trend with further study.

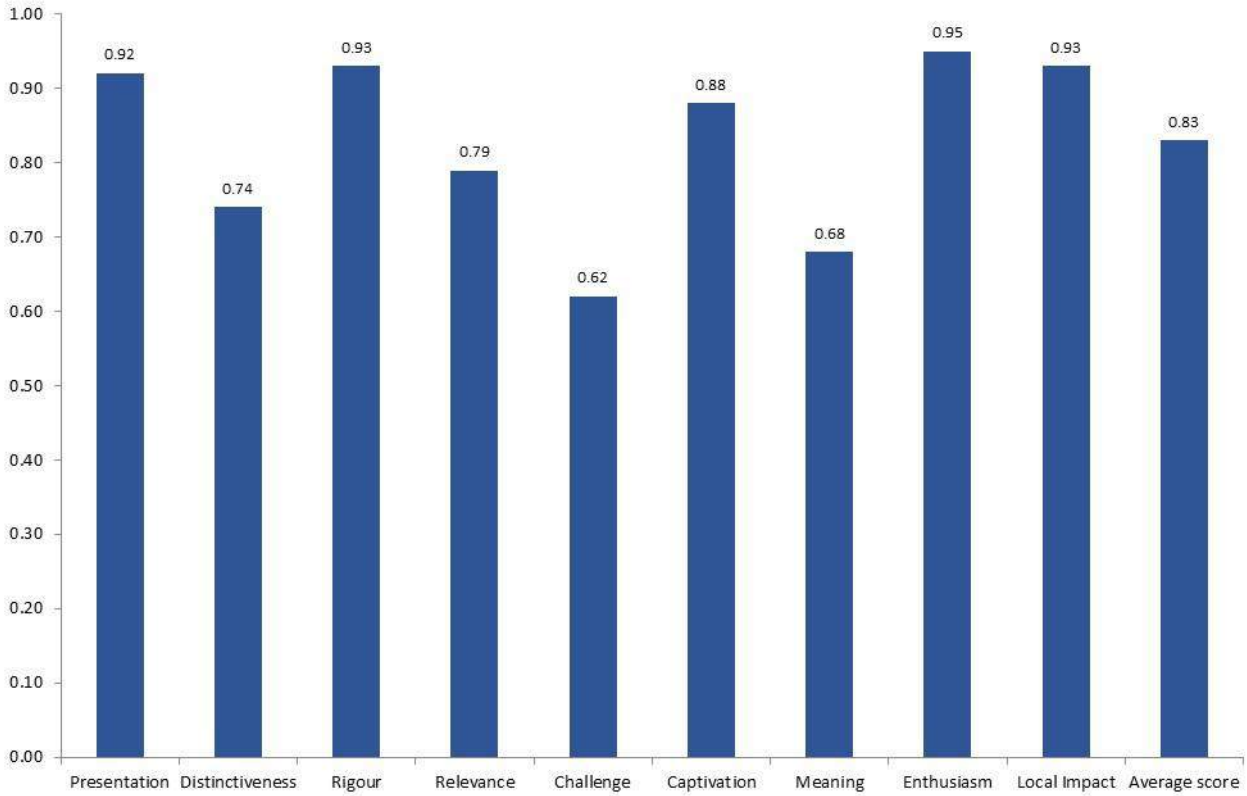
The pilot has given us an interesting insight into the production of our work and this particular product stretched the design of the metric set. We are keen to follow up this work across a broader range of our cultural production.

## 3.5 *Robin Hood* – Octagon Bolton

A total of 57 audience members were surveyed at two matinee and two evening performances of *Robin Hood* at Octagon Bolton on 20 and 21 December 2013. The sample was two-thirds women and the median sample age was 48. Unlike the previous show, *Jack and the Beanstalk*, the Octagon team would not describe *Robin Hood* as a pantomime, but rather a 'play with music written for school and family members.'

Figure 13 shows that the audience response to *Robin Hood* was similar to *Jack and the Beanstalk*, although with slightly higher scores awarded overall (the average public score for *Robin Hood* was 0.83 compared with 0.77 for *Jack and the Beanstalk*). Of all eight Manchester events *Robin Hood* received the highest average score from audience members for 'local impact'; Octagon Bolton's productions may be seen as particularly important to their local area compared with similar events in central Manchester where there is a wealth of leisure activities to choose from.

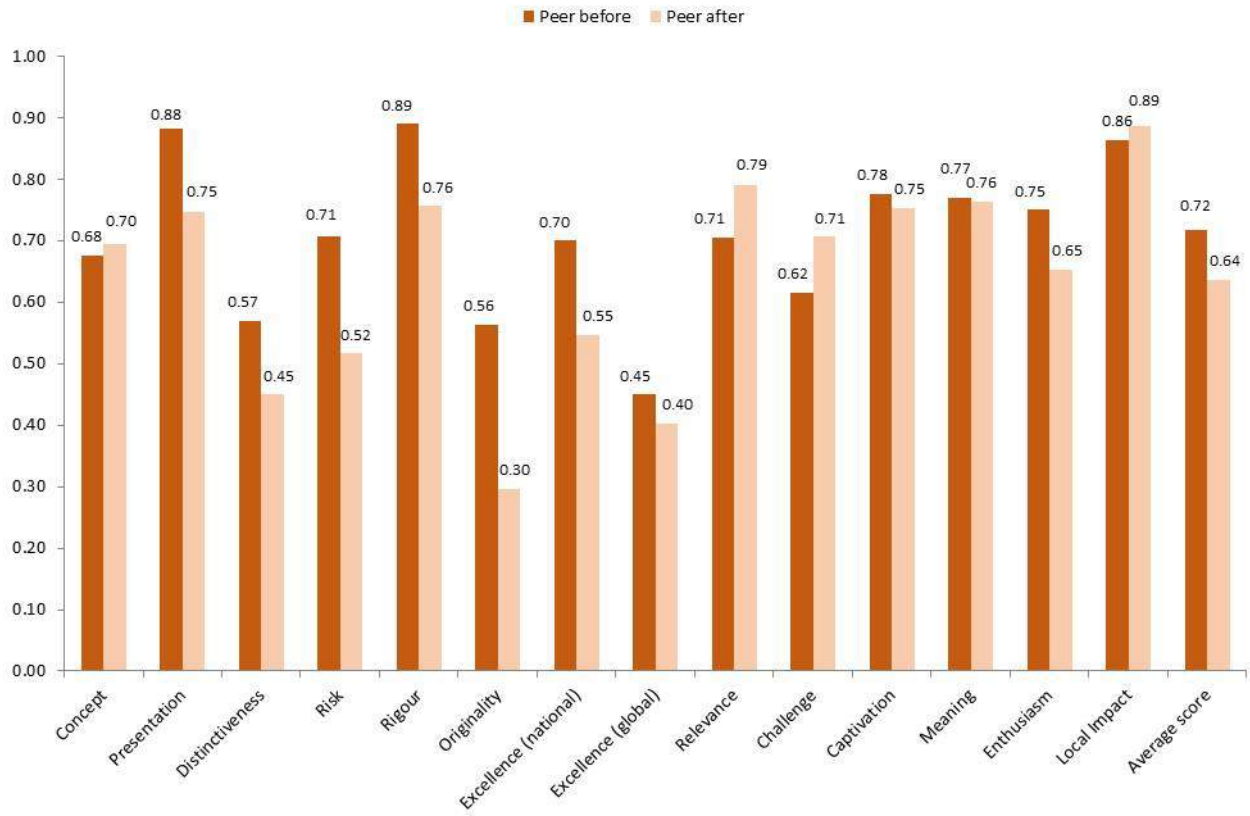
**Figure 13: Average public scores for *Robin Hood***



*n=57*

Peer expectations of *Robin Hood* were fairly high, including the areas of 'risk', 'distinctiveness' and 'originality'. Figure 14 suggests that these expectations may not have been entirely realistic as peer ratings after the event were somewhat lower for a number of dimensions, particularly 'originality' which received an average post-event peer score of 0.30 compared 0.56 beforehand. Post-event peer ratings were relatively high for both 'relevance' (0.79) and 'challenge' (0.71), suggesting that there was quite a lot of substance to the work as well as it being a fun family Christmas event.

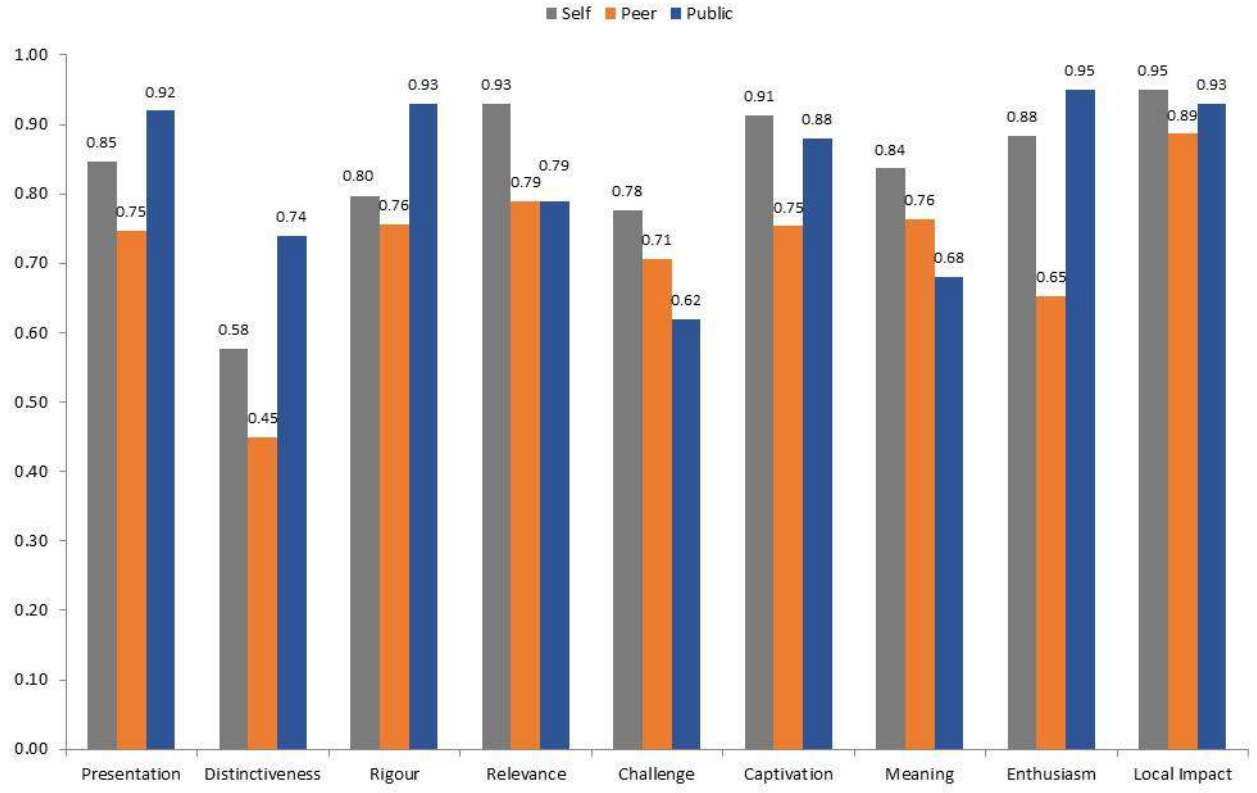
**Figure 14: Average 'before' and 'after' peer scores for *Robin Hood***




*Peer before n=3*
*Peer after n=3*

Figure 15 suggests that the Octagon team, peers and audience members responded to the show in slightly different ways. As is often the case, peers were slightly more critical than artists or the audience for a number of dimensions. Unusually, however, both the Octagon team and peers felt that the show was more thought-provoking and meaningful to them personally than audience members did; at the same time, audience members were more likely to feel that they'd experienced something different.

**Figure 15: Average self, peer and public scores for *Robin Hood* (awarded after the event)**




*Self n=3*
*Peer n=3*
*Public n=57*

**Octagon Bolton says:**

*Robin Hood* was conceived to be a new play with music for schools and family audiences. It follows our tradition of adapting literary classics and other well-known stories as festive shows. Each is commissioned to have a strong Bolton connection and to include a cast of local children and the emphasis is on a highly enjoyable yet meaningful theatre experience. We selected *Robin Hood* as a play which we felt was relevant to current political and social issues in the town, and likely to offer particular substance. This production always attracts a wide socio-economic demographic, particularly as we run targeted work and audience development for it with social housing tenants, for whom issues such as the bedroom tax are important. The festive production is the biggest in the theatre's programme and an essential part of our offer, as for many this will be their only contact with the Octagon. It is perhaps therefore inevitable that the work would not be considered highly distinctive or original given its need to meet traditional expectations and wide appeal, though we need to reflect that we still didn't meet peer expectations on these dimensions.

The dimensions of 'enthusiasm', 'local impact, 'captivation' and 'relevance' were particularly significant for this production and our intended experience for the audience. For the first three of these we scored well with audiences and peers which we were pleased to see. It was also pleasing to see such a high audience score for 'presentation'. That 'enthusiasm' scored so highly is important given that for many, the festive show forms part of an annual family tradition. Both audience and peers scored 'relevance' quite highly, showing our intentions were met though not as highly as we ourselves felt. Public scores for 'relevance' may have been higher had the interviewers been present at some of the cheaper off-peak

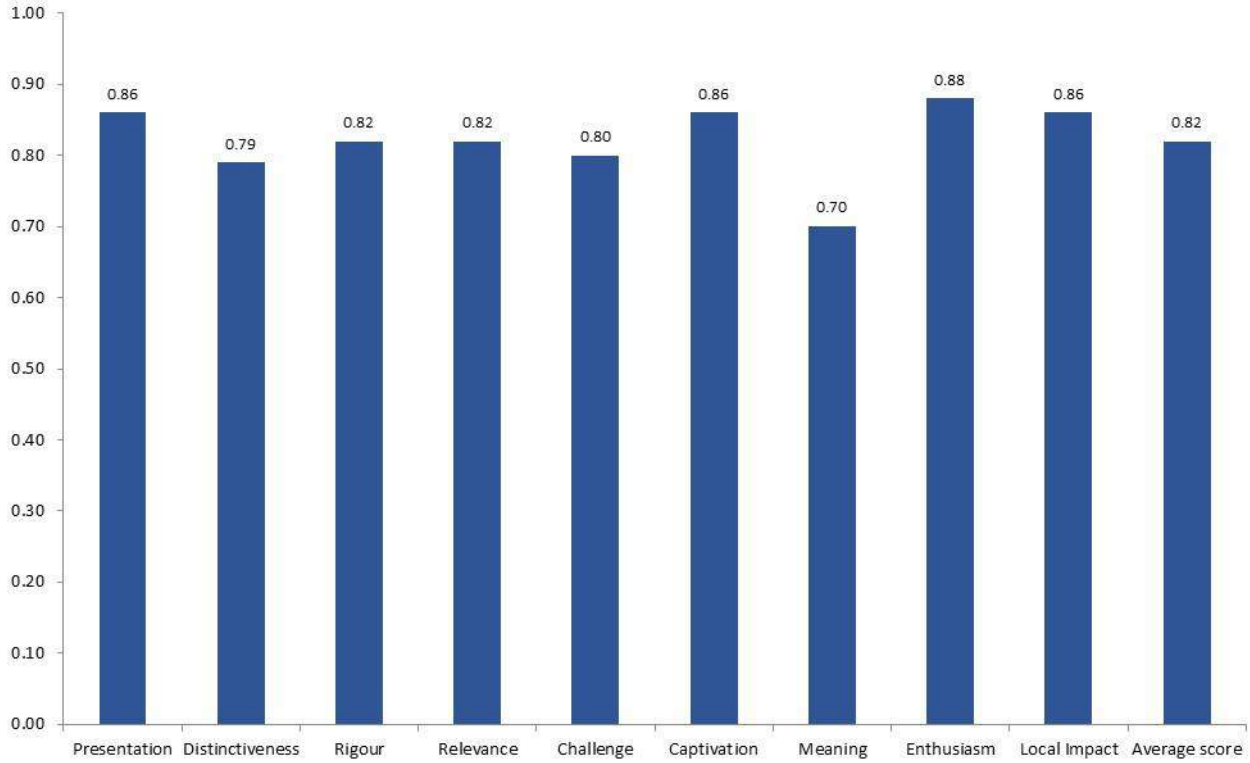
performances when our targeted audience group of social housing tenants were most likely to attend.

## 3.6 *CYAC: Advent Avenue* – Contact

*CYAC: Advent Avenue* was a Christmas show performed by Contact Young Actors Company. As expected the survey sample was much younger for this event: the median age of the 90 people surveyed was 26, the youngest of all the events in the Manchester pilot, and nearly half (47 per cent) of all respondents who provided their age were under 25.

As shown in Figure 16, audience scores for *CYAC: Advent Avenue* were fairly high across all dimensions. However the audience response to this event was slightly different to the other theatre performances in the pilot. *CYAC: Advent Avenue* didn't score as highly as *That Day We Sang*, *Jack and the Beanstalk* or *Robin Hood* in the more technical areas of 'presentation' and 'rigour', but it was seen by its audience to be more relevant and thought-provoking. *CYAC: Advent Avenue* had the lowest level of variation in audience response with a sample standard deviation of 0.19.

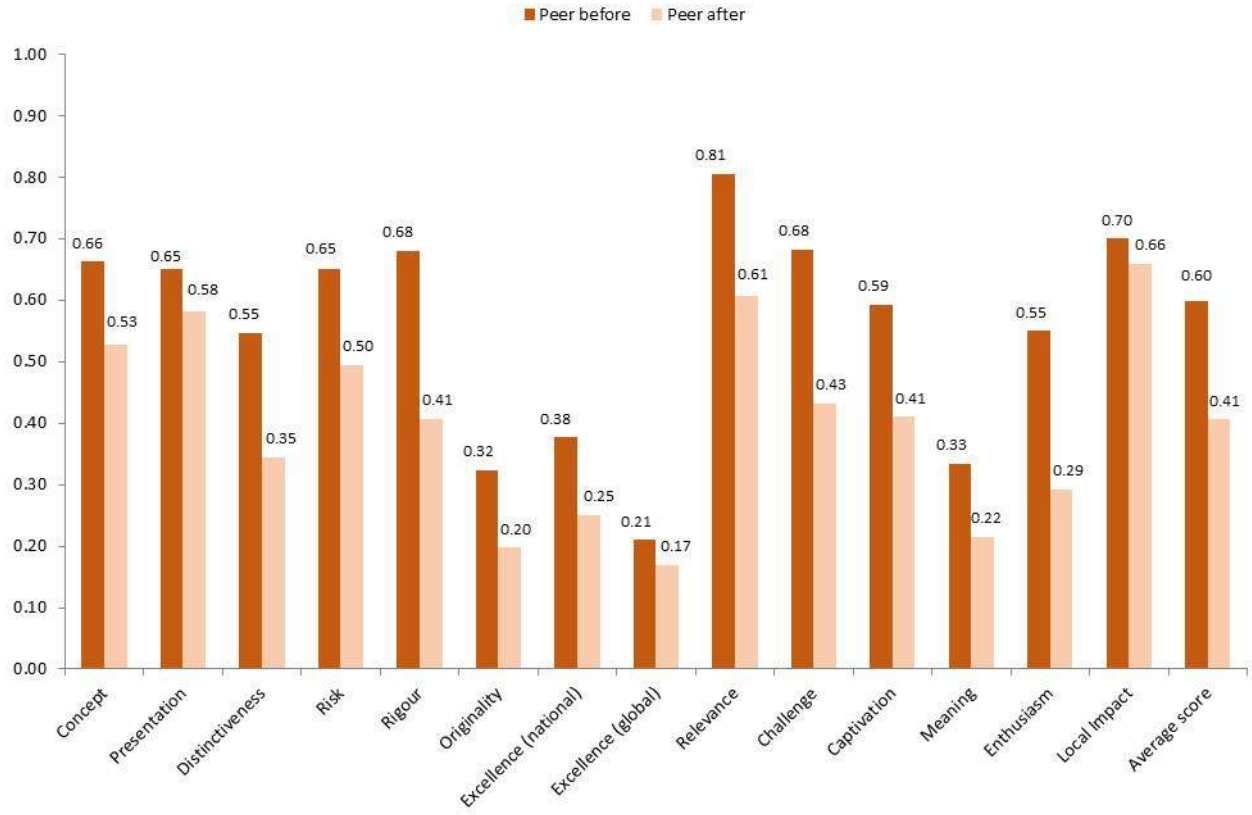
**Figure 16: Average public scores for *CYAC: Advent Avenue***



*n=90*

Peers were less impressed by *CYAC: Advent Avenue* than audience members and the show received the lowest post-event average peer score in the pilot. Figure 17 shows that peer expectations were not particularly high except for 'relevance' and ratings were lower after the event for every dimension. *CYAC: Advent Avenue* received its highest post-event peer scores for 'relevance' (0.61) and 'local impact' (0.66) but was not seen as being amongst the best of its type, receiving an average post-event peer score of 0.25 for 'excellence (national)' and 0.17 for 'excellence (global)'.

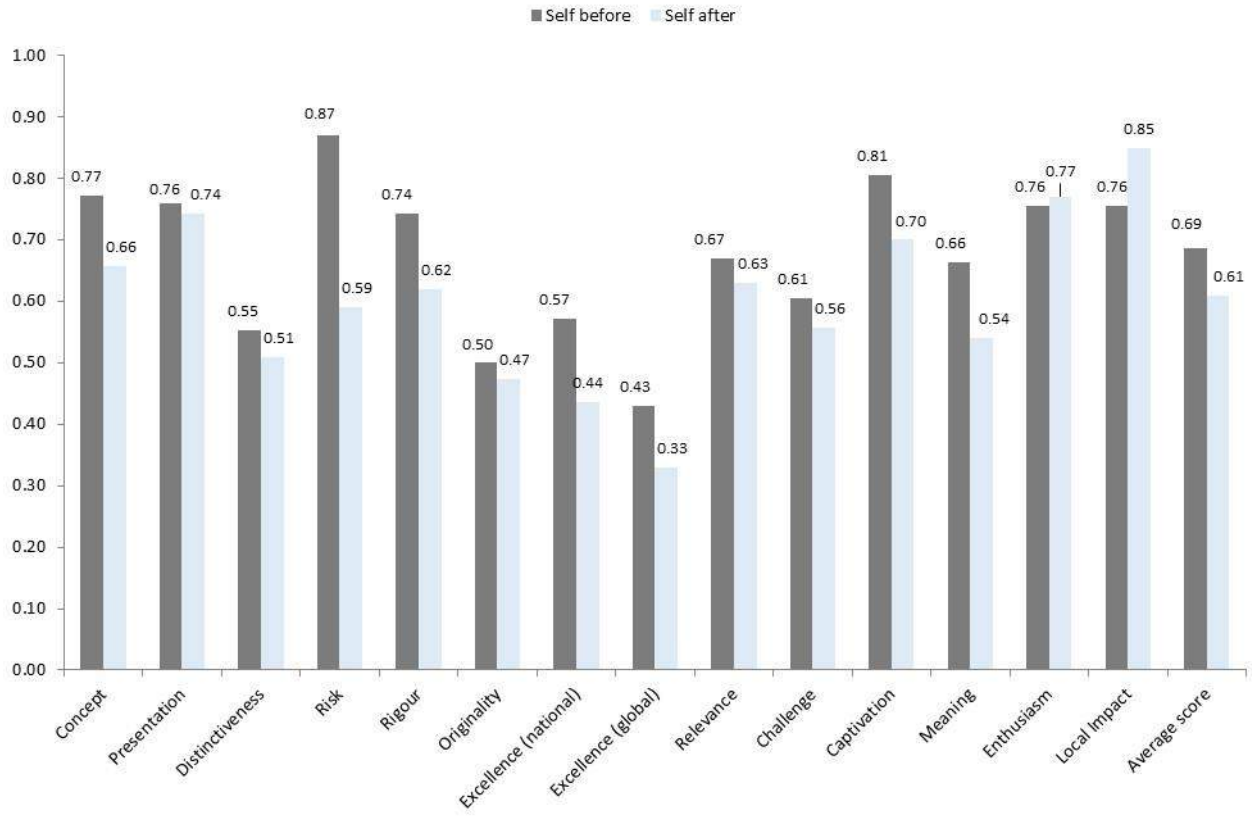
**Figure 17: Average 'before' and 'after' peer scores for *CYAC: Advent Avenue***




*Peer before n=4*
*Peer after n=4*

Figure 18 shows that the self-assessors at Contact tended to agree with their peers, although they were slightly less critical overall. For most dimensions ratings by self-assessors were lower after the event than before and, like peers, the Contact team did not rate their own show particularly highly in relation to national and international standards.

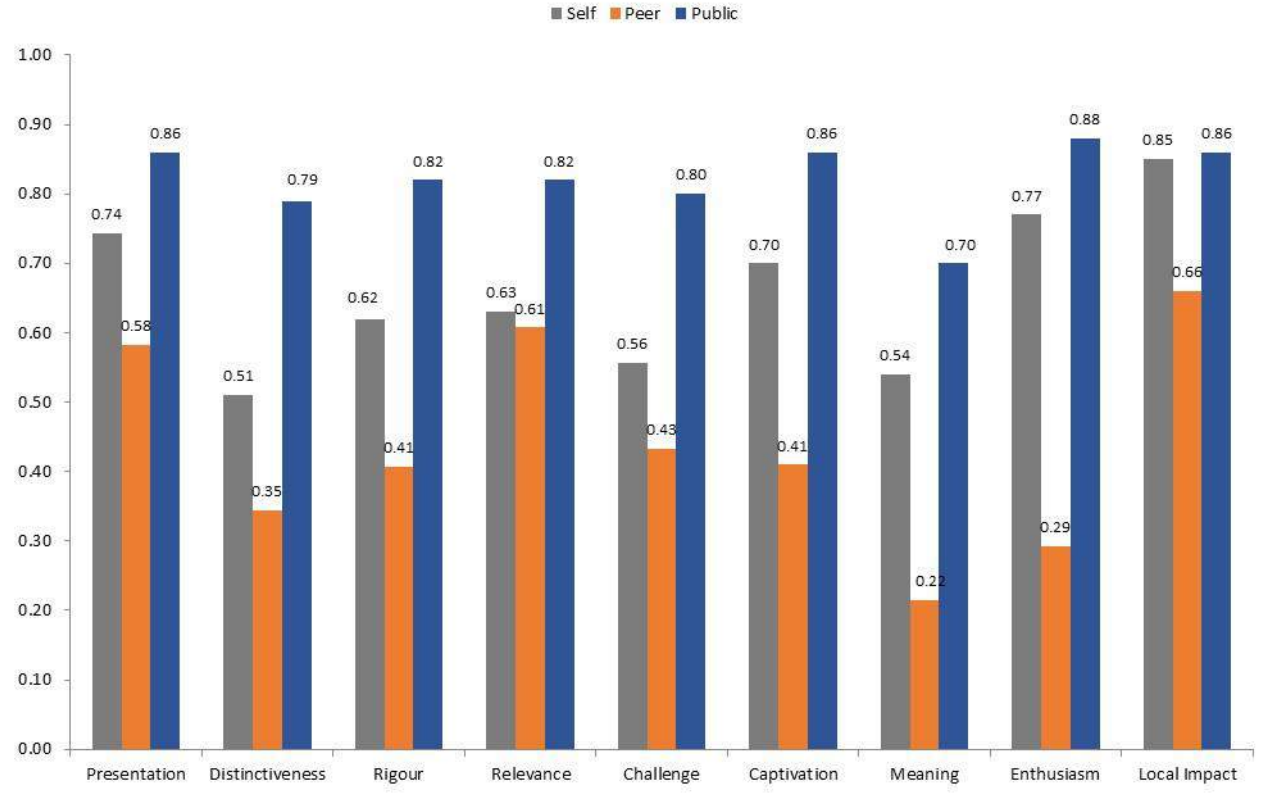
**Figure 18: Average 'before' and 'after' self scores for *CYAC: Advent Avenue***




*Self before n=4*
*Self after n=3*

Figure 19 clearly illustrates the disparity between the enjoyment and appreciation of *CYAC: Advent Avenue* by its young audience and the much more critical views of peers, with self-assessors taking a position somewhere in between. Contact team members were not hugely assertive in rating the quality of their show but perhaps were more aware of what their particular demographic would enjoy than peers.

**Figure 19: Average self, peer and public scores for *CYAC: Advent Avenue* (awarded after the event)**




*Self n=3*
*Peer n=4*
*Public n=90*

**Contact Theatre says:**

The Contact Young Actors Company production of *Advent Avenue* brought together a diverse group of 19 young people (aged 16–24), from a wide range of backgrounds in Greater Manchester, to create a new piece of theatre. Most had not met before, and this was their first experience on a professional stage. The young people worked two evenings per week for eight weeks, followed by a production period, working with a professional facilitator and director to develop their ideas for characters and content, and devising these into a public performance for our main stage. Our aims were: to provide a safe, supportive and creative development process for the young people; to deliver a wide range of confidence building and skills development activities; and to produce a relevant, fun and accessible public Christmas show for teenage and up audiences, made and performed by local young people.

It was clear that the metrics would only provide value judgments of the finished public show (rather than of the project's core purposes around the experience and development of the young people involved); but as our aspiration was also to create a piece of work as close to professional standards as possible, we felt it would be exciting to propose the show for metrics-based assessment.

We hoped that the work would rate as relevant, of local significance and of high production quality, and that the young people's skills, confidence and ownership of the material would come across to audiences and peers alike (perhaps through the 'enthusiasm' category). Our ambition is always that both the participatory process and the resulting product are of a

high national standard for young people's work, and we were slightly disappointed as a team in the final product, and this aligns with peers' responses. We expected audiences to have a higher appreciation of the show, especially as the show was made for and by young people, and the audience was heavily made up of young people – and this was borne out by the results.

We found the metrics results really interesting, especially the wide disparity between the public and peer assessments – one of the largest differences across all the projects in the pilot. Our staff were perhaps more closely aligned with the peer views than the public, in that we considered the final work not to be nationally leading, or of the highest dramaturgical quality, though individual performances and overall production values were strong. The work was made for a young adult audience, and that audience rated the work highly, and found it relevant. We have mixed feelings about the decision to propose a primarily participatory process up for metrics scrutiny, but agree that the peer assessment, though critical, does chime with our internal sense of the strengths and weaknesses of the final production.

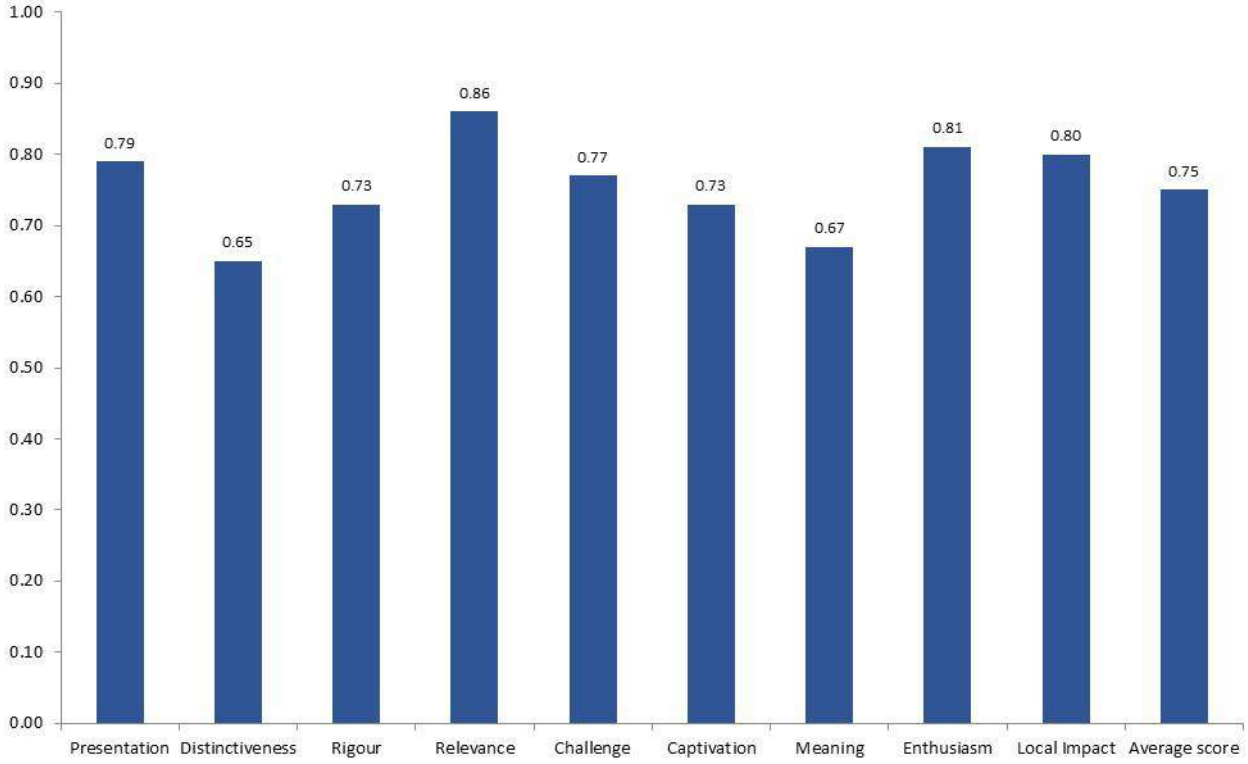
If we were to propose such work again in the future, we think it is important to consider, especially for peers, whether there should be a contextual statement about the participatory and young people's nature of the work, or a specific set of metrics that apply to these areas of practice.

### 3.7 *Jeremy Deller: All That Is Solid Melts Into Air* – Manchester Art Gallery

A total of 133 survey responses were collected from visitors to an exhibition entitled *Jeremy Deller: All That Is Solid Melts Into Air* at Manchester Art Gallery from 9 to 11 December 2013. The sample was relatively young, with a median age of 42, and contained more women (57 per cent of those who recorded their gender) than men (43 per cent).

Figure 20 shows that the public response to the Jeremy Deller exhibition was slightly different to the performing arts events examined so far. The exhibition scored fairly well across the board but was not rated particularly highly for the more technical dimensions, receiving the lowest average scores in the pilot for 'presentation' (0.79) and 'rigour' (0.73). However visitors clearly felt that Jeremy Deller had something meaningful to say about the world today: the exhibition received the highest score in the pilot for 'relevance' and was the only event to gain its highest score for this dimension.

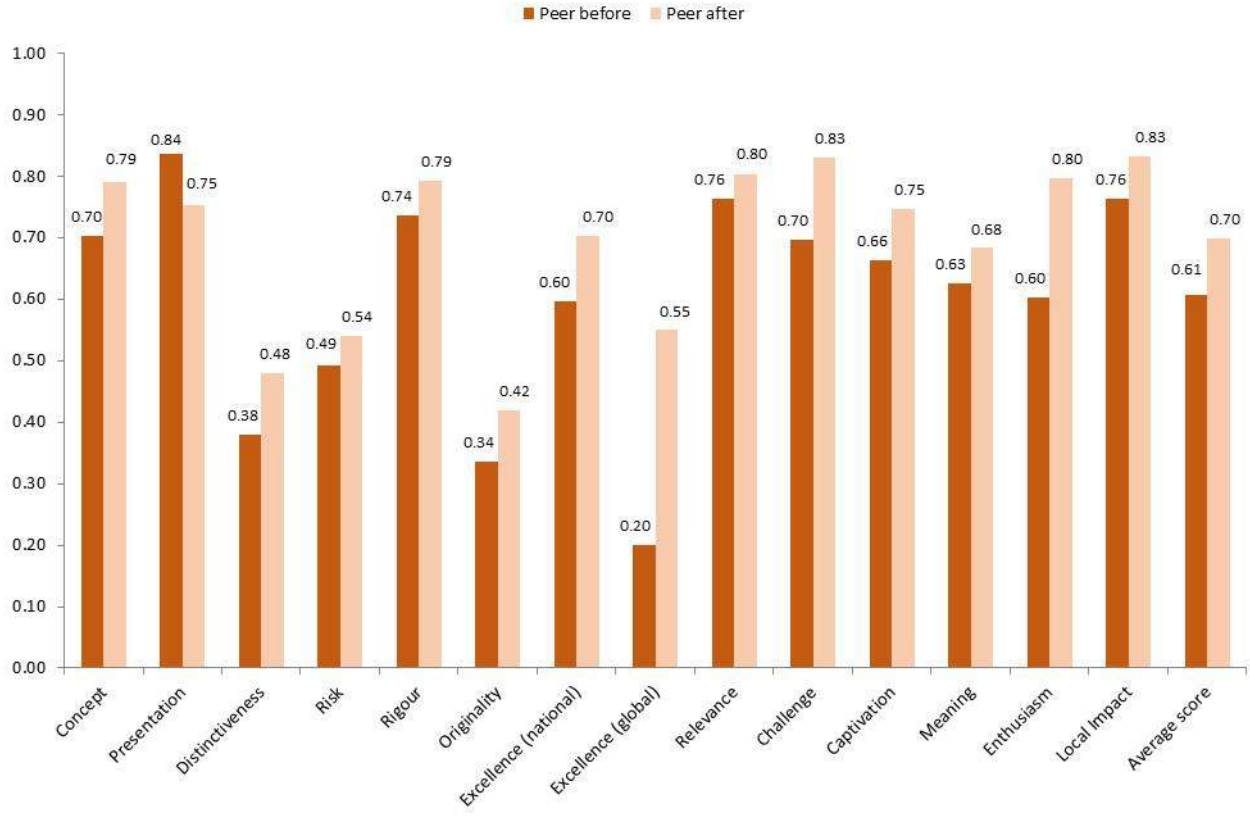
**Figure 20: Average public scores for *Jeremy Deller: All That Is Solid Melts Into Air***




*n=133*

Peers were impressed by the exhibition and awarded higher scores than expected for almost every dimension, as shown in Figure 21. Peers were particularly positive about the quality of the show compared to international benchmarks, with their average score for 'excellence (global)' rising from 0.20 before visiting the gallery to 0.55 afterwards.

**Figure 21: Average 'before' and 'after' peer scores for *Jeremy Deller: All That Is Solid Melts Into Air***

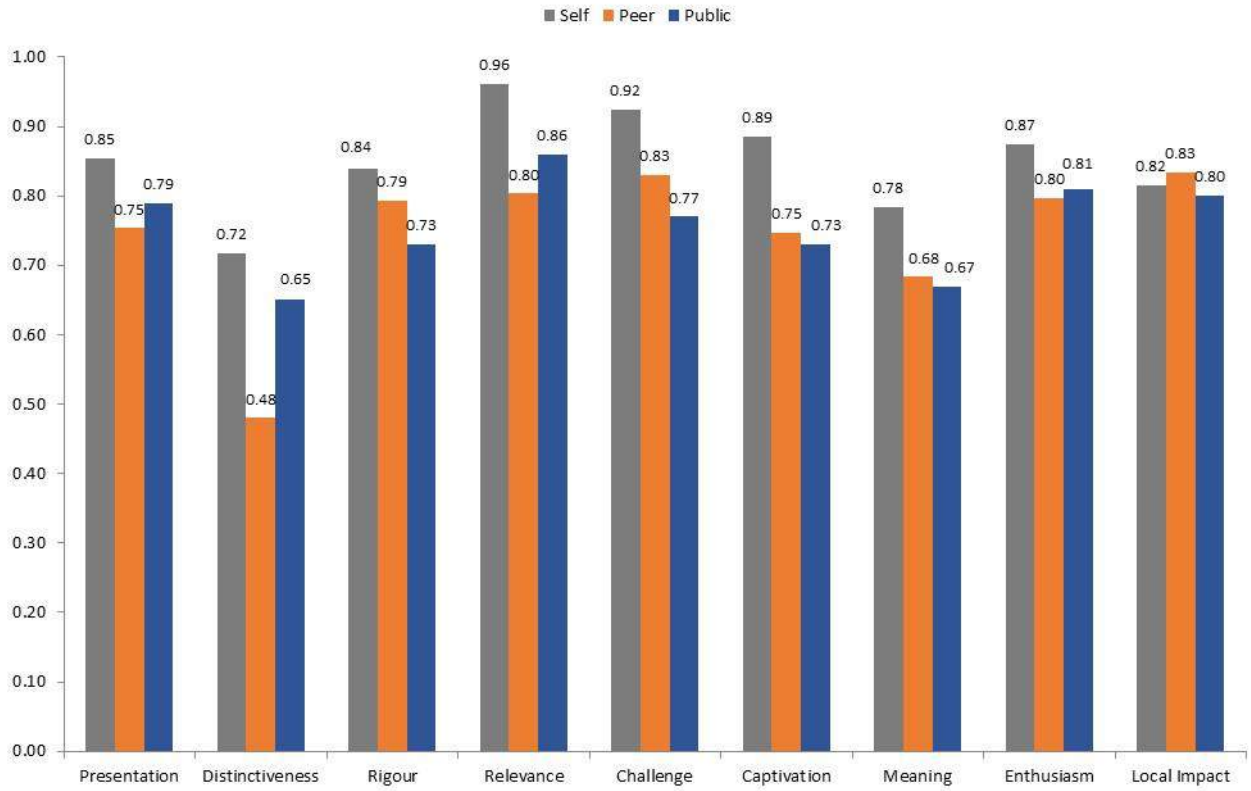



*Peer before n=3*
*Peer after n=3*

The Jeremy Deller exhibition had already opened when the Manchester pilot began and so self-assessors were asked to complete a post-event survey only. In addition, the exhibition was originally conceived and curated by Jeremy Deller working with the Southbank Centre in London, although elements were designed in collaboration with Manchester Art Gallery. As such the self-assessment experience was slightly different for the Manchester Art Gallery team, with one assessor commenting that he was 'by and large, assessing someone else's work!'

Self-assessors at the gallery were slightly more positive about the exhibition than peers or members of the public, as shown in Figure 22. Unlike some of the performing arts events described earlier there was very little difference in the views of peers and regular visitors, although peers were less likely to feel that the exhibition was different to things they'd experienced before, giving an average post-event score of 0.48 for 'distinctiveness' compared to 0.65 given by members of the public.

**Figure 22: Average self, peer and public scores for *Jeremy Deller: All That Is Solid Melts Into Air* (awarded after the event)**




*Self n=5*
*Peer n=3*
*Public n=133*

**Manchester Art Gallery says:**

We collaborated with Hayward Touring to be the launch venue for Jeremy Deller's exhibition about the north and the continuing impact of Industrial Revolution on art, popular culture and social behaviours. We did this as we believed this would present an international quality artist, grappling with a subject that is at the heart of Manchester's identity as a city and a people. This felt like it met our mission to present audacious, challenging, popular work to the widest range of people.

We were hoping that people would recognise that Jeremy is an internationally significant artist (fresh from representing Britain at the Venice Biennale) – so one of the very best artists in the world. At the same time we hoped that we would have a very strong recognition of the relevance of his exhibition to our local history and to individuals in the north today. This was born out by the responses. We were hoping for enjoyment and critical engagement from our audience – we hoped for very high visitor numbers and for a long dwell time in the exhibition, with lots of questions being asked of our gallery attendants. We got this. From peers, we hoped for recognition that this is an international quality show, for the quality of the argument of the exhibition, the sense of good fit for our artistic vision for the Manchester Art Gallery and appreciation of the range of northern art and museum objects we had drawn on. This too was born out by the results.

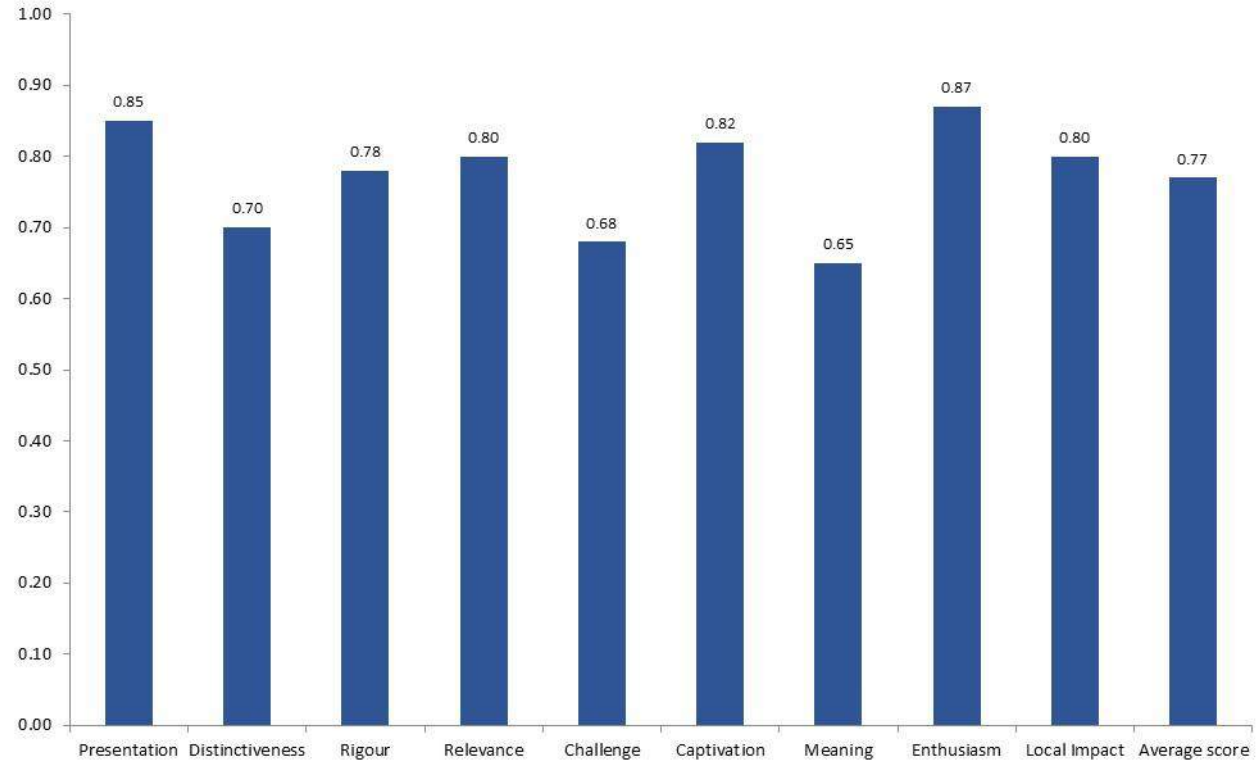
The results made sense to us. The slightly lower scores around technical accomplishment make sense in that Jeremy is a conceptual artist whose intention is not the quality of the objects per se, but rather their relevance. It follows that we were delighted that the exhibition scored highest of any in the study in the 'relevance' dimension. We were particularly pleased to see that visitor and peer views were close – confirming our belief that a wide range of visitors are very happy to engage with conceptual art, if they feel it has wit, relevance and is visually stimulating. It was also heartening to see that peers were persuaded of the international importance of the exhibition after visiting. We are in a period of change at the gallery, which for a long while was not delivering an international quality show. So, the lower expectations are understandable and we are pleased the actual experience changed peers' minds.

## 3.8 Vivarium – Manchester Museum

A total of 66 visitors to the Vivarium exhibition at Manchester Museum were surveyed on 3, 4, 5 and 11 December. The survey sample was young: the median age was 29 and around a third of those who recorded their age were under 25. There were slightly more men (55 per cent) than women (45 per cent).

Figure 23 shows that the visitor response to Vivarium was positive and similar to the perceptions of the Jeremy Deller exhibition, although Vivarium received a higher average score for 'captivation' (0.82 compared to 0.73 for Jeremy Deller) and a lower average score for 'challenge' (0.68 compared to 0.77).

**Figure 23: Average public scores for Vivarium**

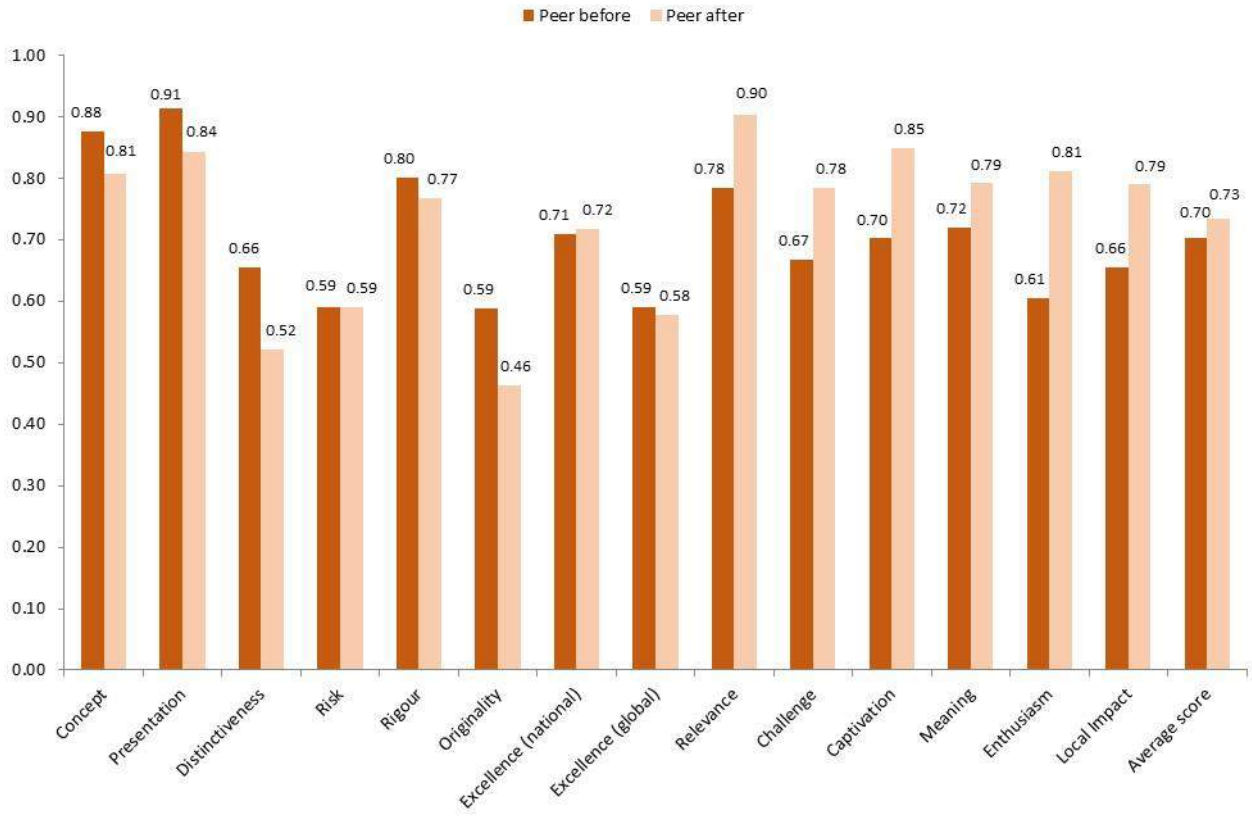



*n=66*

Peer ratings for Vivarium were also high, particularly for 'relevance' (a post-event average score of 0.90), 'captivation' (0.85) and 'presentation' (0.84). Peers found their experience to

be more intellectually and emotionally engaging than they had expected, giving higher post-event scores for ‘relevance’, ‘challenge’, ‘captivation’, ‘meaning’ and ‘enthusiasm’, but did not find the exhibition to be as distinctive or ground-breaking as they thought it might be. It should be noted here that none of the peers appointed by the Arts Council were available to review this event.

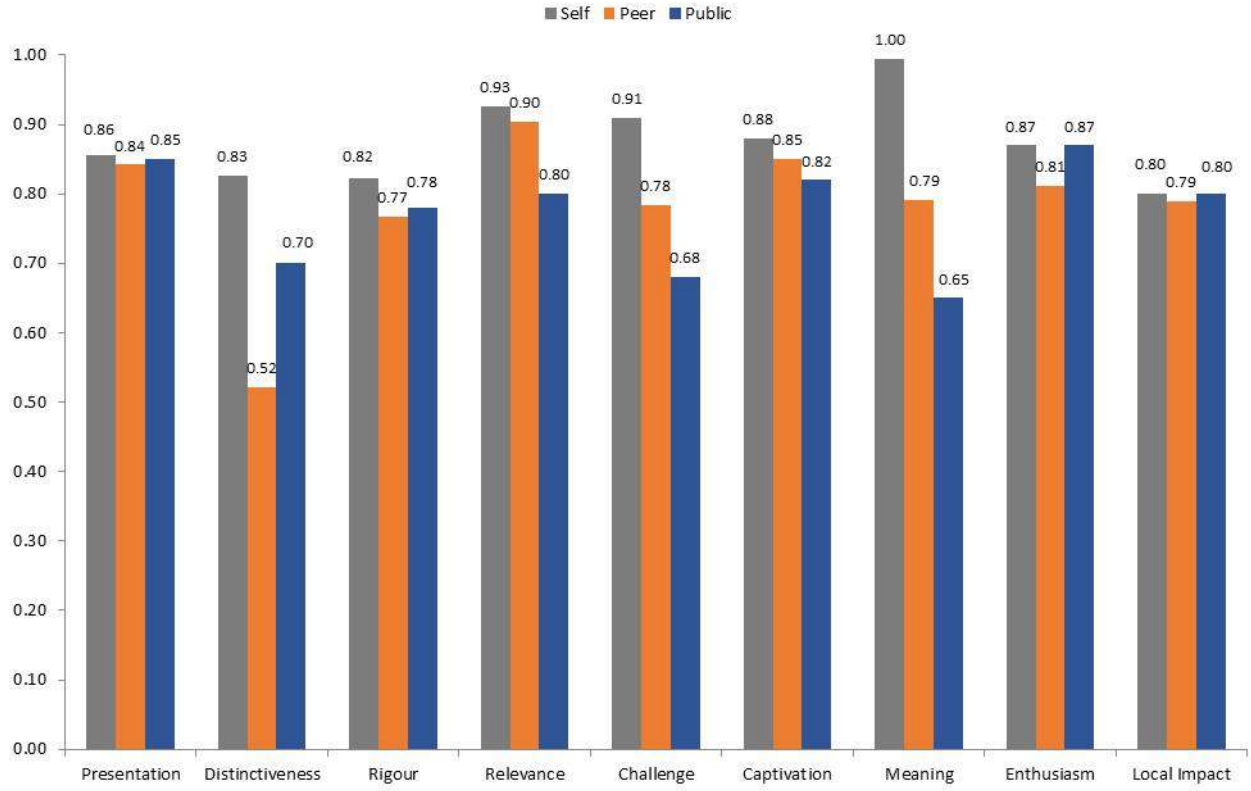
**Figure 24: Average ‘before’ and ‘after’ peer scores for Vivarium**




*Peer before n=5*
*Peer after n=5*

The comparison of self, peer and public scores in Figure 25 shows no major discrepancies, although unusually members of the public were least positive in response to the more personal and subjective measures of ‘relevance’, ‘challenge’ and ‘meaning’. As with the Jeremy Deller exhibition, peers gave a much lower average score for Vivarium’s ‘distinctiveness’ (0.52 compared with 0.70 from exhibition visitors and 0.83 from the Manchester Museum self-assessors).

**Figure 25: Average self, peer and public scores for Vivarium (awarded after the event)**




*Self n=3*
*Peer n=5*
*Public n=66*

**Manchester Museum says:**

With the Vivarium exhibition we were trying to renew the permanent display, which was about 15 years old and had received a lot of wear and tear, in a way that made it fresher and brighter and opened up the back of house captive breeding activities to public view, giving the exhibition a much more explicit conservation message.

Beforehand we would have expected high scores in relation to ‘relevance’, ‘meaning’, ‘challenge’ and ’presentation’. We were also hoping for good scores for ‘originality’ and ‘excellence (national)’ from peers.

We were pleased with the high scores in terms of ‘presentation’ and ‘relevance’ from the public, but slightly disappointed we didn't achieve particularly high scores for ‘meaning’ and ‘challenge’. The former is probably because people may find it difficult to connect with the plight of animals in Costa Rica and elsewhere, and the latter may be because people are slightly saturated with environmental messages. Our highest score was for ’enthusiasm’, which is very pleasing.

In terms of peers, we were pleased with the overall high scores. It was interesting that for half of the measures scores after viewing were slightly lower than prior expectations, while for the other half they were higher. We were particularly pleased at the change in scores for ‘relevance’ from 0.78 to 0.90 and from 0.61 to 0.81 in terms of ’enthusiasm’. We were slightly disappointed about the scores for ‘distinctiveness’ and ‘originality’, as almost no other museums have a similar facility.

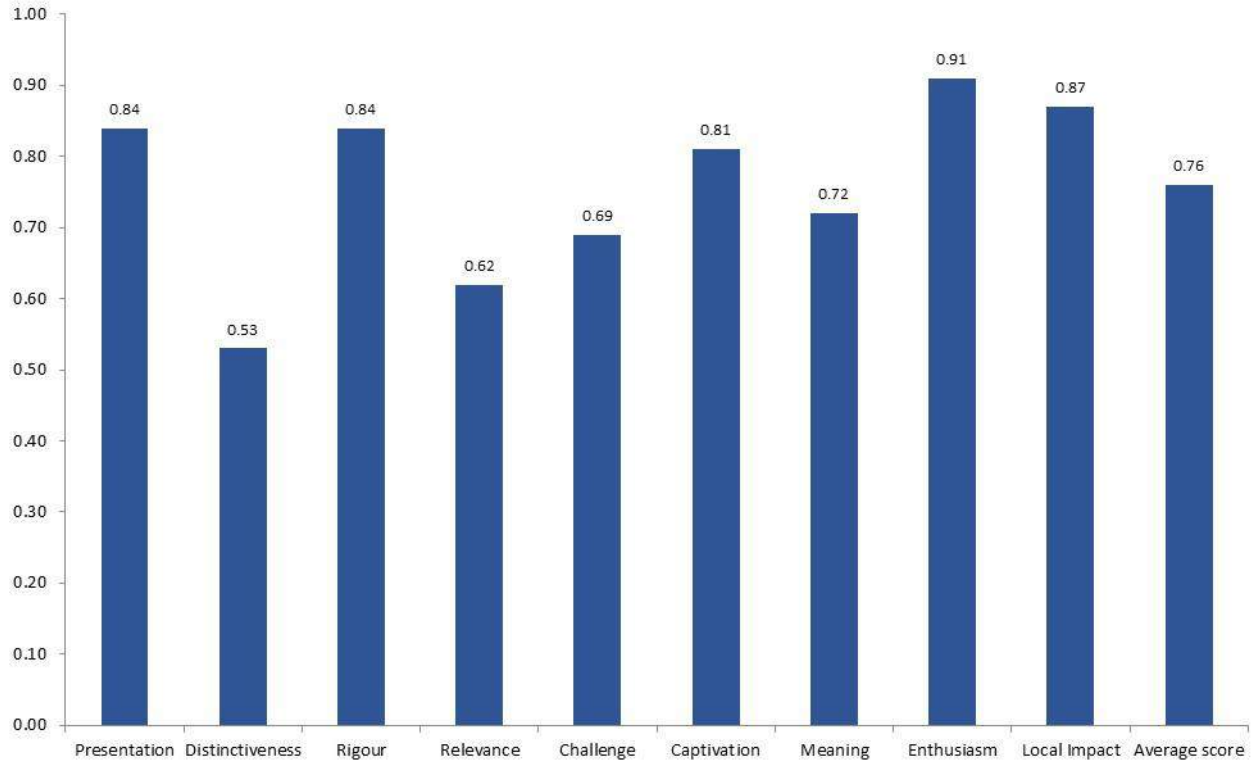
The most insightful aspect of the data was the congruence between self, peer and public on six of the nine measures, which suggests that we are approaching 'truth' on those issues. It was interesting that peers felt that the exhibition was significantly less distinctive than self and public, probably reflecting their wider experience.

## 3.9 *The Radev Collection: Bloomsbury and Beyond* – Abbot Hall Art Gallery

Interviewing was carried out at Abbot Hall Art Gallery in Kendal, Cumbria, on a continuous basis from 10 to 19 December 2013. In total 114 responses were received from visitors to *The Radev Collection: Bloomsbury and Beyond* exhibition. The sample was relatively old, with a median age of 58 and over 70 per cent of respondents over the age of 50, and fairly equally split between men and women.

Figure 26 shows that the public response to *The Radev Collection: Bloomsbury and Beyond* was slightly different to the previous two exhibitions. Visitors were more impressed by the exhibition's 'presentation' and 'rigour', awarding an average score of 0.84 for both, but didn't feel that it offered anything particularly different to things they'd seen before, giving an average score of 0.53 for 'distinctiveness' which was the lowest score received for this dimension across all eight events.

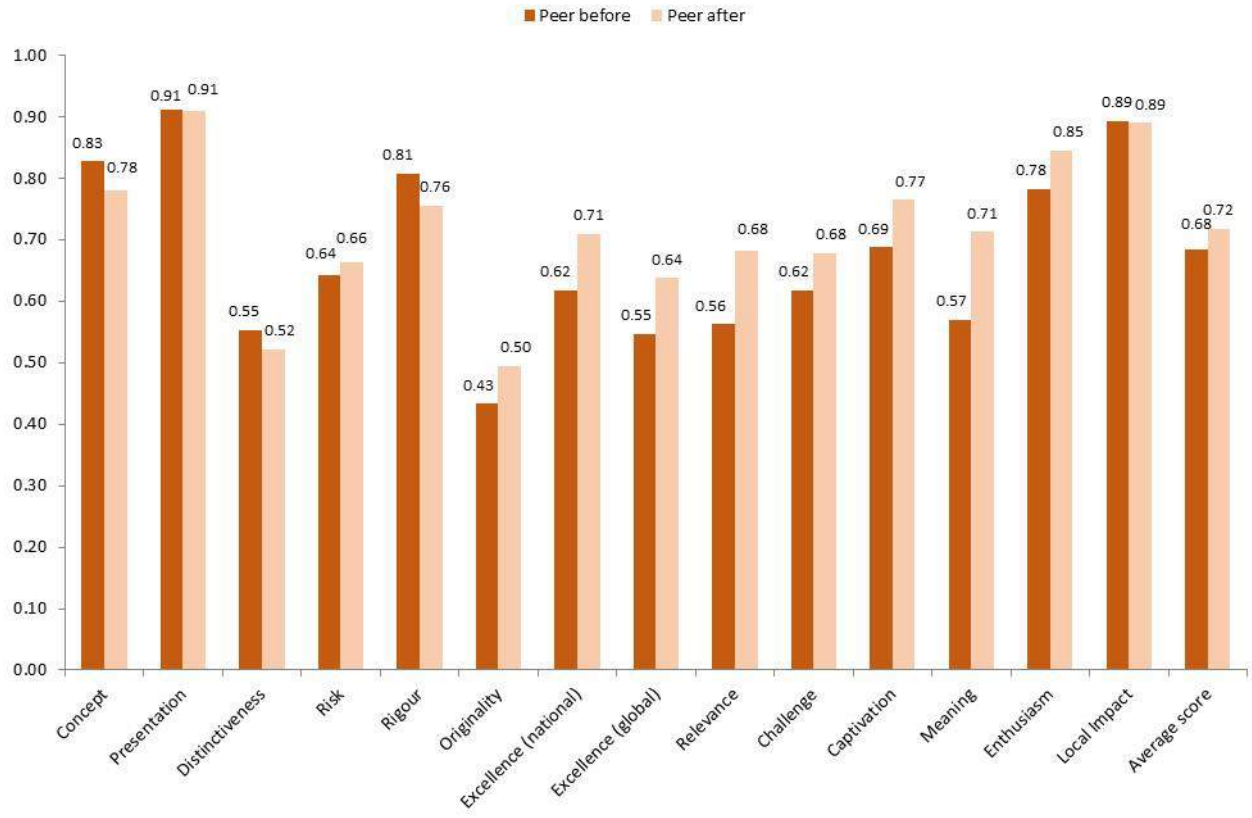
**Figure 26: Average public scores for *The Radev Collection: Bloomsbury and Beyond***




*n=114*

Peer scores were high and *The Radev Collection: Bloomsbury and Beyond* largely met or exceeded their expectations, as shown in Figure 27. The exhibition received particularly good scores from peers for the two measures of excellence and received the highest peer score in the pilot (0.64) for 'excellence (global)'.

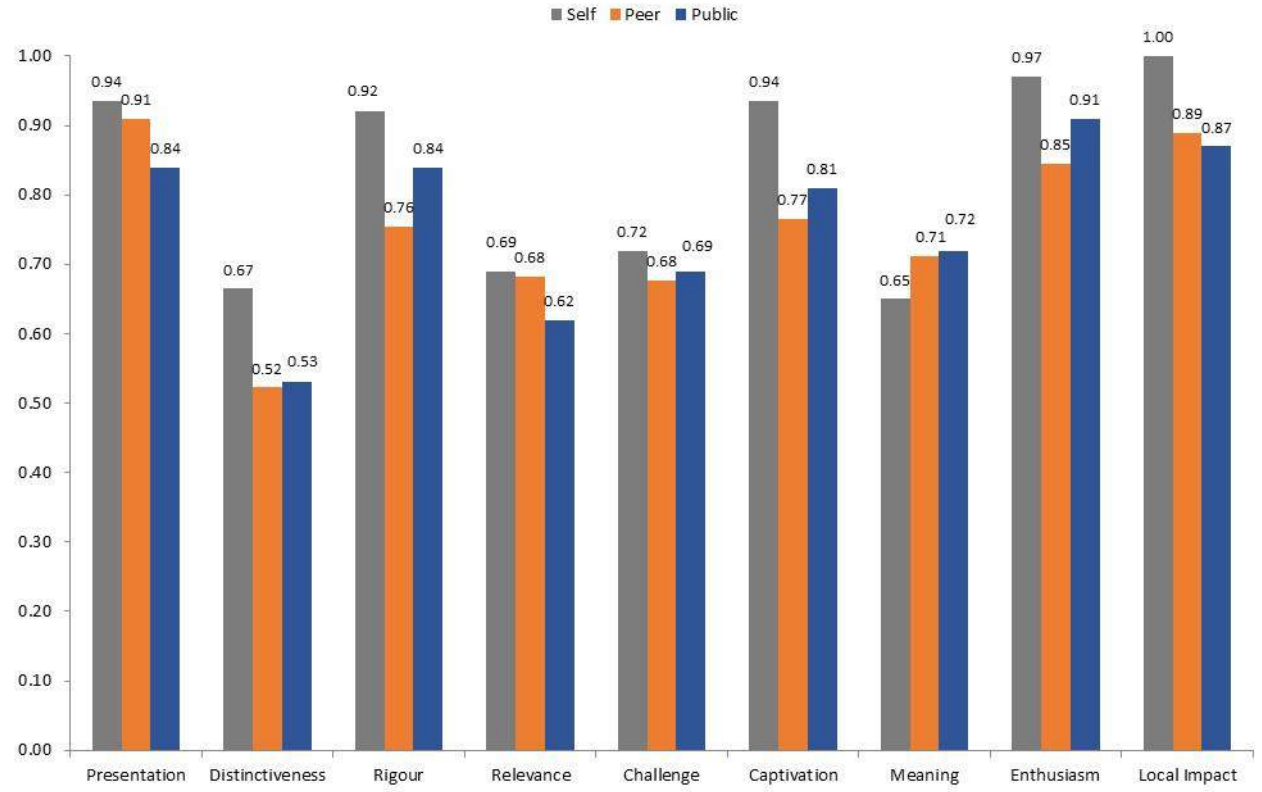
**Figure 27: Average 'before' and 'after' peer scores for *The Radev Collection: Bloomsbury and Beyond***




*Peer before n=4*
*Peer after n=4*

Figure 28 shows that perceptions of the Radev Collection exhibition among peers and visitors were very similar for most dimensions. The two self-assessors at Abbot Hall were generally slightly more positive than the other groups and were particularly proud of the exhibition's importance to Kendal, both giving a maximum score of 1.0 for 'local impact'.

**Figure 28: Average self, peer and public scores for *The Radev Collection: Bloomsbury and Beyond* (awarded after the event)**




*Self n=2*
*Peer n=4*
*Public n=114*

**Lakeland Arts Trust says:**

*The Radev Collection: Bloomsbury and Beyond* exhibition was an excellent opportunity for Abbot Hall Art Gallery to build upon its reputation for exhibiting private collections which in many cases have rarely been shown before in public. The collection featured a great many nationally and internationally recognised artists including Pablo Picasso, Ben Nicholson, Graham Sutherland and George Braque, delivering Lakeland Arts' aim of exhibiting 'art of the highest quality, bringing the work of artists with established national and international reputations to Cumbria.'

We expected the exhibition to appeal to our core audiences and 'Friends of the Trust' and it was specifically programmed at a time (autumn to winter) when our core audiences and Friends are some of our most regular attendees. The result that the audience interviewed 'was relatively old, with a median age of 58 and over 70 per cent of respondents over the age of 50' was therefore in line with our target audiences. We didn't expect the exhibition to break new ground or uncover new research so were not surprised that visitors 'didn't feel that (the exhibition) offered anything particularly different to things they'd seen before'. We wanted our visitors to enjoy the works on show, appreciate the quality of the works and understand the influence of the Bloomsbury Group of artists on future artists, all of which the exhibition achieved; we expected to meet our visitor targets, which we exceeded, and to present a good quality exhibition, which we did. This was reinforced as the exhibition

‘received particularly good scores from peers for the two measures of excellence and received the highest peer score in the pilot (0.64) for ‘excellence (global)’. These results reinforce Abbot Hall’s reputation for excellence.

# 4. Exploring dimensions of quality

In the previous chapter we presented the scores awarded by self, peer and public assessors for all the different quality dimensions for each of the eight events in the Manchester pilot in turn. In this chapter we explore the individual quality metrics in more detail. We take each metric in turn and compare the scores received across all eight events, focusing on peer and public responses for the nine metrics that were assessed by all respondents and peer responses for the five additional metrics that were included in the self and peer surveys only.

In this pilot project we are not so much interested in why one event scored higher than another for any given metric – it's hard to imagine many real-world scenarios where it would make sense to directly compare the quality of a museum exhibition of live animals with the quality of a pantomime or a performance of opera by Verdi. Rather we make comparisons across the eight events to understand more about the meaning and usefulness of the questions being asked. For some dimensions scores were fairly consistent across events while for others there was a lot of variation, and in this chapter we highlight these differences and consider possible explanations. We also draw on feedback from interviewers, peers and the participating cultural organisations to reflect on how individual questions might have been interpreted by respondents, whether questions felt more appropriate in some contexts than others and which dimensions seem to be generating the most insightful data on the quality of cultural events.

## 4.1 Excellence (national and global)

We start by comparing the scores awarded by peers for the two quality dimensions relating to excellence by national and international standards. These dimensions were included for self and peer assessment only – they were not intended to capture an individual's personal response to an experience but rather to gather the views of experts as to how the exhibition or performance compared to similar work taking place around the UK and overseas and to wider developments in the artistic or cultural form.

Figure 29 and Figure 30 show that for most events in the pilot peers awarded reasonably high scores for 'excellence (national)' and, as expected, a lower score for the more challenging criterion of 'excellence (global)'. The two national music critics who acted as peers for the Hallé Orchestra's Verdi bicentenary concert awarded the highest post-event score of 0.88 for 'excellence (national)'; it is perhaps not surprising that Contact Theatre's *CYAC: Advent Avenue* – a new piece of theatre created in just eight weeks by non-professional young performers – received the lowest peer score of 0.25 for this dimension.

A number of peers commented that they had found the question about excellence in an international context difficult to answer, either because it felt inappropriate for the event they were reviewing (particularly the Christmas family shows) or because they didn't feel that they had enough knowledge or experience of overseas work to give an informed response. In fact both of the national critics who provided a post-event assessment for the Verdi bicentenary concert left this question blank.

> ‘I answered the question about how the piece I saw fits in the world arts scene without any real knowledge at all of what’s out there in the rest of the world, which leaves me wondering how useful this question is.’

It may be that ‘excellence (global)’ is not an appropriate dimension for inclusion in any core metric set but rather an additional measure to be used by those cultural events that genuinely aspire to international significance and assessed only by peers with knowledge and experience of global developments in the relevant artistic or cultural form. It would be interesting to know how much work of this nature takes place in the UK every year – and how large the pool is of cultural professionals who feel capable of this kind of international benchmarking.

**Figure 29: Average peer scores for 'Excellence (national): it is amongst the best of its type in the UK'**

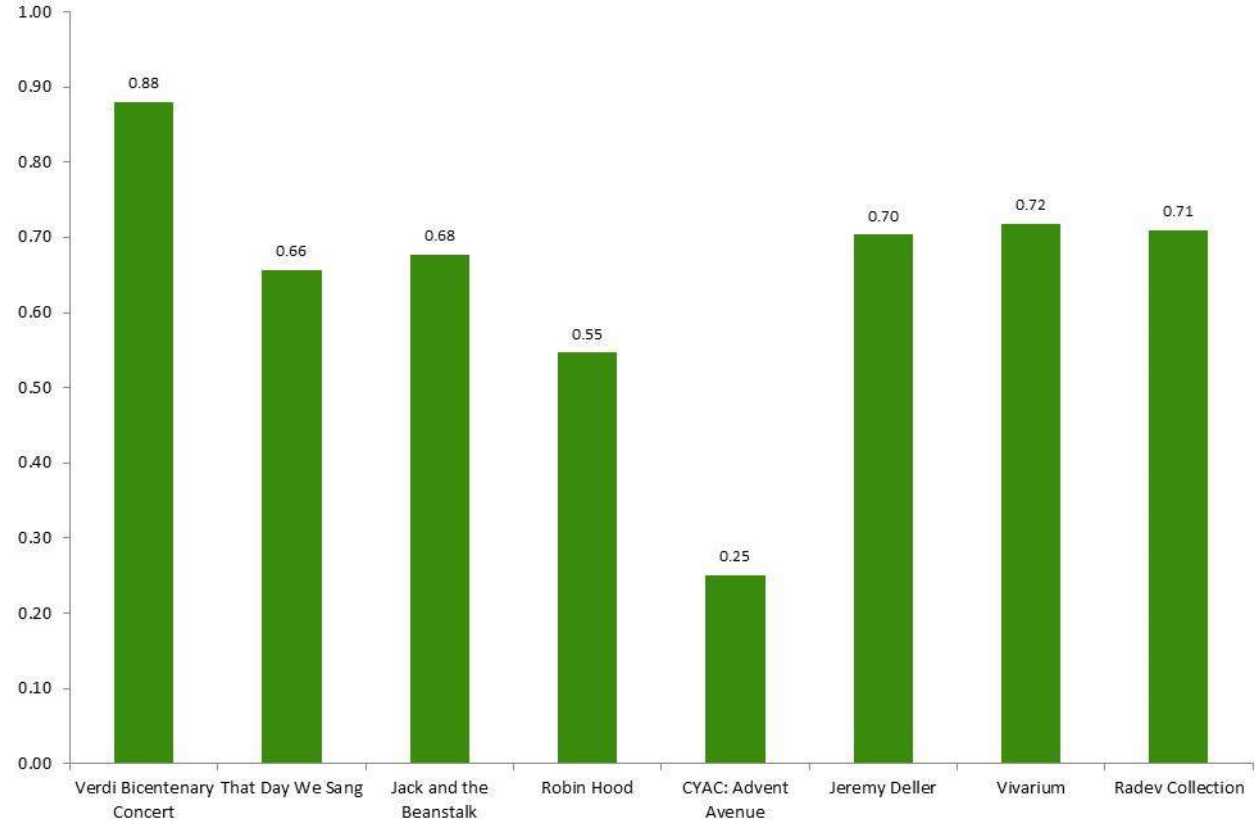
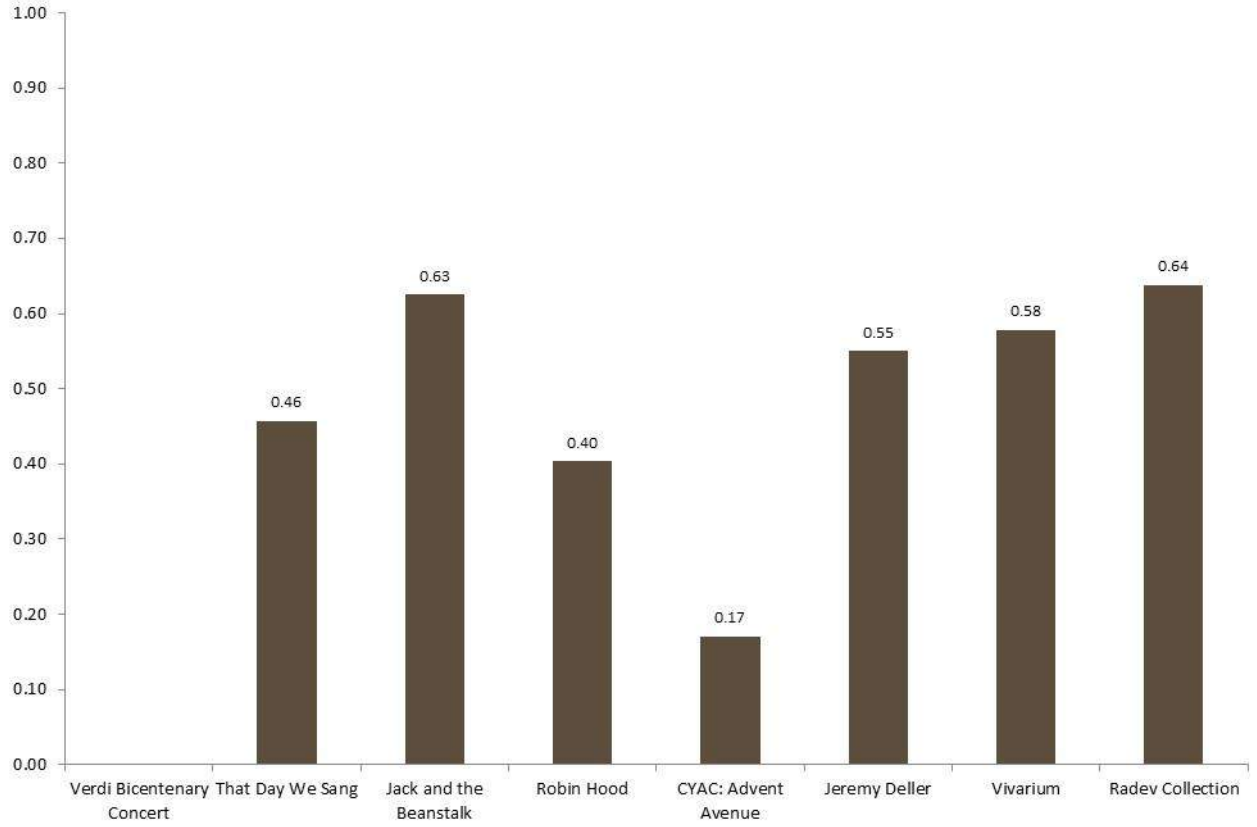


**Figure 30: Average peer scores for 'Excellence (global): it is amongst the best of its type in the world'**




## 4.2 Presentation, rigour and concept

The pilot tested three dimensions that were intended to capture views on the artistic or cultural idea behind a work ('concept'), how well that idea was developed ('rigour') and its actual execution ('presentation'). At the metric development stage if was decided that 'presentation' and 'rigour' would be rated by all respondents while 'concept' would be rated by self and peer assessors only.

Figure 31 and Figure 32 show that there was not a great deal of variation in public responses to the metrics 'presentation' and 'rigour'. Scores were consistently high across events, although the performing arts events generally received slightly higher scores than the museum and gallery exhibitions. Scores may have been high across the board because all the events in the pilot were produced to a high standard by well-respected cultural institutions, and we may have seen more diversity in responses if we had evaluated a wider range of events by less established organisations outside major city centres. However, it may also be that only more knowledgeable audience members and visitors are prepared to give an event a low rating for its 'presentation' or 'rigour' – does it take a degree of confidence or experience to express a view that something is poorly planned or presented?

Peer responses for these dimensions do show more variation and it is worth noting again the particularly big disparity in peer and public perceptions of *CYAC: Advent Avenue*, where peers were much more critical of the show's 'presentation' and 'rigour' than its young audience. For most events peers awarded fairly similar scores for 'presentation', 'rigour' and 'concept', with a couple of interesting exceptions. Peers felt that *Jack and the Beanstalk* wasn't a particularly interesting idea, but that it was well put together and presented, giving

average post-event scores of just 0.54 for ‘concept’ but 0.71 for ‘rigour’ and 0.76 for ‘presentation’. The Radev Collection received fairly average peer scores for its underlying idea and the rigour with which the idea was developed, but it was seen to be particularly well-presented and received the highest peer score in the pilot for this dimension.

Overall we think that ‘presentation’, ‘rigour’ and ‘concept’ are useful as metrics for self and peer assessment, although the definitions may need tightening to ensure that the distinction between them is clear. In any future development of this work we think that it would be worth trialling ‘concept’ as a metric for public as well as self and peer assessment. However, we suggest that in general it will not be necessary to include both ‘presentation’ and ‘rigour’ for public assessment, not because regular audience members and visitors are not capable of understanding the difference between these dimensions but because doing so requires a bit of time and reflection which is not always possible in the busy public environments in which surveying typically takes place.

**Figure 31: Average peer and public scores for 'Presentation: it was well produced and presented’**

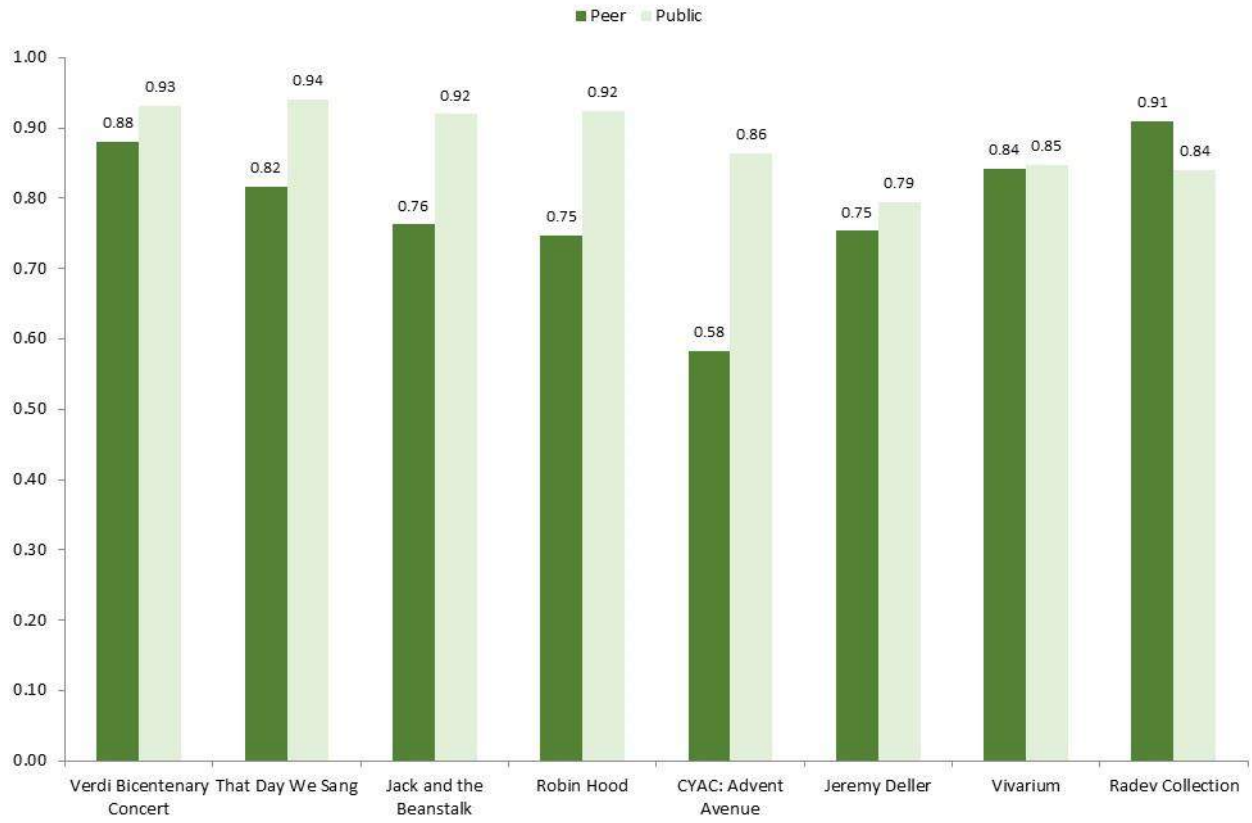


**Figure 32: Average peer and public scores for 'Rigour: it was well thought through and put together'**

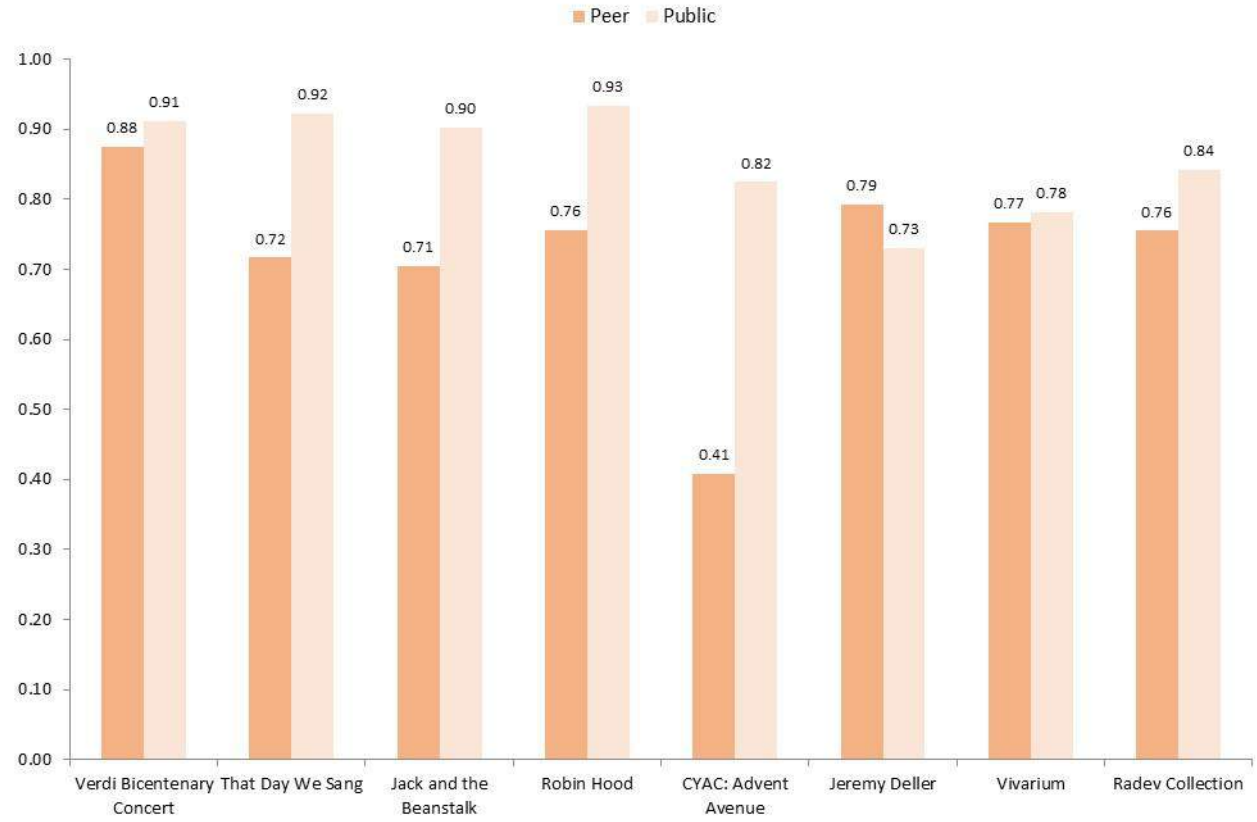
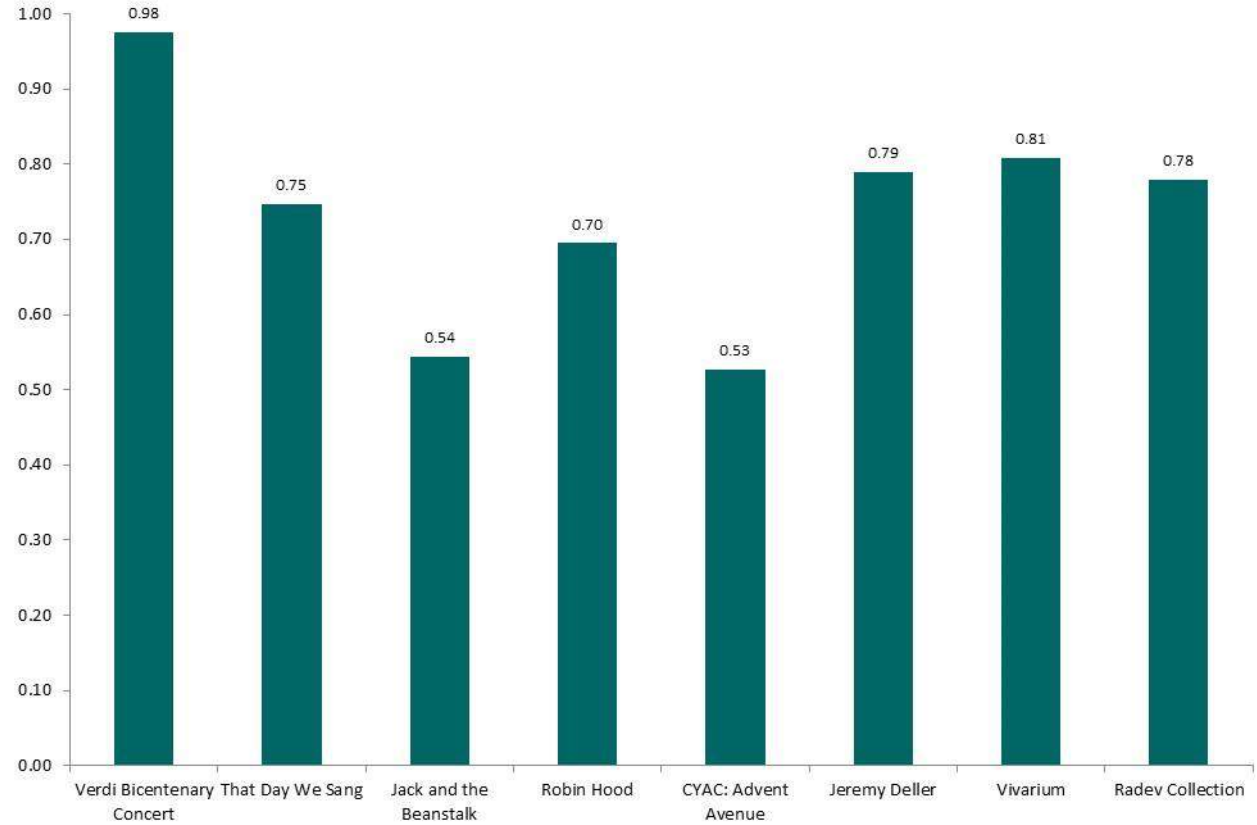



**Figure 33: Average peer scores for 'Concept: it was an interesting idea/programme'**



## 4.3 Captivation and enthusiasm

As with ‘presentation’ and ‘rigour’, public scores for the dimensions ‘captivation’ and ‘enthusiasm’ were consistently high across all the events in the pilot, as shown in Figure 34 and Figure 35. There was more variation in peer responses across the eight events, but for each event the peer scores for ‘captivation’ and ‘enthusiasm’ were fairly similar, and so it is interesting to consider whether the two dimensions are capturing anything different.

In some sense ‘captivation’ and ‘enthusiasm’ are both measures of how much respondents enjoyed their experience; at least, it is hard to imagine someone feeling very enthusiastic about attending a similar event in the future if they found that their attention quickly wandered to other things while watching the performance or walking around the exhibition. In this case ‘enthusiasm’ feels like a less useful measure – interviewees noted that ‘I would come to something like this again’ could have been presented as a yes or no statement and that pretty much everyone would have answered yes, and therefore felt that the dimension may not be contributing a great deal of insight.

However, intention to intend again may reflect a general interest in the artistic or cultural form or loyalty to a particular company or venue, regardless of the experience of a particular work, and as such it may well be a valuable measure for cultural organisations, particularly as it is applicable in pretty much any context. Certainly the commentaries included in chapter three indicate that a number of the organisations that participated in this pilot were particularly pleased and proud to have received high public scores for ‘enthusiasm’.

**Figure 34: Average peer and public scores for 'Captivation: it was absorbing and held my attention’**

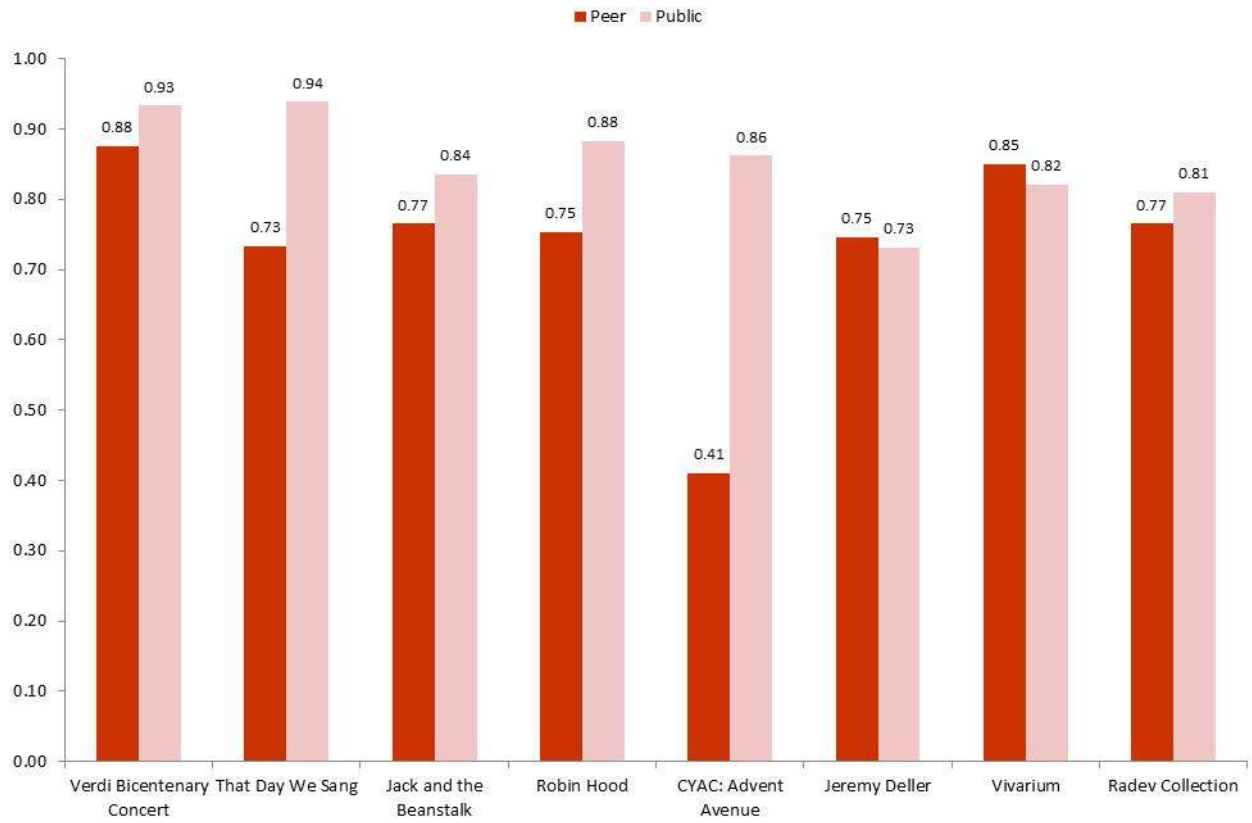
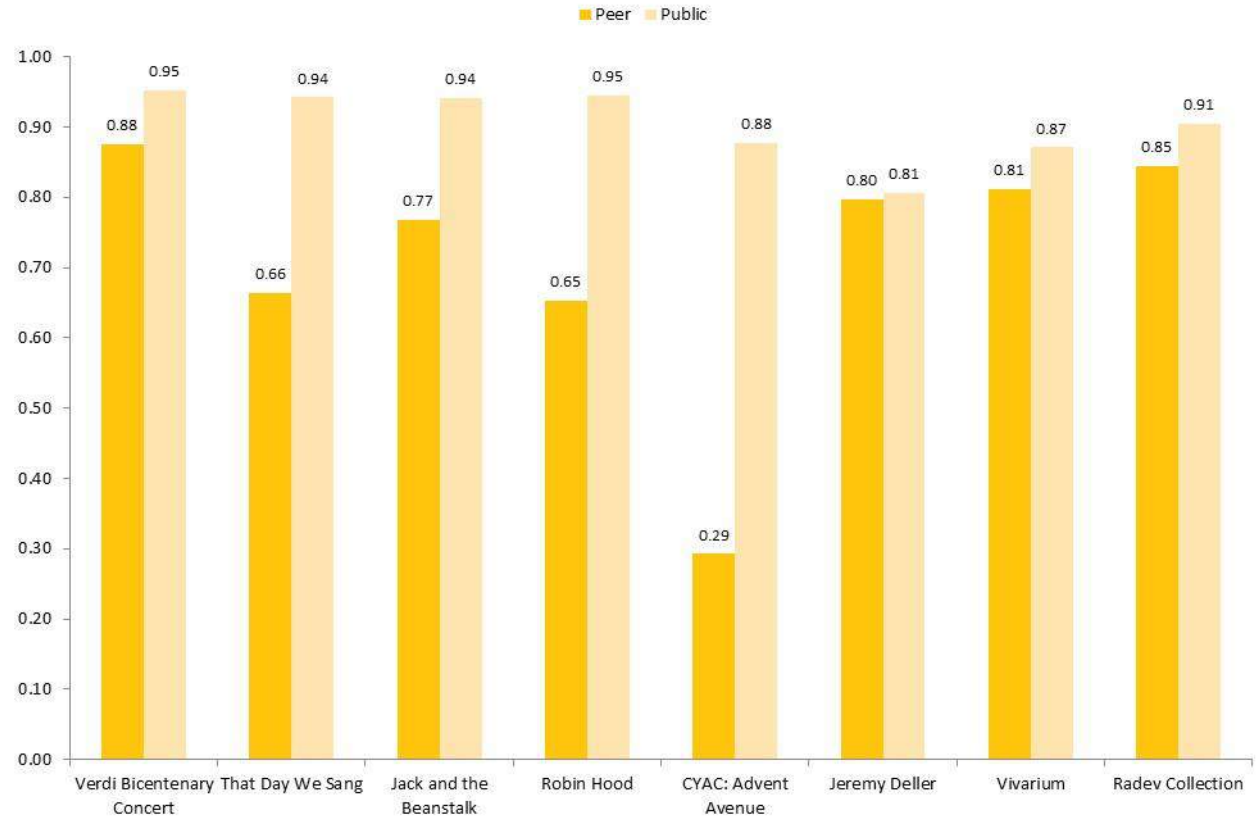


**Figure 35: Average peer and public scores for 'Enthusiasm: I would come to something like this again'**




## 4.4 Distinctiveness, originality and risk

The Manchester Metrics group were keen to include measures of the extent to which a work pushed boundaries, either for the audience, for the artists or curators responsible for the work or for the broader artistic or cultural form. Three separate metrics were developed. 'Distinctiveness' asked self, peer and public respondents whether the work was different from things they'd experienced before. 'Originality' and 'risk' were included as self and peer metrics only, with the former asking whether the work broke new ground and the latter whether it represented a challenge or stretch for the artists or curators involved.

Figure 36 shows that there was considerable variation in both peer and public responses across the eight events for the dimension 'distinctiveness'. The public results make intuitive sense: *That Day We Sang*, a play by Victoria Wood that premiered at Manchester International Festival in 2011, received the highest average score for 'distinctiveness' (0.81) while the Radev Collection exhibition may have been a fairly familiar type of gallery experience for its relatively old visitor base and received the lowest average score of 0.53. Generally, peers were much less likely than audience members or visitors to feel that the work they had assessed was different to things they'd experienced before, with the exception of the Verdi bicentenary concert, which received a surprisingly high score of 0.9 for 'distinctiveness' and the Radev Collection, where peer and public views were very similar. Of all the dimensions tested in the pilot, 'distinctiveness' is perhaps the most dependent on the degree of previous cultural experience of the person responding. It will be much more challenging for a cultural organisation to achieve a high score for 'distinctiveness' for a work that targets loyal subscribers and frequent visitors and that

appeals to an older, more educated market of regular cultural attenders , which may have been the case for the Radev exhibition at Abbot Hall Art Gallery in rural Cumbria.

**Figure 36: Average peer and public scores for 'Distinctiveness: it was different from things I've experienced before'**

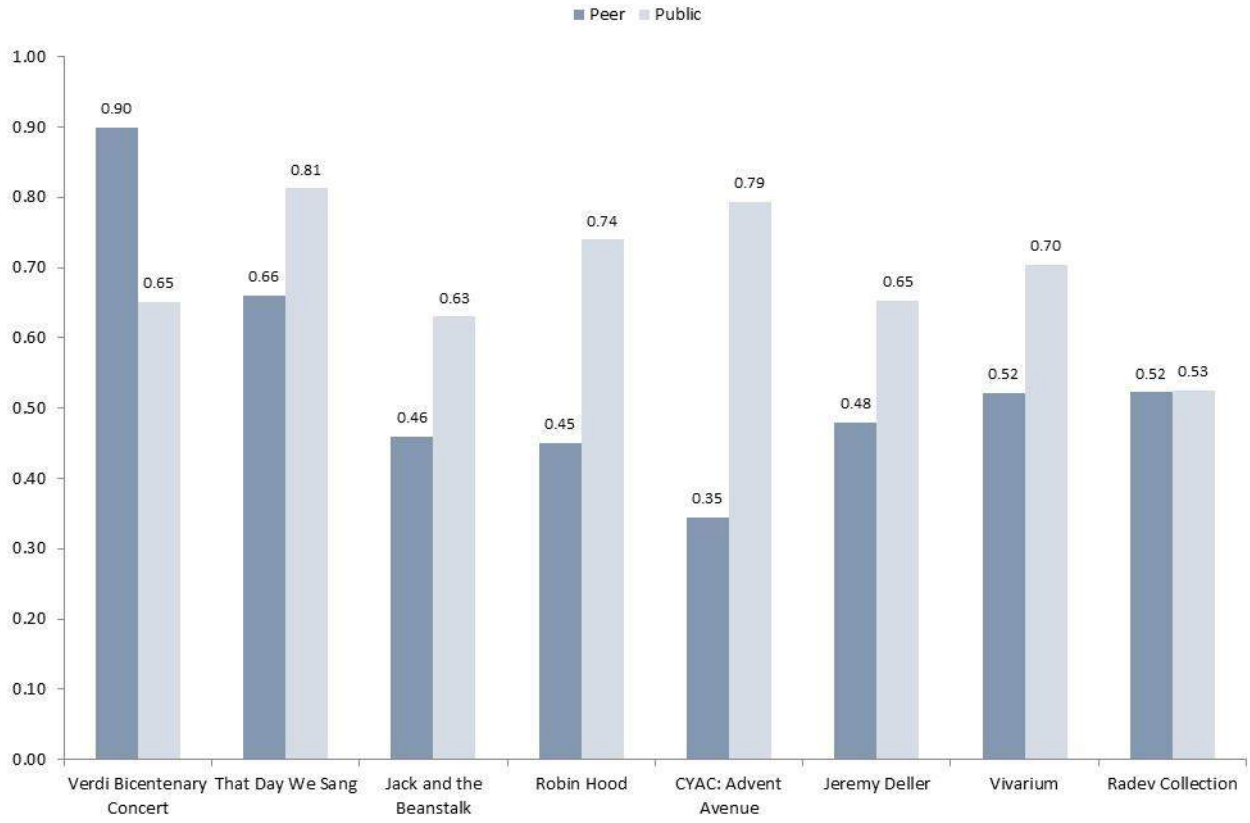
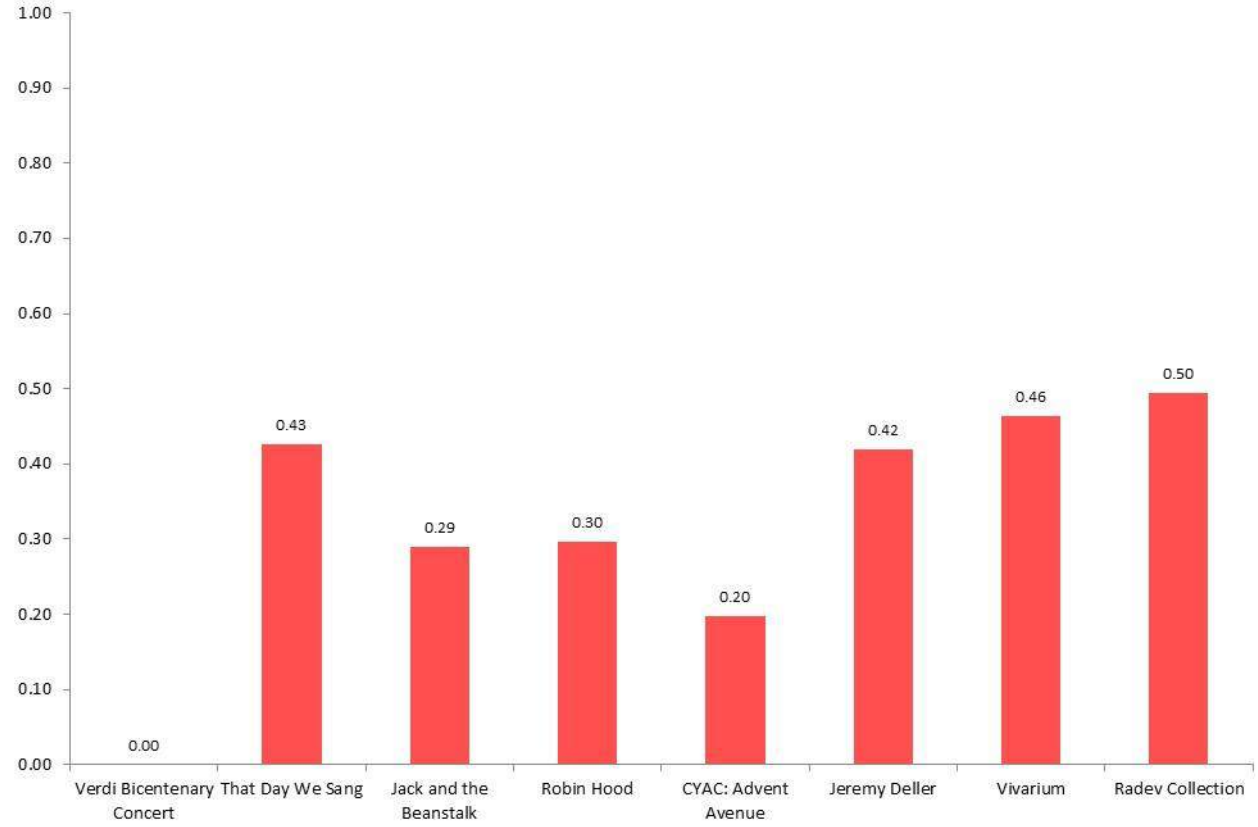


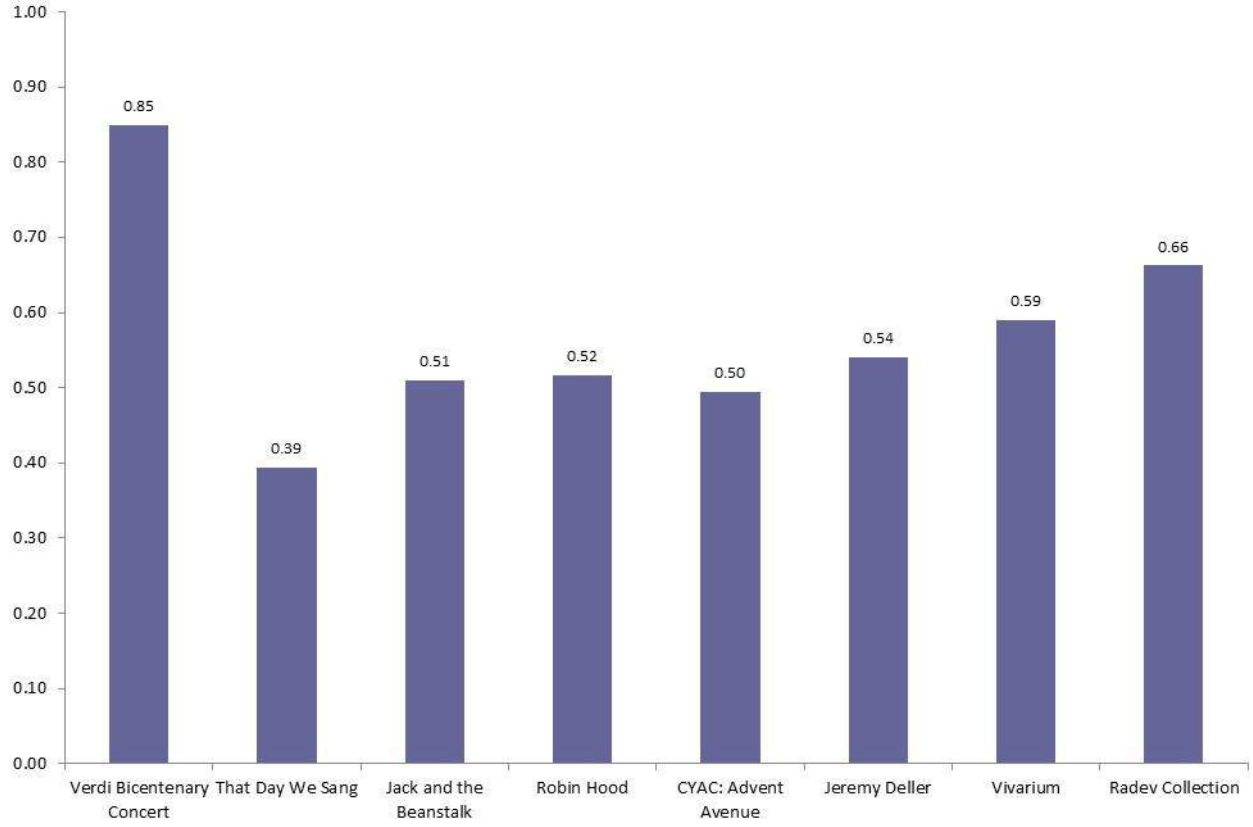

Figure 37 shows that peer scores for ‘originality’ were fairly low across the board. In fact five of the eight events received their lowest peer score for ‘originality’, suggesting that to genuinely break new ground is one of the toughest demands for cultural organisations striving to produce work of the highest quality. It is interesting to note that the Verdi bicentenary concert received a peer score of zero for ‘originality’, despite receiving high scores for both ‘distinctiveness’ and ‘risk’; this may be because while the music itself is 150 years old, it was programmed in a way that made it feel different to other Verdi concerts and challenging for the Hallé to perform it to such a high standard.

**Figure 37: Average peer scores for 'Originality: it was ground-breaking'**




The metric ‘risk’ gives peers an opportunity to assess the degree of stretch involved for the particular artists and curators behind the work. As shown in Figure 38, *That Day We Sang* received the lowest average peer score for ‘risk’, despite being seen as relatively distinctive, and no less ground-breaking than other events in the pilot. This may be because peers had particularly high expectations as to what the Royal Exchange was capable of; conversely, peers recognised to some extent that *CYAC: Advent Avenue* was a challenge for its young performers, even if they didn’t rate the work itself for its ‘distinctiveness’ or ‘originality’.

**Figure 38: Average peer scores for 'Risk: the artists/curators really challenged themselves with this work'**

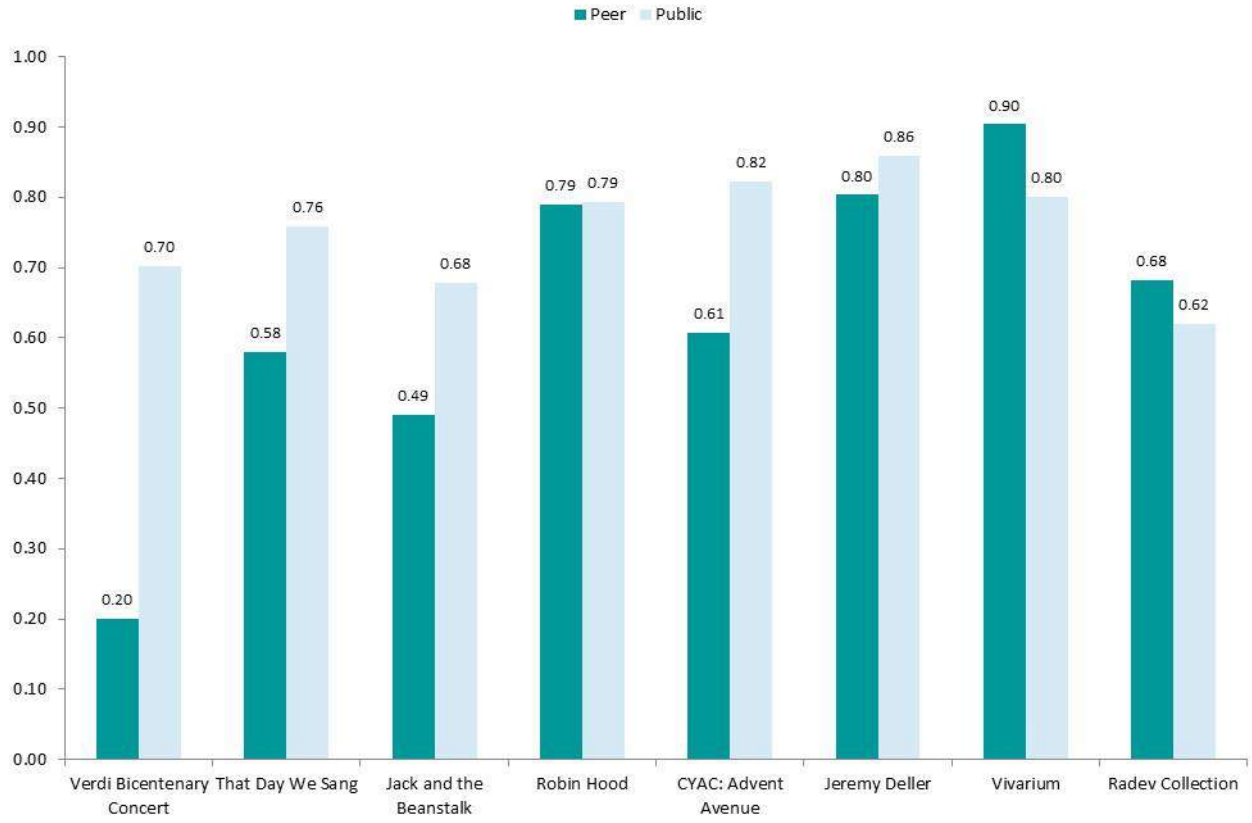



## 4.5 Relevance, challenge and meaning

The dimensions ‘relevance’, ‘challenge’ and ‘meaning’ were developed to give an indication of the extent to which a work connected with its audience and engaged people emotionally and intellectually. These measures were designed to gauge an individual’s personal response to a work and are arguably more subjective than more technical dimensions such as ‘presentation’ and ‘rigour’.

Figure 39 shows that there was quite a lot of variation in both peer and public responses for ‘relevance’. Interviewers felt that members of the public found it difficult to answer this question on some occasions, particularly parents who had brought their children to see a Christmas show. ‘Relevance’ may not always be appropriate in these contexts although it is interesting to note that *Robin Hood*, a family show which explored political and social issues such as the bedroom tax, received the fourth highest score for this dimension. Both peer and public responses for ‘relevance’ were high for the exhibitions at Manchester Art Gallery and Manchester Museum, and as noted previously this metric gives an opportunity for events like the Jeremy Deller exhibition to shine; the exhibition explored the impact of the Industrial Revolution on British society and visitors gave it a higher rating for what it said about the world today than for any other quality dimension.

**Figure 39: Average peer and public scores for 'Relevance: it had something to say about the world in which we live'**




Peer and public responses were also quite mixed for the metric ‘challenge’, as shown in Figure 40. The highest public rating for this dimension was for *CYAC: Advent Avenue*, although interviewers felt that the statement ‘it was thought-provoking’ might have been harder to understand for young audiences. For most other dimensions, peers tended to be more critical than audience members and visitors, but peers were more appreciative than the public of the degree of ‘challenge’ presented by *Robin Hood* and the Jeremy Deller and Vivarium exhibitions. It may be that these events were more thought-provoking for those with a degree of expertise and professional interest in the issues being explored.

Both peers and members of the public commented that it felt odd rating a pantomime for its level of ‘challenge’ and it is perhaps not surprising that *Jack and the Beanstalk* received the lowest peer score of 0.15 and the lowest public score of 0.49 for this dimension. As one interviewer who surveyed audience members at *Jack and the Beanstalk* noted:

> ‘One lady thought the questions were inappropriate for panto. I explained they were benchmark questions for all events, but she still thought it was an unfair reflection for a “cracking good pantomime”.’

**Figure 40: Average peer and public scores for 'Challenge: it was thought provoking'**

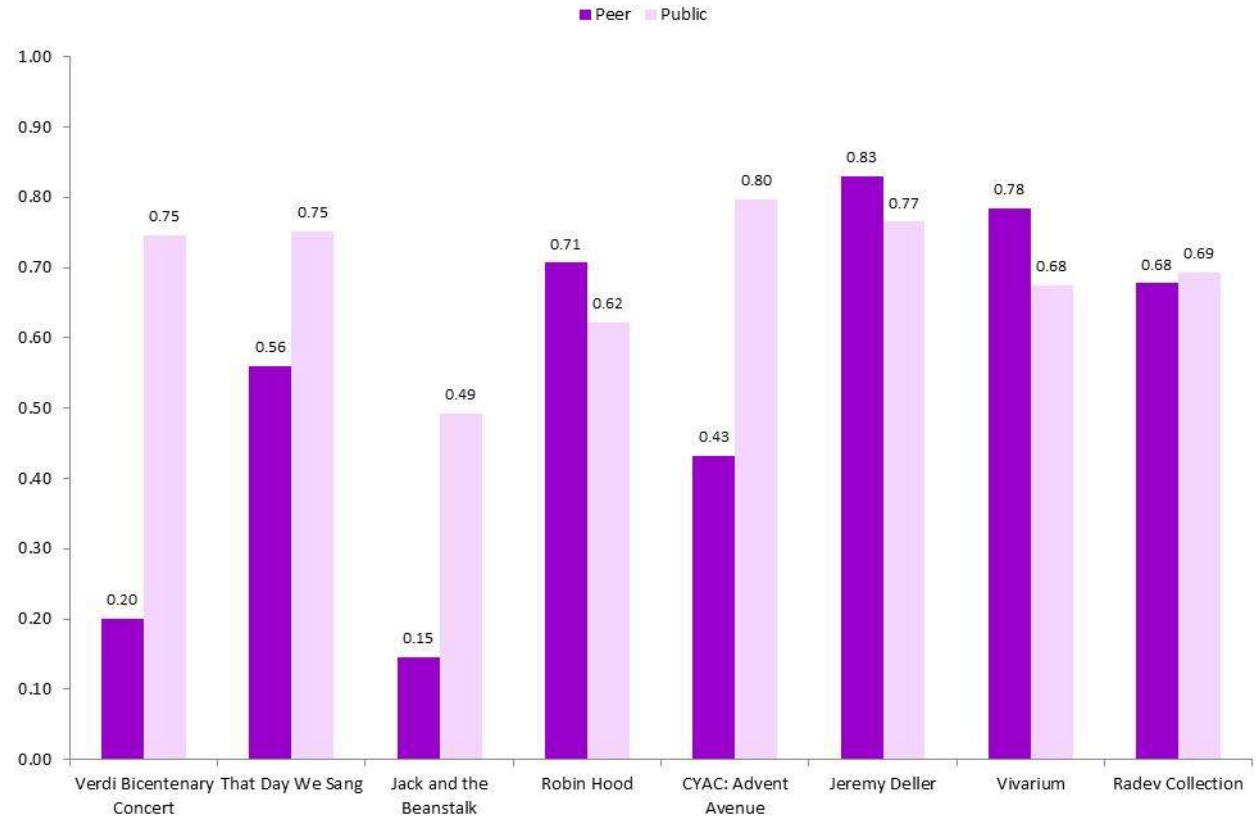
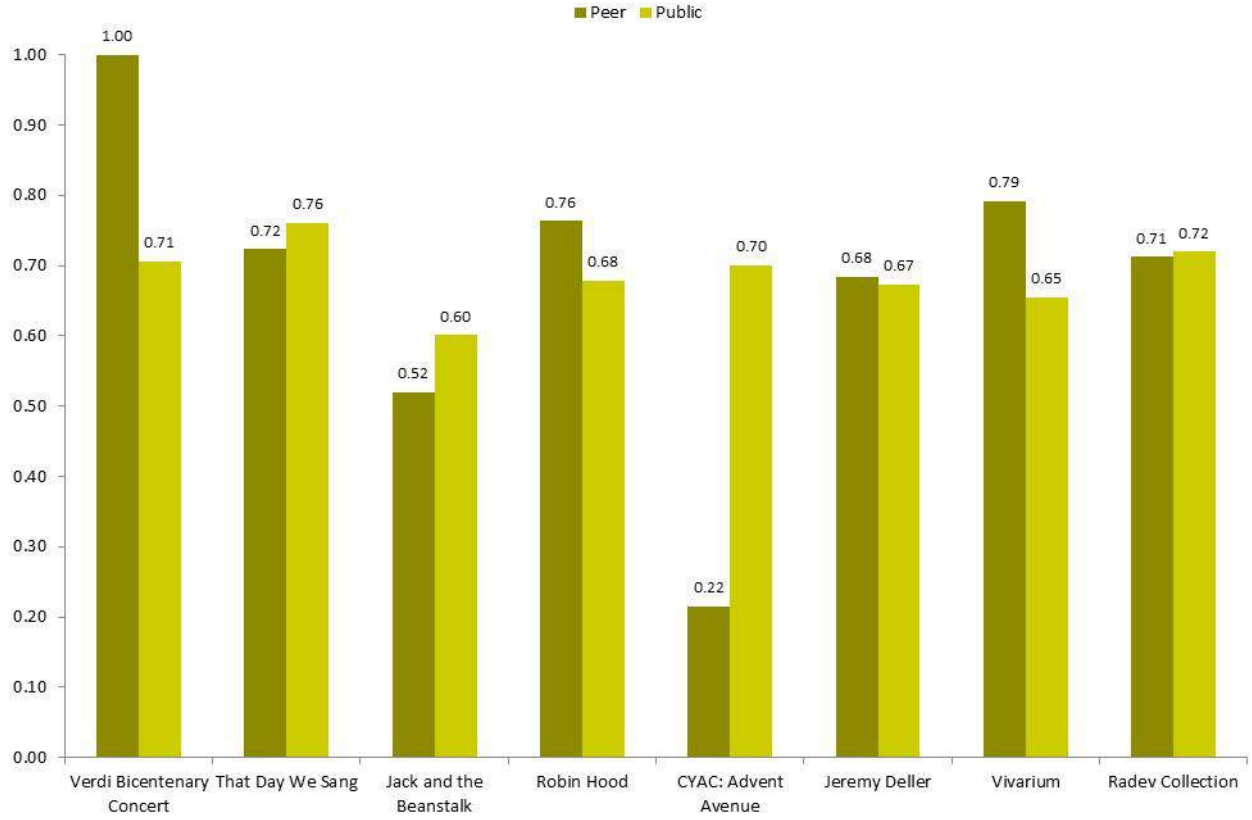



Figure 41 shows that there was less variation in responses for the metric ‘meaning’ and more correspondence between the views of peers and the public than for most other dimensions. Interviewers questioned whether a person’s personal relationship to a piece, which could be explained by a wide range of factors, was really a valid measure of artistic quality. We agree that ‘meaning’ as currently defined is probably the weakest of the more subjective measures and if the metric set is to be developed further we would recommend finding a more concrete way of capturing the extent to which a work resonates with its audience.

**Figure 41: Average peer and public scores for 'Meaning: it meant something to me personally'**



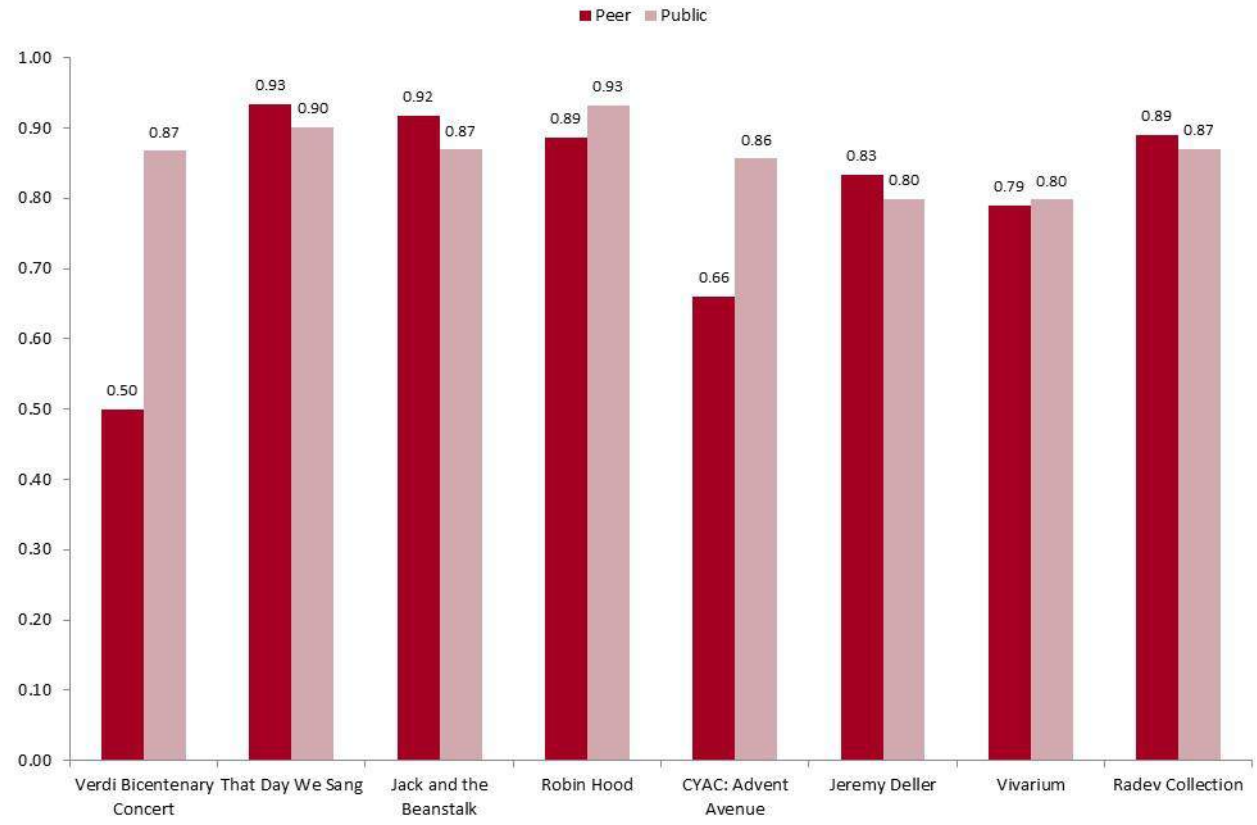

Overall there was clearly more variation in the public scores for the more personal, subjective measures of 'relevance', 'challenge' and 'meaning' than for the more technical dimensions of 'presentation' and 'rigour', and there are a number of possible explanations for this. It may be because there was a lot of variation in the nature and content of the experiences being evaluated – we would not expect the audience at *That Day We Sang*, say, to have the same sort of emotional or intellectual experience as visitors to the Vivarium exhibition, although both events might have been produced to an equivalent high standard. It may also be because people can experience the same work quite differently depending on their background and personality and level of cultural capital – what's relevant and thought-provoking for one person might be little more than good entertainment for another, even if they both agree that the work has been well-produced and performed. And the different levels of variation may be partially explained by the nature of the survey process in that audience members and visitors may give a more considered response to statements about 'relevance', 'challenge' and 'meaning', which relate directly to their own personal experience, and are more assertive about giving a negative rating when one of these measures doesn't quite capture how they felt about an event. It may be that dimensions such as 'presentation', 'rigour' and 'captivation' will always score highly if people feel that in general terms they've had a good day or night out.

It would be useful to carry out further research to explore these issues in more detail but at this stage we observe that, for many cultural organisations, dimensions that ask the public about their personal response to a work are likely to be more challenging measures of quality than dimensions about the production and presentation of the work itself. It's one thing to offer a polished, absorbing, highly enjoyable cultural experience; it's another to make a difference to how people think and feel about the world.

## 4.6 Local impact

To conclude this chapter we look at scores for 'local impact', a dimension that sought to capture the extent to which an event was seen to be of significance and value to the local area. Figure 42 shows that both peer and public responses for this dimension were high for most events, with the highest public rating received by *Robin Hood* at Octagon Bolton. Interviewers felt that this question was problematic at events where many of the people attending weren't local to the venue: they tended to feel that the question wasn't relevant to them and didn't have much of an opinion. This may explain the relatively low peer score for 'local impact' of 0.5 for the Verdi bicentenary concert – the peer assessors for this event were national critics who travelled from outside Manchester to review the performance. We suggest that some measure around local impact will be important for many cultural organisations, especially those outside major city centres that are developing work in response to particular local circumstances, but that the metric definition could be improved to better capture the essence of the contribution that organisations are seeking to make locally.

**Figure 42: Average peer and public scores for 'Local impact: it is important that it's happening here'**



# 5. Process reflections and recommendations

The Manchester Metrics pilot has been designed to help us learn as much as possible about the process of measuring quality, over and above the value of collecting data at the designated events. In this chapter we draw on our own experiences delivering the project and feedback from the participating cultural organisations, peer assessors and interviewers to reflect on the process and note what worked well, what was difficult and what could be improved in the future. We make a number of recommendations both to inform the development of the Culture Counts system and to offer prompts to other researchers involved in collecting data on experiences and perceptions of arts and cultural events.

## 5.1 Improving and expanding the quality metrics

The most significant challenge of this project was presented by the very first task: to develop and agree a set of metrics for assessing the quality of very different arts and cultural events. To achieve this, members of the Manchester Metrics group had to balance their own views on the definitions and data that would be most meaningful and useful to them with the needs of the wider group and the practical requirements of good survey design. Overall, we think that the co-production process worked well: the group approached the challenge with creativity, curiosity and generosity and the process sparked a number of discussions about what constitutes quality in different contexts that were valuable in their own right. The outcomes and metrics developed in Manchester were very similar to those produced in Western Australia, particularly in the areas of 'excellence', 'originality', 'risk', 'rigour' and 'relevance' and we feel confident that any other group of cultural professionals going through a similar process would not end up with a set of metrics that were wildly different to those developed here.

Having said that, the analysis in chapter four shows that if the cultural sector is interested in developing a common set of metrics for assessing quality then there is clearly scope to refine and build on the metrics tested here. A key consideration will be to identify which metrics are 'core' and can be applied in a meaningful way to a diverse range of arts and cultural experiences. We suggest that any core metric set that can be used to benchmark quality across the sector will inevitably be quite small, perhaps containing no more than five to seven measures for use by self, peer and public assessors and a further three to five measures that are used in self and peer assessment only.

In addition to this core set we recommend developing a number of additional modules that organisations can use to tailor surveys to their particular context or to particular events. For example, in the near term it may be appropriate to develop separate modules for the performing arts and for museums and galleries, by applying the same co-production principles and collaborating with professionals in the relevant cultural fields. In the longer term it may be feasible to develop a small number of artform and sub-artform level standardised metrics. So for example, dance professionals could work together to develop shared quality metrics that relate directly to the experience of a dance performance. This approach would create flexibility for cultural organisations in choosing how they evaluate a particular work and enable effective benchmarking at artform and sub-artform level, while still allowing for a degree of consistency and comparability over time and across companies and sites through the core metric set.

The events involved in this pilot were all traditional venue-based experiences and another future development strand would be to develop modules for assessing the quality of festivals, street arts and other outdoor events, and work that takes place in a community context. The Culture Counts method may be particularly useful in these settings because it can capture audience feedback and basic demographic information at events where no box office data is available, and can measure how ratings shift over the course of a day or weekend, revealing how an audience is responding to different elements of a festival programme.

Another useful development for the Culture Counts system would be to include a mechanism for capturing qualitative feedback. While the system is designed primarily as a quantitative tool, and there are many excellent qualitative research techniques for exploring experiences and views already used in the cultural sector, several peers who took part in the project commented that the survey gave them no space to describe their emotional reaction to a work:

> 'The questions/measures were all very "dry" and didn't give an opportunity to rate things within the arts that are important to me – excitement, wonder, joy.'

> 'I was invited to explore how the piece made me think, but not how it made me feel.'

This issue was recognised by the Manchester group during the metric development phase and the group had hoped to include an open-ended question in the self, peer and public surveys along the lines of 'What three words best described how you felt about the work?' Unfortunately it was not possible to programme this additional question into the system within the time available and we recommend that it be tested as part of any further development of this research. An open-ended question of this nature would generate data that could be presented visually (using a tag cloud for example) and add richness to the quantitative results without having too much impact on the speed and simplicity of the survey process. Over time the most common emotional responses could be identified and codified to create a set of standardised options which cultural organisations would be able to tailor depending on the types of response that felt most relevant to a particular work.

As well as wanting more space to voice their personal feelings about a piece, peers also questioned the value of capturing purely quantitative data without asking for any wider interpretive context:

> 'If the questions exist without a narrative, I can't see that the expertise of the peer is being utilised.'

> 'I was expecting there to be more detailed questions about the specifics on artistic quality.'

Again, the purpose of a system such as Culture Counts is not to replace more established approaches to peer review that draw on detailed narrative assessment but rather to offer an additional quantitative component that allows for robust comparison and aggregation of self, peer and public views, over time and across events. However we agree that the full value of quantitative measures of quality can only be realised alongside understanding of what an artist or organisation is trying to achieve with a particular work and how it relates to developments within the wider artistic or cultural form. As such we recommend that a future version of Culture Counts should include space for self and peer assessors to provide

interpretive context; the simplest way to achieve this would be to provide an optional open text field to accompany each metric that enables assessors to explain why they have awarded a particular score for any given quality dimension. An additional text box could be included at the end of the survey to capture any further thoughts or feelings about the work. This would allow for narrative assessment to take place without compromising the structure and integrity of the co-produced metrics framework.

In theory it would be possible to include the same open text fields in the survey for audience members and visitors, effectively turning the public feedback process from a brief questionnaire into more of a structured interview. This would allow for both a quantitative and qualitative exploration of perceptions of quality, but the process would be longer – 20–30 minutes instead of five – and it would be more difficult to achieve a reasonable sample size at the venue itself. As with any market research the choice here is between depth of exploration and breadth of response; cultural organisations wishing to prioritise qualitative feedback for a particular work may want to administer the survey including open text fields to a small sample of audience members after the event via email or telephone, allowing for a more discursive approach to quality assessment while also capturing a small number of data points against the standard metrics.

Our final recommendation on metric development is that further research in this area should include a phase of formal cognitive testing among assessors, particularly members of the public. Feedback from interviewers who administered the public survey during the Manchester pilot suggests that audience members and visitors had few problems understanding the survey questions:

> ‘People got it really well…more than other surveys that I’d done.’

> ‘All the questions were pretty self-explanatory.’

However it would be useful to carry out some follow-up interviews or focus groups with audience members and visitors to explore how they understand and interpret individual metrics, how appropriate the questions feel in different settings and at different types of arts and cultural events and whether there are important aspects of quality that they think are not being fully captured by the survey. This research could also start to explore the fascinating question of how socio-demographic factors such as age, gender, level of education and previous cultural experience affect the ways in which people understand and respond to different metrics. In particular, a strand of testing could be dedicated to young people (say 11 to 18-year-olds) to explore whether there are any differences in how younger audience members and visitors experience the survey and interpret particular questions and whether any additional metrics are required to capture responses to work aimed specifically at a younger demographic. This sort of research would effectively open up the metric co-production process to include audience members and visitors as well as cultural professionals and help ensure that concepts and definitions of quality are relevant to a wide public.

## 5.2 Technical improvements

The pilot revealed that there are a number of ways in which the technical aspects of data collection could be improved. First, using the app to collect feedback from audience members and visitors currently requires a good Wi-Fi or 3G internet connection.

Interviewers had few problems accessing the internet during the Manchester pilot, although the connection was slow at times. However for the system to work reliably in a wide range of settings, including outdoor events and venues in remote locations, it would be helpful to remove the immediate reliance on Wi-Fi or 3G. It should be possible to develop the app so that data is stored locally and uploaded to the server as soon as a working internet connection becomes available.

Another potential technical improvement relates to the way in which scores are recorded. Currently respondents use a sliding scale to indicate how much they agree or disagree with each quality metric. Feedback from interviewers and peers suggests that for the most part people found the slider intuitive to use and that it helped to give the feedback process a dynamic and engaging feel. However interviewers found that there was some variation in the way in which people used the slider. Some respondents moved it quickly and confidently to a point on the scale that felt appropriate, with a tendency to move the slider right to the top of the scale to indicate strong agreement. Others were more cautious and took their time in positioning the slider to give an accurate account of their level of agreement with a particular metric. Currently the survey does not show respondents the actual numeric score being recorded when the slider is moved to the chosen location and displaying a numeric value may enable respondents to give more precise and consistent answers. An alternative would be to ask respondents to select a value from a fixed-point scale from 0 to 10, say, and it would be useful to carry out further research to see whether there are any significant differences in quality ratings awarded by audience members or visitors for a particular event depending on whether a sliding or fixed-point scale is used.

We suggest that it is particularly important for self and peer assessors to be able to record their scores with precision. This is partly because self and peer assessors complete the surveys in their own time and can take as long as they want to determine the exact score they wish to allocate for each dimension. Furthermore, if cultural professionals are to carry out these sorts of assessments on a regular basis for lots of different events then the difference between awarding a score of 0.7 and 0.75, say, will become quite significant. It would be useful to explore the preferences of self and peer assessors for a sliding versus fixed point scale; whichever approach is used, it would be helpful for ‘before’ scores to be visible when assessors complete their ‘after’ survey so that they can provide an accurate account of how their experience of a work compared with their expectations. Self and peer assessors should also be able to access their previously completed surveys so that they can rate a current event relative to scores they have awarded to other events in the past.

## 5.3 A different focus for self-assessment

The Culture Counts system is currently designed to enable self-assessors – typically artists, curators and others working in cultural organisations – to record both ‘before and ‘after’ views on the quality of an event they have been involved in. In the Manchester pilot each participating organisation nominated around five staff members or freelance associates to act as self-assessors, and their scores were very much based on their own perceptions and experiences of the work.

We did not feel that this self-assessment process worked quite as well as it could for a number of reasons:

- self-assessment scores were often higher than those awarded by peers and public and may say more about the degree of confidence, motivation and self-belief required to get an exhibition or production off the ground than they do about the quality of the work itself

- pre and post comparisons of self-assessment scores were not hugely interesting – it seems that the views of artists, curators and other cultural professionals don't tend to change much once a show or exhibition has opened to the public

- self-assessment as it is currently framed is not particularly relevant for venues hosting a touring work that was conceived and developed elsewhere

- prior self-assessment is not possible for permanent museum and gallery exhibitions

The Manchester group suggested that the self-assessment process could be altered so that instead of providing their own personal 'before' and 'after' ratings of a work, cultural professionals complete the survey once only before an evaluation period and record what they *hope* the work will score – what they expect it to achieve given its target audience. For example, a gallery putting on an edgy or provocative exhibition might hope to achieve high public scores for 'distinctiveness', and high peer scores for 'originality' and 'risk', but expect greater variation in public response and hence lower overall scores in relation to dimensions such as 'captivation' and 'enthusiasm'.

In this way the self-assessment process could become a tool to help cultural organisations ask themselves some deeper questions about a work ('Who is this for?' or 'Why did we do it?') and form a useful component of internal planning, reflection and self-evaluation processes. Organisations could use the survey to record their objectives for a single event, or what they hope to achieve in terms of average quality ratings over the course of a season or year. As the Manchester Metrics group pointed out, this kind of self-assessment would to some degree be a reflection of the quality of their cultural leadership – the better an organisation's judgement in tailoring its work to its audience, the narrower the gap will be between the organisation's expected quality scores and peer and public responses.

## 5.4 Who are the right peers?

Generally the peer assessment aspect of the pilot ran smoothly. There were no real problems persuading peers to give up their time to participate in the project – this may have been because of a general willingness to help out their Manchester colleagues, or because they were interested in experimenting with a new kind of peer review process. It probably helped that the assessment process is not particularly onerous (the surveys take no more than five minutes to complete, and the only real time commitment involved is in seeing the work) and informal feedback suggests that peers found the system intuitive and easy to use:

> 'I thought the process, the pre and post show questionnaire worked really well. Simple, quick and user-friendly.'

There is a certain amount of administrative work involved in briefing peers and liaising with them about when to see the work and complete the surveys, and it would be cost-effective to automate as much of this as possible within the online Culture Counts system.

As explained in chapter two, some of the peers who took part in the pilot were recruited by the participating cultural organisations while others were members of the Arts Council's pool of artistic assessors. Overall the peers who were drawn from the Arts Council's central pool tended to be more critical of the events in the pilot than the peers who were nominated by the cultural organisations themselves. This may be because the peers recruited by the Arts Council are experienced assessors who see a large amount of work and over time have developed their own detailed picture of what quality means in different cultural contexts, and as a result have relatively high expectations about what a genuinely 'excellent' work should achieve. However it may also be because the peers who were nominated by an organisation participating in the pilot would have had some sort of existing relationship with that organisation and, consciously or not, may have felt uncomfortable recording a negative view of the work of people they regard as colleagues and friends.

Thus as with any peer review process, cultural organisations wishing to use a system like Culture Counts will need to give careful consideration to who constitutes a 'peer', and whose views they believe to be most valid and valuable in their particular context. One member of the Manchester Metrics group explained that she had selected peers whose 'honesty and critical judgement I hold in high regard', and for many organisations on most occasions this may be the most important criterion. By nominating their own peers, organisations can develop a relationship with a group of respected assessors who provide expert feedback in a consistent way over a long period of time. This approach also enables cultural organisations to administer the peer assessment process themselves, handling all the communication with peers, arranging tickets and dealing with any questions or problems peers might have in using the assessment system.

However, if a group of cultural organisations or a funder wishes to use quality ratings awarded by peers to benchmark performance across organisations then it may be necessary to establish an independent and centrally coordinated process for recruiting and allocating peers. This would no doubt result in a more comparable and therefore fairer assessment of quality. The Manchester Metrics group noted that if this type of approach to measuring quality were to be adopted more widely, there is scope to establish a model like that in the higher education sector, where senior and experienced creative staff employed across the cultural sector would be expected to act as peers within a national peer review system.

## 5.5 Sustainable public data collection

The process of collecting feedback from audience members and visitors was perhaps the most fascinating and also most demanding aspect of the Manchester pilot. At the outset we weren't sure how easy it would be to persuade people enjoying a day or night out to stop and fill in a questionnaire, and we had to give careful consideration to how many interviewers would be needed at each venue and for how long in order to reach our target of 50 public responses per event.

We were very pleased to find that the overall refusal rate was low and interviewers reported that the majority of people they approached were happy to complete the survey, particularly after being told that it would only take a couple of minutes. Obviously this kind of data collection is easier in some settings than others; interviewers felt that the process worked best at museum and gallery exhibitions, where people were less likely to be in a hurry to leave the building. Visitors to the Jeremy Deller exhibition were particularly keen to stop and

talk about the work and a total of 133 responses were achieved at Manchester Art Gallery in around 30 interviewer hours. The process was most pressurised after a late-finishing concert or theatre performance when people were rushing to get home, particularly after the family shows when parents had tired children in tow. Nevertheless interviewers were able to meet the target number of responses at every event except the Verdi bicentenary concert at Bridgewater Hall, where a technical problem meant that the tablet computers were out of action for the first 10 minutes after the concert had finished.

Interviewers felt that using an app on a tablet computer rather than a traditional clipboard and pen made it easier to secure interviews, lending the process a greater degree of 'legitimacy' and creating a less bureaucratic atmosphere and a more intriguing and engaging experience for the respondent (although presumably the novelty will wear off as app-based surveying becomes more mainstream):

> 'The app is really cool. It's easier to convince someone to stop with an app.'

Moreover interviewers felt that many audience members and visitors appreciated being asked for their feedback and were interested in the process and confident in expressing their views about the quality of the event they had just experienced. We sense that the act of giving feedback may add to the value of a cultural experience because it creates an opportunity for people to reflect on and make sense of a performance or exhibition and to think in a structured way about whether they enjoyed it and why. It would be interesting to review the literature in this area and to test the hypothesis more formally as part of any future cognitive testing work (see section 5.1).

At the start of the pilot the Manchester Metrics group had hoped that all the public surveying would be carried out by staff members and volunteers at the participating venues. In the end, this was not possible for most organisations, either because permanent staff were too busy with their day-to-day responsibilities or because volunteers didn't feel they had the right experience or expertise to carry out a market research role. Instead we recruited a central team of causal market researchers and deployed interviewers to different events depending on availability.

In some ways this centralised process worked well as the interviewers developed experience and confidence over the course of the project and were able to provide some interesting insights into how well the survey worked in different settings. The initial two-hour training session was critical in building team spirit as well as developing shared understanding of the pilot, their role as interviewers and the purpose and meaning of individual survey questions. However, it was expensive to recruit experienced market researchers and a significant logistical task to ensure that the right interviewers were sent with enough equipment to the right venues at the right time. If a system like Culture Counts is to be used on a regular basis by a range of different cultural organisations then a more efficient and sustainable approach to public data collection needs to be found. We suggest that there are three main options worth exploring:

1) Resources could be provided to **support cultural organisations to carry out surveying themselves**. As one member of the Manchester Metrics group pointed out, building the capacity and confidence of staff members and volunteers is really 'a matter of training and time'. Useful resources would include online training videos, briefing materials and checklists, FAQs and an online or telephone support desk.

2) The survey could be made available for audience members and visitors to complete on **fixed tablet computers** or on postcards distributed around the venue. This would remove the need for interviewers altogether, which could have a number of effects on the rate and nature of responses. First, it may be that people who have particularly strong (positive or negative) views about their experience are more likely to choose to complete the survey; the presence of an interview approaching audience members and visitors at random helps to reduce this kind of response bias. However the interviewer may also have an influence on the way in which people respond to questions, either by the way in which they present and explain particular questions or because people feel under pressure to respond more quickly or positively than they would do if they completed the survey in private. It will be important to test how responses differ depending on the presence or absence of an interviewer in any further development of this work.

3) Finally, in the long-term there is potential to **make the Culture Counts app publicly available** for anyone to download to a smartphone or tablet computer and use to give feedback at any participating venue or event. Organisations would display a QR code on the programme and around the venue or site for any event they wanted to evaluate, and audience members and visitors would use the app to scan the code and record their views. This presents some exciting possibilities to link the app to social media, and build a digital community of 'lay reviewers' who share experiences and make recommendations, and encourage greater dialogue between artists, cultural organisations, audiences and experts about the meaning and impact of a work. The key unknown here is what would incentivise members of the public to download an app of this nature, and whether its use would ever extend beyond very regular cultural attenders and the loyal friends and subscribers of participating organisations. Again, this is a priority area for further research and development.

## 5.6 Automated reporting and additional analysis

In producing this report we reflected on our approach to analysing and reporting on the data collected during the pilot. One of the main advantages of using an app rather than pen and paper is that the data are available to analyse instantly. By logging onto the Culture Counts administrative interface we were able to access almost real-time feedback on the events being evaluated. However there is currently no automated reporting function within the system – the raw data were provided in Excel files and manipulated manually over a couple of weeks to produce the basic charts included in this report. A major improvement would be to build an automatic report facility so that a chief executive or artistic director could receive a report with headline scores for each dimension, measures of variation and possibly breakdowns by age and gender within 24 hours of a show, exhibition or festival opening (allowing a small window for some basic data checking and cleaning).

The system generates a large amount of data and there are many possibilities for analysis beyond the average scores reported here. For example, some events will polarise opinion more than others, which would be revealed by a thorough analysis of measures of spread in the data such as standard deviations, minimum and maximum scores and quartiles. These sorts of measures are particularly useful when the number of respondents is low; for example, when an event is reviewed by only three or four peers, then particularly strong positive or negative views can be obscured if only the mean scores are reported.

Data collected from members of the public can be analysed to detect differences in response to particular quality dimensions by men and women, and by different age groups. If an organisation were to use the system repeatedly over a number of years, it would be possible to carry out some very interesting analysis of shifts in perceptions of quality over time and to identify significant differences in the ratings received by different types of work.

Perhaps the most exciting analytical possibility is afforded by the collection of postcode data. Not only does this provide some basic information about who attended an event or activity (which is useful in the absence of box office data), it can provide valuable context about audience location and profile in which to position and interpret the quality ratings. For example, a work may not achieve particularly high peer scores for its technical accomplishments, but a cultural organisation may consider it a great success if it were received well by a young audience from communities that don't typically engage with the venue or with the arts more generally. With a large enough sample size – which could be achieved by integrating quality ratings with box office data, or by aggregating data over a number of organisations and events – it would be possible to carry out very detailed analysis of how different types of people respond to different types of work in different places, which would greatly enrich our understanding of the nature and impact of cultural engagement.

## 5.7 Exploring other outcomes

We conclude this section by reflecting that the Manchester pilot focused solely on assessing the quality of a cultural product and how it is experienced by self, peer and public audiences. Of course this is only one way of framing 'quality' in the cultural sector, and quality itself is only one of a number of outcomes that cultural organisations may consider important. This was made particularly evident by the decision to include *CYAC: Advent Avenue* as one of the eight events to be evaluated in the pilot. As pointed out in the commentary by Contact in chapter three, *CYAC: Advent Avenue* was a participatory piece with the primary aim of engaging a group of young, non-professional performers in creating and producing their own work through a process that developed their confidence and skills. Culture Counts was able to show Contact how the finished work was received by audience members and peers, but an additional piece of evaluation would be required to understand how the quality and impact of the participatory creative process was experienced by the young people involved.

As discussed in chapter one, in a preliminary phase of this work the Manchester Metrics group identified a range of outcomes that they felt were important to evaluate. They defined outcomes in relation to quality of creative process and collaboration, as well as quality of product and audience experience; they also identified a number of outcomes relating to reach and organisational health and sustainability. A glance at the research base in this field shows that beyond quality, reach and organisational health, cultural organisations are keen to evaluate any number of wider social and economic impacts of their work. Thus the metric set developed here is by no means a complete account of the various components of cultural value, or a narrow replacement set of KPIs against which the performance of all cultural organisations can be judged; nor is the Culture Counts system a replacement for other evaluative tools, from in-depth peer review to audience focus groups to measures such as Social Return on Investment. Rather Culture Counts is an additional tool available to cultural organisations wishing to provide themselves and others with large-scale standardised data on the quality of their work over the course of a year, and how their own

assessment of quality corresponds with assessments of audiences and informed peers. After our positive experience working on this pilot we feel excited by the possibility of developing new approaches to measuring other aspects of cultural value that abide by the same principles of co-production, commonality of method and comparability of results across different types of work, organisation and setting.

## 6. Conclusions and next steps

This project had the following aims:

- for the participating cultural organisations to work together to agree on a set of outcomes and standardised metric statements to measure the quality of their work and the quality of experience for audiences

- to test the metrics using the Culture Counts system – examining the strengths and weaknesses of this approach, and to identify areas for future development and improvement

- to explore whether audiences would respond favourably to the metrics and the Culture Counts system, and to gain insights into how they can best be engaged in providing clear feedback on their experiences

- to ascertain whether the data being generated across the dimensions is useful and insightful for the participating cultural organisations. Did they recognise their pieces of work in the data responses from peers and public, and do they remain enthusiastic about the further use and development of the metrics and the Culture Counts system?

Overall, the project has successfully met these aims. Whilst the metrics need further refinement, there are very encouraging signs that they are capturing self, peer and public reactions to the quality of the work being produced, and key aspects of audience response in a rigorous way. Audiences have been positive about taking part, and we sense if designed correctly the art of giving feedback may add to the value of a cultural experience, enabling people to reflect on and make sense of a performance or exhibition and to think in a structured way about whether they enjoyed it and why.

The participating cultural organisations have found the data, and discussions about the data, to be useful and they expressed strong support for the overall approach and for the value of the metrics and system as a way of measuring outcomes (via peer and public response) against their creative intentions and expectations for a piece of work (predicting peer and public scores prior to the event). As they note in their creative intention and reflection statements in chapter three, they are keen to test a broader range of work against the metrics to further explore which metrics are proving a sensitive measurement mechanism, and which may need further refinement. They also want to revisit their outcomes and metric statements in the areas of reach and organisational health (including quality of cultural leadership) to see how these might be embraced through the peer response mechanisms within Culture Counts.

As described in the previous chapter, we have already identified a range of improvements and refinements to the metrics and the Culture Counts system that we now plan to make and test in the months ahead. The Manchester Metrics group will continue their involvement in these testing activities, and the aim is to extend the number and range of cultural organisations using the system over the next 12 months.

## 6.1 Looking ahead

Whilst this report is the first public presentation of the emerging metrics and a detailed set of test results, this process began in Western Australia some four years ago. Since then there has been strong and growing interest in this work from cultural organisations, funders, audiences, and researchers interested in the measurement of cultural value.

Cultural organisations have quickly understood the benefit to them of *directly* shaping metrics capturing the quality and reach of cultural activities. Without their direct interventions, they doubt that 'quality' metrics can have high credibility and relevance to the arts and cultural sector. Moreover, they are excited about being able to connect more directly with key communities (peers, the public and specific communities of interest), and use technology platforms to create a rich dialogue around the quality of what they do, and to generate large-scale data relevant to their creative intentions and practices.

The work thus far has also proved that the cultural sector is capable of generating a clear consensus on outcomes and standardised metric dimensions to capture the quality of their work. This could not be assumed as a given at the start of the process, but it is an outcome which has opened up the possibility of creating standardised, high quality, large-scale data on quality and other key aspects of cultural value. This is of huge significance to cultural organisations and funders alike who have a common cause in being able to demonstrate and explain the quality of the work being created, and the depth and significance of audience response and experience.

Further refinement and testing of the metrics and Culture Counts system will quickly provide evidence of sufficient scale and sophistication for cultural organisations, funders, and the research and public policy communities to judge whether the emerging metrics are a powerful dashboard for capturing these vital outcomes, and genuinely offer new and more exciting ways of reporting on cultural value creation.

At this stage we think that this approach has great potential. Placing the cultural sector in the lead, and providing them with technology and delivery platforms that allow them to actively shape metrics that best capture their practices and creative intentions, has opened up the possibility of a measurement approach that:

- allows cost effective generation of large-scale data sets on what the cultural sector believes are the key dimensions of 'quality'
- provides rich insights about the dynamics of cultural experiences, and about how to stage and mediate real-time conversations about cultural value
- creates a rigorous, ubiquitous feedback mechanism that will give the public much greater opportunity to express their views on the quality of their cultural experiences
- allows both the public and the cultural sector to tell a richer story about the wider public value of arts and cultural activity in the UK (recognising that the quality of cultural products and experiences is only one aspect of the full range of value that cultural organisations create and may wish to measure)
- provides cultural organisations with immediate feedback on their work as well as long-term comparisons, enabling them to embed data in both strategic planning and day-to-day decision-making

- allows for the integration of quality measures with a wide range of other 'instrumental' data from the cultural organisations (attendances, box office, earned income, funded income and so on), making it possible to deliver comprehensive value analysis and reporting on a continuous basis

In the end these possibilities will only be fulfilled if the metrics are credible and widely owned by the cultural sector, if the process of giving and receiving feedback (for self, peers and public) is seamless and enjoyable, and if the data being generated provides real insight and value to cultural organisations and audiences alike, and to all those interested in understanding how cultural experiences change and shape us.

Therefore unless the metrics, the methods of collection, and the resulting data are of powerful practical use to the cultural sector, allowing them to refine both their cultural and commercial practices through better data driven decisions, they will not gain currency or acceptance.

We look forward to involving as many cultural practitioners, peers and audiences as possible in the next stage of co-producing metrics fit for the cultural sector's ambitious purposes.

## Appendix A – Manchester Metrics group members

Maria Balshaw (Manchester City Galleries/Whitworth Art Gallery)
Cathy Bolton (Manchester Literature Festival)
Graham Boxer/Russell Miller (Imperial War Museum North)
Matt Fenton (Contact Theatre, from October 2013)
Sarah Fisher (Chinese Arts Centre)
Jean Franczyk (Museum of Science & Industry)
Fiona Gasper (Royal Exchange Theatre)
Roddy Gauld (Octagon Bolton)
David Martin (Coliseum Oldham)
Steve Mead (Manchester Jazz Festival)
Nick Merriman (Manchester Museum)
Dave Moutrey (Cornerhouse/Library Theatre/Home)
John Summers (Hallé Orchestra)

Arts Council England
The Hive
49 Lever Street
Manchester M1 1FN

Email: enquiries@artscouncil.org.uk
Phone: 0845 300 6200
Textphone: 020 7973 6564
artscouncil.org.uk
@ace_national
facebook.com/artscouncilofengland

Charity registration no 1036733


You can get this publication in Braille, in large print,
on audio CD and in electronic formats.

To download this publication, or for the full list of
Arts Council England publications, see artscouncil.org.uk

ISBN: 978-0-7287-1543-1

May 2014

We are committed to being open and accessible. We welcome
all comments on our work. Please send these to:
National Director, Advocacy & Communications,
at Arts Council England, address above.