Quality Metrics Final Report

EXECUTIVE SUMMARY

Quality Metrics National Test

CultureCounts | QUALITY METRICS NATIONAL TEST

by John Knell & Alison Whitaker

Supported using public funding by ARTS COUNCIL ENGLAND

EXECUTIVE SUMMARY

The Key Objectives of the Quality Metrics National Test (QMNT)

The key objectives of the Quality Metrics National Test (QMNT) were to:

- Recruit and support 150 National Portfolio Organisations (NPO) and Major Partner Museums (MPM) to use the quality metrics and Culture Counts platform to evaluate three events, exhibitions or performances between November 2015 and May 2016.

- Recruit and support 10 NPOs to refine and test a set of participatory metrics developed through a previous trial, assessing their alignment with the CYP Quality Principles

- Produce an anonymized aggregated dataset and public facing report providing an analysis of the information collected throughout the QMNT highlighting key trends, points of comparison across the participating organisations, and the most important insights from the project

- Produce a separate public facing report for the participatory metrics strand of the trial highlighting key insights from the project and documenting a revised set of participatory metrics in alignment with the CYP Quality Principles

With support from Arts Council England, Culture Counts put out a call for expressions of interest from organizations to take part. During October 2015, 150 National Portfolio Organisations (NPOs) and Major Partner Museums (MPMs) signed up to test the quality metrics.

The Evaluation Activity Undertaken

During the available timeframe, 418 evaluations were conducted by the NPOs / MPMS, of which 374 evaluations used the quality metrics (as defined in Table 1). The analysis of the 24 evaluations carried out using the participatory metrics is detailed in a separate report.[1] Twenty evaluations were excluded from the trial as they were either incomplete at the data cut-off date or had not used the quality metrics.

If the 150 participating NPOs and MPMS had each reached the targets set for them by the EOI conditions the cohort as a whole would have completed 450 successful quality metric evaluations; based on 13,500 public responses; 450 self assessments, and 2,250 peer assessments.

1 Knell and Whitaker. "Participatory Metrics Report." *Arts Council England (2016)*

The overall outcomes achieved by the participating NPOs and MPMs, against those aspirational targets, were as follows:

374 successful quality metrics evaluations:	(83% of target of 450)
1,358 self assessments:	(302% of target of 450)
921 peer assessments:	(41% of target of 2,250)
19.8K public responses:	(147% of target of 13.5K)

Throughout the life of the project we saw 137 'active' organisations within the Quality Metrics National Test. Eight of the organisations who signed up through the EOI process did not engage at all with Culture Counts and the process, with the remaining 5 'inactive' organisations proving unable to complete any successful evaluations. Taken as whole this means that 91% of the cohort of NPOs and MPMs fully and successfully engaged in the Quality Metrics National Test.

Table 1: Quality Metrics

DIMENSION	STATEMENT	RESPONDENT TYPE		
		SELF	PEER	PUBLIC
Concept	It was an interesting idea	✓	✓	✓
Presentation	It was well produced and presented	✓	✓	✓
Distinctiveness	It was different from things I've experienced before	✓	✓	✓
Captivation	It was absorbing and held my attention	✓	✓	✓
Challenge	It was thought-provoking	✓	✓	✓
Enthusiasm	I would come to something like this again	✓	✓	✓
Local Impact	It is important that it's happening here	✓	✓	✓
Relevance	It had something to say about the world in which we live	✓	✓	✓
Rigour	It was well thought-through and put together	✓	✓	✓
Risk	The artists/curators were not afraid to try new things	✓	✓	-
Originality	It was ground-breaking	✓	✓	-
Excellence	It is one of the best examples of its type that I have seen	✓	✓	-

The Evaluation Data

In addition to the core quality metrics data collected in this QMNT, basic demographic data (age, gender and postcode) was collected for public respondents. In addition, metadata was also assigned to responses or events accordingly.

Our overall approach to constructing metadata was designed with highly sensitive aggregate analysis in mind in order to tell a clear overall story of the top-line aggregated results which does not over simplify the findings. In order to manipulate each data point (i.e. one answer to one question – any individual answer to any individual question) individually with any other, and to create any combination of group comparisons that are meaningful, metadata needs to be assigned to each question response.

The Culture Counts platform does this by design. For this project, we collected additional metadata to that collected via the quality metrics surveys, specifically organisation data, event data and geomapping data.

Structured data for location was applied to each event so that events in particular regions could be viewed in the aggregate, with granularity maintained by postcode for geomapping enabling versatile groups for analysis on a national scale.

Event Choice and Location

Choosing events for the trial was in the hands of the participating cultural organisations. The freedom to choose which events would be suitable to test a new approach to evaluation was important; the nature of many cultural organisations is to approach their work in a unique way and one of the objectives of the QMNT was to see how well this evaluation methodology could support this instinct.

Key Findings

What do self, peer and public responses tell us about the overall quality of work in this evaluation?

The dimension scores for individual organisations, or in aggregate, are not a clapometer, in which a successful piece of work has to be seen to score highly on every single dimension. Where self prior scores (capturing their intentions / expectations for the work) are in close alignment with self post, peer and public responses then the work is delivering against the organisation's creative expectations. Are the cultural organisations in this study adept at making these judgements in alignment with peer and public response? What are the risk and originality profiles of the work they are producing?

Taken together the aggregate results suggest:

- The work presented and analysed in this study received a broadly positive response from peer and public respondents, and largely met the (quite high) prior creative expectations of the creative teams involved in its production (self assessors)
- When it comes to measuring the quality of a cultural experience (for self, peer and public respondents) three dimensions in particular - challenge, distinctiveness and relevance – in the aggregate tended to score lower than the other six dimensions
- The clustering of self, peer and public responses in relation to these metrics suggests that audiences are adept at assessing them, with their judgements showing broad alignment with self and peer responses.
- The participating cultural organisations largely met their creative intentions, as measured by the degree of alignment between their self prior scores for each dimension and the corresponding aggregate scores for peer and public respondents
- Peer responses (as we have seen in all previous evaluations) are consistently lower across all dimensions than self and peer responses

Risk, Originality and Excellence as measured through self and peer aggregate responses for all evaluations

Risk

The aggregated self responses across these three dimensions (risk, originality, and excellence) show that self assessors tend to score themselves more highly than peer assessors; a well-established trend in previous and ongoing evaluations using the quality metrics. Interestingly, at an aggregate level the self assessors perceive themselves to be taking quite high levels of risk (broadly supported by peer scores, which are highest for this metric out of the three). This is encouraging to the extent that it would suggest that taken as a whole the cultural organisations in this study are seeking to stretch themselves with the work they are producing, and that they have a well-developed appetite for creative risk. Another interesting finding is that there is a noticeable variation in self prior risk ratings by artform. This suggests that as more data is gathered across artforms about perceived risk, this could be used to provoke dialogue both within artforms, and across artforms, about what constitutes creative risk. One outcome might be that 'risk' as a dimension measure for self and peers is thickened up with additional metric components.

Originality

Originality is the lowest ranking dimension aggregate score for both peer and self respondents. This would suggest that at an aggregate level self and peer respondents did not consider the work being evaluated in this study to display high levels of originality. Is this a surprising assessment?

The bar is set high by the originality metric, with respondents asked to express their relative support for the notion that the work 'was ground-breaking.'

How Do The Regions Compare?

The analysis cross referenced the location data for each event analysed against the ONS classification[2] for rural and urban areas (which is based on a six-fold categorization moving from strongly rural (rural 1) to strongly urban (urban 6).

To what extent did evaluations in rural and urban areas attract different profiles in terms of dimension scores? Across six of the dimensions (concept, presentation, distinctiveness, rigour, relevance, challenge, and captivation) there is no significant variations in the profile of public dimension scores as you move from the most rural areas (rural 1) to the most urban (urban 6). In other words, distinctiveness is not being rated much higher in urban as opposed to rural areas.

For two of the dimensions, enthusiasm ('I would come to something like this again') and local impact ('it is important that it's happening here'), the differences between public responses in rural and urban areas are of particular interest. Given the differential access to cultural provision in rural as opposed to urban areas, one might expect public ratings for 'enthusiasm' in more rural areas to be high, and they were (higher than in other urban areas). Similarly, one would naturally hypothesise that local impact scores would also attract high public ratings in rural areas and where this was true in some rural areas compared with urban, the rural status alone does not have a specific influence on local impact scores in all cases.

Differences in the data by artform

As one might intuitively expect, different artforms do have distinctive dimension profiles, but this only becomes clear when detailed artforms are considered in the their own right. The variations existing for each artform could be for a variety or reasons, not least the particularities of the work evaluated in this QMNT. Exploring these differences requires dialogue and debate within and across artforms.

Immersive work presents a good example. Interestingly for work defined as 'immersive' in this study, the public and peer ratings were much higher than the aggregate average for challenge, distinctiveness and relevance (the 'lower' scoring aggregate dimensions across all evaluations). With more data it would be interesting to explore whether 'immersive' work is a consistent inflator in peer and public ratings across particular dimensions.

2 http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/2011-rural-urban/index.html

Cohort Engagement, Insights and Data Culture

The evaluative and data culture of the participating organisations as judged by support requests

In addition to Learning and Insight days, much of the contact with the cohort was via in dashboard, phone and email support. In broad terms, just under half the cohort required what we would label as 'high' levels of direct support (in other words the cultural organisations in this group initiating multiple calls and emails across the evaluation period). Around a third of the cohort required moderate levels of support (initiating some calls / emails). Just under a fifth required low levels of support (initiating a few calls / emails).

There are of course some subtleties here, in that some organisations made lots of contact with Culture Counts because they were trying to do imaginative things with their evaluations, as opposed to needing 'hand holding' support to carry out a basic evaluation. Nonetheless, overall it is true to say that the majority of organisations seeking the most help were those with lower capacity and / or evaluative expertise. When cross-referenced with well-configured and executed evaluations, these findings are consistent with our previous depiction of the cultural sector in England as being 80% data shy; 15% data ready; and 5% data driven.[2]

Engagement led to improved evaluation practice and outcomes

What also became very clear during the short span of this study (the majority of evaluations took place in a concentrated period between January and May 2016) was that as the participating organisations grew more familiar with the quality metrics; and the self, peer, public triangulation approach embedded within and facilitated by the Culture Counts platform; this in turn led to more accomplished evaluative practice and better outcomes, as evidenced by:

- Culture Counts observing more accurately configured evaluations (i.e. no mistakes in configuration or in URL attribution for self, peer and public responses) as organisations moved through the project
- A declining number of support calls on how to set up evaluations in the Culture Counts dashboard
- More interest from the cohort in 'value adding' activity, such as creating separate URLs for audience sub groups, and collaborating / sharing data with other organisations
- Improving data collection outcomes in terms of the total number of public responses being achieved by the participating organisations

3 http://artsdigitalrnd.org.uk/wp-content/uploads/2014/06/HOME-final-project-report.pdf

Challenges, Issues and Opportunities identified by the cohort

Culture Counts received feedback on the challenges being encountered by participating organisations both through their direct support request and calls; and through their contributions at the the Learning and Insight sessions. The key points raised were:

Peer Management, Engagement and Building a Peer Community

Planning and managing the peer process was a new experience for all of the participating organisations and represented the greatest challenge in successfully completing their evaluations. Not only do organisations have to select their peers, but they then have to invite them and secure their participation, and follow through with peers checking that they have attended their event and completed their evaluation. Some organisations also mentioned that depending on the location of the particular event the distance required to travel for some productions is a barrier to obtaining peer assessors[3].

The paradox around the peer review element was that whilst it was the most demanding element of the evaluation process, it was also seen as a very positive aspect of the Quality Metrics National Test. Participating organisations welcomed the opportunity to invite their own peers. Strong support was also expressed for the idea that as a result of this national test a peer database is formalised across the arts and cultural sector – in other words support is given to create an open searchable data base of peers for the sector to draw on – in which each peer could list their artform expertise and interests.

Enhancing Peer Continuing Professional Development

One clear potential deficit in the current process identified by some of the participating organisations was that having secured the engagement and participation of their peer evaluators, they had received feedback from peers that they had found the process 'too short'. Peers would have happily answered more questions and would have welcomed more discussion around the results. Clearly, the length of the quality metrics question schedules has no bearing on how the peer community is engaged around the results. Even with the current peer dimensions, and additional open questions, organisations could choose to bring their peer evaluators together on a conference call to discuss their opinions and evaluations, and their reactions to the triangulated self, peer, and public ratings.

3 For example, see http://www.qualitymetricsnationaltest.co.uk/new-blog/2016/5/3/royal-shakespeare-company

These observations notwithstanding, the feedback from the participating organisations suggests that the current evaluation process may be under utilising the insights that could come from the peer evaluation process, both for the organisations, but also in terms of critical reflection and continuing professional development for the peers.

Integration with other databases; online databases; and ticketing / CRM systems

At the Learning and Insight sessions the participating organisations asked a range of questions about how a system like Culture Counts could integrate with the other databases, tools, and CRMs they already use. The organisations were both interested in the future possibilities for integration with existing systems, and informed by the desire to ensure that their evaluation activity in the round is as efficient and as effective as possible both now, and in the future.

Integrating with existing evaluation practices – adaptation and innovation

In both this quality metrics strand, and the participatory metrics strand, those organisations with well developed evaluation frameworks and practices had to think carefully about how best to integrate their quality metrics evaluation work alongside other evaluation activity they already had planned or were committed to.

Sometimes these integration challenges concerned using the metrics within multi-stranded evaluation approaches; or focused on how to design surveys and manage their distribution through a range of URLs targeting different audience segments in ways that added value to existing evaluation activity.

In practical terms these issues of integration and complementarity need to be explored by users in real evaluation examples. For example, in both the quality metrics and participatory metrics strand, one response to this integration challenge saw participating organisations innovating in their survey designs, adding in bespoke questions or picking additional questions from the Culture Counts interface. In total, 485 custom questions were added to surveys in spite of no recommendation from Culture Counts encouraging organisations to do so. The appetite to innovate (but also ask audiences lots of questions) is clearly present.

Staff Turnover and Resource Challenges

As advised by Culture Counts, the majority of participating cultural organisations designated one member of staff to be a super-user of the Culture Counts system. In other words, on behalf of a participating organisation that super-user familiarised themselves with the Culture Counts system. From the outset of the project (in evaluation terms) on November 1st 2015 to the close of the project on May 31st 2016, 14% of the the originally designated super-users of the system either left their job role, or that role disappeared for resourcing issues inside the participating organisation.

Understandably, this was very challenging for the participating organisations in terms of the continuity of their engagement in the trial which definitely impacted on the ability of some organisations to complete the target of three evaluations. This turnover of roughly one seventh of the initially inducted users presented continuity challenges within the organisations evaluating, subsequently impacting on delivery of the overall project.

Accessibility Issues

The evaluation processes highlighted a range of accessibility challenges that need ongoing attention, and the participating organisations also innovated in trying to overcome some of these issues. The specific accessibility issues identified by the cohort were as follows:

i. Those with visual impairment would struggle to complete the survey alone with the Culture Counts interface as it currently stands

ii. Working with children and adults where English is a second language can in some cases pose difficulties in accurately understanding the questions (e.g. 'hard to decipher between some specific words e.g. 'produced' and 'presented')

iii. Specific groups, such as those with dementia, pose very specific challenges (from issues of informed consent to the appropriateness of a survey-based format)

v. The survey response scales are unlikely to be clear enough for participants with ' complex individual needs'

v. Elderly respondents (e.g. sometimes with less familiarity of touch screen interfaces and a greater chance of conditions such as Parkinson's)

vi. For 'early years' participants (0-8) the text base interface is not appropriate

The language of assessment versus evaluation

In a strong mirror of the Quality Metrics National Test work on the participatory metrics, the participating organisations discussed their attitudes to evaluating their work and sharing their findings with peers and other organisations.

Organisations acknowledged that the use of standardised metrics could create anxiety around particular pieces of work being 'judged' in particular ways. Clearly, these types of evaluation approaches will only thrive if cultural organisations are encouraged and supported to explore the resulting data in ways that put the emphasis on critical reflection and improvement, as opposed to a narrow emphasis on 'audit' and 'performance reporting.'

Conclusion

Self-driven scalability

The project has resoundingly confirmed that funded arts and cultural organisations, if offered the right tools and support, can self-drive large scale evaluation activity within a very short time frame, engaging in new ways with peers and audiences about the quality of their work.

This would suggest that the quality metrics and the sector's evident interest in being able to measure their creative intentions, allied to tools that help the arts and cultural sector to collect and analyse data easily and at scale, offers up the prospect of a much richer conversation about cultural value in the future informed by big data. The aggregated data set from this Quality Metrics National Test is available from Arts Council England.[4]

The future potential of this approach

The overall evaluation approach facilitated by the features of the Culture Counts system, mirrored in the design and analysis of the aggregate data set from this Quality Metrics National Test, allows the arts and cultural sector:

- To present a very clear story on quality which does not over simplify the findings

- To use co-produced metadata frameworks, for example relating to artform descriptions, to demonstrate both the variety and plurality of work being produced by the funded portfolio; and to allow a rich analysis of quality by artform and artform attribute.

The approach effectively unites data across the standardised quality metrics, artform, artform attributes, and other open data into a powerful prism through which to better understand quality. As we have seen this will deepen understanding of how artform and certain attributes of work influence quality, and offers up the potential to produce a very wide range of analytical and reflective insights.

Crucially, the interpretation of that data will be driven and widely discussed by the creative professionals that make the work, ushering in an era of co-produced quality metrics, co-produced analytical frameworks, and a co-produced conversation about cultural value informed by big data.

4 Published and managed by Arts Council England

2 Faced Dance, DREAMING IN CODE

Courtesy:
Lawrence Batley Theatre