A Quantitative Approach to Analysing Reliability of Engagement Responses to Dance

Emery Schubert
University of New South Wales

Kim VIncs
Deakin University

Catherine Stevens
University of Western Sydney

The problem of quantification of dance response

Perhaps one of the most important end products of a dance work is how it affects its observers (typically its audience, but also the dancers and choreographers). Of the many ways of discussing and analysing dance, one approach in its infancy is quantification. Our research involves combining continuous response techniques and human response methods to see if we can tease out relationships between continuous, quantitative evaluative responses and the more qualitative choreographer intentions. The aim of this paper is to describe how evaluative responses can be quantified at all, then how they can be related to an unfolding dance work, and finally, how we can isolate 'meaningful' or 'significant' or 'reliable' evaluations of a dance work from those which are no more than a spurious set of not-very-useful numbers presented under the guise of a valid assessment.

For some, quantification immediately brings to mind the necessary limitations that are not so apparent in more qualitative investigations. For example, quantitative research in general has been accused of being reductionist, difficult to define, and having limited application (Cutcliffe & McKenna, 2002; Howe, 2004; Hanson, 2008). These issues form part of the ongoing debates on epistemology, and are far from a conclusive resolution. The choice of method often comes down to the preference and adventurousness of the investigator, as well as political motives (Paul & Marfo, 2001). However, one of the important criticisms of quantitative approaches to dance research is that dance unfolds in time, making the collection of data too simplistic if it suggests that an entire dance can be reduced to a number. Technology has allowed us to make some progress beyond this limitation. For example, in music perception there have been concerted efforts to measure continuous

responses as a piece of music unfolds (Schubert, 2001), and in dance research there are a growing number of studies that examine dancer positions in time as a dance work unfolds using motion capture hardware and software (Calvo-Merino, Glaser et al., 2005; Brown, Martinez et al., 2006; Calvo-Merino, Jola et al., 2008; Stevens, Schubert et al., 2009). These continuous data provide new potential for quantitative analysis, offering different perspectives to the more traditional approaches of understanding dance.

Measuring responses to dance continuously

Responses to dance have been collected continuously in response to dance works in some recent studies (Stevens et al., 2007; Vincs et al., 2007). These have focused on affective and evaluative responses. Affective responses may consist of rating the amount of happiness or sadness expressed by the dance, the amount of arousal or sleepiness and so on. Typically, a rating scale is presented on a computer screen or hand held portable interface, such as a PDA (Stevens, Schubert et al., 2009). The observer of the dance moves an interface, such as a stylus pen or mouse, along the rating scale to best reflect the emotion expressed by the dancer, dancers, or the overall dance environment. What they actually evaluate, and the rating scales themselves, are defined and described before the assessment task begins. For example, they may be asked to move the slider to the top of a vertical scale if the dance work is expressing a high level of arousal, and to the lower part of the slider if sleepiness or restfulness appears to be expressed from the point of view of the observer.

Measuring level of engagement

Another approach is to measure the level of engagement with the dance, where the observer is rating the dancers on a scale that spans from 'engaged' at one end to 'not engaged' at another. In one of our recent studies, the participants were provided with the following definition of engagement:

> compelled, drawn in, connected to what is happening, interested in what will happen next.
>
> (Vincs, 2007)

With rating scales, observers can interpret this definition as they wish and then on the rating scale rate the degree to which they believe they are being engaged. So, how do we know that each observer is using an identical interpretation of the definition (in this case, definition of engagement)? And even if they did, how do we know that the observers agree on the level of engagement at each point in time throughout the dance work? We will argue, that there exist relatively simple (in addition to some more complex quite sophisticated) methods of dealing with the second question. As for the first question, part of the answer is that we do not know whether people are applying an identical definition of, in this case, engagement. However, if the level of agreement, the second question, were found to be good, then we might conclude that the definitions used across participants for 'engagement' were at worst related, and at best identical. From a psychological perspective, we are able to proceed without such precise knowledge, because in

psychological, perceptual research we assume that some error is present, as we shall see below. And this imprecise definition places us in no worse a position than qualitative approaches to the question of how people respond to dance works.

Imagine, then, that a number of observers are watching a dance work while making continuous engagement responses to the same dance work, and that they did so using the continuous response apparatus described above, with the above instructions for rating the engagement of the dance. While we have conducted such a study (e.g. see Vincs, 2007), we discuss the approach in general terms here in an attempt to describe an alternative way of measuring variation in response, and consequently the level of agreement.

Analysing continuous responses

When an observer makes their engagement response continuously, the resulting data make up a time series, consisting of a stream of numbers, each number representing the engagement level at each point in time. We can plot this time series, and examine the shape of the plot then compare it with sections of the dance through analysis of the video of the same performance, or inspection of choreographic notes, or both. The actions occurring at different levels of engagement can then be further inspected. Simplistically, dance movement at peak levels of engagement may be examined, and trough (low) levels may also be identified, with a view to asserting how a dance work could be made more or less engaging. We do not recommend such an approach if for no other reason than that it articulates one of the potential foibles associated with quantitative data – reductionism. Nor is it the intention of cognitive science to necessarily influence the choreographer's artistic approaches and decisions. We know more about the complexities of how a response might be related to a potentially causal dance action. For example, the context of the section of the dance, or a response set off by a contrasting section may be a contributor to the high engagement that was thought to be causally related to the dance motion occurring at the time of the response. Further, the observer has memories and expectations that further affect response (Calvo-Merino, Grezes et al., 2005; Calvo-Merino, Grezes et al., 2006) just as they do when listening to music (Granot & Donchin 2002; Tillmann, Janata et al., 2003). One way this can be investigated in quantitative data is by analysing the 'serial correlation' in the data – looking at whether parts of the engagement response at one point in time can be predicted by, for example, combinations of previous points in time. We will not focus on this issue, although it has been discussed elsewhere (Box, Jenkins et al., 1994; Schubert, 2002). However, even with these techniques, we still cannot be certain that a second observer will respond with the same set of complex patterns (i.e. the same time series).

Spread of response scores at each moment in time

The approach we have taken to address this reliability of response issue is to examine the variability of engagement scores at each point in time across all the participants who provided their engagement responses to the same dance work. If the variability is small, we assume that there is a high level of

agreement between the observers at that point in time (good reliability). However, large levels of disagreement will indicate that the response at that point in time was not reliable (for reasons such as differences in mood, age, gender, economic, social and cultural backgrounds), and any assertions or interpretations about the level of engagement at that point in time in relation to the choreography should be treated with caution. The method of examining variability of scores at each point in time is a simplification we consider necessary to begin to understand what contribution quantitative techniques can make to dance response. But even so, two issues need to be considered here – how to measure the amount of spread of scores at each response moment, and then how to interpret them.

Statistical methods present us with several ways of measuring the spread of scores. One commonly used method is to calculate the standard deviation; this is a method that we have adopted recently (Schubert, Vincs et al., under review). However, in the present paper we will speculate on interquartile values because they provide greater validity and are conceptually simpler than the ubiquitous standard deviation. Indeed, researchers in music perception working with continuous response to music have recently adopted the interquartile method for calculating spread of scores (Korhonen, Clausi et al., 2006; Grewe, Nagel et al., 2007). Interquartiles are simply a collection of all the responses made at a given point in time, sorted in ascending numerical order, and then grouped into four sections. So, for example, on a scale of 0 to 100, where 100 is the highest level of engagement possible on the scale, and 0 the lowest, let us assume that at the twentieth second of the dance, eight participants have their sliders in the following positions (in ascending order): 60, 65, 75, 75, 75, 75, 85, 90. The lowest score at that point in time is 60 and the highest 90. The bottom quarter of responses (two out of the eight) consist of the scores 60 and 65. The next quarter of responses (the next two out of the eight) are 75 and 75. Therefore the lower interquartile value is the boundary between these two quartiles, which is 70 (the value in between 65 and 75). The highest interquartile value is 80, the value falling between 75 and 85 (separating two scores of 75, and the two high scores [upper quartile] of 85 and 90). This is a very simple system for reporting the spread of scores. According to this method, the larger the difference between the upper and lower quartile, the larger the spread of scores. In the present example the interquartile distance at the 20th second of the dance is 10 (= 80-70) 'engagement units'. This measure is called the interquartile distance, and is analogous to standard deviation, but requires a simpler calculation, and does not make the assumptions that the standard deviation calculation makes (which we will not discuss here, but see, for example, Haslam & McGarty, 2003).

We then need to decide what interquartile distance is too large to be able to assume good agreement. The decision is dependent on several parameters, including the number of observers making the responses. The more observers, the better the estimate of the underlying interquartile distance. We can then apply non-parametric statistical analyses (for example Siegel & Castellan, 1988) to assess whether the distance is too great to be considered indicative of good agreement among respondents. We are interested in

determining whether we can empirically determine some rule of thumb for determining a good level of agreement or significance of response – which is at the core of our research interest. Finding such a solution (or principle for solving it) will provide a quantitative technique with a reasonably objective way of measuring when participants' responses are reliable, and subsequently when sections of dance can be asserted as eliciting reliable levels of engagement.

Causes of variability in responses

Finally, we speculate on some of the causes of this variability (poor reliability) in response. There are several factors that contribute to variation within a single participant's responses: their familiarity, mood, personality can all affect their response each time they perform the task (for example, if rating the engagement of an audio/video recording of the same performance on several occasions). It should be evident that these variables can be different across different observers as well, even when observing the same dance at the same time. The level of concentration on the task is another variable. It is unlikely that an observer will be 'on task' for an entire performance: they may be focused on the performance but forget to rate the engagement on the computer, or not feel like rating the engagement, or be reconfiguring their definition of engagement, or simply be resting/unfocussed. These issues will be prevalent for long duration pieces, perhaps even performances longer than a few minutes. The evidence for this comes from the variability in deviation scores over time that we have found in our own data (which in statistical terms is referred to as heteroskedasticity).

If we wish to capture responses to long performances some compromises must therefore be made. So, for the longer performances, these losses in responses might be compensated by a larger number of observers. For example, if we had 100 observers, and observer one went off-task for several seconds, the 99 other responses in that period of time would ensure a negligible effect on the overall ratings. Nevertheless, this moving in and out of focus may be reflected in the deviation scores, and have some accumulation among participants. If many people are not focused on the task at the same time, we may expect larger deviation scores (larger interquartile distance) over those periods of time. We may not be certain whether the large deviations are due to error (the fluctuations indicating how on-task the observers are) or because there is some 'true' lack of agreement in the rated engagement. Deviation scores should therefore be modelled as 'signal' deviation and 'error' deviation, where signal corresponds to 'true' agreement, and error to 'off-task' and other non-task related issues).

Future research will determine how these statistical principles can be applied to address questions in dance perception that have until now been largely monopolised by qualitative introspection and retrospection. By developing quantitative techniques with application to the temporal arts, we are hoping to build up a richer picture of how dance works are perceived by the population represented by the observing participants.

Conclusions

Deviation scores provide great opportunities in analysing and understanding responses to dance. Even applying the simplest techniques, including time series plots of standard deviations (Vincs, Schubert et al., in preparation), or as described in this paper, the interquartile time series can instantly provide a bird's eye view of how agreement in response fluctuates from moment to moment. From this we are able to determine a new order of information that a moment by moment *mean-response* time series could not tell us – the reliability of the responses. We therefore argue that by incrementally applying more sophisticated analytic techniques of time series analysis the otherwise complex field of time series will provide new insights into how responses to dance can be quantitatively investigated. Some of these increasingly sophisticated approaches have been applied to continuous response to music (Schubert, 2001) and should easily translate to the dance medium.

Among the temporally based arts, the methods described here have mainly been applied in the past to music perception and production (Meyer & Palmer, 2003; Palmer & Pfordresher, 2003; Highben & Palmer, 2004). Sophisticated statistical approaches have led to new insights into responses, such as lag structure (the time delay between an action in the dance work and a reaction in the response), and the variability in the time at which a response is made after some causal event (Schubert & Dunsmuir, 1999). In addition, the results can be compared with post performance data collection techniques or more qualitative approaches. We see great potential in applying these complementary and converging techniques in various dance environments to add to our depth of understanding of this complex, temporal art form.

Acknowledgements

References

Box, G. E. P., Jenkins, G. M. et al. (1994). *Time series analysis: Forecasting and control.* New Jersey: Prentice-Hall.

Brown, S., Martinez, M. J. et al. (2006). The neural basis of human dance. *Cerebral Cortex, 16*(8), 1157-1167.

Calvo-Merino, B., Glaser, D. E. et al. (2005). Action observation and acquired motor skills: An fMRI study with expert dancers. *Cerebral Cortex, 15*(8), 1243-1249.

Calvo-Merino, B., Grezes, J. et al. (2006). Seeing or doing? Influence of visual and motor familiarity in action observation. *Current Biology, 16*(19), 1905-1910.

Calvo-Merino, B., Grezes, J. et al. (2005). The influence of visual and motor familiarity during action observation: An fMRI study using expertise. *Journal of Cognitive Neuroscience,* 115-115.

Calvo-Merino, B., Jola, C. et al. (2008). Towards a sensorimotor aesthetics of performing art. *Consciousness and Cognition, 17*, 911-922.

Cutcliffe, J. R., & McKenna, H. P. (2002). When do we know that we know? Considering the truth of research findings and the craft of qualitative research. *International Journal of Nursing Studies, 39*(6), 611-618.

Granot, R., & Donchin, E. (2002). Do Re Mi Fa Sol La Ti–Constraints, congruity, and musical training: An event-related brain potentials study of musical expectancies. *Music Perception, 19*(4), 487-528.

Grewe, O., Nagel, F. et al. (2007). Listening to music as a re-creative process: Physiological, psychological, and psychoacoustical correlates of chills and strong emotions. *Music Perception, 24*(3), 297-314.

Hanson, B. (2008). Wither Qualitative/Quantitative? Grounds for methodological convergence. *Quality & Quantity, 42*(1), 97-111.

Haslam, S. A. & McGarty, C. (2003). *Research Methods and Statistics in Psychology.* UK: Sage.

Highben, Z. & Palmer, C. (2004). Effects of auditory and motor mental practice in memorized piano performance. *Bulletin of the Council for Research in Music Education,* (159), 58-65.

Howe, K. R. (2004). A critique of experimentalism. *Qualitative Inquiry 10*(1), 42-61.

Korhonen, M. D., Clausi, D. A. et al. (2006). Modeling emotional content of music using system identification. *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 36*(3), 588-599.

Meyer, R. K. & Palmer, C. (2003). Temporal and Motor Transfer in Music Performance. *Music Perception, 21*(1), 81-104.

Palmer, C. & Pfordresher, P. Q. (2003). Incremental planning in sequence production. *Psychological Review, 110*(4), 683-712.

Paul, J. L. & Marfo, K. (2001). Preparation of educational researchers in philosophical foundations of inquiry. *Review of Educational Research, 71*(4), 525-547.

Schubert, E. (2001). Continuous measurement of self-report emotional response to music. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 393-414). Oxford: Oxford University Press.

Schubert, E. (2002). Correlation analysis of continuous emotional response to music: Correcting for the effects of serial correlation. *Musicae Scientiae, Spec Issue*, 213-236.

Schubert, E. & Dunsmuir, W. (1999). Regression modelling continuous data in music psychology. In S. W. Yi. (Ed.), *Music, Mind, and Science* (pp. 298-352). Seoul: Seoul National University.

Schubert, E., Vincs, K. et al. (under review). Identifying regions of good agreement among responders in engagement with a piece of live dance. *Dance Research Journal.*

Siegel, S. & Castellan, J. J. (1988). *Nonparametric statistics for the behavioural sciences.* New York: McGraw-Hill.

Stevens, C., Schubert, E. et al. (2009). The Portable Audience Response Facility (pARF): PDAs that Record Real-Time and Instantaneous Data During Live or Recorded Performance. *International Journal of Human-Computer Studies, 67*, 800-813.

Stevens, C., Schubert, E. et al. (2009). Moving With and Without Music: Scaling and Lapsing in Time in the Performance of Contemporary Dance. *Music Perception, 26*(5), 451-464.

Tillmann, B., Janata, P. et al. (2003). Activation of the inferior frontal cortex in musical priming. *Cognitive Brain Research, 16*(2), 145-161.

Vincs, K., Schubert, E. et al. (in preparation). The Gem Moment: Using quantitative techniques to analyse responses to dance. *Journal of Dance Medicine & Science.*

Vincs, K., Schubert, E., & Stevens, C. (2007). *Engagement and the 'gem' moment: How do dance students view and respond to dance in real time?* Proceedings of the 17th Annual Meeting of the International Association for Dance Medicine and Science. Canberra, Australia, October 25-28.

Biographical statements

Emery Schubert is a Leader of the Empirical Musicology Group School of English, Media and Performing Arts at the University of New South Wales, co-editor of Acoustics Australia and President of the Australian Music and Psychology Society (AMPS). His key research areas are in music perception and cognition, with specialisation in continuous measurement of aesthetic and affective responses to music.

Dr Kim Vincs is a choreographer, Associate Professor in Dance, and the Director of the Deakin Motion.Lab, Deakin University's new motion capture studio and research centre. Her research focuses are choreographic process, interactive performance and motion capture, and developing new methodologies for integrating quantitative and practice-based research.

Dr Catherine Stevens applies experimental psychology methods to the study of auditory and temporal phenomena including music, dance, and environmental sounds. She holds BA (Hons) and PhD degrees from the University of Sydney. Kate is an Associate Professor in Psychology, MARCS Auditory Laboratories at the University of Western Sydney.