Free-Improvised Rehearsal-as-Research for Musical HCI

Charles P. Martin and Henry Gardner

October 12, 2017

Abstract

The difficulties of evaluating DMIs (digital musical instruments), particularly those used by ensembles of musicians, are well-documented. These devices are designed to support creative expressions that are often improvised and unexpected. In collaborative settings, interactions are highly complex and it can be almost impossible to directly examine the impact of a new interface. Concert pressure and practicalities limit experimental control but lab studies do not reflect real-world usage.

To address some of these issues, we propose a methodology of rehearsal-as-research to study free-improvisation by ensembles of DMI performers. Sessions are structured to mirror established practices for improvisation training and performance development. These sessions allow controlled, order-balanced studies with extensive data collection in the style of factorial human-computer interaction (HCI) experiments while preserving artistic outcomes. Experiment design, questionnaires, and objective measures such as session duration will be discussed along with two case studies.

Keywords: New Interfaces for Musical Expression (NIME); Evaluation; Methodology; Rehearsal; Improvisation; Collaboration.

1 Introduction

It has been well-documented that systematic evaluation of collaborative creativity support tools is difficult, particularly in performing arts practices (Shneiderman, 2007). Lab-based research into musical HCI can often seem arbitrary and disconnected from genuine artistic creation; live concerts present challenges in formal data collection, and add the pressure of performance and of entertaining an audience. The concept of evaluating digital musical instruments (DMIs) and new interfaces for musical expression (NIMEs) appears to be gaining relevance (Barbosa, Malloch, Wanderley, & Huot, 2015), and calls for evaluation frameworks and methodologies are made with increasing urgency (Jordà & Mealla, 2014).

Free improvisation, however, has a history of practice and pedagogy that is ideally suited to examining new tools for collaborative musical interaction. A potential methodology for investigating the use of these tools is a process of *rehearsal-as-research*. This re-frames the typical improvised rehearsal process as a controlled HCI study. Such a process would provide opportunities to capture subjective data, such as surveys, discussions, and interviews, as well as objective,

such as data logs from instruments, and performance recordings. Crucially, this process can inform interface refinement and still constitutes practical artistic development that leads to concert performances. Thus, new interfaces can be examined from an artistic perspective, in performance, as well as from a scientific perspective, through the analysis of formally collected data.

In this chapter we argue for a rehearsal-focussed methodology in musical HCI and contrast this approach with others in the literature. We present a practical framework for rehearsal-as-research including consideration of experimental design, session structure, questionnaires, and objective measures such as the duration of improvisation. We conclude with two case studies applying this framework in studies of ensemble improvisation with touch-screen digital musical instruments (DMIs).

1.1 Rehearsals with Improvising Ensembles

Free- or non-idiomatic improvisation has no restrictions on style and no predetermination of the music that will be played. Free-improvised performances often take the form of explorations, both of a musical world, and of the affordances of an instrument. For this reason, free-improvisation is used with NIMEs, where the parameters of musical interaction may be unmapped. In ensemble form, this style of music making involves negotiations of musical decisions and game-like interactions that are compelling to both performers and audiences.

The emphasis on exploratory performance and collaborative creativity in free-improvisation has led to its adoption in pedagogies such as Cahn's (2005) Creative Music Making (CMM). Cahn defines a particular style of improvisation characterised by complete freedom for all performers: "performers may play (or not play) anything they wish" (p. 35). Although Cahn suggests that performers should listen carefully to themselves and others it is emphasised that "there is no penalty for breaking this rule" (p. 35). Sessions may have a determined starting point, but as performers are free to stop playing whenever they wish, the end of performances is defined to be "when all of the players have individually decided to stop playing" (p. 41).

CMM sessions typically consist of multiple improvisations, as well as discussions and listening-back sessions similar to the structure of video-cued-recall (Costello, Muller, Amitani, & Edmonds, 2005) sessions. We have previously conducted HCI research using this kind of process to investigate an ensemble of iPad performers (Martin, Gardner, & Swift, 2014).

1.2 Evaluation in Computer Music

In computer music a range of methodologies have been explored for evaluating NIMEs in the lab and on stage, many borrowing concepts from HCI (Wanderley & Orio, 2002). O'Modhrain (2011) argues that there are multiple stakeholder perspectives that could be considered in evaluating a NIME, including audiences, performers, designers, and manufacturers. The most important of these stakeholders, however, are performers as they are "the only people who can provide feedback on an instrument's functioning in the context for which it was ultimately intended" (O'Modhrain, 2011, p. 34). For improvised music, this is particularly important, as the performer is responsible not only for translating

musical intentions into sound with the NIME, but for creating these intentions as well.

Computer music evaluations frequently use qualitative approaches applied to interviews conducted after a period of initial experimentation. Ethnography is often used to make sense of video and interview data from such studies (Krüger, 2008). Longitudinal research has also been advocated to go beyond the first impressions of an interface (Gelineck & Serafin, 2012). Studies such as Xambó et al.'s (2013), have used ethnographic techniques focussed on video footage to investigate ensemble performance over several rehearsals. This study also took into account the collaborations that are well-known to emerge between ensemble performers (Hayden & Windsor, 2007).

Natural rehearsal and development processes were studied in the *Vocal Chorder* project (Unander-Scharin, Unander-Scharin, & Höök, 2014). Here, the design process started in an autobiographical manner (Neustaedter & Sengers, 2012), with the performer evaluating their own artistic outcomes. Indeed, we have also used these methodologies to characterise ensemble performance practices. We suggest that, by borrowing from the process of CMM, these techniques could be complemented by written surveys and other quantitative data collected during sessions.

Other researchers, too, have focussed on the performer perspective to evaluate and improve NIMEs. Notable examples include Gurevich et al.'s (2012) work, where participants practiced and performed with very simple electronic instruments and were studied using a grounded theory approach (Corbin & Strauss, 2008). Performer evaluations of their interface and performance can also be used to iteratively improve a NIME, such as Fiebrink et al's (2011) "direct evaluation" approach to developing musical machine learning systems. Stowell et al. (2009) have interviewed individual NIME performers to evaluate interfaces, as well as groups of performers together.

Longitudinal studies of NIME performers allow observation of how users learn to perform with a new interface, and improvisation ensemble also learn how to interact with each other. Cahn (2005, pp. 37–38) noted that new ensembles overcame initial inhibitions with "severe departures from normal music making", which was followed by more thoughtful and balanced performance. In the present research we argue that even single rehearsal-as-research sessions can produce useful outcomes. Some of our groups had extensive experience so a practice phase was not necessary. For new ensembles, we relied on balanced ordering of experimental conditions, and the use of NIMEs that had already been refined, to control for learning effects. However, it would also be possible to deploy rehearsal-as-research over several sessions.

2 Rehearsal-as-Research Methodology

In the absence of an HCI study, a common format for examining a new musical interface would be for a group of musicians to get together and improvise with it in rehearsals, followed by a live concert. We propose that some of these rehearsals could be formally structured to compare variants of the interface. At the cost of some performative freedom in the sessions, these rehearsals-as-research allow the interface to be probed systematically. Over a number of HCI studies of musical interaction (e.g., Martin et al. (2014); Martin, Gardner, Swift,

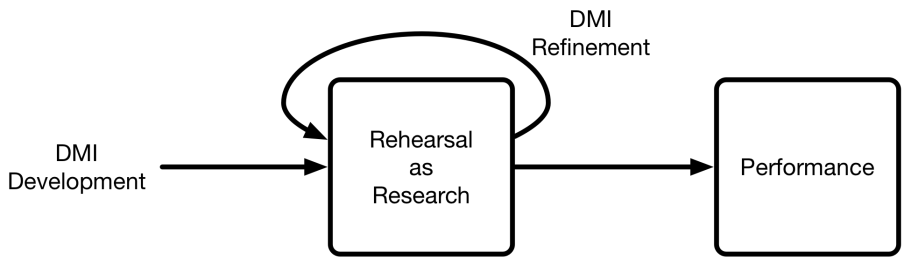

Figure 1: Rehearsal-as-research allows DMIs to be evaluated formally while also developing artistic practice. One or more research session could be part of a rehearsal process leading to interface refinements and concert performance.

and Martin (2015, 2016)), we have developed a methodology for designing and analysing rehearsal-as-research sessions. The sessions consist of sequences of ensemble improvisations where performers explore variations of a DMI design. Research results from each session can inform design refinements that are analysed in further rehearsals. Finally, the artistic results of the sessions can be exhibited in public performances. This process is illustrated in Figure 1.

Our research has involved developing NIMEs to support ensemble performance. A challenge when evaluating these NIMEs has been gathering evidence that technological innovations really do improve aspects of performances. So, our methodology is focussed on comparing different interfaces to assess the impact of particular features or designs. As participants for ensemble performance studies are difficult to recruit, our studies have presented each performer with every interface condition, known as a within groups design. However, rehearsal-as-research could also use a between groups design where each group only uses one or some of the interface conditions.

2.1 Session Structure

The structure of a rehearsal-as-research session mirrors the typical rehearsal practice of repeating performances under different conditions or with different musical instructions. Sessions begin with an orientation of each condition of the musical interface. During this part of the session, the rules (or lack of rules) of CMM style improvisation can be explained. Each experimental condition is then used in CMM improvisation with the researcher outside of the rehearsal studio. After each improvisation, the performers fill in written questionnaires, and at the end of the whole session, an open-ended interview is conducted.

In our sessions, each group improvises with all of the different musical interfaces that form the experimental conditions of the study. So, the number of conditions has an important impact on the length and complexity of sessions. The experience of the participants is a factor in determining the number of improvisations and interfaces to cover. Complete coverage of more experimental conditions requires more improvisations. It is also possible to replicate performances allowing more data to be captured and the development of a more nuanced perspective on the interfaces under test.

The learning effects of new improvising ensembles is well known, and it is possible that such effects could impact comparisons between interfaces. As a

result, performances with experimental conditions should be in balanced order. It may not be possible to account for every permutation of the conditions, but techniques for counter-balancing within groups designs for immediate carry-over effects are available (e.g., Bradley (1958); Williams (1949)). One approach we have taken is to balance the participants' exposure to a visible interface, but not to more subtle background network interactions. This meant that many fewer improvisations were required.

We have explored two different rehearsal structures. In the first, an experienced group performed three replicates of six different interface conditions for a total of 18 improvisations over three hours. In the second, four new groups participated in 90 minute sessions of one improvisations with each of four interface conditions. The sessions began with an introduction to the interfaces and CMM performance. Each improvisation was directly followed by a written survey. An open ended interview was held at the end of each session to compare the performers' experiences. More details of these sessions are explained in the case studies below.

2.2 Questionnaires

In our studies, we have asked performers to fill in written surveys after each improvisation to gain their immediate perspectives. We design our surveys to assess the quality of improvisations and of the musical interfaces under examination. Although free-improvised performance is considered difficult to examine objectively, rating systems have been developed for assessing improvised performances in musical education (Smith, 2009), and in solo improvisations (Eisenberg & Thompson, 2003). While these systems are typically used by expert assessors, we think that performers are also able to evaluate performances, at least from their own qperspective.

Our questionnaires have consisted of multiple ordinal rating scale questions that follow basic aspects of improvised musical interaction: technical proficiency, ensemble interaction, musical structure, creativity, performance quality, and enjoyment. Most of these aspects are clearly linked with the quality of improvised ensemble music; however, the interpretation of technical proficiency differs from that in music education. In the context of comparing NIME performances, technical proficiency connects with ease of use, or expression. Rather than assessing the performer's expertise, we ask them to rate how much an interface enhanced or impeded their ability to explore and express musical ideas.

Short written surveys consisting of ordinal rating scale questions can be administered quickly after each improvisation session without disruption. These surveys present questions such as "How would you rate the creativity in that performance?". Responses are given by selecting one option from a list, e.g, *terrible, bad, neutral, good, excellent*; or by selecting a position between labelled extremes and midpoint, e.g., *terrible, neutral, excellent*. We have observed that participants appear to have little trouble self-assessing an improvisation and are frequently in consensus regarding various aspects of the performances. Rating scale surveys do limit the detail that performers can provide and restrict responses to the researchers' pre-determined questions. However, they are very quick to administer in a session allowing the participants to focus on performance rather than data collection.

As quantitative data, rating scale responses can be subjected to statistical

analysis and significance testing. There are ongoing discussions among HCI and other researchers about the appropriate statistical methods to apply to ordinal (rather than continuous) rating data. Such data may not meet the prerequesites of parametric tests such as ANOVA (Analysis of Variations) and the t-test (Gardner & Martin, 2007). In practice, many researchers apply ANOVA to non-parametric data anyway as this procedure can model the main and interaction effects of experiments with multiple factors and repeated measures. However, newer procedures, such as the ART (Aligned Rank Transform) ANOVA (Wobbrock, Findlater, Gergle, & Higgins, 2011) have been recommended among other options for non-parametric testing (Wobbrock & Kay, 2016).

2.3 Duration: An Objective Measure

The survey methods discussed above directly interrogate performers about the exact we wish to evaluate. However, even the most assiduous performer may not notice every aspect of performances. Furthermore, in the successful improvisations we wish to study, where performers enter a state of creative flow (Csikszentmihalyi, 1991), their awareness of external factors may be lessened. Given that rehearsals-as-research take place in a controlled environment, objective data can be captured during improvisations. Much interaction data can be logged directly from digital musical instruments for later analysis (Martin & Gardner, 2016); however, in this section we focus on our experience using the duration of improvisations as a dependent variable in rehearsal-as-research.

The duration of an improvisation can be influenced by several factors. Some improvisers agree on a performance duration and use a stopwatch to define the finishing point. In many cases the ending of an improvisation is found by a tacit agreement between performers, that is, when all in the group have stopped playing. This process of finding a natural ending can be complicated by the performance context; if the audience looks bored, or one performer signals the end of a performance by fading out or playing slowly, there is pressure to finish an improvisation. The pressure is lessened in a rehearsal situation. As a result, the length of an improvisation could be more related to the performers' level of creative engagement with an interface than other factors.

In our rehearsal sessions we collect individual, as well as group, session durations. The start of a performance is given by the time of the first sound. The individual end times are given for each player by their final sound. The group end time is given by the final sound of the last player to stop as in CMM. We calculate these times automatically from synchronised logs of each players' touch-screen interactions; however, similar calculations could be made from other recordings. By recording individual times for each performer, more precise statistics can be calculated regarding the effects of different interfaces on improvisation.

Although the CMM process could technically allow performers to finish after an extremely short performance, we wish to only record performances that are long enough to fairly evaluate the interface under test. We set a lower-bound of seven minutes on improvisations and use a stage lighting system to give non-intrusive signals to the performers. Two small stage lights have been set on either side of our rehearsal studio within all performers' vision. At the start of each improvisation, the stage lights are set to green to indicate that performers must continue playing. After seven minutes of improvisation, these lights are

remotely faded to blue, indicating that performers can stop when they wish. In practice, we have found that performers usually continue for some time after the change. Further, the performers are generally unaware of the relative length of improvisations. Future studies could investigate the relationship between session length, engagement, creative flow, and performers' subjective ratings.

2.4 Capturing the Performance

Questionnaires and interaction metrics have clear scientific justification but we should bear in mind that the musical performances themselves are still the primary artifact to come out of a rehearsal. We record rehearsal-as-research sessions in the highest practical quality. In our studies of touch-screen instruments, this has meant recording audio directly from each device, as well as from a microphone in the rehearsal studio, video from multiple cameras mounted in the studio ceiling to observe the performers' touch gestures, and logs of time coded touch-interaction data. For other NIMEs, it may also be appropriate to record movement via a motion capture system, using either optical or inertial sensors (Skogstad, Nymoen, & Høvin, 2011).

These data form an extremely detailed record of performances. In some cases, where a rehearsal-as-research process overlaps with studio recording, this record can serve as an artistic output in its own right. From a research perspective it allows the the findings of the study to be supported by the musical performances themselves. For instance, notable occurrences in interaction logs or performer ratings can and should be examined directly in performance recordings. In fact, the whole performance record can also be coded using ethnographic techniques to explore the performers' interactions and musical ideas. Finally, the performance record could be used to validate the findings of a study through external assessors. While there is no substitute for a live performance, using high-quality recordings may be more appropriate for a controlled study of audience perspectives. Given the ability of a recording to act as a primary record for further research, we recommend that (where possible), rehearsal-as-research recordings should be published as an open access archive.

3 Case Studies

In this section we will describe two studies implementing a rehearsal-as-research methodology. Both of these studies involved evaluations, and comparisons, of ensemble-focussed touch-screen digital musical instruments. The participants evaluated and compared a number of interface candidates through a series of rehearsal performances. The objectives were to gather evidence that the interfaces in question enhanced the ensemble improvisations and examine which aspects of these performances were affected. While the objectives of were similar, these studies differed in the performers who participated; the earlier included one group of professional perfomers while the latter used multiple ensembles with mixed experience. In the following sections, each application of the rehearsal-as-research process will be described.

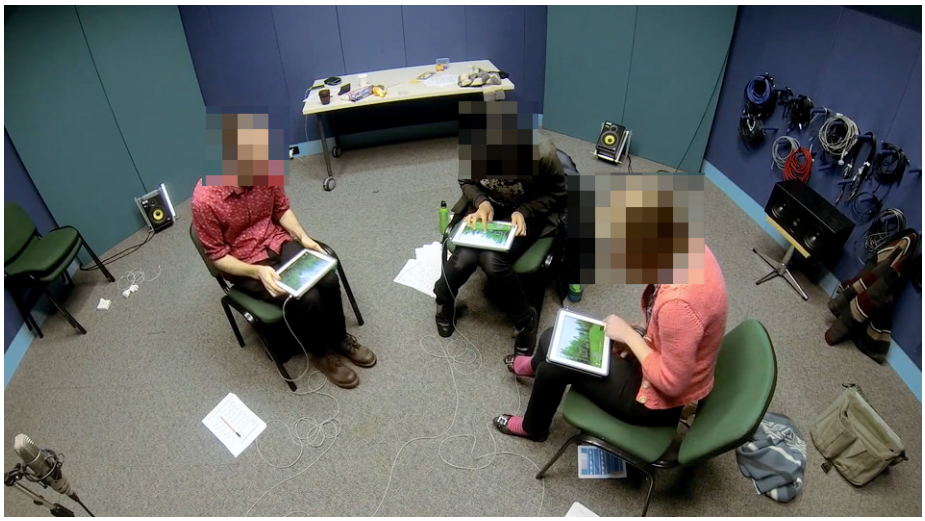

Figure 2: A touch-screen trio in a rehearsal-as-research session. This group performed 18 5-minute performances consisting of three replicates of six interface conditions.

3.1 Trio Performance Study

In this study, a group of three percussionists evaluated three touch-screen musical interfaces and two ensemble director agent systems; a total of six performance conditions. The performers, shown in Figure 2, involved had extensive experience performing as a percussion ensemble so it was feasible to ask the group to participate in an intense rehearsal session of a similar length to professional calls. Significantly, this group had participated in previous rehearsals, studies, and multiple public performances with earlier versions of the touch-screen instruments so they were expert users of the interfaces they were evaluating.

The session was structured around 18 5-minute performances which allowed for three replicates of each of the six performance conditions. These 18 performances were divided into six sets of three, with each set exposing the performers to each of the three interfaces. The performers had a short break in between each set. The ordering of the apps in each set was balanced according to Williams' design (Williams, 1949). The agent systems were alternated between successive performances of the same app. The session opened with an orientation of the experimental procedure and closed with an open-ended interview. The session structure is shown in Table 1.

The participants in this study completed written questionaires directly after each performance. These consisted of seven questions relating to the quality of the improvisation, and the individual and group interaction with the interface and agent. The questions are shown in Table 2 and the responses were given on a five-point Likert-style scale. Interaction data was also collected for this study including video and audio of the session, all touch-screen interactions, and agent-ensemble interactions. The questionnaire data was analysed using the traditional two-way repeated-measures analysis of variance (ANOVA) procedure (Lazar, Feng, & Hochheiser, 2010, §4.5.3) to determine effects due to the two factors (interface and agent). The results of the sessions motivated design

Set	Performance 1	Performance 2	Performance 3
0	orientation		
1	I1, A1	I2, A2	I3, A1
2	I2, A1	I3, A2	I1, A2
3	I3, A1	I1, A1	I2, A2
4	I3, A2	I2, A1	I1, A2
5	I2, A2	I1, A1	I3, A1
6	I1, A2	I3, A2	I2, A1
7	interview		

Table 1: The experiment schedule showing the balanced ordering of interfaces (I1, I2, I3) and agents (A1, A2). The experiment consisted of one session divided into six groups of performances.

Q#	Question Text
Q1	How would you rate that performance?
Q2	How would you rate the level of creativity in that performance?
Q3	How did the agent's impact compare to having it switched off?
Q4	How well were you able to respond to the app's actions?
Q5	How well were you able to respond to the other players' actions?
Q6	How was the app's influence on your own playing?
Q7	How was the app's influence on the group performance?

Table 2: The questionnaire filled out by the trio after each improvisation consisted of these seven questions. Each was answered on a 5-point Likert-style scale (*terrible, bad, neutral, good, excellent*).

refinements in each interface which were used by the group in a subsequent concert. A full report of this study is available in another publication (Martin et al., 2015).

Using expert performers in this study enabled a complicated rehearsal structure to compare a total of six interface conditions. An important decision in this study was to replicate performances under each condition to capture more accurate ratings of how the interfaces impact performance. To fit 18 performances in one session, the improvisations were limited to five minutes; however, this prevented us from measuring performance duration and some performances may have been stopped before the performers' musical ideas had been sufficiently explored. While we believe that useful evaluation data can be gained from three expert performers, the very small sample size limits the power of statistical tests such as ANOVA. The 5-point scale further constrained analysis; performers rarely gave negative ratings, so there were only three effective points in the scale between the neutral mid-point and the positive end.

3.2 Quartet Cross-Ensemble Study

This study examined 16 participants as they improvised with four different touch-screen interfaces designed to help mediate their ensemble interactions. These interfaces were designed specifically for this two-factor study that com-

Figure 3: A quartet of participants engaging in rehearsal-as-research. Four such groups compared four interface variants in this study.

pared a UI and agent-based approach to ensemble mediation. Each system formed one condition along with a no-mediation (control) condition, and the combination of both systems. These participants were split into four quartets, one of which is shown in Figure 3. Each quartet improvised with the same set of interfaces so this was a repeated measures study. The participants in this study were not experts in the NIMEs they were asked to evaluate, however they were skilled musicians recruited from within a conservatorium.

The sessions in this study were designed to be around 90 minutes in duration. As most of the participants were not familiar with the NIMEs used, an orientation at the start of each session included trial performances with each interface condition. Following the orientation, the groups performed improvisations with each interface in counter-balanced order produced by following Bradley's (1958) procedure. The performance order is shown in Table 3. Unlike the previous study, the duration of improvisations was an important metric. The procedure described in Section 2.3 was followed to indicate lower and upper bounds to the groups, and otherwise, performances ended when all performers had stopped playing.

Two questionnaires were used in this study: a performance questionnaire after every improvisation, and a preference survey administered at the end of the session. The performance survey was much more intensive than the previous study with 24 question covering five aspects of improvised performance: technical proficiency, musical interaction, musical structure, creativity, overall quality. Each aspect had a number of questions, and, in particular, asked performers to assess both their individual performance, and that of others in the group. Responses were taken on a 9-point Likert-style scale with labels at the extreme and middle points. The preference survey asked users to choose the interface condition that best supported each of the five aspects as well as their overall preference. The full questionnaires and results are available in a report of this experiment (Martin et al., 2016).

Group	Performance 1	Performance 2	Performance 3	Performance 4
1	I2	I4	I1	I3
2	I1	I2	I3	I4
3	I3	I1	I4	I2
4	I4	I3	I2	I1

Table 3: Schedule for the quartet experiment showing the counter-balanced ordering of interfaces. Each group participated in one session including an induction, the four performances, and a debrief interview.

The performance survey data was analysed using an aligned rank transform (ART) procedure followed by a two-way mixed-effects ANOVA to assess significance. This procedure was chosen to account for ordinal data while allowing for the factorial and within-groups design (Wobbrock et al., 2011). The preference survey data was analysed with Chi-squared tests to determine how significantly the distribution of responses had deviated from chance. Finally, the duration of performances and occurences of ensemble-mediation events was modelled using a two-way within-groups ANOVA procedure. In this study, these analyses revealed that the performance length varied significantly with the experimental conditions. The performance surveys showed which particular aspects of the performance had been affected by the different approaches to ensemble mediation. The application of the objective duration measure turned out to be a critical element in this performance. While the survey responses supported the effectiveness of one main effect, the session duration data supported the other (Martin et al., 2016). This suggests that performer evaluations alone do not tell the whole story about ensemble improvisations. In this case, the controlled environment of the rehearsal-as-research session made it possible to measure this objective data and expose this finding. The results of the study led us to design a new interface to be compared with the both-systems condition and this was evaluated in a follow-up study using the same procedure. Several participants of the study were later invited to perform with the refined systems in concerts and public improvisation workshops.

4 Conclusion

Rehearsals of multiple free-improvised performances present a natural, yet controlled, environment for studying collaborative musical interaction. These sessions permit the application of typical HCI research methods, such as factorial studies with multiple experimental conditions spread over the improvisations, while preserving genuine artistic exploration. Data collection through questionnaires or instrumentation of musical interfaces can be accomplished without disrupting the participants' musical process. In this chapter we have described how this approach of rehearsal-as-research can be applied in formal evaluations of DMIs and NIMEs while retaining natural musical exploration leading to performance.

In our studies, these tools have been used to understand new ensemble-focussed NIMEs. We have been able to compare multiple versions of a musical interface and ensemble director agent systems. The use of both subjective (performer ratings) and objective (improvisation duration and interaction data)

measures have given us multiple viewpoints on how the interfaces were used in performance, and the benefits that features of NIMEs can offer to performers. It has been especially notable when different experimental effects are observed using these different measures, this supports the idea that NIMEs should be evaluated holistically, using all measures that researchers have at their disposal. We propose that controlled rehearsal sessions are an ideal environment to conduct such experiments.

Future rehearsal-as-research studies could extend the ideas presented in this chapter. It may be possible to create more general survey instruments that can be applied to many collaborative interface designs and continuous rating scales may be more appropriate than the ordinal scales we have used. The relationship between improvisation duration, engagement, and the performer's experience has yet to be fully explored. There is also much scope to compare other objective measures, such as motion capture or biometric data, with subjective responses.

Finally, it is notable that even though systematic research may seem to constrain artistic expression, feedback from our participants has emphasised that these were enjoyable and rewarding artistic experiences. Rehearsals-as-research allowed us to examine musical human-computer interaction in great detail, and they were also artistic research sessions resulting in concerts and the development of ongoing musical practices.

References

Barbosa, J., Malloch, J., Wanderley, M., & Huot, S. (2015). What does 'evaluation' mean for the nime community? In E. Berdahl & J. Allison (Eds.), *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 156–161). Baton Rouge, LA: Louisiana State University. Retrieved from `http://www.nime.org/proceedings/2015/nime2015_301.pdf`

Bradley, J. V. (1958). Complete counterbalancing of immediate sequential effects in a latin square design. *Journal of the American Statistical Association*, *53*, 525–528. doi:10.2307/2281872

Cahn, W. L. (2005). *Creative Music Making.* New York, NY: Routledge.

Corbin, J., & Strauss, A. (2008). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory.* London: SAGE Publications. doi:10.4135/9781452230153

Costello, B., Muller, L., Amitani, S., & Edmonds, E. (2005). Understanding the experience of interactive art: Iamascope in Beta_space. In *Proceedings of the Second Australasian Conference on Interactive Entertainment* (pp. 49–56). Sydney: Creativity & Cognition Studios Press. Retrieved from `http://dl.acm.org/citation.cfm?id=1109180.1109188`

Csikszentmihalyi, M. (1991). *Flow: The Psychology of Optimal Experience.* New York, NY: Harper Collins.

Eisenberg, J., & Thompson, W. F. (2003). A matter of taste: Evaluating improvised music. *Creativity Research Journal*, *15*(2–3), 287–296. doi:10.1080/10400419.2003.9651421

Fiebrink, R., Cook, P. R., & Trueman, D. (2011). Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 147–156). New York, NY: ACM. doi:10.1145/1978942.1978965

Gardner, H. J., & Martin, M. A. (2007). Analyzing ordinal scales in studies of virtual environments: Likert or lump it! *Presence: Teleoperators and Virtual Environments*, *16*(4), 439–446. doi:10.1162/pres.16.4.439

Gelineck, S., & Serafin, S. (2012). Longitudinal evaluation of the integration of digital musical instruments into existing compositional work processes. *Journal of New Music Research*, *41*(3), 259–276. doi:10.1080/09298215.2012.697174

Gurevich, M., Marquez-Borbon, A., & Stapleton, P. (2012). Playing with constraints: Stylistic variation with a simple electronic instrument. *Computer Music Journal*, *36*(1), 23–41. doi:10.1162/COMJ_a_00103

Hayden, S., & Windsor, L. (2007). Collaboration and the composer: Case studies from the end of the 20th century. *Tempo*, *61*(240), 28–39. doi:10.1017/S0040298207000113

Jordà, S., & Mealla, S. (2014). A methodological framework for teaching, evaluating and informing nime design with a focus on mapping and expressiveness. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 233–238). London: Goldsmiths, University of London. Retrieved from `http://www.nime.org/proceedings/2014/nime2014_472.pdf`

Krüger, S. (2008). *Ethnography in the Performing Arts: A Student Guide.* York: The Higher Education Academy.

Lazar, J., Feng, J., & Hochheiser, H. (2010). *Research Methods in Human-Computer Interaction.* West Sussex: John Wiley & Sons.

Martin, C., & Gardner, H. (2016). A percussion-focussed approach to preserving touch-screen improvisation. In D. England, T. Schiphorst, & N. Bryan-Kinns (Eds.), *Curating the Digital: Spaces for Art and Interaction.* Cham: Springer. doi:10.1007/978-3-319-28722-5_5

Martin, C., Gardner, H., & Swift, B. (2014). Exploring percussive gesture on iPads with Ensemble Metatone. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1025–1028). New York, NY: ACM. doi:10.1145/2556288.2557226

Martin, C., Gardner, H., Swift, B., & Martin, M. (2015). Music of 18 performances: Evaluating apps and agents with free improvisation. In I. Stevenson (Ed.), *Proceedings of the Australasian Computer Music Conference* (pp. 85–94). Fitzroy: Australasian Computer Music Association. Retrieved from `http://hdl.handle.net/1885/95205`

Martin, C., Gardner, H., Swift, B., & Martin, M. (2016). Intelligent agents and networked buttons improve free-improvised ensemble music-making on touch-screens. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2295–2306). New York, NY: ACM. doi:10.1145/2858036.2858269

Neustaedter, C., & Sengers, P. (2012). Autobiographical design in HCI research: Designing and learning through use-it-yourself. In *Proceedings of the Designing Interactive Systems Conference* (pp. 514–523). New York, NY: ACM. doi:10.1145/2317956.2318034

O'Modhrain, S. (2011). A framework for the evaluation of digital musical instruments. *Computer Music Journal*, *35*(1), 28–42. doi:10.1162/COMJ_a_00038

Shneiderman, B. (2007). Creativity support tools: Accelerating discovery and innovation. *Communications of the ACM*, *50*(12), 20–32. doi:10.1145/1323688.1323689

Skogstad, S. A., Nymoen, K., & Høvin, M. (2011). Comparing inertial and optical mocap technologies for synthesis control. In *Proceedings of the 8th Sound and Music Computing Conference*. Barcelona: SMC Network. Retrieved from `http://smcnetwork.org/system/files/smc2011_submission_124.pdf`

Smith, D. T. (2009). Development and validation of a rating scale for wind jazz improvisation performance. *Journal of Research in Music Education*, *57*(3), 217–235. doi:10.1177/0022429409343549

Stowell, D., Robertson, A., Bryan-Kinns, N., & Plumbley, M. D. (2009). Evaluation of live human-computer music-making: Quantitative and qualitative approaches. *International Journal of Human-Computer Studies*, *67*(11), 960–975. doi:10.1016/j.ijhcs.2009.05.007

Unander-Scharin, C., Unander-Scharin, Å., & Höök, K. (2014). The Vocal Chorder: Empowering opera singers with a large interactive instrument. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1001–1010). New York, NY: ACM. doi:10.1145/2556288.2557050

Wanderley, M. M., & Orio, N. (2002). Evaluation of input devices for musical expression: Borrowing tools from HCI. *Computer Music Journal*, *26*(3), 62–76. doi:10.1162/014892602320582981

Williams, E. J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry*, *2*(2), 149–168. doi:10.1071/CH9490149

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 143–146). New York, NY: ACM. doi:10.1145/1978942.1978963

Wobbrock, J. O., & Kay, M. (2016). Nonparametric statistics in human–computer interaction. In J. Robertson & M. Kaptein (Eds.), *Modern Statistical Methods for HCI* (pp. 135–170). Cham: Springer. doi:10.1007/978-3-319-26633-6_7

Xambó, A., Hornecker, E., Marshall, P., Jordà, S., Dobbyn, C., & Laney, R. (2013). Let's jam the Reactable: Peer learning during musical improvisation with a tabletop tangible interface. *ACM Transactions on Computer-Human Interaction*, *20*(6), 36:1–36:34. doi:10.1145/2530541