What does " Evaluation " mean for the NIME community?

Jeronimo Barbosa, Joseph Malloch, Marcelo M. Wanderley, Stéphane Huot

What does "Evaluation" mean for the NIME community?

Jeronimo Barbosa
IDMIL, CIRMMT,
McGill University
jeronimo.costa@mail.mcgill.ca

Joseph Malloch
Université Paris-Sud,
CNRS (LRI), Inria Saclay
malloch@lri.fr

Marcelo M. Wanderley
IDMIL, CIRMMT,
McGill University
marcelo.wanderley@mcgill.ca

Stéphane Huot
Inria Lille
stephane.huot@inria.fr

ABSTRACT

Evaluation has been suggested to be one of the main trends in current NIME research. However, the meaning of the term for the community may not be as clear as it seems. In order to explore this issue, we have analyzed all papers and posters published in the proceedings of the NIME conference from 2012 to 2014. For each publication that explicitly mentioned the term "evaluation", we looked for: a) What targets and stakeholders were considered? b) What goals were set? c) What criteria were used? d) What methods were used? e) How long did the evaluation last? Results show different understandings of evaluation, with little consistency regarding the usage of the word. Surprisingly in some cases, not even basic information such as goal, criteria and methods were provided. In this paper, we attempt to provide an idea of what "evaluation" means for the NIME community, pushing the discussion towards how could we make a better use of evaluation on NIME design and what criteria should be used regarding each goal.

Author Keywords

Evaluation, Digital Musical Instruments, Metareview, Methodology, Terminology

ACM Classification

A.1 [Introductory and Survey]; H.5.5 [Information Interfaces and Presentation] Sound and Music Computing — Methodologies and techniques; H.5.2 Information Interfaces and Presentation (e.g., HCI): User Interfaces - Evaluation / methodology.

1. INTRODUCTION

> "*In essence, while the search for solid and grounded design and evaluation frameworks is one of the main trends in current NIME research, general and formal methods that go beyond specific use cases have probably not yet emerged. Will these be the El Dorado or the Holy Grail of NIME research?*" [14]

The paragraph above, quoted from Jordà and Mealla's paper published at NIME 2014, illustrates the high expectations often associated with evaluation in NIME research today. This growing interest can also be statistically observed in the conference proceedings. Based on previous works [18, 2], we have performed text analysis on the proceedings of the three last NIME conferences (from 2012 to 2014) and tracked how many publications reported to have performed an "evaluation". Considering oral and posters presentations only: In 2012, 34% of the publications that proposed a NIME evaluated the proposed devices; In 2014, the number has increased to 49% of the publications, as shown in Table 1.

Table 1: Number of "*evaluations*" reported in NIME publications from 2012 to 2014, based on [18, 2].

Evaluates?	2012	2013	2014
Not applicable	24	41	56
No	39	35	41
Yes	20	29	40
Total	34%	45%	49%

However, as the number of evaluations increases, it appears that the meaning of "*evaluation*" in the context of NIME or digital musical instruments (DMIs) may not be as evident as it seems. Initial analyses of the content of evaluation-related papers in NIME literature show us that there are different understandings of the meaning of the term "evaluation". It is common to find papers that use the term to denote the process of collecting feedback from users in order to improve a prototype (e.g., publication 14#A#48 in our corpus[1]). It is also common to find others that use the term to assess the suitability of existing devices for certain tasks [20], or to compare different devices using common characteristics [4]. Describing emerging interaction patterns when using the devices may also be found (e.g., publication 13#O#66). And all in all, these different objectives are all hidden behind the same general term of "evaluation".

Furthermore, there are other complicating factors. As pointed out by [15, 16], there are several stakeholders that might be involved in the design of DMIs and the requirements of one may not intersect those of another. Thus, criteria considered as important for one stakeholder (e.g., playability for the performer [13]) might not be as important for another one (e.g., the audience). In addition, depending on the stakeholder and the goal specified for the evaluation, the time window chosen [10] and the stakeholder's expertise with DMIs [8] might also impact the results. In the case of acoustic instruments, for instance, the criteria for evaluat-

NIME'15, May 31-June 3, 2015, Louisiana State Univ., Baton Rouge, LA.

[1]The identifier follows the format YY#F#ID, where YY denotes the year of publication, F indicates if the publication is a paper ('A') or a poster ('O'), and ID indicates the order in which it was analyzed. The collected data is available at `http://idmil.org/pub/data/dmi_evaluation_nime2012-2014.xlsx`

ing the suitability of a guitar for a beginner might not be the same as for a trained musician.

In this exploratory research, we aim to give insights into how the term evaluation has been more commonly employed in the NIME literature. For this, we have analyzed the proceedings of NIME conference from 2012 to 2014, looking for: a) the most common targets and stakeholders involved in the evaluation; b) the most common goals; c) the most common criteria; d) the most common techniques/methods used for the evaluation; e) the duration of the evaluation.

2. BACKGROUND

The *role of evaluation* has been extensively discussed in the context of HCI [3, 9], Creativity Support Tools [17] and acoustic musical instruments [5]. In the context of DMIs and NIME, discussions are just starting [12] (see, for example, the Workshop on Practice-Based Research in New Interfaces for Musical Expression in NIME 2014[2]). Yet, it is possible to find in literature a large variety of approaches for evaluating DMIs. Here, we provide a brief overview.

Building upon HCI research on the evaluation of 2D input devices [6] and on the comparison of input devices for direct timbre manipulation [19], Wanderley et al. proposed to adapt this knowledge to the context of DMIs [20]. They proposed musical tasks that could allow to quantitatively compare how input controllers perform when considering a certain musical goal.

A different approach, based on the qualitative tradition, was proposed by Stowell and al. [18]. Instead of quantitative comparison, the authors focused on investigating subjective qualities inherent to the musical experience, such as enjoyment, expressivity and perceived affordances. For this, they used semi-structured interviews to collect data with performers, followed by *Discourse Analysis* on the transcribed speech.

Neither do these approaches consider the impact of time on the evaluation (i.e., as time goes by, the more musicians are likely to play and practice with their instruments, and perhaps become better able to express themselves with it). Usually evaluation happens throughout a few sessions, with almost no time interval between them. This issue is addressed by Hunt and Kirk [10]. In their work, they presented an AB Testing based approach (which mixed quantitative and qualitative characteristics) used to evaluate mapping strategies for 3 different DMIs over a period of time.

Another time-related issue is the notion of player's expertise, analyzed both quantitatively and qualitatively by [8], and its perception by the audience, as discussed by [7]. Considering the latter, Barbosa et al. presented an evaluation approach that focuses upon the Audience's perspective [2]. Here, the goal was to assess the participants' comprehension about five components of the instrument, by using an on-line questionnaire.

3. RESEARCH QUESTIONS

In order to assess the context of usage of the term "evaluation" by the NIME community, we have set the following research questions:

Question 1: Which targets are evaluated? For example, the whole DMI, its input module, the mapping module, the output module, or the feedback provided by the DMI. In this process, which stakeholders are usually considered?

Question 2: What are the most common goals for DMI evaluation?

Question 3: What criteria are commonly used for evaluating DMIs?

Question 4: What approaches are used for the evaluation (i.e., quantitative, quantitative or both)? What are the most commonly employed techniques/methods?

Question 5: How long do DMI evaluations last on average (i.e., a single session/experiment, or over time)?

4. METHODOLOGY

We have analyzed all papers and posters available on-line for the last three proceedings of the NIME conference (2012, 2013, 2014). Demos were not considered.

As mentioned before, for each publication we assigned a unique identifier in order to provide practical examples. Then, we collected the following data:

Format: How the work was published (i.e., as oral presentation or poster);

Target: A summary of the main contribution of the publication, using as much as possible the authors' own terminology;

Target category: Classified as: a) DMI; b) Input; c) Mapping; d) Output; e) Feedback; f) Performance. Any other kind of target was classified as "None" as they are outside the scope of this work. One publication can have multiple target categories;

Includes evaluation: Whether or not the authors evaluated the target. For this, we only considered publications in which authors directly used the term "evaluation". If they did not use the term, the publication was not considered.

For those that did evaluate a target (our main interest in this work), we also collected the following data:

Perspective evaluated – According to the stakeholders involved in the design of DMIs [15, 16], what perspective(s) were considered? One publication could address multiple perspectives;

Goal of the evaluation – Here, we tried to use as much as possible the authors' own terminology. However, whenever the name of the target (i.e., the name of the instrument of technology proposed) was mentioned we replaced it with the general term "system";

Criteria considered – Here again, we tried to use as much as possible the authors' own terminology;

Approach – What was the approach chosen towards the evaluation (i.e., quantitative, qualitative, or both)?

Duration – Was the evaluation performed only in a single session/experiment? Or did it occur over time? We did not record specific time durations – if the evaluation lasted several days, weeks, or months, it was categorized as "over time";

Methods – What methods were used to evaluate the target? Here, we tried to add keywords related to the methods employed, with as much details as provided by the authors.

The collected data was gathered in a spreadsheet. For the objective fields (i.e., "Target category", "Evaluates or not", "Perspective evaluated", "Approach", and "Duration") we counted the number of occurrences in order to generate

[2]http://www.creativityandcognition.com/NIMEWorkshop/

tables and graphs. For the more subjective fields (i.e., "Goal of the evaluation", "Criteria considered"), we initially have employed the word cloud technique as provided by Wordle[3]. In order to extract more details from this data, we did further qualitative analysis. This process is described in the next section.

5. RESULTS

From 325 papers analyzed in total, 204 papers were suitable for our purposes (i.e., had DMIs or one of its modules as target). Of these, 89 papers (45 oral presentations & 44 posters) used the term "evaluation" with regards to their target. This result is illustrated in Figure 1. The spreadsheet containing all collected and analyzed data is available on-line[4].

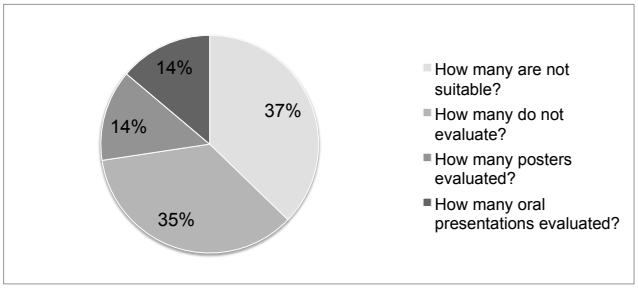

Figure 1: Number of reported evaluations according to format (i.e., oral presentations & posters) in NIME proceedings of 2012, 2013 and 2014.

In this section we present our results according to each of our five research questions.

5.1 Question 1: Evaluated Target

This question regards the most common targets and stakeholders considered in the evaluation. The most common target was the whole DMI (60 publications) and the most common perspective considered was the Performer (52 publications). Results are summarized in Figures 2 and 3 respectively. In both cases, the classification was non-exclusive (i.e., the same publication could assess different targets and perspectives at the same time).

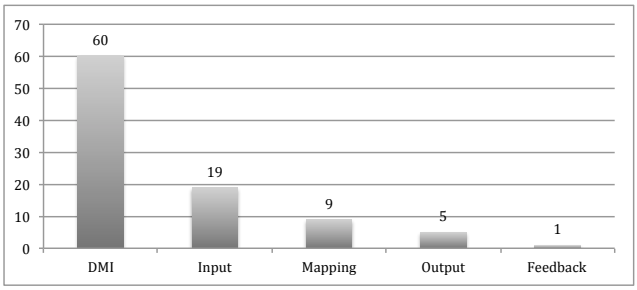

Figure 2: Most common targets used in the evaluations performed.

Regarding the analysis presented in Figure 2, it is interesting to note that the number of mapping strategies and output proposed – and consequently evaluated – are low. This might be due to the fact the conference is more focused on "interfaces", a notion more related to the input module, however these numbers are interesting if we consider that mapping has been stated to play a crucial role in the design of new DMIs [11].

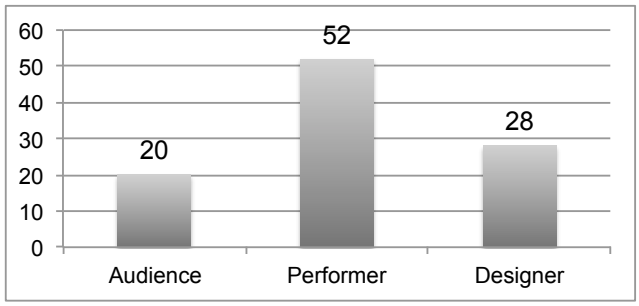

Figure 3: Perspectives considered in the evaluations performed.

As it can be seen in Figure 3, the predominance of the *performer's* perspective support the claim that it is the most important stakeholder in musical performance contexts [4]. The *designer's* perspective is commonly related to the technical aspects of the proposed system (e.g., how effective is a machine learning technique such as in 13#A#47 and 14#O#16, or the frequency response of the sound output such as in 14#O#85). The *audience's* perspective, which is related to how the audience perceives the proposed system (e.g., 13#A#12 and 13#A#11), comes in the last position. These results may indicate that the NIME community tends to under-consider the audience in the design of DMIs, or at least for their evaluation. However, since we consider only papers that report on an evaluation, further investigation is necessary.

5.2 Question 2: Goals of the Evaluation

This question addresses the goals the authors aimed with the evaluation. As it can be seen in a word cloud based on collected data (see Figure 4), a large variety of terms were employed.

This led us to investigate qualitatively the nature of the chosen goals. We came up with six non-exclusive categories related to the general purpose of the evaluation, defined as follows:

A Investigate how the target performs according to specific pre-defined criteria (e.g., 13#A#7);

B Collect feedback in order to improve the target (e.g., 14#A#48);

C Compare the target with similar systems as baseline (e.g., 14#O#39);

D Verify specific hypothesis about the evaluated target (e.g., 14#O#128);

E Describe interesting (emerging) behaviors while testing the target (e.g., 13#O#66);

F Not specified or **different** from the previous (e.g., 13#O#90).

The goals were then classified according to these categories. The same thing was done separately for each stakeholder perspective. The results are presented in Table 2.

We note that 'A' is the most common goal used for all stakeholders. However, it is very common to find goals that are combinations of the above-mentioned categories (e.g., investigate specific predefined criteria and then use this result to compare the target to similar systems, such as in

[3] http://www.wordle.net/

[4] http://idmil.org/pub/data/dmi_evaluation_nime2012-2014.xlsx

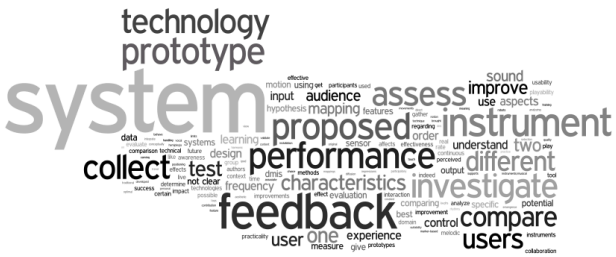

Figure 4: The most common goals found in the reported evaluations. Terms are scaled relative to their frequency in the analyzed text.

Table 2: Goals classified according to the six non-exclusive categories proposed. Results also presented for each stakeholder perspective.

Goal	Occurrences	Perf.	Aud.	Des.
A	47	21	10	23
B	18	15	4	1
C	23	12	4	10
D	12	8	3	2
E	25	20	8	4
F	5	3	1	0

13#A#27 and 14#A#10). We also highlight that the same publication can have multiple stakeholders. Finally, it is interesting to note how diverse are the goals hidden behind the term "evaluation" on NIME literature.

5.3 Question 3: Criteria

This question involves the most common criteria used for the evaluation. At first, for each stakeholder perspective, we have built a word cloud based on the collected data. As shown in Table 5, a large amount of publications omitted this information, and the term "not clear" was very large, hiding the rest of our data. This motivated us to remove it from the word cloud, as presented in Figures 5, 7, 6. At the same time, the fact seems representative, as it illustrates the lack of consistency regarding evaluation criteria in the NIME community.

Figure 5: The most common criteria according to the Performer's perspective ("not clear" excluded).

Considering the Performer's perspective (Figure 5), we can note that some terms emerge despite the large diversity. Most part of them were already addressed in the literature, such as *'engagement'* [21], *'effectiveness'* [13], and *'expressiveness'* [1]. However, these criteria are still subjective in the context of DMIs and there is no consensus on how to measure or analyze them. Considering the Designer's perspective (Figure 6), objective terms like *'precision'*, and *'latency'* emerged. Considering the Audience's perspective (Figure 7), there was no significant difference regarding the usage of the terms (i.e., terms such as *'focus'* and *'intention comprehension'* were mentioned only twice). In all three cases, the large diversity of terms should be highlighted.

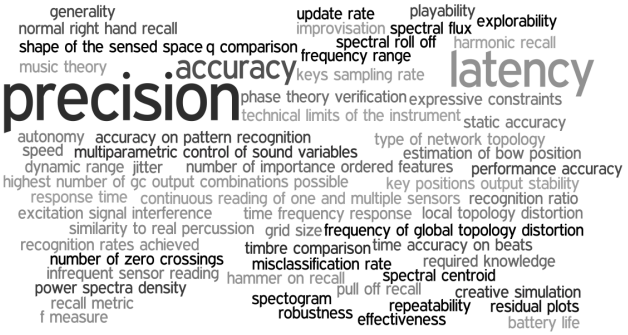

Figure 6: The most common criteria according to the Designer's perspective ("not clear" excluded).

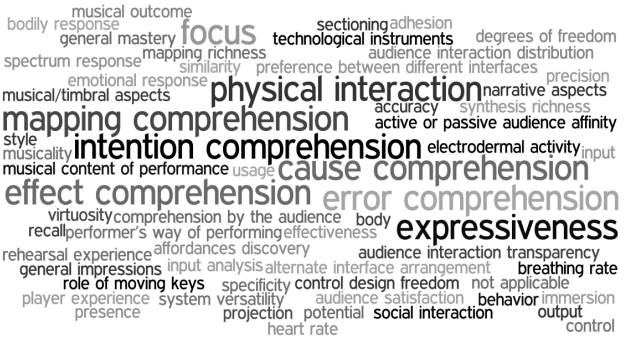

Figure 7: The most common criteria according to the Audience's perspective ("not clear" excluded).

In order to investigate the nature of these criteria, we further performed a qualitative investigation of the data. We classified the criteria as *objective* (i.e., there is a clear understanding on how to measure these criteria), *subjective* (i.e., there is no clear understanding on how to measure these criteria) or *both*. The result is summarized in Table 3. Subjective criteria and "not clear" were the most commonly found categories. Once again, the lack of consistency regarding criteria is apparent.

Table 3: Criteria classified in subjective, objective, both or "not clear" (i.e., not able to determine).

Criteria	Number of occurrences
Subjective	29
Objective	27
Both	7
Not clear	25

5.4 Question 4: Approach

This question investigates the approach chosen (i.e., *qualitative*, *quantitative* or *both*) and the most common techniques

or methods employed. Table 4 summarizes the results. Regarding the techniques/methods, once again, we have built word clouds based on the collected data. The results are presented in Figures 8, 9, 10.

Table 4: Results regarding the evaluation approaches chosen.

Approach	Occurrences	Perf.	Aud.	Des.
Qualitative	28	23	8	1
Quantitative	36	9	5	23
Both	13	11	7	1
Not clear	12	9	0	3

Figure 8: The most common methods/techniques employed according to the Performer's perspective

Although the quantitative approach was the most commonly used, we found less variety among the qualitative related methods (e.g., questionnaire, and interviews). It is also interesting to note that qualitative approaches were more common when evaluating the Performer's perspective. On the other hand, quantitative approaches were preferred when evaluating the Designer's perspective.

5.5 Question 5: Duration

The last question assessed the duration of the evaluation (i.e., single session/experiment, or over time). Regarding this, most part of the evaluation (66%) seems to be performed in a single session. Evaluations over time occurred in some cases (19%), but they were much less common. The remaining (15%) were not clear about the subject.

6. DISCUSSION

Surprisingly, we can notice a significant number of publications that employ the term "evaluation" without giving any detail about criteria (31%) or methods (19%). In some rare cases (4%), even the goal is not clearly stated. This result is shown in Table 5.

Table 5: Quantifying omitted information.

Total of evaluations	89
Do not inform which methods were used	17
Do not inform which criteria were used	28
Do not inform goals for the evaluation	4

We believe that this is the most important issue and that it deserves attention – especially from NIME reviewers. It is completely acceptable to not evaluate the DMIs we create, however, if one wants to "evaluate" something, it is essential to provide basic information such as the goal of the evaluation, how it was performed (e.g., what methods and criteria were used) and what results were achieved. Otherwise, the information provided is not meaningful to the community. More importantly, this is harming the validity of the evaluation and prevents its replicability [9].

Figure 9: The most common methods/techniques employed according to the Audience's perspective.

Figure 10: The most common methods/techniques employed according to the Designer's perspective.

In addition, these results provide us a clearer view of different approaches towards "evaluation" by the NIME community. The data allowed us to picture the profile of a typical evaluation (i.e., evaluates the DMI according to the performer's perspective, in a single qualitative experiment), for which literature offers several different possible approaches. However, how can we address the remaining cases, such as the *audience* perspective, or evaluation over time?

Another interesting issue concerns the lack of agreement about criteria and goals used for evaluation. This provides us with some interesting questions for future research, for example: considering a given goal and stakeholder perspective, would it be possible to find a common understanding about the most important criteria? Will it be possible for us to find a consensus approach to analysing subjective criteria, such as playability, engagement, and expressiveness?

7. PROBLEMS & LIMITATIONS

During our analysis, we faced some issues, the most relevant ones being that:

- It was sometimes hard to classify a target according the categories we were looking for (i.e., DMI, Input, Mapping, Output, and Feedback), such as in 14#A#48;
- The difference between evaluation and experiment (in which hypotheses needed to be demonstrated) is not clear, such as in 14#A#49;
- In the process of extracting subjective fields (i.e., the goals) of the evaluations, some bias may have been introduced since the data we were looking for were not always clearly described (e.g., difficult to say what method/goal/criteria the authors used). We tried to minimize this bias by using the authors' own terminology as much as possible. This problem will be difficult to solve, as it is related to the way the evaluations were reported in the publications;
- Considering publications with multiple stakeholders (such as 12#A#23), we did not differentiate which methods and criteria were set for each stakeholder. This fact introduced some noise to our cross-question analysis (i.e., Questions 3 and 4, in which we have created one word cloud for each stakeholder perspective).

In addition, we stress that the results presented and discussed in this paper are still preliminary. Going forward, further years of NIME proceedings should be considered, as well as other relevant venues, such as the ICMC.

8. CONCLUSION

We have investigated how the term "*evaluation*" has been employed in the NIME literature. The results give us a better idea of: a) the most common targets and stakeholders considered during the evaluation; b) the most common goals set; c) the most common criteria set; d) the most common techniques/methods used for the evaluation; and e) how long the evaluation lasts.

In case one is interested in evaluation within a certain context (e.g., what would be the most used techniques for evaluating mapping considering audience's perspective?), we highlight that cross-relating results (like we did in Questions 3 and 4) can provide a richer analysis scenario.

Finally, although "*there is no one-size-fits-all solution to evaluating DMIs*" [16] and more precisely "*the choice of evaluation methodology - if any - must arise from and be appropriate for the actual problem or research question under consideration*" [9], this work may help us to assess different evaluation profiles in order to find the most suitable techniques considering different goals, criteria and stakeholder's perspectives. Thus, we hope to contribute by going beyond discussing whether the NIME community should or should not evaluate their creations, focusing instead upon how could we make better use of evaluation and what criteria should be used for the evaluation.

9. ACKNOWLEDGMENTS

The authors would like to thank: The NSERC Discovery Grant, Inria/FRQNT/the Embassy of France in Canada (Équipe de Recherche Associée *MIDWAY*), for the funding; the anonymous reviewers, for their valuable comments and suggestions; and Johnty Wang, Carolina Medeiros, and John Sullivan.

10. REFERENCES

[1] D. Arfib, J. M. Couturier, and L. Kessous. Expressiveness and digital musical instrument design. *Journal of New Music Research*, 34(1):125–136, June 2005.

[2] J. Barbosa, F. Calegario, V. Teichrieb, G. Ramalho, and P. McGlynn. Considering Audience's View Towards an Evaluation Methodology for Digital Musical Instruments. In *NIME '12 Proceedings of the 2012 conference on New interfaces for musical expression*, Ann Arbor, Michigan, 2012.

[3] L. Barkhuus and J. a. Rode. From Mice to Men - 24 Years of Evaluation in CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1–16, 2007.

[4] D. Birnbaum, R. Fiebrink, J. W. Malloch, and M. M. Wanderley. Towards a dimension space for musical devices. In *NIME '05 Proceedings of the 2005 conference on New interfaces for musical expression*, pages 192–195, Vancouver, BC, Canada, 2005.

[5] D. M. Campbell. Evaluating musical instruments. *Physics Today*, 67(4):35–40, Apr. 2014.

[6] S. K. Card, J. D. Mackinlay, and G. G. Robertson. A morphological analysis of the design space of input devices. *ACM Transactions on Information Systems*, 9(2):99–122, Apr. 1991.

[7] A. Fyans, M. Gurevich, and P. Stapleton. Examining the spectator experience. In *NIME '10 Proceedings of the 2010 conference on New interfaces for musical expression*, pages 1–4, 2010.

[8] M. Ghamsari, A. Pras, and M. M. Wanderley. Combining musical tasks and improvisation in evaluating novel Digital Musical Instruments. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research*, pages 506–515, Marseille, France, 2013.

[9] S. Greenberg and B. Buxton. Usability evaluation considered harmful (some of the time). In *CHI '08 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 111, 2008.

[10] A. Hunt and R. Kirk. Mapping strategies for musical performance. In Marcelo M. Wanderley and Marc Battier, editor, *Trends in Gestural Control of Music*, volume 21, pages 231–258. IRCAM, Centre Pompidou, Paris, France, 2000.

[11] A. Hunt, M. M. Wanderley, and M. Paradis. The Importance of Parameter Mapping in Electronic Instrument Design. In *NIME '02 Proceedings of the 2002 international conference on New interfaces for musical expression*, pages 429–440, Dublin, Ireland, Dec. 2003. Routledge.

[12] A. Johnston. Beyond Evaluation : Linking Practice and Theory in New Musical Interface Design. In *NIME '11 Proceedings of the 2011 conference on New interfaces for musical expression*, pages 280–283, Oslo, Norway, 2011.

[13] S. Jordà. *Digital lutherie: crafting musical computers for new musics performance and improvisation*. PhD thesis, Universitat Pompeu Fabra, 2005.

[14] S. Jordà and S. Mealla. A Methodological Framework for Teaching, Evaluating and Informing NIME Design with a Focus on Expressiveness and Mapping. In *NIME '14 Proceedings of the 2014 Conference on New Interfaces for Musical Expression*, pages 233–238, London, UK, 2014.

[15] T. Kvifte and A. Jensenius. Towards a Coherent Terminology and Model of Instrument Description and Design. In *NIME '06: Proceedings of the 2006 conference on New interfaces for musical expression*, pages 220–225, Paris, France, 2006.

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

[17] B. Shneiderman. Creativity Support Tools: Accelerating Discovery and Innovation. *Communications of the ACM*, 50(12):20–32, 2007.

[18] D. Stowell, A. Robertson, N. Bryan-Kinns, and M. D. Plumbley. Evaluation of live human-computer music-making: Quantitative and qualitative approaches. *International Journal of Human Computer Studies*, 67(11):960–975, 2009.

[19] R. Vertegaal. *An Evaluation of Input Devices for Timbre Space Navigation*. PhD thesis, University of Bradford, 1994.

[20] M. M. Wanderley and N. Orio. Evaluation of Input Devices for Musical Expression: Borrowing Tools from HCI. *Computer Music Journal*, 26:62–76, 2002.

[21] D. Wessel and M. Wright. Problems and Prospects for Intimate Musical Control of Computers. *Computer Music Journal*, 26(3):11–22, Sept. 2002.