-Figure 3 and 4 show that complexity scores computed by Musipletics generally correlate with those manually graded by musical educators. However, I can still observe some discrepancies. For instance, music pieces with grade 8 has a higher score than those with grade 9, and music pieces with the same grade, 10 have significantly different complexity scores. The illustration in Section 7 sort of explains these discrepancies. But I was wondering how acceptable these results are to musical experts. More data should be gathered to support their arguments about the usefulness of Musipletics, maybe conducting a post survey and sending out the computed results to musical experts for feedback.

**Explain the discrepancies and outliers a little more.**

~~-The complexity score computed by Musiplectics, to a large extent, relies on how well and objectively music experts rank the relative difficulty of individual musical components (e.g., notes, intervals, etc) in the first phase. However, it seems the current complexity settings in Musiplectics are only provided by a single musical expert, and the only description about the expert is that she favors Clarinet. Besides, no rationale is provided to justify the weights associated with each component, which may raise the question like “is it just based on the taste of the single expert?” I think it's necessary to recruit more musical experts and disclose more information~~

~~about the complexity settings for the sake of clarity and objectivity. In~~

~~Section 6.2, the authors mentioned that they were planning to survey about the~~

~~well-established complexity parameters and use statistically significant~~

~~consensus as new complexity parameters. It sounds more reasonable than their~~

~~current settings and I think it might be useful to conduct a survey on this.~~

**~~Point them to the evaluation and future work sections about this (sections 6.2 and 8.1) and explain further how we came up with our complexity settings. What we have is a proof of concept—it is OK to use a limited sample to parameterize the system.~~**

-Figure 3 and 4 are somewhat difficult to interpret, and I wonder whether they could be represented in a different manner, say measuring the coefficient of linear regression results and mean errors of linear regression.

**Perform these calculations and add them into the evaluation and/or discussion sections (sections 6.3 and 7.1).**

~~-Their evaluation demonstrates the usefulness of Musipletics in assessing the~~

~~complexity of music scores of a single instrument. More specifically, they~~

~~only utilized the beginner settings in their tests and did not evaluate or discuss how well Musipletics performs with complexity~~

~~settings on other levels of Clarinet. In addition, since Musipletics is~~

~~designed to be a general approach for any instrument, I wonder how easy to~~

~~adapt this model to other instruments and how accurate it is on other~~

~~instruments. It would be useful to have more thorough evaluation on Musipletics.~~

**~~Point the reviewers to our future work section about this and maybe run the calculations for our other settings (including results as an appendix).~~**

~~- In terms of evaluation, it is unclear whether the authors are measuring the~~

~~capability of OCR of the fitness of complexity score.~~

**~~We are measuring both separately to avoid this confusion. Explain this separation more clearly and why it is necessary.~~**

~~-I was disappointed that the model was defined in an ad-hoc way. Since there seems to be already a lot of music for which the difficulty has been ranked by experts, why not simply learn a model?~~

**~~Music is a very subjective area of human endeavor. Yes, much depends on individual opinions. Some “music experts” may think that a particular performer is a genius, while others may think he is an impostor. Need to clarify this point further. Point them to future work.~~**

~~The key contributions to previous work should be described in a better way. For example [11] build regression models from features for piano music. What features? I also did not understand why the authors’ work has a wider range of applicability. I doubt that there will be a universal model that applies to all music instruments. More details should also be reported about the features used by [15] and [20]. In its current form, it’s difficult for me to tell the key novelty.~~

**~~Provide a more in depth explanation of previous work. (copy from the thesis)~~**

~~-The range of complexity scores is huge, which makes them difficult to understand in practice. I also have the suspicion that the complexity is correlated with the length of pieces.~~

**~~They kind of are and aren’t. Perhaps compare the score to the length of piece (total seconds or measures)? Maybe explain this in the comments section.~~**

-I did find the weights confusing. In Section 5 it sounds like there are different weights for different skill levels, but I did not exactly understand the rationale behind this.

**Explain this further.**

-The model is described on a very abstract level. It be good to include a concrete example of how it is computed for a short piece of music.

**Include the worked out example from my defense as an appendix.**

-This work reminds me a lot of readability of text documents and famous readability formulas. <http://en.wikipedia.org/wiki/Readability> Similar work on source code readability: Raymond P. L. Buse, Westley Weimer: Learning a Metric for Code Readability. IEEE Trans. Software Eng. 36(4): 546-558 (2010).

**Read up on this and maybe list it as related work, but I probably won’t.**

~~-This may sound strange but parts reminded me also about Guitar Hero. I’d expect that they had also models to assess the difficulty of the individual songs/levels.~~

**~~Don’t do anything with this. It indeed sounds strange.~~**

~~-Missing words: “musiplectics this subjectivity”~~

**~~I think it should be “musiplectics embraces this subjectivity”. I’ll change it.~~**

-Figure 5 makes several assumption over the layout of the XML (e.g., one tag per line, same order). I would trust the diff numbers more if the XML format was unified first, e.g., with xmllint.

**Specify our assumptions outright before that experiment.**

-Revise the paper to be a case study as opposed to trying to argue the results generalize (i.e., discuss what you did and how it worked out).

**Change some of the language and tone it down more.**

-When you resubmit your paper in phase two of the review process, by Friday July 17, 2015, you must include a page that describes the changes you have made to the paper to address these comments. It would be very helpful if you used coloured text or some other mechanism to note all the changes made to the paper.

**Include this page listing our changes.**