Response to Reviewers

1. Overall Comments

Reviewer 1: I think the paper is better served by selecting a subset of these topics and focusing on them in more detail, mentioning the others, but leaving a discussion/analysis of them for another venue.

Reviewer 2: The authors have made a thorough effort to address reviewer comments, and the paper is much improved as a result. Most sections are satisfactory now.

Response: The authors appreciate the critique and guidance of the reviewers, thanks to which the paper has been substantially strengthened. With the reviewers' guidance we believe that we have fulfilled our intent to convey to the research community a holistic picture of two major topics: the design and evaluation of BlockPy. In line with this, the introduction section enumerates three contributions (design features/rationale, design issues, and evaluation) and identifies the relevant audience for each. The current version of the paper presents a balanced presentation with approximately 40% of the paper devoted to the first two design-related contributions and 33% of the paper devoted to the evaluation contribution.

2. Research Questions

Reviewer 2: - The authors claimed to have added explicit research questions, but I did not see any (unless the subsection headers of Section 5 were the questions), and the analysis can still feel a bit haphazard without guiding questions.

Response: The introduction now clearly lays out three guiding research *questions* (originally misstated as research hypotheses) at the end of Section 1. The evaluation is organized to be in line with these guiding questions. Each of Section 5.2-5.4 begins with a restatement of the research question addressed in that section.

Reviewer 2: Your evaluation results list is a nice addition, but it seems a bit vague. That the environment is "effective" and novices "can successfully solve problems" in it doesn't convey a concrete idea of what was learned. This seems somewhat symptomatic of the lack of clear research questions. I believe these 3 findings are supposed to relate to three guiding RQs. If so, tha should be stated explicitly, and the RQs should have clear answers by the end of the paper. If you claim your environment is effective, define a clear metric by which "effectiveness" can be answered assessed to answer the question. If that metric is hard to define, consider a more concrete RQ.

Response: The three findings are now clearly stated in response to the three research questions - each section of the revised Evaluation chapter begins with the relevant research questions. Each research question is measured more clearly, using the appropriate data sources (survey data, usage logs).

3. Improve Existing Evaluation

Reviewer 1: My feedback from the first version was that more time/emphasis should be placed on the evaluation section,

Reviewer 2: However, the evaluation, while improved, could still use some minor revisions to the way findings are presented (largely, this should not require additional analysis)

Response: We have elaborated the description of the evaluation and made other changes as noted in this section. With these changes the evaluation material (approximately pages 7-10) is now 33% of the paper in length while the design material (approximately pages 2-6), including features, rationale, and issues, is approximately 40% of the paper in length. We believe that this is a reasonable balance between the design and evaluation themes of the paper.

Reviewer 1: Further, I think the data that is presented should be more closely tied to the claims that the authors are making. For example, in section 5.2, I was expecting some data related to backing up the claim that data sciences was a compelling/effective/authentic context for CS/programming - instead, it presented data showing that the data science context was manageable - which is very different from the claims/justification given in the first portion of the paper. Additionally, the inclusion of the analysis on the appropriateness of the time to transition from BlockPy to Spyder seems like it is in the wrong section, as it does not relate to the data science context.

Response: Our three research questions were revised to more accurately capture the data presented and match our now more properly stated claims. In particular, we have reduced claims about the manageability of the data science context and the curriculum itself, moving this data to the methodology section of the evaluation; its role is now to simply give more guiding information about the participants' use of BlockPy. The new three research questions are now based more centrally around the data that was collected: did students' find the environment helpful (according to the survey data), were the transitioning scaffolds effective (according to the students' perceptions in the survey and the usage data), and was the automatic feedback sufficient (according to the students' perceptions in the survey data and the usage data analyzed using a static analyzer).

Reviewer 1: section 5.1 would benefit from subsections, as consecutive paragraphs are presenting different analyses looking at different aspects of BlockPy. Likewise 5.1 should be longer and provide more detail given the paper is largely about the design and evaluation of BlockPy - only one, relatively short paragraph (the first in 5.1) provides data about the effectiveness of the features discussed in section 3 of the paper (with the exception of the guided feedback, which is given a longer, more detailed analysis in 5.3).

Response: The previous Section 5.1 was divided into two main sub-sections with corresponding headers. These sub-sections relate to (1) student perceptions of the BlockPy features based on student comments reported on the survey (current Section 5.2), and (2) analysis of the scaffolding aspects of BlockPy as revealed through time spent in the blocks vs. text mode in the transition from blocks to Python and observations of issues experienced by students during the transition (current Section 5.3).

Reviewer 1: It is clear from the text that the authors of the paper are also the creators of BlockPy. While it makes sense for BlockPy's creators to write this paper, it also presents a conflict of interest, as the authors are biased towards the tool being evaluated. This comes up a few times in the paper in the wording and phrasing used. For example, in section 5.1 the authors say "students gave positive responses about BlockPy's interface" - this statement is made before data is presented. Later in the section, the authors report on negative responses (or at least frustrations). As much as possible, I encourage the authors to let the data speak for itself (i.e. include quotes and more information about the coding scheme used)., it will make the paper stronger.

Response: Quotations from responses by students on the survey were included throughout Section 5.2 and Section 5.4 in order to make the findings in these two section more objective. Additional information about the coding used in the analysis is presented in the expanded methodology section (Section 5.1).

Reviewer 1: Another examples of this comes from section 5.3 where it reads: "it is clear the students reacted negatively to the reduction in guidance" - did they react negatively to the reduction? or negatively to the lack guidence? The way it is currently worded suggests that learners reacted positively to the feedback and thus wanted more of it - while that may be true, it is not the only explanation and data is not presented in support of this framing. It is these slight word choices and phrasing that contribute to the sense of the authors and researcher were biased towards the environment in their analysis. A careful review of the text and removal of such phrasing would help make this read like a more even evaluation of BlockPy.

Response: The wording in question ("... reacted negatively...") was modified to more accurately portray students' reaction to the lack of guidance - without making claims about the change in guidance, but simply reporting that students wanted more guidance. We have made changes to text throughout these sections in an attempt to present a more objective evaluation.

Reviewer 1: In section 5.1 "more incentives and guidance should be given to direct students to pay attention or take advantage of the text interface" - was this a goal of BlockPy? My understanding was that the goal of BlockPy was to provided a scaffolded, authentic, data-science-driven introduction to programming, not to necessarily transition learners from blocks-to-text. If this was a goal, it should be stated more clearly in the design rationale.

Response: The wording of this section was clarified to make it clear that students should be better directed to take advantage of the text interface *when the question prompts them to*, not necessarily all the time. Although the authors do expect most curricula using BlockPy to transition students (and there is some language about this in Section 3), it is not a major goal. However, it is a desirable behavior for some curricula, such as the one used in this study.

Reviewer 2: The qualitative analysis of the free response data seems half-hearted. I would suggest either detailing a rigorous methodology and including supporting quotations throughout, or cutting this side of the analysis.

Response: The qualitative analysis was given more rigorous treatment, with supporting quotations from students included. In fact, the qualitative data is now the most heavily studied aspect of the evaluation. Additional information about the methodology was included in Section 5.1.

Reviewer 2: 5.1: "Our interpretation from this result is that students appreciated the guidance but wanted more from it." - I see no basis for this conclusion (given the data you've presented so far, anyway). Is this just an intuition from anecdotal evidence? If so, I'm not sure it belongs. Otherwise, please provide justification (quotations, data). I challenge this claim mainly because I have seen thorough analysis of student interviews on help features in programming that suggests many students prefer to work independently. These students may find the help unhelpful because they don't need/want it.

Response: This claim is removed from the paper in favor of presenting the data with less editorialization. Although, there is evidence in our data that students were unsatisfied with the feedback, often specifically because they wanted to know what their program was doing wrong. It may be helpful to recall that the feedback system is not meant as a "hint generation system", but instead the mechanism by which assignments were marked complete - therefore, the feedback was similar more to error messages than hints.

Reviewer 2: 5.2: "Completion rates and student survey responses indicate that the data science context... was manageable.": What do the assignment completion rates themselves tell us about the data science context? The assignments may have simply been easy. Or completion rates may have been higher without the data science context. This is not a hypothesis you can test with your data. Further, you claim survey responses support this conclusion but offer no analysis there. Without providing a rigorous qualitative analysis methodology, your references to the surveys feel anecdotal and rely far too much on the authors' interpretation.

Response: The claim of the context's manageability was removed from the paper. The data on completion rates and work session times was retained as purely descriptive material in the presentation of the methodology (Section 5.1).

4. Guided Feedback

Reviewer 2: 3 - Guided Feedback: This section has been improved with more details and an example of how an instructor might provide help, but I still feel that if it is the "most valuable pedagogical component of BlockPy" as you say, you should give a bit more detail (especially since you don't have a standalone paper on the feature to cite). Specifically, I think you need to make an argument for why this feedback is novel and a contribution. You should give some room in the literature review to other major feedback systems for programming and contrast what you've built to them. You can justify with educational literature why your feedback is theoretically valuable as you claim. Having a professor manually write a template to detect a missing for-loop and then showing a relevant message does not seem novel to me as currently described, as there are other systems that already do this, so I feel that justification is important.

Response: We have moderated the description of the guided feedback mechanism to present it as a “valuable” component of BlockPy rather than “the most valuable” component. The last paragraph of section 2.1 has been revised with more description of the advantages of Guided Feedback and prior work in the field, along with 3 new citations. Further, we have added some text to the first paragraph of Section 3 to better explain the contribution of the guided feedback within BlockPy: "We view the contribution of BlockPy as the synthesis of known features that are carefully integrated into an environment of novice learners, within the technical constraints of client-side execution." So it is not that we believe the guided feedback is a completely novel mechanism, but simply one that we think is a useful, that we have (somewhat) successfully implemented within the constraints of the environment, and bring to bear with other excellent features. We have also added some description (and an accompanying figure) to illustrate the instructor interface. Additional text was added to illustrate the value of the guided feedback mechanism because it can combine static analysis (having access to the student’s code) and dynamic analysis (having access to the variable trace and the program output). The future work section explains how this mechanism extended with additional forms of analysis and feedback to satisfy the need expressed by students for more and more informative feedback.

Reviewer 2: 3 - Guided Feedback: It is not completely clear how feedback is given or requested. Do students get feedback every time they compile and receive traditional syntax error feedback? This is what is sounds like, but explaining your example from the student's perspective, rather than the instructor's, might be beneficial.

Response: We have recast the description in the Guided Feedback sub-section of Section 3 to present some of the examples from the student’s viewpoint. We have also clarified that the feedback is given at “submission” (when the program is run).

Reviewer 2: 4.3: "Rivers and Koedinger explore other approaches that uses prior student submissions for each problem to suggest corrections to the user" - As written, this feels like a bit of a non-sequitur. I suggested this reference less for the data-driven hints (which are interesting and relevant, but perhaps to other sections), and more because of their work repairing uncompilable Python code as a preprocessing step. I'm having trouble finding the citation where this is mentioned, so perhaps I was misremembering. If so, feel free to cut it.

Response: To avoid a tangential issue this discussion and citation was removed.

Reviewer 2: 5.3: Do you have log data of how often the feedback was used/offered, in what circumstances, and whether it prompted changes in code? I would appreciate more insight here into how the students used, or failed to use it. What percentage of students received feedback?

Response: Feedback was administered after every run of the program, so they were constantly receiving feedback as they ran and tested their program.

5. Minor Corrections

Reviewer 1: Table 2 is not referenced in the text.

Response: Table 2 is now correctly referenced in Section 5.1.

Reviewer 1: Figure 5 - this same information was presented in the cited Matsuzawa paper (Figure 4, which uses a matrix where color saturation represents amount of time in Block v. Text). I find that presentation clearer. You might want to consider adopting that *representation.*

Response: Figure 5 was modified to match the reviewer's recommended presentation.

Reviewer 1: Snap is not Scratch's successor as you say in the text, instead it is a parallel effort to replicate Scratch and add new features.

Response: This error has been corrected in the third paragraph in Section 2.1.

Reviewer 1: The most valuable pedagogical component of BlockPy is its Guided Feedback system." - Do you have data to back up this claim?

Response: This statement was a bit overzealous as originally written, and therefore we decided to reduce the claim made in this statement - instead, we describe the Guided Feedback system as "another valuable component", rather than "*the most* valuable component".

Reviewer 1: "usually in the form of a Parsons' problem" - Parson's problems are probably not familiar to all readers, a description or citation should be added.

Response: Due to space constraints (and the need to expand the evaluation sections), we remove the reference to Parson's problems, and simply describe the problems as having starting code.

Reviewer 1: "the curriculum continued to the Spyder IDE, as a way to transition students into a more authentic settings" - this claim seems counter to the argument that BlockPy was designed to be authentic.

Response: The paragraph in question was clarified to avoid diminishing the authenticity of BlockPy.

Reviewer 2: 2.2: "has a similar goal to the CORGIS project" - You haven't previously mentioned/introduced CORGIS at this point.

Response: The introduction of CORGIS has been reworked to introduce CORGIS before its comparison to BRIDGES.

Reviewer 2: 5: "using a grounded theory approach" - It may be more appropriate to say "we performed qualitative, open coding on the free response questions, converging on a set of generalized tags...". Grounded theory is a much more holistic methodology, from data collection to analysis to theory building, and (unless you did these aspects as well), the open coding portion is just an associated technique. Moreover, there does not seem to be much use of these tags in the analysis at all, so why explain them at all? When you make an explanatory claim, that would be a good time to include supporting quotations discovered in your qualitative analysis.

Response: As previously described, the methodology was more correctly and clearly described in Section 5.1. Example quotations from student responses were added to further support the explained tags.

Reviewer 2: 5.1: "Consistent with the criticism of the automatic guidance" - what criticism? If this was earlier in the paper, I missed it.

Response: This misleading transition was removed, to better clarify the author's intent: the students' had criticisms of the guidance, as described in 5.4.

Reviewer 2: 5.1: "One effect observed... was that students were confused what the keywords of Python were": Your later analysis suggests that it was a fairly minor problem problem, so if you need to cut something, this section could probably be reduced.

Response: This section was shortened slightly to conserve space, although we retained mention of this keyword confusion because it was an observed issue, albeit a transitory one.

Reviewer 2: Fig 8 does not seem to be referenced anywhere. You should probably incorporate it into the analysis or cut it.

Response: Figure 8 was moved to Figure 5 as part of the revisions to the Evaluation chapter. It is now correctly referenced from 5.1 as part of the description of the students' use of the BlockPy system.

Reviewer 2: Table 3: I'm not sure why this is in the Feedback Quality section. What does it have to do with BlockPy's feedback? Are you suggesting it should include static analysis, or that it's feedback was unsuccessful due to these problems? The connection should be more explicit.

Response: Table 3 is now more fully explained in the last part of section 5.4 - the intent is to show that students commit many problems detectable by a static analyzer, and that such a system needs to be included in BlockPy.

Reviewer 2: Abstract: "computing curricula need[]..." and "We seek to improve [a] curriculum" (or "improve curricula").

Response: The typo has been corrected to "improve curricula".

Reviewer 2: Double-check the references for capitalization (e.g. [29], [36], [39]).

Response: The capitalization of [29], [36], and [39] has been corrected.