SUMMARY - ICSP Special Issue Review Form

Dear David. Please thank the reviewers for their excellent, detailed and insightful comments. We have changed much in this draft and we hope that it meets s their approval. Note that our replies to the reviewers are marked in yellow. Note that our comments are numbered [1], [2] etc and when multiple reviewers refer to the same issue, we fix it once and refer to that fix multiple times.

From the Editor: Dear Tim, There appear to be a variety of opinions about this paper from the reviewers. The reviewers seem to ask for more clear explanations around how you discuss uncertainty, the presentation of results early in the paper and their implications, how your experiments are reported as well as the benefits of your approach. There seems to be some questions about your evaluation of other methods (in particular process simulation).

[1] This current draft takes great care not to repeat the mistake of draft one which confused COCOMO with conventional process simulations. While both are models, we think the third reviewer's point is well taken: process simulations are characterized by elaborate visual environments that are missing in COCOMO. Instead, the current draft calls COCOMO a model from which we can make estimates about project details.

Please utilize the suggested revisions by the reviewers to improve your paper. In particular, it would improve the understandability of the paper if the method were described more fully and clearly.

[2] Following your direction, what we've done here is separate that description into a high-level summary of the algorithm in section 4 (p8) and a technical appendix (at the end). Further, the experiment description is now greatly clarified (see pages 10..13). This greatly clarifies and simplifies the exposition.

[3] Also, the introduction is not substantially rewritten, based on a suggestion from reviewer 3 who pointed out that we were conflating decision making with reducing the tuning variance. The current introduction clarifies that point.

As far as your question about page limits, thank you for being a "good little author". Let me know how many more pages you would require. I would say that it would be better to invest extra space in explaining the process implications of your analysis and results.

[4] At your excellent suggestion, we've moved and expanded the implications section into two pages and have moved these to the front of the paper (see p 4,5). Also, we have buried the AI technical details out the back.

It would probably be less interesting to our audience for you to spend more time talking about SEESAW and the AI methods being used. And, that will give you the opportunity to write another paper ☺

General Information

Title: Accurate Estimates Without Calibration

Author Name: Tim Menzies

Section I.

1. Which category describes this manuscript?
   ___Practice/Application/Case Study/Experience Report
   _3_Research/Technology
   ___Survey/Tutorial/How-To

2. Is the manuscript technically sound? Please explain your answer.
   ___Yes
   _3_Appears to be - but didn't check completely
   ___Partially
   ___No

3. Are the title, abstract, and keywords appropriate? Please comment.
   _1_Yes
   _2__No

Reviewer 2: There are no keywords provided.

[5] Keywords added on p1: AI, decision making, software engineering, model-based project management, search

Reviewer 3: The results need to be better connected to the conclusions of the paper.

[6] The conclusion now comes back to the introduction in a much clearer way (see the last set of dot points on p14)

4. Does the manuscript contain sufficient and appropriate references? Please comment.
   _3_References are sufficient and appropriate
   ___Important references are missing; more references are needed
   ___Number of references are excessive

5. Does the introduction state the objectives of the manuscript in terms that encourage the reader to read on? Please explain your answer.
   _2__Yes
   _1_Could be improved
   ___No

Reviewer 1: For explanation, please see "Detailed Comments" below.

Reviewer 2: The introduction explains the area where the research is conducted, the problems addressed, and the findings. It makes you want to read the rest of the paper.

[7] Many thanks.

6. How would you rate the organization of the manuscript? Is it focused? Is the length appropriate for the topic? Please comment.

_1_Satisfactory
_1__Could be improved
_1__Poor

Reviewer 2: The method description lacks details that would support its understanding. Empirical validation is missing essential elements. For more comments please refer to Section III.

[8] The previous draft was very confusing on the terminology, and this greatly confused the exposition. This draft simplifies and unifies the terminology. Combined with the other changes you proposed in section III, this draft better supports understanding.

7. Please rate and comment on the readability of this manuscript.

___Easy to read
_1.5_Readable - but requires some effort to understand
_1.5__Difficult to read and understand
___Unreadable

Reviewer 2: For comments please refer to Section III.

Section II. Summary and Recommendation

A. Evaluation

Please rate the manuscript. Explain your choice.

___Award Quality
___Excellent
_1_Good
__1_Fair
_1__Poor

B. Recommendation

Please make your recommendation and explain your decision.

___Accept with no changes
_2_Accept if certain minor revisions are made
_1_Author should prepare a major revision for a second review
___Reject

Section III. Detailed Comments

A. Public Comments (these will be made available to the author)

Reviewer 1: This is a very interesting manuscript definitely relevant and of sufficient quality for a journal publication (after several minor revisions as outlined in the following).

Footnote *: "Parts of this research as carried …" → "… was carried …"

[9] Many typos fixed, including this one.

Abstract: The concept of "process model" is defined extremely inclusive. In the process modeling literature, models such as COCOMO or COQULMO are usually not considered process models but prediction models.

[9] Agreed. See comment [1].

Abstract: What is the difference between "local tuning" and "local calibration"? What is the difference between "local calibration" and plain "calibration"? If there is no difference, I suggest using simply "calibration" (for both "local tuning" and "local calibration") since this is the term used in the title.

[10] We agree that this was a major problem with the previous draft. This draft has discarded the term “calibration” and unified the rest. Now, this paper consistently talks about “reducing tuning options” and “reducing project options”.

Introduction: There is a problem with the terminology "local tuning" and "local calibration". Both terms seem to refer to the same concept. Using just one term avoids confusion. However, if there is a difference, please explain!

[11] Fixed. See [10] .

Related Work: This section defines the term "calibration" for the first time in the manuscript. Please provide the definition earlier. Also, clarify what is different to "local calibration" (and "local tuning").

[12] This terminology choice was a major source of confusion in the prior draft and has been fixed here. See [10] .

Related Work: "… *project* choices." → "… *project* variables." ???

Related Work: "… methods like simulated annealing." → "… methods like simulated annealing (SA)." Alternatively, avoid using "SA" altogether.

[13] Use of "SA" removed. Now it is all "simulated annealing".

Related Work – last sentence: Again, "tuning" is used without clarifying what is different to "local tuning" and "calibrating".

[14] Fixed. See [10]

Section 4 – Figure 4: What do the various lines on the right-hand side represent? How many lines are there in total? Is the answer "20", i.e., one line per "repeat"? What exactly does one data point in each of the lines represent? This would then explain what you mean by "All values".

[15] Fixed. That figure's caption is now "Right- hand-side shows MRE1s generated over the NASA93 data set for ten case studies (one study per line). ". Note that this is now Figure 5, not Figure 4… see p25

Section 4 – last paragraph: In "Since we failed to generate precise tunings that yield exact estimates, …", what do you mean by "precise tunings" and "exact estimates'? Please clarify.

[16] Fixed. That text is now " Since we failed to reduce estimation variance by constraining the tuning variables, we took another approach. Rather than constrain the tuning variance, we developed the SEESAW system to explore the effects of just constraining the projects variance."

Section 4 – last paragraph: In "Perhaps, we argued, it was time to explore the space of possible tunings.", do you mean "space of possible project choices (variables)"?

[17] Fixed. See [16].

Section 6 (Experiments): "SEESAW"S" → "SEESAW's". Section 6 (Experiments) – sentence before Δ formula: "… these ten cases studies runs …" → "… these ten case study runs …" ???

[18] The above typos are all now fixed. Thanks for the careful proofread.

Section 6 (Experiments): You define Δ as the relative difference between SEESAW and LC estimates under point 7 of the explanation of Figure 6. This is consistent with the right-hand side of Figure 7. However, towards the end of page 12, you say that Δ = … = {20, …, 26}%, where {20, …, 26}% are the median values of MRE2. Shouldn't therefore the correct formula be: median(Δ) = {20, …, 26}% ?

[19] You are 100% right. Fixed.

Section 6 (Experiments) – caption of Figure 7: "come from Figure 1" → "come from Figure 4".
[20] You are 100% right. Fixed.

Section 6 (Experiments) – first sentence after Figure 7: "delta Δ value" → "Δ value".
[21] Fixed.

Section 6 (Experiments): I don't understand the argumentation with regards to the discussion of "small" Δ values. First of all, do you mean small "median Δ values"?
[22] Yes. Fixed.

If you mean "Δ values", what does the right-hand side of Figure 7 tell us? Doesn't it tell us that MRE2 varies between 0 and something close to 80%? You don't provide any other study that tells us that these differences between LC and SEESAW estimates are "small". I don't think it is possible to "show" that SEESAW estimates are approximately as good as LC estimates by presenting relative differences that vary between 0% and 80%. Similarly, I don't see how the median delta values (varying between 20 and 26%) can be "shown" to be small by saying that the MRE2 values are (mostly) very close to the MRE1 values. I propose showing the SEESAW MRE1 values, i.e., showing the differences between estimates and actual values, and then comparing the MRE1(LC) with MRE1(SEESAW).

[23] It is not possible to do MRE1(seesaw) because, as shown in Fig6, MRE1 are the results from training and testing on historical NASA data. The only other comparison is to train on NASA data and test on data generated from SEESAW- and that is the MRE2 results

[24] But stepping back from the particulars of this comment, we quite take the reviewer's more general point that the experiments were poorly described. In this new version, the experimental section is divided into GOALS, METHODS, RESULTS, SIGNIFICANCE TESTS, and SUMMARY OF RESULTS. Further, in the GOALS section, we take care

to formally state the two hypotheses (H0,H1) explored in this experiment. See pages 10..13.

Reviewer 2: # Comments on the method #
The method described in the paper seems to be more of a project optimization than a prediction approach. A prediction method typically takes as its input actual or estimated (certain or uncertain) characteristics of a project and provides a predicted output (effort, time, defect content, or risk). The method provided in the paper, on the other hand, takes fixed project characteristics and searches for such values of unknown or modifiable project characteristics that result in minimal estimates (effort, time, defect content, and risk).

[25] Reviewer 3 made a similar comment. Now, this draft takes care to exactly describe current practice (see steps1, steps2, steps3 on page 4) as well as how this work is different to that process.

Moreover, the description of the SEESAW method and algorithm is quite rough and very difficult to understand. More systematic and detailed description of the method's steps would be required. For example, it is not clear whether SEESAW constrains the values of project choices or whether it is also able to remove project choices from an estimation model (as irrelevant ones).

[26]. This is a good point and it was one that we missed. The new draft has words on exactly this issue in the second paragraph of section 3.4 (p8)

Comment on the validation
The design of the empirical validation is not clear. The (mostly qualitative) results of the validation are quite vague. On the one hand, the reader would expect a detailed description of the experiment design including quantitative definition of research hypotheses and a clear specification of the experimental procedure (whether characteristics of the experiment, such as independency and randomness, are clearly recognizable). On the other hand, the reader would expect quantitative results of an experiment including: statistical significance and power of applied statistical tests.

[27] Agreed, See point 24, above

Detailed comments
Section 2, page 4: The paper provides references for estimating uncertainty using Bayesian analysis. Since the paper mainly deals with effort and quality estimation it would by probably useful to provide a reference to Fenton's paper where BBN were applied for effort estimation (in addition to a paper where BNN was applied for defect prediction).
[28] Done. See the related work (and also, a quote from Fenton in the introduction).

Section 3.3, page 7: Equations 2, 3, and 4 are actually tables (as compared to equation 1). On the other hand, some equations (e.g., page 6, page 8) are missing any caption. This inconsistency increases the difficulty of reading the paper.
[29] Fixed

Section 4, Page 9: Figure 4 is missing units. For example, regarding the right-hand-side figure, the text refers to its values as given in %. Moreover, the definition of MRE provided in the paper is not consistent with the values given in text (MRE definition does not indicate %).
[30] % added. Definition fixed to include %.

Section 6, page 12: The abbreviation "lc" in Figure 6 is not defined.
[31] Fixed

Some more minor comments:

- Section 2, 1st list, "Calibration: import an log…" → import a log (remove the n)
- Section 2, 2nd list, "Prediction is used to create one point…" → to create a one point… and further: add a space between the comma and PRICE-S
- Section 2, paragraph following 2nd list: I guess you mean Structured Analysis with SA – maybe you should say this somewhere, there are too many two-letter abbreviations around
- Section 2, 3rd paragraph following 2nd list: "Harmon's writing inspired us try…" → "Harmon's writing inspired us to try…"
- Section 2, last paragraph: "…tuning and validation one…" → tuning and validating one…"
- Page 10, 1st sentence: why did you use 100 calls? Is there a specific reason (i.e., 50 was not enough, but 200 did not get you any better results), or are there practical constraints of some sort?
- Section 6, 1st sentence: SEESAW"S → SEESAW's
- Section 8, 3rd list item: Consider do draw conclusions instead of making them?

[32] The above typos are all now fixed. Thanks for the careful proofread.

Reviewer 3:

1. The COCOMO family of models are more accurately classified as cost estimation models. The term "process models" in general and "software process models" in particular refer to graphical depictions (i.e. models) of the software development process. To this reviewer's knowledge, the COCOMO family of models have never included graphical depictions of the process and it is not correct for them to be referred to as "process models" or "software process models".

[33] We 100% agree. See comment [1], above

2. The authors are asked to provide more distinctions around uncertainty. They describe 2 types of uncertainty in the abstract – a) uncertainty associated with process/model inputs and b) uncertainty associated with "internal parameters that control the conversion of inputs to outputs." Please tell us more. What do you mean by the conversion of inputs to outputs? Are you specifically referring to the Tuning Variance? Please make this distinction clear in the introduction or early in the paper. Please make the link between these types of uncertainty and Project variance as well.

[34] Agreed. We've addressed this above in comment [10]

3. The authors should clearly state that preset COCOMO parameters were used in this analysis and were developed based upon years of research

[35] We have added that text at the end of section 2 (p5)

4. How long does it take to select among the preset COCOMO parameter choices? Was this time included in assessing the cost of your method?

[36] Boehm took years to do it (see [35]). That work is now $0 for this approach since we lever that work.

5. The authors state in the abstract that, "Our conclusion is that, (a) while local tuning is always the preferred option, there exist some process models for which local tuning is optional; and (b) when building a process model, we should design it such that it is possible to use it without tuning."

   Do the authors show that for each of their case studies, the same decision would be made about the process whether or not local tuning was done to their

COCOMO models? If the authors' claim is true, one conclusion could be that local tuning is not necessary when using reasonable preset COCOMO model parameters because the structural process issues captured by COCOMO are significant enough to dominate the effects of local tuning.

[37] The actual project decisions are NOT listed in this paper, for reasons of space. Rest assured that all these project yield different project recommendations.

6. The authors state that, "after 26 years of trying, we have only collected less than 200 sample projects for the COCOMO database. Also, even after two years of effort we were only able to add 7 records to a NASA-wide software cost metrics repository ". Are the authors are going to count the time spent collecting host organization data into their approach? Are they going to count the amount of time spent selecting preset COCOMO variables? If not, should then that amount of time should be disregarded from other approaches the author is comparing their approach to.

[38] Not clear on this comment. The local collection time here is just the time required to build fig2 (page 22). We've done this many times and it takes hours (not days) if the business user needs a COCOMO pre-briefing (and minutes if they don't).

7. The authors present their main result early in the paper:

   "The range of estimate errors seen after constraining the project choices (but not the tuning variables) is almost identical to the range seen after constraining just the tuning variables."

   The author needs to better explain why this result is important to both academics and practitioners. Although it is a good result to have variance reduce, doesn't variance usually reduce when input values to a COCOMO type model are constrained?

[39] Agreed. See the new "Implications" section on page 4. The value in reducing variance is that it enables the defense of processes options

8. Can the authors explain what is difference in the result doing when both local calibration to reduce variance in tuning variables combined with constraining project choices? What is the result of doing a little bit of local calibration combined with constraining the project variables? Is there a "sweet spot" for how much local calibration is done? Is that given in the other papers? One could imagine that the first bit of local calibration could go a long way.

[40] This is another paper, under development. Currently it would appear, as the reviewer suggests that here is no need to completely ignore local data. A little bit of tuning goes a long way! However, this paper is not about that work and this paper is making the case that decision making can proceed without local data.

9. Given the problem of different definitions of the data and different understandings of the meanings of the parameters, isn't at least some minimum amount of calibration required?

[41The results of this paper assume zero calibration. So we can generate baseline results using zero data and we are checking , in the other paper, if a little calibration takes us to a better place..

10. The authors claim that, "From a business perspective, this result means that certain process models can be used for decision making in one of two ways:
    1. Either constrain the tuning variance using historical data;
    2. Or constrain the project choices using an AI search engine like SEESAW.

    The authors need to explain why they have stated the implications of the results this way. To this reviewer, the implication of the result as stated by the authors doesn't seem to follow. The reason is that it would appear that one is doing two different things when constraining tuning variance and when constraining project choices. Constraining project variance has implications on decisions and project performance. Constraining tuning variance does not. Ultimately, we want to use the model to help in making process decisions. This implies that we need to constrain the project variables no matter what. If it is true that these process decisions dominate the variability associated with tuning variance and the tuning doesn't matter, so be it. Great! That's a result people want to know. And, if variance in project performance also happens, again great. You have just proven that best practices indeed have a substantial impact on project performance. And, let's not forget that the range of tuning variance has already been constrained by past research done throughout the years on COCOMO. The authors seem to say this more clearly in the conclusions of the paper. Perhaps similar language should be used earlier. P>> T.

    [42] This is a very important comment and lead to a total reorganization of the introduction along the lines you mention above. And yes, we make more use of the P, T terminology.

11. Given recent papers presented at ICSP and elsewhere, and given the newer object based process simulation techniques, the process simulation community has advanced well beyond earlier efforts. The author's comment about it taking 2 years to develop useful models is simply inaccurate. The time frame required to create these models is more on the order of 1 to 2 months.

    [43] Yes. Based on this suggestion, this paper makes a comment at the start of the conclusion that the techniques of this paper are not required for data rich domains.

12. This paper shows interesting work. However, it is unclear to this reviewer how the conclusions follow or make sense. The author is asked to explain the results and benefits and costs of their approach better in the early part of the paper. I believe that for the audience of SPIP, a further discussion of the AI tools and approach would have limited value. However, greater explanation of the results and their implications would be extremely helpful. The authors should also explain their experiments and what they tested for and compared.

    [44] Yes. Absolutely. This comment made us add in the new "Implications" section on p45 and significantly reorganize the experiment section (which now has GOALS, METHODS, RESULTS, SIGNIFICANCE TESTS, and SUMMARY OF RESULTS sections, p10-13).Also, based on this suggestion, we have moved as much of the AI stuff as we can to an appendix.