More importantly than the above are the discussions we have about the models. The following section is not a formal analysis. Rather, it summarizes the common dialogues we have when we present the above results.

One reaction we get is *over-acceptance*. We have found one audience that are uncritical of our work and want to implement our results, immediately. Typically, this audience is formed of managers who are nervous of their current practices and aver (overly) eager to change to any other approach.

While this is a gratifying response, we have found that members of this audience may not fully appreciate a important part of the above conclusions; namely that any policy that looks useful in general may not be useful in particular cases. Before implementing any of the above, we strongly recommend that users use our tools to find specialized policies that do best for their particular site.

Another reaction we get is our conclusions come from untuned COCOMO models and *the above policies only make sense after local tuning*. In reply, we offer separate comments on the mean and variance of the estimates generated by our approach.

Issues with Mean: The $experiment = 1$ results of Figure **??** shows the differences between the estimates generated *with* and *without* tuning on local data (the no-tuning estimates came from our search-based tool and the tuning-results can from on-line versions of the standard USC models[1]). The $experiment = 2$ results of that figure show the mean errors seen when estimates are generated using conventional methods. Note that the mean differences in the $experiment = 1$ results are small (36% or less). Further, the mean effort values of the conventional methods is larger (never less than 42%). That is, Figure **??** shows no evidence that the differences between our methods and conventional methods is inordinately large.

Issues with Variance: Our tools sample across the space of possible tunings. The variance of the conclusions reached in that sample could be reduced by the use of local data to pre-tune our models, before we perform any sampling. However, it must be noted that the variances of our conclusions are quite small (see Figure **??** and Figure **??**). In summary, while we believe it may be useful to tune with local data, the overall benefit on reducing the variance or the mean not be large.

[1]Phil: what URLs?

	experiment	data set		
		flight	OSP	OSP2
defects	1	30 ± 23	39 ± 33	32 ± 18
months	1	26 ± 17	23 ± 17	12 ± 7
effort	1	36 ± 31	20 ± 17	18 ± 13
effort	2	41 ±40	44 ± 63	42 ± 57

Figure 1: Results (Mean ± standard deviations) from our experiments. For an explanation of the method by which this was generated, see Figure **??**.

Figure 2: fig:delta

The lines labeled $experiment = 1$ in Figure **??** were generated as follows:

1. Impose the constraints C of Figure **??** on our search-based tool;
2. Learn a policy P;
3. Impose the combination of $C \cup P$ on our search-based tool;
4. Generate 100 *estimates* (without further searching);
5. Generate *baselines* from the standard on-line USC models[2] by picking values at random from within C and setting any unknown values to "nominal";
6. Combine the *estimates* from step 4 with the *baselines* from step 5 using $delta = \frac{estimate - baseline}{baseline}$.

The line labeled $experiment = 2$ in Figure **??** was generated by taking flight systems/OSP/OSP2 systems and estimating them using a COCOMO model tuned to some NASA COCOMO, as follows:

- Using the NASA93 COCOMO data set[3] we tuned a COCOMO model using using Boehem's LC method [**?**];
- Using the constraints C of Figure **??** we generated 100 random projects at random consistent with C;
- Effort estimates for the generated projects were generated using the tuned COCOMO model.
- The means and standard deviations for those estimates were recorded in Figure **??**.

Figure 3: Generation method for the Figure **??** results.