**Crossover Experiment**

Graphical user interface, application

Description automatically generated

I added another factor of complexity: **Phantom type**

* **Normal** (factory density inserts and no air)
* **Expanded** (factory inserts + new adapted inserts + air)

Chart, box and whisker chart

Description automatically generated

**Step1**

**Model Fitting with all data to explore:**

**Text

Description automatically generated**

ANOVA Type III indicates no difference between crosses, but effect of phantom type used and effect of different colonies

**Checking for multi-collinearity:**

Text

Description automatically generated

After removing multi-collinearity by standardizing our response variable using colony surface area, I filtered data to have only Type1 and Type2 crosses to better isolate effects. Then I proceeded to fit a similar model to check if significant effects remain

Text

Description automatically generated

Text

Description automatically generated

Visualizing the data

Chart

Description automatically generated

**Removing outliers:**

**Text

Description automatically generated**

Model with no outliers and with multicollinearity removed:

It seems effect of colony and Phantom Type remain, and cross type is not an issue,

*which is good as we want to prove that offsets from type1 and type2 crosses are not different.*

Text

Description automatically generated

**Step 2:**

Checking ANOVA assumptions in fitted model before proceeding with figures, conclusions and statistical reporting :

1. **Homogeneity of variance assumption**

Chart, scatter chart

Description automatically generated

Found visual outliers and some minor ‘parabolic’ appearance but we are just about good.

It can be useful to remove to ensure assumptions. However, if we run Levene’s test (see below) it give us reassurance that we can proceed.

Text

Description automatically generated

According to Levene’s test we’re good. p-value is not less than the significance level of 0.05. This means that there is no evidence to suggest that the variance across groups is statistically significantly different. Therefore, we can assume the homogeneity of variances in the different treatment groups.

1. **Normality:**

Chart, scatter chart

Description automatically generated

As all the points, but a few remaining outliers, fall approximately along this reference line, we can assume normality.

The conclusion above, is supported by the **Shapiro-Wilk test** on the ANOVA residuals which finds no indication that normality is violated.

Text

Description automatically generated

Now final model with Type III ANOVA to account for unbalanced design

Text

Description automatically generated

Final ANOVA indicates Standardized offset is significantly affected by PhantomType and Coral but not by Crosstype

**Step 3:**

Exploratory figures to better visualize effects

Chart, box and whisker chart

Description automatically generated

Chart, line chart

Description automatically generated

**Step 4:**

**Post hoc analyses to report pairwise effects**

Contrast between Type 1 and Type 2 🡪 not statistically different

*(good – it means we can do internal calibration, but as shown below – PhantomType matters).*Text

Description automatically generated

Contrast between Phantom Normal and Expanded 🡪

*statistically different – PhantomType matters.*

Text

Description automatically generated

So let’s have a look at pairwise contrasts:  
Text

Description automatically generated

We see above that

Within Type1 – the kind of Phantom matters

Within type2 - the kind of phantom does not matter *(but borderline – which also gives us a hint that we should have Expanded phantom from now on as I will show later).*

**Results/ Take home message:**

**1 )** Final ANOVA shows that effects of colony and phantom type on weight offset are predominant over cross type.

**2)** Type1 crosses produce mean offsets that are statistically different between Phantom Types (p-value=0.0013). Leads to think that a type1 cross is not reliable, because if you adopt it you will get different values based on the phantom you use *(good ! that’s what I wanted to show and proves we internal calibration is better)*

**3)** When adopting internal calibration (Type2 cross), we found weak evidence to support that an Expanded Phantom is better over a Normal Phantom (Contrast p-value=0.053)

*(maybe also helps justify scans we did in the past where we did not have an adapted phantom at hand - but we should still prefer Expanded Phantom to cover a wide range of possible density values, thus better constraining calibration curves).*

**Going back to hypotheses on page 1 (for curiosity purposes):**

Text

Description automatically generated

Calendar

Description automatically generated Chart, line chart

Description automatically generated

Data above summarized in one figure:

H1: Offset largest for Type4 cross

**Answer**: Seems true, their boxplots show largest ranges.

H2: Offset is the **smallest** from **Type1 cross**, as Coral and Phantom do not interact. ​

**Answer**: False. Type2 offsets are comparable according to stats (see above, at least when using Expanded Phantom)

H3. The difference between Type1 and Type2 is statistically insignificant. ​

**Answer**: True (at least when using Expanded Phantom, good so can do internal calibration)

H4: Offset is **intermediate**from Type3. I’d expect the offset from Type3< Type4, as phantom volume is less affected by interaction with coral. ​

**Answer**: Not tested (not useful).

1. Am I correct that for the extended ‘vs normal phantom experiment it was in the same scan just extracted differently? Yes  This means that the high density standard (Ti) is present in the scan whether it is used or not for the calibration?  If so its effect will be felt on both calibrations whether it is used to create it or not. So need to be careful here. I did ask the same question to Amsterdam colleagues when I went there. Based on the chats I had with them, my impression was that the high-density insert would have no effect on the other inserts, but rather the prevailing effect on phantom would come from presence or absence of coral.   
     
   This experiment is not testing the phantom impact fully *(it would be rather the coral impact on phantom, no?),* nor a prefect replication of the extended v’s factory phantom calibration use in previous scans, nor is it statistically independent. I am absolutely not suggesting that you should do this part of the experiment again – just saying that your arguments are on very thin ice.

I would leave out this comparison in the paper/chapter. Simply quantify what % of voxels are missing in standard calibration v’s extended.  It is clear on the density offset graph where this becomes an important effect and so I think this can be added as a visual annotation more powerfully than further testing.

1. Correct me if I am wrong but key result we are interested in is whether scanning a coral with the calibration phantom gives a more accurate and more precise prediction of true weight (n= 10 colonies + variation of scan conditions in this experiment) compared to separate scans but correctly matched scan conditions (n = 10). *I’d say we are not aiming for more precise, but rather equivalent (i.e., no difference amongst the two cross types so we could argue there’s no way one is better over the other).*

But wouldn’t this statistical comparison be much clearer if the density offset effect was removed by simply doing a paired-t test for offset value for each colony pairing type1 and type 2? Currently you don’t know how much of the density offset effect is spreading out the populations adding noise. It could be the dominant cause of the distributions seen in the box-plot figures. *Agreed, a paired t-test haven’t even crossed my mind, I was so focused on average offsets that I totally forgot about the DIFFERENCES.*

1. Is the weight offset a fraction or a % in the figures?  It is small even including the density offset effect.
2. In terms of the impact of these effects on colony density (remembering that this is your key scientific parameter that you want to measure), even including the worst case type 4 extremes you are looking at ±0.06 (1SD) g/cm^3.  This is 2-5x smaller than of population variation measured by Lough and Barnes (2000) at a single reef.  The value is exaggerated, so need to calculate like for like.  However this is great.  It’s not important for the science – go measure! 😊
3. I’d suggest a simple paired-t test for the instruments, but the calibration really is at the noise level that isn’t going to change the science story.  The extended calibration improves predictions, but that isn’t surprising.  I do think we are at write-up stage now.  YAY 😊