A reply to Bergh: FISHERIES/2020/AUG/SWG-PEL/84

Richard B. Sherley[1,2]

1. Centre for Ecology and Conservation, College of Life and Environmental Sciences, University of Exeter, Penryn Campus, Cornwall, TR10 9FE, United Kingdom.
2. FitzPatrick Institute of African Ornithology, DST-NRF Centre of Excellence, University of Cape Town, Rondebosch 7701, South Africa.

Email: r.sherley@exeter.ac.uk

Note: For readers' ease, responses (**R**) are in blue text, have been numbered in a manner that corresponds to the numbering used in FISHERIES/2020/AUG/SWG-PEL/84 (e.g. **R1**) and are inserted after each point from the original document below. Italics are used for direct quotes.

Key Summary Point:
Given the need to make a decision during 2020, I would suggest a more useful way forward than these constant rebuttals is to focus on where multiple sets of analyses lead to the same broad inference (see Table 1) and move forward with management in an adaptive framework. That is not to say that the points below cannot be revisited, although I suspect that there is more to gain from addressing the issue of how these observed effects translate into population-level change (though that too is non-trivial in some cases, see **R14** below).

1. A debate that has been circulating for a number of years is the merit of the use of individual bird data versus standardised aggregated data. This debate has not been resolved in FISHERIES/2020/JUL/SWG-PEL/***53REV***. This dichotomy is one of the important differences between the results reported in FISHERIES/2020/JUL/SWG-PEL/***53REV***, and those reported in FISHERIES/2020/JAN/SWG-***PEL/09*** (and FISHERIES/2019/NOV/SWG-***PEL/27rev***). Further deliberations on the management actions required for island closures need to see this debate resolved by methods that occur in fisheries management deliberations in South Africa. This involves different researchers establishing 'common ground', viz. using the same data and methods to produce the same analytical results. Thereafter, researchers should demonstrate how departure from such 'common ground' leads to differences in the analytical results, and to provide step by step justification of the merits of these departures. This is both a scientifically defensible approach and it gives those on the periphery of the technical deliberations confidence to base management recommendations on a particular analytical result. Allusion to general theoretical support for departures from 'common ground' may have some merit but it is necessary to back all this up by comprehensive numerical analyses. Management deliberations are impeded when different methods are applied without the 'common ground' approach, and especially when they also produce different results. There are further differences between the methods in FISHERIES/2020/JUL/SWG-PEL/***53REV*** and FISHERIES/2020/JAN/SWG-***PEL/09*** (see next point). But at this stage, without the use of 'common ground' to facilitate deliberations, it is very difficult to accept one or another approach as a sound basis for decision making.

R1: Agreed, finding common ground is useful. But this comment rather overlooks the peer-reviewed scientific literature on this issue (see comments in Sherley 2020a, FISHERIES/2020/AUG/SWG-PEL/83). And it overlooks the point that I have already submitted documents attempting to provide this common ground on more than one occasion – see Sherley (2016) and Sherley and Winker (2019). In the interests of doing so again, please see Sherley (2020b), FISHERIES/2020/SEP/SWG-PEL/86 and Table 1 overleaf.

Given the need (which I believe the SWG-PEL has expressed) to make a decision during 2020, I would suggest, rather than further technical debates a more useful way forward is to focus on where the two sets of analyses lead to the same broad inference and move forward with management in an adaptive framework. In Table 1, blues can be taken to mean the signal to noise ratio in the data is too poor to detect effects. Greens can be taken to mean strong evidence for a negative effect of the closure on the penguins. And Oranges and Reds tentative and strong evidence for a positive effect of the closure on the penguins, respectively. In both columns, Oranges and Reds are more common than Greens. Both parties find 5 positive effects with tentative or strong evidence spilt across 3 islands. In other words, there is broad agreement between these two sets of analysis.

Table 1: Common ground between FISHERIES/2020/JUL/SWG-PEL/53REV (and associated analyses), and FISHERIES/2020/JAN/SWG-PEL/09 and FISHERIES/2019/NOV/SWG-PEL/27rev). Positive (+) and negative (−) signs are the direction of the effect: + here means positive for the penguins in the sense of Sherley (2020c; FISHERIES/2020/SEP/SWG-PEL/89, so indicates a negative δ value in FISHERIES/2020/JAN/SWG-PEL/09 and FISHERIES/2019/NOV/SWG-PEL/27rev. – means the reverse. The colour scheme denotes how meaningful each effect is and is based on the colour scheme used in Table 2 of Ross-Gillespie and Butterworth (2020; FISHERIES/2020/JAN/SWG-PEL/09. For entries in the R-G&B column: Green 🟩 = "There is no evidence in the current data to support a biologically meaningful fishing effect, as the lower bound of the normal distribution, $\delta^{EM*}_{data} - 2(se)$, lies above the threshold". Blue = "The experiment needs to continue for more than 10 years before a biologically meaningful fishing effect is likely to be detected, if it is present". Orange 🟧 = "The experiment needs to continue for 2 to 5 years before a biologically meaningful fishing effect is likely to be detected, if it is present". Red 🟥 = There is evidence in the current data of a biologically meaningful fishing effect because $X > P_{min}$. For entries in the Sherley et al. column: Blue 🟦 = There is no evidence at present for a closure effect either way in this dataset; < 95% of the posterior distribution has the same sign as the mean. Orange 🟧 = There is a 95% probability of a closure effect in the direction indicated by the sign in this dataset. Red 🟥 = There is a greater than 97.5% probability of a closure effect in the direction indicated by the sign in this dataset. Green 🟩 is not used in the Sherley et al. column as there are no cases where there is evidence for a negative impact of the closures that is credibly different from zero at the 97.5% level.

Island	Data type	R-G&B	Sherley et al.	Sources
Dassen	Chick growth	🟩 −	🟦 +	[1,2]
	Chick Condition	🟦 +	🟦 +	[1,3]
	Fledging success	🟦 −	NA	[1]
	Path Length	🟩 −	NA*	[1]
	Trip Duration	🟩 −	NA*	[1]
	Max. Distance	🟦 −	🟦 +	[1,3]
	Chick survival	🟥 +	🟥 +	[1,3]
Robben	Chick growth	🟦 −	🟦 +	[1,2]
	Chick Condition	🟥 +	🟧 +	[1,2]
	Fledging success	🟥 +	NA	[1]
	Path Length	🟦 −	NA*	[1]
	Trip Duration	🟦 −	NA*	[1]
	Max. Distance	🟦 −	🟦 +	[1]
	Chick survival	🟧 +	🟥 +	[1,3]
Bird	Chick Condition	NA	🟦 −	[3]
	Path Length	🟦 −	🟦 +	[3,4]
	Trip Duration	🟦 −	NA*	[4]
	Max. Distance	🟦 −	🟦 +	[3,4]
St. Croix	Chick Condition	NA	🟦 −	[3]
	Path Length	🟦 +	🟧 +	[3,4]
	Trip Duration	🟦 +	NA*	[4]
	Max. Distance	🟥 +	🟥 +	[3,4]

Notes: NA = Not analysed by that group. R-G&B = Ross-Gillespie and Butterworth. *Not analysed due to issues of heterogeneity of variance (see R8 below). Sources: [1] Table 2b of Ross-Gillespie and Butterworth (2020; FISHERIES/2020/JAN/SWG-PEL/09). [2] The last time growth was analysed using a disaggregated approach was in Hagen et al. (2014; MARAM/IWS/DEC14/Peng/A3). This result is used here as the dataset has not been updated since. [3] Figure 1 of Sherley (2020c; FISHERIES/2020/SEP/SWG-PEL/89). [4] Based on option 2 (Closure model, 2018 foraging data) in Figure 1 of Ross-Gillespie and Butterworth (2019; FISHERIES/2019/NOV/SWG-PEL/27rev).

2. The international stock assessment review panel convened in 2015 (*International Review Panel Report for the 2015 International Fisheries Stock Assessment Workshop 30 November – 4 December 2015, UCT. A Dunn M Haddon, A M Parma, A E Punt*) proposed a methodology to be incorporated into any method that is applied in estimating the island closure effect. The relevant excerpts are points 3 and 4 of page 3, point 15 on page 4, point A3 of page 5, Table 1 and Appendix A. These points and recommendations have not been addressed or followed in FISHERIES/2020/JUL/SWG-PEL/53REV

R2: The top of page 3 of Dunn et al. (2015) states: "In relation to next steps for a power analysis to evaluate closure effects on penguins:" and points 3 and 4 then immediately follow and pertain to conducting a power analysis, as does point 15 on page 5.

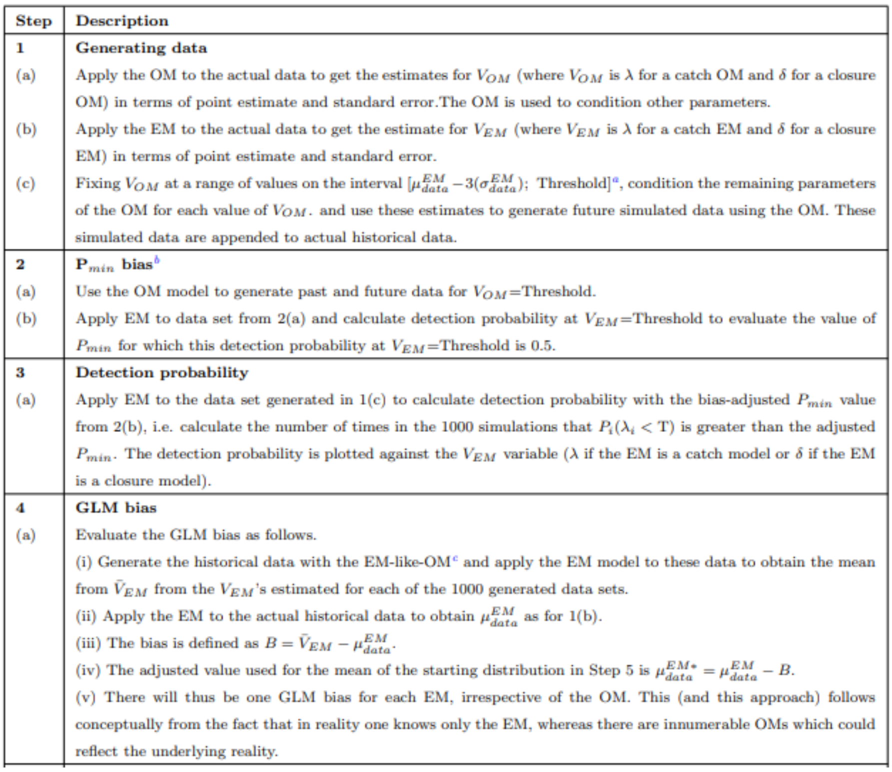
Table 1 still pertains to a power analysis and contains recommendations to fit models using catch rather than closure; these recommendations were later superseded by the recommendations of the 2016 panel: "*that the Closure estimator form the basis for decision making in the short-term*" (Dunn et al. 2016). Thus, it would be odd to now go back to the 2015 recommendation and implement models using catch again – particularly given the issues outlined by Bergh (2014) about the biases that arise when both catch and biomass are in the same models.

Appendix A of Dunn et al. (2015), "OUTLINE OF THE PROCESS OF CONDUCTING A POWER ANALYSIS FOR AFRICAN PENGUINS", again pertains to conducting a power analysis.

Sherley (2020d), FISHERIES/2020/JUL/SWG-PEL/53REV is not conducting a power analysis.

3. The 2016 International Review Panel Report (*International Fisheries Stock Assessment Workshop 28 November – 2 December 2016, UCT, A Dunn , M Haddon , A M Parma , A E Punt*) noted that "the use of disaggregated data in an estimator would require that steps 1-4 of Table 2 of MARAM/IWS/DEC16/Peng_Clos/P1a be followed" – these four steps are reproduced below:

Table 2: Summary of the power analysis procedure.



Step	Description
1	**Generating data**
(a)	Apply the OM to the actual data to get the estimates for V_{OM} (where V_{OM} is λ for a catch OM and δ for a closure OM) in terms of point estimate and standard error.The OM is used to condition other parameters.
(b)	Apply the EM to the actual data to get the estimate for V_{EM} (where V_{EM} is λ for a catch EM and δ for a closure EM) in terms of point estimate and standard error.
(c)	Fixing V_{OM} at a range of values on the interval $[\mu^{EM}_{data} - 3(\sigma^{EM}_{data});$ Threshold$]^{a}$, condition the remaining parameters of the OM for each value of V_{OM}. and use these estimates to generate future simulated data using the OM. These simulated data are appended to actual historical data.
2	**$\mathbf{P}_{min}$ bias**[b]
(a)	Use the OM model to generate past and future data for V_{OM}=Threshold.
(b)	Apply EM to data set from 2(a) and calculate detection probability at V_{EM}=Threshold to evaluate the value of P_{min} for which this detection probability at V_{EM}=Threshold is 0.5.
3	**Detection probability**
(a)	Apply EM to the data set generated in 1(c) to calculate detection probability with the bias-adjusted P_{min} value from 2(b), i.e. calculate the number of times in the 1000 simulations that $P_i(\lambda_i < \mathrm{T})$ is greater than the adjusted P_{min}. The detection probability is plotted against the V_{EM} variable (λ if the EM is a catch model or δ if the EM is a closure model).
4	**GLM bias**
(a)	Evaluate the GLM bias as follows. (i) Generate the historical data with the EM-like-OM[c] and apply the EM model to these data to obtain the mean from $\bar{V}_{EM}$ from the V_{EM}'s estimated for each of the 1000 generated data sets. (ii) Apply the EM to the actual historical data to obtain μ^{EM}_{data} as for 1(b). (iii) The bias is defined as $B = \bar{V}_{EM} - \mu^{EM}_{data}$. (iv) The adjusted value used for the mean of the starting distribution in Step 5 is $\mu^{EM*}_{data} = \mu^{EM}_{data} - B$. (v) There will thus be one GLM bias for each EM, irrespective of the OM. This (and this approach) follows conceptually from the fact that in reality one knows only the EM, whereas there are innumerable OMs which could reflect the underlying reality.

These four steps have not been followed in FISHERIES/2020/JUL/SWG-PEL/***53REV***.

R3: Again, the Table 2 above is labelled "*Summary of the power analysis procedure*" and so pertains to the process for conducting a power analysis. It isn't at all clear why it ought to be necessary to go through a process designed for a power analysis when simply using a (generalised) linear mixed model (GLMM) to determine whether a difference between two means is credibly (or statistically) different from zero. Indeed, as FISHERIES/2020/JUL/SWG-PEL/53REV is not, and has never been, conducting a power analysis, suggesting that it is necessary to follow a procedure designed for when conducting a power analysis (in a specific context) before the results can be accepted is tantamount to suggesting that anyone, anywhere must follow the procedure in Table 2 before using a GLM(M) in any analysis. It further suggests that all papers using GLM(M)s need to have their analysis reconsidered until such time that they apply the procedure in Table 2. Surely, that is not what is being suggested?

4. The methods referred to in point (2) and (3) above include methods that need to be applied to deal with estimation bias (see in particular point 4 in the extract above) as is extensively elaborated and enlarged upon in MARAM/IWS/DEC15/PengD/P1. Such estimation methods have not been incorporated into FISHERIES/2020/JUL/SWG-PEL/***53REV*** although they were incorporated into FISHERIES/2020/JAN/SWG-PEL/09. This shortcoming needs to be addressed.

R4: I'm unclear on the point being made here. MARAM/IWS/DEC15/PengD/P1 appears to deal predominately with establishing whether there is a different bias for the Catch estimator when used in a GLM versus the Closure estimator. The 2016 panel concluded that there was a bias in the catch estimator and favoured the Closure estimator: "*that the Closure estimator form the basis for decision making in the short-term*" (Dunn et al. 2016) [although the Catch estimator continues to be used until at least November 2019 in Ross-Gillespie and Butterworth (2019), FISHERIES/2019/NOV/SWG-PEL/27rev]. And points 2 and 3 above pertain to conducting a power analysis.

I request that we go back to the 2019 panel and ask them whether they are of the opinion that FISHERIES/2020/JUL/SWG-PEL/53REV needs to be amended to take account of the table above.

5. It is surely necessary to test whether effects which are being introduced as random effects satisfy the requirements for assuming that they are random. Or can one simply assign certain effects that need to be accounted/allowed for as random, and others for which specific estimates are require as fixed?

R5: The diagnostic plots presented FISHERIES/2020/JUL/SWG-PEL/53REV make this check and that is why they were asked for by the 2019 panel: "*This should include standard residual analysis as well as residual analyses that are tailored to the problem at hand (e.g., temporal, spatial or within-season plots of residuals)*" (Die et al. 2019). The residual plots presented show no evidence of deviation from the assumptions underlying the use of the relevant mixed effects models.

6. The use of different random effects has a large impact on the standard error of certain of the closure effects reported in FISHERIES/2020/JUL/SWG-PEL/***53REV*** (see for example M1 of Table 1). The reasons for this result needs explanation since it leads to a conclusion that the island closure effect is not statistically significant.

R6: As outlined in FISHERIES/2020/JUL/SWG-PEL/53REV, one possibility is that this is an issue with having Island in both the fixed and random components of the model. M1 in all cases in the maximal model (the most complex possible random effect structure); maximal models are "generally wasteful and costly in terms of statistical power for testing hypotheses" (Stroup 2012, pg. 185) and maximal models – even when they converge – can result in overparameterization that leads to uninterpretable models (Bates et al. 2018). The maximal model may actually trade-off power for some conservatism beyond the nominal Type I error rate, even in cases where the maximal model matches the generating process exactly (Matuschek et al. 2017).

7. The Kaplan Meier results suggests that at Robben Island chick survivorship is dependent on time, or chick age (see Figure A4.13 of FISHERIES/2020/JUL/SWG-PEL/***53REV***). The potential that this has biased the closure effect estimate because of the selection of chicks at different ages/times for estimating chick survivorship needs to be fully explored. This has not been done in FISHERIES/2020/JUL/SWG-PEL/***53REV***.

R7: Please see Response 18 in Sherley (2020e), FISHERIES/2020/SEP/SWG-PEL/85. In principle, this could be explored further, but it is far from clear that this is necessary.

8. FISHERIES/2020/JUL/SWG-PEL/***53REV*** and FISHERIES/2020/JAN/SWG-***PEL/09*** focus on different penguin response variables and data – the data that were used are indicated in shaded blocks.

	FISHERIES/2020/JUL/SWG-PEL/53REV			
53Rev	**Dassen Island**	**Robben Island**	**St Croix**	**Bird Island**
Chick condition	■	■	■	■
Max distance	■	■	■	■
Chick survival	■	■		
Chick growth				
Forage length				
Forage duration				
Fledging success				

	FISHERIES/2020/JAN/SWG-PEL/09		FISHERIES/2019/NOV/SWG-PEL/27rev	
	Dassen Island	**Robben Island**	**St Croix**	**Bird Island**
Chick condition	■	■		
Max distance	■	■	■	■
Chick survival	■	■		
Chick growth	■	■		
Forage length	■	■	■	■
Forage duration	■	■	■	■
Fledging success	■	■		

It would be useful to obtain a justification for which data and results are, and/or should be, relevant for future management deliberations. The 2015 international panel report made some recommendations in this regard, specifically:

i. All six response variables should be assessed with respect to how reliably they are sampled and how informative they are regarding potential fishery effects on population growth rates.

ii. A response variable should not be considered further if there is no (objective) way to determine a threshold for it.

iii. If a particular response variable is sub-ordinate or directly correlated with another then there may be little to be gained by considering it further.

Points (i) to (iii) need to be applied to all the variables listed above to produce an agreed set of variables for use in management deliberations to pre-empt and avoid different researchers making difference choices about which variables to include in analyses.

R8: I can only comment on why I have used the datasets that I have analysed.

Chick survival and fledging success are correlated (when compared like for like) – see Sherley (2020e; FISHERIES/2020/SEP/SWG-PEL/85). Fledging success cannot be analysed using an individual data-based approach, so I have favoured using chick survival since 2012 (Sherley 2012a, FISHERIES/2012/SWG-PEL/ICTT/06; Hagen et al. 2014, MARAM/IWS/DEC14/PENG/A3) because chick survival estimates come with measures of precision (Sherley et al. 2015) while fledging success estimates do not acknowledge this uncertainty (essentially observation error). Chick survival is one of the key data sources that I have argued should be used because they can be analysed using an individual data-based approach, as endorsed by the 2019 IWS panel: "*Given the nature of the experiment, use of individual data is to be preferred*" (Die et al. 2019).

Max. distance, foraging path length and forage trip duration are all broadly correlated with one another, though the exact strengths of the correlations differ. Indeed, they are all derived from exactly the same sampling process (each of the three datasets are derived from an individual GPS logger deployment, see the Supporting Information in Campbell et al. 2019 if you need further information on this), so two of the three could certainly meet the definition of being sub-ordinate. Maximum distance generally has a more straight forward empirical link to resource availability in the literature, in that it has long been known to be responsive to colony size (e.g. Wakefield et al. 2013, Jovani et al. 2016), itself an obvious predictor of resource competition. Maximum distance (not trip duration or path length) was selected for the most recent analyses (Sherley et al. 2019 and FISHERIES/2020/JUL/SWG-PEL/53REV) for several reasons.

First, because it has recently been found to respond to local prey abundance (measured by the small-scale surveys) around Robben Island (Campbell et al. 2019), while foraging path length and forage trip duration have been found to show threshold-like, or bell-shaped responses to local prey abundance at St Croix and Bird Island (McInnes et al. 2019). In other words, it is difficult to say conclusively whether a 'long' path length means there is lots of competition for food (or little food around) and the penguins are working hard to find food, or the opposite there is little competition for food (or lots of food around) and the penguins are taking advantage to 'stock up' by staying out a long time and foraging a lot (see also R14 below for more context from the literature on this issue). See Figure 1 below. McInnes et al. (2019) found almost no difference in path lengths at the very highest local prey abundance and the very lowest (Figure 1, bottom left panel) and only identified thresholds in response to the very highest local availability or abundance. Remember, penguins are feeding themselves while at sea too, not only provisioning for their chicks. So not all the time at sea necessarily results in a net energy loss.

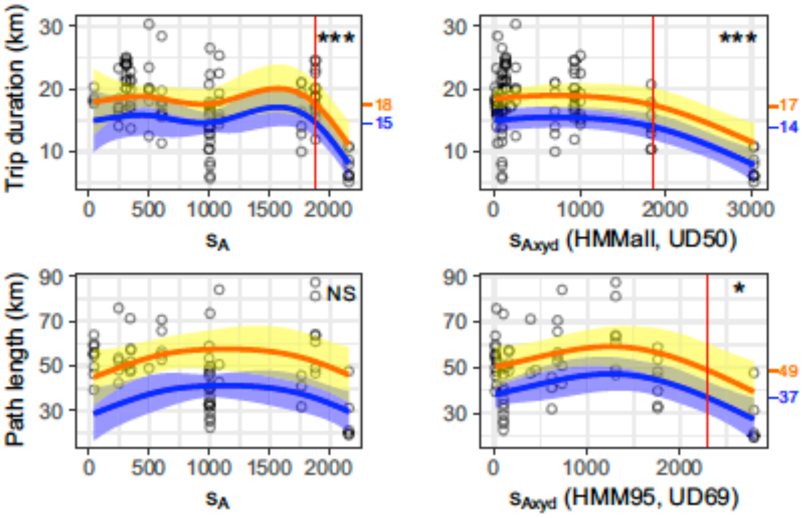


Figure 1: An excerpt from Figure 3 of McInnes et al. (2019) showing relationships between foraging effort indices of African penguins (trip duration: n = 94, path length: n = 55) and prey abundance. Prey abundance estimates: sA includes all prey, sAxyd for targeted prey. Critical thresholds (vertical red lines) in the response are given on the right axis (blue—Bird Island, orange—St Croix Island). Shaded ribbons denote 95% confidence intervals: blue—Bird Island, yellow—St Croix Island. Asterisks denote nominal p-values: *p < 0.05, ***p < 0.001.

Second, the sample sizes are always smaller for path length and (to a lesser extent) trip duration than maximum distance because of the way the GPS devices sample. They obviously loose signal when the penguins are underwater, and take some time to acquire it again when the penguins surface. So sometimes large portions of the trip can be lost, particularly the commute in and out from the colony when the penguins are ‘porpoising’ and moving quickly (sometimes the batteries may die and lose the return commute too, though this is rare). Path length is generally the most problematic of the three because the tracks can often end up with many missing locations. To accurately calculate path length, these gaps must be interpolated across somehow. Lots of methods exist to do this (some of the best using Hidden Markov Models), but to date the island closures dataset has only been analysed using linear interpolation (due to time and personnel constraints). So, I have generally not used path length and trip duration as they require more assumptions than maximum distance.

Third, and most importantly, the path length and trip duration datasets both violate the assumptions of homogeneity of variance (substantially in the case of trip duration) at Dassen Island and trip duration does so at both Eastern Cape islands (although the variance ratios are not actually that bad in this case) based on the Fligner-Killeen test (Fligner and Killeen 1967), which is one of the tests for homogeneity of variances which is most robust against departures from normality (the data are Gamma, or log-normally distributed). The maximum distance dataset does not have this issue. This makes comparisons of the mean difference between open and closed years highly questionable for path length and trip duration, as it certainly violates one of the central assumptions underlying the generalised linear model (that the groups being compared are drawn from distributions with the same variance):

Western Cape, Trip Duration
Dassen Island: Closed years variance = 338.2 (CV = 82.1%), Open years variance = 51.0 (CV = 51.7%), variance ratio = 6.6. Fligner-Killeen test result: χ^2= 35.7, df = 1, p < 0.001.
Robben Island: Closed years variance = 92.6 (CV = 56.1%), Open years variance = 70.7 (CV = 52.3%), variance ratio = 1.31. Fligner-Killeen test result: χ^2= 3.8, df = 1, p = 0.052.

Eastern Cape, Trip Duration
Bird Island: Closed years variance = 37.6 (CV = 41.1%), Open years variance = 61.1 (CV = 45.5%), variance ratio = 0.61. Fligner-Killeen test result: χ^2= 8.5, df = 1, p = 0.004.
St. Croix Island: Closed years variance = 99.8 (CV = 44.6%), Open years variance = 50.0 (CV = 32.5%), variance ratio = 2.0. Fligner-Killeen test result: χ^2= 6.3, df = 1, p = 0.012.

Western Cape, Path Length
Dassen Island: Closed years variance = 3427.5 (CV = 83.4%), Open years variance = 829.1 (CV = 51.5%), variance ratio = 4.1. Fligner-Killeen test result: χ^2= 11.0, df = 1, p < 0.001.
Robben Island: Closed years variance = 805.7 (CV = 59.4%), Open years variance = 437.7 (CV = 43.9%), variance ratio = 1.84. Fligner-Killeen test result: χ^2= 0.61, df = 1, p = 0.435.

Eastern Cape, Path Length
Bird Island: Closed years variance = 289.8 (CV = 39.3%), Open years variance = 307.2 (CV = 40.6%), variance ratio = 0.94. Fligner-Killeen test result: χ^2= 2.4, df = 1, p = 0.12.
St. Croix Island: Closed years variance = 457.2 (CV = 36.4%), Open years variance = 455.4 (CV = 28.9%), variance ratio = 1.0. Fligner-Killeen test result: χ^2= 0.06, df = 1, p = 0.81.

Western Cape, Max. Distance

Dassen Island: Closed years variance = 231.3 (CV = 89.5%), Open years variance = 69.8 (CV = 49.8%), variance ratio = 3.3. Fligner-Killeen test result: χ^2= 0.34, df = 1, p = 0.56.

Robben Island: Closed years variance = 58.0 (CV = 62.4%), Open years variance = 66.4 (CV = 65.8%), variance ratio = 0.87. Fligner-Killeen test result: χ^2= 0.25, df = 1, p = 0.61.

Eastern Cape, Max. Distance

Bird Island: Closed years variance = 66.2 (CV = 55.4%), Open years variance = 46.6 (CV = 49.3%), variance ratio = 1.42. Fligner-Killeen test result: χ^2= 0.24, df = 1, p = 0.63.

St. Croix Island: Closed years variance = 66.1 (CV = 38.6%), Open years variance = 88.5 (CV = 31.4%), variance ratio = 0.75. Fligner-Killeen test result: χ^2= 3.65, df = 1, p = 0.06.

So, an argument could be made to replace max. distance iteratively with foraging path length and forage trip duration and recalculate the Overall Effect Size in FISHERIES/2020/JUL/SWG-PEL/53REV, or include all three together, but not give them all the same weighting as other metrics. The latter would require a degree of subjectivity and the former would require an approach that can test for a difference while accounting for heterogeneity of variance (e.g. GLS models).

But, it would also seem to be that if there is an effect on maximum distance at St. Croix that is credible, biologically meaningful and agreed between FISHERIES/2020/JUL/SWG-PEL/53REV and FISHERIES/2019/NOV/SWG-PEL/27rev, it would be a mistake to ignore that unless there were equally credible and equally biologically meaningful effects on path length and forage trip duration that went in the opposite direction; according to FISHERIES/2019/NOV/SWG-PEL/27rev that is not the case.

Chick condition and chick growth could also be argued to be somewhat sub-ordinate to one another. Chick condition has been used in recent analyses primarily because the chick growth dataset has not been updated since 2014 (~halfway through the experiment). The primary reason for this was disturbance to the birds when collecting it (they are an endangered species, so we try to minimise this). Chick growth requires weighing and measuring the same chicks repeatedly; this appeared to result in more birds moving their chicks away from their original nest site. Chick condition takes a random sample of chicks and each nest is only used once per season. Chick growth data is also difficult to analyse in the sense that a growth model, usually Gompertz or logistic for birds, needs to be applied to each of n individual pair of measurements to generate a growth coefficient (see e.g. Sherley 2010, Sherley 2012b or Bonato et al. 2013 for methods applied in this case). The $n(n-1)/2$ growth coefficients for each chick are then aggregated by taking their median to give average growth coefficients that describes the chicks growth across all the time period it was measured (Sherley 2012b). And then in the analysis in FISHERIES/2020/JAN/SWG-*PEL/09*, these medians are then aggregated again (by taking the median of the medians) to give annual medians for each island-year in the dataset. So, analysing growth the way it is done in FISHERIES/2020/JAN/SWG-*PEL/09* requires doing statistics on statistics on statistics (or a 3-step approach) without taking account of any of the error associated with the different estimates at each hierarchical sampling level (see comments on the issue with a two-stage approach in Sherley 2020e). A more appropriate approach would require building a Gompertz or logistic growth model directly into the likelihood of a hierarchical model to compare open and closed years. This could be done within a Bayesian framework, but would be time consuming to develop and has yet to be implemented (this is one of the reasons the growth analysis in Sherley 2010 remains unpublished in the peer-reviewed literature). An easier alternative would be to use a Bayesian approach to propagate the uncertainty at each stage through to the final analysis (*sensu* Sherley et al. 2015), though again this has not yet been implemented.

9. In relation to the tables of data used in FISHERIES/2020/JUL/SWG-PEL/***53REV***, the possibility that particular results have been favoured because they provide a particular result needs either to be dispelled, or a Bonferroni adjustment should be used in assessing the statistical significance of results.

R9: See R8 above about why particular datasets have been used, including their empirical links to prey availability in the literature, ability to model them in an individual framework and other analytical difficulties.

And I am not clear on how a Bonferroni adjustment would apply in a Bayesian framework, but I am happy to learn if you could explain this in more detail?

Note also, that I am not testing statistical significance in the traditional, frequentist sense in FISHERIES/2020/JUL/SWG-PEL/53REV but presenting entire posterior distributions to make the uncertainty associated with effects clear and evident. The problem of multiplicity (if that is the issue here?) is "*usually discussed within the hypothesis testing paradigm, and most solutions that have been proposed (e.g. Bonferroni correction) are designed for p values*" (Sjölander and Vansteelandt 2019). The move away from p-values to communicating uncertainty is in keeping with the direction of travel in modern-day statistics (see the recent special issue of 43 papers on this topic in the American Statistician, headed with the editorial "Moving to a World Beyond "p < 0.05" (Wasserstein, Schirm and Lazar 2019).

While it might be possible to use a single, hierarchical model, with a single hyperprior specifying the prior probability of a correlation between each of these metrics as a way to correct for the fact that the prior probabilities (and so the posterior probabilities) of these metrics are in some sense dependent (e.g. Westfall et al. 1997), we have little direct information on the prior correlation of these metrics with which to specify that prior (and there is no guarantee such a model would be tractable). We could use the correlation between the observed datasets, but this is somewhat circular and, I would guess that we would not get consistent answers across colonies, necessitating the need for multiple, very time-consuming model runs to determine sensitivity. These procedures are sensitive to the prior choice of the correlation coefficient prior (Scott and Berger 2006), again suggesting the need for many model runs, and may be more trouble than they are worth with Bayesian approaches; Scott and Berger (2006) note that "One of the attractions of the Bayesian approach is that there is no need to introduce a penalty term for performing thousands of simultaneous tests; Bayesian testing has a built-in penalty or 'Ockham's razor effect'" and that this approach does not "provide a magic answer available for the multiple testing problem".

Moreover, I would hazard that the SWG-PEL would be unlikely to agree as to whether the prior probability of fishing having a negative impact on penguins should be given the same weight as the prior probability of fishing having a positive impact on penguins. I suspect it would be hard to find agreement because one SWG-PEL member has in the past tried to use a purely theoretical and abstract mathematical argument to contend that fishing should benefit penguins by breaking up shoals of fish into smaller shoals (see e.g. Butterworth et al. 2011, Robinson 2013, Robinson and Butterworth 2014, Cherry 2014). Some observers have, on the other hand, collected empirical data showing that penguins can corral large shoal of fish (Ryan et al. 2012, McInnes et al. 2017), forage more effectively on shoals than single-prey items (McInnes et al. 2017), that larger foraging groups of penguins are more likely to be associated with larger shoals of fish (Ryan et al. 2012), and that penguins forage more efficiently when there is more prey available around their colonies (Campbell et al. 2019, McInnes et al. 2019). So, here again, we are likely to run into yet more debate about how that correlation should be specified, or we would end up with a relatively weak prior on the correlation structure, which would ultimately be over-ridden by the data.

Given the need (which I believe the SWG-PEL has expressed) to make a decision during 2020, I would suggest a more useful way forward is to focus on where multiple sets of analyses lead to the same broad inference and move forward with management in an adaptive framework. That is not to say that these points cannot be revisited, but I suspect there is more to gain from address the issue of how these observed effects translate into population-level change. Although that too is non-trivial in some cases (see R14).

10. The use of an overall closure effect, as presented in Figure 5 of FISHERIES/2020/JUL/SWG-PEL/***53REV***, needs more care. The international panel report of 2016 submitted some ideas about this, pointing out that some variable are relevant to fledgling success while other may be relevant to adult survival ("*…while chick condition and chick growth are likely correlated, chick condition/growth and fledgling success affect processes that are sequential in the life history of penguins, which means that a fishery effect on each of chick condition/growth and fledgling success in combination could lead to a biologically meaningful population effect. Moreover, increases in forage trip length due to fishery impacts may have negative consequences for adult survival.*"). Integration would seem to require at least some demographically sensible basis for their combination and a framework for establishing where in penguin life history their effects are active, whether on adult survival or on chick survival and fledgling success.

R10: The Overall Closure Effect does not claim to integrate the impact of these variables on penguin demography. To do that would, I agree, require some framework for establishing where in penguin life history their effects are active (see also R14) and a model of the kind presented in Sherley et al. (2018), where we did link chick condition to juvenile survival objectively. But this is not the key point. The key point is that there is, in this situation, a pre-agreed management threshold for what constitutes a biologically meaningful effect: –0.1 in the MARAM analyses and 10% in FISHERIES/2020/JUL/SWG-PEL/53REV (Cochrane 2016). It is in that context that Figure 5 in FISHERIES/2020/JUL/SWG-PEL/53REV and the Overall Closure Effect should be viewed – as a tool to support decision making by visualising the effect sizes and the uncertainty around them, relative to that specific management threshold. The Overall Closure Effect basically says, "without making any additional assumptions, what is the probability, given all the uncertainty in the penguin responses, that, overall, this pre-agreed threshold of 10% has been exceeded". It is certainly more honest than a simple account of positive and negative (or null effects), as it directly incorporates the underlying uncertainty in each estimate (i.e. a very precisely estimated effect does not carry the same weight as a very poorly estimated trend in the opposite direction; Jefferys and Berger 1992).

11. The combination of posterior distributions in the manner proposed in Figure 5 of FISHERIES/2020/JUL/SWG-PEL/***53REV*** needs to address the opposite result for 'Condition Index' at St Croix versus the result for 'Max Distance' at St Croix, before attempting to integrate these results. Invoking methods used in local fisheries management deliberations, two opposite results should not be 'averaged out'. Firstly, the reasons for this would be explored insofar possible. Failing resolution at this level, the full management implications of each would be considered and any compromise position would be developed at a late stage once the management implications were known. A similar approach is applicable here. In

addition, it is inappropriate to add up very different response variables – the demographic implications of different variables may imply the use of different weightings.

R11: This comment seems to stem from a misunderstanding; the Overall Closure Effect in FISHERIES/2020/JUL/SWG-PEL/53REV does not claim to show demographic implications, it merely reports on whether the effect overall (taking account of all uncertainty in the model outputs) meets a pre-agreed threshold for management action (Cochrane 2016). Second, it offers no evidence (from either peer-reviewed literature or even from previous SWG-PEL documents) as to why the approach in FISHERIES/2020/JUL/SWG-PEL/53REV is apparently inappropriate. It is very difficult to respond to the charge that the approach in FISHERIES/2020/JUL/SWG-PEL/53REV does not invoke "*methods used in local fisheries management deliberations*" when no example is cited of such "*deliberations*". An argument of this nature should be backed up by citations to peer-reviewed literature. And third, this comment seems to overlook the key point that the Overall Closure Effect in FISHERIES/2020/JUL/SWG-PEL/53REV is based on an approach that has been peer-reviewed and has been used to support management decisions, including in CITES (Sherley et al. 2020).

The above notwithstanding, different weightings could certainly be considered and discussed. For example, Sherley (2020c; FISHERIES/2020/SEP/SWG-PEL/89) offers an updated Overall Closure Effect in which Max. Distance and Path Length are both presented for the Eastern Cape, so both given half the weighting of the other metrics. But this approach requires either a degree of subjectivity, or clear frameworks to link these metrics to demographic change (on that see R10 above and R14 below). The approach in FISHERIES/2020/JUL/SWG-PEL/53REV actually makes fewer assumptions (and thus is more parsimonious) than using different weightings.

I also agree that the reasons for opposite results should be explored insofar as possible, where those opposite results are equally meaningful or statistically robust. On that point, see Sherley (2020c; FISHERIES/2020/SEP/SWG-PEL/89). Updating the models in MARAM/IWS/2019/PENG/P4 for chick condition in the Eastern Cape indicates that the negative effect at St Croix is not as strong or well sampled as the Max Distance effect at St. Croix. Further analysis could be done to try to understand this difference, but further analysis at this point should not be used to delay a decision (that would not be in keeping with the precautionary principle).

12. Figure 5 of FISHERIES/2020/JUL/SWG-PEL/***53REV*** duplicates the posterior distribution for the Robben/Dassen Island chick survival based island closure effect. If, as is the case here, the closure effects at Robben Island and Dassen Island have been constrained to be equal, then there is only one closure effect estimate and only one posterior distribution, not two. Therefore, notwithstanding other comments here about approaches used to combine results from difference response variables, only the one posterior distribution should contribute to the overall posterior distribution in Figure 5. This will change the overall posterior distribution presented in Figure 5 (bottom left panel).

R12: The posteriors are not duplicated. Because this is a Bayesian analysis, it is trivial to derive a posterior distribution for the difference in survival at both islands, even though the interaction is not in the model. This is possible in much the same way that it would be possible in a frequentist framework to derive a mean estimate for both islands in both closure cases with an additive only model (you can do this just by adding/subtracting the summary output in R, for example). What is usually difficult in a frequentist framework (requiring the use of the delta method, for example) is estimating the combined SE. But in a Bayesian framework, it is a simple additive expression within the model to derive a posterior probability for the percentage closure effect at each island. So, they are not identical, just very similar.

But, in the interests of full transparency, I have refit M1 in FISHERIES/2020/JUL/SWG-PEL/53REV, but including the Island × Closure Interaction within the fixed effects structure. This model (let's call it M1.I) yields the following results (results for M1 in FISHERIES/2020/JUL/SWG-PEL/53REV in square parentheses; as a reminder the closure main effect (95% HPDI) was 0.38 (021–055 for M1)):

Robben Island Closure effect (95% HPDI) = 0.37 (0.16–0.57), percentage increase = 9.8% (4.1–15.7%) [9.9% (1.1–18.2%)], percentage of posterior > 0 = 99.9% [97.9%], percentage of posterior > 10% = 46.8% [49.0%].

Dassen Island Closure effect (95% HPDI) = 0.41 (0.11–0.70), percentage increase = 11.4% (2.4–20.4%) [10.1% (0.2–19.7%)], percentage of posterior > 0 = 99.7% [97.7%], percentage of posterior > 10% = 59.8% [51.0%].

So, whether model M1 is used (Island + Closure) or M1.I is used (Island × Closure), the Overall Effect Size and the inference about whether these effects are biologically meaningful would be unchanged.

And see Sherley (2020c; FISHERIES/2020/SEP/SWG-PEL/89) for an example of the Overall Effect Size refit to include the island specific effects on chick survival at Robben and Dassen.

13. FISHERIES/2020/JAN/SWG-***PEL/09*** provides a different island closure impact on chick survivorship for Robben and Dassen Islands. Figure 5 of FISHERIES/2020/JUL/SWG-PEL/***53REV*** provides estimates under the constraint that the closure effect is equal across the two islands (at least in its Figure 5). There seems no reason for these estimates to be equal. The separate estimates for the closure effect at these islands should be presented.

R13: FISHERIES/2020/JAN/SWG-PEL/09 uses a dataset spanning a different time period to FISHERIES/2020/JUL/SWG-PEL/53REV, so the results should not necessarily be expected to be similar. FISHERIES/2020/JAN/SWG-PEL/09 used data only until 2015, FISHERIES/2020/JUL/SWG-PEL/53REV uses data until 2018. FISHERIES/2020/JAN/SWG-PEL/09 also uses the data from OLSPS (2020), which differ slightly from the annually aggregated data that would be derived from the individual data used in FISHERIES/2020/JUL/SWG-PEL/53REV. So, we should not expect the estimates to necessarily be comparable.

The above notwithstanding, R12 above, provides the separate estimates requested here. It is worth noting at this point that the estimated effect size at Dassen Island is −0.13 (or 13%, see Butterworth 2020, FISHERIES/2020/AUG/SWG-PEL/82) in FISHERIES/2020/JAN/SWG-PEL/09 and 11.4% above. These are basically comparable. Moreover, **both sets of analyses** (this document, FISHERIES/2020/JAN/SWG-PEL/09, FISHERIES/2020/JUL/SWG-PEL/53REV and MARAM/IWS/2019/PENG/P4) **agree that the effects on chick survival are biologically meaningful** (Table 1). The corresponding estimates at Robben Island are −0.04 (or 4%) in FISHERIES/2020/JAN/SWG-PEL/09 and 9.8% above. So, while there is a difference in magnitude here, FISHERIES/2020/JAN/SWG-PEL/09 concluded that "*the experiment needs to continue for 2 to 5 years before a biologically meaningful fishing effect is likely to be detected, if it is present*" for chick survival at Robben Island and FISHERIES/2020/JUL/SWG-PEL/53REV added 3 years of data to the dataset used in FISHERIES/2020/JAN/SWG-PEL/09. Thus, **these two analyses broadly agree** (Table 1). And Sherley (2020b) provides a comparable model fit to the aggregated and individual data (albeit without the interaction) and shows that both yield positive closure effects with greater than 99% probability.

14. The use of maximum forage distance as a relevant and appropriate response variable needs more justification. Intuitively, foraging energy expenditure is relevant to population level outcomes such as adult survival. On that basis energy expenditure does not equate to the maximum foraging distance, but rather to total distance travelled (i.e. forage length). Additionally, it is necessary to translate this result to a value which is demographically meaningful.

R14: See also R8. Maximum foraging distance has a strong link to resource availability in seabirds and has recently been found to respond to directly estimated local prey abundance (measured by the small-scale surveys) around Robben Island (Campbell et al. 2019). Path length and trip duration are potentially difficult to interpret for a number of reasons (including the GPS issue outlined above), not least because seabirds (including penguins) can alternate between long and short foraging trips (both in path length and duration), depending on whether they are predominately provisioning their chicks or predominately building up their own body reserves (e.g. Saraux et al. 2011). However, these long trips in time and distance travelled need not necessarily take them further from their colony if prey resources are available within their normal foraging range. Furthermore, travelling further from the colony is a well-established response to increased competition for prey (Wakefield et al. 2013, Jovani et al. 2016).

I agree that foraging energy expenditure is relevant to population level outcomes; certainly, to chick survival (see Wilson et al. 2018) and almost certainly to adult survival as well. However, no framework actually exists at present to link foraging energy expenditure directly to adult survival (though, a potential framework has been proposed for Wandering Albatross, probably the most data-rich seabird species in the world; Weimerskirch 2018). The framework in Wilson et al. (2018) might offer the potential to link foraging energy expenditure to chick survival, but a) this framework was not available the last time an attempt was made to objectively link these metrics (in the table in 8 above) to penguin demographics (Sherley et al. 2018) and we have only truly been collecting foraging *energy expenditure* data (using accelerometers and cameras, see McInnes et al. 2017) in the last few years of the experiment (since ~2017).

References cited in reply:

Bates D, Kliegl R, Vasishth S and Baayen 2018. Parsimonious Mixed Models. arXiv:1506.04967v2.

Bergh MO. 2014. Further clarification of the biases in and interpretation of regressions where catch is a predictor of penguin response. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/DEC14/Peng/A10. Pp. 1–21.

Bonato M, Evans MR, Hasselquist D, Sherley RB, Cloete SWP and Cherry IC. 2013. Ostrich chick humoral immune responses and growth rate are predicted by parental immune responses and paternal colouration. Behavioral Ecology and Sociobiology 67: 1891–1901.

Butterworth DS. 2020. A response to Sherley: FISHERIES/2020/JUL/SWG-PEL/53REV. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/AUG/SWG-PEL/82. Pp. 1–38.

Butterworth DS, Brandão A, de Moor CL and Robinson W. 2011. Claim by Pichegru et al. that marine no-take zone benefits penguins remains premature. http://rsbl.royalsocietypublishing.org/content/6/4/498.abstract/reply#roybiolett_el_82

Campbell KJ, Steinfurth A, Underhill LG, Coetzee JC, Dyer BM, Ludynia K, Makhado AB, Merkle D, Rademan J, Upfold L and Sherley RB. 2019. Local forage fish abundance influences foraging effort and offspring condition in an endangered marine predator. Journal of Applied Ecology 56: 1751–17650.

Cherry M. 2014. African penguins put researchers in a flap. Nature 514: 283.

Cochrane K. 2016. Chair's Introduction to documents from the Technical Team on the Penguin Island Closure Experiment. Department of Agriculture, Forestry and Fisheries Report: MARAM/IWS/DEC16/Peng Clos/P6. Pp. 1–4.

Dunn A, Haddon M, Parma AM and Punt AE. 2015. INTERNATIONAL REVIEW PANEL REPORT FOR THE 2015 INTERNATIONAL FISHERIES STOCK ASSESSMENT WORKSHOP 30 November–4 December 2015, UCT. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/DEC15/General/8.

Dunn A, Haddon M, Parma AM and Punt AE. 2016. INTERNATIONAL REVIEW PANEL REPORT FOR THE 2016 INTERNATIONAL FISHERIES STOCK ASSESSMENT WORKSHOP 28 November–2 December 2016, UCT. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/DEC16/General/7.

Fligner MA and Killeen TJ. 1976. Distribution-Free Two-Sample Tests for Scale. Journal of the American Statistical Association 71: 210–213.

Hagen C, Sherley RB, Steinfurth A, Pichegru L, Robinson KJ, Jarre A, Crawford RJM, Winker H, Altwegg R, van der Merwe L and McInness A. 2014. Annex 2: An evaluation of the evidence of the impact of fishing closures around breeding colonies of African Penguins. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/DEC14/Peng/A3. Pp. 1–20.

Jefferys WH and Berger JO. 1992. Ockham's Razor and Bayesian Analysis. American Scientist 80: 64–72.

Jovani R, Lascelles B, Garamszegi LZ, Mavor R, Thaxter CB and Oro D. 2016. Colony size and foraging range in seabirds. Oikos 125: 968–974.

Matuschek H, Kliegl R, Vasishth S, Baayen H and Bates D. 2017. Balancing Type I error and power in linear mixed models. Journal of Memory and Language 94: 305–315.

McInnes AM, McGeorge C, Ginsberg S, Pichegru L and Pistorius P. 2017. Group foraging increases foraging efficiency in a piscivourous diver, the African penguin. Royal Society Open Science 4: 170918.

McInnes AM, Ryan PG, Lacerda M and Pichegru L. 2019. Targeted prey fields determine foraging effort thresholds of a marine diver: Important cues for the sustainable management of fisheries. Journal of Applied Ecology 56: 2206–2215.

OLSPS. 2020. A simple restructuring of penguin chick survival data to estimate survivorship at Dassen and Robben Islands. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/JAN/SWG-PEL/06. Pp. 1–3.

Robinson WML. 2013. Modelling the impact of the South African small pelagic fishery on African penguin dynamics. PhD Thesis, University of Cape Town.

Robinson WML, Butterworth DS. 2014b. Comments on the benefit to penguins of fishing restrictions around Robben Island predicted by Weller et al. Robben Island penguin model simulations. FISHERIES/2014/MAR/SWG–PEL/ICTT/7: 4 pp.

Ross-Gillespie A and Butterworth DS. 2019. 2019 Updated GLMM results for the South Coast penguin colony foraging data. Department of Environment, Forestry and Fisheries Report: FISHERIES/2019/NOV/SWG-PEL/27rev. Pp. 1–12.

Ross-Gillespie A and Butterworth DS. 2020. Updated implementation of the Algorithm recommended by the Panel for the 2016 International Stock Assessment Workshop for assessing whether or not to continue with the penguin island closure experiment. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/JAN/SWG-PEL/09. Pp. 1–17.

Ryan P, Edwards L and Pichegru L. 2012. African Penguins *Spheniscus demersus*, bait balls and the Allee effect. Ardea 100: 89–94.

Saraux C, Robinson-Laverick SM, Le Maho Y, Ropert-Coudert Y and Chirardia A. 2011. Plasticity in foraging strategies of inshore birds: how Little Penguins maintain body reserves while feeding offspring. Ecology 92: 1909–1916.

Scott JG and Berger JO. 2006. An exploration of aspects of Bayesian multiple testing. Journal of Statistical Planning and Inference 136: 2144 –2162.

Sherley RB. 2010. Factors influencing the demography of Endangered seabirds at Robben Island, South Africa: Implications and approaches for management and conservation. PhD Thesis, University of Bristol, U.K.

Sherley RB. 2012a. Update on the breeding productivity of African Penguins (*Spheniscus demersus*) on Robben Island. Department of Environment, Forestry and Fisheries Report: FISHERIES/2012/SWG-PEL/ICTT/06. Pp. 1–8.

Sherley RB. 2012b. Report on the analysis of growth in African Penguin chicks for the Island Closure Feasibility Study in South Africa, 2012. Department of Environment, Forestry and Fisheries Report: FISHERIES/2012/NOV/SWG-PEL/ICTT/03. Pp. 1–11.

Sherley RB. 2016. Additional analysis suggested in response to differences in variance estimates between Sherley (2016) and Ross-Gillespie & Butterworth (2016). Department of Environment, Forestry and Fisheries Report: MARAM/IWS/DEC16/Peng Clos/P4. Pp. 1–4.

Sherley RB. 2020a. Some comments on FISHERIES/2020/JAN/SWG-PEL/08. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/AUG/SWG-PEL/83. Pp. 1–5.

Sherley RB. 2020b. Some observations on comparisons of fitting to the annually aggregated and the individual data, this time using JAGS and for the cases considered in FISHERIES/2020/JUL/SWG-PEL/53REV. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/SEP/SWG-PEL/86. Pp. 1–4.

Sherley RB. 2020c. Model selection results for the remaining penguin metrics that can currently be analysed using an individual data approach. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/SEP/SWG-PEL/89. Pp. 1–7.

Sherley RB. 2020d. Revisiting the key results in MARAM/IWS/2019/PENG/P4 in light of the 2019 Panel recommendations. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/JUL/SWG-PEL/53REV. Pp. 1–27.

Sherley RB. 2020e. A response to Butterworth: FISHERIES/2020/AUG/SWG-PEL/82. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/SEP/SWG-PEL/85. Pp. 1–23.

Sherley RB and Winker H. 2019. Some observations on comparisons of fitting to the annual means and the observation-level data for the cases in MARAM/IWS/DEC19/Peng/P4 that support a positive effect of the island closures experiment on African penguins. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/2019/PENG/WP3. Pp. 1–5.

Sherley RB, Winker H, Altwegg R, van der Lingen CD, Votier SC and Crawford RJM 2015. Bottom-up effects of a no-take zone on endangered penguin demographics. Biology Letters 11: 20150237.

Sherley RB, Barham BJ, Barham PJ, Campbell KJ, Crawford RJM, Grigg J, Horswill C, McInnes A, Morris TL, Pichegru L, Steinfurth A, Weller F, Winker H and Votier SC. 2018. Bayesian inference reveals positive but subtle effects of experimental fishery closures on marine predator demographics. Proceedings of the Royal Society B: Biological Sciences 285: 20172443.

Sherley RB, Barham BJ, Barham PJ, Campbell KJ, Crawford RJM, de Blocq A, Grigg J, Le Guen C, Hagen C, Ludynia K, Makhado AB, McInnes A, Meyer A, Morris T, Pichegru L, Steinfurth A, Upfold L, van Onselen M, Visagie J, Weller F and Winker H. 2019. A Bayesian approach to understand the overall effect of purse-seine fishing closures around African penguin colonies. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/2019/PENG/P4. Pp. 1–25.

Sjölander A and Vansteelandt S. 2019. Frequentist versus Bayesian approaches to multiple testing. European Journal of Epidemiology 34: 809–821.

Stroup WW. 2012. *Generalized linear mixed models: Modern concepts, methods and applications*. Boca Raton: CRC Press.

Wakefield ED, Bodey TW, Bearhop S, Blackburn J, Colhoun K, Davies R, Dywer RG, Green JA, Grémillet, Jackson AL, Jessopp MJ, Kane A, Langston RHW, Lescroël A, Murray S, Le Nuz M, Partick SC, Péron C, Soanes LM, Wanless S, Votier SC and Hamer KC. 2013. Space Partitioning Without Territoriality in Gannet. Science 341: 68–70.

Wasserstein RL, Schirm AL and Lazar NA. 2019. Moving to a World Beyond "$p < 0.05$". The American Statistician 73: Suppl. 1, 1–19.

Weimerskirch H. 2018. Linking demographic processes and foraging ecology in wandering albatross—Conservation implications. Journal of Animal Ecology 87: 945–955.

Westfall PH, Johnson WO and Utts JM. 1997. A Bayesian perspective on the Bonferroni adjustment. Biometrika 84: 419–427.

Wilson RP et al. 2018. Luck in Food Finding Affects Individual Performance and Population Trajectories. Current Biology 28: 3871–3877e.5.