When you are ready to resubmit, please upload the following:  
  
[1] A letter containing a detailed list of your responses to the review comments and a description of the changes you have made in the manuscript. Please note while forming your response, if your article is accepted, you may have the opportunity to make the peer review history publicly available. The record will include editor decision letters (with reviews) and your responses to reviewer comments. If eligible, we will contact you to opt in or out.  
  
[2] Two versions of the revised manuscript: one with either highlights or tracked changes denoting where the text has been changed; the other a clean version (uploaded as the manuscript file).

# Reviewer #1:

The description of the models used is not very comprehensive. Perhaps a supplementary information appendix exists, but this wasn’t available to download from the website. In particular, I couldn’t find how the density-dependent fecundity was implemented. There a several references, but it needs to be made explicit in the text. In addition, there’s not much information about where parameter values come from. The priors seem generally uninformative, but the parameters of egg production (h, r and g) are not attributed to any source and the fact that they are all powers of 10 suggests that they might be ad-hoc. Is this the case? In reality, these parameters would come with fairly broad priors, given the considerable difficulty in estimating them experimentally.

I’m sure that the authors know that the aggregation of observed egg burdens investigated in the first section of the methods section is not the aggregation of the worms in the population or the aggregation of eggs from a single individual. I see that the values quoted could be valuable for comparison between datasets. However, for readers who are not familiar with the technicalities, it needs to be pointed out that the egg burden aggregation is the result of both those mechanisms, and depends on the mean worm burden and any density-dependent limitations on fecundity as well. It doesn’t have a clear simple interpretation in terms of the underlying mechanisms of worm distribution and egg production. In particular, the abstract mentions the numerical values for this aggregation, calling it the ‘dispersion parameter’. Anyone not reading the paper in detail or without sufficient technical background would probably assume this refers to the aggregation of worms in the population.  
  
For the estimation of worm aggregation and mean worm burden, I think there is a significant flaw. As observed above, the individual data points are generated from two stochastic processes; the number of worms (male and female) that an individual has and the egg count generated, given that worm population. The mean observed egg count depends on the mean shehia worm burden, its aggregation and h, the mean egg production by a pair (also the fecundity mechanism). The egg count variance is a function of these values and the aggregation of egg counts, r. As a result, any inference regarding the values of mean worm burden and worm aggregation are underpinned by the parameters h, r and g. As a result, if these values are ad-hoc, so are all the absolute values of k and W. If these values do have a source, the uncertainty in the values should certainly be accounted for in the estimation process. Given that worm expulsion is not possible with schistosomes in humans, it’s certainly the  
case that the uncertainties would be large. It may be possible to ascribe significance to changes in the inferred values, but the relationship between parameters of worm distribution, parameters of egg production (h,r,g) and the statistics of egg counts in a population are quite nonlinear, so it’s not clear that that would work either.   
  
The title of the paper states that aggregation increases under intense control efforts. However, the analysis seems to focus only on comparisons between the different control regimes. Have the authors looked at the changes over time in the estimates? There are 5 or 6 time points potentially available for each shehia. It would be interesting to see if the data shows a decrease in the estimated mean worm burden over time, either over all shehias or in individual shehias.  
On the same point, I’m not sure how the changes in aggregation at the egg count across the population should be interpreted. Assuming a negative binomial distribution for worms in the population and eggs from an individual, the mean and variance of the egg counts of individuals can be directly calculated in closed form. From these, the aggregation of egg counts can be approximated using the relationship between mean and variance of the negative binomial (assuming the compound egg count distribution is near to negative binomial). Using this approximation, the aggregation parameter of egg counts from a population falls as mean worm burden falls (that is, the egg counts become more aggregated), even when all other parameters are constant (I can point the authors in the direction of the calculation if necessary). Hence an increase in aggregation in egg counts across individuals is expected when mean worm burden is reduced. The authors need to show that the increase in  
aggregation exceeds the ‘null’ increase to show that anything as happened to the aggregation of worms in the population.  
  
I assume that Figure 1A shows the maximum likelihood estimators for each of the individual shehias fitted with the size of the shapes representing uncertainty in the estimates. As a understand it, these are just independent estimates of the mean and aggregation parameters of negative binomial distributions for each shehia. Couldn’t the uncertainties be represented with credible intervals drawn from the MCMC chain or the posterior? This would give a much clearer idea of the uncertainties involved.  
  
I interpreted Figure 3 as showing the ML estimates for worm burden and aggregation for each useable shehia and time point in the data. I think the authors should include the uncertainties in these parameters recovered from the ABC algorithm. It seems possible that the uncertainties in these parameters may be pretty large, possibly on the order of the variability of the worm burden/aggregate estimates across the shehias. In that case, the egg data collected would not even be sufficient to differentiate the different worm burdens and/or aggregation parameters in different shehias.

I think the paper would be improved by a more comprehensive exposition of the summary statistics of the data and of the structure of the study that generated them. It would be informative to see perhaps the mean count plotted against the prevalence in each shehia or how the mean count changes across time under treatment. A table would also be useful to show how much data is available for each shehia. Maybe the distribution of the number of individuals sampled in each shehia and the number of egg-positives. From the small amount of summary data provided, my back-of-the-envelope calculation suggested there were only a few tens of counts for each shehia, which would be very few given the uncertainty inherent in the negative binomial distributions, but perhaps this isn’t correct. Also, given that the authors are looking at the impact of interventions on the parameters they are inferring, it would be useful to know more about the interventions. For example, what  
levels of coverage were achieved in shehias, both overall and in the adult and child age categories? Did levels of coverage change across time, etc?  
  
Reviewer #2:

The analysis performed matches their plan, and the figures are well presented.  
  
Minor point - in figure 5 I don’t understand the basis of k=f(E) line derived from the GEE. The derivations of mating probability use aggregation of the worms, so not sure why aggregation of eggs (which is complicated by additional variance from diagnostics etc) is relevant for the mating probability?  
  
  
The conclusions are supported by the data, and the relevance to the design of control strategies is thoughtfully discussed.  
  
Minor point on line 380: “The fact that Case 3 provides the best fit suggests that capturing variability and susceptibility is important to capture aggregation dynamics at very low transmission intensities”. I agree with this, although from a modelling/data perspective it may be challenging practically to separate the contribution of both. It might be useful to mention this. Also, it is not clear to what extent treating them separately or combined as one variable would have on the quality of fit to data.

The different models/cases are described statistically, but further description of the assumptions underpinning the statistical models would improve the paper, and extra clarity is needed in a few areas (see below). Otherwise, the statistical methodology and model fitting is solid.  
  
For clarity the following points should be addressed:  
  
1. The total number of observations is given, but it may be useful to also give an idea of the minimum number of people in a single sample, because the ML method for estimating k is not robust with small sample sizes (ideally >100 samples is needed for each ML estimate – estimates of k are known to be biased upwards if the sample size is too small).  
  
2. In the Bradley & May paper (ref [25]) there is a nice explanation of the situations that might give rise to Case 1 and Case 2, and it would be good to summarise these situations in the Methods or in the Discussion. In particular, Bradley&May mention that Case 1 is better suited to high transmission and Case 2 for low.  
  
3. Case 3 is described as “mechanistic in that it considers susceptibility and exposure as distinct processes with independent distributions”. It’s not clear from the explanation as to what the exact mechanism is – a modelled distribution of worm burdens should mechanistically arise from a stochastic birth-death process. I would say the model is still phenomenological, as in case 1 and 2, but with added detail.  
  
4. You should include an explanation as to why you’ve chosen the hypergeometric distribution – i.e. it is appropriate when you’re sampling from a small population and there is no replacement.  
  
5. What was the reason for fixing alpha and beta parameters to be those values in the gamma distribution? Were they chosen because they fit well? If that is the case, then perhaps it wouldn’t be surprising that Case 3 fits the data better because effectively it has two more parameters?  
  
6. What is the functional form of the DDF, I can’t find it in table 1?  
  
7. Figure 3 - As I understand it, for Case 1, kw is just the (median?) estimate of the model parameter. In Case 2, kw is again the fitted model parameter (although technically this is aggregation of each sex, aggregation of worms is double this because the sum of the two negbins is negbin with ks added). In case 3 (possibly 4?) the distribution of worm burdens isn’t negbin, and the k parameter in the model is for cercaria. Does that mean the estimates of k\_w for case 3 and 4 are from ML fits assuming a negbin distribution?  
  
8. There is a miss-match in notation for egg burden which switches between greek epsilon and latin E in the text and figures.  
  
9. The citations should precede the full stops – a lot of them are placed after a full stop.  
  
10. The literature generally uses latin ‘k’ instead of Greek kappa for the negbin aggregation parameter. Unless there is good reason for using kappa, it might be better to revert to ‘k’ for consistency.  
  
The paper presents a novel analysis of egg count data from the SCORE project, inferring the aggregation of schistosome burdens which is important for understanding transmission dynamics as elimination is approached. It is well written and will be of interest to those coordinating control efforts and modellers looking refine their models.  
  
Reviewer #3:

This is an interesting article studying aggregation of S. haematobium infection as a function of mean population egg and worm burdens. The authors are right that it is questionable whether the classic analytic estimates of the mating probability and breakpoint apply in elimination settings, especially when low prevalences are derived after many rounds of MDA. Although the paper is well-written, there are some lacunes in the description of the methodology and some analyses are incomplete.  
  
Major comments  
  
A major omission is that the GEE analyses do not account for the MDA duration, i.e. the number of rounds. Now the authors basically consider a low mean community egg/worm burden at baseline the same as one after several years of MDA. However, the process of people systematically missing treatment rounds will likely have caused the observed increase of aggregation, as these individuals maintain having their worms and worm pairs. Only by including the number MDA rounds in the GEE analyses (such as in Figure 1B), possibly in interaction with intervention type and age group, it can indeed be demonstrated that aggregation is increased by intensive control; i.e. natural stable low endemic situations may still very well have a low degree of aggregation. In fact, accounting for the number of MDA rounds in the analyses may also help to improve Figure 5, by distinguishing results for shesias at baseline from those after several years of MDA. Furthermore, as the measurements from the  
same community during MDA are to some degree dependent, it seems necessary to account for repeated measurements by including shesia as a random factor.  
  
An important weakness of the second part of the study concerns the fact that parameters h, r and g (Table 1) are considered fixed and known. First of all, no supporting information/evidence about their values is given. Also, it is not explained (e.g. in Table 1) how parameter g influences egg production per worm pair. Nevertheless, the text (and the non-negative value of g) indicates that density dependent egg production is assumed, whereas the existence of density dependent fecundity in schistosomes is still far from certain. Medley & Anderson [32] were the first to suggest this phenomenon, based on analysis of autopsy data (Cheever 1968), but they had used an incorrect statistical approach (e.g. see Gryseels & De Vlas, Parasitology Today 1996). It is essential that the current study includes a thorough sensitivity analysis by using alternative values for these and other parameters, certainly including the option of no density dependent fecundity (g=0).  
  
The methodology and quantifications underlying Figure 5 are poorly described. It is unclear why the analytical mating probabilities for Case 2 are so much different from the other three. Why would it be 0.25 maximally, where the others approximate 1.0? Case 4 does not seem to suffer from the same phenomenon, even though it is a hybrid of Case 1 and Case 2. This needs to be explained in more detail in the caption or supplements. Also, which constant value of k was used to arrive at the solid lines? And how exactly did the authors use the GEE derived association between aggregation and mean in community egg burdens for a mating function (concerning worms)? I seriously doubt whether this association, k(E), can be used in the first place.  
  
Other methodological comments  
  
There is no information about the b\_0 and b\_1 parameters (line 162) of any of the GEE derived associations between the aggregation parameter and the mean (for egg burden and estimated worm burdens). These should be provided in a supplementary table.  
  
Case 1 and 2 are nested in Case 4, with Case 4 having an extra free parameter (p). Still Case 4 is not performing better than Case 1. How is this possible?  
  
Figure 4. Is it reasonable to compare the k values of Case 1 and Case 3, as they have a different meaning: aggregation in the distribution of worms and cercarial exposure, respectively. This is also reflected in their notation in Table 1, using a W and C. By the way, this W is missing in the lower row (Worms) of the same table.  
  
The authors correctly indicate that limited diagnostic performance of egg detection is a limitation of this study. To which extent is it also a limitation that those regularly not showing up at MDA may be underrepresented in the data collection? As a consequence, I think that the authors may even have underestimated the degree of aggregation after multiple rounds of MDA.  
  
The better performance of Case 3 is now used to suggest that assuming variability in both exposure and susceptibility is important to capture aggregation dynamics at low transmission dynamics (line 380-382). However, the assumption of worms distributed together (Case 1) and an additional (binomial) process due to repeated treatment rounds, including non-participation, may likely be the real mechanism explaining that aggregation increases when approaching elimination situations. This mechanism could be studied by an extra Case 5, but it should at least be discussed, including the need to identify and treat those that miss (several) rounds of treatment.  
  
Minor comments  
  
Abstract (line 38-39): “… associated with an interquartile range increase in the mean shehia egg burden.” This is too cryptical for an abstract and should be said in simpler words.  
  
Introduction (line 82). The definition of aggregation is not fully correct. It is not just the distribution, but the extent to which a distribution is overdispersed.  
  
The introduction (line 108) indicates that the study is relevant for other helminthiases, but this is not really worked out in the discussion. For example, for onchocerciasis and lymphatic filariasis, one adult male worm is considered to be sufficient to fertilize all present female worms. I invite the authors to shed some light on the implications of their study for these infections, instead of leaving this to the reader.  
  
It may be more logical to swap Case 3 and Case 4, so the hybrid of Case 1 and 2 follows these immediately. In fact, the same order was used in their presentation in the text (lines 195-213). Moreover, the theta numbers in Table 1 already show this order, and are now not in agreement with the corresponding case numbers.  
  
Figure 1. The first line of the caption (line 280) should end with: for each shehia, , and year, . This will help to make clear that each dot reflects a shehia and time point. Also, there is a typo (line 281): Inveerse.  
  
Figure 1A. The y-axis is difficult to read and should better be presented on a log-scale, just as in Figure 3. This will also help to compare the (linear) trends in both figures. Furthermore, it would be helpful to visualize how the unstratified marginal (point) estimate of k\_IQR = 0.012 (Figure 1B) results from Figure 1A, e.g. by indicating dashed horizontal and vertical 25% and 75% quartile lines.  
  
Figure 2. The adjusted prevalence needs to be shortly explained in the caption to help the reader, who should otherwise return to the methods.  
  
The mean Bayes factor (297-300) and how to interpret it is not explained.  
  
Line 305. The reference to Fig 1, which is about mean egg burdens, is confusing here, as the text is about estimated mean worm burdens.  
  
The y-axis of Figure 5 is easier to read by using steps of 0.20 instead of 0.25.  
  
Reference 26. There is a typo in the initials of both authors. It should be RM, not RMRM.