We would like to thank the associate editor and reviewers for the many helpful comments and suggestions. This is how we addressed each comment:

AE Comments/Suggestions

General Comments

1. There's a general problem here with the assessment of normality from QQ plots and from standard hypothesis tests (AD, CVM, KS). Even for the classical LM, the latter are overly sensitive to minor departures from strict normality which does not generally compromise scientific inference, while standard QQ plots tend to be sensitive mainly to differences in the tails, and it is harder to distinguish between a general pattern of departure and one compromised by a few aberrant points. One simple way to correct this last is to plot a detrended version of the QQ plot, plotting the difference between the data quantile and the standard normal quantile on the ordinate, which has the effect of making the reference line horizontal and magnifying the vertical scale, making the pattern of departure more visible. As well, there is no discussion of the visual criteria for rejecting normality based on the QQ plots, and the simulation studies make use only of the AD tests without further comment.

   Response: We are aware of the perceptual challenges of using Q-Q plots for distributional assessment, and have added additional discussion of the assessment of normality based on Q-Q plots and their relationship to formal tests of normality as a way to help clarify their use.

   Additionally, we have looked into the use of detrended Q-Q plots, however, our initial study of their utility found that Q-Q plots with point-wise bands outperformed their detrended counterparts. Based on our initial results, we decided to present "traditional" Q-Q plots in this paper, but plan to follow-up our results in a different study, which seems outside the scope of this paper.

   Finally, we do mention in Section 4.2 that the CVM and KS tests performed similarly to the AD test and point the reader to the supplemental materials for full simulation results.

2. More generally in connection with scientific inference—tests and confidence intervals for the random effects themselves—there needs to be some discussion or results related to the impact of non-normality on these tests, for which normality may be necessary theoretically but perhaps of lesser practical importance. In connection with the simulation study, including cases with $\sigma^2_{b_0} = 0,\ \sigma^2_{b_0} = 0$ might be one way to address this in the current study; or else cite some literature on this question. This is similar to the point raised by Reviewer 2 regarding consequences of violation of normality.

   Response: We hope that we clarified the consequences by pointing to some existing literature and summarizing the findings more clearly. The main consequence of misspecification of the random effects distribution is that it can invalidate prediction intervals for quantities involving the random effects (such as mixed effects in small area estimation which was discussed by Jiang, 2001) and inference for the variance components in the covariance matrix of the random effects (which was discussed, along with an adjustment, by Verbeke and Lesaffre, 1997).

3. In the discussion (or earlier), you need to provide some more insight about why reduced-rank rotated effects are a reasonable thing to do.

   Response: we have shown that reduced-rank rotated effects do decrease the amount of confounding between levels of a hierarchical model. The approach is similar to what is standard practice in factor analysis. We have tried to emphasize this aspect more in the current form of the manuscript.

Specific Comments

1. p3 / l3: define "shrinkage"

   Response: We have added a brief definition.

2. p6 / l5: observed Q-Q plot (panel $12 + 2^2$) – why 2^2? Anyway, this panel does seem to stand out vs. the rest, with a few exceptions.

   Response: We used $12+2^2$ to inform the reader which panel was created using the observed random slopes. We use this notation in an effort to allow readers who explore figure 4 first to do so without knowing which panel is constructed using the observed random effect. The expression simply hides the answer until after referring to the plot, or solving this simple expression. We believe that this technique is important to the presentation of these lineup plots, so no change was made.

   Upon further reflection, we realized the potential problem with the expression $12 + 2^2$, since panels 12 and 4 both deviate further from unity than the true panel, 16. We have changed the expression to $3^2 + 7$ to avoid any unintentional bias that may have imposed on the reader.

3. p6 / l -2 & fig 4: "none of the null plots conform ..." Well, all are + skewed, but some panels (1, 2, 17, 18, 19) seem quite close, at least with this representation. There seems to be an implicit, but unstated criterion here that none of the points can be outside the (undefined, but probably pointwise) envelope used in this figure.

   Response: We did overstate this and have softened our language to state that most null plots do not conform to normality. One of the previous lineups we had used in a presentation we gave on this topic had no null plots conforming to normality, which explains this small mistake.

4. p 7 / Table 1 & discussion: The table provides a caution against using these standard tests for random effects, but it might be well to give some comparative results for alternatives (Lange-Ryan) designed to overcome these problems.

   Response: We decided to include those results in Section 1.5 of the supplemental materials as we are aware that our paper is a approaching the 25 page limit. We did add a reference to the supplementary materials so that a reader knows where to find such results.

5. p8 / l -4 : "equations reveals the inherent dependence" – need to be a bit more explicit here

   Response: We have added additional clarification here.

6. p 11/ Algorithm 1: This is the standard canonical transformation used widely in multivariate regression, MANOVA, canonical correlation, and need not take up space by being stated as an algorithm.

Response: Thank you for this suggestion. We have moved this procedure to the appendix so it does not take up additional space. It could also be moved to the supplemental materials.

7. p 11 / l -1: eigenvalues of A^*, U_s, i.e., minimizing (8), ~~making~~ taking

   Response: We rewrote much of this section, so this is no longer an issue.

8. p 14 / Fig 6 caption: "calculated where Q for a simulated random ..." — something wrong here.

   Response: We have fixed this issue.

9. Figs 7-9: The symbols in the legend include λ (which doesn't appear in Fig 7). Perhaps this is a font error?

   Response: We were unable to find a λ in the legend of Figure 7, so we did not make a change here. We hope that this was a one time error, but if it appears again please tell us specifically where it appears.

10. Supplementary materials: Perhaps mention HLMdiag here

    Response: In our first submission we had mentioned HLMdiag in the supplement, but it appeared at the top of page 24, while the other two items appear on page 23. We have added a link to the development version of the package on github, but made no substantial changes here since we had already included a reference.

Review 1

The discussion is of procedures to test the normality of random effects. But there seems to be a big confusion here. The distribution of the actual (unobserved) random effects is not the same as the distribution of the estimated random effects.

Consider an extreme case, with 50 groups with 1000 observations each and 50 groups with 1 observation each. For the groups with 1000 observations, the estimated random effect will be approximately equal to the true random effect. For the groups with 1 observation, the estimated random effect will be approximately 0. Thus, if the underlying distribution of random effects is normal, the distribution of estimated random effects will look like mixture of a normal with a spike at 0 containing 50% of the mass.

Response: We agree with the reviewer that the distribution of the predicted random effects and actual (unobserved) random effects differ, and we do not claim that they are in fact the same. The approach taken in our paper is to use the predicted random effects to assess the distributional assumptions made on the random effects, which, for the reason pointed out by the reviewer, is not trivial. We support this fact with the small simulation study in Section 2. The remainder of the paper discusses how we can frame an assessment of the distributional assumptions in a familiar graphical setting using the rotated random effects.

Review 2

General Revisions

1. It should briefly contextualize the role of distributional tests among other types of diagnostics based on residuals.

   Response: We have added a brief discussion of tests for normality.

2. It would be desirable to address the issue of the consequences of violating normality on inference in the model – not just testing normality for its own sake. Under what circumstances would it be crucial to verify normality and when would its violation be relatively innocuous.

   Response: We hope that we clarified the consequences by pointing to some existing literature and summarizing the findings more clearly. The main consequence of misspecification of the random effects distribution is that it can invalidate prediction intervals for quantities involving the random effects (such as mixed effects in small area estimation which was discussed by Jiang, 2001) and inference for the variance components in the covariance matrix of the random effects (which was discussed, along with an adjustment, by Verbeke and Lesaffre, 1997).

3. The theory in this paper is based on the linear model

   $$Y = X\beta + Zb + \varepsilon,\ b \sim N(0, D) \perp \varepsilon \sim N(0, R)$$

   The implementation of diagnostics in this paper is based on the lme4 package whose particular strength, for models with normal errors, is that it can fit crossed random effects. Apart from this, however, it is very restrictive in the structures of D and R matrices it can model.

   The lme function in the nlme package is much more general. On the R side, the lme4 package only allows $R = \sigma^2 I$ while the nlme package allows very rich R-side modeling including ARMA models, spatial correlation and heteroskedastic models, etc. Also, lme allows a great deal of flexibility in the G(D)-side model (non-parametric smoothing splines, heteroscedastic models, etc). Tantalizingly, the methods developed in this paper appear to be fully applicable to these lme models. It's a pity that the diagnostics are not extended to lme since whitening would work even for non-independent observations within clusters. Normality diagnostics for the random effects generating smoothing splines would be particularly interesting where they could serve as a diagnostic for the appropriateness of the spline basis. Thus implementing the methods for lme as well as lmer would extend their applicability considerably.

   Response: We absolutely agree that implementing the methods for lme in addition to lmer extends their applicability. We had not implemented many of the methods provided in HLMdiag for lme objects at the time of our first submission. Since then, we have extended most of the diagnostic functions in HLMdiag to work with lme objects, including `rotate_ranef`. We still have a few more functions to write that do not relate to the rotated random effects before we push this version to CRAN, but it is available on github.

4. The paper is quite well written except that, I think, it would need extensive restructuring to introduce concepts logically in a way that requires much less second guessing. There are too many places where the purpose of concept becomes obvious a page or two later. For example, the idea and the purpose of considering W'b should be discussed before introducing Definition 1.

Response: We have greatly restructured Section 3.2 in an effort to make the discussion clearer and more rigorous. We hope that this addresses the reviewer's comment.

5. Definition 1 and equation (6) require some statement of the conditions required. The discussion preceding equation (6) is unclear. If the s-dimensional space is 'given' then $J_1(s)$ is a constant, i.e. any two matrices W that span the same column space will yield identical values for $J_1(s)$. What is meant is 'for any given dimension'.

   Response: We have made the suggested change.

6. There are many ways of defining a fraction of confounding using different symmetric functions of the relative eigenvalues of WAW' to WBW'. There should be a discussion of the decision to use the trace and the possibility of extending the notion using different symmetric functions.

   Response: Fraction of confounding has been introduced by Hilden-Minton as the trace of the fraction, but we do agree that there are many other ways of measuring confounding between levels. We have added a discussion to that effect.

7. The approach takes D and R as known although, generally, there might be very little information on some aspects of D. What are the consequences of using this approach given that D and R are estimated?

   Response: Our original simulation study did not assume that D and R were fixed, but we have since run a simulation study where this was the case. The results between the studies are quite similar, so we decided to retain the original simulation study in the paper and add the additional simulation results to the supplement. Additionally, we added a statement explaining this choice at the bottom of page 16.

Specific Comments

1. p. 8 l. 52: V has not been previously defined.

   Response: We defined $\mathbf{V}$.

2. p. 9, l. 15: This might not be true with a small cluster size.

   Response: We reworded this sentence so that it indicates that this is the case in situations that we have considered, which have included some small cluster sizes but was not an exhaustive search.

3. p. 9, l. 38ff: The exposition would benefit greatly from some motivation at this point. Introducing the random vector $W'\widehat{b}$ whose properties are the subject of the concept of fraction of confounding, would make the Definition 1 much easier to understand.

   Response: We have greatly restructured Section 3.2 in an effort to make the discussion clearer and more rigorous. We hope that this addresses the reviewer's comment.

4. p.9,l.52: the fact that $Trace((W'BW)^{-}W'AW)$ is well defined is not trivial but it would follow from subsequent discussions. I think this should be mentioned.

   Response: We have added such a statement.

5. p. 10 l. 4: 'by definition' is misused here.

Response: We deleted "by definition" to fix this issue.

6. p. 10 l. 34: The development here is not rigorous. Depending on the value of s and on the rank of B, $(W'BW)^{-1}$ might not exist. I don't think that the development makes it clear why this all does work. The key fact is that $span(A) \subseteq span(B)$ (indeed B-A is non-negative definite) which makes everything work, following de Leeuw (1982), even though A, B, $W'AW$ and $W'BW$ might be singular.

   Response: We have rewritten Section 3.2 to be more rigorous and made sure to explicitly discuss this issue.

7. p. 11 l. 16: Here too, the fact that $span(A) \subseteq span(B)$ makes things work but that is not clear in the description of the algorithm.

   Response: We have since moved this to the appendix, but did add a clarification to resolve this issue.

8. p. 15 l.14: It might be helpful for many readers to explain why this is reasonable: i.e. that the predictor variables can be transformed affinely to make this true. However, this also suggests that these methods would work best if X is transformed to achieve a nearly diagonal D matrix. Perhaps this should be discussed.

   Response: We did not feel like we had the space necessary to go into this discussion, since, as the you have suggested, there would be two points of discussion and we are already "up against" the page limit.