Author's response to reviews

Title: Fitting multilevel models in complex survey data with design weights: Recommendations

Authors:

    Adam C Carle (adam.carle@unf.edu)

Version: 2 Date: 6 February 2009

Author's response to reviews: see over
February 6\textsuperscript{th}, 2009

Re: MS: 6698644172274584

Melissa Norton, MD Editor-In-Chief
BMC Medical Research Methodology

Dear Dr. Norton,

Attached please find a copy of my revised manuscript entitled “Fitting multilevel models in complex survey data with design weights: Recommendations.” I am submitting this revised manuscript for you to consider publishing in \textit{BMC Medical Research Methodology}. Four reviewers indicated that the original manuscript reported important findings. Additionally, several reviewers indicated that I had written the paper well and at an appropriate level for the intended audience. For example, Reviewer 2 felt “this article is a solid foundation and will be a good reference. The author has a very structured/organized layout that has his story flowing nicely. As a statistician, I feel the author does a good job at balancing between too technical and not technical enough. So he provides sufficient formulas and explanation of model development but does not go overboard, which otherwise could intimidate the novice reader.” Likewise, Reviewer 3 noted that I used a “compelling literature review” and that “any analyst who is working with complex designed data will need to read this paper and consider adopting one of the three recommended strategies.” However, some concerns tempered some of the reviewers’ enthusiasm. I appreciated these critiques, felt they would strengthen the study, and incorporated all of them into this revision. Let me detail these changes:

\textbf{Reviewer 1 Sophia Rabe-Hesketh:}

1. …The word “strata” should not be mentioned here (p. 3) because the paper clearly does not address any issues related to strata.

   I have removed strata from this section and the paper entirely.

2. The first paragraph of Section 1.1 should be improved considerably by mentioning that the clusters are typically sampled in the first stage, followed by sampling of units within clusters in the second stage, etc., actually describing some survey designs.

   I have done this. The third sentence (p. 3 line 6 in the revision) of the first paragraph now reads,
“For example, a survey may first identify clusters (e.g., all counties within an area), sample the clusters (i.e., select some but not all of the counties), and then select units within the clusters (e.g., people within a county).”

3. The traditional approaches treating the “clustered nature of the data as a nuisance variable” should be explained as adjusting the standard errors for the sampling design.

I have modified the sentence to provide further explanation. The sentence (p. 3 line 13 in the revision) now reads,

“Analysts have traditionally used techniques that treat the clustered nature of complex survey data as a nuisance variable by adjusting the standard errors for the sampling design.”

4. The last sentence of p.3 seems to imply that only MLM produces correct standard errors, but so do the traditional survey approaches.

I have clarified this section. I did not intend to imply that only MLM procedures produce correct standard errors. I now clearly state the traditional survey approaches also produce correct standard errors. This sentence (p. 3 line 15 in the revision) now reads,

“This method delivers correct standard errors and properly accounts for non-independence.” [“this method” refers to the traditional survey techniques discussed in the preceding sentence.]

5. It is not correct that the traditional survey approaches “fail to allow analysts to investigate what predicts variation across clusters… MLM adds … estimates of the between-cluster variance not accounted for by the covariates.

I have made this sentence more precise. I now indicate that MLM allow analysts to investigate what predicts variation across clusters and estimate the between-cluster variance not accounted for by the covariates. This sentence (p. 3 line 16 in the revision) reads,

“However, it fails to allow analysts to investigate what predicts variation across clusters and examine the amount of between-cluster variance unaccounted for by predictors included in the model.”

[“it” refers to the traditional survey techniques discussed in the paragraph.]

6. A .. point made by … papers cited in Section 1.2.1 is that the scaling of weights does not matter if the cluster sizes are large….This should at least be mentioned….

I now mention this. In the second to last and last full sentences on p. 6 line 20 in the revision, I include the following two sentences:
“For example, as cluster sizes increase, the estimates generally become less biased.[1, 9, 10] This suggests that with sufficiently sized clusters, an analyst may worry less about scaling the weights.”

7. On page 6, line -6: It is not true that failing to scale the weights consistently resulted in biased estimates. ….

I have improved this section as indicated. I have removed “consistently.” Additionally, I have added language to indicate that bias becomes particularly pronounced with small cluster sizes. This section (p. 7 line 5 in the revision) now reads,

“… simulations point to a need for some type of scaling if using weights, especially with small cluster sizes...Including the weights but failing to scale them (i.e., including them as “raw” weights) results in biased parameters and standard errors, especially with small cluster sizes.[10]”

8. For this same reason, it is also not correct that method A “increases its advantage as cluster size increases” (p.7, line -3).

I clarified this statement to indicate that method A has an advantage in moderately sized cluster sizes. I now state (p. 8 line 12 in the revision),

“…as cluster sizes increase (n > 20), method A appears to increase its advantage,[1] though bias decreases substantially for all methods as cluster sizes become sufficiently large.[1, 9, 10] Thus, when working with cluster sizes larger than n = 20 and a concern that insufficient cluster size may lead to biased estimates, analysts may wish to report method A’s results.”


I now include this paper in my review. See for example pp. 4 and 6.

10. The author should read all the papers cited in Section 1.2.1 again carefully to make sure that this section is a faithful review of the literature.

I have reread each paper cited in 1.2.1 again. I have also read each new paper I include in this section at the reviewer’s request (e.g., Grilli & Pratesi, 2004). To the best of my knowledge, my revised manuscript faithfully review the literature as it relates to the purpose of this paper.

11. The paper should say something about the estimation methods used by different software. Which [estimation] method was used in which software? If quadrature was used, how many quadrature points were used?
I have included this information for each program in section 2.1.1. There, I note that, given that many analysts will use their chosen program’s default settings, I used each program’s default settings in the analyses reported in my manuscript. This section (p. 12, line 6 in the revision) now reads,

“Thus, for Mplus, I used MLR for both the continuous and categorical analyses. MLR delivers maximum likelihood parameter estimates with robust standard errors computed using a sandwich estimator. For categorical outcomes, MLR uses numerical integration and adaptive quadrature using 15 integration points per dimension.[21] For MLwiN, I used the default iterative generalized least squares (IGLS) estimator for both continuous and categorical outcomes. For the categorical outcome, I used default 1st order marginal quasi likelihood (MQL) estimation. By default, MLwiN provides robust standard errors for models incorporating design weights. For unweighted analyses, MLwiN does not provide robust standard errors by default. I used these settings.[22] For GLLAMM, I used adaptive quadrature with 8 quadrature points (the default). By default, GLLAMM provides robust standard errors computed using a sandwich estimator for models incorporating design weights. For unweighted analyses, GLLAMM does not provide the robust standard errors by default.[23, 29] I used these settings.”

12. ...the idea of a pseudolikelihood should be mentioned.

I now mention the idea of a pseudolikelihood approach. And briefly describe how design weights lead to problems under this approach. I now include the following (p. 4 line 18 in the revision),

“...MLM analyses that incorporate sampling weights use a pseudomaximum likelihood estimation approach.[19] The pseudomaximum likelihood estimator is a function of sums across level-1 and level-2 units rather than a straightforward sum of level-1 units.[20] Thus, one must take special care to include design weights.”

13. There are many errors in the list of references.

I have reviewed and updated the entire reference list. The errors in the initial submission occurred because I used a reference manager for the first time for this article (at BMC’s request). I failed to check the format of the references output by the reference manager because (honestly) it did not occur to me that the output list would differ from y database. My database correctly listed the references, however, the list output by the manager did not reflect what I expected. I apologize. This will not occur. Again, I have reviewed and completely updated the entire reference list.

14. p.7, line 1: Insert “scaling of” before “level-2 weights”

I have inserted “scaling of” before “level-2 weights” (p. 7 line 16 in the revision).
15. p.8, line 9: Grilli and Pratesi (2004) trick SAS into using the weights correctly. This should be mentioned here.

   I now mention Grilli and Pratesi (2004) here. This section (p. 9 line 3 in the revision) now includes the,

   “…though Grilli and Pratesi[20] developed a relatively complicated method to “trick” SAS NLMIXED into properly handling weights under some conditions (e.g., models with no more than two levels)…”

16. p.21, line 1: “Several estimators”: Be more specific.

   I now specify the estimators (p. 18, line 12 in the revision)

   “…e.g., iterative generalized least squares (IGLS), restricted IGLS, and Markov chain Monte Carlo (MCMC)…”


18. P.21, line -3: It is not true that GLLAMM offers no diagnostic tools “of its own”. The predict command provides residuals of different kinds, as well as influence diagnostics. An advantage of GLLAMM is that it runs in a package with professional graphics and other important tools making diagnostics easy.

   I have changed this. I have combined this statement with the previous statement and now state that GLLAMM offers “few” diagnostic tools, rather than “no diagnostic tools,” and, subsequently, I highlight that users familiar with STATA can easily incorporate STATA commands. This sentence (p. 18 line 22 in the revision and reads,

   “….GLLAMM offers few automatic features (e.g., automatic grand or group mean centering) and diagnostic utilities. However, users familiar with STATA will find it easy to incorporate STATA commands, data manipulation, and diagnostic tools when using GLLAMM....”
Reviewer 2: Robert Gallop
Overall, I believe this article is a solid foundation and will be a good reference. The author has a very structured/organized layout that has his story flowing nicely. As a statistician, I feel the author does a good job at balancing between too technical and not technical enough. So he provides sufficient formulas and explanation of model development but does not go overboard, which otherwise could intimidate the novice reader.

Major Compulsory Revisions:
1. ... it feels like the reader may still be caught in the middle of the street trying to decide between software and weighting/not weighting. You provide solid information about the pros and cons. You discuss how sometimes the effects differ and sometimes they don't. ... I think the average reader is looking for guidance/recommendations. It would be good to summarize with what your standard practice is: such as weighting method do you use (A or B), the software you use, do you ever consider unweighted rather than weighted. ... I think the reader would benefit on your decisions when faced with analyzing such data.

I have added a new section toward the paper's end entitled “6.1 Applied Summary Recommendations.” Here, I discuss the software program I use, the scaling method I adopt, and the scaling method I report. I also discuss the conditions under which I would fit and report the results of an unweighted model. Finally, I give my reasons for these decisions. This section begins on p. 21 line 8 in the revision and reads,

“Given the breadth of findings discussed and presented and the various strengths and weakness of each approach and software program, the reader might now wonder, “what do I do in practice?” In my work, I take the following standard approach. First, in terms of software, I generally use Mplus. I do this because Mplus offers the most flexibility relative to speed. I frequently fit models that MLwiN cannot estimate (e.g., MLM multiple group structural equation models) and I rarely fit models with more than two levels (which Mplus currently cannot estimate). Analysts fitting the types of models discussed in this paper will generally find MLwiN more than meets their needs. Second, in terms of scaling the weights, I always fit the models using each scaling technique (methods A and B). I do this to examine any inferential discrepancies. If I find no inferential discrepancies, I generally report the findings from method A. I do this because I frequently work with cluster sizes larger than n = 20 and I am interested in both point estimates and variance-covariance discussions. If I worked with smaller cluster sizes (n < 20) and had an interest primarily in variance-covariance estimates, I would report the results of method B. If I had an interest primarily in point estimates, I would report the results of method A. Finally, if I encountered a model I could not estimate for some reason using scaled weighted data, I would take the following approach. I would fit a simpler model using scaled weighted and unweighted data. If I did not observe a difference in the inferential conclusions across these approaches, I would then fit the more complex model using unweighted data. However, I would include a note in my reporting of the unweighted findings highlighting that I used unweighted data and that readers should interpret the results with caution.”
Minor Essential Revisions:

2. On p. 2. Line 5 of the abstract "This article examines the performance....". Would it be more clear to describe when you say variety of MLM you are really talking about the weighting structures within the model and not various MLM models?

   I have clarified my description of the study. This line, the final line of the Background section in the abstract (p. 2 line 7 in the revision), now reads,

   “This article examines the performance of scaled-weighted and unweighted analyses across a variety of MLM and software programs.”

3. On p. 2 first line of Results and conclusions, "Scaled estimates...." should it be "Scaled weighted estimates....". Otherwise it seems confusing.

   I have made this change (p. 2 line 14 in the revision).

4. On p. 3 line 8 "This in turn, can lead to inferential errors". Namely you’re talking about Type I error, right? You should discuss what this means ....

   I have clarified that I am discussing Type I errors and that this leads to incorrect rejection of null hypotheses. This sentence (p. 3 line 12 in the revision) now reads,

   “This, in turn, can lead to increased Type I errors, causing analysts to incorrectly reject null hypotheses.”

5. On p. 4 line 9 from the top "This results for several reasons" this line does not seem grammatically correct.

   I have modified this sentence to improve its readability. The sentence (p. 4 line 15 in the revision) now reads,

   “Limited adoption occurs for several reasons.”

6. Last line on p. 5 perhaps say "For further discussion on the mathematical derivations associated with this work, readers should consult...."

   I made this change. This sentence (p. 6 line 10 in the revision) (now at the top of page 6) currently reads,

   "For further discussion on the mathematical derivations associated with this work, readers should consult....”

7. With page 6 or even somewhere else, does there need to be a discussion concerning missing data within each cluster, whether its missing outcomes or predictors or a combination of both. How
does "missing data" impact the weighting procedures? Missing data is so evident in our data, how
is this handled with the weighting.

The reviewer highlights an important problem. The work I review here has not addressed
missing data. A well developed literature does not yet exist regarding missing data, design
weights, MLM, and estimation issues simultaneously. To address this concern, I have added a
section in the limitations and discuss missing data as a remaining issue to be addressed.
Additionally, I offer one possible solution. This addition begins on p. 20, line 18 in the revision
and reads,

“Finally, little work addresses missing data’s role in MLM with design weights. It remains
unclear how to best handle missing data within the context of MLM, complex survey data,
and design weights. Analysts might take a zero-weighting approach for missing data,[33]
treating individuals with complete data as a subgroup, to address missing data. If one uses
this approach, the analyst should take special care to scale the weights using the full set of
weights. Future work should explore missing data’s role and develop and test solutions to
handle it.”

8. On p.8 in discussion of the software packages, does SAS SurveySelect, SurveyMeans, SurveyReg
accomodate any of this?

No. I have added a sentence to indicate this (p. 9 line 5 in the revision). This sentence reads,

“One cannot fit MLM using SAS Survey procedures.”

9. Tables on p. 13 -18 are a lot of information. Is there an easy way to summarize the difference
between Software packages and weighting schemes within software packages? In simulating data
and running statistical models, one measure we worry about is %bias. Could there be a
comparable index to summarize these tables more efficiently, as compared to just "eyeballing" the
regression coefficients?

I appreciate the reviewer’s concern regarding the volume of information presented. To my
knowledge, no developed summary index exists for the type of information I present in these
tables. Thus, I cannot summarize the tables with an index. Additionally, Reviewer 4 felt
strongly about the detail presented in these tables (in that the detailed information delivered
important points) and recommended including additional information. This makes it even
more difficult to reduce the information. However, in an effort to make the information more
accessible, I have included additional language in the results and discussion section describing
and summarizing the tables’ results. There, I include more language highlighting and
discussing the main observed differences in the tables. The additional language begins on p.
13, line 2 in the revision and continues to p. 15, line 20.

10. In the software strengths and weakness portion, I think it would be better to first layout all the
strengths for the software packages followed by the weaknesses....
I have made this change. Using the same language, but in a different order, I now layout the strengths of each software package. I follow with the weaknesses. This change begins on p. 16, line 17 and ends on p. 19, line 4.

11. Your discussion of subgroup analyses on pp.22-23, can interaction models be fit where contrasts are done within the interaction, to focus on hypotheses within the specified subgroup.

Yes. One can include interaction terms to focus on hypotheses within specified subgroups. However, analysts often wish to examine subgroups of the sample and exclude members of the sample entirely. My discussion refers to the situation when analysts wish to select a subset of individuals from the sample and conduct analyses within this subgroup only. I have clarified this text. The section (p. 20 line 5 in the revision) now reads,

“…analysts often wish to investigate relationships within a certain subgroup. Although analysts can use interaction terms to investigate hypotheses within the specified subgroup, analysts may wish to examine a subgroup of the sample excluding other sample members entirely. For this situation, where analysts wish to investigate hypotheses among a specific subgroup only, no established guidelines exist regarding a best practice method for estimating MLM in complex survey data with design weights. When using complex surveys, one should include the entire sample in the analyses. This leaves the sample design structure whole and leads to proper estimation of variances and standard errors. However, it presents a problem when analysts would like to select a subgroup and examine a MLM for this subgroup of individuals in a sample. Analysts should not simply subset the data to the desired group of interest.[32, 33] ….”
Reviewer 3: Byron Gajewski

Using a compelling literature review, the author of this submitted paper argues for incorporating both MLM and scaled weights when analyzing data from complex surveys. I congratulate the author on applying this approach to an actual dataset from the National Survey of Children with Special Health Care needs using an array of popular MLM software programs...Any analyst who is working with complex designed data will need to read this paper and consider adopting one of the three recommended strategies.

Discretionary Revisions (which are recommendations for improvement but which the author can choose to ignore)


   In the original and revised manuscript, I define level-1 and level-2 where I first mention them (at the bottom of page 4 in the revision). In the section to which the reviewer refers (on page 7 of the revision) (p. 7 line 12 in the revision), I have inserted parentheses as follows to re-define these terms and remind the reader of their previous definition,

   “This confounding of level-1 (individual) and level-2 (cluster) design...”

2. Page 12, line -8. I think it should be “…the design weights ‘are’ correlated…”

   I have fixed this sentence. It should have (and now does) read (p. 15 line 13 in the revision),

   “…where the design weights correlate with the outcome...”

3. Page 23. I don’t know what the sentence “When using complex surveys, one should include the entire population in the analyses” is referring to.

   I corrected this statement. I should have (and now does) read (p. 20, line 10 in the revision),

   “When using complex surveys, one should include the entire sample in the analyses.”

4. Page 25. Appendix A. The author mentions using “multiple imputation (MI).” MI traditionally implies that there are several imputed datasets. Why weren’t standard Rubin formulas used for combing several datasets? Please explain.

   As suggested by Pedlow, Luke, and Blumberg (2007), I used a single imputed value included on the data file that NCHS has entitled “multiple imputation” for these analyses. In my revision, I have clarified that NCHS labels the file from which I selected, that I use a single imputed value provided by the National Center for Health Statistics, and that I am following a suggestion indicated by the authors of this data. Because this paper focuses on MLM and because Pedlow, et al. support using a single imputed value for these data, for simplicity’s sake, I chose not to complicate the discussion with a treatment of using multiply imputed
data. I include the reference to Pedlow, et al. in the revision (p. 27, line 8) and this section now reads,

“….available on the data file NCHS entitles “multiple imputation”. NCHS labels the family income variable POVLEVEL_I. For these analyses, I used a single imputation, imputation 1, as suggested by Pedlow, et al.[34] “

5. Please provide an appendix that defines the parameters in Tables 1 & 2 using traditional MLM equations.

Appendix C, (starting on p. 30, line 1 in the revision), now presents these equations. The following sentence on p. 12, line 1 in the revision directs readers to the Appendix.

“Appendix C presents traditional MLM equations for each model I estimate.”

6. The writing, at times, is verbose … the author is a fine writer, but just could use some tightening… the author can be trusted to correct [this]…

I have reviewed the entire manuscript with a professional writer and made several minor revisions to tighten the writing. I do not review these in detail here as the reviewer indicated his trust in my ability to handle these minor editorial revisions.
Reviewer 4: Alastair H Leyland
The paper has the potential to provide a worthwhile contribution....

Major compulsory revisions:
1. The weighting systems need to be described more fully; in particular, the methods section should give formulae for the weights. Without these it is not possible to gauge, for example, the appropriateness of leaving level 2 unweighted. (This seems strange, given the different sizes of the states, but the plausibility of this will depend on the level 1 weights.)

Appendix B now gives the equations for the weight scaling method systems. I present the formulas in Appendix B (p. 29, line 7) rather than in the methods section because, as Reviewer 2 notes, I “provide sufficient formulas and explanation of model development but do not go overboard, which otherwise could intimidate the novice reader.” Reviewer 1 also highlighted this paper’s utility in making readers aware of issues related to design weights and MLM. Both highlight and support my intention to approach these concepts from an introductory non-mathematical approach. By including the equations in the Appendix, advanced readers can evaluate them, but the equations will not off-put readers seeking a non-mathematical approach like the one I have intentionally adopted.

Finally, I further clarify and support leaving level-2 unweighted. I expand upon my earlier language and indicate that each state was sampled with certainty. Thus, states were not selected with unequal probability and do not need weighting. I now include that NCHS does not provide level-2 weights and I clarify that the level-1 weights account for unequal probability of selection given different population sizes within states. This section (starting on p. 10, line 15 in the revision) reads,

“The NS-CSHCN sampled each state with certainty. Thus, states were not selected with unequal probability and do not need nor does NCHS include weights. As described, the level-1 weights account for unequal probability of selection given different population sizes within states. Thus, I left level-2 unweighted.”

2. Five models have been fitted using each weighting scheme (including unweighted) in each package... This is done without explanation... one [model] that should be included, is that for random intercepts only....if the author is considering random slopes why not also let the coefficient of the level-2 predictor vary across States ....?

I now present more information regarding how I chose these models. I did this because they represent the basic models presented by major texts on MLM (e.g., Raudenbush and Bryk, 2002), the models form the building blocks for more complicated models, and because the models in each series represent typical types of models analysts would explore in MLM. This addition occurs on p. 9, line 20 in the revision and reads,

“I chose these models because they represent the basic models presented by major texts on MLM (e.g., Raudenbush and Bryk[16]), the models form the building blocks for more
complicated models, and because the models in each series represent typical types of models analysts would explore in MLM.[16, 28]"

Additionally, I clarify that the unconditional model is a random intercepts only model and that all subsequent models allow the intercept to vary across the states. Also, as Raudenbush and Bryk (2002) note, I indicate that the level-2 coefficients are presumed to vary. I added the following sentence, which begins on p. 11 line 20 in the revision,

“For each series, all models allowed the intercept to vary across the states. Additionally, as Raudenbush and Bryk note,[16] for models that include level-2 coefficients, the level-2 coefficients are presumed to vary.”

Finally, at the request of Reviewer 3 (see Reviewer 3, point 5 above), I now include the traditional MLM equations for each model. In this way, advanced readers can further appreciate each model’s details. I now indicate to readers that Appendix C (p. 30, line 1 in the revision) includes the equations for each estimated model. The following sentence on p. 12, line 1 in the revision addresses this,

“Appendix C presents traditional MLM equations for each model I estimate.”

3. The results show several … differences … which should be … presented in more detail … to convince [readers] that they have not arisen due to technical differences (e.g. determination of convergence) or mis-specification of one or more models. The author should consider:

(a) why do MLwiN and GLLAMM give such low estimates of the SE of the residual variance in the unweighted models?

GLLAMM and MLwiN provide a low estimate of the SE of the residual variance because these programs do not provide robust standard errors by default with weighted data. One must request robust standard errors. Mplus provides robust standard errors by default. At the suggestion of this reviewer (see below) and Reviewer 1 (see Reviewer 1, point 11, above), I now explicitly indicate the estimation methods I use in the Methods section. There (p. 12, line 6 in the revision) I indicate that I used each program’s default settings. I use the defaults in the analyses reported in this paper because many beginning MLM analysts (to whom I have directed this paper) will use their chosen program’s default settings. However, to verify that the difference in the standard errors result from using and not using robust standard errors, I ran each set of analyses in MLwiN and GLLAMM requesting robust standard errors. In each case, the residual robust standard error (0.26) matched that of Mplus. I also reran the analyses in Mplus turning off the default robust standard error estimation process. In that case, the residual variance output by Mplus (0.026) matched MLwiN and GLLAMM. I now clearly indicate the reason for the difference in the estimate of the standard error of the residual variance in the unweighted models in the paper. Further, I offer to share the results of these analyses with interested readers. I do not present the results because two other reviewers (Reviewer 2 and Reviewer 3) requested that I condense the paper and amount of data
presented. But, again, I assure the reader that the programs converge when using robust standard errors. This addition (beginning on p. 13, line 20 in the revision) reads,

“This transpires because, by default, Mplus outputs robust standard errors. In separate analyses (not shown here), I reran the analyses requesting robust standard errors in MLwiN and GLLAMM. The robust standard errors converged across programs (SE = 0.26) all other parameters remained consistent across programs. I will gladly share these results with the interested reader.”

(b) Why does MLwiN appear to provide such a poor estimate of the residual variance (and its SE) under weight method B for all models?

It is unclear why MLwiN provides a relatively discrepant estimate of the residual variance and its standard error under method B for all models. To insure that this did not transpire because of convergence criteria, I re-estimated the models using several increasingly stringent convergence criteria. In each case, MLwiN arrived at the same estimate of the residual variance and its standard error. Using RIGLS also led to the same result. Finally, to insure that this did not result from model misspecification or error in creating the dataset. I recreated the dataset entirely using the exact same data used in the GLLAMM runs. I then setup the model “fresh” in MLwiN. Again, MLwiN arrived at the same discrepant estimate of the residual variance and its standard error. This suggests that the difference does not occur because of the convergence criteria, but rather occurs because of estimation differences. I now clearly indicate in the manuscript that I tested several different convergence criteria to insure that the discrepancy does not result from convergence issues, but appears to result from estimation differences. This addition (p. 14, line 4 in the revision) reads,

“MLwiN estimated a smaller residual variance and residual variance standard error using weight method B than either Mplus or GLLAMM. Likewise, MLwiN’s estimate of the slope for state poverty and its standard error diverged slightly (but consistently) from Mplus and GLLAMM under all scaled weighting analyses. To investigate the source of these differences, I reran these analyses with increasingly stringent convergence criteria. In all cases, MLwiN arrived at the same estimate of the residual variance. This suggests that the discrepancy does not result from convergence issues, but results from estimation differences.”

(c) Why does GLLAMM appear to provide such a poor estimate of the variance of the intercepts in unweighted models that include the individual level predictor (but not when the cross-level interaction is included)? And, why does GLLAMM appear to provide such a poor estimate of the covariance between intercept and slope (and its SE) in the model containing a level-1 predictor only in the unweighted analysis?

To examine the source of these discrepancies, I re-estimated the models in GLLAMM and increased the number of quadrature points from 8 (the default) to 16. This brought the estimate of the variance in the intercepts in the unweighted model nearly completely in line with Mplus and MLwiN. It also brought the estimate of the covariance between intercept and slope (and its SE) in the model containing a level-1 predictor more in line with the other
programs, though not completely. As a result, I have added language in the paper that the differences appear to occur because of the estimation process (i.e., the number of quadrature points used in the analysis). I have also included language indicating that the failure to agree perfectly with the other programs may occur because of the larger cluster sizes used in this data. Quadrature works less well with large cluster sizes (Rabe-Hesketh, et al., 2005). I also recommend that, in their own analyses, readers should check the estimation’s adequacy by estimating models with increasing numbers of quadrature points, as noted by Rabe-Hesketh, et al. (2005). As above, I do not present the results because two other reviewers (Reviewer 2 and Reviewer 3) requested that I try and condense the paper and amount of data presented. I do include a note that I will share detailed results with the interested reader. These additions (p. 13, line 4 in the revision) read,

“For example, in unweighted models that included the individual level predictor but not the cross-level interaction, GLLAMM’s estimate of the variance of the intercepts diverged from Mplus and MLwiN. Likewise, GLLAMM provided a discrepant estimate of the covariance between the intercept and slope (and its standard error) in the model containing a level-1 predictor only. To examine the source of this discrepancy, I re-estimated the models in GLLAMM and increased the number of quadrature points from 8 (the default) to 16. This brought GLLAMM’s estimates nearly in line, though not perfectly, with Mplus and MLwiN. Thus, it appears that the differences between GLLAMM and the other programs largely occur because of differences in estimation (I will happily share the details of the analyses with the interested reader). The fact that the estimates did not converge perfectly across GLLAMM and the other programs may have occurred because of the relatively large cluster sizes in these data. Large cluster sizes may limit the performance of quadrature estimation, and MQL methods (like those used by MLwiN) may work better.[23] As Rabe-Hesketh, et al. suggest,[23] analysts should check the adequacy of the quadrature points in any given situation by estimating models with increasing numbers of quadrature points.”

(d) Why does MLwiN appear to provide such a poor estimate of the covariance between intercept and slope in the model containing a level-1 predictor only under weighting scheme A (or is this a typo)?

I made a typographical error. The correct estimate should have been and now is -0.008. I and an independent reader have rechecked every single estimate reported in the tables against the program output. Other than the missing value noted below (point 10) which I fixed, we found no other typos.

(f) Why does MLwiN appear to provide such poor estimates of the slope for State poverty (and its SE) in all models in which this is included and under all weighting schemes?

Similar to the above (point b), it is unclear why MLwiN provides a estimates of the slope for State poverty (and its SE) in all models in which this is included and under all weighting schemes. As above, to insure that this did not transpire because of convergence criteria, I re-estimated the models using increasingly stringent convergence criteria. In each case, MLwiN arrived at the same estimate and standard error. Using RIGLS also led to the same result. I also
used the recreated data (see above) and setup the models “fresh” in MLwiN. Again, MLwiN arrived at the same discrepant estimate and standard error. This suggests that the difference does not occur because of the convergence criteria, but rather occurs because of estimation differences. I now clearly indicate in the manuscript that I tested several different convergence criteria to insure that the discrepancy does not result from convergence issues, but appears to result from estimation differences. My addition addressing points (b) and (f) (p. 14 line 4 in the revision) reads,

“MLwiN estimated a smaller residual variance and residual variance standard error using weight method B than either Mplus or GLLAMM. Likewise, MLwiN’s estimate of the slope for state poverty and its standard error diverged slightly (but consistently) from Mplus and GLLAMM under all scaled weighting analyses. To investigate the source of these differences, I reran these analyses with increasingly stringent convergence criteria. In all cases, MLwiN arrived at the same estimate of the residual variance. This suggests that the discrepancy does not result from convergence issues, but results from estimation differences.”

Minor essential revisions:
4. The first paragraph of the introduction reads as though the strata in a stratified sample should be treated as a level in a multilevel model…This should be clarified.

At the suggestion of Reviewer 1, I have removed a discussion or mention of strata from this section and the paper entirely.

5. P4 para 1: “Unlike traditional programs … MLM directly model complex sampling designs…” It is not clear what is meant by “directly” modelling complex sampling designs – this paragraph needs to be rewritten to clarify what the differences are between the “traditional” and “multilevel” programs.

I have clarified the difference between “traditional” and “multilevel” approaches here. At the suggestion of Reviewer 1, I more clearly state that traditional approaches correct the standard errors for the sampling design, but do not fully allow analysts to investigate variance across clusters. I also more clearly indicate that in MLM, one expresses the sampling design as part of the MLM equations. And, I direct readers to Appendix C (p.30, line 1, added in the revision) that includes examples of MLM equations. This section on p. 4, line 3 in the revision now reads,

“Thus, they allow analysts to describe which variables predict individual differences, they allow analysts to describe which variables predict cluster level differences (e.g., state level differences), and they allow analysts to explore variation across and within clusters. Moreover, because MLM explicitly model the clustered nature of the data, MLM can correctly estimate standard errors and lead to more accurate inferential decisions.[16, 17] Traditional programs for analyzing complex survey data (e.g., SUDAAN)[18] use program commands to correct the standard errors for the sampling design, but treat the sampling design as a nuisance variable. In MLM, one expresses the sampling design as part of the equations in the model (Appendix C presents a series of MLM equations), rather than expressing the design outside the model.[17]
In this way, analysts can examine predictors at both the cluster and individual levels and investigate variance within and across clusters.”

6. P9 para 2: The publication by Blumberg … should be included in the references and cited appropriately.

I have corrected this reference and cite it appropriately.

7. P9 para 3: What is the “200% poverty level”?

200% poverty level refers to a family income that is no greater than twice the federal poverty level. I have clarified this in the revision (beginning on p. 10, line 22).

 “…the proportion of families in the state with an income no greater than twice the US federal poverty level (i.e., 200% poverty level)....”

8. Methods: The analytical methods are not provided in sufficient detail to permit replication. For example, for MLwiN the author should state whether IGLS or RIGLS estimation and, for the dichotomous response, whether a 1st or 2nd order approximation was used and whether MQL or PQL.

I have included this information for each program in section 2.1.1. There, I note that, given that many analysts will use their chosen program’s default settings, I used each program’s default settings in the analyses reported in my manuscript. This section (p. 12 line 6 in the revision) now reads,

“Thus, for Mplus, I used MLR for both the continuous and categorical analyses. MLR delivers maximum likelihood parameter estimates with robust standard errors computed using a sandwich estimator. For categorical outcomes, MLR uses numerical integration and adaptive quadrature using 15 integration points per dimension.[21] For MLwiN, I used the default iterative generalized least squares (IGLS) estimator for both continuous and categorical outcomes. For the categorical outcome, I used default 1st order marginal quasi likelihood (MQL) estimation. By default, MLwiN provides robust standard errors for models incorporating design weights. For unweighted analyses, MLwiN does not provide robust standard errors by default. I used these settings.[22] For GLLAMM, I used adaptive quadrature with 8 quadrature points (the default). By default, GLLAMM provides robust standard errors computed using a sandwich estimator for models incorporating design weights. For unweighted analyses, GLLAMM does not provide the robust standard errors by default.[23, 29] I used these settings.”

9. …. it would be helpful to present the results of unweighted and weighted single level analyses alongside the results of multilevel analyses. (These need only be presented for one software package given that they should provide standard results.)…
I now present the results of unweighted and weighted single level analyses in MLwiN for the continuous and categorical outcome in Table 3. I direct the reader to Table 3 (p. 15, line 11 in the revision) with the following sentence,

“Table 3 presents the results of single level analyses ignoring sampling design for comparison.”

10. P13 Table 1: The SE is missing for the MLwiN estimate of the variance in the slopes between States under weight method B.

I have added the missing estimate.

11. P22 para 2: If MLM is used as an abbreviation for multilevel models you cannot say “creating a three-level MLM...”. Similarly p23 para 1: “…to examine a MLM...”.

I fixed this. The first sentence (p. 19, line 21 in the revision) now reads, “...creating a three-level model....”

As a result of revisions requested by Reviewer 2 (see Reviewer 2, point 11, above), I no longer include the second sentence that included “…to examine a MLM....”

12. P30 References 18-20 & 23 should be provided in full.

I have reviewed and updated the entire reference list. The errors in the initial submission occurred because I used a reference manager for the first time for this article (at BMC’s request). I failed to check the format of the references output by the reference manager because (honestly) it did not occur to me that the output list would differ from y database. My database correctly listed the references, however, the list output by the manager did not reflect what I expected. I apologize. This will not occur. Again, I have reviewed and completely updated the entire reference list.

I believe these changes address the reviewers’ concerns. If you or they have any questions about these revisions, my response, or require any further information, please contact me at adam.carle@unf.edu or 904-620-3573. Thank you for your continued consideration.

Sincerely,

Adam C. Carle, PhD