Evans, C., Leckie, G., & Merlo, J. (2020). Multilevel versus Single-Level Regression for the Analysis of Multilevel Information: The Case of Quantitative Intersectional Analysis. *Social Science and Medicine*, *245*, 112499. [112499]. https://doi.org/10.1016/j.socscimed.2019.112499

Multilevel versus Single-Level Regression for the Analysis of Multilevel Information: The Case of Quantitative Intersectional Analysis

Clare R. Evans (corresponding author)[1]

George Leckie[2]

Juan Merlo[3]

[1] Department of Sociology, 1291 University of Oregon, Eugene, OR 97403, USA

ORCID ID: https://orcid.org/0000-0002-9862-9506

[2] Centre for Multilevel Modelling and School of Education, University of Bristol, UK

ORCID ID: https://orcid.org/0000-0003-1486-745X

[3] Research Unit of Social Epidemiology, Faculty of Medicine, University of Lund, Sweden

ORCID ID: https://orcid.org/0000-0001-8379-9708

Email address for corresponding author: cevans@uoregon.edu

Funding: GL & JM are in part supported by The Swedish Research Council (Vetenskapsrådet) through the project "Multilevel Analyses of Individual Heterogeneity: innovative concepts and methodological approaches in Public Health and Social Epidemiology": (#2017-01321, PI: JM).

Keywords: Intersectionality; Multilevel Models; Health Inequality; Linear Regression; Quantitative Methods; Social Determinants

Multilevel versus Single-Level Regression for the Analysis of Multilevel Information: The Case of Quantitative Intersectional Analysis

Abstract

Intersectional MAIHDA involves applying multilevel models in order to estimate intercategorical inequalities. The approach has been validated thus far using both simulations and empirical applications, and has numerous methodological and theoretical advantages over single-level approaches, including parsimony and reliability for analyzing high-dimensional interactions. In this issue of SSM, Lizotte, Mahendran, Churchill and Bauer (hereafter "LMCB") assert that there has been insufficient clarity on the interpretation of fixed effects regression coefficients in intersectional MAIHDA, and that stratum-level residuals in intersectional MAIHDA are not interpretable as interaction effects. We disagree with their second assertion; however, the authors are right to call for greater clarity. For this purpose, in this response we have three main objectives.

First, in the LMCB commentary the authors incorrectly describe model predictions based on MAIHDA fixed effects as estimates of "grand means" (or the mean of means), when they are actually "precision-weighted grand means." We clarify the differences between average predicted values obtained by different models, and argue that predictions obtained by MAIHDA are more suitable to serve as reference points for residual/interaction effects. This further enables us to clarify the interpretation of residual/interaction effects in MAIHDA and conventional models. Using simple simulations, we demonstrate conditions under which the precision-weighted grand mean resembles a grand mean, and when it resembles a population mean (or the mean of all individual observations) obtained using single-level regression, explaining the results obtained by LMCB and informing future research. Second, we construct a modification to MAIHDA that constrains the fixed effects so that the resulting model predictions provide estimates of population means, which we use to demonstrate the robustness of results

reported by Evans et al. (2018). We find that stratum-specific residuals obtained using the two approaches are highly correlated (Pearson corr=0.98, p<0.0001) and no substantive conclusions would have been affected if the preference had been for estimating population means. However, we advise researchers to use the original, unconstrained MAIHDA. Third, we outline the extent to which single-level and MAIHDA approaches address the fundamental goals of quantitative intersectional analyses and conclude that intersectional MAIHDA remains a promising new approach for the examination of inequalities.

Keywords: Intersectionality; Multilevel Models; Health Inequality; Linear Regression; Quantitative Methods; Social Determinants

1. Introduction

An innovative new approach to quantitative intersectional analysis has recently been proposed: intersectional MAIHDA (multilevel analysis of individual heterogeneity and discriminatory accuracy) (Evans 2015; Evans et al. 2018; Green et al. 2017; Merlo 2018). Intersectional MAIHDA involves applying multilevel models in order to estimate intercategorical inequalities. Individuals are viewed as nested within intersectional strata. This framing treats strata as a type of context, analogous to how neighborhoods or other physical contexts would be modeled in a multilevel framework. Strata are defined using relevant axes of marginalization and inequality such as gender, race/ethnicity, and socioeconomic status. This approach offers many theoretical and methodological advantages over conventional single-level approaches. Among those, MAIHDA provides an improved and parsimonious way of analyzing high-dimensional interactions as compared with single-level regression, which typically entails inclusion of interaction parameters (Evans et al. 2018; Jones et al. 2016; Merlo 2018). Furthermore, intersectional MAIHDA has answered calls from intersectional scholars for innovative new quantitative methods to study intercategorical inequalities (Bauer 2014; Bowleg 2012; McCall 2005; Nash 2008) and calls from social epidemiologists for innovative new eco-epidemiologic approaches (Merlo 2014; Merlo & Wagner 2012). While MAIHDA has been validated thus far using simulations (Bell et al. 2019; Evans 2015; Evans et al. 2018; Jones et al. 2016) and has been applied in a number of empirical scenarios (Axelsson Fisk et al. 2018; Evans 2019; Evans & Erickson 2019; Evans et al. 2018; Hernández-Yumar et al. 2018; Persmark et al. 2019), there remain many subtleties to this new approach that are worthy of further exploration and explication.

In this issue of SSM, Lizotte, Mahendran, Churchill and Bauer (2019) (hereafter “LMCB”) put forward a critique of intersectional MAIHDA as a tool for estimating residual “interaction effects” for intersectional social strata. They agree that MAIHDA provides unbiased estimates of the strata-specific population means, and that the decomposition of these means by Evans et al.

(2018) and others into additive (fixed) effects and interaction (residual) effects is mathematically correct. However, they take issue with the interpretation these authors have assigned to the additive fixed effect regression coefficients in intersectional MAIHDA and the model predictions which result from these. Their argument proceeds as follows. First, they assert that researchers have erroneously interpreted model predictions based on the fixed effect regression coefficients in intersectional MAIHDA as population means. Second, using simulations they demonstrate that "*fixed effects in MAIHDA do not represent population average effects; rather, they reflect effects under an implicit re-weighting of the data given all intersections are of equal size*." Though they do not use the term "grand mean" at any point, based on this description they imply that the fixed effects regression coefficients estimated by MAIHDA always produce predictions of grand means (i.e., a mean of stratum-specific means). They also note that grand means are equal to the population means (i.e., mean of all individual observations) obtained using single-level models when strata are of equal size (as in their simulation Scenario 1). Third, because they assert that the fixed effects in MAIHDA do not result in population-average inferences, they "do not provide a meaningful reference point for intersectional effects" in the form of residual interaction parameters. Ultimately this leads the authors to conclude that MAIHDA is unsuitable for estimation of intersectional interaction effects. Instead they argue for the use of single-level approaches and suggest that the issue of small sample sizes in some strata could be addressed through penalized regression or other techniques.

While the authors are right to call for greater clarity on the interpretation of MAIHDA fixed effects, they are incorrect in their claim that Evans et al. (2018) interpreted these as population means. In supporting their point, they refer only to an (admittedly) ambiguously worded passage by Evans (2019) that, despite their assertion, did not use the term "population average" and was intended to illustrate a different point entirely (namely, the fact that additive fixed effects in models which include interactions have a fundamentally different interpretation than the analogous fixed effects in models which do not include interactions). Evans & Erickson (2019)

instead used the term “global mean” and Hernández-Yumar et al. (2018) referred to population averages where the “population” was of strata, not individuals. The authors’ larger point that the ambiguity of terminology in this literature may have created some confusion is valid, and so in this response we define clearly the differences between population means, grand means, and precision-weighted grand means, and we demonstrate the conditions under which these values are similar or different.

In response to their second point regarding an implicit reweighting of the data, the simulation scenarios they consider illustrate a known and well-understood difference between single-level models and multilevel models: namely, that single-level models estimate population means (or averages across individuals) while multilevel models estimate precision-weighted grand means (or precision-weighted means of cluster-specific means). Despite this, the authors instead describe fixed parameters from intersectional MAIHDA as being grand means and imply that this is generally true. While it is true that in their simulation scenarios the MAIHDA models’ estimates of precision-weighted grand means happen to be equal to grand means, this is a reflection of the parameter space these simulations are situated within and especially the large number of individuals per stratum in the simulated samples. The difference between a grand mean and a precision-weighted grand mean is meaningful, because in parameter spaces more often encountered in intersectional research the precision-weighted grand mean will tend to fall somewhere between the population mean and the grand mean. In other words, the simulation scenarios outlined in the LMCB commentary are unrealistic and this exaggerates the differences seen between the two approaches more than will typically be encountered. Furthermore, we stress that there is nothing inherently wrong with choosing to estimate either population means or precision-weighted grand means. Rather, these choices are rooted in the specific research question at hand.

Finally, we disagree with LMCB’s assertion that because intersectional MAIHDA fixed effects do not result in estimates of population means that this makes it inadvisable to interpret

the stratum residuals as interaction effects. In fact, we contend, the estimation of precision-weighted grand means in MAIHDA is actually an advantage of the approach, and therefore the residual "interaction effect" estimates are both interpretable and meaningful.

To address these points, our response proceeds as follows. First, we define clearly the differences between the various approaches to making population-average inferences and discuss their relative advantages and disadvantages in the context of intersectional analysis. Using simulations, we outline the conditions under which precision-weighted grand means more closely resemble grand means, and when they more closely resemble population means. Based on this exercise we explain why the simulation scenarios outlined in the LMCB commentary make it appear that intersectional MAIHDA provides estimates of grand means rather than precision-weighted grand means. This further enables us to clarify the interpretation of residual/interaction effects in MAIHDA and conventional models, and the differences between estimates of "interaction effects" between the two approaches. Second, we address the concerns of the authors that the results of prior publications might appear different if the fixed effects in intersectional models had estimated population means rather than precision-weighted grand means. Though we ultimately recommend the continued use of intersectional MAIHDA as it was originally proposed, we provide an alternative version of MAIHDA (a "constrained MAIHDA") which constrains the fixed effects such that they produce population means. We apply this constrained approach to both the simulation scenarios outlined by LMCB and to the empirical data used by Evans et al. (2018) to demonstrate the robustness of these results. Finally, we outline the extent to which single-level and MAIHDA approaches satisfy the fundamental goals of quantitative intersectional analyses and conclude that intersectional MAIHDA remains a more suitable new approach for the examination of inequalities.

2. Population Means, Grand Means, and Precision-Weighted Grand Means

Terms such as “population average” and “grand mean” have been used very differently across the literature, and at times this has resulted in confusion about their meaning. Acknowledging differences in terminology, we now define what is meant by them in the *present* study so as to provide clarity. As a simple example, suppose survey respondents are randomly sampled from 100 neighborhoods. In a null single-level model (a linear regression with no covariates), the intercept represents the *population mean*, or the mean value of outcome y_i across the sample regardless of neighborhood. A *grand mean* would be obtained by calculating the mean value of outcome y_i in each neighborhood, then taking the mean of the neighborhood-specific means. While this statistic will tell us something about the average neighborhood value, it does not take into account the differences across neighborhoods in the reliability of neighborhood-specific estimates. Some neighborhoods might have many observations, and as such their sample averages will provide more reliable estimates of their neighborhood population means than would be the case for for neighborhoods with fewer observations. The grand mean implicitly treats all neighborhoods as being of equal sample size and reliability. To clarify further, the population mean can be understood as a frequency-weighted mean of the neighborhood means, while the grand mean is an unweighted mean of the neighborhood means. In a null multilevel model, the intercept estimates a *precision-weighted grand mean*, or a weighted average of the neighborhood-specific means that takes into account the reliability of the neighborhood values. For simplicity this value has sometimes been referred to as a “global mean” (e.g., Evans & Erickson 2019), but this may have contributed to the confusion surrounding interpretation identified in the LMCB commentary.

It is important to note that the descriptions of the three types of mean values above are made in the context of null models. In a null single-level model the intercept is an estimate of the population mean, while in a null multilevel model the intercept is an estimate of a precision-weighted grand mean. In models that include fixed effects these interpretations become slightly more complex. For instance, in a single-level model which includes the gender variable

"woman" (a dummy variable where 1=woman, 0=man) the intercept estimates the population mean among men, while the regression coefficient on the woman dummy variable estimates the population mean difference between women and men. So this model is still estimating "population means." We therefore refer more generally to single-level models as estimating population means and MAIHDA as estimating precision-weighted grand means, without restricting this interpretation to the null case alone.

In their commentary, LMCB use simulations to demonstrate that when interactions exist between two axes of marginalization (such as gender and immigration status), then in an additive-only model the population mean effect for "gender" will reflect the frequency of immigrants in the sample, whereas a grand mean will be based on the assumption that all strata are of equal size and therefore that strata of immigrants and non-immigrants should be weighted equally in the sample. The result is a difference between population means and grand means. They assert that the population mean has a more intuitive interpretation, however they provide no deeper justification to support this assertion beyond the claim that grand means are reflecting a "artificial population where all intersections are of equal size." However, as noted previously, this assertion is based on both a misstatement of what average values are obtained using MAIHDA and the arbitrary assumption that residuals in population mean models are inherently superior for intersectional analysis.

Because they do not discuss precision-weighted grand means, LMCB's commentary frames the question of which value has a more intuitive interpretation for intersectional analysis as being a contest between population means and grand means. The core question when considering a baseline additive effect for intersectional strata is: *what is the additive effect in the average stratum?* (Or in the case of non-null models, what is the additive effect for the average strata of women, or the average strata of immigrants?) But what counts as an "average stratum"? A population mean's estimate assigns more weight to larger strata and less weight to smaller strata when estimating outcomes in a hypothetical average stratum. A grand mean, on

the other hand, weights all strata equally. Neither value is inherently more intuitive, and they represent subtly different aspects of reality. Population means provide a window on the average experiences in a population where all individuals are weighted equally (and therefore strata are weighted according to size). The downside of this is that a form of erasure happens, where what is considered “average” is more a reflection of outcomes for the majority than the minority. In intersectional analysis the purpose is often to reveal points of erasure, and to acknowledge that an average based on everyone will inherently be less likely to represent reality for minorities. On the other hand, grand means provide estimates where all strata are weighted equally, which deals with the issue of erasure, but may also fail to address the issue of reliability. Some strata have estimates that are more reliable than others, so why wouldn’t we want to factor this into our assessment of what an “average stratum” is experiencing? Critically, the intersectional MAIHDA approach is generally a compromise between these two views of “average stratum” because it estimates precision-weighted grand means. We therefore argue that this is ultimately a more satisfying way to describe predicted outcomes for an “average stratum.” What is less clear is under what conditions the precision-weighted grand mean will resemble the population mean, and when it will resemble the grand mean. In this study we explore this issue more fully.

While there is nothing inherently wrong with estimating population means, grand means, or precision-weighted grand means, LMCB are right to call for greater clarity on this issue and to challenge us to explore the issue more fully.

3. Simulations

At the heart of this first simulation exercise is a question about when the precision-weighted grand mean (obtained using intersectional MAIHDA) will resemble the population mean and when it will resemble the grand mean. We base this simulation on the Scenario 2 outlined in the LMCB commentary, and therefore briefly repeat the details of this simulation below.

3.1 Simulated Data

The data consist of $N = 100000$ simulated individuals nested within 32 strata formed by the unique combinations of five independent binary variables. Let y_{ij} denote the outcome for individual i ($i = 1, \dots, n_j$) in stratum j ($j = 1, \dots, 32$) and $x_{1j}, x_{2j}, x_{3j}, x_{4j}, x_{5j}$ the five independent binary variables from which the strata are defined. Further, let $p_1 = 0.7, p_2 = 0.7, p_3 = 0.5, p_4 = 0.5, p_5 = 0.5$ denote the marginal probabilities of positive values for each variable. The data generating process for y_{ij} can then be written as the following linear regression model:

$$y_{ij} = 0 + 1x_{1j} + 1x_{2j} + 1x_{3j} + 1x_{4j} + 1x_{5j} + 1x_{1j}x_{2j} + v_{ij} \quad \text{(Eq.1)}$$

$$v_{ij} \sim N(0,1)$$

where we have included the interaction between x_{1j} and x_{2j} as well as each covariate as a main effect. The resulting population mean and variance in the simulated data is 3.39 and 2.99. For simplicity and clarity, we fit all models using restricted maximum likelihood estimation (REML).

3.2 Model Specifications

LMCB first fit a conventional single-level regression model that includes all five additive main effects but excludes the interaction parameter:

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4j} + \beta_5 x_{5j} + r_{ij} \quad \text{(Eq.2)}$$

$$r_{ij} \sim N(0, \sigma_r^2)$$

We refer to this single-level model as Model 1. The intersectional MAIHDA model (Model 2), which has unconstrained fixed parameters (in order to contrast later with constrained MAIHDA), can be written as:

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4j} + \beta_5 x_{5j} + u_j + e_{ij} \quad \text{(Eq.3)}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

where u_j represents the residual for stratum j and e_{ij} represents the residual for individual i in stratum j. The variance partition coefficient (VPC) can be calculated as:

$$\text{VPC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \quad \text{(Eq.4)}$$

Table 1 provides the estimates for Model 1 (single-level model without interaction) and Model 2 (unconstrained intersectional MAIHDA). Interaction parameters are not included in Model 1 in order to explore the differences in fixed parameter estimates between Models 1 and 2 due to frequency vs. precision-weighting of the data provided on each stratum, rather than to fixed effect specification differences. In Model 1 we expect that the regression coefficient for x_{1j}, for example, will reflect the frequency in the sample of x_{2j} because of the omitted interaction between the two variables and the subsequent (expected) omitted variable bias. Because $x_{2j} = 1$ is present for 70% of the sample, the estimated regression coefficient for x_{1j} reflects this: $\beta_1 = 1.7$. As we will show, in this parameter space the precision-weighted grand mean is equal to the grand mean, and a grand mean would be estimated under the assumption that $x_{2j} = 1$ is present in 50% of the population (i.e., all intersections are assumed to be of equal size), and therefore $\beta_1 = 1.5$.

For each model, we calculated the predicted mean outcome:

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + \hat{\beta}_3 x_{3j} + \hat{\beta}_4 x_{4j} + \hat{\beta}_5 x_{5j} \quad \text{(Eq.5)}$$

And the predicted total interaction (or residual) effect:

$$\hat{r}_j = \bar{y}_j - \hat{y}_j \quad \text{(Eq.6)}$$

where $\bar{y}_j$ denote the sample average outcome in stratum j.

Thus, the predicted total interaction effect for each stratum is simply the difference between the mean observed and predicted outcome for that stratum. These predictions are provided in Table 2. In the multilevel model, it is then standard to additionally apply a shrinkage factor to these predictions as:

$$\hat{u}_j = \hat{s}_j \hat{r}_j, \qquad \hat{s}_j = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_j}} \tag{Eq.7}$$

where $\hat{s}_j$ denotes the shrinkage factor. The resulting shrunken effects $\hat{u}_j$ are often referred to as the empirical Bayes predictions of the random effects. However, in the current simulated data n_j is large for all strata and so $\hat{s}_j$ is effectively 1 in all strata (i.e., no shrinkage). The shrunken effects $\hat{u}_j$ are therefore effectively the same as the unshrunken effects $\hat{r}_j$. It is also for this reason that the precision-weighted grand mean estimates equal the grand mean. It is very relevant to note that in empirical data applications where some or many strata consist of far fewer individuals, the effect of shrinkage will be notable, and the precision-weighted grand mean will diverge from the grand mean.

3.3 Simulation: When does the precision-weighted grand mean resemble the grand mean, and when does it resemble the population mean?

Because the main difference between a precision-weighted grand mean and a grand mean is in the shrinkage factor, which depends on both the sample size per stratum (and by extension the total sample size, N) and the variance within- and between-strata (and therefore the VPC) it is the parameter space at varying levels of N and VPC that we explore.

All simulations were based on Scenario 2 parameters, except for the total sample size and VPC. VPC in the original Scenario 2 was an extremely high 66.7%, far higher than that estimated in any application of MAIHDA to real data. We therefore modified the VPC to more

realistic values of 5% and 1% by increasing the relative size of the individual-level variance. This enabled us to leave the stratum-level variance alone and to match the fixed parameter values for each variable in the data generating process. For each VPC value a simulated data set containing 100000 observations was constructed. Then a random sample of the appropriate size (N) was taken from this data set. Both null single-level and null unconstrained MAIHDA models were fit to this subsample, and the intercept estimate for each model was estimated. A total of 100 iterations of the sampling procedure were conducted for each combination of sample size and VPC, enabling the estimation of average intercept values for each model. Figure 1 presents the results of this simulation.

Across all parameter conditions, the null single-level model provided estimates of the population mean. When the total sample size was large (and therefore the average stratum size was large) and/or the VPC was large, the unconstrained MAIHDA estimate resembled the grand mean. As the N decreased and the VPC decreased the precision-weighting feature of unconstrained MAIHDA meant that the estimates obtained moved away from the grand mean and converged on the population mean.

While the values selected by LMCB were reasonable choices for the purposes of clarity in demonstrating the underlying difference between population and (precision-weighted) grand means, the sample size and VPC values were considerably higher than those reported in all applications of MAIHDA to date. In the null models, the VPC was 66.77% in Scenario 1, 66.69% in Scenario 2, and 50.57% in Scenario 3. In most applications of MAIHDA, the VPC is <5%, and rarely exceeds 10%. Furthermore, the total sample size of 100000, when distributed across only 32 strata, resulted in large sample sizes in even the very smallest strata (>1,000 in Scenario 2). This resulted in shrinkage factors for all strata of effectively 1 (no shrinkage), and therefore the precision-weighted grand mean was effectively equal to the grand mean in all simulated conditions. This explains why the authors erroneously concluded that MAIHDA in general provides estimates of grand means.

The results of LMCB's simulation Scenario 1 are also mathematically intuitive: when the sample size in all strata is equal, then the population mean will be equal to the grand mean. But what might cause the population mean and the grand mean to diverge further (other than unequal strata sizes)? One situation where the two values would be very different is when there is a systematic correlation between the size of strata and stratum-specific means. In Scenario 2, for instance, there is a systematic positive correlation (corr=0.778) between stratum size and stratum means (see Table 2). Population means place more weight on strata with more respondents than grand means do (because grand means weight all strata equally regardless of size), and if these strata also happen to have higher means, then this will result in the population mean being significantly higher than the grand mean. As expected, this was the case in Scenario 2, where the population mean=3.39 and the grand mean=2.75. In simulation Scenario 3 outlined by LMCB the correlation between stratum size and stratum mean values was negative (corr=−0.202) and so the population mean was lower than the grand mean (population mean=1.82; grand mean=2.00). If, on the other hand, there were no correlation between size of strata and stratum-specific means then the population and grand means would coincide. LMCB imply that grand means and population means are equal only when strata are of equal size; This discussion reveals additional conditions when this may be true.

This simulation exercise has provided some general insights into the conditions under which the population mean, grand mean, and precision-weighted grand mean will be similar or different. When strata are of equal size and/or there is no correlation between stratum means and stratum size, all three values will be similar. When strata are of unequal size, if the strata are large and/or the VPC is large, shrinkage will be minimal and the precision-weighted grand mean will resemble the grand mean. However, as sample size (in total and per stratum) decreases and the VPC decreases, the precision-weighted grand mean will converge to the population mean. While this simulation exercise explored the difference between these three

approaches in the context of null models, as we demonstrate empirically below these insights also extend to models that include covariates.

3.4 Clarifying the Interpretation of Interaction Effect Estimates

Residuals in intersectional MAIHDA models that include additive fixed effects and fixed effect interaction parameters in conventional single-level models can only be interpreted as "interaction effects" under the assumption of no omitted variable bias. This has been clearly stated with respect to MAIHDA (Evans et al. 2018), however the same is also true for the conventional approach and this subtlety is not always acknowledged.

The clarifications we have made about the average values obtained by the two approaches enable us to now clarify the interpretation of the interaction effects in each approach. In both models the additive component represents what would have been predicted for a stratum (or, consequently, an individual belonging to that stratum) based on additive effects alone, whereas the residual/interaction parameter represents how much that stratum's total predicted value differs from this additive approximation. There are two important differences between the models and interpretations, however. First, as we have noted, the additive baseline in the conventional model is an estimate of a population mean, and therefore the interaction effect modifies this estimate of the population mean. In intersectional MAIHDA the additive baseline is an estimate of a precision-weighted grand mean, and therefore the residual interaction effect modifies this value. Second, as discussed more extensively by Evans (2019), the additive parameters in the two models are making fundamentally different comparisons. As a simple example, in a conventional single-level model containing fixed dummy parameters "woman" (1=woman; 0=man) and "Black" (1=Black; 0=White) and the dummy interaction parameter "woman *and* Black" the parameter for "Black" would have the interpretation of the difference in the population means between Black men and White men. In a similar MAIHDA model, the fixed interaction parameter would not be included and therefore the

parameter for “Black” would represent the difference in the precision-weighted grand means between all Black strata and all White strata, with no specificity with respect to gender.

While this point may perhaps be obvious, we nevertheless stress that determining whether a particular stratum is relatively advantaged or disadvantaged based on residual (or interaction parameters) alone is not possible. Interaction effects tell us only the direction and magnitude of the difference between the total predicted value for a stratum and what might have been predicted for it based on additive effects alone. In other words, the interaction effect tells us whether the stratum is advantaged or disadvantaged *relative to what might have been predicted for it based on additive effects alone*. It is therefore necessary to consider the magnitude of the additive effects as well as the direction and magnitude of the interaction effects when determining relative (dis)advantage between strata, or more simply to compare the total predicted values (which sum across the additive and interaction effects).

4. Examining the robustness of results: A demonstration using constrained intersectional MAIHDA

For the purposes of demonstration only, we outline an adaptation of linear intersectional MAIHDA as it has been applied to date. This involves continuing to conduct a multilevel analysis of individuals nested in intersectional strata, but one where we effectively constrain the fixed effects regression coefficients to equal those of a single-level linear regression. This leads to estimation of population means while still arguably retaining the advantages of multilevel modeling, such as the calculation of variance partition coefficients (VPCs) and the shrinkage adjustment being applied to predicted interaction effects. Though we ultimately recommend the continued use of intersectional MAIHDA as it was originally proposed (i.e., unconstrained MAIHDA) for the reasons provided above, we provide this alternative in order to explore a concern expressed by LMCB: namely, that results and conclusions of prior studies might have been different had intersectional MAIHDA estimated population means.

We illustrate this proposed alternative first in the context of simulation Scenario 2 (similar results were obtained for simulated Scenarios 1 and 3, available upon request) and then apply it to the empirical case used by Evans et al. (2019).

4.1 Constrained intersectional MAIHDA

In practice, the easiest way to conduct a constrained intersectional MAIHDA is through a two-step process: (1) fit a single-level model to the data that includes all additive main effects, but excludes interaction parameters (as in Eq.2), then (2) fit a two-level variance components model to the residuals from the step 1 model. The intercept should be omitted from this second model. The second model decomposes the total residual into stratum- and individual-level residual components. Similarly, the total residual variance is decomposed into separate stratum- and individual-level variance components. The step 2 model can be written as:

$$\hat{r}_{ij} = u_j + e_{ij} \qquad \text{(Eq.8)}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

The estimates for this constrained multilevel model (Model 3) are presented in the third column of Table 1. The predicted mean and interaction effects are presented in Table 2. As in the case of unconstrained MAIHDA in this simulation, the constrained model has no relevant shrinkage acting on the stratum-level residuals. Therefore, while constrained MAIHDA provides the (expected) identical predicted mean outcomes as the single-level approach, it also provides identical residuals in this case. In more realistic data sets, residuals would be expected to differ between Models 1 and 3 due to shrinkage.

Results for null versions of all three models are available in Supplemental Tables 1 and 2. As can clearly be seen from these comparisons across models, Model 2 (unconstrained MAIHDA) produces subtly different estimates for fixed effect parameters than Model 1

(conventional single-level regression) due to the reweighting of estimates. Model 3 (constrained MAIHDA), however, is a hybrid approach that estimates population mean values while maintaining the advantages of MAIHDA. The Pearson correlation between stratum residuals from Models 2 and 3 is strong (corr=0.872, p<0.0001; see Table 3), but low enough that, at least in these simulated data, the two approaches would lead to somewhat differing predictions for the stratum specific means.

4.2 Empirical Application

In their commentary, LMCB particularly highlight the original contribution of Evans et al. (2018), and therefore we return to this empirical example in order to determine the extent to which the results originally published might have been different using the constrained approach.

4.2.1 Empirical Data

Data comes from Wave 2 of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) (collected in 2004-2005) (Grant and Kaplan, 2005). A final sample of *N*=32,788 respondents was analyzed after deletions for missingness on necessary variables. The dependent variable was a continuous measure of body mass index (BMI). Social strata were defined by all combinations of: gender (2 categories), race/ethnicity (three categories), education (4 categories), income (4 categories), and age (4 categories) for a total of 384 strata (382 of which contained respondents in the sample). For further details, see Evans et al. (2018).

4.2.2 Models and Results

For comparison purposes, four models were fit: a conventional single-level model (Model 1), unconstrained intersectional MAIHDA to mirror analyses performed by Evans et al. (Model 2), the constrained intersectional MAIHDA as proposed in this article (Model 3), and a “grand means” model (Model 4). The grand means model (also sometimes called a “between-model”) was a single-level regression performed on a data set where each value was one stratum and the dependent variable was the mean value for all observations in that stratum. This forces the

parameter estimates to be equal to grand means, rather than precision-weighted grand means. As in the simulation scenarios, Model 1 does not include interaction parameters in order to show differences between Models 1 and 2 that are due to the issue of frequency vs. precision-weighting of the stratum-specific samples, rather than to differences in fixed effect specification.

Fixed and random effect parameters are compared across models in Table 4, and stratum-specific estimates for Models 1-3 are available in Supplemental Table 3. As expected, differences between models in fixed parameter interpretations as population means (Models 1 and 3) and precision-weighted grand means (Model 2) results in slight differences in parameter estimates. Of interest to the present discussion is the extent to which stratum residuals obtained in Models 2 and 3 resemble each other after adjustment for additive main effects (Supplemental Table 3). As reflected in Figure 2, this correlation is very high, with a Pearson correlation=0.98 ($p<0.0001$) (Table 3). Figure 3 presents substantive results for Model 2 (unconstrained MAIHDA) that mirrors a similar figure in Evans et al. (2018) showing patterns in higher/lower than expected residuals for strata at varying levels of education and age, grouped by gender, race/ethnicity and high/low income. For comparison, the same figure setup is used for results from Model 3 (constrained MAIHDA).

From these comparisons and correlation values we conclude that the substantive conclusions previously published by Evans et al. (2018) would not have been meaningfully different if constrained MAIHDA had been used. Indeed, the two sets of results are very similar, suggesting that the potential concerns raised by the LMCB matter little in practice when the different approaches are applied in real intersectional analyses.

5. Simulation Parameter Spaces: Clarity versus Realistic Values

While the correlation between residuals obtained using unconstrained (Model 2) and constrained (Model 3) MAIHDA was still strong in the simulated scenarios (e.g., Scenario 2 corr=0.872 ($p<0.0001$)), it was substantially stronger in the empirical case (corr=0.977

(p<0.0001)). On first glance this would suggest that we should sometimes see meaningful differences between the two approaches. However, based on results from the simulations conducted to determine when the precision-weighted grand mean resembles the population mean, and when it resembles the grand mean, the reason for the lack of any meaningful difference is clear. In the empirical example from NESARC, the VPC was <2% and though the total sample size was large it was averaged over a larger number of strata, thus reducing the average stratum sample to 85.8 respondents. This places the null unconstrained MAIHDA in the parameter space where the precision-weighted grand mean from a null model will be converging on the population mean.

The extension of the logic that precision-weighted grand means fall between population means and grand means is more complex in the case of additive models, however it can be seen to mostly hold in the empirical example outlined in Table 4. With the exception of two fixed parameters for levels of age, all fixed parameters obtained using unconstrained MAIHDA (precision-weighted grand means) fell between values for the population mean (Model 1) and the grand mean (Model 4).

In summary, though there is nothing inherently problematic with choosing to estimate grand means, in most empirical cases the estimates produced by an unconstrained MAIHDA will more closely resemble those from a single-level model than the simulations outlined in the LMCB commentary would seem to imply. These results inform our understanding of other studies which have used unconstrained MAIHDA to date.

6. Fundamental Analytic Goals of Quantitative Intersectional Approaches

The MAIHDA framework fits well with the concept of proportionate universalism discussed by Marmot within the literature on resource allocation in public health (Carey et al. 2015; Marmot & Bell 2012). That is, health actions must be universal, not targeted, but with a scale and intensity that is proportionate to the level of disadvantage. The MAIHDA approach

can be used to inform decisions regarding the degree to which public health interventions need to be universally targeted (Merlo et al. 2019). The use of MAIHDA may also enable a better understanding of the heterogeneity underlying traditional uni-dimensional studies of health inequalities (Merlo 2018). When applying intersectionality to issues in public health, however, it is also essential to maintain the critical theory orientation inherent to the approach in order to avoid blunting intersectionality's "critical edge and transformative aims" (May 2015, p.141).

From a methods perspective, a basic analytic framework for performing quantitative analyses of intersectional inequalities in public health should embrace at least three analytical goals: (1) mapping average health outcomes across intersectional strata; (2) quantifying the share of the total variation in the outcome that is at the between-intersectional strata level as opposed to the within-stratum level; and (3) quantifying interaction of effects between the dimensions that define the intersectional strata.

6.1 Mapping Averages Across Intersectional Strata

The first analytic goal represents traditional epidemiological descriptive analyses of the distribution of the average health outcome values across strata. This procedure helps to identify strata of individuals that – on average – are over or under a pre-specified reference (e.g., the whole population, or another intersectional stratum). Mappings of risk support informed public health decision-making, however an over-reliance on comparisons of averages ("the tyranny of the averages" (Merlo 2003; Merlo 2014; Merlo et al. 2017; Merlo & Wagner 2012)) results in a failure to consider the heterogeneity that exists within strata. For instance, mapping averages is not sufficient to assess whether it would be preferable to direct resources only toward certain intersectional strata or to adopt an approach that will shift population distributions (Rose 1992). For this purpose, we need measures of discriminatory accuracy.

6.2 Quantifying Heterogeneity Within and Between Strata

The second goal is focused on the measurement of components of variance and discriminatory accuracy (e.g., VPC, ICC). This focus is still unusual in traditional epidemiology, which is mostly dedicated to the analysis of measures of association (Merlo 2014), yet this information is relevant for many reasons. Discriminatory accuracy refers to the ability of a measure (in this case, intersectional strata classifications) to distinguish between individuals who have a health outcome and those who do not. Even in the case of larger average differences between strata, when discriminatory accuracy is low focusing on specific strata (i.e., a “high risk” strategy) may lead to inefficient interventions and raise ethical issues related to risk communication and the perils of stigmatization of individuals from specific strata. Despite strong condemnation, many continue to use markers of social position and identity such as categories of gender, race/ethnicity and socioeconomic status as reified, essentialized labels, and sometimes attribute sources of between-stratum differences along these axes to biology rather than to social determinants. Consideration of the (often low) discriminatory accuracy of stratum labels and the (frequently high) within-stratum variability provides a valuable check on this unfortunate tendency. Some scholars have suggested referring to cases of relatively high discriminatory accuracy as being supportive of a “categorical” perspective (Merlo 2018; Wemrell et al. 2017a), by which they mean a situation in which category labels have been particularly well-suited to sorting those who are affected by adverse outcomes from those who are not. Symmetrically, cases of low discriminatory accuracy encourage us to remain skeptical of the use of such labels for determining individual risk, particularly in clinical settings. These cases encourage a more “anticategorical” perspective (Hernández-Yumar et al. 2018; Merlo 2018; Wemrell et al. 2017a). It is critical to note that the terms (inter)categorical and anticategorical are meant here very differently than how they are frequently used to define distinct approaches within intersectional scholarship. McCall (2005), for instance, defines intercategorical approaches as those that adopt categorical labels (while remaining aware of their limitations) in

order to document inequality, while anticategorical approaches are those that seek to illuminate the extent to which these labels are “simplifying social fictions” (p.1773). While anticategorical intersectional scholarship approaches the issue from a very different perspective and uses a very different set of methods, the “anticategorical” conclusion yielded by cases of low discriminatory accuracy ultimately arrives at a similar argument: namely, that labels are insufficient to understand the complex realities of the situation and we should remain cognizant of this fact.

6.3 Estimating Interaction Effects

The third goal is well-known in social epidemiology, as it focuses on the identification of interaction effects. While in qualitative intersectionality the term “interaction” is metaphoric, here it is operationalized as an estimated quantity (either a fixed effect interaction term or a random effect residual) that enables final “total” estimates for particular strata to differ from what might have been expected based on some combination of additive effects. In the conventional single-level approach this is accomplished by selecting *a priori* a single reference level (often high SES White men) and building from that a set of between-stratum comparisons that modify the expected value. This means that interaction terms are only estimable for certain strata, and these values represent comparisons with other strata expected values. In intersectional MAIHDA this is accomplished by controlling for all additive effects in the fixed part of the model and leaving all interactions to be encompassed by the stratum-level residual estimates. While the unconstrained and constrained approaches to MAIHDA now introduce some differences in interpretation of these fixed parameters, the ultimate goal of identifying interaction terms for all social strata simultaneously remains.

6.4 To what extend do the different approaches satisfy the fundamental goals of quantitative intersectional analysis?

A side-by-side comparison of the modeling approaches is presented in Table 5. As shown, both approaches—conventional single-level and unconstrained intersectional MAIHDA—are capable of providing estimates of mean or frequency values in all strata combinations (Goal #1). However, the conventional single-level approach is far less parsimonious than the MAIHDA approach under high-dimensionality, and unless additional analytic steps are taken to adjust standard errors, strata with smaller sample sizes may have unreliable estimates. On the other hand, shrinkage in multilevel models enables us to obtain reasonable estimates for strata with fewer respondents (Evans et al. 2018), and while this shrinkage may not always be complete it is substantially less likely to yield erroneously significant effects than the single-level approach (Bell et al. 2019).

For Goal #2, MAIHDA automatically provides estimates of variability between- and within-strata, enabling consideration of heterogeneities at all levels and assessments of discriminatory accuracy. Single-level approaches, on the other hand, do not directly estimate these values. A recent proposal for how to adapt single-level approaches to evaluate heterogeneity and discriminatory accuracy using AUC analysis (which in this case approximates the VPC) holds considerable promise (Wemrell et al. 2017b), but this is still far from customary.

For Goal #3, both approaches are capable of estimating interaction effects, though the interpretation of these effects differs between the approaches (Evans 2019). Furthermore, as mentioned previously, interaction effects are obtained for only some strata in the single-level approach, whereas they are obtained for all strata in MAIHDA. Which approach is most suitable will depend on a number of factors, including what the research question is and whether sufficient data is available to make some of the practical advantages of MAIHDA irrelevant.

7. Conclusion

Intersectional MAIHDA answers calls from intersectional scholars for innovative new quantitative methods to study intercategorical inequalities (Bauer 2014; Bowleg 2012; McCall 2005; Nash 2008) and calls from social epidemiologists (Merlo 2014; Merlo & Wagner 2012) for innovative eco-epidemiologic approaches. Methodological development requires engagement from many scholars to test the claims made about any new approach, and to examine the conditions under which different approaches either have advantages or disadvantages relative to better-known approaches. This exchange has helped to clarify one of the estimation subtleties of intersectional MAIHDA which affects the interpretation of stratum-level residual estimates. In this response we clarify the interpretation of fixed and residual parameters in MAIHDA, demonstrate the conditions under which precision-weighted grand means will more closely resemble population means or grand means, and discuss factors that could cause population and grand means to diverge. Under conditions most frequently encountered by researchers (i.e., lower VPC and smaller sample size per stratum) intersectional MAIHDA estimates of precision-weighted grand means will result in fixed effect estimates that tend to fall between population means and grand means. We argue that this is another advantage of the intersectional MAIHDA approach. While the single-level approach defines an "average stratum" as being whatever is experienced by the majority (thus treating majorities as the "average" or "default" against which all other strata are assessed), and a between-stratum (grand means) approach would fail to consider differences in reliability of stratum-specific estimates due to differences in sample size, the intersectional MAIHDA approach is a compromise that allows the experiences of minority strata to contribute more to a vision of what an "average stratum" experiences while simultaneously adjusting for reliability of estimates.

We argue that fundamentally the multilevel approach is more suitable for modeling intersectional inequalities, both for theoretical and methodological reasons. Intersectionality is a theoretical framework that draws attention to the social processes constructing interlocking systems of oppression and inequality; Intersectionality does not locate the causes of concern

with individuals. A multilevel modeling framework that situates individuals within intersectional social strata, and estimates parameters associated with strata rather than individual-level variables, is strongly aligned with this theoretical perspective. Methodologically, the multilevel approach also acknowledges a basic truth often apparent empirically—namely, that statistical clustering by social strata occurs and the data is therefore multilevel in structure. While single-level approaches can still be used successfully to evaluate inequalities, and we do encourage their future use, intersectional MAIHDA is the more natural framework for modeling and interpreting intersectional effects.

In the future, researchers have a variety of analytic tools at their disposal for intersectional analysis, and we advise them to select the most appropriate approach for their particular research question(s) while taking steps to address all of the fundamental analytic goals of quantitative intersectional analysis.

References

Axelsson Fisk, S, S Mulinari, M Wemrell, G Leckie, RP Vincente and J Merlo. 2018. "Chronic Obstructive Pulmonary Disease in Sweden: An Intersectional Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy." *Social Science & Medicine (Population Health)* 4:334-46.

Bauer, Greta R. 2014. "Incorporating Intersectionality Theory into Population Health Research Methodology: Challenges and the Potential to Advance Health Equity." *Social Science & Medicine* 110:10-17.

Bell, AJ, D Holman and K Jones. 2019. "Using Shrinkage in Multilevel Models to Understand Intersectionality: A Simulation Study and a Guide for Best Practice." *Methodology* 15(2):88-96.

Bowleg, L. 2012. "The Problem with the Phrase Women and Minorities: Intersectionality— an Important Theoretical Framework for Public Health." *American Journal of Public Health* 102(7):1267-73.

Carey, G, B Crammond and E De Leeuw. 2015. "Towards Health Equity: A Framework for the Application of Proportionate Universalism.". *Int J Equity Health* 14:81.

Evans, C.R. 2015. "Innovative Approaches to Investigating Social Determinants of Health - Social Networks, Environmental Effects and Intersectionality." Doctoral dissertation, Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health. Accessed at: http://nrs.harvard.edu/urn-3:HUL.InstRepos:23205168.

Evans, CR. 2019. "Adding Interactions to Models of Intersectional Health Inequalities: Comparing Multilevel and Conventional Methods." *Social Science & Medicine* 221:95-105. doi: https://doi.org/10.1016/j.socscimed.2018.11.036.

Evans, CR and N Erickson. 2019. "Intersectionality and Depression in Adolescence and Early Adulthood: A MAIHDA Analysis of the National Longitudinal Study of Adolescent to Adult Health, 1995–2008." *Social Science & Medicine* 220:1-11.

Evans, CR, DR Williams, JP Onnela and SV Subramanian. 2018. "A Multilevel Approach to Modeling Health Inequalities at the Intersection of Multiple Social Identities." *Social Science & Medicine* 203:64-73. doi: 10.1016/j.socscimed.2017.11.011

Grant, BF and KD Kaplan. 2005. *Source and Accuracy Statement for the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC)*. Rockville, Maryland: National Institute on Alcohol Abuse and Alcoholism.

Green, MA, CR Evans and SV Subramanian. 2017. "Can Intersectionality Theory Enrich Population Health Research?". *Social Science & Medicine* 178:214-16. doi: 10.1016/j.socscimed.2017.02.029.

Hernández-Yumar, A, M Wemrell, IA Alessón, BG López-Valcárcel, G Leckie and J Merlo. 2018. "Socioeconomic Differences in Body Mass Index in Spain: An Intersectional Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy." *PLoS One* 13(12):e0208624.

Jones, K, R Johnston and D Manley. 2016. "Uncovering Interactions in Multivariate Contingency Tables: A Multi-Level Modelling Exploratory Approach." *Methodological Innovations* 9:1-17.

Lizotte, DJ, M Mahendran, SM Churchill and GR Bauer. 2019. "Math Versus Meaning in MAIHDA: A Commentary on Multilevel Statistical Models for Quantitative Intersectionality." *Social Science & Medicine*.

May, VM. 2015. *Pursuing Intersectionality, Unsettling Dominant Imaginaries*. New York, NY: Routledge.

Marmot, M and R Bell. 2012. "Fair Society, Healthy Lives." *Public Health* 126 Suppl 1:S4-S10.

McCall, L. 2005. "The Complexity of Intersectionality." *Signs* 30(3):1771-800.

Merlo, J. 2003. "Multilevel Analytical Approaches in Social Epidemiology: Measures of Health Variation Compared with Traditional Measures of Association." *Journal of Epidemiology & Community Health* 57(8):550-52.

Merlo, J. 2014. "Invited Commentary: Multilevel Analysis of Individual Heterogeneity—a Fundamental Critique of the Current Probabilistic Risk Factor Epidemiology." *American Journal of Epidemiology* 180(208-212).

Merlo, J. 2018. "Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA) within an Intersectional Framework." *Social Science & Medicine* 203:74-80.

Merlo, J, S Mulinari, M Wemrell, SV Subramanian and B Hedblad. 2017. "The Tyranny of the Averages and the Indiscriminate Use of Risk Factors in Public Health: The Case of Coronary Heart Disease." *SSM-Population Health* 3(684-698).

Merlo, J and P Wagner. 2012. "The Tyranny of the Averages and Indiscriminate Use of Risk Factors in Public Health: A Call for Revolution." *Eur J Epidemiol* 28:148.

Merlo, J, P Wagner and G Leckie. 2019. "A Simple Multilevel Approach for Analysing Geographical Inequalities in Public Health Reports: The Case of Municipality Differences in Obesity." *Health & Place* 58:102145.

Nash, JC. 2008. "Re-Thinking Intersectionality." *Feminist Review* 89:1-15.

Persmark, A, M Wemrell, CR Evans, SV Subramanian, G Leckie and J Merlo. 2019. "Intersectional Inequalities and the U.S. Opioid Crisis: Challenging Dominant Narratives and Revealing Heterogeneities." *Critical Public Health*. doi: 10.1080/09581596.2019.1626002.

Rose, G. 1992. *The Strategy of Preventive Medicine*. Oxford England, New York: Oxford University Press.

Wemrell, M, S Mulinari and J Merlo. 2017a. "An Intersectional Approach to Multilevel Analysis of Individual Heterogeneity (MAIH) and Discriminatory Accuracy." *Social Science & Medicine* 178:217-19.

Wemrell, M, S Mulinari and J Merlo. 2017b. "Intersectionality and Risk for Ischemic Heart Disease in Sweden: Categorical and Anti-Categorical Approaches." *Social Science & Medicine* 177:213-22.