Variance partitioning in multilevel models for count data

George Leckie, William Browne, Harvey Goldstein,

Centre for Multilevel Modelling

School of Education

University of Bristol, UK

Juan Merlo

Unit for Social Epidemiology, Faculty of Medicine

Lund University

Malmö, Sweden

Peter Austin

Institute of Health Management, Policy and Evaluation,

University of Toronto

Ontario, Canada

Acknowledgements of support

This research was funded by UK Economic and Social Research Council grant ES/R010285/1. This work contains statistical data from Office for National Statistics (ONS), UK, which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data.

Address for correspondence

Professor George Leckie

Centre for Multilevel Modelling

School of Education

University of Bristol

35 Berkeley Square

Bristol

BS8 1JA

United Kingdom

g.leckie@bristol.ac.uk

Variance partitioning in multilevel models for count data

Abstract

A first step when fitting multilevel models to continuous responses is to explore the degree of clustering in the data. Researchers fit variance-component models and then report the proportion of variation in the response that is due to systematic differences between clusters or equally the response correlation between units within a cluster. These statistics are popularly referred to as variance partition coefficients (VPCs) and intraclass correlation coefficients (ICCs). When fitting multilevel models to categorical (binary, ordinal, or nominal) and count responses, these statistics prove more challenging to calculate. For categorical response models, researchers frequently appeal to their latent response formulations and report VPCs/ICCs in terms of latent continuous responses envisaged to underly the observed categorical responses. For standard count response models, however, there are no corresponding latent response formulations. More generally, there is a paucity of guidance on how to partition the variance. As a result, applied researchers are likely to avoid or inadequately report and discuss the substantive importance of clustering and cluster effects in their studies. A recent article drew attention to a little-known algebraic expression for the VPC/ICC for the special case of the two-level random-intercept Poisson model. In this article, we make a substantial new contribution. First, we derive VPC/ICC expressions for the more flexible negative binomial model that allows for overdispersion, a phenomenon which often occurs in practice with count data. Then we derive VPC/ICC expressions for three-level and random-coefficient extensions to these models. We illustrate all our work with an application to student absenteeism.

Introduction

Multilevel models (random effects, mixed-effects or hierarchical linear models) are now a standard generalization of conventional regression models for analyzing clustered and longitudinal data in the social, psychological, behavioral and medical sciences. Examples include students within schools, respondents within neighborhoods, patients with hospitals, repeated measures within subjects, and panel survey waves on households. Multilevel models have been further generalized to handle a wide range of response types, including, continuous, categorical (binary or dichotomous, ordinal, and nominal or discrete choice), count, and survival responses. Standard introductions to these models can be found in textbooks by Goldstein (2011b), Hox et al. (2017), Raudenbush and Bryk (2002), Skrondal and Rabe-Hesketh (2004), and Snijders and Bosker (2012).

A natural first step in any multilevel analysis is to report the degree of clustering in the response since the greater the degree of clustering, the greater the need for a multilevel approach for both statistical and substantive reasons. Consider a study of the relationships between a continuous student math score and a range of student and school characteristics. Here the concern is student math scores will correlate within schools (within-cluster dependence), even after adjusting for the covariates, thereby violating the linear regression assumption of independent residuals. Such clustering is envisaged to be due to unmodelled student and school influences on math scores that vary between schools (between-cluster heterogeneity). A two-level linear regression attempts to account for these influences, and therefore clustering, by including a school random-intercept effect. The total residual variance is then decomposed into separate within- and between-school components. The proportion of response variance which lies between schools (conditional on any covariates) can then be estimated and reported using the variance partition coefficient (VPC; Goldstein et al., 2002). This statistic is calculated as the ratio of the estimated between-school variance to the total residual variance. The VPC therefore summarizes the 'importance' of the clusters in influencing the response, in our case the importance of schools in influencing student outcomes above and beyond modelled student and school characteristics. As such, the VPC is widely quoted in multilevel studies. In the case of the current random-intercept model, the VPC can also be interpreted as the expected correlation between two students from the same school (conditional on any covariates), in which case it is referred to as the intraclass correlation coefficient (ICC). As a result, the two terms are often used interchangeably though for more complex random-coefficient models and for models with non-continuous responses the terms diverge in meaning (Goldstein et al., 2002). In this article, however, we shall focus primarily on the VPC interpretation.

VPCs can also be calculated in three- and higher-level models for continuous responses as well as in models with more complex random effect structures (cross-classified, multiple membership, spatial, and dyadic structures). In these settings, VPCs are often used to ascertain the relative importance of different sources of clustering in influencing the response. VPCs can also be calculated after fitting models with random coefficients. In all these cases, the VPC expressions become more complex and don't necessarily continue to have parallel ICC interpretations, but these issues are well described in the literature (see textbook references).

For multilevel models for categorical responses (e.g., binary, ordinal, and nominal logistic regression), calculating VPCs is less straightforward as these statistics typically no longer have closed-form expressions (they involve integrating out the random effects which can only be achieved via numerical integration). The standard approach to this problem is to appeal to the latent-response formulation of these models and to report VPCs in terms of an unobserved continuous variable envisaged to underlie the observed categorical responses (see textbook references). For example, in a study of whether students pass an exam, we would appeal to the notion of a continuous exam score scale underlying the observed binary pass or fail status and we would report the VPC in terms of this latent variable, that is, in terms of the propensity for the student to pass the exam. An appealing feature of this approach is that it allows one to calculate VPCs for categorical responses using essentially the same expressions as those derived for continuous responses.

For multilevel models for count responses (e.g., Poisson model and negative binomial regression), however, there are no corresponding latent-response formulations (Skrondal and Rabe-Hesketh, 2004) and so it is less obvious how one should calculate the VPC. More generally there is a paucity of guidance in the multilevel literature on how to partition the response variance. As a result, applied researchers are likely to avoid or inadequately report and discuss the substantive importance of clustering and cluster effects in their studies.

A notable exception is the work by Stryhn et al. (2006) and later Austin et al. (2017) who show that the VPC for the special case of a two-level random-intercept Poisson model does have a closed-form and so can be calculated with a simple algebraic expression. However, for many applications, researchers now routinely apply multilevel negative binomial models to account for overdispersion (the phenomenon whereby the variance of the observed counts is larger than that implied by the expectation) and so algebraic expressions for the VPC are also needed for these more flexible models. Likewise, many researchers now routinely fit count models allowing for three or more levels or random coefficients and so these VPC expressions must also be extended to account for these modelling extensions.

In absence of these algebraic expressions for calculating VPCs in count response models, one approach currently available to researchers is to apply the simulation method proposed by Goldstein et al. (2002) and further illustrated by Browne et al. (2005). While these articles discuss this method in the context of two-, three- and four-level binary response models, the method readily extends to the count models discussed here. Indeed, Austin et al. (2017) used this simulation method to confirm that the algebraic expression introduced there for the VPC for the special case of a two-level random-intercept Poisson model gives the correct value. The simulation method involves using the fitted model to simulate new count responses and to then calculate the within- and between-cluster variances for these simulated data, averaging over the data to approximate integrating out the random effects. The VPC can then be calculated in the usual way. The principal disadvantage of this simulation method is that it is computationally intensive. It has also not been implemented in software, forcing researchers to write their own code which is error prone, especially for models with complex random effect structures.

In this article, we instead derive algebraic expressions for the VPC for four different multilevel count response models and their extensions to multiple levels of random effects and random coefficients. We first focus in the sections below on the two-level random-intercept Poisson and negative binomial model (mean dispersion or NB2 version) as these are most widely applied. However, we additionally present VPC expressions in the Supplemental materials for the Poisson model with an overdispersion random effect and the constant dispersion or NB1 version of the negative binomial model. We then extend these VPC expressions so that they can be applied to more complex three-level and random-coefficient models. These expressions render the simulation method redundant and so we do not consider it further here. We illustrate our work with an application to student absenteeism.

The rest of this article proceeds as follows. In the next section we review two-level random-intercept versions of the Poisson and negative binomial models. We then derive the VPC for each model. Next we discuss three-level random-intercept versions of these models and present their VPC expressions. We then extend these models and VPC expressions to their random-coefficient versions. Next we present our application. We end by summarizing our findings and place them in the context of the existing literature.

Review of two-level random-intercept models for count responses

We will start by reviewing the two most widely applied multilevel models for count responses, namely the two-level random-intercept Poisson and negative binomial models (mean dispersion or NB2 version). We restrict our attention to the standard versions of these models which use the canonical log link. The negative binomial model allows for overdispersion with the Poisson model is the special case where the variance equals the expectation. These models can be fitted in standard software to give maximum likelihood estimates (via adaptive quadrature) and this is the approach we shall use in our application. Postestimation, the cluster effects can then be calculated via empirical Bayes predictions. These models can alternatively be fitted by Markov chain Monte Carlo methods and we note some of the potential advantages of this approach in the Discussion.

Poisson model

Let y_{ij} denote the count for unit i ($i = 1, \dots, n_j$) in cluster j ($j = 1, \dots, J$). In terms of our application, the units will be students, the clusters schools, and the count will be the number of days each student is absent from school over the course of the school year. We can then write the two-level random-intercept Poisson model for y_{ij} as follows

$$y_{ij}|\mu_{ij} \sim \text{Poisson}(\mu_{ij})$$
$$\ln(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_j \quad (1)$$
$$u_j \sim N(0, \sigma_u^2)$$

where μ_{ij} denotes the expected count, $\mathbf{x}_{ij}$ denotes the vector of unit- and cluster-level covariates (including the intercept, any cross-level interactions, and where relevant an offset), $\boldsymbol{\beta}$ is the associated vector of regression coefficients, and u_j is the cluster random intercept effect, assumed normally distributed with zero mean and variance σ_u^2. The exponentiated regression coefficients can be interpreted as incidence-rate ratios (IRRs) or ratios of expected counts.

The conditional expectation of y_{ij} (given $\mathbf{x}_{ij}$ and u_j) is given by

$$\mu_{ij}^C \equiv \text{E}(y_{ij}|\mathbf{x}_{ij}, u_j) = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j) \quad (2)$$

which in this model is simply equal to μ_{ij} in Equation 1, but this will not be the case in the next model hence the introduction here of the “C” superscript. The conditional variance is given by

$$\omega_{ij}^C \equiv \text{Var}(y_{ij}|\mathbf{x}_{ij}, u_j) = \mu_{ij}^C \quad (3)$$

Thus, the conditional variance of the counts is assumed to equal the conditional expectation. In practice, this equi-dispersion assumption often fails, with the variance of the observed counts in many clusters being larger or smaller than that implied by the mean, phenomena known as overdispersion (extra-Poisson variability) or underdispersion, respectively. Overdispersion is far more common than underdispersion and is typically attributed to unobserved unit-level covariates.

Negative binomial model

The negative binomial model (mean dispersion or NB2 version) is an extension of the Poisson model that adds a normally distributed unit-level overdispersion random effect to represent omitted unit-level variables that are envisaged to be driving any overdispersion. In contrast to the conventional cluster random intercept effect, this does not induce any dependence among the units. The model can be written as

$$
\begin{aligned}
y_{ij} | \mu_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
\ln(\mu_{ij}) &= \mathbf{x}'_{ij}\boldsymbol{\beta} + u_j + e_{ij} \\
u_j &\sim N(0, \sigma_u^2) \\
\exp(e_{ij}) &\sim \text{Gamma}\left(\frac{1}{\alpha}, \alpha\right)
\end{aligned}
\tag{4}
$$

where e_{ij} denotes the overdispersion random effect. Thus, in this model, two units with the same covariate and random intercept effect value may nonetheless differ in their expected counts μ_{ij}, with such differences attributed to the two units differing in terms of their values on the omitted unit-level variables. The exponentiated overdispersion random effect $\exp(e_{ij})$ is assumed gamma distributed with shape and scale parameters $1/\alpha$ and α and are therefore distributed with mean 1 and variance or overdispersion parameter α. The larger α is, the greater the overdispersion. When $\alpha = 0$, the model simplifies to the Poisson model (Equation 1) and so we can conduct a likelihood-ratio test to compare the two models to see whether the estimated overdispersion is statistically significant.

In this model, we can again calculate the conditional expectation and variance of the response. However, here we must integrate out the overdispersion random effect since this is not typically of substantive interest. The conditional expectation of y_{ij} (given $\mathbf{x}_{ij}$ and u_j but averaged over e_{ij}) has the same form as in the Poisson model (Equation 2), with

$$\mu_{ij}^{C} \equiv E\left(y_{ij} \middle| \mathbf{x}_{ij}, u_j\right) = \exp\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + u_j\right) \tag{5}$$

and we see that, in contrast to the previous model, $\mu_{ij}^{C} \neq \mu_{ij}$. The conditional variance is then given by

$$\omega_{ij}^{C} \equiv \mathrm{Var}\left(y_{ij} \middle| \mathbf{x}_{ij}, u_j\right) = \mu_{ij}^{C} + \left(\mu_{ij}^{C}\right)^2 \alpha \tag{6}$$

Thus, the conditional variance is now a quadratic function of the conditional expectation and is larger than the conditional expectation if $\alpha > 0$. Therefore, the usual variance-mean relationship for the Poisson model is relaxed, allowing overdispersion with respect to the conditional expectation ($\omega_{ij}^{C} > \mu_{ij}^{C}$).

3. Marginal statistics: Marginal expectation, variance, covariance, correlation, ICC and VPCs

Multilevel models for count responses are conditional (cluster-specific) models in the sense that they condition on the values of the random effects. In order to understand what has been assumed for the observed counts y_{ij} in the two-level random-intercept Poisson and negative binomial models presented in the previous section, we now present the marginal (population-averaged) expectations, variances, covariances and correlations of the responses (given the covariates but averaged over the random effects). We then present the ICC and VPC for each model. The expression for the VPC for the Poisson model was published in Stryhn (2006) and Austin et al. (2017). However, we are not aware of any publications presenting the VPC for the negative binomial model and so this is an important new result. More generally, while expressions for the other marginal statistics can be found in the literature, they are rarely found in one place and so we hope that our treatment below provides a further useful resource for readers. We provide full derivations for all these expressions in the Supplemental materials (Section S2).

We note that when desired, interval estimates can also be calculated for VPCs (Goldstein et al., 2002). When models are fitted by maximum likelihood estimation, a 95% confidence interval for the VPC can be constructed via the delta method (the VPC is a non-linear combination of the model parameters) or via multilevel bootstrapping (e.g., fitting the model to 1000 bootstrapped samples to obtain a sampling distribution for the VPC) (Goldstein, 2011a). When the model is fitted by MCMC methods, a 95% credible interval can be calculated using the MCMC chain for the posterior distribution of the VPC.

Poisson model

The marginal expectation of y_{ij} (given $\mathbf{x}_{ij}$ but now averaged over u_j) is given by

$$\mu_{ij}^{M} \equiv \mathrm{E}\left(y_{ij} \middle| \mathbf{x}_{ij}\right) = \exp\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + \sigma_u^2/2\right) \tag{7}$$

Given the additional term $\sigma_u^2/2$ is a constant and doesn't depend on the covariates, the intercept is the only parameter which is not the same as in the expressions for the conditional expectation (Equation 2; the conditional intercept is β_0 whereas the marginal intercept is larger, $\beta_0 + \sigma_u^2/2$; the remaining regression coefficients are the same). Thus, we note that in contrast to random-intercept logistic regression and other categorical response models, the regression coefficients have both cluster-specific and population-average interpretations (see textbook references).

The marginal variance of y_{ij} (given $\mathbf{x}_{ij}$ but averaged over u_j) is given by

$$\omega_{ij}^M \equiv \mathrm{Var}\left(y_{ij} \middle| \mathbf{x}_{ij}\right) = \mu_{ij}^M + \left(\mu_{ij}^M\right)^2 \{\exp(\sigma_u^2) - 1\} \tag{8}$$

The marginal variance is therefore a quadratic function of the marginal expectation and is larger than the marginal expectation if there is clustering, $\sigma_u^2 > 0$.

The marginal covariance of y_{ij} and $y_{i'j}$ (given $\mathbf{x}_{ij}$ and $\mathbf{x}_{i'j}$ but averaged over u_j) is a function of the covariate values for unit i and i' in cluster j. We focus on the special case where the covariate values are the same $\mathbf{x}_{ij} = \mathbf{x}_{i'j}$ as this results in a simpler expression for the ICC and one that coincides with the VPC presented below. This approach is also applied when calculating the ICC in random-coefficient models for continuous responses when the ICC is again a function of the covariate values of two different units (Goldstein et al., 2002). The resulting covariance is given by

$$\mathrm{Cov}\left(y_{ij}, y_{i'j} \middle| \mathbf{x}_{ij}\right) = \left(\mu_{ij}^M\right)^2 \{\exp(\sigma_u^2) - 1\} \tag{9}$$

The marginal correlation of y_{ij} and $y_{i'j}$ (given $\mathbf{x}_{ij} = \mathbf{x}_{i'j}$ but averaged over u_j) can then be calculated in the usual way to give

$$\text{Corr}\left(y_{ij}, y_{i'j} \middle| \mathbf{x}_{ij}\right) = \frac{\left(\mu_{ij}^{M}\right)^2 \{\exp(\sigma_u^2) - 1\}}{\mu_{ij}^{M} + \left(\mu_{ij}^{M}\right)^2 \{\exp(\sigma_u^2) - 1\}} \quad (10)$$

This marginal correlation can be interpreted as the ICC as it is the response correlation between two units in the same cluster (with the same covariate values). Note that as this expression depends on the marginal expectation, $\mu_{ij}^{M} = \exp\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + \sigma_u^2/2\right)$, it will take different values for different units within a cluster that have different covariate values. Specifically, the ICC is an increasing function of the marginal expectation.

The expression for the level-2 VPC – the proportion of the marginal response variance which lies between clusters – also coincides with that for the ICC. To see this, reconsider the expression for the marginal variance (Equation 8). The expression consists of the summation of two terms which can be shown to capture the within- and between-cluster variance in y_{ij} (Supplemental materials: Section S2.2). Specifically, the first term μ_{ij}^{M} captures the average variance within clusters in the observed unit-level counts y_{ij} around the expected counts μ_{ij}^{C} (Equation 2), while the second term $\left(\mu_{ij}^{M}\right)^2\{\exp(\sigma_u^2) - 1\}$ captures the variance between clusters in their expected counts μ_{ij}^{C} attributable to the cluster random intercept effect u_j The expression for the VPC can then be derived in the usual way: as the ratio of the level-2 component of the marginal variance divided by the summation of the level-2 and -1 components to give

$$\text{VPC}_{ij} = \frac{\overbrace{\left(\mu_{ij}^{M}\right)^{2}\{\exp(\sigma_u^2)-1\}}^{\text{level}-2\text{ variance}}}{\underbrace{\left(\mu_{ij}^{M}\right)^{2}\{\exp(\sigma_u^2)-1\}}_{\text{level}-2\text{ variance}}+\underbrace{\mu_{ij}^{M}}_{\text{level}-1\text{ variance}}} \tag{11}$$

and this expression is identical to that for the marginal correlation or ICC given in Equation 10. Thus, like the ICC, the VPC is an increasing function of the marginal expectation $\mu_{ij}^{M} = \exp\left(\mathbf{x}_{ij}' \boldsymbol{\beta} + \sigma_u^2/2\right)$. We also see the usual result whereby the higher the cluster variance σ_u^2, the higher the VPC. The level-1 VPC – the proportion of the marginal response variance which lies within clusters – is simply equal to one minus the level-2 VPC.

The dependence of the VPC on μ_{ij}^{M} means one should inspect how the VPC varies as a function of μ_{ij}^{M} and potentially also as a function of individual covariates. A simple approach is to compute the VPC for every unit in the data based on the covariate pattern for that unit. Typically, one will want to summarize this distribution. A natural choice is to report the mean of these VPC value a (or perhaps the median accompanied by the interquartile range to communicate the variability). Alternatively, one might calculate the VPC at specific meaningful values of the covariates. For example, at the covariate values associated with prototypical units and clusters. Finally, we note that when an offset is also included in the model, the VPC will additionally be a function of this variable as the offset can be viewed as just another covariate in $\mathbf{x}_{ij}'\boldsymbol{\beta}$, but whose regression coefficient is constrained to equal 1.

Negative binomial model

The marginal expectation of y_{ij} (given $\mathbf{x}_{ij}$ but now averaged over u_j as well as e_{ij}) is given by

$$\mu_{ij}^{M} \equiv \mathrm{E}\left(y_{ij} \middle| \mathbf{x}_{ij}\right) = \exp\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + \sigma_u^2/2\right) \quad (12)$$

which is the same as that for the Poisson model (Equation 7).

The marginal variance of y_{ij} (given $\mathbf{x}_{ij}$ but averaged over u_j and e_{ij}) is given by

$$\omega_{ij}^{M} \equiv \mathrm{Var}\left(y_{ij} \middle| \mathbf{x}_{ij}\right) = \mu_{ij}^{M} + \left(\mu_{ij}^{M}\right)^2 \{\exp(\sigma_u^2)(1+\alpha) - 1\} \quad (13)$$

which differs from that for the Poisson model (Equation 8) via the inclusion of the additional multiplicative term $(1 + \alpha)$. Thus, in this model, the marginal variance is larger than the marginal expectation if there is clustering $\sigma_u^2 > 0$ *or* overdispersion $\alpha > 0$.

The marginal covariance of y_{ij} and $y_{i'j}$ (given $\mathbf{x}_{ij} = \mathbf{x}_{i'j}$ but averaged over u_j, e_{ij} and $e_{i'j}$) is given by

$$\mathrm{Cov}\left(y_{ij}, y_{i'j} \middle| \mathbf{x}_{ij}\right) = \left(\mu_{ij}^{M}\right)^2 \{\exp(\sigma_u^2) - 1\} \quad (14)$$

and is the same as that for the Poisson model (Equation 9).

The marginal correlation of y_{ij} and $y_{i'j}$ (given $\mathbf{x}_{ij} = \mathbf{x}_{i'j}$ but averaged over u_j, e_{ij} and $e_{i'j}$) is then given by

$$\mathrm{Corr}\left(y_{ij}, y_{i'j} \middle| \mathbf{x}_{ij}\right) = \frac{\left(\mu_{ij}^{M}\right)^2 \{\exp(\sigma_u^2) - 1\}}{\mu_{ij}^{M} + \left(\mu_{ij}^{M}\right)^2 \{\exp(\sigma_u^2)(1+\alpha) - 1\}} \quad (15)$$

and differs from that for the Poisson model (Equation 10) only in the inclusion of the additional multiplicative term $(1 + \alpha)$ in the denominator.

As in the Poisson model, the marginal variance (Equation 13) can again be rearranged to equal the summation of two terms which can be shown to capture the within- and between-cluster variance in y_{ij}. The resulting level-2 VPC is given by

$$\text{VPC}_{ij} = \frac{\overbrace{\left(\mu_{ij}^{M}\right)^{2}\{\exp(\sigma_u^2) - 1\}}^{\text{level}-2\text{ variance}}}{\underbrace{\left(\mu_{ij}^{M}\right)^{2}\{\exp(\sigma_u^2) - 1\}}_{\text{level}-2\text{ variance}} + \underbrace{\mu_{ij}^{M} + \left(\mu_{ij}^{M}\right)^{2}\exp(\sigma_u^2)\alpha}_{\text{level}-1\text{ variance}}} \tag{16}$$

and this expression is identical (after rearranging terms) to that for the marginal correlation or ICC given in Equation 15.

Studying Equation 16, we see that as in the Poisson case the VPC is again an increasing function of the marginal expectation $\mu_{ij}^{M} = \exp\left(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \sigma_u^2/2\right)$. We also again see that the higher the cluster variance σ_u^2, the higher the VPC. However, we now additionally see that the higher the overdispersion parameter α, the lower the VPC. This makes sense. As the overdispersion increases, all else equal, the more unmodelled variation there is at level-1 and so the VPC decreases.

Comparing the two VPC expressions (Equations 11, 16), we see that the expression for the level-2 component of the marginal variance is the same and so it is only the expression for the level-1 component which varies across models. This makes sense as the models differ only in their treatment of overdispersion, which is viewed as a level-1 phenomenon. The overdispersion parameters in the negative binomial model leads the expression for the level-1 variance to exceed that of the Poisson model. The marginal variance is simply the summation of the level-2 and -1 variances and so is also expected to be higher in the negative binomial model compared to that of the Poisson model.

Three-level random-intercept models for count data and calculation of VPCs

In this section we focus on the more flexible negative binomial model and on presenting its level-specific VPC expressions. We provide full derivations for these expressions as well as those for the other marginal statistics in the Supplemental materials (Section S3). Note that as the Poisson model is simply the special case of the negative binomial model with no overdispersion, the level-specific VPC expressions for the Poisson model can be obtained by setting $\alpha = 0$ in the expressions below.

Let y_{ijk} denote the count for unit i ($i = 1, \dots, n_j$) in cluster j ($j = 1, \dots, J_k$) in supercluster k ($k = 1, \dots, K$). In terms of our application, the units will be students, the clusters schools, and the superclusters school districts. The three-level random-intercept negative binomial model can then be written as

$$
\begin{gathered}
y_{ijk} | \mu_{ijk} \sim \text{Poisson}(\mu_{ijk}) \\
\ln(\mu_{ijk}) = \mathbf{x}'_{ijk} \boldsymbol{\beta} + v_k + u_{jk} + e_{ijk} \\
v_k \sim N(0, \sigma_v^2) \\
u_{jk} \sim N(0, \sigma_u^2) \\
\exp(e_{ijk}) \sim \text{Gamma}\left(\frac{1}{\alpha}, \alpha\right)
\end{gathered}
\qquad (17)
$$

where v_k is the new supercluster random-intercept effect assumed normally distributed with zero mean and variance σ_v^2 and all other terms are defined as before.

With two higher levels there is now interest in reporting the relative importance of both superclusters and clusters as separate sources of response variation. As in the simpler two-level setting we can decompose the marginal variance ω_{ijk}^M into level-specific variance components and we use these to construct different VPC statistics (Supplemental materials: Section S3.2).

The level-3 VPC can then be written as

$$\mathrm{VPC}(3)_{ijk} = \frac{\overbrace{\left(\mu_{ijk}^M\right)^2\{\exp(\sigma_v^2)-1\}}^{\text{level}-3\text{ variance}}}{\underbrace{\left(\mu_{ijk}^M\right)^2\{\exp(\sigma_v^2)-1\}}_{\text{level}-3\text{ variance}}+\underbrace{\left(\mu_{ijk}^M\right)^2\exp(\sigma_v^2)\{\exp(\sigma_u^2)-1\}}_{\text{level}-2\text{ variance}}+\underbrace{\mu_{ijk}^M+\left(\mu_{ijk}^M\right)^2\exp(\sigma_v^2+\sigma_u^2)\alpha}_{\text{level}-1\text{ variance}}} \quad (18)$$

and is interpreted as the proportion of response variance which lies between superclusters. This expression can also be interpreted as an ICC as it also gives the response correlation between two units in the same supercluster, but different clusters.

The level-2 VPC can be written as

$$\mathrm{VPC}(2)_{ijk} = \frac{\overbrace{\left(\mu_{ijk}^M\right)^2\exp(\sigma_v^2)\{\exp(\sigma_u^2)-1\}}^{\text{level}-2\text{ variance}}}{\underbrace{\left(\mu_{ijk}^M\right)^2\{\exp(\sigma_v^2)-1\}}_{\text{level}-3\text{ variance}}+\underbrace{\left(\mu_{ijk}^M\right)^2\exp(\sigma_v^2)\{\exp(\sigma_u^2)-1\}}_{\text{level}-2\text{ variance}}+\underbrace{\mu_{ijk}^M+\left(\mu_{ijk}^M\right)^2\exp(\sigma_v^2+\sigma_u^2)\alpha}_{\text{level}-1\text{ variance}}} \quad (19)$$

and is interpreted as the proportion of response variance which lies within superclusters, between clusters. This VPC does not have a corresponding ICC interpretation. This can be seen by realizing that the implied correlation would be between two units in different superclusters, but the same cluster and this is not a possibility in hierarchical data.

We can also calculate the proportion of response variance collectively attributable to superclusters and clusters. This VPC is calculated by replacing the numerator in the previous equations with the sum of the level-3 and level-2 variances.

$$\text{VPC}(2,3)_{ijk} = \frac{\overbrace{\left(\mu_{ijk}^{M}\right)^{2}\{\exp(\sigma_v^2)-1\}}^{\text{level}-3\text{ variance}}+\overbrace{\left(\mu_{ijk}^{M}\right)^{2}\exp(\sigma_v^2)\{\exp(\sigma_u^2)-1\}}^{\text{level}-2\text{ variance}}}{\underbrace{\left(\mu_{ijk}^{M}\right)^{2}\{\exp(\sigma_v^2)-1\}}_{\text{level}-3\text{ variance}}+\underbrace{\left(\mu_{ijk}^{M}\right)^{2}\exp(\sigma_v^2)\{\exp(\sigma_u^2)-1\}}_{\text{level}-2\text{ variance}}+\underbrace{\mu_{ijk}^{M}+\left(\mu_{ijk}^{M}\right)^{2}\exp(\sigma_v^2+\sigma_u^2)\alpha}_{\text{level}-1\text{ variance}}} \quad (20)$$

This expression once again has an ICC interpretation, namely, the response correlation between two units in the same supercluster and the same cluster. This correlation will be stronger than that between two units in the same supercluster but different clusters (Equation 18) as here the units share not only unobserved supercluster influences (captured by the level-3 variance in the numerator), but now additionally share unobserved cluster influences (capture by the level-2 variance in the numerator). The level-1 VPC – the proportion of the marginal response variance which lies within clusters – is simply equal to one minus this joint level-3 and level-2 VPC.

Random coefficient models and calculation of VPC

In this section we focus on the two-level random-coefficient negative binomial model (i.e., regression coefficients, not just the intercept, vary across clusters), but the Poisson model and three-level versions of both models can be easily extended to include random coefficients in a parallel fashion (see Sections S4 in the Supplemental materials).

The two-level random-coefficient negative binomial model for y_{ij} can be written as follows

$$y_{ij}|\mu_{ij} \sim \text{Poisson}(\mu_{ij})$$

$$\ln(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_j \qquad (21)$$

$$\mathbf{u}_j \sim N(0, \boldsymbol{\Omega}_{\mathbf{u}})$$

$$\exp(e_{ij}) \sim \text{Gamma}\left(\frac{1}{\alpha}, \alpha\right)$$

where $\mathbf{z}_{ij}$ denotes a vector of unit- and cluster-level covariates (typically an intercept and a subset of the unit-level covariates in $\mathbf{x}_{ij}$) and $\mathbf{u}_j$ is the associated vector of cluster random coefficient effects, assumed multivariate normally distributed with zero mean vector and covariance matrix $\boldsymbol{\Omega}_{\mathbf{u}}$.

The expression for the VPC is as in Equation 16, but where σ_u^2 is replaced by the cluster variance function $\mathbf{z}'_{ij}\boldsymbol{\Omega}_{\mathbf{u}}\mathbf{z}_{ij}$. For simplicity, consider the special case of a model with a random intercept and one random coefficient associated with the first predictor x_{1ij}. In this case we have $\mathbf{z}'_{ij}\mathbf{u}_j = u_{0j} + u_{1j}x_{1ij}$ and $\mathbf{z}'_{ij}\boldsymbol{\Omega}_{\mathbf{u}}\mathbf{z}_{ij} = \sigma_u^2 + 2\sigma_{u01}x_{1ij} + \sigma_{u1}^2x_{1ij}^2$. The expression for the VPC is then given by

$$\text{VPC}_{ij} = \frac{\overbrace{\left(\mu_{ij}^M\right)^2\left\{\exp\left(\sigma_u^2+2\sigma_{u01}x_{1ij}+\sigma_{u1}^2x_{1ij}^2\right)-1\right\}}^{\text{level-2 variance}}}{\underbrace{\left(\mu_{ij}^M\right)^2\left\{\exp\left(\sigma_u^2+2\sigma_{u01}x_{1ij}+\sigma_{u1}^2x_{1ij}^2\right)-1\right\}}_{\text{level-2 variance}}+\underbrace{\mu_{ij}^M+\left(\mu_{ij}^M\right)^2\exp\left(\sigma_u^2+2\sigma_{u01}x_{1ij}+\sigma_{u1}^2x_{1ij}^2\right)\alpha}_{\text{level-1 variance}}} \qquad (22)$$

where the marginal expectation is defined as follows

$$\mu_{ij}^M = \exp\left\{\mathbf{x}'_{ij}\boldsymbol{\beta} + \left(\sigma_u^2 + 2\sigma_{u01}x_{1ij} + \sigma_{u1}^2x_{1ij}^2\right)/2\right\}$$

The VPC is now a function of both the marginal expectation and the cluster variance function.

When researchers discuss the VPC in the case of random-coefficient models for continuous responses, they typically illustrate how the VPC varies as a function of the covariate with the random coefficient (Goldstein et al., 2002). For example, if the effect of student socioeconomic status is allowed to vary randomly across schools, one could report the VPC as a function of socioeconomic status. In the current case, matters are complicated by the additional dependence on the marginal expectation which also includes all other covariates. One way to study the VPC in this model is to therefore first evaluate it at the covariate values for each unit in the data and then plot the resulting VPC values against the marginal expectation separately for different values of the covariate with the random coefficient. This plot can then be compared to the same plot based on for the simpler random-intercept model (which will show a single relationship between the VPC and marginal expectation). We will illustrate this approach in our application.

Application: Student absenteeism from school

Student absenteeism and its detrimental effects on student learning are ongoing concerns in the US (EPI, 2018), UK (DfE, 2019a) and many other countries. In response, school accountability systems are increasingly monitoring student absenteeism rates alongside more traditional attainment and progress measures. Student absenteeism is known to vary by student demographic and socioeconomic characteristics, factors which also typically vary across school intakes. It would therefore seem important to adjust student absenteeism rates for school differences in student composition before making any attempt to potentially hold schools accountable for their performance. In this application, we explore these ideas using multilevel models for count data. These models and notions of adjusting student outcomes for student characteristics when comparing schools are analogous to those used to estimate school value-added effects on student attainment (Castellano and Ho, 2013; Goldstein, 1997; Leckie and Goldstein, 2019; OECD, 2008; Raudenbush and Willms, 1995).

Our data relate to students in London schools who completed their compulsory secondary schooling at the end of the 2016/17 academic year (students aged 15/16 in UK school year 11; equivalent to US 10th grade). The data are drawn from the national pupil database, a census of all students in state-maintained schools in England (DfE, 2019b). The data have a three-level hierarchy consisting of 66,955 students (level-1) nested in 434 schools (level-2) nested in 32 school districts (level-3). The response is a count of the number of days students were absent from school during the academic year. Figure 1 presents a frequency distribution. The distribution is positively skewed, with students, on average, absent for 8.41 days over the academic year, but with an interquartile range spanning from 2 to 10 days (Variance 124.39; Min 0; Max 156). Figure 2 illustrates variation in the mean number of days absent by district (left; Mean 8.36; IQR = 7.56, 9.10) and by school (right; Mean = 8.46; IQR = 6.56, 9.76). We see meaningful differences in student absence rates, especially between schools. While in this application we analyze these counts using Poisson and negative binomial models, other possibilities would be to analyze them using binomial count responses models (as we know the total number of days in the academic year; approximately 155) or to first log transform the counts (adding 0.5 or some other value to all counts prior to transforming to avoid logarithms of zero counts) then analyze them using continuous response models.

We start by fitting “empty” or “null” single-, two-, and three-level random-intercept Poisson and negative binomial models with no covariates. In the continuous response case, these models are referred to as variance-component models and we use that terminology here. The aim of these initial analyses is to quantify the degree of clustering and overdispersion in the data and we do this using our VPC expressions. We will then pick a preferred model and extend it by entering student characteristics as covariates. Here our aim is to not just study the predictors of student absenteeism, but to adjust for these factors so that the resulting predicted school random effects provide more meaningful estimates of school influences on students that are plausibly related to factors within schools' control. At this point, we recalculate the VPCs and explore how they now vary as a function of the marginal expectation and covariates. We will investigate the predicted school effects. Last, we shall fit a random coefficients model, in which we allow the effect of one of the student covariates to vary across schools, again exploring the implications for calculating and interpreting the VPC.

We fit all models using maximum likelihood estimation (adaptive quadrature) as implemented in the "mepoisson" and "menbreg" commands in Stata (StataCorp, 2019). The models are nested and so we compare models using likelihood-ratio tests. These

Variance-component models

Table 1 presents the results. Model 1 is a conventional single-level Poisson model, but with no covariates. The estimated intercept of 2.129 implies a marginal expectation of 8.41 (= exp(2.129)) and this equals the sample mean number of days absent. The model-based marginal variance is also 8.41 due to the single-level Poisson model assumption that the marginal variance equals the marginal expectation. The sample variance, however, is 124.39 and so the single-level Poisson model fails to capture the true extent of variability in the data.

Model 2 is a two-level Poisson model (Equation 1). The model includes a school random intercept to investigate and account for school clustering. The model estimates the school variance to be 0.100 and a likelihood-ratio test confirms that this variation in student absence rates between schools is statistically significant (Model 2 vs. Model 1: $\chi_1^2 = 53194, p < 0.001$). What is less clear is the practical or substantive importance of this variation. Put simply, is a value of 0.100 big? Does school clustering matter? Do we care? We can answer these questions using the estimated marginal statistics. The estimated marginal expectation changes little, from 8.41 to 8.46 (application of Equation 7). The model-based marginal variance, however, almost doubles, from 8.41 to 15.98 (application of Equation 8). This implies that school clustering accounts for approximately half the variation captured by the model. This is confirmed by the decomposition of the estimated marginal variance. The school component equals 7.52 while the student component equals 8.46. The resulting school VPC equals 0.47 and so 47% of the marginal variance is due to systematic differences between schools (application of Equation 11). This suggests that school level factors, or at least school-level variation in student characteristics, account for half the variation in student absenteeism between students. The marginal variance of 15.98, however, is still far below the sample variance of 124.39 and so this model also proves inadequate for these data. The low marginal variance is likely due to the level-1 component of the marginal variance still being constrained to equal the marginal expectation, hence their identical estimate of 8.46.

Model 3 is a two-level negative binomial model (Equation 4). The model includes a student overdispersion random effect to account for any within-school variation due to omitted student influences. The overdispersion parameter estimated to be 0.877 and a likelihood-ratio test confirms that there is significant overdispersion (Model 3 vs. Model 2: $\chi_1^2 = 363096, p <$

0.001). What is less clear is the practical or substantive importance of this additional variation. Put simply, is a value of 0.877 big? Does overdispersion matter? Do we care? Here too, we can answer these questions using the estimated marginal statistics. Most importantly, the estimated marginal variance now increases from 15.98 to 84.10 (application of Equation 13). As expected, the school component remains approximately stable and so the increase in the marginal variance is brought about by the student component which increases nine-fold from 8.46 to 77.15. This increase indicates that, even within schools, student absenteeism is far from a random Poisson process, rather the models suggest that there is substantial within school variability driven by omitted student characteristics. This difference in the student component in turn has a dramatic impact on the estimated VPC. The model estimates the school VPC to be 0.08 (application of Equation 16), suggesting that it is in fact omitted student-specific factors rather than omitted school-specific factors that are likely the dominant cause of the variation in absenteeism. This estimated VPC is far lower than the estimate of 0.47 reported for Model 2. Thus, an important finding is that by ignoring overdispersion, the Poisson model grossly overestimates the true importance of schools in these data. More generally, the Poisson VPC is biased upwards in the presence of overdispersion.

Model 4 is a three-level negative binomial model (Equation 17). The model includes a district random intercept effect to investigate and account for potential superclustering by district. The estimated intercept is effectively unchanged. The model estimates the district variance to be 0.006 and a likelihood-ratio test confirms that this district variation in student absence rates is statistically significant (Model 4 vs. Model 3: $\chi^2_1 = 7.29, p < 0.001$). The school variance in turn decreases by 0.006 from 0.093 to 0.087. The overdispersion parameter is also unchanged. This is expected as overdispersion is treated as a level-1 phenomenon and so is unaffected by whether the school variation is decomposed into separate within and between district components, as we have done here, or not. Here, we see immediately that districts are of little practical or substantive importance when studying student absenteeism, as the estimated variance is only around a tenth of the magnitude of the school variance. We can confirm this using the estimated VPC statistics for three-level models (Equations 18, 19, 20). Districts, schools, and students account for 0.5%, 8% and 92% of the variations in days absent. We will therefore not consider three-level models further in this application.

Crucially, all of the substantive insights we have given have only been made possible by estimating and interpreting the VPC expressions and various marginal statistics presented throughout the paper.

Random-intercept model with student covariates

Table 2 presents the results. Model 5 is a two-level negative binomial model where we include seven student covariates: prior attainment (test score quintile, based on tests taken five years earlier just before the start of secondary schooling), age (season of birth: Autumn, Winter, Spring, Summer; note that grade retention and acceleration is not a feature of the UK education system so children vary only in their month of birth, not their academic year of birth), gender, ethnicity (white, mixed, Asian, black, other), language (English or not), special educational needs (SEN), and free school meals (FSM). Table S1 in the Supplemental materials presents variable definitions and summary statistics.

A likelihood-ratio test confirms that the current model is preferred to its empty counterpart (Model 5 vs. Model 3; $\chi_1^2 = 6609$, $p < 0.001$) and so adding the covariates improves the fit of the model. The current model also continues to be preferred to its single-level counterpart (Model

5 vs. a single-level negative binomial model with covariates; results not shown; $\chi_1^2 = 5653$, $p < 0.001$) indicating that significant school clustering remains even after adjusting for the covariates. Similarly, the current model continues to be preferred to its Poisson counterpart (a two-level random-intercept Poisson model with covariates; results not shown; $\chi_1^2 = 314336$, $p < 0.001$) and so overdispersion also remains in the residual variation.

Examining the parameter estimates, we see that all predictors are statistically significant. Student absenteeism in London is, on average, higher among lower prior attainers, older students, girls, white students, those who speak English as a first language, those with SEN, and those on FSM. The exponentiated regression coefficients can be interpreted as incidence-rate ratios (IRRs) or ratios of expected counts. Consider, for example, the FSM estimate of 0.377. The estimated IRR is 1.46 (= exp(0.377)). Thus, FSM students are, on average, predicted to miss almost one and a half days for every day missed by otherwise equivalent non-FSM students. This differential is substantial. Interestingly, introducing the student characteristics does not lead the school variance to decrease as we move from Model 3 to Model 5 suggesting that school differences in student absenteeism are not predicted by school differences in student characteristics. Indeed, the school variance increases from 0.093 to 0.103. However, little should be read into this change (e.g., possible suppression effect) as this increase of 0.010 is very small relative to its standard error of 0.007. The overdispersion parameter decreases from 0.877 to 0.782.

Next consider the marginal statistics. Each of these quantities is now an increasing function of the marginal expectation. To explore this, we compute the VPC for every student in the data based on their covariate pattern. Thus, we first compute the predicted marginal expectation for each student (based on Equation 12). We then calculate the VPC for each student (based on Equation 16). Figure 3 plots these predicted VPC values against the predicted marginal expectation values (top panel). The figure also plots the distribution of predicted marginal expectation values in the sample (bottom panel). The VPC increases from approximately 0.085 to 0.105 as we increase the marginal expectation from its minimum to maximum predicted values. In models where the covariates have greater explanatory power, the marginal expectation and therefore VPC would be expected to vary more. Rather than plot the VPC distribution, researchers will often prefer to report a single summary statistic. A natural choice is to report the mean, in our case 0.10 (or perhaps the median accompanied by the interquartile range to communicate the variability). This VPC is slightly higher than that in the variance-components model, 0.08, suggesting that the covariates have explained a higher proportion of student variation in the data compared to the school variation which perhaps might be expected given the covariates are defined at the student level.

Table 2 presents sample mean values for the various other estimated marginal statistics. These estimates are broadly similar to those reported for the two-level variance-components model (Table 1, Model 3). This suggests the covariates have low explanatory power. The most important predictors for student absenteeism would appear to lie beyond those available to us here. Nonetheless, as detecting outlying schools was one of the motivations for this illustrative application, Figure 4 presents a scatterplot of the predicted school random effects from the current model (Model 5) against those based on the variance-components model (i.e., the null model; Model 3) (left panel). The figure also presents the scatterplot in terms of the ranks of these two sets of predicted school effects (right panel). The corresponding correlations are both 0.99, which are very high and confirm that adjusting for school differences in student composition, at least with respect to the factors examined here, does not lead to a substantial reordering of schools.

An important final question is: What would be the consequences of ignoring the significant and substantial residual overdispersion seen in these data? We can answer this question by fitting a Poisson version of the current two-level random-intercept model (not shown). The regression coefficients are almost identical. The standard errors, however, are approximately one third those in the negative binomial model. The Poisson estimates are therefore spuriously precise because they ignore the overdispersion. Accordingly, they should not be trusted. Note, however, that were the models to additionally include school-level covariates, then the standard errors on these regression coefficient would be expected to be far more similar across the Poisson and negative binomial versions of the model, as the precision with which the coefficients of cluster-level covariates are estimated is determined primarily by the cluster-level variance and the estimate of this parameter is similar in both models.

Random coefficient model

The previous model predicted FSM students in London miss, on average, almost one and a half days for every day missed by otherwise equivalent non-FSM students. In Model 6 we now allow this average effect to vary across schools by introducing a random coefficient for FSM. The estimates are presented in Table 2. The model has two extra parameters, a random slope variance σ_{u1}^2 and a random intercept-slope covariance σ_{u01}. A likelihood-ratio test confirms that this model is statistically preferred to the simpler random-intercept model (Model 5 vs. Model 4: $\chi_2^2 = 170, p < 0.001$). The random FSM effects are assumed normally distributed with an estimated mean of 0.372 (IRR = 1.45) and an estimated variance of 0.035. The 95% limits of this distribution are 0.005 and 0.739 (IRR = 1.00, 2.09). Thus, the FSM gap in student absenteeism varies substantially across London schools with FSM students in some schools missing no more days, on average, than otherwise equivalent non-FSM students in these schools, but with FSM students in other schools missing over two days, on average, for every day missed by otherwise equivalent non-FSM students in those schools.

Figure 5 explores the relationship between the predicted VPC and the marginal expectation. This plot should be contrasted with that based on the simpler random-intercept model shown earlier in Figure 3. The predicted VPC now varies not only as a function of the marginal expectation, but also as a function of the estimated school variance function $\sigma_u^2 + 2\sigma_{u01}x_{1ij} + \sigma_{u1}^2 x_{1ij}^2$, where x_{1ij} denotes the dummy variable for FSM (Equation 22). The school variance function simplifies to σ_{u0}^2 for non-FSM students and $\sigma_{u0}^2 + 2\sigma_{u01} + \sigma_{u1}^2$ for FSM students and thus results in two predicted values: 0.116 for non-FSM students and 0.097 for FSM students. This in turn leads to two distinct relationships between the predicted VPC and the marginal expectation and these are plotted in the figure. The relationship for non-FSM students lies above that for FSM students. Thus, for any given predicted number of days absent, the estimated VPC is higher for non-FSM students than for FSM students. This suggests that the influence of school attended on student absenteeism is more pronounced for non-FSM students than FSM students. That is, non-FSM students appear more sensitive and susceptible to their environments with respect to being absent from school than FSM students. This is an interesting finding worthy of further explanation and again highlights the additional insights provided when one calculates VPCs in count models.

Discussion

In this article, we have derived algebraic expressions for variance partition coefficients (VPCs) in multilevel Poisson and negative binomial models (mean dispersion or NB2 version) for count data. We have presented expressions for two- and three-level random-intercept versions of these models and we have shown how these can be extended to accommodate models with random coefficients. Parallel derivations are provided in the Supplemental materials for two further count models: the Poisson model with an overdispersion random effect and the constant dispersion or NB1 negative binomial model. This work significantly extends that of Stryhn (2006) and Austin et al. (2018) who focus only on the special case of a two-level random-intercept Poisson model. We are not aware of any other publications presenting the VPC expressions for the three more general count models that we consider in this article and that allow for overdispersion and so all of these results are new. We view these extensions as important as overdispersion a phenomenon which typically occurs in practice.

While the presented VPC expressions have the same general form as those for continuous and categorical responses (when expressed in terms of the latent responses underlying the observed binary, ordinal, and nominal responses), they are nonetheless more complex as they are increasing functions of the marginal expectation. Consequently, VPCs are not comparable across studies unless the marginal expectations are the same. Furthermore, where models include covariates, the dependence on the marginal expectation leads the VPC to be a function of the covariates. One must therefore choose the covariate values at which to evaluate the VPCs. This is also the case in random-coefficient models for continuous and categorical responses, so this is not a new idea for readers familiar with those models (Goldstein et al., 2002). A natural choice is to report the mean (or perhaps the median accompanied by the interquartile range to communicate the variability). There is no reason why such calculations cannot be automated in software as standard postestimation commands and we encourage software developers to do this.

While we have focused on deriving algebraic expressions for VPCs in different count models, our research also suggests some recommendations for applied researchers analyzing multilevel count data.

First, when count data exhibit overdispersion, as is frequently the case, the standard multilevel Poisson model will prove inadequate. The regression coefficients and cluster variance are not expected to be systematically affected by the overdispersion. The VPC, however, will be biased upwards in the standard Poisson model and so the degree of clustering and the cluster effects will appear more important than they truly are. Furthermore, the standard errors of the regression coefficient for unit-level (level-1) covariates, but not cluster-level (level-2) covariates, will be biased downwards, potentially leading to Type I errors of inference and incorrect research conclusions.

Second, we have also found the Poisson model to sometimes run into computational difficulties when fitted to data with substantial overdispersion. We therefore recommend negative binomial models since these account for overdispersion and so avoid the problems associated with the standard Poisson model. Of these, the mean dispersion is the more widely applied and is the version we focused on in this article. However, in software where negative binomial models are not implemented, or where the researcher is more familiar with the Poisson model, one alternative would be to simply add a unit-level overdispersion random effect to the Poisson model (see Section S1 in the Supplemental materials). The resulting model is very similar in form to the mean dispersion negative binomial model and would be expected to lead to similar results. However, in contrast to the negative binomial model, this overdispersed Poisson model proves computationally burdensome and will potentially prove prohibitive in large data settings as it requires integrating out the unit-level overdispersion random effect.

Third, we have found that as all the models we have discussed become more complex, fitting them by maximum likelihood estimation proves challenging, both in terms of increased convergence difficulties and increased computation time. Thus, researchers planning to fit count models allowing for three or more levels or random coefficients, especially in large datasets where there are often many clusters or superclusters, may be better off using MCMC methods.

Our most important recommendation, however, is that researchers should always explore competing models on their data, just as we have done so here. One learns more from the data doing this than by restricting attention to any one model.

An additional benefit of our work is that we derive and present expressions for not just the VPC, but the marginal expectation, variance, covariance, correlation, and ICC for all the different models considered in the article (see Sections S2, S3 and S4 of the supplemental materials). These expressions are rarely found in one place and so we hope that our treatment provides a useful resource for readers. For example, researchers can use these expressions when simulating count data or when designing simulation studies to choose true parameter values such that they imply a certain marginal expectation and variance and degree of clustering or overdispersion in the population.

Our research also suggests some areas for further work. First, we have explained how our algebraic expressions for the VPC for two-level models extend to three-level models and the same steps can be followed to further extend these expressions to four and higher-level settings. However, it is less obvious how they extend to cross-classified and multiple membership models and more work is required here.

Second, we have restricted our attention to the standard versions of the Poisson and negative binomial models which use the canonical log link. In some applications, researchers may wish instead to use the identity or power link functions and these will lead to different VPC expressions. More generally, there are a range of more complex multilevel count models which we have not explored, but for which it should be possible to extend the algebraic VPC expressions presented here to incorporate those modelling extensions. These include generalized negative binomial models, with-zeros or zero-inflated models, truncated and censored models, hurdle or two-part models, and mixture models (Cameron and Trivedi, 2013). We leave these extensions for future research.

Third, count data can be modelled as ordinal data, possibly after some grouping of counts to limit the number of observed categories. An advantage of this approach is that we can then appeal to the latent response formulation of ordinal models and their associated VPC expressions. Thus, we can then model the counts as being due to a latent continuous process that on crossing progressively higher thresholds leads to progressively higher values of the observed count. This approach would seem most useful when there are just a few low observed counts, say 0, 1, and 2 or more, as when there are many categories, as there would be in our application, this would lead to many additional threshold parameters to be estimated. One solution would be to merge adjacent categories, but the resulting coarsening of the data will often prove unappealing. Goldstein and Kounali (2009) present an alternative solution which is to apply a smoothing function to the threshold parameters. Relevant to the current work, they show how with sufficient structure imposed on these threshold parameters the ordered probit model reduces to the standard Poisson model. They refer to this formulation of the Poisson model as the ‘Poison latent normal transformation’. Thus, in future work we shall explore whether this alternative formulation of the Poisson model in terms of a latent continuous response variable leads to a simple and easy to interpret VPC expression which can complement those shown here for the observed count data.

References

Austin, P. C., Stryhn, H., Leckie, G., & Merlo, J. (2018). Measures of clustering and heterogeneity in multilevel Poisson regression analyses of rates/count data. *Statistics in Medicine*, *37*, 572-589.

Browne, W. J., Subramanian, S. V., Jones, K., & Goldstein, H. (2005). Variance partitioning in multilevel logistic models that exhibit overdispersion. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*, 599-613.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data (2nd ed.)*. Cambridge university press.

Castellano, K. E., & Ho, A. D. (2013). *A Practitioner's Guide to Growth Models*. Council of Chief State School Officers.

DfE (2019a). A guide to absence statistics: March 2019. Department for Education, London.

DfE (2019b). National pupil database. Department for Education, London. URL: https://www.gov.uk/government/collections/national-pupil-database.

EPI (2018). Student absenteeism. Who misses school and how missing school matters for performance. Economic Policy Institute, Washington DC.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement*, *8*, 369–395.

Goldstein, H. (2011a). Bootstrapping in Multilevel Models. In: Hox J. J., Roberts JK (eds). *Handbook of Advanced Multilevel Analysis*. Routledge: New York, NY,163-171.

Goldstein, H. (2011b). *Multilevel statistical models (4th ed.)*. Chichester, UK: Wiley.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *1*, 223-231.

Goldstein, H., & Kounali, D. (2009). Multilevel multivariate modelling of childhood growth, numbers of growth measurements and adult characteristics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*, 599-613.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications (3rd ed.)*. Routledge.

Leckie, G., & Goldstein, H. (2019). The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability in England. *British Educational Research Journal*, *45*, 518-537.

OECD (2008). Measuring improvements in learning outcomes: Best practices to assess the value - added of schools (Paris, Organisation for Economic Co - operation and Development Publishing & Centre for Educational Research and Innovation).

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata (3rd ed., Vol. 2: Categorical, count and survival responses)*. College Station, TX: Stata Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.)*. Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Willms, J. (1995). The estimation of school effects, *Journal of Educational and Behavioral Statistics*, *20*, 307-335.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modelling (2nd ed.)*. London: Sage.

StataCorp. (2019). Stata statistical software: Release 16 [Computer software]. College Station, TX: StataCorp LLC. URL: http://www.stata.com.

Stryhn, H., Sanchez, J., Morley, P., Booker, C., & Dohoo, I. R. (2006). Interpretation of variance parameters in multilevel Poisson models. In *Proceedings of the 11th Symposium of the International Society for Veterinary Epidemiology an Economics*, 702–704. Cairns, Australia.

Table 1. Estimates for variance components models fitted to the student absenteeism data.

	Model 1: Single-level variance-components Poisson model	Model 2: Two-level variance-components Poisson model	Model 3: Two-level variance-components negative binomial model	Model 4: Three-level variance-components negative binomial model
		Parameter estimates		
β_0 – Intercept	2.129 (0.001)	2.085 (0.015)	2.088 (0.015)	2.086 (0.020)
σ_v^2 – District variance				0.006 (0.003)
σ_u^2 – School variance		0.100 (0.007)	0.093 (0.007)	0.087 (0.007)
α – Overdispersion			0.877 (0.005)	0.877 (0.005)
		Marginal statistics		
Marginal expectation	8.41	8.46	8.45	8.44
Marginal variance	8.41	15.98	84.10	83.79
District (level-3) component				0.42
School (level-2) component		7.52	6.95	6.50
Student (level-1) component		8.46	77.15	76.87
District (level-3) VPC				0.005
School (level-2) VPC		0.47	0.08	0.08
Student (level-1) VPC		0.53	0.92	0.92
		Fit statistics		

Deviance	838336	785142	422046	422039

Note. Number of districts: $K = 32$; number of schools: $J = 434$; number of students: $N = 66{,}955$. Standard errors in parentheses.

Table 2. Estimates for two-level random-intercept and -coefficient models fitted to the student absenteeism data.

	Model 5: Two-level random-intercept negative binomial model	Model 6: Two-level random-coefficient negative binomial model
	Parameter estimates	
β_0 – Intercept	2.126 (0.021)	2.126 (0.021)
β_1 – Prior attainment: Quintile 2	-0.051 (0.012)	-0.048 (0.012)
β_2 – Prior attainment: Quintile 3	-0.118 (0.012)	-0.116 (0.012)
β_3 – Prior attainment: Quintile 4	-0.222 (0.012)	-0.219 (0.012)
β_4 – Prior attainment: Quintile 5	-0.330 (0.014)	-0.326 (0.014)
β_5 – Age: Spring born	0.026 (0.011)	0.026 (0.011)
β_6 – Age: Winter born	0.077 (0.011)	0.078 (0.011)
β_7 – Age: Autumn born	0.112 (0.011)	0.112 (0.011)
β_8 – Female	0.122 (0.009)	0.122 (0.009)
β_9 – Ethnicity: Mixed	-0.073 (0.014)	-0.074 (0.014)
β_{10} – Ethnicity: Asian	-0.194 (0.013)	-0.198 (0.013)
β_{11} – Ethnicity: Black	-0.422 (0.011)	-0.421 (0.011)
β_{12} – Ethnicity: Other	-0.194 (0.017)	-0.195 (0.017)
β_{13} – Language not English	-0.244 (0.009)	-0.242 (0.009)
β_{14} – Special Educational Needs (SEN)	0.267 (0.011)	0.267 (0.011)
β_{15} – Free school meal (FSM)	0.377 (0.008)	0.372 (0.013)

σ^2_{u0} – School intercept variance	0.103 (0.007)	0.116 (0.009)
σ^2_{u1} – School FSM variance		0.035 (0.005)
σ^2_{u01} – School intercept-FSM covariance		-0.027 (0.005)
α – Overdispersion	0.782 (0.005)	0.775 (0.005)
	Marginal statistics	
Marginal expectation	8.50	8.52
Marginal variance	87.05	87.20
School (level 2) variance	8.71	9.04
Student (level 1) variance	78.34	78.17
School (level 2) VPC	0.10	0.10
Student (level 1) VPC	0.90	0.90
	Fit statistics	
Deviance	415438	415268

Note. Number of schools: $J = 434$; number of students: $N = 66{,}955$. Reference categories. Prior attainment: Quintile 1 (lowest prior attainment); Age: Summer born (youngest in year); Ethnicity: White. Standard errors in parentheses. Sample average values are reported for the marginal statistics as each statistic is a function of the covariates.

Figure 1. Distribution of number of days absent over the academic year.

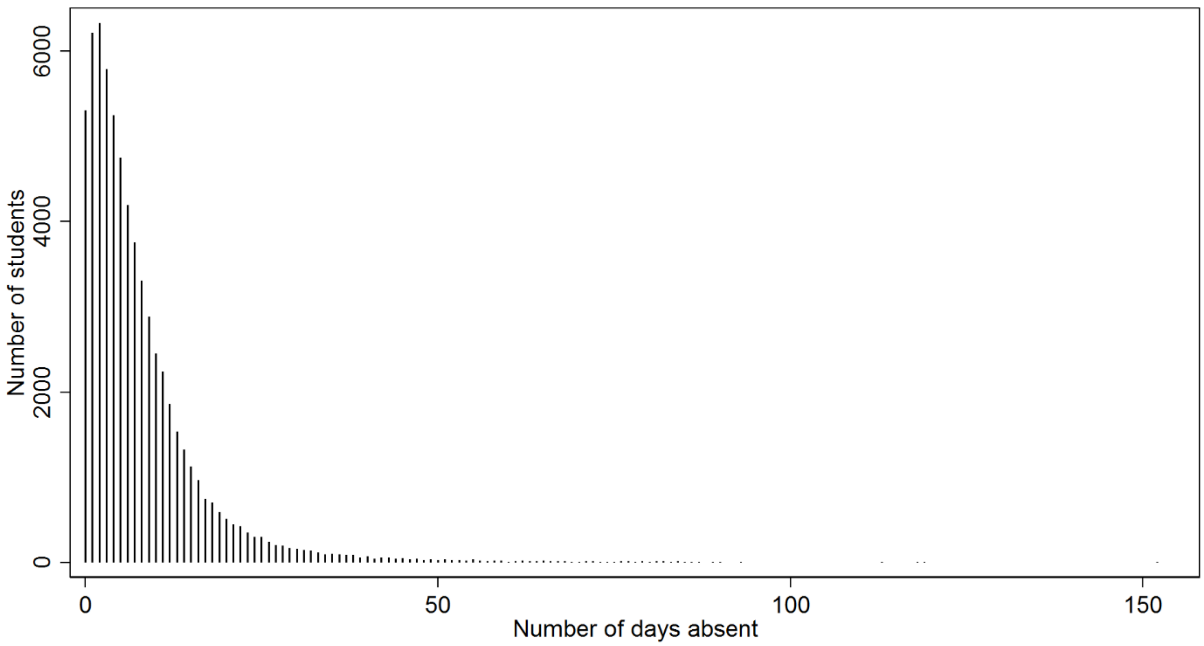


Figure 2. Mean number of days absent by district (left panel) and school (right panel). The horizontal line depicts the student sample mean.

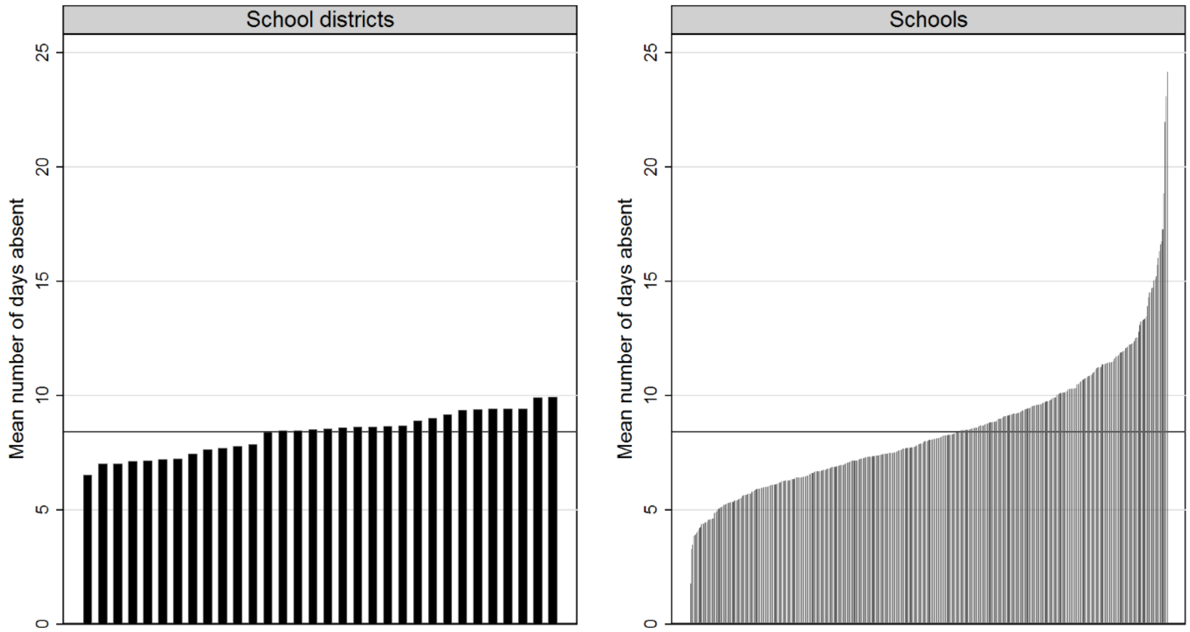


Figure 3. Relationship between the predicted school VPC and marginal expectation (top panel). Distribution of predicted student-level marginal expectation values (bottom panel). Plots are based on Model 3.

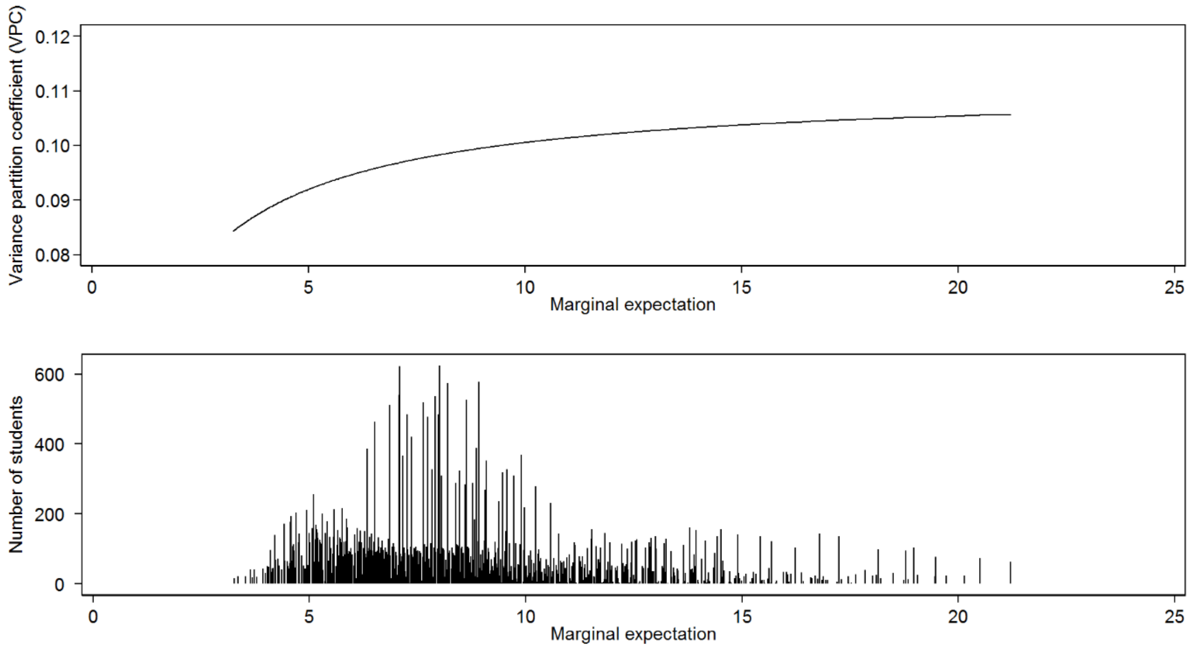


Figure 4. Scatterplots of predicted school effect values (left panel) and ranks (right panel) from unadjusted and covariate adjusted two-level random-intercept negative binomial models (Models 3 and 5).

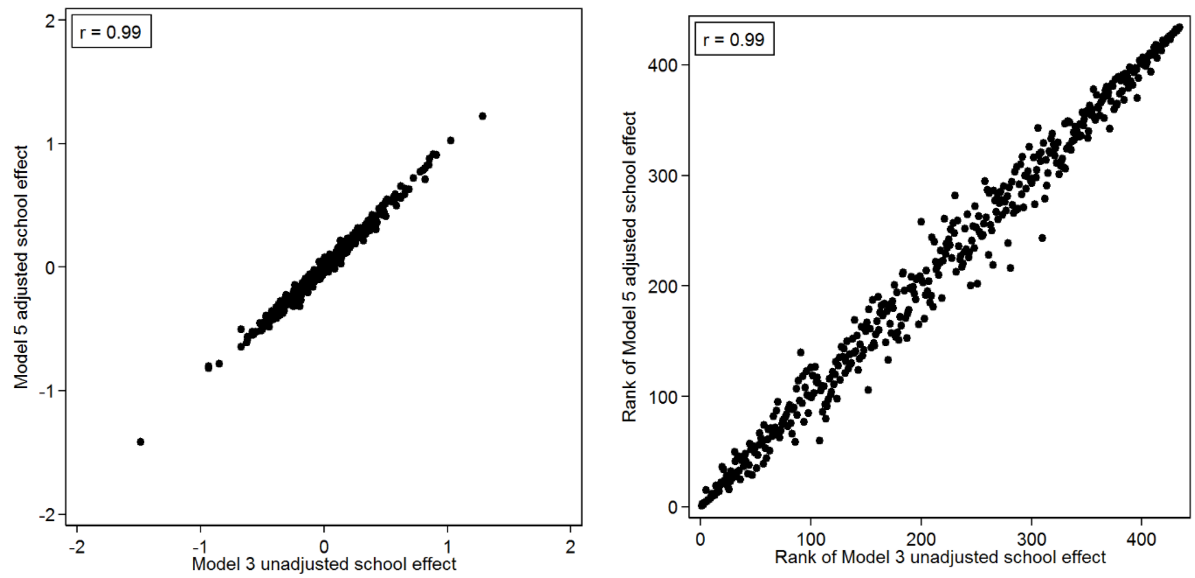
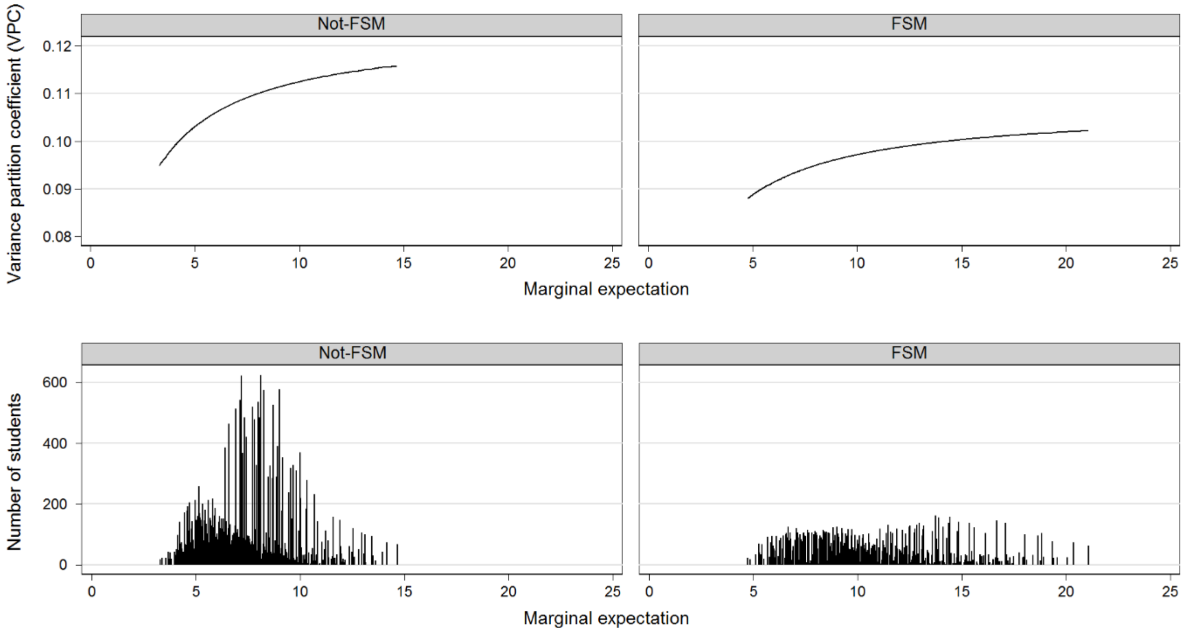


Figure 5. Relationship between the predicted school VPC and marginal expectation by student FSM status (top panel). Distribution of predicted student-level marginal expectation values by student FSM status (bottom panel). Plots are based on Model 6.



Supplemental materials

S1. Review of two-level random-intercept models for count data

In Section 2 of the article, we review the two most widely applied multilevel models for count responses, the Poisson model and the negative binomial model (mean dispersion or NB2 version). In this section, we review two further count models, the Poisson model with an overdispersion random effect and the constant dispersion or NB1 version of the negative binomial model. Table S3 presents a summary table allowing readers to compare the conditional expectations and variances across all four models. Table S2 presents the corresponding table for conventional single-level versions of these models.

S1.1 Poisson model with overdispersion random effect

The Poisson random-intercept model (Equation 1) includes a normally distributed cluster random intercept effect to account for the clustering in the data. A natural way to deal with overdispersion in this model is to therefore add a normally distributed unit-level overdispersion random effect to represent omitted unit-level variables that are envisaged to be driving any overdispersion. In contrast to the conventional cluster random intercept effect, this does not induce any dependence among the units. The new model can be written as

$$y_{ij}|\mu_{ij} \sim \text{Poisson}(\mu_{ij})$$

$$\ln(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_j + e_{ij} \qquad \text{(S1)}$$

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

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

where e_{ij} denotes the overdispersion random effect. Thus, in this model, two units with the same covariate and random intercept effect value may nonetheless differ in their expected counts μ_{ij} with such differences attributed to the two units differing in terms of their values on the omitted unit-level variables. This is also a feature of the two negative binomial models (Equation 4 and Equation S4). The overdispersion random effect is assumed normally distributed with zero mean and variance or overdispersion parameter σ_e^2. The larger σ_e^2 is, the greater the overdispersion. When $\sigma_e^2 = 0$, the model simplifies to the Poisson model (Equation 1) permitting a likelihood-ratio test for whether any estimated overdispersion is statistically significant.

In this model we can again calculate the conditional expectation and variance of the response. However, here we must integrate out the overdispersion random effect since this is not typically of substantive interest. To do this, we exploit the fact that μ_{ij}^C is log normally distributed and so its expectation and variance have known forms. Thus, the conditional expectation of y_{ij} (given $\mathbf{x}_{ij}$ and u_j but averaged over e_{ij}) is now given by

$$\mu_{ij}^C \equiv E\left(y_{ij} \middle| \mathbf{x}_{ij}, u_j\right) = \exp\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + u_j + \sigma_e^2/2\right) \tag{S2}$$

and we see that, in contrast to the Poisson model (Equation 2), $\mu_{ij}^C \neq \mu_{ij}$. The conditional variance is then given by

$$\omega_{ij}^C \equiv \mathrm{Var}\left(y_{ij} \middle| \mathbf{x}_{ij}, u_j\right) = \mu_{ij}^C + \left(\mu_{ij}^C\right)^2\{\exp(\sigma_e^2) - 1\} \tag{S3}$$

Thus, the conditional variance is now a quadratic function of the conditional expectation and is larger than the conditional expectation if $\sigma_e^2 > 0$. Therefore, the usual variance-mean relationship for the Poisson model (Equation 3) is relaxed, producing overdispersion with respect to the conditional expectation ($\omega_{ij}^C > \mu_{ij}^C$).

To help see the similarities between the mean dispersion negative binomial model (Equation 4) and the current model, recall that in the former we assume $\exp(e_{ij}) \sim \text{Gamma}(1/\alpha\,, \alpha)$ while in the latter we assume $e_{ij} \sim N(0, \sigma_e^2)$, implying $\exp(e_{ij}) \sim \text{logN}(0, \sigma_e^2)$. Thus, the two models differ only in the distribution they assume for the exponentiated overdispersion random effect (gamma vs. log-normal).

S1.2 Negative binomial model: Constant dispersion version

The constant dispersion version of the negative binomial model cannot be derived from a version of the Poisson model with a unit-level overdispersion random effect. The model is instead written as

$$y_{ij}|\mu_{ij} \sim \text{Poisson}(\mu_{ij})$$

$$\mu_{ij}|\mathbf{x}_{ij}, u_j \sim \text{Gamma}\left\{\frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j)}{\delta}, \delta\right\} \quad \text{(S4)}$$

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

where we now assume the expected count μ_{ij} has a conditional gamma distribution (given $\mathbf{x}_{ij}$ and u_j) with shape and scale parameters $\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j)\,\delta^{-1}$ and δ and therefore expectation $\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j)$ and variance $\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j)\,\delta$ where δ is the overdispersion parameter. The larger δ is, the greater the overdispersion. When $\delta = 0$, the variance of this gamma distribution is equal to zero and the model simplifies to the Poisson model (Equation 1), once again permitting a likelihood-ratio test for whether the estimated overdispersion is statistically significant.

It follows that the conditional expectation of y_{ij} (given $\mathbf{x}_{ij}$ and u_j) is again the same as in the Poisson model (Equation 2), with

$$\mu_{ij}^{C} \equiv E\left(y_{ij} \middle| \mathbf{x}_{ij}, u_j\right) = \exp\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + u_j\right) \tag{S5}$$

However, the conditional variance of y_{ij} is now

$$\omega_{ij}^{C} \equiv \operatorname{Var}\left(y_{ij} \middle| \mathbf{x}_{ij}, u_j\right) = \mu_{ij}^{C}(1 + \delta) \tag{S6}$$

Thus, in contrast to the quadratic form seen in the two unit-level random effect overdispersion models (Equation 6 and S3), the variance in the current model is a constant multiple of the conditional expectation.

To help see the similarities between the two versions of the negative binomial model, note that we can rewrite the mean dispersion model (Equation 4) as follows

$$\begin{gathered} y_{ij}|\mu_{ij} \sim \text{Poisson}\left(\mu_{ij}\right) \\ \mu_{ij}|\mathbf{x}_{ij}, u_j \sim \text{Gamma}\left\{\frac{1}{\alpha}, \alpha \exp\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + u_j\right)\right\} \\ u_j \sim N(0, \sigma_u^2) \end{gathered} \tag{S7}$$

Thus, both versions of the negative binomial model assume the expected count μ_{ij} follows a conditional gamma distribution, but the models differ in terms of the shape and scale parameters of this distribution. In the mean dispersion model (Equation 4), the shape and scale parameters are chosen such that the conditional expectation of μ_{ij} is $\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j)$ and conditional variance is $\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j)^2 \alpha$ whereas in the constant dispersion model (Equation S4) the shape and scale parameters are chosen such that the conditional expectation of μ_{ij} is $\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j)$ and conditional variance is $\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j)\,\delta$. Thus, the difference between the two versions of the negative binomial model can be viewed as a difference in the relationship between the conditional variance and conditional expectation of μ_{ij}. This difference then leads the models to differ in terms of the conditional variances of the response ω^C_{ij} (Equation 6 vs. S6).

S2. Derivation of the marginal statistics in two-level random-intercept models: Marginal expectation, variance, covariance, correlation, VPCs and ICC

We now present general derivations of the marginal expectation, variance, covariance and correlation for two-level random-intercept models (McCulloch & Searle, 2001). We then show that the ICC is simply the marginal correlation. We then decompose the marginal variance into components of variance at each level and show that the VPC can be calculated as the ratio of the level-2 component to the sum of the level-1 and level-2 components. We then apply all these results to each of the four count models considered in this article. Table S3 presents these quantities.

S2.1 Marginal expectation

The marginal expectation of y_{ij} (given $\mathbf{x}_{ij}$ but averaged over u_j) can be derived by exploiting the law of total expectations ($\mathrm{E}(A) = \mathrm{E}\{\mathrm{E}(A|B)\}$; law of iterated expectations)

$$\begin{aligned} \mu_{ij}^{M} \equiv \mathrm{E}(y_{ij}|\mathbf{x}_{ij}) &= \mathrm{E}\{\mathrm{E}(y_{ij}|\mathbf{x}_{ij}, u_j)|\mathbf{x}_{ij}\} \\ &= \mathrm{E}(\mu_{ij}^{C}|\mathbf{x}_{ij}) \end{aligned} \tag{S8}$$

In all four models, μ_{ij}^{C} follows the log normal distribution and so we exploit the closed form solution for the expectation of a log normally distribute variable.

S2.2 Marginal variance

The marginal variance of y_{ij} (given $\mathbf{x}_{ij}$ but averaged over u_j) can be derived by exploiting the law of total variance ($\mathrm{Var}(A) = \mathrm{Var}\{\mathrm{E}(A|B)\} + \mathrm{E}\{\mathrm{Var}(A|B)\}$; see also McCulloch and Searle, 2001, p.12)

$$\begin{aligned} \omega_{ij}^{M} \equiv \mathrm{Var}(y_{ij}|\mathbf{x}_{ij}) &= \mathrm{Var}\{\mathrm{E}(y_{ij}|\mathbf{x}_{ij}, u_j)|\mathbf{x}_{ij}\} + \mathrm{E}\{\mathrm{Var}(y_{ij}|\mathbf{x}_{ij}, u_j)|\mathbf{x}_{ij}\} \\ &= \underbrace{\mathrm{Var}(\mu_{ij}^{C}|\mathbf{x}_{ij})}_{\text{level}-2\text{ variance}} + \underbrace{\mathrm{E}(\omega_{ij}^{C}|\mathbf{x}_{ij})}_{\text{level}-1\text{ variance}} \end{aligned} \tag{S9}$$

It is instructive to realize that the law of total variance decomposes the marginal variance into separate level-2 and level-1 specific variance components. The level-2 variance component $\mathrm{Var}(\mu_{ij}^{C}|\mathbf{x}_{ij})$ captures between cluster variation in the cluster specific expected counts μ_{ij}^{C} attributable to the cluster random intercept effect u_j (i.e., given $\mathbf{x}_{ij}$). The level-1 variance component $\mathrm{E}(\omega_{ij}^{C}|\mathbf{x}_{ij})$ captures within cluster variation in y_{ij} around these cluster specific expected counts μ_{ij}^{C} (i.e., given $\mathbf{x}_{ij}$) averaged across all clusters (as ω_{ij}^{C} given $\mathbf{x}_{ij}$ still varies across clusters as function of u_j).

S2.3 Marginal covariance

To derive the marginal covariance of y_{ij} and $y_{i'j}$ (given $\mathbf{x}_{ij}$ and $\mathbf{x}_{i'j}$ but averaged over u_j) , we exploit the law of total covariance ($\mathrm{Cov}(A,B) = \mathrm{Cov}\{\mathrm{E}(A,B|C)\} + \mathrm{E}\{\mathrm{Cov}(A,B|C)\}$; see also McCulloch and Searle, 2001, p.12).

$$\begin{aligned}\mathrm{Cov}\left(y_{ij}, y_{i'j} \middle| \mathbf{x}_{ij}, \mathbf{x}_{i'j}\right) &= \mathrm{Cov}\left\{\mathrm{E}\left(y_{ij} \middle| \mathbf{x}_{ij}, u_j\right), \mathrm{E}\left(y_{i'j} \middle| \mathbf{x}_{i'j}, u_j\right) \middle| \mathbf{x}_{ij}, \mathbf{x}_{i'j}\right\} \\ &\quad + \mathrm{E}\left\{\mathrm{Cov}\left(y_{ij}, y_{i'j} \middle| \mathbf{x}_{ij}, \mathbf{x}_{i'j}, u_j\right) \middle| \mathbf{x}_{ij}, \mathbf{x}_{i'j}\right\} \\ &= \mathrm{Cov}\left(\mu_{ij}^{C}, \mu_{i'j}^{C} \middle| \mathbf{x}_{ij}, \mathbf{x}_{i'j}\right) + \mathrm{E}\left(0 \middle| \mathbf{x}_{ij}, \mathbf{x}_{i'j}\right) \qquad \text{(S10)}\end{aligned}$$

The second term equals zero due to the assumption of conditional independence among the responses for the same cluster given the covariates. Namely, y_{ij} and $y_{i'j}$ are assumed conditionally independent (given $\mathbf{x}_{ij}$, $\mathbf{x}_{i'j}$ and u_j) (i.e., assuming independent Poisson sampling variation).

S2.4 Marginal correlation

The marginal correlation or ICC can then be calculated in the usual way

$$\mathrm{Corr}\left(y_{ij}, y_{i'j} \middle| \mathbf{x}_{ij}, \mathbf{x}_{i'j}\right) = \frac{\mathrm{Cov}\left(y_{ij}, y_{i'j} \middle| \mathbf{x}_{ij}, \mathbf{x}_{i'j}\right)}{\sqrt{\mathrm{Var}\left(y_{ij} \middle| \mathbf{x}_{ij}\right)}\sqrt{\mathrm{Var}\left(y_{i'j} \middle| \mathbf{x}_{i'j}\right)}}$$

$$= \frac{\mathrm{Cov}\left(\mu_{ij}^{C}, \mu_{i'j}^{C} \middle| \mathbf{x}_{ij}, \mathbf{x}_{i'j}\right)}{\sqrt{\mathrm{Var}\left(\mu_{ij}^{C} \middle| \mathbf{x}_{ij}\right) + \mathrm{E}\left(\omega_{ij}^{C} \middle| \mathbf{x}_{ij}\right)}\sqrt{\mathrm{Var}\left(\mu_{i'j}^{C} \middle| \mathbf{x}_{i'j}\right) + \mathrm{E}\left(\omega_{i'j}^{C} \middle| \mathbf{x}_{i'j}\right)}} \tag{S11}$$

S2.5 Variance partition coefficients (VPCs)

Variance partition coefficients report the proportion of the total variation in the observed counts (given the covariates) which lies between clusters. This is equal to the proportion of marginal variance operating at each level of analysis. These can be calculated in the usual way, as ratios of each variance component to the marginal variance. The level-2 VPC is therefore calculated as

$$\mathrm{VPC}(2)_{ijk} = \frac{\overbrace{\mathrm{Var}\left(\mu_{ij}^{C} \middle| \mathbf{x}_{ij}\right)}^{\text{level-2 variance}}}{\underbrace{\mathrm{Var}\left(\mu_{ij}^{C} \middle| \mathbf{x}_{ij}\right)}_{\text{level-2 variance}} + \underbrace{\mathrm{E}\left(\omega_{ij}^{C} \middle| \mathbf{x}_{ij}\right)}_{\text{level-1 variance}}} \tag{S12}$$

The level-1 VPC is calculated as

$$\mathrm{VPC}(1)_{ijk} = \frac{\overbrace{\mathrm{E}\left(\omega_{ij}^{C} \middle| \mathbf{x}_{ij}\right)}^{\text{level-1 variance}}}{\underbrace{\mathrm{Var}\left(\mu_{ij}^{C} \middle| \mathbf{x}_{ij}\right)}_{\text{level-2 variance}} + \underbrace{\mathrm{E}\left(\omega_{ij}^{C} \middle| \mathbf{x}_{ij}\right)}_{\text{level-1 variance}}} \tag{S13}$$

S2.6 Intraclass correlation coefficients (ICCs)

The ICC is simply the marginal correlation. We focus on the special case where the two units have the same covariate values $\mathbf{x}_{ij} = \mathbf{x}_{i'j}$ in which case $\mu_{ij}^{C} = \mu_{i'j}^{C}$ and $\omega_{ij}^{C} = \omega_{i'j}^{C}$ and so the ICC simplifies to

$$\text{Corr}\left(y_{ij}, y_{i'j} \middle| \mathbf{x}_{ij}\right) = \frac{\overbrace{\text{Var}\left(\mu_{ij}^{C} \middle| \mathbf{x}_{ij}\right)}^{\text{level-2 variance}}}{\underbrace{\text{Var}\left(\mu_{ij}^{C} \middle| \mathbf{x}_{ij}\right)}_{\text{level-2 variance}} + \underbrace{\text{E}\left(\omega_{ij}^{C} \middle| \mathbf{x}_{ij}\right)}_{\text{level-1 variance}}} \quad \text{(S14)}$$

The expression for the ICC is the ratio of the level-2 variance component defined earlier to the total or marginal variance. This is also the definition of the VPC (Equation S12).

S2.7 VPCs for specific count models

For each model, Table S3 presents the level-1 and -2 components of the marginal variance used in the calculation of the VPC. The VPC (and ICC) for the Poisson model and the mean dispersion version of the negative binomial model are presented in Equations 11 and 16. The VPC (and ICC) for the Poisson model with overdispersion random effect is given by

$$\text{VPC}_{ij} = \frac{\overbrace{\left(\mu_{ij}^{M}\right)^{2}\{\exp(\sigma_{u}^{2})-1\}}^{\text{level-2 variance}}}{\underbrace{\left(\mu_{ij}^{M}\right)^{2}\{\exp(\sigma_{u}^{2})-1\}}_{\text{level-2 variance}} + \underbrace{\mu_{ij}^{M} + \left(\mu_{ij}^{M}\right)^{2}\exp(\sigma_{u}^{2})\{\exp(\sigma_{e}^{2})-1\}}_{\text{level-1 variance}}} \quad \text{(S15)}$$

The VPC (and ICC) for the constant dispersion version of the negative binomial model is given by

$$\text{VPC}_{ij} = \frac{\overbrace{\left(\mu_{ij}^{M}\right)^2\{\exp(\sigma_u^2)-1\}}^{\text{level}-2\text{ variance}}}{\underbrace{\left(\mu_{ij}^{M}\right)^2\{\exp(\sigma_u^2)-1\}}_{\text{level}-2\text{ variance}}+\underbrace{\mu_{ij}^{M}(1+\delta)}_{\text{level}-1\text{ variance}}} \tag{S16}$$

Comparing all four VPC expressions (Equations 11, 16, S15, S16), we see that the expression for the level-2 component of the marginal variance is always the same and so it is only the expression for the level-1 component which varies across models. This makes sense as the models differ only in their treatment of overdispersion and this is viewed as a level-1 phenomenon. The overdispersion parameters in the three overdispersion models lead the expressions for the level-1 variance to exceed that of the Poisson model. The marginal variance is simply the summation of the level-2 and -1 variances and so is also expected to be higher in the three models which allow for overdispersion compared to that of the Poisson model. It can be shown that as with the mean dispersion negative binomial model, the VPC for these three models is an increasing function of μ_{ij}^{M} and σ_u^2, but a decreasing function of the overdispersion parameter, where this has been included (σ_e^2 or δ).

S3. Derivation of the marginal statistics in three-level random-intercept models: Marginal expectation, variance, covariance, correlation, VPCs and ICCs

We now present general derivations of the marginal expectation, variance, covariance and correlation for three-level random-intercept models. We then decompose the marginal variance into components of variance at each level and show that the different VPCs can be calculated as ratios of the level-specific variance components. We then apply all these results to each of the four count models considered in this article. Table S4 presents these quantities.

S3.1 Marginal expectation

The marginal expectation of y_{ijk} (given $\mathbf{x}_{ijk}$ but averaged over v_k and u_{jk}) can be derived by exploiting the law of total expectations ($\mathrm{E}(A) = \mathrm{E}\{\mathrm{E}(A|B)\}$; law of iterated expectations)

$$\mu_{ijk}^{M} \equiv \mathrm{E}\left(y_{ijk}\middle|\mathbf{x}_{ij}\right) = \mathrm{E}\left\{\mathrm{E}\left(y_{ijk}\middle|\mathbf{x}_{ijk}, v_k, u_{jk}\right)\middle|\mathbf{x}_{ijk}\right\}$$
$$= \mathrm{E}\left(\mu_{ijk}^{C2}\middle|\mathbf{x}_{ij}\right) \qquad \text{(S17)}$$

In all four models μ_{ij}^{C2} follows the log normal distribution and so we exploit the closed form solution for the expectation of a log normally distribute variable.

S3.2 Marginal variance

The marginal variance of y_{ijk} (given $\mathbf{x}_{ijk}$ but averaged over v_k and u_{jk}) can be derived by repetitively exploiting the law of total variance ($\mathrm{Var}(A) = \mathrm{Var}\{\mathrm{E}(A|B)\} + \mathrm{E}\{\mathrm{Var}(A|B)\}$

First decompose the marginal variance of y_{ijk} (given $\mathbf{x}_{ijk}$ but averaged over v_k and u_{jk}) into a between supercluster and within supercluster components

$$\underbrace{\mathrm{Var}\left(y_{ijk}\middle|\mathbf{x}_{ijk}\right)}_{\text{Total variance}} = \underbrace{\mathrm{Var}\left\{\mathrm{E}\left(y_{ijk}\middle|\mathbf{x}_{ijk}, v_k\right)\middle|\mathbf{x}_{ijk}\right\}}_{\text{level}-3\text{ variance}} + \underbrace{\mathrm{E}\left\{\mathrm{Var}\left(y_{ijk}\middle|\mathbf{x}_{ijk}, v_k\right)\middle|\mathbf{x}_{ijk}\right\}}_{\text{level}-2\text{ and}-1\text{ combined variance}} \qquad \text{(S18)}$$

Next decompose the conditional variance of y_{ijk} (given $\mathbf{x}_{ijk}$ and v_k but averaged over u_{jk}) into a between cluster and within cluster component (i.e., decompose the contents of the expectation in the second term of the above expressions).

$$\underbrace{\mathrm{Var}\left(y_{ijk}\middle|\mathbf{x}_{ijk}, v_k\right)}_{\text{level}-2\text{ and}-1\text{ combined variance in supercluster }k} = \underbrace{\mathrm{Var}\left\{\mathrm{E}\left(y_{ijk}\middle|\mathbf{x}_{ijk}, v_k, u_{jk}\right)\middle|\mathbf{x}_{ijk}\right\}}_{\text{level}-2\text{ variance in supercluster }k} + \underbrace{\mathrm{E}\left\{\mathrm{Var}\left(y_{ijk}\middle|\mathbf{x}_{ijk}, v_k, u_{jk}\right)\middle|\mathbf{x}_{ijk}\right\}}_{\text{level}-1\text{ variance in supercluster }k} \quad \text{(S19)}$$

Substitute Equation S19 into Equation S18 to give

$$\underbrace{\mathrm{Var}\left(y_{ijk}\middle|\mathbf{x}_{ijk}\right)}_{\text{Total variance}} = \underbrace{\mathrm{Var}\left\{\mathrm{E}\left(y_{ijk}\middle|\mathbf{x}_{ijk}, v_k\right)\middle|\mathbf{x}_{ijk}\right\}}_{\text{level}-3\text{ variance}} + \underbrace{\mathrm{E}\left\{\mathrm{Var}\left\{\mathrm{E}\left(y_{ijk}\middle|\mathbf{x}_{ijk}, v_k, u_{jk}\right)\middle|\mathbf{x}_{ijk}\right\}\middle|\mathbf{x}_{ijk}\right\}}_{\text{level}-2\text{ variance}} + \underbrace{\mathrm{E}\left\{\mathrm{Var}\left(y_{ijk}\middle|\mathbf{x}_{ijk}, v_k, u_{jk}\right)\middle|\mathbf{x}_{ijk}\right\}}_{\text{level}-1\text{ variance}} \quad \text{(S20)}$$

This expression can be written more concisely as

$$\underbrace{\mathrm{Var}\left(y_{ijk}\middle|\mathbf{x}_{ijk}\right)}_{\text{Total variance}} = \underbrace{\mathrm{Var}\left(\mu_{ijk}^{C3}\middle|\mathbf{x}_{ijk}\right)}_{\text{level}-3\text{ variance}} + \underbrace{\mathrm{E}\left\{\mathrm{Var}\left\{\mu_{ijk}^{C2}\middle|\mathbf{x}_{ijk}\right\}\middle|\mathbf{x}_{ijk}\right\}}_{\text{level}-2\text{ variance}} + \underbrace{\mathrm{E}\left(\omega_{ijk}^{C2}\middle|\mathbf{x}_{ijk}\right)}_{\text{level}-1\text{ variance}} \quad \text{(S21)}$$

The level-3 variance component $\mathrm{Var}\left(\mu_{ijk}^{C3}\middle|\mathbf{x}_{ij}\right)$ captures between supercluster variation in the supercluster specific expected counts μ_{ijk}^{C3} attributable to the supercluster random intercept effect v_k (i.e., given $\mathbf{x}_{ij}$). The level-2 variance component $\mathrm{E}\left\{\mathrm{Var}\left\{\mu_{ijk}^{C2}\middle|\mathbf{x}_{ijk}\right\}\middle|\mathbf{x}_{ijk}\right\}$ captures the within supercluster, between cluster variation in the cluster specific expected counts μ_{ijk}^{C2} attributable to the cluster random intercept effect u_{jk} (i.e., given $\mathbf{x}_{ij}$). The level-1 variance component $\mathrm{E}\left(\omega_{ijk}^{C2}\middle|\mathbf{x}_{ijk}\right)$ captures within cluster variation in y_{ij} around these cluster specific expected counts μ_{ijk}^{C2} (i.e., given $\mathbf{x}_{ij}$) averaged across all clusters (as ω_{ijk}^{C2} given $\mathbf{x}_{ij}$ still varies across clusters as function of v_k and u_{jk}).

S3.3 Variance partition coefficients (VPCs)

Variance partition coefficients report the proportion of the total variation in the observed counts (given the covariates) which lies between clusters. This is equal to the proportion of marginal variance operating at each level of analysis. These can be calculated in the usual way, as ratios of each variance component to the marginal variance. The level-3 VPC is therefore calculated as

$$\text{VPC}(3)_{ijk} = \frac{\overbrace{\text{Var}\left(\mu_{ijk}^{C3} \middle| \mathbf{x}_{ijk}\right)}^{\text{level−3 variance}}}{\underbrace{\text{Var}\left(\mu_{ijk}^{C3} \middle| \mathbf{x}_{ijk}\right)}_{\text{level−3 variance}} + \underbrace{\text{E}\left\{\text{Var}\left\{\mu_{ijk}^{C2} \middle| \mathbf{x}_{ijk}\right\} \middle| \mathbf{x}_{ijk}\right\}}_{\text{level−2 variance}} + \underbrace{\text{E}\left(\omega_{ijk}^{C2} \middle| \mathbf{x}_{ijk}\right)}_{\text{level−1 variance}}} \quad \text{(S22)}$$

The level-2 VPC is calculated as

$$\text{VPC}(3)_{ijk} = \frac{\overbrace{\text{E}\left\{\text{Var}\left\{\mu_{ijk}^{C2} \middle| \mathbf{x}_{ijk}\right\} \middle| \mathbf{x}_{ijk}\right\}}^{\text{level−2 variance}}}{\underbrace{\text{Var}\left(\mu_{ijk}^{C3} \middle| \mathbf{x}_{ijk}\right)}_{\text{level−3 variance}} + \underbrace{\text{E}\left\{\text{Var}\left\{\mu_{ijk}^{C2} \middle| \mathbf{x}_{ijk}\right\} \middle| \mathbf{x}_{ijk}\right\}}_{\text{level−2 variance}} + \underbrace{\text{E}\left(\omega_{ijk}^{C2} \middle| \mathbf{x}_{ijk}\right)}_{\text{level−1 variance}}} \quad \text{(S23)}$$

The level-1 VPC is calculated as

$$\text{VPC}(3)_{ijk} = \frac{\overbrace{\text{E}\left(\omega_{ijk}^{C2} \middle| \mathbf{x}_{ijk}\right)}^{\text{level−1 variance}}}{\underbrace{\text{Var}\left(\mu_{ijk}^{C3} \middle| \mathbf{x}_{ijk}\right)}_{\text{level−3 variance}} + \underbrace{\text{E}\left\{\text{Var}\left\{\mu_{ijk}^{C2} \middle| \mathbf{x}_{ijk}\right\} \middle| \mathbf{x}_{ijk}\right\}}_{\text{level−2 variance}} + \underbrace{\text{E}\left(\omega_{ijk}^{C2} \middle| \mathbf{x}_{ijk}\right)}_{\text{level−1 variance}}} \quad \text{(S24)}$$

S3.4 Variance partition coefficients (VPCs) for specific count models

For each model, Table S4 presents the level-1, -2 and -3 components of the marginal variance used in the calculation of the three VPCs. The VPCs for the mean dispersion version of the negative binomial model are presented in Equations 18, 19 and 20. The VPCs for each of the three remaining count models can be calculated by substituting the relevant expressions for the different components of the marginal variance in these equations.

S4. Derivation of the marginal expectation, variance, covariance and correlation in two- and three-level random-coefficient models

All expressions derived in Section S2 were for two-level random-intercept models. The corresponding expressions for two-level models with random coefficients are obtained by replacing u_j and σ_u^2 in all expressions with $\mathbf{z}'_{ij}\mathbf{u}_j$ and $\mathbf{z}'_{ij}\mathbf{\Omega}_\mathbf{u}\mathbf{z}_{ij}$.

All expressions derived in Section S3 were for three-level random-intercept models. The corresponding expressions for three-level models with random coefficients are obtained by replacing u_j and σ_u^2 in all expressions with $\mathbf{z}'_{\mathbf{u}ijk}\mathbf{u}_{jk}$ and $\mathbf{z}'_{\mathbf{u}ijk}\mathbf{\Omega}_\mathbf{u}\mathbf{z}_{\mathbf{u}ijk}$ and by replacing v_k and σ_v^2 in all expressions with $\mathbf{z}'_{\mathbf{u}ijk}\mathbf{v}_k$ and $\mathbf{z}'_{\mathbf{v}ijk}\mathbf{\Omega}_\mathbf{v}\mathbf{z}_{ijk}$.

S5. References

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York, USA: John Wiley & Since, Inc.

Table S1.

Covariate distributions, including mean days absent.

Student characteristic	N	Percent	Mean days absent
Prior attainment (quintile)			
1 (lowest prior attainment)	14366	21%	10.1
2	13288	20%	9.0
3	12703	19%	8.3
4	16000	24%	7.4
5 (highest prior attainment)	10598	16%	7.0
Age			
Summer (youngest in year)	17032	25%	8.1
Spring	16495	25%	8.2
Winter	16551	25%	8.5
Autumn (oldest in year)	16877	25%	8.8
Gender			
Male	33628	50%	8.0
Female	33327	50%	8.8
Ethnicity			
White	27803	41%	9.6
Mixed	6181	09%	9.6
Asian	13914	21%	7.1
Black	14409	22%	7.0
Other	4648	7%	7.9

Language			
English	40529	61%	9.2
Not English	26426	39%	7.2
SEN			
Not SEN	57157	85%	7.9
SEN	9798	15%	11.4
FSM			
Not FSM	41606	62%	7.3
FSM	25349	38%	10.2

Note.

Number of school districts: $K = 32$; number of schools: $J = 434$; number of students: $N = 66{,}955$.

Prior attainment quintiles are based on an average test score across separate tests in English and maths taken five years earlier at the end of primary schooling, just before the start of secondary schooling.

Table S2.

Single-level count response models: Expressions for the marginal expectation and variance of the response.

Description	Notation	Poisson model	Poisson model with overdispersion random effect	Negative binomial model: Mean dispersion or NB2 version	Negative binomial model: Constant dispersion or NB1 version
Marginal expectation	$\mu_{ij}^{M} \equiv \mathrm{E}(y_{ij}\|\mathbf{x}_{ij})$	$\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})$	$\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma_e^2/2)$	$\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})$	$\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})$
Marginal variance	$\omega_{ij}^{M} \equiv \mathrm{Var}(y_{ij}\|\mathbf{x}_{ij})$	μ_{ij}^{M}	$\mu_{ij}^{M} + (\mu_{ij}^{M})^2\{\exp(\sigma_e^2) - 1\}$	$\mu_{ij}^{M} + (\mu_{ij}^{M})^2\alpha$	$\mu_{ij}^{M}(1 + \delta)$

Table S3.

Two-level count response models: Expressions for the conditional and marginal expectations and variances of the response and the level-1 and -2 components of the marginal variance used in the calculation of the VPCs.

Description	Notation	Poisson model	Poisson model with overdispersion random effect	Negative binomial model: Mean dispersion or NB2 version	Negative binomial model: Constant dispersion or NB1 version
Conditional expectation	$\mu_{ij}^{C} \equiv \mathrm{E}(y_{ij} \mid \mathbf{x}_{ij}, u_j)$	$\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j)$	$\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j + \sigma_e^2/2)$	$\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j)$	$\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_j)$
Conditional variance	$\omega_{ij}^{C} \equiv \mathrm{Var}(y_{ij} \mid \mathbf{x}_{ij}, u_j)$	μ_{ij}^{C}	$\mu_{ij}^{C} + (\mu_{ij}^{C})^2\{\exp(\sigma_e^2) - 1\}$	$\mu_{ij}^{C} + (\mu_{ij}^{C})^2\alpha$	$\mu_{ij}^{C}(1+\delta)$
Marginal expectation	$\mu_{ij}^{M} \equiv \mathrm{E}(y_{ij} \mid \mathbf{x}_{ij})$	$\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma_u^2/2)$	$\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma_u^2/2 + \sigma_e^2/2)$	$\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma_u^2/2)$	$\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma_u^2/2)$
Marginal variance	$\omega_{ij}^{M} \equiv \mathrm{Var}(y_{ij} \mid \mathbf{x}_{ij})$	$\mu_{ij}^{M} + (\mu_{ij}^{M})^2\{\exp(\sigma_u^2) - 1\}$	$\mu_{ij}^{M} + (\mu_{ij}^{M})^2\{\exp(\sigma_u^2)\exp(\sigma_e^2) - 1\}$	$\mu_{ij}^{M} + (\mu_{ij}^{M})^2\{\exp(\sigma_u^2)(1+\alpha) - 1\}$	$\mu_{ij}^{M}(1+\delta) + (\mu_{ij}^{M})^2\{\exp(\sigma_u^2) - 1\}$
Level-2 variance	$\mathrm{Var}(\mu_{ij}^{C} \mid \mathbf{x}_{ij}) \equiv \mathrm{Var}\{\mathrm{E}(y_{ij} \mid \mathbf{x}_{ij}, u_j) \mid \mathbf{x}_{ij}\}$	$(\mu_{ij}^{M})^2\{\exp(\sigma_u^2) - 1\}$	$(\mu_{ij}^{M})^2\{\exp(\sigma_u^2) - 1\}$	$(\mu_{ij}^{M})^2\{\exp(\sigma_u^2) - 1\}$	$(\mu_{ij}^{M})^2\{\exp(\sigma_u^2) - 1\}$
Level-1 variance	$\mathrm{E}(\omega_{ij}^{C} \mid \mathbf{x}_{ij}) \equiv \mathrm{E}\{\mathrm{Var}(y_{ij} \mid \mathbf{x}_{ij}, u_j) \mid \mathbf{x}_{ij}\}$	μ_{ij}^{M}	$\mu_{ij}^{M} + (\mu_{ij}^{M})^2 \exp(\sigma_u^2)\{\exp(\sigma_e^2) - 1\}$	$\mu_{ij}^{M} + (\mu_{ij}^{M})^2 \exp(\sigma_u^2)\,\alpha$	$\mu_{ij}^{M}(1+\delta)$

Note.

The above expressions are for two-level random-intercept models. The corresponding expressions for the two-level models with random coefficients are obtained by replacing u_j and σ_u^2 in all expressions with $\mathbf{z}'_{ij}\mathbf{u}_j$ and $\mathbf{z}'_{ij}\boldsymbol{\Omega}_{\mathbf{u}}\mathbf{z}_{ij}$.

Table S4.

Three-level count response models: Expressions for the conditional and marginal expectations and variances of the response and the level-1, -2 and -3 components of the marginal variance used in the calculation of the VPCs.

Description	Notation	Poisson model	Poisson model with overdispersion random effect	Negative binomial model: Mean dispersion or NB2 version	Negative binomial model: Constant dispersion or NB1 version
Conditional expectation (level-2)	$\mu_{ijk}^{C2} \equiv \mathrm{E}\left(y_{ijk} \mid \mathbf{x}_{ijk}, v_k, u_{jk}\right)$	$\exp\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + v_k + u_{jk}\right)$	$\exp\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + v_k + u_{jk} + \sigma_e^2/2\right)$	$\exp\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + v_k + u_{jk}\right)$	$\exp\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + v_k + u_{jk}\right)$
Conditional variance (level-2)	$\omega_{ijk}^{C2} \equiv \mathrm{Var}\left(y_{ijk} \mid \mathbf{x}_{ijk}, v_k, u_{jk}\right)$	μ_{ijk}^{C2}	$\mu_{ijk}^{C2} + \left(\mu_{ijk}^{C2}\right)^2\{\exp(\sigma_e^2) - 1\}$	$\mu_{ijk}^{C2} + \left(\mu_{ijk}^{C2}\right)^2\alpha$	$\mu_{ijk}^{C2}(1+\delta)$
Conditional expectation (level-3)	$\mu_{ijk}^{C3} \equiv \mathrm{E}\left(y_{ijk} \mid \mathbf{x}_{ijk}, v_k\right)$	$\exp\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + v_k + \sigma_u^2/2\right)$	$\exp\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + v_k + \sigma_u^2/2 + \sigma_e^2/2\right)$	$\exp\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + v_k + \sigma_u^2/2\right)$	$\exp\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + v_k + \sigma_u^2/2\right)$
Conditional variance (level-3)	$\omega_{ijk}^{C3} \equiv \mathrm{Var}\left(y_{ijk} \mid \mathbf{x}_{ijk}, v_k\right)$	$\mu_{ijk}^{C3} + \left(\mu_{ijk}^{C3}\right)^2\{\exp(\sigma_u^2) - 1\}$	$\mu_{ijk}^{C3} + \left(\mu_{ijk}^{C3}\right)^2\{\exp(\sigma_u^2 + \sigma_e^2) - 1\}$	$\mu_{ijk}^{C3} + \left(\mu_{ijk}^{C3}\right)^2\{\exp(\sigma_u^2)(1+\alpha) - 1\}$	$\mu_{ijk}^{C3}(1+\delta) + \left(\mu_{ijk}^{C3}\right)^2\{\exp(\sigma_u^2) - 1\}$
Marginal expectation	$\mu_{ijk}^{M} \equiv \mathrm{E}\left(y_{ijk} \mid \mathbf{x}_{ijk}\right)$	$\exp\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + \sigma_v^2/2 + \sigma_u^2/2\right)$	$\exp\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + \sigma_v^2/2 + \sigma_u^2/2 + \sigma_e^2/2\right)$	$\exp\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + \sigma_v^2/2 + \sigma_u^2/2\right)$	$\exp\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + \sigma_v^2/2 + \sigma_u^2/2\right)$
Marginal variance	$\omega_{ijk}^{M} \equiv \mathrm{Var}\left(y_{ijk} \mid \mathbf{x}_{ijk}\right)$	$\mu_{ijk}^{M} + \left(\mu_{ijk}^{M}\right)^2\{\exp(\sigma_v^2 + \sigma_u^2) - 1\}$	$\mu_{ijk}^{M} + \left(\mu_{ijk}^{M}\right)^2\{\exp(\sigma_v^2 + \sigma_u^2 + \sigma_e^2) - 1\}$	$\mu_{ijk}^{M} + \left(\mu_{ijk}^{M}\right)^2\{\exp(\sigma_v^2 + \sigma_u^2)(1+\alpha) - 1\}$	$(1+\delta)\mu_{ijk}^{M} + \left(\mu_{ijk}^{M}\right)^2\{\exp(\sigma_v^2 + \sigma_u^2) - 1\}$
Level-3 variance	$\mathrm{Var}\{\mu_{ijk}^{C3} \equiv \mid \mathbf{x}_{ijk}\}$	$\left(\mu_{ijk}^{M}\right)^2\{\exp(\sigma_v^2) - 1\}$	$\left(\mu_{ijk}^{M}\right)^2\{\exp(\sigma_v^2) - 1\}$	$\left(\mu_{ijk}^{M}\right)^2\{\exp(\sigma_v^2) - 1\}$	$\left(\mu_{ijk}^{M}\right)^2\{\exp(\sigma_v^2) - 1\}$
Level-2 variance	$\mathrm{E}\{\mathrm{Var}\left(\mu_{ijk}^{C23} \mid \mathbf{x}_{ijk}, v_k\right) \mid \mathbf{x}_{ijk}\}$	$\left(\mu_{ijk}^{M}\right)^2 \exp(\sigma_v^2)\{\exp(\sigma_u^2) - 1\}$	$\left(\mu_{ijk}^{M}\right)^2 \exp(\sigma_v^2)\{\exp(\sigma_u^2) - 1\}$	$\left(\mu_{ijk}^{M}\right)^2 \exp(\sigma_v^2)\{\exp(\sigma_u^2) - 1\}$	$\left(\mu_{ijk}^{M}\right)^2 \exp(\sigma_v^2)\{\exp(\sigma_u^2) - 1\}$
Level-1 variance	$\mathrm{E}\{\omega_{ijk}^{C23} \mid \mathbf{x}_{ijk}\}$	μ_{ijk}^{M}	$\mu_{ijk}^{M} + \left(\mu_{ijk}^{M}\right)^2 \exp(\sigma_v^2 + \sigma_u^2)\{\exp(\sigma_e^2) - 1\}$	$\mu_{ijk}^{M} + \left(\mu_{ijk}^{M}\right)^2 \exp(\sigma_v^2 + \sigma_u^2)\,\alpha$	$\mu_{ijk}^{M}(1+\delta)$

Note.

The above expressions are for three-level random-intercept models. The corresponding expressions for the three-level models with random coefficients are obtained by replacing u_j and σ_u^2 in all expressions with $\mathbf{z}'_{\mathbf{u}ijk}\mathbf{u}_{jk}$ and $\mathbf{z}'_{\mathbf{u}ijk}\boldsymbol{\Omega}_{\mathbf{u}}\mathbf{z}_{\mathbf{u}ijk}$ and by replacing v_k and σ_v^2 in all expressions with $\mathbf{z}'_{\mathbf{u}ijk}\mathbf{v}_k$ and $\mathbf{z}'_{\mathbf{v}ijk}\boldsymbol{\Omega}_{\mathbf{v}}\mathbf{z}_{ijk}$.