Title: Modelling correlated data: multilevel models and generalized estimating equations and their use with data from research in developmental disabilities.

Authors:

Dimitrios Vagenas[1] and Vasiliki Totsika[2]

[1] Institute of Health and Biomedical Innovation, Queensland University of Technology, Australia

[2] Centre for Educational Development, Appraisal, and Research (CEDAR) and Centre for Education Studies (CES), University of Warwick, UK; Centre for Developmental Psychiatry and Psychology, Department of Psychiatry, Monash University, Australia

Corresponding Author: Dimitrios Vagenas

Correspondence address: dimitrios.vagenas@qut.edu.au

Running title: MLMs and GEEs in IDD research

Keywords: clustered, longitudinal, multilevel, Generalized Estimating Equations, repeated, developmental disability

Abstract

Background: The use of Multilevel Models (MLM) and Generalized Estimating Equations (GEE) for analysing clustered data in the field of intellectual and developmental disability (IDD) research is still limited.

Method: We present some important features of MLMs and GEEs: main function, assumptions, model specification and estimators, sample size and power. We provide an overview of the ways MLMs and GEEs have been used in IDD research.

Results: While MLMs and GEEs are both appropriate for longitudinal and/or clustered data, they differ in the assumptions they impose on the data, and the inferences made. Estimators in MLMs require appropriate model specification, while GEEs are more resilient to misspecification at the expense of model complexity. Studies on sample size seem to suggest that Level 1 coefficients are robust to small samples/clusters, with any higher-level coefficients less so. MLMs have been used more frequently than GEEs in IDD research, especially for fitting developmental trajectories.

Conclusions: Clustered data from research in the IDD field can be analysed flexibly using MLMs and GEEs. These models would be more widely used if journals required the inclusion of technical specification detail, simulation studies examined power for IDD study characteristics, and researchers developed core skills during basic studies.

What this paper adds?

Research data cease to be independent when a super-ordinate structure or repeated measurements create correlation amongst individual data points. Ignoring this correlation in model specification leads to a bias in standard errors that is proportionate to the magnitude and direction of the correlation. Multilevel models (MLM) and Generalized Estimating Equations (GEEs) model the data taking into account this correlation. The two approaches differ in the way they handle this correlation, and selecting between the two relies on the research aims and study characteristics. The paper discusses some of the core features of MLMs and GEEs for researchers who are considering how to analyse their longitudinal or clustered data. We review how IDD researchers have used the models so far, in the hope that other researchers will consider using them. We believe their use would be more widespread if researchers were taught these models as part of their studies, if journals required researchers to include more technical details on how MLM or GEE models were

fitted, and if further research focused on examining how powerful these models could be for IDD studies that often rely on modest sample sizes, few clusters or large cluster to participant ratios.

- Ignoring data dependence can lead to inappropriate inferences and conclusions
- MLMs and GEEs are appropriate for clustered/longitudinal data
- IDD research could benefit from increased use of these models

1.1 Introduction

Research in the field of intellectual and developmental disabilities (IDD) often generates longitudinal and/or correlated data. One source of correlation comes from clustering such as data from mothers and children who share the same household; data from parents (couples) of children with IDD. Another often encountered source of correlation is repeated measurements obtained from a group of participants over time.

Researchers have three options when they analyse correlated data (e.g. longitudinal data on the same individuals, or data clustered within a hyper-ordinate structure): (1) ignore the correlation, (2) bypass it by withholding one part of the data, or (3) deal with the correlation using appropriate analytic techniques. The first two approaches are not really efficient or appropriate since they (1) either result in inappropriate inferences or (2) do not make full use of the data. Statistical expertise to deal with clustered and/or longitudinal data is required at an advanced level, one which often exceeds the training researchers in the IDD field may have. In addition, IDD research presents some unique challenges: (a) low prevalence of condition examined (typically resulting in small sample size), (b) often a high number of super-ordinate clusters (for example, a relatively small number of children within a large number of genetic syndromes). In addition, where research is applied and the focus is on informing educational or social policy, there is often an interest in drawing conclusions about the population with a particular need, and not about the way individuals' diagnostic labels/ clinical services/educational settings are clustered. In other words, the clustering is not always part of the substantive research question. This often results in a higher than expected frequency of the first two options, i.e., either ignoring the clustering or bypassing it by not using part of the data.

With the present paper, we would like to encourage IDD researchers to use appropriate modelling techniques when having correlated data. We focus on two analytic techniques: Multilevel Models (MLM; also known as Mixed Models proposed by Laird and Ware 1982) and Generalized Estimating Equations (GEE); proposed by Zeger and Liang, 1986). Using non-technical terms, we will discuss: (i) why it is important to account for the correlation and (ii) what MLMs and GEEs do and how they could be used in research in IDD drawing on examples from published research in the field.

1.2 Generalized Linear Models (GLM)

The standard ANOVA and regression models are part of a bigger family called the Generalized Linear Models (GLMs). GLMs assume that one variable – referred to as the outcome or dependent variable - is explained by or depends on some other variables called the explanatory or independent variables. The outcome and explanatory variables are assumed to be related ("linked") with a function called the "link function". One of the purposes of a GLM is to estimate a unique combination of the explanatory variables which explain as much variation of the outcome variable as possible. This is done by estimating different weights called "regression coefficients". Up to this point the whole process is a mathematical process referred to as optimization, as its aim is to minimize the unexplained variance of the outcome (i.e. the amount of variation not accounted for by the explanatory variables). This is usually done by using the so called "least squares" optimization method. Additionally, we want to know how certain we are about these estimated coefficients given that their estimation was based on a sample rather than a population. This is achieved by estimating confidence intervals and the associated *p* values (i.e., hypothesis testing). The usual test checks if the coefficients are significantly different from zero; for this we rely on distributional assumptions, and the issue now becomes a statistical one.

1.3 Assumptions of GLMs

A standard GLM assumes: (i) a distributional assumption (i.e. the distribution of residuals of the regression has a particular shape), needed for estimating confidence intervals (ii) a link function which connects the outcome with the explanatory variables (iii) constant variance, so that the inferences we make are valid for all the range of the dependent and independent variables and (iv) independence of the individual measurements. This last assumption is violated when analyzing longitudinal and clustered data in standard ANOVAs or regressions. We demonstrate the effect of this violation with an example in 1.3.4

The distributional assumption is used for mathematical calculations and approximations during the optimization procedure. Importantly, it is also used for making inferences and especially for estimating the confidence intervals. The confidence intervals are used to determine if an estimated coefficient is different from zero (hence statistically significant) or not.

1.3.1 Linear relationship

In every GLM we assume a mathematical relationship between the dependent variables and the independent variables. In a linear regression, we assume that the link function takes the form f(y)=y, and thus in most cases it is not stated. On the other hand, in Poisson regression for example, we assume the log() link function and thus f(y)=log(y) and the Poisson regression can be written as log(y)=a+bx+e. The equivalent form of the normal linear regression is Y=a+bx+e.

1.3.2 Homoscedasticity

Homoscedasticity is a term of Greek origin and means "equal variance". This is more of a practical assumption since it allows us to use the same standard error for the range of the values we have available. Otherwise, we would have to estimate a separate standard error for each value.

1.3.3 Independence

For estimating standard errors of regression coefficients we need two components: (i) a properly estimated residual variance and (ii) the number of observations that this residual variance is based upon, the so called degrees of freedom. The estimation of the residual variance is where the main estimation bias occurs when we incorrectly use a GLM for analysing clustered or longitudinal data.

1.3.4 Violation of independence with clustered and longitudinal data

Let us assume a simple longitudinal design with two time points. Using data from Totsika, Hastings, Vagenas & Emerson (2014), where child behaviour problems were measured in 516 children with intellectual disability at three (Time 1) and five years (Time 2), we will estimate the standard error of the mean difference, given different levels of correlation between the repeated measures. For simplicity, we assume that scores were normally distributed. Mean total behaviour problems at Time 1 were 16.18 and 12.27 at Time 2. In statistical notation this is represented as: $\widehat{\mu_1} = 16.18$ and $\widehat{\mu_2} = 12.27$, where the hat stands for estimated, μ stands for mean and the subscript 1 and 2 denotes the times of measurement. Equivalently, the standard deviations can be denoted as $\widehat{\sigma_1} = 6.32$ and $\widehat{\sigma_2} = 8.11$. The mean difference between the two time points is $\widehat{\mu_2} - \widehat{\mu_1} = 12.27 - 16.18 = -3.91$ units. Our objective is to determine if this difference is statistically significantly different from zero on the 5% level of significance, and for this we need to estimate the standard error of the mean difference. If we assume that there is no correlation between the two measurements the appropriate formula to determine the standard error of the mean difference is:

$$s.e. = \sqrt{(\widehat{\sigma_1} + \widehat{\sigma_2})/n} \quad (1)$$

However, if we assume that the two measurements are related we have to take into account the covariance of the two means, symbolized as σ_{12} and estimated by:

$$s.e. = \sqrt{(\widehat{\sigma_1} + \widehat{\sigma_2} - 2 \times \sigma_{12})/n} \qquad (2)$$

The effect of the omission of the covariance term depends on the relative size of the covariance compared to the sum of the variances. For demonstration purposes, we plot the relationship between the covariance term and the standard error of the mean difference, for different levels of correlation using the Totsika et al (2014) data (see Figure 1).

----------Please insert Figure 1 here----------

As it can be seen in Figure 1, ignoring the correlation between the two measurements will lead to biased inferences about the standard error of the mean difference. The level of bias depends on the strength of the correlation. For behaviour problems, we have good evidence to indicate that over-time correlations tend to be at least moderate both in children and adults with IDD (Emerson et al., 2014; Totsika, Toogood, Hastings & Lewis, 2008; Totsika et al., 2013; 2014). As Figure 1 suggests, the presence of positively correlated measurements is associated with smaller standard errors compared to non-correlated measures (i.e., a correlation coefficient of zero), whereas a negative correlation is associated with higher standard errors than would have been estimated under the assumption of independence. Thus, if we were to test if a particular difference between two time points was different from zero, a positive correlation between the two time points would result in a tighter confidence interval compared to zero correlation, whereas a negative correlation would result in a wider confidence interval. As a corollary, we would need a smaller sample size for finding a statistically significant difference when we have positive correlation compared to a negative one when the difference if of similar magnitude. As an example, using the above numbers of Totsika et al. (2014) and assuming a correlation of -0.5 and 0.5 would have led to a standard error that would have been obtained if they had 67% and 190% of the participants (n)

respectively, compared to that obtained if there was no correlation between the two time points.

2.1 Multilevel Models and Generalized Estimating Equations

There are several approaches for accounting for the correlation between measurements in longitudinal and clustered data. Repeated measures ANOVA is one of the better known approaches. Whilst it is part of the GLM family, it has some drawbacks which make it unsuitable for many studies (Fitzmaurice et al., 2004; pp.73-78). Repeated measures ANOVA assumes a so called compound symmetry correlation structure for the data. This means that each measurement time/cluster unit has the same correlation with every other time/cluster unit. This might be a valid assumption for cluster units, but it is generally not true for longitudinal data. Typically, the closer the time points of measurement, the higher the correlation amongst them. For example, parenting stress on the day of measurement will have a higher correlation with parenting stress experienced yesterday, compared to a year ago. Therefore, the assumption of compound symmetry is more difficult to satisfy in applied research with people, whereas it was easier to assume in agricultural data that ANOVAs were originally developed to analyse. A further difficulty with repeated measures ANOVA is the treatment of missing data. All the data for individuals who have missing values are excluded from the analysis (i.e., listwise deletion). This is not the most efficient use of the data, and modern methods have been developed to use all available data.

MLMs and GEEs are two approaches for dealing with correlated data as they estimate the appropriate correlation and take it into account when estimating standard errors for the regression coefficients. Both can be viewed as extensions of GLMs with the main difference being the optimization method used for getting estimates of the regression coefficients and/or the partition of variance. MLMs were developed to analyse dependent variables where the independent variables were measured at different levels, or where data violated the assumption of independence. These models have been used extensively with continuous outcomes, although modern software now enables the analysis of categorical, binary or ordinal outcomes (e.g., Heck, Thomas & Tabata, 2012). GEEs were developed as an extension of GLMs for analysing outcomes that violate the assumption of independence, and are most often used with count, binary, ordinal or just non-normally distributed outcomes, although they can also be used with normally distributed data. In sections 2.2 to 2.5, we highlight

particular areas of interest within each modelling approach, and we give examples of their use in IDD research (3.1). By necessity, the present paper cannot cover all aspects of MLMs and GEEs; our aim is to provide readers with some initial information on these approaches to support them when considering adopting them.

2.2 Multilevel Models

2.2.1 Fixed and random effects in MLMs

Multilevel models are also known as mixed models because they mix two types of effects (variables): fixed and random effects. All independent variables used in an ANOVA/regression are effectively equivalent to fixed effects in MLMs. At least five different definitions of fixed effects have been proposed, but, at its simplest, Gelman and Hill (2007, p.245) suggest that fixed effects are constant across individuals, whereas random effects vary. When variables are specified as fixed effects in MLMs, they are assumed to: (i) be measured without error (ii) not vary, i.e., all their levels have been included in our study (the levels assumed to affect the outcome of interest). Gender is a typical example of a variable usually modelled as a fixed effect.

On the other hand, random effects: (i) are thought to represent a *sample* of all possible values of this characteristic in the population and (ii) have an underlying distribution that in the simplest case is the standard normal distribution. Thus, if we were conducting the same study again we could have, potentially, sampled different levels of the random effects. Examples of variables that fit into this description are individuals, time points, schools, wards etc. When we recruit individuals into a study we are not interested in these specific individuals but rather in what they represent as a (random) sample from the wider population of such individuals. For example, when we recruit children with ASD in one study, we are interested in these children, insofar as they represent a sample of children from all the population of children with ASD. When we sample a clinic or a classroom we are not interested (usually) in this specific clinic or classroom but rather on what this classroom/clinic represents as a member of the wider clinical or educational population. Time points can also be considered as random effects. Researchers are usually interested in change over some time period, rather than in the exact time points that they are taking the measurements at. Thus, although in a study the researchers might have planned to take measurements every 7 days, the same results might

have been obtained if measurements were obtained at 5- or 10-day intervals. The distributional assumption is also an important one and complements the issue of "sampling" point. We assume that random effects (individuals, patients, families, etc.) come from a standard normal distribution. For reasons that will not be discussed in this paper, random effects create a correlation amongst the measurements. For an explanation of how the random effect create this correlation the reader is referred to Fitzmaurice et al. (2004, pp.198-199).

2.2.2 Specifying random and fixed effects in MLM

Deciding on which variable to treat as fixed and which as random can be tricky, as one might have guessed given the variation in definitions. This is not helped by the fact that published papers do not always report whether independent variables were treated as fixed or random effects in their models. Here, we provide some guidelines on how to make a decision. The reader should be aware that, depending on the research question, a variable treated as fixed in one study may be treated as random in another. In other words, a variable's characterization as fixed or random is largely contextual.

In general, we treat as random: (i) independent variables whose values are a (random) sample from a wider population of values, and whose specific values (in our study) are not of intrinsic interest; (ii) independent variables that we believe create some correlation in our study (e.g., cluster units; repeated measures). These are general guidelines and the choice will depend on the specific research question. Below, we discuss selection by giving some examples.

We discussed the first criterion in 2.2.1 above. As long as the sample is randomly selected from the population, it will provide an unbiased estimate of the underlying population parameter. The point here is that individual participants are not of intrinsic interest. What is of interest is that they represent the population of interest. A very similar case can be made for classrooms, schools, districts, households etc. They are entities representing the wider population of classrooms, schools, districts, households. The specific entity (e.g. "household 1011") is not of intrinsic interest, but the fact it represents the wider population of such entities is.

In longitudinal studies, two variables typically fitted as random are the participants' identity (ID) and time. Specifying ID as a random effect essentially indicates that measurements with the same ID could be correlated and thus the (co)variance associated with this variable will be estimated as a result of the mathematical formulation of the model (Fitzmaurice et al., 2004). In this example, the variable participant ID is specified as a random effect following criterion (ii) in the paragraph above.

Time is another variable which is often specified as a random effect in longitudinal studies. Although time per se does not 'cause' a correlation between individuals, it usually is an essential component of correlation for longitudinal data since time is what creates the within individual component. Please note, however, that where we specify time as a random effect the underlying assumption (which ought to be reflected in the research question) is that we are not interested in the exact times measured in the study. In most cases, the time component in longitudinal research has an element of convenience and randomness (in the sense of choosing convenient times to suit the researcher), and researchers are not interested in the exact times chosen. When the latter is true, however (e.g., in intervention evaluation which is a special case of longitudinal design), the research question directly dictates measuring the effect of a specific time point, e.g., post or follow up. For example, De Boer and colleagues examined the effect of a school-based intervention on attitudes towards students with disabilities one week post and 12 months after the intervention (De Boer et al., 2014). To determine whether there were significant group (intervention vs control) differences at post and follow up (i.e. "Time") De Boer et al., (2014) fitted fixed effects of interaction terms group * Time. Another type of longitudinal design where time is meaningful as a random effect is developmental trajectories studies. We consider the specification of time as a random effect in such studies in detail in 2.2.4 below but before this we review the issue of levels in MLMs.

2.2.3 Levels in Multilevel Modelling

Multilevel modelling is called so because the random effects can be viewed in levels which are nested i.e. one level contain another level. Educational data illustrate this point well. We are interested in analyzing data collected from different schools, within educational districts on individual students. So, we have students within a class, classes within a school, and

schools within educational districts. Starting numbering from lowest to the highest nested level we have: (i) individual students within a class are at level 1 (ii) classrooms within schools are at level 2, (iii) schools within an educational district are at level 3, and (iv) educational districts are level 4. A 4-level multilevel model is quite complicated but in theory we could have as many levels as we want. In this example, if we had repeated measures of each individual student, the "time" variable becomes level 1, the individual becomes level 2, the classroom level 3 and so on. A variance component will be estimated for each level (four in total for the original 4 level model), with six covariances as well (n x (n-1)/2 where n = number of levels).

An intuitive way of thinking about the levels in MLM is by using the two stage random effects formulation. Let us consider a simple two level model: an outcome is measured on participants over time, along with another explanatory variable with two levels (e.g. Group 1 vs. Group 2). For each individual participant we could perform a regression of time on the outcome variable as a first step. If this is a simple linear regression with an intercept, for each individual we have estimated two parameters: (i) an intercept and (ii) a slope for time. We can then fit a regression line for each parameter (intercept and slope) using the independent variable (not used in the first step) and an intercept. This last intercept will represent the overall mean of the population for the outcome variable, and the estimated Group effect will represent the deviation of one group from the other. The first step is about random effects whereas the second one is for fixed effects. This formulation can be generalized to further random and fixed effects. Indeed, some software require a specification of this type for fitting MLMs (e.g., MLwin, HLM).

2.2.4 Intercepts and slopes in longitudinal MLMs that study growth

In longitudinal studies where the focus is on growth or developmental trajectories, there are two aspects of time that need to be considered: where people are at the start of the study or a crucial fixed time point (i.e., intercept); and how or how much people change over time (slope). Each participant is assumed to have a specific intercept (e.g., starting point) which is modelled as a random effect because it is considered a deviation from the overall population mean. Using IQ scores as an example, we estimate a baseline mean across the sample, and then estimate the intercept of each individual as the difference between each individual's

baseline score and that mean. Intercepts are usually assumed to be normally distributed with a mean of zero and a variance which is estimated from the data. Note that the mean of intercepts is by default zero, since they are effectively measures of deviation from the overall mean. Intercepts specified as random effects may be referred to as random intercepts.

For continuous variables in longitudinal studies, a random slope is also fitted for each individual. Similarly to the random intercepts, the change over time is modelled by fitting an overall slope for the average of the population, and each person's change over time is modeled as a deviation from the overall slope. Slopes are assumed to be normally distributed with a mean of zero and a variance which is estimated from the sample. Figure 2 provides a visual representation of two individuals (broken lines) and their mean (solid line) as an example of: (i) random intercepts and random slopes (ii) random slopes fixed intercepts and (iii) fixed slopes random intercepts.

----------Please insert Figure 2 here----------

Of interest is the specification of intercepts and slopes in a study by Mervis and colleagues (2012) who examined the trajectory of IQ scores in children with Williams syndrome. In their study, 40 children with Williams Syndrome completed the same IQ test, four to seven times over a period of five years on average. Researchers wanted to examine the trajectory of IQ in this population.

We will use the example of Mervis et al., 2012 to describe the most used MLM, that of longitudinal measurements taken on a cohorts/group(s) of individuals. This is described by a 2-level model. The lowest more "granular" level 1 describes the change of each individual over time (i.e. within individual), whereas level 2 describes differences between individuals. Thus, each individual's personal trajectory is described by two variables: (i) the individual and (ii) the time within each individual. These two variables will be used as random effects since they are sufficient for accounting for the correlation between measurements of the same individual. This will be fully described and explained in this section. Assuming that the reader is familiar with simple linear regression, we could describe these two levels as two separate (but linked) regressions: (i) one at level 1 where a linear regression is fitted for every

individual, separately. Thus, for each individual *j* we have the following regression line for each measurement over time *i*:

$$Yij = aj + bj*time + eij \quad (1)$$

In the above equation, we have sacrificed mathematical accuracy (no index I for "a" and "b") to highlight the fact that the above regression line is specific for each individual. In the example of Mervis et al (2012), Yij are the IQ measurements for individual j at time point i. Thus, for each individual we get an intercept for the IQ (aj), regression coefficient bj for the change over time and a residual (i.e. unexplained part of the IQ) eij. For the level 2 part of the MLM, there are two different but equivalent ways we can think. The first way is to think of the above regression coefficients being further split into an "average" and an "individual" part:

$aj = a0 + kj + aej$

$bj = b0 + dj + bej$

where a0 and b0 are the "average" intercept, and regression coefficient which are the same for each individual; kj and dj are deviations from the mean intercept and slope for each individual; and aej and bej are the respective residuals. At this level, we can introduce the between individual variables, which by definition will have the same value no matter what the timepoint (i.e. the within individual) is. In Mervis et al. (2012), these were the child gender and the maternal education variables. The second but equivalent way to think about level 2 is a second regression where we predict the population average values for the intercept and slopes from individual values, rather than decomposing the individual intercept and slope: one could think of it like taking the average of aj for intercepts and bj for slopes from the previous notation. Thus, we have:

$a0 = aj + aej$

$b0 = bj + bej$

This second way of thinking about this level is the so called 2-stage formulation of MLMs. Thus, in the case of Mervis et al. (2012) we estimate an overall intercept and slope for the average population from the estimates of the individuals' IQ intercepts and slope of the participants, adding the child sex and maternal education.

2.3 Generalized Estimating Equations (GEEs)

Generalized Estimating Equations (GEEs) take a different approach to MLMs: GEEs do not distinguish between fixed and random effects but rather require the specification of a clustering variable which is assumed to account for the covariance between measurements either on the same individual or on members of the same cluster. They use a different optimization procedure for achieving this compared to MLMs. A result of this different optimization procedure, which is often understated, is that the interpretation of the output is different compared to that of MLMs: MLMs essentially make inferences about the individual whereas GEEs make inferences about the population average (hence an alternative name for GEEs is "marginal models"). Mathematically, this difference is not so important when the underlying assumed distribution is symmetric (i.e., normal distribution) but becomes important when this distribution is non symmetrical (e.g., binomial distribution for logistic regression). However, this difference is crucial when we consider what the aims of the research are (e.g., individual prediction vs population description/prediction).

2.4 Model Specification in Analysis of Clustered Data

Finally, the two approaches to modelling clustered data differ in the assumptions they impose on the data. A full list of assumptions for each approach will not be provided here as relevant textbooks have available information (Garson, 2013; Fitzmaurice et al. 2004, pp.187-200, p. 294). Here, we will focus on two issues: model specification, and adequacy of sample size. Model specification relates to the method MLMs and GEEs use to estimate coefficients and their standard errors from their data. MLMs are particularly sensitive to model specification: in other words, if the distribution of error terms is inappropriately specified, any inference based on the estimated coefficients may be biased, because the standard errors of these coefficients will be biased (Garson, 2013). GEEs do not model random effects and by extension they do not need to model the covariance structure of these effects, hence they impose fewer assumptions on the data (Zeger, Liang, & Albert, 1988). Researchers suggest that GEEs can be more robust to misspecification of the distribution of the outcome variable or the link function, especially when the sample size is not very small (Hubbard et al., 2010).

2.4.1 Estimators in Multilevel Models

There are two iterative optimization methods that can be used with MLMs: (i) maximum likelihood (ML) and (ii) Restricted Maximum Likelihood (REML). In a statistical package, users typically have the option to select either of them. From a user's perspective, it suffices to know that ML will give biased results (albeit a very small bias) for the random effects (but unbiased for the fixed effects), whereas the REML will give biased results for the fixed effects (but unbiased for the random effects) (Zuur et al. 2009, pp.116-119). Thus, the question becomes how to select which one to use. Zuur et al. (2009, pp. 121-122) present a protocol for selecting the most parsimonious MLMs, and recommend the selection of the random part using REML first, and then the selection of the fixed part using ML. We believe this is sound approach. ML and REML are relatively sensitive to distributional misspecifications, and thus care should be taken to specify the correct underlying distribution. This is more so with small sample sizes.

2.4.2 Estimator in Generalized Estimating Equations

A very appealing property of the GEEs is the estimator used for obtaining the regression coefficients which is known as the sandwich estimator. We will not provide mathematical details but we will highlight some important properties of this estimator. The sandwich estimator tends to result in unbiased and consistent estimates for the regression coefficients and their standard errors, even when the covariance (i.e., correlation) structure is mispecified (Hubbard et al., 2010). What is required is that the mean is appropriately modelled (i.e. the regression equation) rather than the covariance. This is quite an appealing property, unlike in MLMs where a correct specification of the covariance is also important. However, one should be aware that this is the case when the following conditions hold (Fitzmaurice et al. 2004, pp.304-305): (i) the sample is large, (ii) the design is balanced (i.e., equal number of observations for each group), (iii) there are enough observations for the covariances to be appropriately estimated, (iv) the number of individuals observed is relatively large compared to the number of timepoints/clusters. If these conditions are not met, the standard errors obtained from the sandwich estimator will usually be underestimated (Fitzmaurice et al., 2004).

2.5 Sample Size Considerations

In terms of sample size, researchers need to consider not just the overall N, but also the number of available clusters, and the size of the clusters. From a statistical perspective

without a constraint on available resources (e.g. funds, subjects, time etc.) or ethical issues (e.g., we do not know if the new treatment is really beneficial, hence the test) more of everything is better in terms of sample size, but this is not the case in applied research where constraints such as the above are present. Research on sample size has examined parameters that need to be taken into account when determining optimum sample size within each of the two modelling approaches (e.g., Gibbons et al., 2010; Teerenstra et al., 2010). Research of this type is important for determining study design before the data is actually collected. "One-size fits all" methods for determining sample size/cluster size are not available; rather guidance is specific to design, i.e., whether the study is a trial, the level of randomisation (Level 1 or Level 2), whether the model includes any/how many random factors, if the outcome is continuous, if the link function is linear, etc.

In the field of IDD research, we are often faced with three common scenarios: (i) small N because of low prevalence of the condition (which means that more investment in recruitment is unlikely to result in a linear increase in N); (ii) access to a limited number of clusters (partly, this is due to low prevalence of IDD [if a small proportion of children have IDD than IDD-clinics would be fewer than, say, clinics for children without IDD], and partly to issues related to level of funding and ethics); and (iii) large cluster:participant ratios (i.e., a small-er number of participants in a large number of super-ordinate clusters): for example, when considering genetic syndrome clustering, or recruiting from mainstream classrooms or mainstream clinical services. Research that considers all three of these scenarios in determining sample size is limited, especially research that considers different types of outcomes (continuous and non-continuous), and/or research that compares the performance of GEEs and MLMs under these scenarios. Moving away from rules of thumb that have at times been suggested (e.g., 30 clusters with 30 participants each) and discredited (Bell et al., 2010a), the question in the IDD field is how low is too low for using either of these methods.

Overall, MLMs are fairly robust to small numbers of clusters and participants. Huang (2016) in a simulation study found that Level 1 estimated coefficients were not biased even with samples as low as five for Level 1 participants and 10 for Level 2 clusters, while Level 2 estimates required a minimum of 30 Level 2 clusters with five Level 1 participants each to avoid bias. Apart from the estimated coefficient, the impact of sample size has also been

examined in relation to the standard errors of the estimated parameters. Overall, standard errors for Level 1 variables seem to be unaffected by small participant or cluster numbers (Bell et al., 2010a, 2010b; Huang, 2016). In terms of the standard errors of the Level 2 coefficients, the pattern of findings differs among simulation studies. Huang (2016) in a simulation involving one Level 2 independent variable and two Level 1 independent variables found that Level 2 standard errors from MLMs were without bias even for the smallest sample size condition he examined (five Level 1 participants and 10 Level 2 clusters). Standard errors of Level 2 predictors were not biased even in models with more independent variables, binary or continuous, that also included interaction terms (Bell et al., 2010a). However, it appears that standard errors of Level 2 predictors are more prone to bias (higher possibility of Type I error) as the ratio cluster:participant increases (fewer participants in each cluster unit) when the number of clusters is small (50 or less; Bell et al., 2010b). A recent simulation study indicated that using GEEs with a small number of clusters (about 10) that have either few (ranging between 7 and 14 participants) or many participants (ranging 17 to 34 participants) was associated with a higher risk for Type I errors, despite the fact that estimated coefficients were fairly robust (McNeish & Stapleton, 2016).

A third area of consideration with regard sample size is power. It appears more challenging to achieve a desired level of power (e.g. conventional .80) in MLMs especially when they are fairly complex. Bell et al (2010a) found that achieving .80 power was possible with about 30 clusters with 20 to 40 Level 1 participants. When the number of clusters is below 10 and effect sizes are anything other than large, both MLMs and GEEs will be underpowered (NcNeish & Stapleton, 2016). With a small level of clusters (i.e., below 10), researchers need to consider whether modelling their effect is actually of substantive interest. Another way to account for the clustering effect without explicitly modelling it is to use a fixed effects regression, i.e., a standard regression model that includes dummy codes for the cluster levels (termed the fixed effects approach to clustering; Cohen, Cohen, West & Aiken, 2003, pp. 539-541). This has been shown to produce unbiased estimates and standard errors for Level 1 variables (Huang, 2016). This approach can be used even within a MLM environment to reduce model complexity. For example, in a MLM used to account for repeated measurements of activity levels in adults with ID, the clustering of adults within settings (a Level 3 hyper-ordinate factor) was not part of the research question and, moreover, there were only four settings

(van der Putten et al., 2017). Researchers thus fitted three dummy coded variables to represent the settings specified as fixed effects in a 2-level MLM (van der Putten et al., 2017).

Overall, simulation findings seem to suggest that variation in sample size at either individual- or cluster-level does not affect the estimated coefficients or Level 1 standard errors, but it may affect Level 2 standard errors (Bell et al., 2010a; 2010b; Huang, 2016; McNeish & Stapleton, 2016). Any conclusions regarding the direction of bias (and hence our ability to conclude accurately what is significant and what not) is restricted to the different combination of parameters examined in simulation studies to date. Given the large number of parameters that need to be taken into account to determine adequacy of sample size for analysis using a specific approach (N, cluster size, cluster:participant ratio, intraclass correlation, number of covariates, predictors binary or continuous, outcome distribution, design, etc.), it is important that more simulation studies, and in particular simulation studies with data from IDD research are conducted to facilitate field-specific recommendations. Our suggestion is that researchers determine the modelling approach guided by the research question primarily (and the considerations outlined in 2.2.1- 2.4.2). Any sample size considerations should include reference to model convergence (Bell et al., 2010a, 2010b; Nooraee, Molenberghs, & van den Heuvel, 2014).

For further guidance the reader is referred to the following resources which the authors find helpful. As a basis and starting point we recommend the paper by Rutterford et al, (2015) which summarizes in a very accessible way the methods for sample size determination in cluster randomized trials. An easy to read book which is focused on this issue is that of Ahn et al. (2015). Practical advice on how to estimate the various parameters needed for sample size calculations along with reference to relevant software are given by Guo et al. (2013).

3.1 Use of these modelling approaches in IDD research

An overview of the available literature suggests that MLMs have been used more frequently than GEEs when it comes to the analysis of data in the field of IDD research. GEEs are a relatively recent development in statistical analysis methods. They are also less likely to be in the training curriculum of social science researchers so it is likely that researchers in IDD are less exposed to them.

MLMs and GEEs have both been used to account for clustering caused by a hyper-ordinate structure. For example, MLMs have been used to analyse data clustered because of nesting of parents in couples (Garcia-Lopez et al., 2016; Hartley & Schultz, 2015; Jones et al., 2014; Langley et al., 2017; Pottie et al., 2009), families in households (Pottie et al., 2009), individuals with ID in community homes (Qian et al., 2015), support staff in organizations (Knotter et al., 2016). GEEs have been used to account for clustering caused by carers nesting within households (Totsika et al., 2017), multiple births nesting within women (Brown et al., 2016), or twins nesting within families (Cheng et al., 2015). A key difference in these two approaches reflected in the studies is that those studies that modelled their data using MLMs wanted to describe how much of the outcome variance could be attributed to the factors causing the clustering, whereas in GEE analyses this was not a main consideration, but a characteristic of the design that had to be controlled for prior to the interpretation of estimated parameters.

With repeated measures, researchers are often interested in modelling growth/developmental trajectory, or evaluating interventions with multiple evaluation points. MLMs tend to be used more often than GEEs in IDD research for either of these repeated measures scenarios, although overall there are more developmental trajectory studies than intervention evaluations in our field. A body of work has examined developmental trajectories or growth or change over time in cognitive, social, behavioural, or psychological outcomes of individuals with IDD or their carers using MLMs (e.g., Benson, 2014; Jenni et al., 2015, Hartman et al., 2014; Mervis et al., 2002; Wong et al., 2014, Woodruff-Borden et al., 2010). In such designs, MLMs are sometimes called growth curve models. Growth curve models as a term encompass longitudinal analyses fitted within a MLM framework or a structural equation framework (Curran, Obeidat, & Losardo, 2010). They may also be called hierarchical linear models, which refers to a group of MLMs that analyses data where the relationship between predictors and outcome is linear. Repeated measures studies with data analysed by GEEs have focused on modelling longitudinal outcomes in this population, especially outcomes that are not continuous (e.g., counts, binary or ordinal outcomes: Downes et a., 2015; Miller et al., 2017; Lin et al., 2007; Shoushtari et al., 2014), although not always (e.g., continuous adaptive skills scores analysed by GEEs in van Schie et al., 2013). Sample sizes vary considerably among these studies. Similarly, the number of repeated measures available also varies, but three is the minimum. While three is the

minimum number of repeated measures for a longitudinal analysis, some studies analysed data sets where not everyone had the required three time points (e.g., Benson, 2014, van Schie et a., 2013), whereas others exclude participants who did not have at least three repeated measures available (e.g., Woodruff-Borden et al., 2010). Overall, GEEs and MLMs cope well with unbalanced designs and missing data (and this is one of their main attractions compared to ANOVAs). However, when the aim of a MLM is to describe a developmental trajectory or growth line, three data points should be the minimum considered as anything less than three would only be able to identify linear trajectories.

Last, intervention evaluations have used MLMs to account for clustering caused by repeated measures (typically at least three evaluation points; pre, post and follow up) and hyper-ordinate structures (e.g., Level 2: intervention vs control group, Level 3: schools, clinics). Where the lack of association between a hyper-ordinate structure and outcomes could be established, researchers have dropped Level 3 to reduce model complexity and analysed 2-level models, even if randomization was clustered (e.g., de Boer et a., 2014; Hassiotis et al., 2009). GEEs have been used in intervention evaluations when intervention outcomes are not continuous (e.g., psychiatric diagnosis present/absent in Hassiotis et al., 2009) or not normally distributed (e.g., Shu & Lung, 2005; Shu, Lung, & Huang, 2002), though not always (e.g., Wei et al., 2012).

Intervention evaluations are still being analysed using repeated measures ANOVAs or MANOVAs or even change scores analysis with the pitfalls we described in 1.3.4. In addition to correctly accounting for data non-independence, MLMs and GEEs offer a lot more flexibility as they can cope with missing data, unequally spaced time points and outcomes that are not continuous or normally distributed. The more flexible modeling options can in turn offer more powerful interpretation. To highlight this, we note the slopes-as-outcomes MLMs Dykens and colleagues (2014) used to evaluate the effectiveness of a mindfulness intervention for parents of children with IDD. Their cluster randomized trial with six evaluation points focused not on between group differences at post/follow up controlling for pre-scores (which is what a mixed ANOVA model would do), but on the between group differences of the change over time (i.e., the course of change in mothers during the study; in other words mothers' slopes), thus

allowing a more interesting insight to the effects of the intervention, while also using all of the data available.

4.1 Conclusion

In this paper, we discussed some important features of MLMs and GEEs, two statistical methods for analyzing clustered/longitudinal data. Our aim was to encourage the take up of these modelling approaches where it is appropriate. We believe three factors could increase the use of such models. Firstly, we propose the inclusion of model specification details in published papers. A full description of how MLMs or GEEs were fitted, and why, is likely to increase use by other IDD researchers, who will be able to read, understand, and replicate the approach. Usually model specification details are sacrificed on the altar of word count restrictions. However, researchers and publishers have recognized the need to publish technical details for other methods (e.g., systematic reviews), and found ways to include them in papers (e.g., use of online appendices). Second, future simulation studies need to draw on data from IDD research to generate more knowledge about sample size and power that is specific to IDD research characteristics (for example, small N, limited clusters, large cluster:participant ratios). This information is invaluable during study design, which is the best time for deciding what statistical model to use (and in some cases, e.g., randomised trials, the only time when analysis can be planned). Last, training social sciences researchers in the use of MLMs and GEEs at undergraduate or postgraduate level will ensure that they have some core skills in place when they decide to embark on research in IDD. This is easier nowadays when statistical software has become more user-friendly.

Funding: The research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

References

Ahn, C., Heo, M., & Zhang, S. (2015). *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research*. CRC Press.

Bell, B.A., Morgan, B.G., Kromrey, J.D., & Ferron, J.M. (2010b). The impact of small cluster size on multilevel models: A Monte Carlo examination of two level models with binary and continuous predictors. *JCM Proceedings Survey Research Methods Section*, 4057-4067. Retrieved from https://ww2.amstat.org/sections/srms/Proceedings/y2010/Files/308112_60089.pdf

Bell, B.A., Morgan, B.G., Schoeneberger, J.A., & Loudermilk, B.L. (2010a). Dancing the sample size limbo with mixed models: How low can you go? *SAS Global Forum*, 4, 11-14. Retrieved from http://support.sas.com/resources/papers/proceedings10/197-2010.pdf

Benson, P.R. (2014). Coping and psychological adjustment among mothers of children with ASD: An accelerated longitudinal study. *Journal of Autism and Developmental Disorders*, 44, 1793-1807. doi: 10.1007/s10803-014-2079-9

Brown, H.K., Kirkham, Y.A., Cobigo, V., Lunsky, Y, & Vigod, S.N. (2016). Labour and delivery intervention in women with intellectual and developmental disabilities: a population-based cohort study. *Journal of Epidemiology and Community Health*, 70, 238-244.

Cheng, E.R., Palta, M., Poehlmann-Tynan, J., & Witt, W.P. (2015). The influence of childen's cognitive delay and behavior problems on maternal depression. *Journal of Pediatrics*, 167, 679-686. doi: 10.1016/j.jpeds.2015.06.003

Cohen, J., Cohen, P. West, S.G., & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (* 3rd ed). Mahwah, New Jersey: Lawrence Erlbaum.

Curran, P.J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve models. *Journal of Cognition and Development*, 11, 121-136. doi: 10.1080/15248371003699969

De Boer, A., Pijl, S.J., Minnaert, A., & Post, A. (2014). Evaluating the effectiveness of an intervention program to influence attitudes of students towards peers with disabilities. *Journal of Autism and Developmental Disorders*, 44, 572-583/ doi:10.1007/s10803-013-1908-6

Downes, A., Anixt, J.S., Esbensen, A., Wiley, S., & Meinzen-Derr, J. (2015). Psychotropic medication use in children and adolescents with Down Syndrome. *Journal of Developmental and Behavioral Pediatrics,* 36, 613-619. doi:10.1097/DBP.0000000000000179

Dykens, E.M., Fisher, M.H., Taylor, J.L., Lambert, W., & Miodrag, N. (2014). Reducing distress in mothers of children with autism and other disabilities: A randomized trial. *Pediatrics*, 134, E454-E463. doi:10.1542/peds.2013-3164

Emerson, E., Blacher, J., Einfeld, S., Hatton, C., Robertson, J., & Stancliffe, R. (2014). Environmental risk factors associated with the persistence of conduct difficulties in children with intellectual disabilities and autistic spectrum disorders. *Research in Developmental Disabilities, 35*, 3508-3517. http://dx.doi.org/10.1016/j.ridd.2014.08.039

Fitzmaurice G.M., Laird, N.M., & Ware J.H. (2004). *Applied Longitudinal Analysis. Wiley Series in Probability and Statistics*. New Jersey: John Wiley & Sons.

Garcia-Lopez, C., Sarria, E., & Pozo, P. (2016). Multilevel approach to gender differences in adaptation in father-mother dyads parenting individuals with autism spectrum disorder. *Research in Autism Spectrum Disorders*, 28, 7-16. doi: 10.1016/j.rasd.2016.04.003.

Garson, D. (2013). *Generalized Linear Models/Generalized Estimating Equations.* Statistical Associates Publishing. ISBN-10: 1626380155.

Gelman, G. & Hill, G. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gibbons, R.D., Hedeker, D., & DuToit, S. (2010). Advances in analysis of longitudinal data. *Annual Review of Clinical Psychology, 6,* 79-107. doi:10.1146/annurev.clinpsy.032408.153550.

Guo, Y., Logan, H.L Glueck, D.H., & Muller, K.E. (2013). Selecting a sample size for studies with repeated measures. *BMC Medical Research Methodology*, 13:100. Doi: 10.1186/1471-2288-13-100

Hanley, J.A., Negassa, A., Edwardes, M.D.deB., & Forrester, J.E. (2003). Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology,* 157, 364-375. doi: 10.1093/aje/kwf215

Hartley, S.L., & Schultz, H.M. (2015). Support needs of fathers and mothers of children and adolescents with autism spectrum disorders. *Journal of Autism and Developmental Disorders,* 45, 1636-1648. doi: 10.1007/s10803-014-2318-0

Harman, E., Smith, J., Westendrop, M., & Visscher, C. (2015). Development of physical fitness in children with intellectual disabilities. *Journal of Intellectual Disability Research*, 59, 439-449. doi: 10.1111/jir.12142

Hassiotis, A., Robotham, D., Canagasabey, A., Romeo, R., Langridge, D. Blizard, R., …King, M. (2009). Randomized, single-blind, controlled trial of a specialist behaviour therapy team for challenging behavior in adults with intellectual disabilities. *American Journal of Psychiatry*, 166, 1278-1285. doi:10.1176/appi.ajp.2009.08111747

Heck, R.H., Thomas, S.L., & Tabata, L.N. (2012). *Multilevel modelling of categorical outcomes using IBM SPSS.* New York: Routledge.

Hubbard, A. Ahern, J., Fleischer, N.L., Van der Laan, M., Lippman, S.A., Jewell, N…Satariano, W.A. (2010). To GEE or Not to GEE: Comparing population average and mixed models for estimating the associations between neighbourhood risk factors and health. *Epidemiology, 21,* 467-474. doi:10.1097/EDE.0b013e3181caeb90

Huang, F.L. (2016). Alternatives to multilevel modelling for the analysis of clustered data. *The Journal of Experimental Education*, 84, 175-196. http://dx.doi.org/10.1080/00220973.2014.952397

Jenni, O.G., Fintelmann, S., Calfisch, Latal, B, Rousson, V., & Chaouch, A. (2015). Stability of cognitive performance in children with mild intellectual disability. *Developmental Medicine and Child Neurology, 57*, 463-469. doi: 10.1111/dmcn.12620

Jones. L., Totsika, V., Hastings, R.P. & Petalas, M. (2013). Gender differences when parenting children with Autism Spectrum Disorders: a multilevel modeling approach. *Journal of Autism and Developmental Disorders,* 43, 2090-2098.

Knotter, M.H., Stams, G.J.J.M., Moonen, X.M.H., & Wissink, I.B. (2016) Correlates of direct care staffs' attitudes towards aggression of persons with intellectual disabilities. *Research in Developmental Disabilities,* 59, 294-305. doi: 10.1016/j.ridd.2016.09.008

Laird, N.M., & Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.

Langley, E., Totsika, V., & Hastings, R. P. (2017). Parental relationship satisfaction in families of children with Autism Spectrum Disorder (ASD): A multilevel analysis. *Autism Research,* 10, 1259-1268.

Lin, J.D., Loh, C.H., Choi, I.C., Yen, C.F., Hsu, S.W., Wu, J.L., & Chu, C.M. (2007). High outpatient visits among people with intellectual disabilities caring in a disability institution in Taipei: A 4-year survey. *Research in Developmental Disabilities*, 28, 84-93. doi: 10.1016/j.ridd.2005.12.003

McNeish, D., & Stapleton, L.M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research, 51*, 495-518, doi:10.1080/00273171.2016.1167008

Mervis, C.B., Kistler, D.J., John, A. E., & Morris, C.A. (2012). Longitudinal assessment of intellectual abilities of children with Williams Syndrome: multilevel modeling of performance on the Kaufman Brief Intelligence Test-2. *American Journal on Intellectual and Developmental Disabilities, 117,* 134-155.

Miller, M., Iosif, A-M., Hill, M., Young, G.S., Schwichtenberg, A.J., & Ozonoff, S. (2017). Response to name in infants developing autism spectrum disorder: A prospective study. *The Journal of Pediatrics,* 183, 141-146.e. doi: 10.1016/j.jpeds.2016.12.071

Nooraee, N., Molenberghs, G., & van der Heuvel, E.R. (2014). GEE for longitudinal ordinal data: Comparing R-geepack, R-multgee, R-repolr, SAS-GENMOD, SPSS-GENLIN. *Computational Statistics and Data Analysis,* 77, 70-83. https://doi.org/10.1016/j.csda.2014.03.009

Pottie, C.G., Cohen, J., & Ingram, K.M. (2009). Parenting a child with autism: Contextual factors associated with enhanced daily parental mood. *Journal of Pediatric Psychology*, 34, 419-429. Doi: 10.1093/jpepsy/jsn094

Qian, X., Ticha, R., Larson, S., Stancliffe, R.J., Wuorio, A. (2015). The impact of individual and organizational factors on engagement of individuals with intellectual disability living in community group homes: a multilevel model. Journal of Intellectual Disability Research, 59, 492-505. doi: 10.1111/jir.12152

Rutterford, C., Copas, A., & S. Eldridge. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, 44(3): 1051-1067. DOI: 10.1093/ije/dyv113

Shooshtari, S., Brownell, M., Dik, N., Chateau, D., Yu, C.T., Mills,…Wetzel, M. (2013). A population-based longitudinal study of depression in children with developmental disabilities in Manitoba. *Journal of Mental Health Research in Intellectual Disabilities, 7*, 191-207. DOI: 10.1080/19315864.2013.798389

Shu, B.C., & Lung, F.W. (2005). The effect of support group on the mental health and quality of life of mothers with autistic children. *Journal of Intellectual Disability Research,* 49, 47-53. doi:10.1111/j.1365-2788.2005.00661.x

Shu, B.C., Lung, F.W., & Huang, C. (2002). Mental health of primary family caregivers with children with intellectual disability who receive a home care programme. *Journal of Intellectual Disability Research,* 46, 257-263. doi: 10.1046/j.1365-2788.2002.00370.x

Teerenstra, S., Lu, B., Preisser, J.S., van Achterberg, T. & Borm, G.F. (2010). Sample size considerations for GEE analyses of three-level cluster randomized trials, *Biometrics, 66,* 1230-1237. doi:10.1111/j.1541-0420.2009.01374.x.

Totsika, V., Toogood, S., Hastings, R.P., & Lewis, S. (2008). Persistence of challenging behaviours in adults with intellectual disabilities over a period of 11 years. *Journal of Intellectual Disability Research,* 52, 446-457.

Totsika, V., Hastings, R.P., Emerson, E., Lancaster, G.A., Berridge, D.M., & Vagenas, D. (2013). Is there a bidirectional relationship between maternal well-being and child behavior problems in Autism Spectrum Disorders? Longitudinal analysis of a population-defined sample of young children. *Autism Research, 6,* 201-211. doi: 10.1002/aur.1279

Totsika, V., Hastings, R.P., Vagenas, D., & Emerson, E. (2014). Parenting and the behavior problems of young children with an intellectual disability: Concurrent and longitudinal relationships in a population-based study. *American Journal of Intellectual and Developmental Disabilities, 119*, 422-435. doi: 10.1352/1944-7558-119.5.422

Totsika, V., Hastings, R.P., & Vagenas, D. (2017). Informal caregivers of people with an intellectual disability in England: health, quality of life and impact of caring. Health and Social Care in the Community, 25, 951-961. doi: 10.1111/hsc.12393

Van Schie, P.E.M., Sciebes, R.C., Dallmeijer, A.J., Schuengel, C., Smits, D.W., Gorter, JW., & Bercher, J.G. (2013). Development of social functioning and communication in school-aged (5-9 years) children with cerebral palsy. *Research in Developmental Disabilities*, 34, 4485-4494. doi: 10.1016/j.ridd.2013.09.033

Van der Putten, A.A.J., Bossink, L.W.M., Frans, N., Houwen, S., & Vlaskamp, C. (2017). Motor activation in people with profound intellectual and multiple disabilities in daily practice. Journal of Intellectual and Developmental Disability, 42, 1-11. doi:10.3109/13668250.2016.1181259

Wei, Y.S., Chu, H., Chen, C.H., Hsueh, Y.J., Chang, Y.S., Chang, L.I., & Chou, K.R. (2012). Support groups for caregivers of intellectually disabled family members: effects on physical-psychological health and social support. *Journal of Clinical Nursing*, 21, 1666-1677. doi: 10.1111/j.1365-2702.2011.04006.x

Wong, J.D., Mailick, M.R., Greenberg, J.S., Kong, J., & Coe, C.L. (2014). Daily work stress and awakening cortisol in mothers of individuals with autism spectrum disorders or Fragile X. *Family Relations*, 63, 135147. doi: 10.1111/fare.12055

Woodruff-Borden, J., Kistler, D.J., Henderson, D.R., Crawford, N.A., & Mervis, C.B. (2010). Longitudinal course of anxiety in children and adolescents with Williams Syndrome. *American Journal of Medical Genetics, Part C-Seminars in Medical Genetics, 154C*, 277–290. doi:10.1002/ajmg.c.30259

Zeger, S.L., & Liang, K-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrics, 42, 121-130.

Zeger, S.L., Liang, K-Y., & Albert, P.S. (1988). Models for longitudinal data: A Generalized Estimating Equation Approach, *Biometrics, 44*, 1049-1060.

Zuur, A.F., Ieno, E., Walker, N., Saveliev, A.A., & Smith, G.M. (2009). Mixed Effects Models and Extensions in Ecology with R. New York: Springer. Doi: 10.1007/978-0-387-87458-6

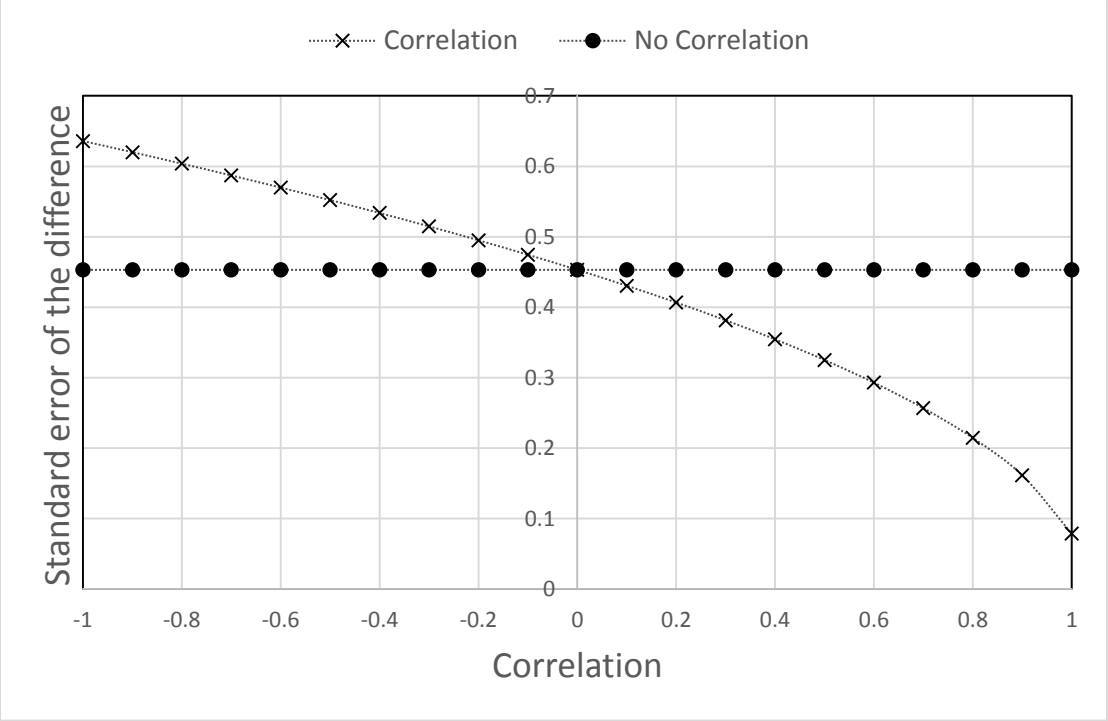
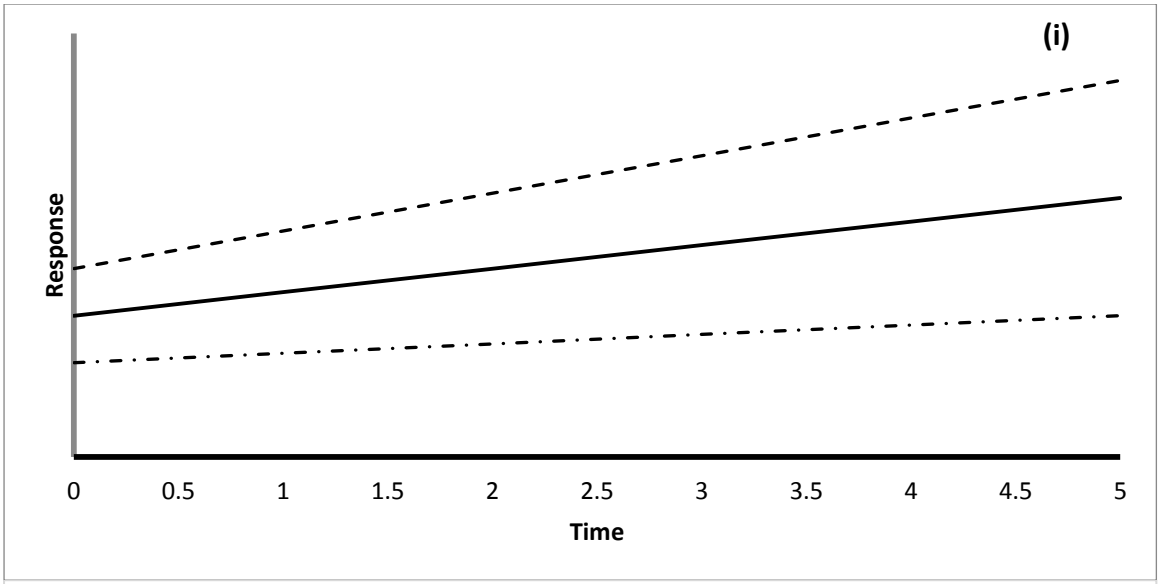
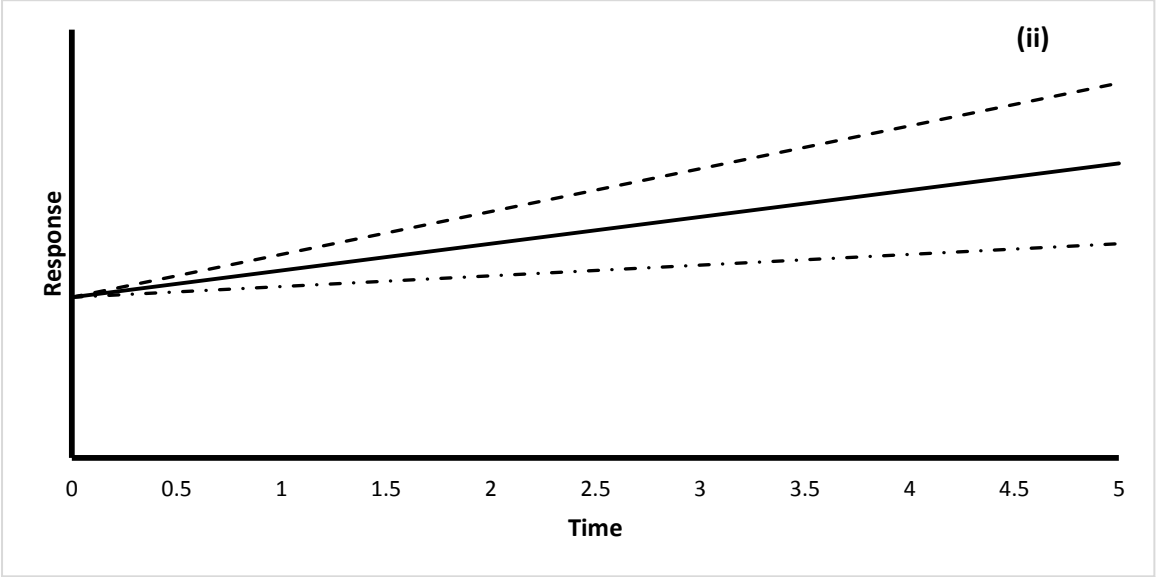


Figure 1.

Figure 2


(i)
Response
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Time


(ii)
Response
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Time

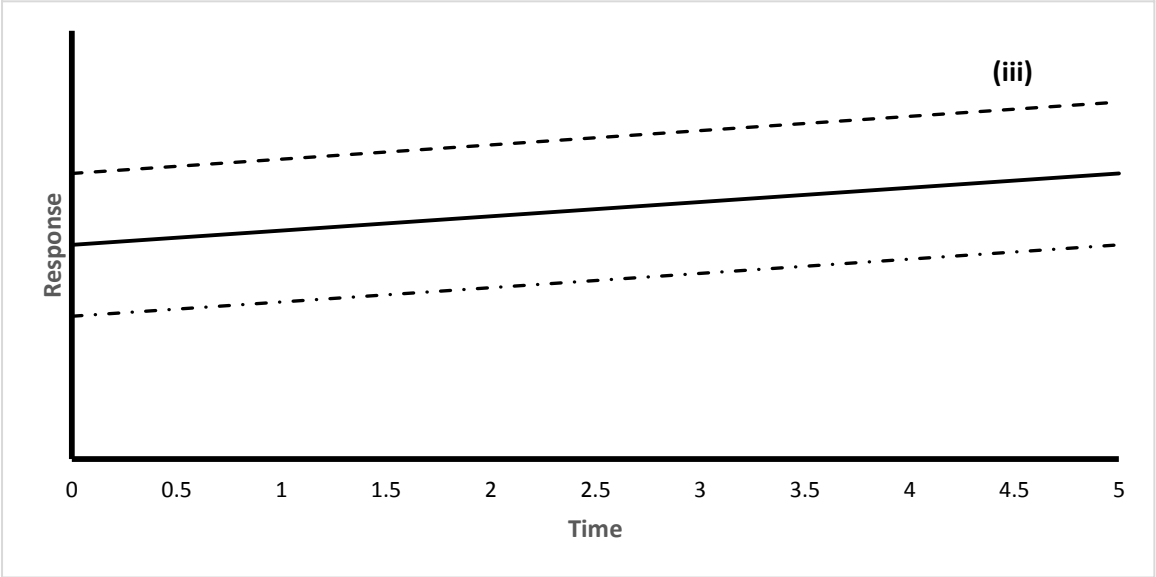

(iii)
Response
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Time

Figure 1. Comparison of standard error of mean difference, assuming no correlation (bullet points) and correlations from -1 to 1 (crosses).

Figure 2. Examples of multilevel models with: (i) random intercepts and random slopes (ii) random slopes fixed intercepts and (iii) fixed slopes random intercepts, for two individuals (broken lines) and their average (solid line).