MULTIVARIATE BEHAVIORAL RESEARCH
2017, VOL. 52, NO. 5, 630–647
https://doi.org/10.1080/00273171.2017.1354758

Routledge
Taylor & Francis Group

Double Decomposition of Level-1 Variables in Multilevel Models: An Analysis of the Flynn Effect in the NSLY Data

Patrick O'Keefe and Joseph Lee Rodgers

Vanderbilt University, Peabody College, Nashville, TN, USA

ABSTRACT

This paper introduces an extension of cluster mean centering (also called group mean centering) for multilevel models, which we call "double decomposition (DD)." This centering method separates between-level variance, as in cluster mean centering, but also decomposes within-level variance of the same variable. This process retains the benefits of cluster mean centering but allows for context variables derived from lower level variables, other than the cluster mean, to be incorporated into the model. A brief simulation study is presented, demonstrating the potential advantage (or even necessity) for DD in certain circumstances. Several applications to multilevel analysis are discussed. Finally, an empirical demonstration examining the Flynn effect (Flynn, 1987), our motivating example, is presented. The use of DD in the analysis provides a novel method to narrow the field of plausible causal hypotheses regarding the Flynn effect, in line with suggestions by a number of researchers (Mingroni, 2014; Rodgers, 2015).

KEYWORDS

Centering; Flynn effect; multilevel modeling; within- versus between-level variance

Introduction

When data are collected from clustered observations emerging from a nested design structure – members of families, children in classrooms, or repeated measures – multilevel data analysis is a recommended analytic method (Snijders & Bosker, 2012). Researchers routinely use variables drawn from multiple levels. In a family study, for example, a researcher might use data about children, the lowest level of analysis (level 1), as well as data about the family itself, the cluster level (level 2), to predict an outcome. Information about higher level clusters (e.g., parental education) provides a context for observations at the lower levels (e.g., child achievement scores). It was acknowledged early that ignoring the context of an observation in multilevel data could result in the attribution of effects to the wrong level of analysis (Cronbach & Webb, 1975), such as attributing family characteristics to child characteristics and vice versa. Context effects can be any variable measured at the cluster level; however for this paper, we will focus on the context effects derived from lower level variables (e.g., cluster means).

A common context effect used in multilevel models is the cluster mean. The cluster mean is the mean of all observations in a given cluster, the higher level unit (e.g., the mean achievement scores of students in a classroom, with classroom being the level-2 unit). The cluster mean is particularly useful because variables centered using the cluster mean are uncorrelated with variables at higher levels (Raudenbush, 1989; see also Appendix A). Cluster mean centering can help in increasing the interpretability of model parameters because the lack of correlation helps to isolate the level at which an effect occurs. However, for some research questions, the cluster mean does not address the question of interest, whereas other context effects derived from the same lower level variables (e.g., baseline scores) do address the question. The importance of context effects beyond the cluster mean is not a novel suggestion (Cronbach et al., 1976; Plewis, 1989), but previous work has not presented a way to obtain the mathematical benefits of cluster mean centering (e.g., uncorrelated higher and lower level variables) when using alternative context variables derived from lower level variables. This paper presents a method of cluster mean centering that also allows for the use of alternative context effects. This method decomposes a single independent variable into multiple independent variables, with the goal of increasing model descriptiveness and interpretability. We call the method "double decomposition (DD)," because one decomposition is the standard mean centering approach (across levels) followed by additional decompositions in relation to other contextual variables within the same level, measured using the same units (e.g., "years"). Under

CONTACT Patrick O'Keefe patrick.okeefe@vanderbilt.edu Department of Psychology and Human Development, Vanderbilt University, Nashville, TN 37240-0002, USA.

certain circumstances, this results in variables with clearer substantive interpretations than cluster means.

The substantive problem that motivated the development of this method, and which we will return to throughout the paper, is the Flynn effect. The Flynn effect is the name for systematic increases in IQ scores over the past century in countries all over the world (Flynn, 1987). This effect has been observed in many different data sets across time and culture, using a variety of research methods. Relevant to the current study, the Flynn effect was observed in patterns of Peabody Individual Achievement Test-Math (PIAT-Math) scores from 1986 to 2000 among children from the National Longitudinal Survey of Youth-Children survey (NLSYC; Rodgers & Wänström, 2007). A current version of this data set, with data from 1986 to 2012, will be used here.

The purpose of the analysis presented in this paper is not to prove the existence of the Flynn effect in the NLSYC (which has already been demonstrated), but rather to identify the location of the Flynn effect. Does the increase in PIAT-Math scores over time emerge from an annual increase in an individual's IQ score, an increase in scores between birth cohorts, or an increase in scores between family cohorts? Most previous analyses have focused on a single level in analyses. In fact, most prior studies of the Flynn effect lacked the appropriate data necessary to study this effect in a multilevel context (see Sundet (2014)). All studies of the Flynn effect have the same basic independent variable: year of testing. This is the variable on which we will focus in our demonstration, with additional context variables to answer more nuanced questions than in the previous Flynn effect research.

The DD model

Before considering DD, it helps to look at the two alternatives, cluster mean centering and a decomposition that uses substantively interesting variables (but ones that lack the properties of cluster means). When using multilevel models, it is recommended that the researchers utilize cluster mean centering, which can result in more interpretable parameters and can distinguish the effects of lower level predictors from those of higher level predictors (Curran, Lee, Howard, Lane, & MacCallum, 2012; Enders & Tofighi, 2007; Hoffman, 2014; Hoffman & Stawski, 2009; Raudenbush, 1989; Wang & Maxwell, 2015). Cluster mean centering is thoroughly explained by a number of authors in a variety of fields of research (Hoffman, 2007; Hoffman & Stawski, 2009; Kreft, 1995; Paccagnella, 2006; Wang & Maxwell, 2015; Wu & Wooldridge, 2005) and in several popular textbooks (Aiken & West, 1991; Hoffman, 2014; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). From the NLSYC, we can create a multilevel design, with repeated measures (the first level) nested within child (the second level), and multiple children nested within family (the third level). Table 1 shows what cluster mean centered (CMC) data would look like for two families with multiple children and multiple observations for the children. The variable "year of testing" is cluster mean centered by each child's mean year, and then the child means are centered by each family's mean. The resulting family means and mean-centered variables sum up to the original year of testing variable (the child mean is, itself, mean-centered). Each of these variables could be then entered into a model and slopes could be used to estimate their individual effects on a given outcome. This is how cluster mean centering is typically applied, and it can help to determine if an effect is consistent across levels; consistency cannot be determined without cluster mean centering. However, in some research settings, the cluster mean is not a particularly interpretable variable.

Frequently, we directly measure the variables that are of interest. In the cases when these variables use the same unit of measurement as lower level variables (e.g., "years"), deviation scores can be created in the same way as with cluster means. Because these context variables can be measured independently of the original independent variable, and can arise naturally as the variables of interest, we use the term "natural context variable" to emphasize that these variables are not calculated (as cluster means are). We illustrate in Table 2 how several

Table 1. Cluster mean centering of year of testing.

Family	Child	Year of testing	Child mean year of testing	Year of testing centered by child mean year of testing	Family mean year of testing	Child mean year of testing centered by family mean year of testing
1	1	1996	1997	−1	1998	−1
1	1	1998	1997	1	1998	−1
1	2	1998	1999	−1	1998	1
1	2	2000	1999	1	1998	1
2	1	1990	1992	−2	1992.8	−.8
2	1	1994	1992	2	1992.8	−.8
2	2	1992	1993	−1	1992.8	.2
2	2	1994	1993	1	1992.8	.2
2	3	1994	1994	0	1992.8	1.2

Table 2. Decomposition of year of testing using natural context variables.

Family	Child	Year of testing	Birth year of child	Birth year of first child in family	Years current child born after first child in family	Age of current child at year of testing
1	1	1996	1988	1988	0	8
1	1	1998	1988	1988	0	10
1	2	1998	1990	1988	2	8
1	2	2000	1990	1988	2	10
2	1	1990	1984	1984	0	6
2	1	1994	1984	1984	0	10
2	2	1992	1985	1984	1	7
2	2	1994	1985	1984	1	9
2	3	1994	1987	1984	3	7

natural context variables can provide a perfect decomposition in the NLSYC data. The third column denotes the year of testing for several actual observations within the NLSYC data. The last three columns (columns 5–7) show three interpretable context variables that fully decompose the year of testing. The fourth column, birth year of child, is redundant in the final decomposition, although it can be thought of as being centered by the birth year of the first child in the family giving column 6, "Years current child born after first child in family." It is easily verified that for each row, the three final variables sum up to the year of testing. Thus, entering those three variables as IVs, instead of the traditional year of testing as an IV, allows a much more nuanced evaluation of the location of the Flynn effect. In this paper, we will refer to variables that can be decomposed like this as "composite" variables because we can think of them as being composed of multiple related variables. Note that we have only illustrated in Table 2 a decomposition using natural context variables, which leaves us vulnerable to attributing effects to the wrong level of the model. Further decomposition is required to assist in the interpretation of the location of the Flynn effect in these data, as we will illustrate in our expanded Flynn effect analysis later in this paper.

The major advantage of cluster means over other context variables (such as the birth year in our example) is that centering using the cluster mean produces variables that are orthogonal across levels, something no other contextual variable produces (Raudenbush, 1989). For example, using cluster mean centering in our Flynn effect example produced two variables (child mean year of testing and deviations from that mean; columns 4 and 5 in Table 1) that were both orthogonal and uncorrelated. The same result is not obtained by centering using child's birth year or other natural context variables. These correlation structures can be verified by referring Tables 1 and 2.

As researchers, we may have substantive questions unanswered by using cluster means, but the questions are confounded if we simply use the natural context variables related to our questions. Unfortunately, if we include both the cluster mean and the natural context variable in our analysis, the cluster mean may be substantially correlated with our natural context variable. This analysis is not necessarily wrong; however, it can complicate the interpretation of effects. The current paper introduces an extension of cluster mean centering called Double Decomposition ("DD"), which handles this problem, providing results that are more directly interpretable and intuitive. This centering method separates a given variable into orthogonal between-level components, which achieves the same effect as cluster mean centering, and then defines within-level components (e.g., a natural context variable and a remainder) that help to reduce the correlation between, and disaggregate effects due to, cluster means and natural context variables. Our method is a special case of cluster mean centering. It is useful when context effects (beyond the cluster mean) which utilize the same unit of measurement as the lower level variable exist and when the research question is not answered by using cluster means. In the Flynn effect example, the birth year of a child is that kind of context effect. The purpose of this paper is to present DD as an extension of cluster mean centering, the one that allows a researcher to distinguish the effects of natural context variables from cluster means.

The notation is important. Because we are splitting a variable into at least three parts, we need a way to distinguish between the parts and also to indicate that they are all related. In this paper, subscripts are used to distinguish across levels, whereas marks above the variable name (e.g., $\cdot$ and $\cdot\cdot$) are used to indicate the component being discussed. For example, a generic level-1 variable, $generic_{ij}$ in a multilevel model has a variable name "generic" as well as subscripts indicating to which level the variable applies. The "*i*" and "*j*" subscripts tell us that this is observation *i* in cluster *j*. Further subscripts (*k*, etc.) can be added to denote higher level clusters. In the empirical example of the Flynn effect, we use the name of the variable and three additional subscripts, *i*, *j*, and *k*, to denote the interview within child, child within family, and family levels. Because the original variable, its cluster mean, and potentially the natural context effect will all share the same name, an additional notation is needed to

distinguish these variables. In this paper $\dot{}$ and $\ddot{}$ ("dots") denote CMC variables, $\bar{}$ and $\breve{}$ indicate cluster means, and $\tilde{}$ represents the difference between an original variable and a natural context variable.

In cluster mean centering, a level-1 variable, x_{ij}, can be thought of as a linear combination of two separate variables, a level-1 variable $\dot{x}_{ij}$ and a level-2 variable $\bar{x}_{.j}$, where $\bar{x}_{.j}$ is the mean of x_{ij} for cluster j (the level-2 unit), and the "." subscript indicates that $\bar{x}_{.j}$ is the mean of all i observations in the level-2 unit j. Thus, $\dot{x}_{ij} + \bar{x}_{.j} = x_{ij}$. The end result is two variables that can be used to model the effect of overall differences on x between clusters, as well as the effect of individual differences on x within a cluster. The centered variable and the resulting mean variable do not co-vary and may or may not measure separate effects (Cronbach et al., 1976; Raudenbush, 1989). Both the cluster mean and the individual deviations from that mean can be used as predictors in a multilevel model, with slopes representing the effects of these variables on the dependent variable (DV). The interpretation of slopes in which researchers have not included the cluster mean can result in the attribution of effects to the wrong level, the ecological fallacy (Curran et al., 2012). In the present case, an additional error can be made: failing to properly separate the natural context effect from the cluster mean can result in an improper attribution of effects, but within the same level instead of across levels.

With a composite variable, like year of testing in the Flynn effect example, there is a level-2 natural context variable z_j, and $x_{ij} - z_j = \tilde{x}_{ij}$. The z_j variable, unlike cluster means, can be measured independently of x_{ij}, and represents a distinct but related variable, that is measured using the same units (note that these "units" are the same measured unit, but they are not on the same statistical scale, i.e., they do not necessarily have the same variance). The difference component, $\tilde{x}_{ij}$, of a composite variable requires cluster mean centering. Cluster mean centering gives $\tilde{x}_{ij} - \breve{x}_{.j} = \ddot{x}_{ij}$, where $\breve{x}_{.j}$ is the cluster or group mean of $\tilde{x}_{ij}$. This method results in one level-1 variable $\ddot{x}_{ij}$ and two level-2 variables $\breve{x}_{.j}$ and z_j such that $\ddot{x}_{ij} + \breve{x}_{.j} + z_j = x_{ij}$. The cluster mean of x_{ij} is $\bar{x}_{.j} = \sum \frac{x_{ij}}{n}$ in a given cluster j. This equation can be rewritten in terms of the level-2 variables z_j and $\tilde{x}_{ij}$, giving $\bar{x}_{.j} = \sum \frac{\tilde{x}_{ij}+z_j}{n}$. Because the summation is confined to a single cluster, z_j is simply a constant giving $\bar{x}_{.j} = \sum \frac{\tilde{x}_{ij}}{n} + z_j$. The expression $\sum \frac{\tilde{x}_{ij}}{n}$ is the cluster mean of $\tilde{x}_{ij}$ which is $\breve{x}_{.j}$, giving $\bar{x}_{.j} = \breve{x}_{.j} + z_j$. Because $\dot{x}_{ij} = x_{ij} - \bar{x}_{.j}$ and $\ddot{x}_{ij} = x_{ij} - (\breve{x}_{.j} + z_j)$, we see that $\ddot{x}_{ij} = \dot{x}_{ij}$. The subtraction of z_j from x_{ij} leaves some level-2 variance so long as $z_j \neq \bar{x}_{.j}$. The level-2 variance not removed by z_j is removed by $\breve{x}_{.j}$. The two level-2 variables are nonorthogonal components of the original cluster mean variable, $\bar{x}_{.j}$. The result is that we decompose both across levels and within the same level in a way that may improve interpretability and help to avoid incorrect attributions of effects to $\bar{x}_{.j}$ if indeed the effects of $\breve{x}_{.j}$ and z_j are different. Because the DD creates two variables at level 2, out of a single variable, and each is measured using the same units, the potential for high collinearity between double decomposed variables may raise concerns. In general, this is unlikely unless the natural context variable is closely related to the level-2 cluster mean. A more complete mathematical explanation can be found in Appendix A.

Table 3. Raw score multilevel model.

	Model 1
Level 1	$y_{ij} = \beta_{0j} + \beta_{1j} * x_{ij} + e_{ij}$
Level-1 intercept	$\beta_{0j} = \gamma_{00} + u_{0j}$
Level-1 slope	$\beta_{1j} = \gamma_{10}$
Reduced form	$y_{ij} = \gamma_{00} + u_{0j} + \gamma_{10} * x_{ij} + e_{ij}$

The potential importance of DD can be seen if we compare three simple models: the raw score (RS) model, the CMC model, and the DD model. The RS model forms the basis of the other two and a simple version of an RS model would be one with a single independent variable with a random intercept and fixed slope as seen in Model 1 (Table 3).

It has been pointed out by many previous authors (Curran et al., 2012; Kreft, de Leeuw, & Aiken, 1995; Snijders & Bosker, 2012) that the RS model assumes that the effects of the level-2 component $\bar{x}_{.j}$ and the level-1 component $\dot{x}_{ij}$ are identical. The CMC model does not make this assumption. By cluster mean centering and reintroducing the cluster means as a level-2 variable, researchers are able to test the hypothesis that the effects of the cluster mean and the level-1 deviations from the cluster mean are the same. A CMC alternative to our RS model is Model 2 (Table 4).

This model allows for the cluster mean to have its own slope γ_{01}. More complex models could have the cluster mean of x predicting the slope of the lower level components of x.

The DD model takes this one step further. In addition to splitting across levels as the CMC model does, the DD

Table 4. Cluster mean centered multilevel model.

	Model 2
Decomposition	$\bar{x}_{.j} + \dot{x}_{ij} = x_{ij}$
Level 1	$y_{ij} = \beta_{0j} + \beta_{1j} * \dot{x}_{ij} + e_{ij}$
Level-1 intercept	$\beta_{0j} = \gamma_{00} + \gamma_{01} * \bar{x}_{.j} + u_{0j}$
Level-1 slope	$\beta_{1j} = \gamma_{10}$
Reduced form	$y_{ij} = \gamma_{00} + \gamma_{01} * \bar{x}_{.j} + u_{0j} + \gamma_{10} * \dot{x}_{ij} + e_{ij}.$

Table 5. Double decomposition multilevel model.

	Model 3
Decomposition	$\check{x}_{.j} + z_j + \dot{x}_{ij} = x_{ij}$
Level 1	$y_{ij} = \beta_{0j} + \beta_{1j} * \dot{x}_{ij} + e_{ij}$
Level-1 intercept	$\beta_{0j} = \gamma_{00} + \gamma_{01} * \check{x}_{.j} + \gamma_{02} * z_j + u_{0j}$
Level-1 slope	$\beta_{1j} = \gamma_{10}$
Reduced form	$y_{ij} = \gamma_{00} + \gamma_{01} * \check{x}_{.j} + \gamma_{02} * z_j + u_{0j} + \gamma_{10} * \dot{x}_{ij} + e_{ij}$

Table 6. Cluster mean centered model with a natural context variable included.

	Model 4
Decomposition	$\bar{x}_{.j} + \dot{x}_{ij} = x_{ij}$
Level 1	$y_{ij} = \beta_{0j} + \beta_{1j} * \dot{x}_{ij} + e_{ij}$
Level-1 intercept	$\beta_{0j} = \gamma_{00} + \gamma_{01} * \bar{x}_{.j} + \gamma_{02} * z_j + u_{0j}$
Level-1 slope	$\beta_{1j} = \gamma_{10}$
Reduced form	$y_{ij} = \gamma_{00} + \gamma_{01} * \bar{x}_{.j} + \gamma_{02} * z_j + \gamma_{10} * \dot{x}_{ij} + u_{0j} + e_{ij}.$

model splits within the higher level. A DD version of the previously presented CMC model appears in Model 3 (Table 5).

The CMC model assumes that $\check{x}_{.j}$ and z_j share the same slopes, a potential error similar to the one made by the RS model. Because z_j is a "natural" variable, one that could be overlooked as being a component of x_{ij}, it is possible that a researcher might inadvertently include it in their models along with cluster means (or in an uncentered model x_{ij}). Because $\bar{x}_{.j} = \check{x}_{.j} + z_j$, such a model will not provide a direct estimate of the effect of z_j; instead, it will estimate the effect conditioned on the fact that $\bar{x}_{.j}$ has "duplicate" variance exactly equal to that of z_j. This effect will be the difference between the effect of $\bar{x}_{.j}$ and the effect of z_j. This will be demonstrated by simulation later in this paper.

For more than two levels, the process of DD can be repeated as necessary. It should be noted at this point that DD is not limited to subtracting only one variable of the type z_j from the previous levels. In the empirical example presented later in this paper, we carried the decomposition out across three levels with a time variable, year of test, being decomposed. At the third level, two natural level-3 components existed, mother's age at first birth and mother's birth year; both were included in the model without difficulty.

In all cases of DD, the final variables included in the model should sum up to the original variable. In our empirical example to be presented, even with seven different components, the sum of all the components is exactly equal to the original variable. In a model in which the variables do not add back to the original value, the researcher will have omitted variance present in the original variable.

Finally, when using DD, researchers should recognize that complexity is added by DD, and the goal of parsimony in model building may be threatened. Unlike cluster mean centering, which is almost universally applicable, researchers should not feel compelled to find ways to doubly decompose their data. In fact, it is likely that there will not necessarily be a sensible way to doubly decompose data in many cases. However, in cases where DD is possible, if the effects of the natural context variables differ from those of the other context variables (e.g., the cluster mean), then DD is necessary to separately identify those effects (similar to the logic behind CMC).

Simulation

The following simulation is designed to demonstrate the potential utility of the DD formulation. The goal is to demonstrate the utility of the model through a proof-in-principle, rather than a complete exploration of the effects of DD. We will use a two-level data structure with a single outcome variable and one independent variable to be doubly decomposed. The models to be tested include the three models just presented in Tables 3–5, as well as an additional model presented in Table 6. Data were simulated in R using the MASS package (R Core Team, 2016; Venables & Ripley, 2002) and analyzed using the nlme package (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2016).

Because of the nature of the decomposition, it was necessary to create the individual components of the variables separately and then combine them in subsequent analyses as needed. To simulate the level-2 data, three variables were drawn from a standard multivariate normal distribution. One was the outcome variable ($\bar{y}_{.j}$), one was the natural context variable (z_j) and one was the cluster mean deviations from the natural context variable ($\check{x}_{.j}$). The level-2 population correlations were $r = .5$ between the outcome and the natural context variable and $r = -.5$ between the cluster means and the other two variables. It should be noted here that the partial correlation and standardized regression coefficient among these three variables would be $\pm.33$ in the population. One hundred level-2 observations were selected and each was duplicated 30 times (the level-2 observation is duplicated across all members of a group), giving 100 clusters of size 30 each. This step created the level-2 observations, and the level-1 components needed to be created separately.

For the level-1 components of the variables, 3000 observations were independently drawn from a standard normal distribution. These observations were divided into groups of 30. Each group was CMC, and the cluster means from this procedure were discarded. Because CMC variables are uncorrelated with higher level variables, it

was not necessary to use the cluster means from the lower level observations, i.e., we can replace them with the cluster means generated independently at level 2 for the purposes of this simulation. The resulting variable measured individual deviations from the cluster mean, the level-1 variable ($\dot{x}_{ij}$) in the subsequent analyses. This variable was then used to create an outcome variable, $\dot{y}_{ij}$. The two level-1 variables had $r = .5$. The two outcome variables were added together to create an outcome variable for the multilevel model ($\bar{y}_{.j} + \dot{y}_{ij} = y_{ij}$). The independent variable was created by adding the natural context variable, the cluster means, and the CMC observations ($\breve{x}_{.j} + z_j + \dot{x}_{ij} = x_{ij}$). A cluster mean that would result from a typical CMC model was created by adding the natural context and cluster mean variables ($\breve{x}_{.j} + z_j = \bar{x}_{.j}$). The resulting variables were used in four models. One model used the summed independent variable, without any centering (Model 1; Table 3), the next model was a standard CMC model (Model 2; Table 4), the third model was the DD model (Model 3; Table 5), and the fourth model used the cluster means from the CMC model, but also included the natural context variable from the DD model (Model 4; Table 6). All models estimated random intercepts because the process of creating $\dot{y}_{ij}$ introduced random intercepts. All slopes were fixed. This process was repeated 5000 times. R code for the simulation can be found in Appendix B.

The first model (Table 3), which used the sum of the three component variables (x_{ij}) to predict the outcome (y_{ij}), gave an average slope estimate of .49, with a mean t value of 30.77 across the 5000 replications. This model would suggest that a one unit increase in the independent variable is associated with a .49 unit increase on the dependent variable. This model does not answer any questions about how the effect of the dependent variable differs across levels.

The second model (Table 4) denotes a typical CMC model. The lower level variable is the deviation of individuals from their cluster mean ($\dot{x}_{ij}$). That variable had a slope of .5 with an average t value of 31.07. The cluster mean, which in our case is a sum of the natural context variable and the mean deviation from that natural context variable ($\bar{x}_{.j} = \breve{x}_{.j} + z_j$), has an average slope of zero with an average t value of $-.01$ (as opposed to the population correlations of $-.5$ and .5 for $\breve{x}_{.j}$ and z_j, respectively). If a researcher was to use this model to assess the data we have used here, they might erroneously conclude that there was no substantial relationship between the level-2 component of the independent variable and the dependent variable.

The third model (Table 5) uses the DD formulation. The level-1 results are unchanged from the CMC model; however, the level-2 results are different. The slope for the cluster mean deviation ($\breve{x}_{.j} = \bar{x}_{.j} - z_j$) is $-.33$ and the slope for the natural context variable (z_j) is .33. These values accurately reflect the population level regression coefficients. The corresponding t values are -3.48 and 3.46, respectively. This model reveals significant associations between components of the independent variable and the dependent variable at level 2 that were invisible when using the more typical CMC model.

The fourth model (Table 6) used the CMC variable at level 1 ($\dot{x}_{ij}$), the natural context variable at level 2 (z_j), and the raw cluster mean ($\bar{x}_{.j}$). In effect, this model includes the natural context variable twice, first as the natural context variable (z_j), and again as a part of the cluster mean ($\bar{x}_{.j} = \breve{x}_{.j} - z_j$). The slopes and t values remained unchanged for the level-1 variable ($\dot{x}_{ij}$) and the cluster mean ($\bar{x}_{.j}$); however, the slope and the t value doubled for the natural context variable (z_j; $B = .67$; $t = 6.94$). The estimate is starkly different from previous models. The coefficient for the natural context variable represents the context effect and denotes the difference between the effect of the cluster mean and the natural context variable. Here, the unique effect of both $\bar{x}_{.j}$ and z_j was estimated, because 100% of the variance associated with z_j was also present in $\bar{x}_{.j}$ and the coefficient for z_j had to account for this and doubled in size. If a researcher attempted to interpret the slope directly, without fully considering the impact of including the other variables, s/he could misinterpret the effect as being far stronger than it actually is.

These results are not meant to be a complete analysis of the potential effects of various model formulations when DD is possible. They are meant to highlight the potential importance under certain conditions of the DD model, which is demonstrated here to have the potential to identify an effect that was built into the data but that the traditional multilevel analysis would not have identified. Specifically, failing to doubly decompose in this case led to the effects at level 2 going unnoticed, and additionally, simply adding the natural context variable to the CMC model artificially doubled the corresponding slope, which could lead to erroneous conclusions if researchers are not careful.

Example applications

An analysis of the Flynn effect was the initial motivation for the development of this method. We hope that other researchers will find their own uses for DD. The basic requirements for the use of DD are a multilevel data set to be analyzed and natural context variables of substantive interest (other than the cluster mean) measured using the same units as a lower level variable. In this section, we provide some conceptual examples and motivation, and

then in the next section, we present our own example analytically in some detail.

DD will likely be particularly useful in longitudinal studies. For example, in a hypothetical clinical intervention, a researcher might be interested in the effect of perceived social support on depression over the course of treatment. In our hypothetical experiment, the researcher has daily questionnaire data detailing a person's perception of social support and their levels of depression. This is a multilevel data set with observations nested in persons. A DD approach may compare the effects of baseline (e.g., day 1) levels of social support, the mean level of social support minus the baseline (the mean overall change from baseline for a given person), and day-to-day changes in perceived social support. This analysis could be used to determine if initial levels of social support are more important than changes in social support over the course of depression treatment. This could be modeled as an interaction effect between time in treatment and each social support component (three interaction effects). In addition, the researcher could see the effect of day-to-day changes on perceived social support and depression, and the effect of baseline and mean change in social support on mean depression scores.

Plewis (1989) raised the issue of context variables unrelated to the mean, and suggested that the mode or other measures of central tendency might be useful. His suggestion can be applied to a DD model while retaining the benefits of cluster mean centering. Consider a model predicting student achievement test (SAT) scores from parental income. The median (or modal) income could be included at the classroom level alongside the mean deviation from the median (or mode); within-class incomes would still be a given student's deviation from the class mean income (as in Raudenbush (1989)). Income is used in this example because it is well documented that at the population level, incomes are not normally distributed (Bandourian, McDonald, & Turley, 2003), and a researcher might be motivated to use a measure of central tendency less affected by skewness. To be clear, a lack of normality in data may introduce its own difficulties, and should be dealt with accordingly. This example is meant only to illustrate a potential application of DD and is not meant as a way to deal with nonnormal data.

Finally, the empirical example in the next section is a longitudinal application. This example uses family data and decomposes year of testing into seven components across three levels. There are four natural variables, age at testing, year of birth relative to the oldest sibling, mother's age at first birth, and mother's birth year. Age at testing is cluster mean centered within children and within families, and the birth year variable is centered within families. The next section details how this decomposition allows for certain hypotheses to be explicitly addressed that could not otherwise be evaluated.

An empirical example: DD of the Flynn effect

The Flynn effect is the name given to the systematic rise in IQ scores over time, which has occurred at a rate of approximately three IQ points per decade on standard IQ tests (Flynn, 1987; Pietschnig & Voracek, 2015). The Flynn effect is primarily associated with fluid intelligence measures, which assess problem solving and real-world reasoning; smaller Flynn effect patterns have occasionally been observed in crystallized intelligence, associated with verbal facility and memory. A recent meta-analysis identified the Flynn effect as a roughly linear relationship between IQ and time, although with identifiable periods of deviation from linearity (Pietschnig & Voracek, 2015). Many different hypotheses have been proposed to explain the Flynn effect; previous researchers have suggested that steps should be taken to evaluate and reduce the number of plausible hypotheses (Mingroni, 2014; Rodgers, 1998). For the present analysis, we use the DD model to identify the location of the Flynn effect within a broad and flexible data set, and in doing so eliminate particular classes of hypotheses and allow focus on other classes that remain logically plausible.

Within the design structure of the NLSYC, the causal processes underlying the Flynn effect can emerge from three possible locations: within-person, within-family and between-family. A within-person process would be caused by systematic processes leading to changes throughout a person's lifetime. Hypotheses such as increased exposure to tests (Tuddenham, 1948), improvements in niche picking (Dickens & Flynn, 2001) and slowed life history (Woodley, 2012) could plausibly cause within-person increases. In niche picking, for example, a person's ability to pick an intellectually facilitating niche would likely improve over the course of their life, both for developmental reasons and because of gradual improvements in society. This improved ability to pick a niche would act on a person's cognitive ability, which could appear as a within-person effect on fluid intelligence. Within-family effects of interest would primarily be those due to birth cohort differences between siblings, or natural birth-order effects. Improvements in neonatal nutrition (Lynn, 2009) is one example of a birth cohort effect. Between-family effects are the ones that vary between families, but not within families. Heterosis (Mingroni, 2007) is an example of a between-family hypothesis. Heterosis, or hybrid vigor, is the idea that as human mating has expanded geographically, increases in genetic variability have provided children with genetic advantages, including in fluid intelligence (Mingroni, 2007; see Woodley (2011) for criticisms). An alternative

between-family hypothesis that might be considered is that increases in maternal education allow for mothers to provide a better cognitive environment for their children.

The developers of some of these hypotheses might disagree with our classification, and it is worth noting that hypotheses operating at lower levels can naturally filter to higher levels. For example, if niche picking occurs partly as a function of parental guidance in the child's niche picking, that effect would also show up as a between-family process. Lower level hypotheses must generally add a constraint to move up a level. Higher level hypotheses cannot easily move down to operate at lower levels. It would be logically difficult, for example, to imagine genetic effects such as heterosis having systematic within-family effects on a meaningful scale.

The data

The NLSYC (Bureau of Labor Statistics, 2012) provides a nearly ideal data set to test the Flynn effect hypotheses and apply DD. This data set comprises biennial observations of all biological children of the approximately 6500 women surveyed in the National Longitudinal Survey of Youth-1979 (NLSY79; Bureau of Labor Statistics, 2012) sample. There are a number of demographic, economic, behavioral and cognitive measures collected in each round of the NLSYC. The survey began in 1986 and continues till the present. At the time of this analysis, data were available through the 2012 survey year. In its initial years, the NLSY79 (maternal) sample included an oversample of poor whites, minorities and military personnel; however, the military and poor white oversamples were subsequently dropped due to funding constraints. The overall data sets (including the oversample) were used here. There have been approximately 11,500 children born to the NLSY79 females, and who have taken part in the NLSYC.

For the present analysis, only families with at least two children born in different years are included, for two reasons. The first is because of cluster mean centering at the family level. If families with single children were used, single children would be given scores of 0 on family mean-centered time variables, roughly equivalent to those of middle children in larger families (presuming children are approximately evenly spaced). Given the plausible existence of meaningful differences between families with only one child and larger families, as well as plausible differences between only children and middle children, the decision was made to drop only children and avoid conflating them with middle children and likewise avoid conflating families with single children and multiple children. The second reason families with only children were omitted was to ensure variability within all families to identify potential within-family patterns. The resulting data set has 2881 families, with 7822 children, almost all of whom have been measured multiple times across the longitudinal survey process, for a total of 29,921 observations. There was a mean of 10.39 observations per family ($SD = 3.83$) and a mean of 3.83 observations per child ($SD = 1.22$).

Measures

The measure used in our analysis is the PIAT-Math subscale (Dunn & Markwardt, 1970). This measure is particularly well suited to studies of the Flynn effect for two reasons. First, the Flynn effect has already been observed in the PIAT-Math in NLSYC data through 2000 (Rodgers & Wänström, 2007; also see Ang, Rodgers, & Wänström (2010) for the replication of these findings in gender, race, and urbanicity subsamples). The second is that the PIAT-Math was designed to test children's ability to apply math concepts in the real world (Dunn & Markwardt, 1970), making it an excellent measure of fluid, or problem solving, intelligence (Flynn, 2000). PIAT-Math administrations occurred between the ages of six and 14 in the NLSYC. The same 1968 version of the PIAT-Math has been administered since the beginning of the study in 1986. The PIAT-Math standardized scores have a mean of 100 and a standard deviation of 15, and the Flynn effect previously identified in the PIAT-Math was approximately equal in a raw IQ scale to the Flynn effect using other measures of IQ (Rodgers & Wänström, 2007). These standardized scores are used here to allow for comparisons across ages.

A control variable for mother's cognitive ability, the mother's Armed Forces Qualifying Test (AFQT), was calculated from their Armed Services Vocational Aptitude Battery (Ree, Mathews, Mullins & Massey, 1982). The AFQT is a test designed for adults, so rescaling due to age was necessary for NLSY79 mothers, who ranged in age from 15 to 23 when the AFQT was administered in 1980. The score used here was a version created by the NLS staff that adjusts for age differences among the mothers taking the test. Note that it is plausible (even likely) that there exists a Flynn effect in the mother's AFQT scores themselves. It should be noted however that a Flynn effect in mother's AFQT scores would likely bias the Flynn effect in children downward because the AFQT scores would have variance associated with time as well as cognitive ability adjusting out some of the time effect on children's PIAT-Math scores.

Time, measured in years, is the variable of primary concern in nearly all Flynn effect research. In this DD

analysis, time is the independent variable to be decomposed. In the context of a longitudinal family study, such as the NLSYC, time can be partitioned into at least four "natural" parts: age of the child at testing, the difference between a given child's birth year and the oldest child's birth year, mother's age at first birth, and mother's birth year. These components correspond to within-person (child's age at testing), within-family (child birth year) and between-family (mother's age at first birth and mother's birth year) components. The between-family component could be the year that the first child was born, summing up the mother's year of birth and the mother's age at first birth. However, there are two advantages of splitting this variable. First, it presents a good demonstration of how a variable may have more than two "natural" contextual components at a given level. Second, mother's age at first birth has often been an effective between-family variable in past demographic and behavior genetic studies (Neiss, Rowe, & Rodgers, 2002; Rodgers et al., 2008), and has often been more informative of family differences than the year of the mother's birth.

Within this example, there are four "natural" variables; of those four, two need to be further mean-centered. Age is mean-centered within person and within family, and the differences in sibling birth years are centered within family. The resultant cluster means are then reintroduced to the model. This decomposition into age, birth year, and family components allows us to test multiple hypotheses concerning the Flynn effect simultaneously. For example, if the Flynn effect is due to ongoing changes in a person's overall environment (e.g., improvements to the family, to educational settings, etc.) it should appear as a within-person change, as a part of the individual age effect. If the Flynn effect is due to changes between birth cohorts, it should appear as the within-family effect associated with the differences in birth years. If the Flynn effect is due to between-family differences, it should appear in one or more of the multiple between-family components for time.

DD is necessary to adequately distinguish between the different hypotheses tested here. For example, consider that if simple cluster mean centering was used and a within-family effect was found, this result would not allow us to distinguish the between-effect cause by birth cohort differences and age differences at the within-family level. Furthermore, without cluster mean centering, it is impossible to test that the effect for age or birth cohort is consistent across levels. Furthermore, many (most) of the natural/contextual variables provide substantive interpretations. DD therefore supports distinguishing between and properly testing the hypotheses.

The goal of the current DD analysis is not to test specific Flynn effect hypotheses per se, but rather to identify which slopes are statistically different from zero, their magnitude and direction. Such an analysis is pseudo-exploratory. The mathematical model to be fit is clearly defined, as is the sample, and we expect at least some of the slopes to be positive, because we know that there is a Flynn effect within the NLSYC data (see Rodgers & Wänström (2007)). Additionally, we know that the Flynn effect has generally been observed to be approximately three points per decade in a standard IQ metric (Pietschnig & Voracek, 2015), giving a "ballpark" estimate of the expected effect sizes. (We note that because our PIAT-Math measure is a more pure form of fluid intelligence than that contained in a standard IQ metric, which combines fluid and crystallized components, we might expect our effect sizes to be a bit larger than .3 points per year.) However, we have no a priori hypothesis about which slopes in particular will be statistically significant and/or have meaningful effect sizes, as contributions to the Flynn effect. We do have constraints on which slopes must be significant for a particular time component (e.g., age) to be considered as a candidate for the Flynn effect. Results here should be explicitly tested in other data sets, using DD as previously outlined. The NLSY97 data set (an approximate 18-year replication of the NLSY79) contains PIAT-Math scores, and is a data source that could be used for such a replication.

Analysis

There are two analytic models used. The first is a CMC model, and the second is a DD model. The CMC model is presented for comparison purposes; the primary analysis relies on the DD model. In addition to the time component present in both models, the mother's AFQT score, standardized by age, is included as a control variable to account for maternal cognitive ability in both models. The rationale for including maternal IQ differences in the model is that not quite all NLSYC respondents have reached the age of 15, and so there are selection effects due to the systematic differences between children born to younger and older mothers. Including maternal AFQT scores within the model adjusts for this selection bias (see Rodgers & Wänström (2007) and Ang et al. (2010) for further discussion and examples). We also note, however, that maternal IQ is a between-family variable that can function as more than just a simple control variable. First, in addition to controlling for the selection bias, maternal IQ may partially control for genetic similarity between mothers and their children (see Rodgers, Rowe, & May (1994), for a heritability analysis of the NLSYC, including PIAT-Math, which shows moderate levels of heritability in the NLSY cognitive ability measures). Second, maternal IQ is a between-family measure,

and may also be a part of the explanatory process underlying the Flynn effect. Thus, though we treat it primarily as a variable controlling for the selection bias, we also keep in mind that it can function as a substantively important measure in defining the location of the Flynn effect as well.

The true difference between the models is in the treatment of the time variable. In the CMC model (Model 5; Table 7), there are three time components included in the model. First, $\ddot{year}_{ijk}$ denotes the person mean centered level-1 component (Model 5, Equation [1]). The level-2 within-family component, denoted as $\ddot{year}_{.jk}$, is the family mean-centered person mean year. (Model 5, Equation [2]). The level-3 between-family component denotes the family mean year, $\overline{year}_{..k}$. In both the CMC and the DD models, all random effects were included and allowed to co-vary. In the DD model, the random effects' model fit better, $\chi^2_{11} = 796.24, p < .001$, so random effects were retained in both models to facilitate comparisons. Tables of the random effects are provided in the Results section but they are not the primary focus of this analysis. In the CMC model, the family mean-centered person mean year was used to predict the person intercept, β_{0jk} (Model 5, Equation [4]), and the family mean year of testing and mother's AFQT score were used to predict the family intercept, θ_{00k} (Model 5, Equation [5]). The model with cluster mean centering, fixed and random effects, and their covariances is presented in Table 7. $Math_{ijk}$ is the age-standardized PIAT-Math measure, and $year_{ijk}$ is the year in which a respondent took the PIAT-Math test.

The DD model (Table 8) is naturally more complex than the CMC model and has seven time components included in the model. Age, the measure of respondents' age at testing within the context of a given level, is split into three components: $\dddot{Age}_{ijk}$, $\ddot{Age}_{.jk}$, and $\overline{Age}_{..k}$ (Model 6, Equations [1] and [2]). As established previously, the level-1 components of a DD and CMC model are identical, so $\dddot{Age}_{ijk} = \ddot{year}_{ijk}$. The other two components for age are the family mean-centered person mean age and the family mean age, respectively. The sibling birth year variable is split into two components: $\ddot{SiblingBirthYear}_{jk}$ and $\overline{SiblingBirthYear}_{.k}$ (Model 6, Equation [3]). $\ddot{SiblingBirthYear}_{jk}$ denotes the family mean-centered component, and $\overline{SiblingBirthYear}_{.k}$ denotes the family mean. Because the two family mean-centered person fixed components of age and sibling birth year decomposed the family mean-centered person mean year of test, $\ddot{SiblingBirthYear}_{jk} + \ddot{Age}_{.jk} = \ddot{year}_{.jk}$. The final two time components denote the mother's age at first birth, $MotherAgeFirstBirth_k$, and mother's birth year, $MotherBirthYear_k$. As with the level-2 components, these two family level components combine with the family mean age and the family mean sibling birth year components such that $MotherBirthYear_k + MotherAgeFirstBirth_k + \overline{Sibling\ Birth\ Year}_{.k} + \overline{Age}_{..k} = \overline{year}_{..k}$.

From this example, it is clear how multiple valid contextual variables can be contained within a cluster mean variable, or in other words, how a cluster mean can be further decomposed within the DD approach. From this understanding, the full model flows directly from the CMC model. While $\overline{year}_{..k}$ appears in the CMC model, it is replaced by the four family-level components in the DD model, each component having its own slope. A similar substitution occurs for $\ddot{year}_{.jk}$, although here the slopes also have their own random component as well. The level-1 components $\dddot{Age}_{ijk}$ and $\ddot{year}_{ijk}$ are identical and interchangeable. The final DD model in mathematical notation is presented in Model 6 (Table 8).

Results

Models were fit in Mplus using the MLR robust standard errors. Both sets of results are presented to demonstrate and compare the differences and similarities between the CMC model and the DD model. The CMC model results are presented first.

The mean PIAT-Math score was 100.42 with a standard deviation of 14.42 across 29,921 observations. The proportion of variance at level 3, between families, was .36, and between children, it was .24; all remaining variance (.40) was within person at level 1. For the purpose of the demonstrations in this paper, fixed and random slopes provided approximately equivalent results; however, the random slopes provided a statistically significantly better fit.

CMC analysis

A summary of the fixed effects is presented in Table 9; a summary of the random effects is presented in Table 10. The CMC model had an AIC of 226,983.9 and a BIC of 227,108.5. At the between-family level, a one-point increase in mother's AFQT score contributed approximately two-tenths of a point to the family mean PIAT-Math score. The between-family component of time had a slope similar in size to those previously found for the Flynn effect with approximately one-third of a point increase in the family mean PIAT-Math scores for each additional year. The within-family slope for time was statistically significant, but one-third of that normally found in the Flynn effect literature. The within-person time component had a slope roughly half of what is expected from previous Flynn effect research, slightly larger than the within-family slope. Using pseudo-standardized coefficients (Hoffman, 2014 p. 342) allows some comparison

Table 7. Cluster mean centered Flynn effect model.

Model 5	
eq1: centering of level-1 variable, year	$\ddot{year}_{ijk} = year_{ijk} - \ddot{year}_{.jk} - \overline{year}_{..k}$
eq2: centering of level-2 cluster mean variable, year	$\ddot{year}_{.jk} = \overline{year}_{.jk} - \overline{year}_{..k}$
eq3: Level 1	$Math_{ijk} = \beta_{0jk} + \beta_{1jk} * \ddot{year}_{ijk} + e_{ijk}$
eq4: Level-1 intercept	$\beta_{0jk} = \theta_{00k} + \theta_{01k} * \ddot{year}_{.jk} + u_{0jk}$
eq5: Level-2 intercept	$\theta_{00k} = \gamma_{000} + \gamma_{001} * \overline{year}_{..k} + \gamma_{002} * AFQT_k + u_{00k}$
eq6: Level-2 slope for family mean centered, child mean year	$\theta_{01k} = \gamma_{010} + u_{01k}$
eq7: Level-1 slope for child mean centered year of testing	$\beta_{1jk} = \theta_{10k} + u_{1jk}$
eq8: Level-2 slope for child mean centered year of testing	$\theta_{10k} = \gamma_{100} + u_{10k}$
eq9: Reduced form	$Math_{ijk} = \gamma_{000} + \gamma_{100} * \ddot{year}_{ijk} + \gamma_{010} * \ddot{year}_{.jk} + \gamma_{001} * \overline{year}_{..k} + \gamma_{002} * AFQT_k + u_{10k} * \ddot{year}_{ijk} + u_{1jk} * \ddot{year}_{ijk} + u_{01k} * \ddot{year}_{.jk} + u_{00k} + u_{0jk} + e_{ijk}$
Level-1 residual distribution	$e_{ijk} \sim N(0, \sigma^2)$
Leve-2 residual distribution	$U_2 \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau^2_{00} & \\ \tau_{10} & \tau^2_{11}\end{bmatrix}\right)$
Level-3 residual distribution	$U_3 \sim N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\tau^2_{000} & & \\ \tau_{100} & \tau^2_{110} & \\ \tau_{200} & \tau_{210} & \tau^2_{220}\end{bmatrix}\right)$

Table 8. Double decomposition Flynn effect model.

Model 6	
eq1: centering of level-1 variable, age	$\ddot{Age}_{ijk} = Age_{ijk} - \ddot{Age}_{.jk} - \overline{Age}_{..k}$
eq2: centering of level-2 cluster mean variable, child mean age	$\ddot{Age}_{.jk} = \overline{Age}_{.jk} - \overline{Age}_{..k}$
eq3: centering of level-2 context variable, sibling birth year	$\ddot{SiblingBirthYear}_{jk} = SiblingBirthYear_{jk} - \overline{SiblingBirthYear}_{.k}$
eq4: Level 1	$Math_{ijk} = \beta_{0jk} + \beta_{1jk} * \ddot{Age}_{ijk} + e_{ijk}$
eq5: Level-1 intercept	$\beta_{0jk} = \theta_{00k} + \theta_{01k} * \ddot{Age}_{.jk} + \theta_{02k} * \ddot{SiblingBirthYear}_{jk} + u_{0jk}$
eq6: Level-2 intercept	$\theta_{00k} = \gamma_{000} + \gamma_{001} * \overline{Age}_{..k} + \gamma_{002} * \overline{SiblingBirthYear}_{.k} + \gamma_{003} * MotherAgeFirstBirth_k + \gamma_{004} * MotherBirthYear_k + \gamma_{005} * AFQT_k + u_{00k}$
eq7: Level-1 slope for child mean centered age	$\beta_{1jk} = \theta_{10k} + u_{1jk}$
eq8: Level-2 slope for child mean centered age	$\theta_{10k} = \gamma_{100} + u_{10k}$
eq9: Level-2 slope for family mean centered child mean age	$\theta_{01k} = \gamma_{010} + u_{01k}$
eq10: Level-2 slope for family mean centered sibling birth year	$\theta_{02k} = \gamma_{020} + u_{02k}$
eq11: reduced form	$Math_{ijk} = \gamma_{000} + \gamma_{001} * \overline{Age}_{..k} + \gamma_{002} * \overline{Sibling\ Birth\ Year}_{.k} + \gamma_{003} * Mother\ Age\ First\ Birth_k + \gamma_{004} * Mother\ Birth\ Year_k + \gamma_{005} * AFQT_k + \gamma_{100} * \ddot{Age}_{ijk} + \gamma_{010} * \ddot{Age}_{.jk} + \gamma_{020} * \ddot{Sibling\ Birth\ Year}_{jk} + u_{10k} * \ddot{Age}_{ijk} + u_{1jk} * \ddot{Age}_{ijk} + u_{01k} * \ddot{Age}_{.jk} + u_{02k} * \ddot{SiblingBirthYear}_{jk} + u_{00k} + e_{ijk}$
Level-1 residual distribution	$e_{ijk} \sim N(0, \sigma^2)$
Level-2 residual distribution	$U_2 \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau^2_{00} & \\ \tau_{10} & \tau^2_{11}\end{bmatrix}\right)$
Level-3 residual distribution	$U_3 \sim N\left(\begin{bmatrix}0\\0\\0\\0\end{bmatrix}, \begin{bmatrix}\tau^2_{000} & & & \\ \tau_{100} & \tau^2_{110} & & \\ \tau_{200} & \tau_{210} & \tau^2_{220} & \\ \tau_{300} & \tau_{310} & \tau_{320} & \tau^2_{330}\end{bmatrix}\right)$

Table 9. Fixed effects from the cluster mean centered Flynn effect model.

Level	Variable	Slope	Pseudo-standardized slope	Standard error	Slope/S.E.	p Value, 2 tailed
Within-person	Year	0.16	0.05	0.26	6.03	<.001
Within-family	Year	0.11	0.05	0.03	3.59	<.001
Between-family	Year	0.33	0.27	0.03	10.27	<.001
	AFQT	0.20	0.98	0.01	34.62	<.001

Table 10. Random effects from the cluster mean centered Flynn effect model.

Level-2 variances and correlations		
Intercept	Slope for level-1 time component	
48.76***		
.29***	.87***	
Level-3 variances and correlations		
Intercept	Slope for level-1 time component	Slope for level-2 time component
33.57***		
.17**	.43***	
.15	.13	.21***

***$p \leq .001$.
**$p \leq .01$.

between the effects. The between-family effect is nearly six times the size of the other two time effects.

Substantively, the CMC model suggests that future research on the causal process of the Flynn effect might be best focused on the family level. However, a case could be made for both within-family and within-person processes because both these slopes were statistically significant as well. However, the lower level slopes were much smaller in effect size magnitude than the traditional magnitude of the Flynn effect, and their pseudo-standardized effects were substantially smaller than the between-family effect. The DD model that we will present next will better assist us in pinpointing the location of the Flynn effect in the data.

DD analysis

A summary of the fixed effects is presented in Table 11; a summary of the random effects is presented in Table 12. The DD model had an AIC of 226,879.6, and a BIC of 227,070.6, both lower than the CMC model. Of particular note, at the between-family level, the mother's AFQT had a nearly identical effect size as in the previous model. Mother's age at first birth was associated with a nearly half-point increase in the family average PIAT-Math score for every additional year in which the mother waited to have her first child. This effect was, by far, the largest effect found in both the CMC and the DD analyses presented, and more than double of the raw slope of any other effect in the DD analysis. However, although the slopes for the time effects can be compared in the sense that they all measure what would happen for a one-year increase in a given time-related variable, the slopes are not comparable from a statistical standpoint because the scale of the variables is different (i.e., the variance of the variables is different, and the variance of the outcome is different across levels). Converting slopes into pseudo-standardized slopes (Hoffman, 2014 p. 342), the effect of mother's age at first birth was eight times as large as the next largest time-related effect. Taken as a whole, this result strongly suggests that mother's age at first birth is the most important time-related effect included in this model.

The sibling birth year variable was broken into two components. The first varied between families and denotes the mean number of years children in a given family were born after the oldest child in that same family weighted by the number of observations for each child. Results suggested that families who were (on average) born later did not differ from families born earlier. The second component of this variable was within families and was a given child's deviation from their family's mean. This component was associated with an approximately .1 point increase on the average PIAT-Math score for each additional year.

The last "natural" variable included was the age of the child at testing, which was split into within-person, within-family, and between-family components. The between-family component was the average age at which children took the test in a given family, weighted by the number of observations for each child. This slope was not statistically significant, suggesting no meaningful differences between families whose children were older on average compared to families whose children were younger on average. The next component was within families, and was the deviation of the child's mean age from their family's mean age. This fixed component of this slope was just significant using a one-tailed test. This result suggests that children with an average age one year greater than their siblings would have a roughly .2 point advantage over their siblings on the PIAT-Math score. The last component denoted the deviations of a child's age at a given observation from their mean age across tests. As in the CMC model, this slope was approximately half of the magnitude expected for the Flynn effect.

Overall, the DD model allows us to conclude that not only the Flynn effect is largely a family-level process, but it also appears to be closely related specifically to mother's age at first birth. This is a stronger and more precise claim than we were able to make with the CMC model. It is difficult to argue, from the results of the DD model, that the Flynn effect is related to an aging process; older families were not statistically significantly different from younger families, and older children were only marginally so. Likewise, it is difficult to argue for a birth cohort process because families that were born later were not significantly different from families born earlier.

Discussion

We have presented an alternative method for cluster mean centering variables in multilevel analyses. This

Table 11. Fixed effects from the doubly decomposed Flynn effect model.

Level	Variable	Slope	Pseudo-standardized slope	Standard error	Slope/S.E.	p Value, 2 tailed
Within-person	Age	0.16	0.05	0.03	5.95	<.001
Within-family	Age	0.19	0.02	0.11	1.67	0.09
	Sibling birth year	0.10	0.05	0.03	3.21	<.001
Between-family	Age	0.11	0.02	0.15	0.74	0.46
	Sibling birth year	0.08	0.04	0.06	1.22	0.22
	Mother's age at first birth	0.48	0.41	0.04	13.05	<.001
	Mother's birth year	− 0.03	− 0.01	0.07	− 0.46	0.65
	Mother's AFQT	0.18	0.89	0.01	30.55	<.001

method results in variables that can be linearly combined to give CMC variables, resulting in potentially more explanatory and more interpretable models. The DD model retains many of the benefits of cluster mean centering, particularly the orthogonality of higher and lower level variables (Raudenbush, 1989), but allows for the consideration of contextual variables beyond the cluster mean such as those suggested by previous authors (Plewis, 1989). We demonstrated the potential effects of failure to use the DD model in a brief simulation study. We then used this method to analyze the Flynn effect in an attempt to winnow the field of plausible explanatory hypotheses.

The simulation study was presented as a proof-in-principle. Future analysis should look at the effects of different magnitudes and directions of the relationships between variables at the second level of analysis. Furthermore, research should also examine the effects on power and the type-one error rate. Finally, because a single variable is being split into three components, it will be important to examine the effects of multiplicity. Given the results of the present study, particular care should be taken when researchers do not use the DD model and include the natural context variables we have discussed alongside cluster means.

Table 12. Random effects from the doubly decomposed Flynn effect model.

Level-2 variances and correlations			
Intercept	Slope for level-1 age effect		
46.27***			
.30***	.87***		
Level-3 variances and correlations			
Intercept	Slope for level-1 age effect	Slope for level-2 age effect	Slope for level-2 birth year effect
32.45***			
.16**	.43***		
.20	.63***	1.84**	
.18*	.021	− .01	.20***

***$p \leq .001$.
**$p \leq .01$.
*$p < .05$.

When analyzing the Flynn effect, the first model fit was a standard CMC model. This model was of limited use in the present study. The hypotheses regarding the Flynn effect tend to work either across everyone every year, across birth cohorts, or between families. A between-family variable is present in this analysis but it is a mixture of between-family components of birth cohort and age as well as the mother's birth year and age at first birth. Birth cohort effects are not explicitly modeled and age effects are mixed at higher levels with other between-person and between-family effects. However, we know that at level 1, the person mean-centered year is identical in value to the person mean-centered age, and thus that variable is present in both models. Overall, the highest slope, at .33, was for the between-family effect and is in line with the general magnitude of the effect size associated with the Flynn effect. The other two slopes are weaker, but also in the expected direction. This model is of limited utility in narrowing the potential causes of the Flynn effect because many different hypotheses are consistent with the patterns of results (although CMC results would appear to tilt in the direction of a between-family explanation).

Utilizing DD avoids some of the pitfalls of the CMC model, and allows us to interpret the results more precisely. The results do not support an effect for the mother's year of birth, but do support an effect for mother's age at first birth (the strongest effect size that emerged from this analysis). Based on the other results, this leads us to conclude that the location of the Flynn effect is primarily in the between-family part of the NLSYC data. We want to be cautious in our interpretation, however. Although the results support that the Flynn effect is primarily a between-family effect and that the mother's age at first birth is closely related to the Flynn effect, we cannot simply state that the mother's age at first birth is the causal explanation of the Flynn effect. Rather, we believe it to be related to the true causes of the Flynn effect, but any explanation of the effect is likely related to a number of disparate factors.

The effect of birth year was significant within families but not between families. Such a result is not fully consistent with an overall birth cohort effect, however.

The lack of a family level effect indicates that families whose birth years were later on average did not have higher PIAT-Math scores. An explanation for the small within-family effect is not immediately apparent. The present birth cohort effect findings are also inconsistent with much of the research on birth-order effects (Zajonc & Sulloway, 2007), although previous research has suggested that the birth-order effect may be spurious (Wichman, Rodgers, & MacCallum, 2006, 2007), or a byproduct of the Flynn effect (Rodgers, 2014); however, the literature related to birth-order effects has a long history, and the treatment is beyond the scope of the present study.

The effect for age was significant within person, marginally significant within families, and not significant between families. This final result is inconsistent with a Flynn effect due to age, or more precisely due to a constant effect on all individuals, for reasons similar to that for the birth cohort effect. As with the birth cohort effect, the lack of a statistically significant effect at the family level is difficult to interpret. If there was truly an effect on individuals across time (the age effect), families that were on average older ought to have higher means than families who are younger, all else equal. Similarly to the birth cohort effect, the age effect, where significant, was substantially smaller than the overall Flynn effect, or the maternal age at first birth effect. Unlike the within-family birth cohort effect, an explanation of the age effect within person is readily available. A longitudinal study such as the NLSYC provides repeated measures of the same individual, and it is likely that a small but significant practice effect may be occurring because of repeated administration of the PIAT-Math score. In summary, the most strongly supported class of hypotheses to explain the Flynn effect was the between-family class, in particular those related to maternal IQ and maternal age at first birth variables.

The purpose of presenting this example was to highlight the utility of DD, and the results fully justify it. Based on these results, it would appear that the Flynn effect emerges primarily from between-family processes, and furthermore, given the effect of mother's age at first birth and IQ, factors specific to parents are likely causal candidates. This result would suggest a rather substantial re-orientation in thinking about the Flynn effect, because most previous work has focused on the child and the individual, including recent calls to focus on within-family explanations (Mingroni, 2014; Sundet, 2014); these results suggest that a re-focus on the role of the mother/parents, and how they impact intellectual development within the family, would be appropriate.

The Flynn effect analysis, as an illustration of the DD model, is meant to demonstrate the utility of DD, as an expansion of cluster mean centering, in a setting in which there are many potential sources of interesting variance. In the CMC model, there are natural contextual variables of clear substantive importance (e.g., mother's age at first birth) that are masked when using cluster mean centering. Using DD allows for a more nuanced examination of potential effects. In the CMC model, there were significant effects at all levels, but it was impossible to determine how the results matched different sources of explanation, in relation to past theory. To its credit, the CMC did show that lower level effects were substantially smaller than both the level-3 effect and the Flynn effect generally (Pietschnig & Voracek, 2015). Conversely, in a model based on uncentered "natural" variables, it would be impossible to know if the results were due to consistent effects across levels, or alternatively if effects that should be present were lacking as indicated in the final results. The DD model overcomes both shortcomings. In the DD model, it was apparent that the results at level 3 were not due to the effect of lower level variable cluster means. This lack of effect for cluster means ruled out the lower level variables because their effects were not consistent across levels. This is a question that could only be adequately answered using an approach like the DD model.

Beyond its application to the Flynn effect presented here, DD is more generally meant to allow researchers the flexibility to examine the effects of cluster contexts at a level of detail not provided by cluster means. The example applications provided, such as baseline effects, cluster modes, and the Flynn effect analysis are examples that the authors were readily able to develop. Presumably, other researchers will identify additional useful applications, as they identify natural contextual variables relevant to their own research.

There are of course limitations to this method. For example, Plewis (1989) suggested the cluster standard deviation as one possible context variable; this context variable may be important in some applications, but it is difficult to see how it would be implemented via DD. The resulting variables would be the standard deviation of a cluster, the mean deviation of scores from the standard deviation of that cluster, and deviations within cluster from the cluster mean. Although those variables might be mathematically sound, it is difficult to grasp the substantive meaning of the mean difference of a cluster from that cluster's standard deviation. In summary, although many contextual variables that were not accommodated by CMC models are accommodated by DD models, some contexts are beyond the application of DD models. Two other limitations should be considered by researchers. First, there is a tradeoff in model parsimony; if a simpler model serves the research purpose, it should be used. However, as with any model, differing models should be compared to evaluate meaningful statistical differences. The second issue is that the within-level effects may be correlated. This, however, is a concern

for any two variables included in any model at the same level.

Future research regarding the centering method should examine its effects on random components of multilevel models in comparison to more typical cluster mean centering. The effects of DD on power and type I error should also be examined. Future research regarding the Flynn effect should apply this decomposition to other Flynn effect data sets to evaluate whether the between-family results found in the NLSYC are consistent across samples. Our research is not the first to use family data to study the Flynn effect and especially the large samples of Norwegian conscript data (Sundet, 2014) may be well adapted to the present method. It is hoped that DD will prove similarly useful in other multilevel modeling applications.

Article information

Conflict of interest disclosures: Each author signed a form for the disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported by a grant.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank Dr. Kris Preacher for a comment early in the course of this research that helped in motivating some of their thinking.

The authors would like to thank Conor Dolan, Ellen Hamaker, and Lesa Hoffman for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and the endorsement by the authors' institution is not intended and should not be inferred.

References

Aiken, L., & West, S. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage, Newbury.

Ang, S., Rodgers, J. L., & Wänström, L. (2010). The Flynn effect within subgroups in the U.S.: Gender, race, income, education, and urbanization differences in the NLSY-children data. *Intelligence, 38*(4), 367–384. https://doi.org/10.1016/j.intell.2010.05.004.

Bandourian, R., McDonald, J. B., & Turley, R. S. (2003). Income distributions: an inter-temporal comparison over countries. *Estadistica*, *55*(1), 135–152. https://doi.org/10.2139/ssrn.324900.

Bureau of Labor Statistics, U.S. Department of Labor. National Longitudinal Survey of Youth 1979 cohort, 1979-2012 (rounds 1-25). (2012). *Produced and distributed by the center for human resource research*. Columbus, OH: The Ohio State University.

Bureau of Labor Statistics, U.S. (2012). Department of labor, and national institute for child health and human development. Children of the NLSY79, 1979-2012. *Produced and distributed by the center for human resource research*. Columbus, OH: The Ohio State University.

Cronbach, L. J., & others. (1976). *Research on classrooms and schools: Formulation of questions, design and analysis*. Retrieved from http://eric.ed.gov/?id = ED135801.

Cronbach, L. J., & Webb, N. (1975). *Between-class and within-class effects in a reported aptitude* treatment interaction: Reanalysis of a study by GL Anderson*. https://doi.org/10.1037/0022-0663.67.6.717. Retrieved from http://psycnet.apa.org/journals/edu/67/6/717/.

Curran, P. J., Lee, T. H., Howard, A. L., Lane, S. T., & MacCallum, R. C. (2012). Disaggregating within-person and between-person effects in multilevel and structural equation growth models. In J. Harring (Ed.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 217–253). Charlotte, NC: Information Age Publishing.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review, 108*(2), 346–369. https://doi.org/10.1037/0033-295X.108.2.346.

Dunn, L. M., & Markwardt, F. C. (1970). *Peabody individual achievement test manual* (1st ed.). Circle Pines, MN: American Guidance Service, Inc.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*(2), 171. https://doi.org/10.1037/h0090408.

Flynn, J. R. (2000). IQ gains and fluid g. *American Psychologist, 55*(5), 543. https://doi.org/10.1037/0003-066X.55.5.543.

Hoffman, L. (2007). Multilevel models for examining individual differences in within-person variation and covariation over time. *Multivariate Behavioral Research, 42*(4), 609–629. https://doi.org/10.1080/00273170701710072.

Hoffman, L., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development, 6*(2–3), 97–120. https://doi.org/10.1080/15427600902911189.

Hoffman, L. (2014). *Longitudinal Analysis: Modeling within-person fluctuation and change*. New York, NY: Routledge. https://doi.org/10.4324/9781315744094.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781849209366.

Kreft, I. (1995). *The effects of centering in multilevel analysis: Is the public school the loser or the winner*? Retrieved from http://eric.ed.gov/?id = ED392837.

Kreft, I., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30*(1), 1–21. https://doi.org/10.1207/s15327906mbr3001_1.

Lynn, R. (2009). What has caused the Flynn effect? Secular increases in the development quotients of infants. *Intelligence, 37*(1), 16–24. https://doi.org/10.1016/j.intell.2008.07.008.

Mingroni, M. A. (2014). Future efforts in Flynn effect research: Balancing reductionism with holism. *Journal of Intelligence, 2*(4), 122. https://doi.org/10.3390/jintelligence2040122.

Mingroni, M. A. (2007). Resolving the IQ paradox: Heterosis as a cause of the Flynn effect and other trends. *Psychological Review, 114*(3), 806–829. https://doi.org/10.1037/0033-295X.114.3.806.

Neiss, M., Rowe, D. C., & Rodgers, J. L. (2002). Does education mediate the relationship between IQ and age of first birth? A behavior genetic analysis. *Journal of Biosocial Science, 34*, 259–275. https://doi.org/10.1017/s0021932002002596.

Paccagnella, O. (2006). Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. *Evaluation Review, 30*(1), 66–85. https://doi.org/10.1177/0193841X05275649.

Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains: A formal meta-analysis of the Flynn effect (1909–2013). *Perspectives on Psychological Science, 10*(3), 282–306. https://doi.org/10.1177/1745691615577701.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team. (2016). *_nlme: Linear and nonlinear mixed effects models_*. R package version 3.1-128, http://CRAN.R-project.org/package = nlme.

Plewis, I. (1989). Comment on "centering" predictors in multilevel analysis. *Multilevel Modeling Newsletter, 1*(3), 6, 11.

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raudenbush, S. W. (1989). A response to Longford and Plewis. *Multilevel Modeling Newsletter, 1*(3), 8–11.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Ree, M. J., Mullins, C. J., Mathews, J. J., & Massey, R. H. (1982). Armed Services Vocational Aptitude Battery: Item and factor analyses of Forms 8, 9, and 10 (Rep. No. AFHRL-TR-81-55). *Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division. (NTIS No. AD-A113465)*.

Rodgers, J. L. (1998). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence, 26*(4), 337–356. https://doi.org/10.1016/S0160-2896(99)00004-5.

Rodgers, J. L. (2014). Are birth order effects on intelligence really Flynn effects? Reinterpreting Belmont and Marolla 40 years later. *Intelligence, 42*, 128–133. https://doi.org/10.1016/j.intell.2013.08.004.

Rodgers, J. L. (2015). Methodological issues associated with studying the Flynn effect: Exploratory and confirmatory efforts in the past, present, and future. *Journal of Intelligence, 3*(4), 111. https://doi.org/10.3390/jintelligence3040111.

Rodgers, J. L., Kohler, H. P., McGue, M., Behrman, J., Petersen, I., Bingley, P., & Christensen, K. (2008). Education and IQ as direct, mediated, or spurious influences on female fertility outcomes: Linear and biometrical models fit to Danish twin data. *American Journal of Sociology, 114*(Supplement), S202–S232. https://doi.org/10.1086/592205.

Rodgers, J. L., Rowe, D. C., & May, K. (1994). DF analysis of NLSY IQ/achievement data: Nonshared environmental influences. *Intelligence, 19*, 157–177. https://doi.org/10.1016/0160-2896(94)90011-6.

Rodgers, J. L., & Wänström, L. (2007). Identification of a Flynn effect in the NLSY: Moving from the center to the boundaries. *Intelligence, 35*(2), 187–196. https://doi.org/10.1016/j.intell.2006.06.002.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Sundet, J. M. (2014). The Flynn effect in families: Studies of register data on Norwegian military conscripts and their families. *Journal of Intelligence, 2*(3), 106. https://doi.org/10.3390/jintelligence2030106.

Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *American Psychologist, 3*(2), 54. https://doi.org/10.1037/h0054962.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. Fourth Edition. New York: Springer. ISBN 0-387-95457-0. https://doi.org/10.1007/978-0-387-21706-2.

Wang, L. (Peggy), & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods, 20*(1), 63–83. https://doi.org/10.1037/met0000030.

Wichman, A., Rodgers, J. L., & MacCallum, R. C. (2006) A multilevel approach to the relationship between birth order and intelligence. *Personality and Social Psychology Bulletin, 32*, 117–127. https://doi.org/10.1177/0146167205279581.

Wichman, A., Rodgers, J. L., & MacCallum, R. C. (2007). Birth order has no effects on intelligence: A reply and extension of previous findings. *Personality and Social Psychology Bulletin, 33*, 1195–2000. https://doi.org/10.1177/0146167207303028.

Woodley, M. A. (2011). Heterosis doesn't cause the Flynn effect: A critical examination of Mingroni (2007). *Psychological Review*, *118*(4), 689–693. https://doi.org/10.1037/a0024759.

Woodley, M. A. (2012). A life history model of the Lynn–Fynn effect. *Personality and Individual Differences, 53*(2), 152–156. https://doi.org/10.1016/j.paid.2011.03.028.

Wu, Y.-W. B., & Wooldridge, P. J. (2005). The impact of centering first-level predictors on individual and contextual effects in multilevel data analysis. *Nursing Research, 54*(3), 212–216. https://doi.org/10.1097/00006199-200505000-00009.

Zajonc, R. B., & Sulloway, F. J. (2007). The confluence model: Birth order as a within-family or between-family dynamic? *Personality and Social Psychology Bulletin, 33*(9), 1187–1194. https://doi.org/10.1177/0146167207303017.

Appendix A

Further examining the correlation between the level two variables, it can be shown that because we mean centered the level-1 variable after removing the level-2 variable, there can be no systematic covariation between the level 1 and either of the resultant level-2 variables. Prior to cluster mean centering, there is a covariation between $\tilde{x}_{ij}$

and z_j, and this covariance is given as follows:

$$
\begin{aligned}
&cov(\tilde{x}_{ij}, z_j) \\
&= \sum \frac{(\tilde{x}_{ij} - e(\tilde{x}_{ij}))(z_j - e(z_j))}{n} \\
&= \sum \frac{(\breve{x}_{.j} + \ddot{x}_{ij} - e(\breve{x}_{.j} + \ddot{x}_{ij}))(z_j - e(z_j))}{n} \\
&= \sum \frac{(\breve{x}_{.j} + \ddot{x}_{ij} - e(\breve{x}_{.j}) - e(\ddot{x}_{ij}))(z_j - e(z_j))}{n} \\
&= \sum \frac{(\breve{x}_{.j} + \ddot{x}_{ij} - e(\breve{x}_{.j}) - 0)(z_j - e(z_j))}{n} \\
&= \sum \frac{\breve{x}_{.j}z_j + \ddot{x}_{ij}z_j - e(\breve{x}_{.j})z_j - \breve{x}_{.j}e(z_j) - \ddot{x}_{ij}e(z_j) + e(\breve{x}_{.j})e(z_j)}{n} \\
&= \sum \frac{\breve{x}_{.j}z_j + 0 - e(\breve{x}_{.j})z_j - \breve{x}_{.j}e(z_j) - 0 + e(\breve{x}_{.j})e(z_j)}{n} \\
&= \sum \frac{\breve{x}_{.j}z_j - e(\breve{x}_{.j})z_j - \breve{x}_{.j}e(z_j) + e(\breve{x}_{.j})e(z_j)}{n}
\end{aligned}
$$

If we calculate $cov\left(\breve{x}_{.j}, z_j\right)$, we see that it is $\sum \frac{(\breve{x}_{.j} - e(\breve{x}_{.j}))(z_j - e(z_j))}{n} = \sum \frac{\breve{x}_{.j}z_j - e(\breve{x}_{.j})z_j - \breve{x}_{.j}e(z_j) - e(x_{.j})e(z_j)}{n}$ which is precisely equal to $cov(\tilde{x}_{ij}, z_j)$ calculated previously. The entire covariance between the level-1 variable $\tilde{x}_{ij}$ and the level-2 variable z_j is included in $\breve{x}_{.j}$. However, although $cov\left(\breve{x}_{.j}, z_j\right) = cov\left(\tilde{x}_{ij}, z_j\right)$, the respective correlations are not equal because the standard deviations of $\tilde{x}_{ij}$ and $\breve{x}_{.j}$ are not equal. This correlation is of some concern. If it is equal to 1, the two variables cannot be included in the model simultaneously; however, this will only occur if $\breve{x}_{.j} = c * z_j$, where c is some constant. In such a scenario, this implies that the original variable x_{1j} comprised solely $\ddot{x}_{ij} + (1 + c) * z_j$. In our example of using patient baseline scores, this would imply that every patient's mean deviation from their baseline was some multiple of their baseline, and that this multiple was exactly the same for every patient in the study. It is unlikely that such an event would occur in practice.

If a researcher is concerned about excessive collinearity, it is possible to impose $cov\left(\breve{x}_{.j}, z_j\right) = 0$ by residualizing $\breve{x}_{.j}$ on z_j and multiplying z_j by the resulting regression coefficient $\beta + 1$. Although this might be a technically correct way to manage the collinearity, because $\breve{x}_{.j}$ and z_j form a linear combination equal to $\bar{x}_{.j}$, if $\breve{x}_{.j}$ and z_j are problematically collinear, simply replacing the two variables with $\bar{x}_{.j}$ is likely a better course of action. In the case of high collinearity, the high positive correlation implies that $\breve{x}_{.j}$ and z_j are equivalent variables and because $\bar{x}_{.j}$ is a linear combination of two equivalent measures, it would be a legitimate substitute. Furthermore, using this regression method instead of the simple subtraction method may reduce the interpretability of the variables, defeating the purpose of double decomposition. However, if a researcher desires to use double decomposition, and also wants to remove all correlations between $\breve{x}_{.j}$ and z_j, this regression method is a means of achieving that goal. If the assumptions of regression have been met, $\breve{x}_{.j}$ and z_j will not be correlated as $\breve{x}_{.j}$ is a residual variable.

Appendix B

```
library(MASS)
library(nlme)
set.seed(420)

##Strong correlations, negative for group mean##
i<-1
Whole_Tvalues<-c()
CMC_Tvalues<-c()
DD_Tvalues<-c()
CMC_Context_Tvalues<-c()

Whole_pvalues<-c()
CMC_pvalues<-c()
DD_pvalues<-c()
CMC_Context_pvalues<-c()

Whole_Coeff<-c()
CMC_Coeff<-c()
DD_Coeff<-c()
CMC_Context_Coeff<-c()
while(i<5001){
  Datums<-c()
  Group<-rep((1:100),each = 30)
```

```
  ID<-c(1:3000)
  Remainder<-rnorm(3000, sd = 1)
  X<-as.data.frame(mvrnorm(100,c(0,0,0),matrix(c(1,-.5,.5,-.5,1,-.5,.5,-.5,1),3)))
  names(X)<-c("Outcome","Group_mean","Context")
  Outcome<-rep(X$Outcome,each = 30)
  GMR<-rep(X$Group_mean, each = 30)
  Context<-rep(X$Context, each = 30)
  Datums<-as.data.frame(cbind(Group, ID, Remainder, Outcome, GMR, Context))
  Datums<-Group_function(Datums, "Remainder",levels = "Group", center = T, append = T)
  Datums$Original_Var<-Datums$Remainder+Datums$GMR+Datums$Context
  Datums$Original_cluster_mean<-Datums$GMR+Datums$Context
  Datums$Outcome_lower<-Datums$Remainder*.5+rnorm(3000,sd = sqrt(.75))
  Datums$Outcome<-Datums$Outcome+Datums$Outcome_lower

  Fit_whole<-lme(Outcome~Original_Var, data = Datums, random = ~1|Group)
  Fit_CMC<-lme(Outcome~Remainder+Original_cluster_mean, data = Datums, random = ~1|Group)
  Fit_DD<-lme(Outcome~Remainder+GMR+Context, data = Datums, random = ~1|Group)
  Fit_CMC_Context<-lme(Outcome~Remainder+Original_cluster_mean+Context, data = Datums, random =
~1|Group)

  Whole_Tvalues<-rbind(Whole_Tvalues,summary(Fit_whole)$tTable[,"t-value"])
  CMC_Tvalues<-rbind(CMC_Tvalues,summary(Fit_CMC)$tTable[,"t-value"])
  DD_Tvalues<-rbind(DD_Tvalues,summary(Fit_DD)$tTable[,"t-value"])
  CMC_Context_Tvalues<-rbind(CMC_Context_Tvalues,summary(Fit_CMC_Context)$tTable[,"t-value"])

  Whole_pvalues<-rbind(Whole_pvalues,summary(Fit_whole)$tTable[,"p-value"])
  CMC_pvalues<-rbind(CMC_pvalues,summary(Fit_CMC)$tTable[,"p-value"])
  DD_pvalues<-rbind(DD_pvalues,summary(Fit_DD)$tTable[,"p-value"])
  CMC_Context_pvalues<-rbind(CMC_Context_pvalues,summary(Fit_CMC_Context)$tTable[,"p-value"])

  Whole_Coeff<-rbind(Whole_Coeff,summary(Fit_whole)$tTable[,"Value"])
  CMC_Coeff<-rbind(CMC_Coeff,summary(Fit_CMC)$tTable[,"Value"])
  DD_Coeff<-rbind(DD_Coeff,summary(Fit_DD)$tTable[,"Value"])
  CMC_Context_Coeff<-rbind(CMC_Context_Coeff,summary(Fit_CMC_Context)$tTable[,"Value"])
  i<-i+1
  if(i%%50 == = 0){
    timestamp()
    print(i)
  }
}
```