Multiple Imputation in Complex Survey Settings: A Comparison of Methods within the Health Behaviour in School-aged Children Study

By:

Laura Holder

In Collaboration with Supervisors:

Dr. Michael McIsaac

Dr. William Pickett

Biostatistics Practicum Report

Department of Public Health Sciences

Queen's University

(September 2015)

Acknowledgements

First and foremost, I would like to thank my primary supervisor Dr. Michael McIsaac for his ongoing advice, guidance, and encouragement. I would also like to thank my co-supervisor Dr. William Pickett for his valuable insight and feedback, as well as the staff and faculty at the Department of Public Health Sciences for their help and support.

Funding and Support

The Public Health Agency of Canada and Health Canada funded Cycle 7 of the Health Behaviour in School-aged Children study in Canada. Additional support for this analysis included operating grants from the Canadian Institutes of Health Research and the Heart and Stroke Foundation of Canada (MOP 97962; PCR 101415). The Canadian principal investigators of the HBSC study are Dr. John Freeman and Dr. William Pickett at Queen's University, Canada, and its national coordinator is Matthew King at Queens University, Canada. The international coordinator of the HBSC survey is Dr. Candace Currie at University of St Andrews, Scotland, and the international databank manager is Dr. Oddrun Samdal at the University of Bergen, Norway. Laura Holder was additionally supported by a Queen's University Graduate Award.

Abstract

Missing data due to non-response is pervasive in large-scale survey research. Failing to appropriately account for these missing values can lead to erroneous findings and false conclusions. Multiple Imputation (MI) has become a highly recommended approach to handling non-response in surveys since it is flexible, easy to implement, and can effectively regain efficiency and reduce bias. Features of complex surveys such as clustering, unequal probability sampling, categorical variables, and multi-item scales, can necessitate the use of complex and tailored MI procedures. These more complex methods detract from the desirability of MI since its ease of implementation and general applicability are among its main advantages. Although simulation studies have demonstrated consequences of ignoring these features during MI, in a more practical sense, it is not clear what advantage these complex methods offer over those that may be implemented in simpler settings. The present investigation involved exploration of this problem through application MI within the Health Behaviour in School-aged Children study (HBSC). The current literature was examined and best practices were ascertained. A set of MI methods ranging in complexity were identified and applied. In particular, applied MI methodology differed by: i) the extent to which the clustered nature of the HBSC data was incorporated in the imputation procedure; ii) the approach used to impute a multi-item measurement; and iii) the parametric assumptions involved. MI was effective at regaining efficiency and reducing concerns about bias, however, no particular MI method resulted in substantially different findings or conclusions. It was concluded that more simple MI methodology is adequate in the context of the investigated research questions. Situations where more complex MI methods may be required were identified, in order to make recommendations for circumstances which extend outside of the analysis goals in the present investigation.

List of Acronyms

CCA	Complete-case analysis
FCS	Fully conditional specification
HBSC	Health Behaviour in School-aged Children
ICC	Intra-class correlation
KNN	k nearest neighbour
MI	Multiple imputation
MAR	Missing at Random
MCAR	Missing Completely at Random
MPML	Multi-level psuedo-maximum likelihood
MNAR	Missing Not at Random
PS	Psychosomatic symptoms

Contents

List of Tables

List of Figures

1 Introduction

Missing data is a nearly inevitable problem in large-scale survey research. Individuals often fail to respond to particular elements of a survey, which leads to scattered missingness throughout a data set. Ignoring these individuals during analysis usually results in throwing out substantial amounts of observed data, as well as potentially biasing parameter estimates (Little and Rubin, 2002). The pervasiveness of missing data has motivated extensive research on developing and evaluating methods for missing data treatment. One of these methods in particular, multiple imputation (MI) as described by Rubin (1987), has received extensive attention due to its flexibility, ease of use, and its ability to provide unbiased and efficient estimators (Little & Rubin, 2014; Reiter & Raghunathan, 2007; Schafer & Graham, 2002; White, Royston, & Wood, 2011).

Multiple imputation involves filling in missing values to create a complete data set, in a way that accounts for both the natural variability in the data and the uncertainty involved in imputing values. The goal of imputation is not to generate accurate predictions of missing values, but rather to replace them in a way that maintains the relationships in the data set in order to exploit the available data from a partially-observed individual (Little & Rubin, 2014). For this reason, it is of utmost importance that the procedures used during MI are *congenial* to the model that will be used for analysis (Meng, 1994). In other words, the imputation procedure must maintain the relationships of interest in any analyses to be conducted on the imputed data.

Ensuring congeniality when utilizing MI in complex survey settings poses unique challenges (Andridge, 2011; Carpenter & Kenward, 2012; Kim et al., 2006; Reiter et al., 2006; Seaman et al., 2012). Therefore, as the popularity of large-scale survey research based on complex designs has increased, research based on more involved and tailored MI procedures has pro-

ceeded as well (Carpenter, Goldstein, & Kenward, 2011; Schafer, 2009; van Buuren, 2011; Zhao & Yucel, 2009). However, these complex procedures are more computationally intensive, more challenging to implement, and less widely available in popular software (Drechsler, 2015). Clustered data, unequal probability sampling, and multi-item measurements, seem to necessitate the use of these more complex MI procedures thereby detracting from the desirability of MI. It is not clear how much is gained, in a practical sense, by using more complex MI procedures in order to account for these features over what may be implemented in a less complex setting.

This practicum involved exploration of this problem through application of MI in an applied data set: the 2009-2010 Canadian Health Behaviour in School-aged Children study (HBSC). The HBSC is a cross-national survey conducted in collaboration with the World Health Organization (WHO) every four years in over 40 countries (Freeman et al. 2011). The 2009-2010 Canadian HBSC study is characterized by many typical complex survey features including multi-stage cluster sampling, disproportionate stratified sampling, multi-item scale measures, and many categorical variables. Many variables of interest for ongoing research questions within the HBSC data set have portions of missing values due to non-response (Freeman et al., 2011). The missing data within these variables must be handled in a way that is congruent with the complex features of the data set. Comparative application of MI methodology within the HBSC was directed by two main goals:

1. To evaluate how the methodological complexity of MI (i.e. the extent that the MI procedure accounts for particular complex survey features) impacts the results of a specific set of analyses within the HBSC; and

2. To evaluate the practical implications of these results as they may extend to other analyses.

Comparative application of such methodology in a realistic context is limited in current

literature and can provide valuable information not gained from the restrictive settings in simulation studies. Further, it will serve to inform recommendations for methods which may be applied in the immense number of projects that involve data from the HBSC study. In particular, this comparison of MI methods focused on the following complexities:

i. The HBSC utilizes a multi-stage sampling design resulting in data that is clustered in nature; this complex survey feature must be considered during MI (Andridge, 2011; Carpenter & Kenward, 2012; Kim et al., 2006; Reiter et al., 2006; Seaman et al., 2012).

ii. The HBSC involves several multi-item measures, which could be imputed either as a total score, or as a series of individual variables which make up the total score (Eekhout et al., 2014; Gottschall et al., 2012).

1.1 Report Structure

The remainder of this report is structured as follows. Section 2 includes a general summary of some necessary missing data concepts and an overview of some common missing data methodology. Section 3 introduces the HBSC in detail, along with an overview of survey analysis in the complete data setting with particular attention to the challenges present in the HBSC. In Section 4, MI will be overviewed in more detail, along with a review of the literature surrounding the application of MI in complex survey settings. Section 5 is dedicated to presenting the comparative application of MI methods within the HBSC, including a preliminary complete-case analyses informed by the literature review, a description of the implemented imputation methodology, and a comparison of results from the complete-case analyses and the various MI methods. Finally, Section 6 involves a more in-depth comparison of the imputation methods, along with conclusions and recommendations.

2 Background

In survey research, two scenarios result in missing data: *unit non-response* and *item non-response* (Little, 1988). Unit non-response describes missing data that occurs when a sampled individual fails to complete any of the survey. Data available for this individual are usually restricted to demographic or geographic characteristics and, therefore, there is a more restricted range of methods to handle such missing data (Little & Rubin 2014). In contrast, item non-response occurs when an individual completes part of a survey but fails to respond to specific questions. This may occur when an individual fails to notice a question, or chooses not to answer it for any number of reasons. Item non-response is scattered throughout the data set, and the remaining data from partially-observed individuals allows for implementation of a wider range of missing data analysis techniques. The present investigation focuses on the methodology directed towards handling item non-response.

2.1 Missing Data Mechanisms

Understanding the cause (or mechanism) leading to the missing data is a key component in determining a suitable analysis strategy. Rubin (1976) formalized the concept of missing data mechanisms which have since become referred to in the literature as three distinct groups (Little & Rubin, 2014; Schafer & Graham 2002). Consider a data set $Y = (y_1, y_2, ..., y_k)$ with k variables, some of which have missing values. This data set can be partitioned into its observed and unobserved components referred to as Y_{obs} and Y_{mis} respectively. Let R be the vector of response indicators, so that R_i is 1 when element i is observed and 0 when it is missing. When $P(R|Y;\gamma) = P(R|Y_{obs}, Y_{mis};\gamma) = P(R|\gamma)$, the missing data mechanism is called Missing Completely At Random (MCAR). When $P(R|Y;\gamma) = P(R|Y_{obs}, Y_{mis};\gamma) = P(R|Y_{obs};\gamma)$, then missing data mechanism is called Missing at Random (MAR). When

$P(R|Y;\gamma) = P(R|Y_{obs}, Y_{mis};\gamma) \neq P(R|Y_{obs};\gamma)$, the missing data mechanism is called Missing Not At Random (MNAR). Roughly speaking, if the propensity to respond does not depend on any observed or unobserved data, then the data are MCAR; if the propensity to respond is related to other observed responses but unrelated to the missing values then data are MAR; if propensity to respond is related to the missing values even after controlling for observed values then data are MNAR.

Consider, for example, a survey which asks participants to report their yearly income. Suppose the page on which this question was printed was accidentally left out of a random set of surveys. These missing values can be considered a random sample of the (theoretical) full data set and, therefore, MCAR. Suppose instead, those who have certain professions are more likely to refuse to answer the question of yearly income and also tend to have lower incomes. Following conditioning on "profession" the likelihood of an individual refusing to report their yearly income will be independent of the missing values and, therefore, the data are considered MAR. Finally, suppose participants with a lower yearly income simply less inclined to report it. The likelihood of response now depends on the missing values themselves even after controlling for profession and, therefore, the data are considered MNAR. Missing data mechanisms are important to consider, since common missing data approaches are only valid under particular missing data mechanisms and will lead to bias if the assumed mechanism is not appropriate (Little & Rubin, 2014).

2.2 Complete-case Analysis

When item non-response occurs in survey settings, it remains common practice to implement what is referred to as *complete-case analysis* (CCA). This strategy simply discards all individuals with incomplete variables using *list-wise deletion*. Although this is perhaps the simplest approach to handling missing data, this strategy has potential short comings

in terms of efficiency and, more importantly, bias. Firstly, discarding incomplete cases is potentially inefficient and could lead to a loss of statistical power since observed data are discarded. Secondly, if the complete cases differ systematically from those with missing data, then the reduced data set can be biased and erroneous conclusions can arise (Little & Rubin, 2014). If data are MCAR, then list-wise deletion of individuals with incomplete data is often considered an adequate strategy despite the loss of efficiency because it will not introduce bias. Additionally, CCA can be unbiased in some circumstances even if data are not MCAR. Specifically, if only the outcome variable is incompletely observed and the missingness is can be considered MAR given the covariates included in the model, then results will not be biased (Sterne et al., 2009). Similarly, if the occurrence of missing data in predictor variables is unrelated to the outcome, CCA will not introduce bias (Sterne et al., 2009). Unfortunately, the assumption that the data are MCAR or adhere to one of the above criteria is often implausible. It is more reasonable to assume that data are MAR after controlling for variables which are part of the substantive analysis as well as on *auxiliary variables*. Therefore, missing data treatment methods which allow for incorporation of auxiliary variables are advantageous. In particular, MI is consistent for MAR data, so determining suitable variables to make this assumption plausible becomes an important task (Carpenter & Kenward, 2012). MI can readily be extended to MNAR settings by making untestable assumptions about the nature of the relationship between missingness and the unobserved data; this approach is important for the implementation of sensitivity analysis (Carpenter & Kenward, 2012). Such sensitivity analyses are important in practice, but they are not a focus of this practicum.

2.3 Inverse Probability Weighting

Inverse probability weighting can be considered an extension of CCA that attempts to address the issues of efficiency and bias by incorporating partially-observed or auxiliary data in the creation of weights in an attempt to render the missingness MAR (Little & Rubin, 2014; Seaman & White, 2013). The inverse probability weighting method assigns respondents weights equal to the inverse of their probability of being completely observed by estimating the probability of response within class based adjustment cells or through parametric modelling of propensity to respond. Inverse probability weighting is typically applied to handle unit non-response, as application when the pattern of missingness is scattered throughout the data set is often infeasible (Carpenter et al., 2006; Seaman & White, 2013). Furthermore, applying inverse probability weighting in the case of item non-response may still not be most efficient use of the information from partially-observed individuals (Carpenter et al., 2006). Inverse probability weighting can also be implemented to account for unequal probabilities of selection into the sample (*sampling weights*) or to ensure that response distributions adhere to the known distribution of values in the population (*post-stratification weights*); sampling, after all, is simply unit missingness that occurs by design (Chambers & Skinner, 2003). Typically, each of these characteristics is considered as a separate component of the probability of inclusion in a sample and are combined to form one overall weight.

2.4 Imputation

The complete case and weighting methods discussed above do not make efficient use of available data from partially-observed individuals. Missing data methods which fill in (or *impute*) the missing variables with suitable replacement values have the advantage of avoiding the deletion of partially-observed individuals. However, some common imputation approaches

have important limitations. If all missing values are replaced with a single constant value (as can be done in *mean imputation*) then bias can be introduced and the variability in the data set will be artificially reduced, which can result in underestimated variances (Rubin, 1987). Model-based imputation methods improve upon mean imputation by allowing the inclusion of auxiliary variables during the imputation process. When chosen appropriately, inclusion of auxiliary variables can render the data MAR and reduce bias due to non-response. For example, *regression model imputation* involves regressing the variable with missing values on a suitable set of fully observed variables in the data set. The missing values are then imputed from predicted values of the regression model, usually augmented with random draws from the residual distribution to model the natural variation within the missing data. Although improving upon single value imputation, these methods fail to acknowledge the missing data as a source of uncertainty (Rubin, 1987). Multiple imputation, first proposed by Rubin (1987), is motivated by this idea. MI accounts for this uncertainty by generating multiple, say m, values for each missing data point, resulting in m complete data sets. After performing analysis on each of these data sets individually, the results can be pooled to get a single point estimate, and a single variance estimate which incorporates both the variation within data sets (within-imputation uncertainty) and the variation between data sets (between-imputation uncertainty).

Multiple imputation requires specification of an *imputation model* in order to take advantage of auxiliary variables. Little and Rubin (2002) classify imputation methods into two general groups based on whether imputed values are derived from either an *explicit* or *implicit* model. Explicit approaches involve imputing values that are generated by applying a formal model with explicit assumptions, while implicit approaches use an algorithm to identify an individual that is "similar" to one with missing data and draw values for imputation from the observed values of this individual. Explicit multiple imputation methods are typically

based on parametric models, such as the regression model-based imputation described above, and therefore rely on the associated assumptions. On the other hand, implicit models relax the requirements of modelling assumptions. Therefore, these methods can be advantageous when the data may involve higher-order and non-linear relationships that are not known or are very challenging to model, and when efficiency is not a primary concern.

Multiple Imputation can be a powerful and flexible tool for reducing bias due to missing data and retaining the efficiency that may be lost through CCA or inverse probability weighting (Little & Rubin, 2014; Reiter & Raghunathan, 2007; Schafer & Graham, 2002; White, Royston, & Wood, 2011). Despite these advantages, however, there are many potential challenges and complications which may arise during MI, especially in complex survey settings such as the HBSC (Andridge, 2011; Carpenter & Kenward, 2012; Kim et al., 2006; Reiter, Raghunathan, & Kinney, 2006; Seaman et al., 2012). Ensuring *congeniality* (i.e. ensuring the imputation procedure maintains all relationships present in the subsequent analysis) is particularly challenging in complex survey settings. When the imputation methodology and the subsequent (or *substantive*) analysis are uncongenial, the results from the analysis may be biased. If, for example, the observed outcome of interest was left out of the imputation model for a missing exposure, then this uncongeniality would result in an attenuation of the true relationship (Carpenter & Kenward, 2012). In that setting, the direction of the bias is clear because the outcome and exposure were explicitly assumed to be unrelated within those individuals requiring imputations. In complex survey settings, analyses require complex modelling structures (Asparouhov & Muthen, 2006; Carle, 2009; Rabe-Hesketh & Skrondal, 2006). The impact of uncongenial imputation models when analyses involve these high-order structures is less clear. Composite variables (e.g. multi-item measurement scales) also present an interesting challenge as they may be imputed in various ways and the ideal method is often uncertain (Eekhout, 2014; Gottschall et al., 2012). Prior to discussing these

challenges in more detail, it is necessary to first overview the analytical challenges of survey data present in the complete data setting.

3 Literature Review: Analysis of Complex Survey Data

3.1 Health Behaviour in School-aged Children Study

The HBSC is a study of health and health risk behaviours in adolescent populations (Freeman et al. 2011). It is conducted every 4 years in collaboration with the World Health Organization and collects data through written health surveys administered in classroom settings. Following a common international protocol, the 2009-2010 Canadian study implemented a multi-stage sampling strategy in which schools were sampled from each of 11 Canadian provinces and territories (New Brunswick and Prince Edward Island chose not to participate in the 2009-2010 cycle of the HBSC). In most provinces, a systematic sample of schools was performed. However, in the Northern territories, a census of all students was attempted. Initially stratified by province, the primary sampling unit was schools, which were further stratified by type of school board (public versus separate), urban and rural geographic status, and language of institution. A systematic sampling procedure was then used to select schools so that they were sampled proportionally to the size of the school, as estimated by the number of classrooms. Of the 765 randomly selected schools, 436 (57.0%) chose to participate. When a school chose not to participate a neighbouring school with consistent characteristics was contacted to participate instead. Generally, one or two classrooms were selected from each sampled school, and all students from selected classrooms were asked to participate. Of the 33868 students in the sampled classrooms, 26078 (77.0%) were present and chose to fill out the questionnaire on the appointed day. The number of students who participated from each school ranges from 2 to 519 students.

Although sampling probabilities or weights reflective of the sampling procedure are not available with the HBSC data set, weights were calculated based on population distributions by province and grade in order to account for the disproportionate sampling between provinces and grades. That is, all students from the same province within the same grade are assigned a specific weight based on the ratio of the size of the sample selected from this stratum to the size of the strata within the population of Canadian school children (i.e. a single post stratification weight for each combination of province and grade).

The unit non-response in the study will not introduce bias if the missingness is not related to differences in the relationships of interest–an assumption that is made in the present investigation (Little & Rubin, 2014). Item non-response is the focus of this investigation, as the partial information available on those individuals who completed at least part of the questionnaire can be exploited to reduce bias (Little & Rubin, 2014). Accommodating item non-response will be explored in Section 4 and 5 while the remainder of Section 3 is focused on appropriate analyses of the complex survey data in absence of incomplete data.

In summary, the complex survey features of the HBSC include a multi-stage sampling design leading to clustered data, and crude survey weights which reflect disproportionate sampling between provinces and grades.

3.2 Survey Weighting

When elements in a survey are sampled with unequal probability and the probability of sampling is related to the outcome (after potentially controlling for covariates), then the sample design is considered *informative* (Asparouhov, 2006). In this case, the sampling process (usually reflected in a data set through the presence of sampling weights or more general survey weights which may be adjusted for unit non-response and post-stratification)

cannot be ignored without leading to biased estimators. A suitable strategy to account for the sampling process depends on the inferential goals of an analysis (Chambers & Skinner, 2003). In *design-based inference*, the goal is to estimate features of a population which are considered fixed (e.g. means and ratios), and the only source of variability is the sampling distribution (Sterba, 2009). That is, the goal is to estimate what would have been observed if a census was taken (in which case, there would be no variability for the parameters of interest). Sampling weights cannot be disregarded when inferential goals are design-based, otherwise, unequal probabilities of inclusion will not be accounted for and estimated values will not be representative of the target population values (Chambers & Skinner, 2003).

In contrast to design-based inference, the use of sampling weights in *model-based inference* has been debated (e.g. Little, 1993, 2004; Pfeffermann, 1993, 1996). The target of model-based inference is the model parameters which generate the outcome, as opposed to finite population characteristics (so there would be variability around our estimator even if we had a census because we are not interested in the characteristics of this finite population, but rather the parameters generating the finite population; Sterba, 2009). If the model between the outcome and covariates is correctly specified and is consistent across sampling groups, then excluding sampling weights will not introduce bias (Gelman, 2007; Little, 1993, 2004). Along these lines, a fully model-based approach to handling sampling design is to incorporate suitable *design variables* (characteristics of the population on which sampling depends, such as strata) into the model as covariates (Little, 1993, 2004, 2004; Sterba, 2009). This approach may be inadequate, however, if there is insufficient auxiliary information to fully describe inclusion probabilities or if the inclusion of these additional covariates substantially modifies model interpretation (Little, 1993, 2004; Pfeffermann, 1996). In these circumstances, *hybrid* methods which compromise between design and model-based methods by using weighted estimation, are considered preferable (Pfeffermann, 1996; Sterba, 2009).

These approaches incorporate the sampling weights into the likelihood expression, using what is called *pseudo-likelihood estimation* (Pfeffermann, 1993). Utilizing these hybrid approaches provides robustness to model misspecification, since incorporation of the weights allows for *design consistent* estimation of model parameters (Pfeffermann, 1993, 1996). That is, the parameters still have a meaningful interpretation as consistent estimators of finite population values in the presence of model misspecification, while without weights they may just be biased (Pfeffermann, 1996). It is relevant to note, however, that if sampling weights are uninformative (i.e. not related to the outcome after controlling for covariates), it is recommended that they not be used in the analysis (Asparouhov, 2006). If weights are uninformative, the reduction in the bias achieved by weighting will be overwhelmed by the increase in the variance of the parameter estimates and should not be used (Asparouhov, 2006).

Consideration of the HBSC weights can be informed based on the above discussion. If an analysis is to proceed based on a design-based inference framework (i.e. the goal is to estimate finite population values, such as proportions) then the weights should necessarily be included in analysis (Chambers & Skinner, 2003). Otherwise, the disproportionate sampling between provinces present in the HBSC will not be captured, and estimates will not be representative of the targeted population values. However, if the analysis is better suited to a model-based inference framework (e.g. associative relationships), the importance of the weights is less clear. A weighted and un-weighted analysis should be compared, and if there are no substantial differences between results, the weights can be considered uninformative and are better off being left out of the analysis (Asparouhov, 2006).

3.3 Clustered Data

The HBSC uses a multi-stage sampling strategy resulting in data with a clustered structure in which students are nested within classrooms nested within schools. With clustered data

the units of observation are not independent, which violates assumptions of many traditional statistical techniques such as multiple regression. Therefore, clustered data must be handled appropriately in order to account for the homogeneity of individuals belonging to the same cluster (Snijders, 2011). Within design-based analyses, clustering is typically treated as a nuisance and taken into account through the use of robust estimation of the variance of parameters (Chea, 2009). This post hoc procedure only affects standard error estimates, however, and does not adjust the parameter estimates themselves (Cheah, 2009). While this is suitable for simpler population estimates (e.g. means, ratios), a model-based approach is often used to account for data clustering for more complex estimates (e.g. regression parameters) even when a design-based inferential framework may be preferred for analysis goals (Asparouhov 2006; Rabe-Hesketh & Skrondal, 2006 ; Cheah, 2009).

Model-based methods incorporate the clustered structure of the data into the modelling process through the inclusion of random parameters, in *mixed* or *multi-level* models. Covariates and factors in such models are typically referred to as *fixed* or *random* effects, although Gelman and Hill (2006) advise against the use of these terms and recommend to instead focus on description of the model itself. Along these lines, it is more appropriate to describe the parameters in multi-level models as fixed or random where appropriate. A random parameter is allowed to vary for each level-two unit (cluster), thereby allowing for cluster specific estimates which deviate from a mean estimate (intercept or slope). On the other hand, fixed parameters are not estimated uniquely for each cluster. For example, a fixed intercept would imply that the mean of the dependent variable is the same in each cluster. Allowing for a random intercept involves estimation of both a fixed intercept component, assumed to be invariant across clusters, and a randomly varying component which represents the deviation of each cluster from this fixed component. For example, a linear multi-level model with random intercepts of the relationship between Y_{ij}, an outcome variable for individual i within

cluster j, and the corresponding predictor x may be expressed as:

$$Y_{ij} = \beta_{0j} + \beta_1 x_{ij} + e_{ij}, \tag{3.1}$$

$$\beta_{0j} = b_{00} + U_{0j}, \tag{3.2}$$

where we assume

$$e_{ij} \sim N(0, \sigma^2), \tag{3.3}$$

$$U_{0j} \sim N(0, \tau^2). \tag{3.4}$$

The level-one model in (3.1) describes the relationship between the outcome variable Y and a level-one predictor x through the fixed parameter β_1. The level-two model in (3.2) describes the intercept β_{0j} as a fixed component b_{00}, and a random component U_{0j} that represents the cluster specific deviation from the mean intercept b_{00}. The residual variability in the data is decomposed into the component attributable to variance between clusters (τ^2), and the portion attributable to variance between individuals (σ^2). The total residual variability is then considered the sum of these two components ($\tau^2+\sigma^2$). In some settings, the assumption in (3.1c) is relaxed to allow the within-cluster variance to differ across clusters, thus replacing σ^2 by σ_j^2 (Van Burren, 2011).

Estimates of random parameters themselves are usually not of direct interest. Instead interest lies in the extent to which the random components vary between level-two units (i.e. the level-two variance of the random parameter τ^2; Nezlek, 2011). In particular, *intraclass correlations* (ICCs), functions of the variances, are often of interest (Nezlek, 2011). ICCs are a measure of the extent to which variability in the outcome is attributable to the variability between clusters. This measure is can be estimated by fitting a model with no predictors, so the estimated variance components are purely measures of the variability in the outcome. The

ICC is then calculated as ratio of between-group residual variance and total residual variance:

$$ICC = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{3.5}$$

A large ICC demonstrates high similarity between individuals in the same cluster and suggests the need for multi-level modelling (Nezlek, 2011). For this reason, ICCs are often used to determine whether multi-level modelling is necessary, however, ICCs capture only one aspect of differences between clusters and may not be well suited to this task (Nezlek, 2011). That is, ICCs reveal little about how the actual relationships under study vary between clusters.

Generalized linear multi-level models are extensions of linear multi-level models that allow for modelling of binary, categorical, or other non-normally distributed responses through the application of a link function (Rabe-Hesketh & Skrondal, 2001). Notably, the level-one error variance for models with certain link functions (e.g. logit and probit) is typically assumed to be fixed (Breslow & Clayton, 1993; Demidenko, 2013; Rabe-Hesketh & Skrondal, 2001, 2006). For example, in a logistic multi-level model the level-one error variance is fixed as $\sigma^2 = \frac{\pi^2}{3}$ by assumption and is not estimated (Rabe-Hesketh & Skrondal, 2006). Instead, the total residual variance is estimated as $(\tau^2 + \frac{\pi^2}{3})$. Therefore, the inclusion of random effects increases the overall residual variance by allowing $\tau^2 > 0$, but unlike with linear multi-level models, no corresponding change occurs in the individual level variance σ^2. To account for this, the dependent variable is rescaled and model parameter estimates increase in absolute magnitude. Consequently, the effects estimated in a single-level logistic model, β^{SL}, are approximately related to the effects estimated in a multi-level logistic model with random intercepts, β^{MM}, (in which the fixed parameter estimates are conditional on the random

parameter estimates) as follows:

$$\beta^{SL} \approx \sqrt{\frac{1}{1 + \frac{\pi^2}{3}\ \tau^2}}\ \beta^{MM} \tag{3.6}$$

This makes accurate estimation of τ^2 in multi-level logistic models imperative (Rabe-Hesketh & Skrondal, 2001, 2006). Furthermore, given that the multi-level model is correct, estimates from ordinary logistic regression which disregard random components will be biased toward the null (Breslow & Clayton, 1993; Demidenko, 2013; Neuhaus et. Al., 1990).

Using a multi-level model in the presence of clustered data (even when clustering appears minimal) can be beneficial due to the effect of *partial pooling* (Gelman & Hill, 2006). This is particularly true when cluster sizes are unbalanced. When the multi-level structure of the data is ignored with use of single-level model, point estimates are based on relationships across the whole sample. With a random intercept model, a separate mean is estimated for each cluster, specifically, the overall mean b_{00} plus the cluster specific deviation U_{0j} in equation (3.2). This could also be achieved by including fixed parameters for cluster status indicators, however, with a multi-level model these cluster specific means are constrained to come from a normal distribution with mean b_{00} (due to the assumption in equation (3.3); Gelman & Hill, 2006). Therefore, individual cluster estimates are represented in the final point estimates, but estimates for smaller clusters (which are less certain) are "shrunk" towards the overall mean b_{00} and, therefore, towards the effects estimated for the larger clusters (which are more certain; Gelman & Hill, 2006). In contrast, in single-level models, the extent to which point estimates are skewed toward the relationships within the largest clusters is more extreme. Therefore, the relationships present in small clusters may not be adequately represented (Gelman & Hill, 2006). This does not necessarily mean fixed parameter estimates from a single-level model will be incorrect, but it does point to an advantage of using multi-level models with unbalanced clusters: they allow for relationships

occurring in smaller clusters to be more represented in point estimates (Gelman & Hill, 2006).

3.4 Survey Weighting in Multi-level Models

The HBSC involves both clustered data and strata with different survey weights. A fully model-based approach may be attempted by incorporating relevant design features into the multi-level model. Fixed parameters could be used to account for stratification with unequal selection probabilities, while random parameter could be used to account for school-level clustering as described above. However, as with single-level models, if this approach does not sufficiently account for the probability of inclusion or if it results in unwanted changes to model interpretation, a hybrid approach is favoured. (Pfeffermann, 1996; Sterba, 2009). Adaptations to the hybrid method have been developed for use with multi-level models, although suitable implementation of this hybrid approach in the multi-level context is somewhat more challenging than the single-level case (Asparouhov, 2006; Rabe-Hesketh and Skrondal, 2006). Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006) propose the *multi-level pseudo-maximum likelihood* method (MPML), which estimates the population likelihood function by weighting the sample likelihood function at each level of the multi-level model. This is the method for incorporating weights into multi-level models that is implemented in many popular software packages including SAS (in SAS/STAT 13.1 and later), Mplus, and Stata (Asparouhov & Muthen, 2006; Rabe-Hesketh and Skrondal, 2006; Zhu 2014). It is essential that the software be used correctly when implementing the MPML method, in particular, sampling weights must be constructed and specified differently than the weights used for single-level analysis (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006). The method requires both a level-two weight w_j (the inverse of the probability that cluster j is selected), and a conditional level-one weight $w_{i|j}$ (the probability of selecting a

level-one unit i, given that the level-two cluster j was selected).

There are two difficulties that arise in this setting. The first difficulty is that, unlike with single-level models, scaling of level-one weights for use in MPML estimation can influence point estimates of the model (Asparouhov, 2006; Asparouhov & Muthen, 2006; Rabe-Hesketh & Skrondal, 2006). In multi-level models, the distribution of random effect estimates is affected by the ratio of the cluster sample size and the sum of the weights within the cluster (for a detailed explanation of this effect see: Rabe-Hesketh and Skrondal, 2006). Typically, the weights are scaled so they sum to the cluster sample size n_j, which generally achieves the most accurate results (Asparouhov, 2006, 2008; Asparouhov & Muthen, 2006). The second difficulty is that secondary data sets typically only include an overall unconditional survey weight (representing the overall probability of selection). Neglecting to use level-two weights implies clusters were selected with equal probability, which may lead to biased estimators. In this circumstance, Asparouhov (2006) suggests a single-level model be used instead, or if weights are uninformative, that they be excluded from the analysis. Alternatively, approximations of the cluster level weights have been proposed (Goldstein, 2003; Kovacevic and Rai 2003). A simulation study evaluating the performance of these approximations showed that the method suggested by Kovacevic and Rai (2003) may perform adequately in some scenarios, however, it requires that the number of clusters in the population is known (Stapleton, 2012). Stapleton (2012) warns against using weight approximation methods when neither cluster-level or conditional level-one weights are available as no approximation methods have shown to perform reliably in this setting.

The weights in the HBSC are constructed based on the sampled proportion of level-one units (students) in subpopulations determined by province and grade (i.e. post-stratification weights). These weights are not conditional weights, dependent on belonging to a given cluster and cannot be attributed to either cluster-level or conditional level-one weights.

Rabe-Hesketh and Skrondal (2006) caution that post-stratification weights are not suitable for use in MPML estimation due to their unconditional nature. Further, Stapleton (2012) suggests that further studies are required to understand the appropriateness of weight approximation methods in the setting of post-stratification weights. Therefore, there is no clear approximation method that would be suitable for the HBSC weights. The bias which may be introduced by using these weights in multi-level analysis is not clear, and may be minimal in circumstances which only fixed parameters are of interest. However their use in multi-level models is not supported by the current literature (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2012).

3.5 Summary

Analysis of complex survey data, such as that from the HBSC, must be carefully considered, and should be directed by the appropriate inferential framework for the question at hand. When estimating features of a finite population through design-based inference, it is necessary that weights be incorporated in analysis to avoid bias (Chambers & Skinner, 2003). On the other hand if the goal is to describe in a way that is more readily generalizable the process by which the data were generated, a model-based approach is appropriate and weighting may not be necessary (Little, 1993, 2004; Gelman, 2007). With the HBSC data, a model-based analysis should involve the use of a multi-level model in order to account for the dependence of individuals within clusters. However, the weights provided with the HBSC data set are not appropriate for use with the MPML method for weighted multi-level models. Instead, a model-based approach that accounts for the information present in the weights though inclusion of dummy variables for grade and province strata as covariates should be adequate in most scenarios. A weighted and unweighted single-level analysis can be performed to assess if weights are informative beyond this model-based approach. If substantial differences

are observed between results from weighted and unweighted analyses (assuming a correctly specified model), it may be preferable to use a hybrid approach to include the weights in a single-level model with SEs adjusted for clustering (Asparouhov, 2006; Pfeffermann, 1993, 1996; Sterba, 2009).

Nonetheless, neglecting to use a multi-level model in the presence of clustered data such as the HBSC may have important consequences. Firstly, when analysis involves a logistic regression, failing to include an existing random effect does not yield unbiased estimators as with the linear regression case (Breslow & Clayton, 1993; Demidenko, 2013; Rabe-Hesketh & Skrondal, 2001, 2006). The inherent link between random component variances and point estimates in generalized linear multi-level models means that, given the model which includes a random component is correct, the corresponding single-level logistic model will bias parameter estimators towards the null (Breslow & Clayton, 1993; Demidenko, 2013). This also highlights the importance of accurate estimation of the random variance components in generalized linear multi-level models, which may be compromised with the use of the unconditional weights provided in the HBSC data set (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2001, 2006; Stapleton, 2012). Secondly, since the clusters in the HBSC data set are unbalanced (ranging from 2 to 519 students per school), the partial pooling which occurs during multi-level analysis provides the advantage of ensuring that small clusters are represented in point estimates (Gelman & Hill, 2006). In a single-level model, the relationships occurring in the largest clusters will dominate the results.

4 Literature Review: Multiple Imputation in Complex Surveys

4.1 Multiple Imputation

Multiple imputation (introduced by Rubin, 1987, and discussed in detail in Rubin, 1987, 2004) has become a very popular approach for analyzing incomplete data, owing to its ease of use, flexibility, and potential to reduce bias and improve efficiency in data sets with missing values (Little & Rubin, 2014; Schafer & Graham, 2002; White, Royston, & Wood, 2011). It is widely available for use thanks to its implementation in popular statistical software programs such as PROC MI, MIANALYZE, and IVEware in SAS (SAS Institute Inc., 2008; Raghunathan et al., 2002); the packages mice (van Buuren & Groothuis-Oudshoorn, 2011), Amelia II (Honaker, King & Blackwell, 2011), and mi (Su et al., 2011) in R; ICE and MI commands in STATA (Royston, 2011; Royston & White, 2011); and NORM, MIX, CAT, and PAN packages in S-plus (also available in R; Schafer, & Olsen, 1998; Schafer, 2012). MI is motivated by the goal of preserving the advantages of imputation while allowing the uncertainty due to imputation to be assessed. To achieve this a MI strategy must i) fill in missing values with plausible replacements that preserve the relationships present in the observed data, while incorporating random variation; and ii) use independently drawn imputations to generate multiple imputed data sets; the variation across these data sets reflects the uncertainty about the imputations (Little & Rubin, 2014).

Carrying out this process m times results in m imputed data sets. Each completed data set can then be analysed separately as if it were the true data, and the m resulting estimators can be combined for inference (Rubin, 1987). By generating multiple imputed data sets, the within-imputation uncertainty and the between-imputation uncertainty can be examined in order to calculate appropriate standard errors. The combining rules put forth by Rubin

(1987), known as *Rubin's rules*, are as follows. Let $\hat{\theta}_i$ denote the estimate of the parameter of interest θ from the i^{th} out of m imputed data sets and let $\hat{U}_i$ denote the variance associated with $\hat{\theta}_i$. The combined point estimate of the parameter can be calculated as the mean of the m point estimates,

$$\bar{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i. \tag{4.1}$$

The variance of $\bar{\theta}$ is estimated as

$$T = \bar{U} + (1 + \frac{1}{m})B, \tag{4.2}$$

which is a weighted sum of the within-imputation variance

$$\bar{U} = \frac{1}{m} \sum_{i=1}^{m} \hat{U}_i, \tag{4.3}$$

and the between-imputation variance

$$B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_i - \bar{\theta})^2. \tag{4.4}$$

These rules are easy to apply and, for the most part, generally applicable, adding to the flexibility of MI.

In the Bayesian terminology which underlies the MI framework, values for imputation should be repeated independent samples from their *posterior predictive distribution* (Rubin, 1987). Generally, this involves specifying a parametric imputation model for the data and using this to derive (or approximate) the conditional distribution of the missing data given the observed data. Care must be taken in specifying this imputation model in order ensure that it is appropriate for the desired analysis; this idea is discussed further in the next subsection and forms the basis of the explorations throughout this report.

More specifically, consider a substantive analysis model based on a simple linear regression of a incomplete variable $Y = (Y_{mis}, Y_{obs})$ on a complete variable X. Suppose an auxiliary variable V has been identified which renders the data MAR (using simplified notation from Section 2.1) so that $P(R|Y, X, V) = P(R|X, V)$ and, equivalently, $P(Y|X, V, R) = P(Y|X, V)$. An imputation model, $P(Y|X, V; \zeta)$, is fit to the observed data to yield the vector of parameter estimates $\hat{\zeta}$. Although ζ is not of interest, this fitted model is used to generate a complete data set which is analysed with the substantive model to estimate the parameter of interest θ. This is done by randomly drawing the parameter vector, $\hat{\zeta}^*$, from it's posterior distribution, followed by drawing Y_{mis} conditionally given $\hat{\zeta}^*$ to yield the first of m imputations. This two-step procedure is used to ensure appropriate sampling variability between the m sets of imputations so they can be considered *proper*; imputations must be proper for Rubin's rules to hold (see Rubin, 1987, for a formal definition). In contrast, *improper* imputations will lead to insufficient variability between imputations and, therefore, underestimation of the total variance of the parameter of interest (Rubin, 1987).

In practice, there are many ways to implement MI, however, one of three main approaches is typically used:

1. *Joint modelling* (Rubin & Schafer, 1990; Schafer, 1997): The joint modelling method involves specifying a joint distribution for multivariate data, estimating the parameters for this distribution, and then drawing imputed values from this distribution. Although this refers to any MI procedure in which the data are assumed to follow a joint probability distribution, in practice, the joint modelling method nearly always involves the assumption that the data follow a multivariate normal distribution. It can be challenging to make the necessary draws directly from the specified joint model, so a special case of a Markov Chain Monte Carlo (MCMC) procedure called a Gibbs Sampler is typically used. During joint-multivariate normal model imputation this procedure alternates between estimating the

means, variances and covariances of the distribution, and drawing values for imputation (the Gibbs sampling procedure and other commonly used approximation algorithms are described in detail by Carpenter & Kenward, 2012, Appendix A).

2. *Fully conditional specification* (Raghunathan et al., 2001; van Buuren et al, 2006): In contrast to the joint modelling approach, the FCS method does not aim to model all incomplete variables jointly. Instead, FCS imputes data on a variable-by-variable bases, with the goal of specifying a full multivariate distribution for the variables through a set of conditional distributions for each incomplete variable. Imputation then proceeds iteratively across all of these conditional imputation models for a specified number of times, in order to converge to a theoretical joint distribution. For example, suppose that both Y and X variables in the substantive model described above have missing values. Using the described imputation model, the FCS algorithm involves the following steps:

   (a) Fill in X_{mis} and Y_{mis} with some starting values (e.g. simple mean imputation) to generate the imputed complete variables X_{imp} and Y_{imp}

   (b) Fit a simple linear regression of the incomplete variable Y on X_{imp} and V in order to estimate the associated parameters $\hat{\zeta}_Y$

   (c) Randomly draw new parameters, $\hat{\zeta}_Y^*$ based on the fitted model in step (b)

   (d) Use the parameters $\hat{\zeta}_Y^*$ to generate stochastic imputations for Y_{mis} to update Y_{imp}

   (e) Fit a simple linear regression of the incomplete variable X on Y_{imp} and V in order to estimate the associated parameters $\hat{\zeta}_X$

   (f) Randomly draw new parameters, $\hat{\zeta}_X^*$ based on the fitted model in step (e)

   (g) Use the parameters $\hat{\zeta}_X^*$ to generate stochastic imputations for X_{mis} to update X_{imp}

   (h) repeat steps (b)-(g) for a specified number of cycles (until convergence) in order to generate the first set of imputations.

(i) repeat steps (b)-(h) m times in order to generate m imputed data sets

Since the conditional models in this example are simple linear regressions (and variables are assumed to be conditionally normally distributed), they are known to adhere to a joint-multivariate normal distribution. Therefore, in this case, FCS is equivalent to the joint-multivariate normal modelling approach (Hughes et al., 2014). Alternatively, each conditional imputation model can be chosen based on the type of variable that is being imputed. For example, a logistic regression may be used to impute a binary variable. Although this flexibility of FCS may provide an advantage when imputing discrete data (van Buuren, 2007), there is no guarantee that the conditional models will be compatible with an existing joint model (Raghunathan et al., 2001; van Buuren et al, 2006). Despite this theoretical limitation, there are many simulation and empirical studies which show good performance of this method even when each conditional model is not clearly compatible with a joint distribution (Lee & Carlin, 2010; Van Burren et al., 2006; Van Burren, 2007; White et al., 2011)

3. *Non-parametric (implicit)*: To this point, discussion has focused on specifying an explicit parametric MI model in order to generate imputations. In contrast, implicit model (non-parametric) MI algorithms allow for the relaxation of these modelling requirements and can be advantageous in many settings. Broadly, these methods involve imputing missing values with observed values from other individuals. This is usually accomplished by identifying a suitable donor individual or set of donors (a donor "pool") from which to draw imputations for an individual with missing values (Andrige & Little, 2010). Non-parametric imputation exploits auxiliary information from the observed data in the formation of donor pools in the hopes of rendering the missingness mechanism MCAR within each donor pool (Andrige & Little, 2010). This may be accomplished through formation of cross-classification cells based on key variables or the estimated propensity to respond. Alternately, nearest

neighbour methods are based on identifying a single donor or set of k donors that minimize a specified measure of distance to the case with missing values (Andridge & Little, 2010; Carpita & Manisera, 2011; D'Orazio , 2011; Siddique & Belin, 2008; Jonsson Wohlin, 2004). Randomly selecting donors from these defined donor groups allows for preservation of the natural variability in the data set.

In order to incorporate the uncertainty associated with imputing values, donors can be repeatedly selected from within the donor pools to perform multiple imputation. Care must be taken, however, to ensure that this repeated sampling constitutes "proper" multiple imputation so that Rubin's rules hold (Rubin, 1987); the commonly-employed hot deck imputation (Andrige & Little, 2010) for example, does not result in "proper" MI, but with only slight modifications to this approach, *Approximate Bayesian Bootstrapping* can be employed and Rubin's rules will be valid (Koller-Meinfelder; 2009; Rubin & Schenker, 1986).

Other procedures which allow for relaxation of parametric modelling requirements, such as random forest procedures, have been implemented (Shah et al., 2014; Stekhoven & Buhlmann, 2012). Random forest imputation is an extension of classification and regression trees, which allow flexible, non-linear modelling by findings optimal cut points in predictor variables to recursively subdivide the data (Shah et al., 2014). Since these algorithms do not rely on assumptions about specified parametric models, they are more capable of capturing complex non-linear relationships and interactions between variables (Stekhoven & Buhlmann, 2012; Shah et al., 2014).

4.1.1 Congeniality During Multiple Imputation

Establishing a suitable model (implicit or explicit) for imputation is imperative for achieving unbiased results following MI (Meng, 1994). Firstly, the imputation model should be cor-

rectly specified, which can be challenging as the data do not usually conform to a convenient model. Secondly, the imputation model must accurately maintain all relationships within the data that will be part of any subsequent analyses (i.e. the model must be congenial). Fortunately, with moderate amounts of missingness, imputation methods are typically quite robust to violations of the underlying assumptions of the imputation model (Allison, 2001; Schafer, 1997). Uncongeniality, on the other hand, can readily lead to biased results and emphasis should be placed on ensuring that the imputation model reflects necessary data features and relationships (Carpenter & Kenward, 2012; Meng, 1994; Schafer, 1997). Therefore, an imputation model should minimally include all variables and relationships present in the substantive analysis (i.e. the response model of interest). This includes interactions and other non-linear relationships, as well higher-order data structures. For example, data with a clustered structure may be analysed appropriately with a multi-level model and, therefore, an appropriate imputation model should also be multi-level (this topic is addressed in more detail in the following sections; Andridge, 2011; Reiter et al., 2006).

A key advantage of MI is the ability to utilize auxiliary variables to make the assumption that the data are MAR more plausible; inclusion of auxiliary variables can also serve to improve the precision of the imputations (Carpenter & Kenward, 2012). However, when auxiliary variables are included in the imputation model, the model for imputation and the model used for analysis will not match; the imputation model is no longer considered strictly congenial (Meng 1994). An imputation model is considered "richer" than the substantive model when it allows for the relationships of interest within the substantive model as well as incorporating auxiliary variables or relationships, so that it contains the congenial imputation model nested within it (Carpenter & Kenward, 2012; Schafer, 2003). Alternatively, an imputation model may also be "poorer" than the substantive model when variables or relationships in the substantive model are missing from the imputation model. Uncongeniality from a richer

imputation model is not a practically important problem and is, in fact, considered beneficial due to the increase in plausibility of the MAR assumption and the increase in efficiency from added information (Meng 1994; Rubin, 1996; Schafer, 2003). In contrast, if uncongeniality results from a poorer imputation model, MI will not adequately maintain the relationships between variables and may introduce new biases rather than reducing biases inherent in the complete-case data (Meng 1994; Rubin, 1996; Schafer, 2003). A rich imputation model with a potentially large number of variables, interactions, and higher-order relationships can be challenging to capture with a parametric model (Sterne et al., 2009). Consequently, non-parametric MI can often be advantageous; non-parametric models do not require distributional assumptions so they can implicitly account for non-linear terms and interactions, and have shown superior performance in cases of high dimensional complex data (Carpita & Manisera, 2011; Liao et al., 2014; Shah et al., 2014; Stekhoven & Buhlmann, 2012).

Note that, inclusion of too many auxiliary variables can lead to problems with identifiability or convergence of estimation procedures and can increase small-sample variance (Carpenter & Kenward, 2012; Schafer, 2003). Therefore, auxiliary variables should be chosen carefully; they can generally be selected if they satisfy either of two criteria described by Carpenter and Kenward (2012, pg 72):

1. The variable is predictive of missingness as well as with the variable being imputed, or
2. The variable is predictive of the variable being imputed.

4.2 Multiple Imputation in Complex Surveys

Using MI to handle missing data in complex survey settings introduces additional challenges for ensuring congeniality (Andridge, 2011; Carpenter & Kenward, 2012; Reiter et al., 2006, 2012; Seaman et al., 2012). It has been repeatedly demonstrated that failure to account

for survey design features (e.g. survey weights and clustered data) when employing MI may result in biased estimators and poor confidence interval coverage (Andridge, 2011; Carpenter & Kenward, 2012; Reiter et al., 2006; Seaman et al., 2012). Beyond sampling features, complex survey data often have other characteristics that make it challenging to specify congenial imputation models. These include, collections of mixed variable types (continuous and discrete), sets of multi-item scales, and potential variable restrictions or skip patterns. These features add additional challenges when using MI for complex surveys and must also be carefully considered (Carpenter & Kenward, 2012; Eekhout, 2014; Gottschall et al., 2012; Lee & Carlin, 2010; Yucel, 2011).

4.2.1 Categorical Data

Survey data commonly consist of a mix of categorical and continuous variables, however, except in special circumstances where the number of variables is small, methodological development of MI has typically focused on the optimality of the joint-multivariate normal modelling approach (Carpenter & Kenward, 2012). Fortunately, many simulation studies have shown that imputation procedures are quite robust to departures from the assumption of joint normality (Bernaards et al., 2007; Kropko et al., 2013; Lee & Carlin, 2010; Yucel, 2011). Categorical variables can be imputed using models for continuous variables by assuming that they are discretized realizations of an underlying latent normally-distributed continuous random variable. For example, dichotomization of a latent normal variable results in binary data. This approach has been shown to perform well for binary variables (or nominal variables considered as a set of binary variables) and ordinal variables, especially when limited constraints are applied and no post-imputation rounding is used (Bernaards et al., 2006; Demirtas, 2009; Horton et al., 2003; Jia & Enders, 2015; Rodwell et al., 2014; von Hippel, 2013). Specialized rounding procedures following imputation of categorical data as

continuous variables can be employed when the variables must maintain to their categorization, although not rounding is preferable (Bernaards et al., 2006; Demirtas, 2009; Horton et al., 2003). In some cases, the latent normal approach may distort non-linear relationships between outcomes and exposures (Lee et al., 2012). In these circumstances, imputing an ordinal variable as a set of binary dummy variables (as with nominal data) may be preferable (Lee et al., 2012).

The FCS approach allows for imputation of categorical variables through specification of logistic, multinomial, or other imputation models for non-normally distributed data (Van Buuren, 2006). The FCS approach may modestly out-perform the joint modelling approach when imputing discrete variables (Van Buuren, 2007; Yu et al., 2007). In general, however, differences in the performance of the two approaches are negligible if the joint modelling approach appropriately employs the techniques described in the previous paragraph (Bernaards et al., 2006; Carpenter & Kenward, 2012; Demirtas, 2009; Horton et al., 2003; Jia & Enders, 2015; Rodwell et al., 2014; von Hippel, 2013). It should also be noted that using the FCS approach with categorical regression models involves the risk of perfect prediction. Perfect prediction, occurs when, for a certain combination of covariate values there is no variation in observed values of the outcome variable (White et al., 2010, 2011). Perfect prediction (or near-perfect prediction) prevent the iterative FCS approach from converging and manifest through unpredictable parameter estimates and inflated standard errors (White et al., 2010, 2011). Reducing the number of categorical auxiliary variables in the imputation model and, where possible, collapsing variables with a large number of categories can reduce the risk of perfect prediction. When these solutions are not practical, simple adjustments such as adding a few extra observations to the incomplete variable (known as data augmentation) can successfully avert perfect prediction (White et al., 2010, 2011).

Finally, non-parametric imputation methods, such as the aforementioned nearest neighbour

and random forest procedures, can appropriately handle categorical variables without making strong assumptions about the underlying distributions (Andridge & Little, 2010; Carpita & Manisera, 2011; Jonsson & Wohlin, 2004; Liao et al., 2014; Stekhoven & Buhlmann, 2012). Since values for imputation are selected directly from the observed values, these methods ensure that all imputations are within a correct range and adhere to categorical boundaries. Some studies have shown superior performance of these methods over fully parametric approaches in cases with large numbers of mixed variables (Carpita & Manisera, 2011; Liao et al., 2014; Shah et al., 2014)

4.2.2 Composite Measurements

Another common feature of survey data is the presence of multi-item scales and measures (Eekhout et al., 2014; Gottschall et al., 2012; Shrive, 2006). These generally are summed or combined in some other manner, in order to provide a single score for analysis (van Ginkel et al., 2015). Missing responses for individual measurement items impair the calculation of the summary score, and therefore individuals who fail to respond to some of the items are usually treated as completely missing (van Ginkel et al., 2015). Although it is common for researchers to replace missing score items by the mean of the remaining items for an individual (or a variants on this method; van Ginkel et al., 2015), MI is a more suitable approach (Eekhout et al., 2014; Gottschall et al., 2012; Shrive, 2006).

There are two clear options for imputing these types of measures: either imputing the final composite score or imputing each item individually and combining to get the final score. Imputing the items themselves allows for available information from partially-observed measures to be used both in the calculation of the composite score and in the imputation of any missing items. Therefore, compared to imputing the final composite measure, the "impute by item" approach can improve efficiency, and could also reduce bias if the partially-

observed items help render the data MAR (Eekhout et al., 2014; Gottschall et al., 2012).

The "impute by item" approach can be uncongenial to the substantive model when it is based on a composite measure that is a non-linear combination of its individual items. Consider for example, *passive* imputation of BMI, accomplished by imputing height and weight variables then calculating BMI from these imputed values (Morris et al., 2014). Since BMI is not a linear function of height and weight, bias may be introduced when BMI is passively calculated from imputations of height and weight, instead of being *actively* imputed. However, if log BMI is of interest, then it may be passively calculated from imputations of log height and log weight since the relationship between these variables is linear on the log scale (Morris et al., 2014).

Despite recommendations for the "impute by item" approach, this strategy can inevitably complicate the imputation model. Multi-item measures are often comprised of ordinal Likert-type scales which necessitate appropriate categorical variable considerations. Various methods such as FCS imputation, normal distribution truncation, ad hoc rounding, and transformations have been attempted during the imputation of Likert-type variables (Jia & Enders, 2015; Rodwell et al., 2014; Lee et al., 2012). It has generally been found, that the most accurate imputations and most suitable maintenance of parameter estimates are achieved by imputation under a normal model without any transformation or truncation (Jia & Enders, 2015; Rodwell et al., 2014; von Hippel, 2013). The inferior performance of the FCS approach for Likert type items may be due to the sensitivity of categorical regression models to sparseness in ordinal data (Jia & Enders, 2015). Furthermore, imputing the Likert-type items with a set of polytomous regression models which are then summed together may perform well at maintaining the marginal distribution of the individual items, but may perform poorly at maintaining a particular linear relationship involving the composite score. As before, non-parametric imputation methods offer an alternative to the complex modelling required

during parametric imputation of each item of a multi-item measure (Liao et al., 2014; Shah et al., 2014; Carpita & Manisera, 2011; Jonsson & Wohlin, 2004).

The challenges surrounding fitting a model with each individual measurement item can be avoided in parametric MI if the composite score of interest is imputed directly. A naive application of this strategy will results in a loss of information and a corresponding loss of efficiency since it fails to exploit the observed items in partially-observed scores (Gottschall et al., 2012; Eekhout et al., 2014). A potential way to regain the information lost from this approach is through the use of rejection sampling. Rejection sampling consists of repeatedly drawing potential values for imputation until a particular criteria or boundary is satisfied (Carpenter & Kenward, 2012). Therefore, imputed values which are not compatible with the observed items for a partially-observed individual can be rejected and resampled. That is, imputations for the composite scores are actively drawn and any imputed values that are impossibly large or small given the partially-observed data are rejected. More details about this approach are given by Carpenter and Kenward (2012, pg. 143 & 201).

4.2.3 Survey Weighting

An imputation model which ignores sampling weights will be uncongenial to a weighted substantive model and may lead to biased results (Andrige & Little, 2009; Carpenter & Kenward, 2012; Kim et al., 2006; Seaman et al., 2012). Ideally, imputation would proceed separately in each strata defined by the weights, ensuring that the weights would no longer be informative in the imputation model. However, this would require specification of many different imputation models in strata with potentially small sample sizes. Although seemingly a reasonable alternative, simply weighting the imputation model is not sufficient to eliminate potential biases if the weights are informative and are related to the missing data (Kim et al., 2006; Carpenter & Kenward, 2012). Instead, relationships in the imputation

model should be allowed to differ within strata defined by the weights.

Building on this notion, Seaman et al., (2012) note that in order for MI estimates to be consistent in the presence of sampling weights, the imputation model should include the weights as covariates and be correctly specified. This implies that is it necessary to include the sampling weights in the imputation model, along with all relevant interactions with the sampling weights. This can quickly lead to the inclusion of a large number of variables and interactions, increasing the risk of near-perfect prediction and identifiability problems (Carpenter & Kenward, 2012; Seaman et al., 2012). Therefore, a compromise is necessary between the including all possible design information in the imputation model and fitting a robust model with a more reasonable amount of key design variables (Carpenter & Kenward, 2012; Seaman et al., 2012).

An approach that aims to resolve the issues of complex model specification in the presence of weights, is based on the weight smoothing methods utilized by Elliott and Little (2000). During analysis of complete data, large weight ranges can result in high variance estimates. As a resolution, Elliott and Little (2000) use multi-level models in which strata formed based on sampling weights are treated as level-two units. Carpenter & Kenward (2012) propose extending this approach for MI, by allowing strata based on sampling weights to define level-two units in a multi-level imputation model. Although their preliminary research shows good performance of this method, there is no clear way to implement it when it is also desired to use a multi-level imputation model to account for clustered data (a method which is discussed in more detail in Section 4.2.4).

Several methods of incorporating sampling weights during implicit (non-parametric) imputation have been considered. For example, sampling weights may be used to modify the probabilities of donor selection, or to place a restriction on the number of times a respondent can act as a donor (Andrige, 2009; Andrige & Little, 2010). However, analogously to

weighting an explicit imputation model, neither of these approaches are ideal since each will fail to eliminate bias if the weights are informative and response probability is not constant within an adjustment cell (Andrige, 2009; Andrige & Little, 2010). Since the goal of implicit imputation methods is to create imputation cells which are homogeneous with respect to the outcome and probability of response, a more suitable approach is to use the sampling weights (or sufficient design variables) along with other auxiliary variables in formation of the donor pools (Andrige, 2009; Andrige & Little, 2010).

4.2.4 Clustered Data

The higher-level structure of clustered data should be accounted for during imputation in order to ensure congeniality. Ideally, MI would proceed independently within each cluster in order to avoid the need to capture the complex data structure in the imputation model (Graham, 2009). However, this approach is usually not feasible due to small cluster sizes, and involves the added challenges of fitting a separate imputation model for each cluster. From a theoretical standpoint, using a multi-level imputation model is an ideal approach to ensure congeniality (Carpenter & Kenward, 2012; Drechsler, 2015). However, researchers often ignore the clustered structure entirely during imputation, or attempt to account for it with the inclusion of fixed parameters for cluster status indicators in the imputation model (Drechsler, 2015). Although these methods are less computationally intensive and can be readily implemented in standard software packages, the consequences of these uncongenial approaches can be challenging to predict. (Andridge, 2011; Carpenter & Kenward, 2012; Diaz-Ordaz et al., 2014; Dong, 2014; Drechsler, 2015; Reiter et al., 2006; Taljaard et al., 2008).

First, consider disregarding the clustered structure entirely by imputing under a single-level model (without cluster indicators as covariates). The imputed values will not reflect

the homogeneity within clusters, and will therefore have an inappropriate variance structure. Cluster specific imputed values will be biased towards the grand overall mean leading to overestimation of the variability attributable to within-cluster relationships and underestimation the variability attributable to between-cluster relationships. This results in underestimated variance of random components and the related measure, ICC (Andridge, 2011; Diaz-Ordaz et al., 2014; Drechsler, 2015; Reiter et al., 2006; Taljaard et al., 2008; Van Buuren, 2011). The inappropriate variance structure of imputed values resulting from this approach also manifests in biased estimators of fixed parameter variances, the direction of which depends on the extent to which the estimator is based on within-cluster or between-cluster effects (Carpenter & Kenward, 2012). Variances for relationships which largely occur between clusters will be underestimated while variances for relationships which occur primarily within clusters will tend to be overestimated (Carpenter & Kenward 2012; Reiter et al., 2006). Finally, in the case of unbalanced cluster sizes, use of a single-level imputation model introduces similar problems as the use of a single-level model for analysis with unbalanced clusters: fixed parameter estimates following imputation may be unnecessarily skewed towards the relationships seen in the largest clusters (Carpenter & Kenward, 2012; Gelman & Hill, 2006).

Including fixed parameters for cluster indicators variables in the imputation model may improve on disregarding the clustered structure entirely by allowing for estimation of cluster specific means. However, it remains uncongenial to a multi-level analysis model as it represents the limiting case where ICC tends to 1 (Andridge, 2011; Drechsler, 2015; Van Buuren, 2011). Therefore, this approach biases imputed values towards cluster specific means, thereby inflating the differences between clusters leading to overestimated variance of random components and overestimated ICCs (Andrige, 2011; Drechsler, 2015; Van Buuren, 2011). Likewise, fixed parameters and their variances may be biased in an opposing fashion to that

described above during imputation under a single-level model without cluster indicators as covariates.

To correctly account for clustered data, a multi-level imputation model should be used (Andridge, 2011; Diaz-Ordaz et al., 2014; Drechsler, 2015; Reiter et al., 2006; Taljaard et al., 2008; Van Buuren, 2011). With the joint modelling approach, parameters of the imputation model can be generated with a MCMC algorithm as described earlier with an additional step to draw random effects at each iteration (Schafer, 2001; Schafer & Yucel, 2002). Studies implementing multi-level MI with the joint normal modelling approach have shown suitable maintenance of ICC values and nominal confidence interval coverage of parameter and variance estimates (Andridge, 2011; Dong 2014; Mistler 2015; Zhao & Yucel, 2009).

The FCS approach has a natural extension to multi-level imputation, where it involves fitting a separate conditional multi-level model for each incomplete variable. For example, multi-level binary variables may be imputed through specification of conditional logistic multi-level models (Yucel et al., 2006; Zhao, & Yucel, 2009). Simulation studies comparing the FCS and joint modelling approaches during multi-level imputation, have shown generally equivalent performance of the two methods (Dong, 2014; Mistler, 2015; Zhao & Yucel, 2009). Currently, implementation of multi-level FCS imputation in publicly available software is limited to conditional specification of linear models.

A potential disadvantage to the use of multi-level imputation models is the necessity of making more extensive modelling assumptions compared to single-level models. Specifically, the assumption of normally distributed random effects is required. This may be challenging to justify in some cases, but both the joint modelling and the FCS approach have been shown to be quite robust to violations of this assumption (Dong, 2014; Yucel and Demirtas, 2010).

4.3 Summary

Multiple imputation has become a very popular approach for dealing with item non-response in survey settings (Little & Rubin, 2014; Reiter et al., 2007; Schafer & Graham, 2002; White, Royston, & Wood, 2011). It is clear that the model used for imputation must be congenial to the model used for analysis, which becomes more challenging in complex survey settings (Andridge, 2011; Carpenter & Kenward, 2012; Kim et al., 2006; Meng, 1994; Reiter et al., 2006; Seaman et al., 2012). Inclusion of survey weights may be necessary, which can complicate model fitting (Kim et al., 2006; Seaman et al., 2012). Multi-level imputation models are necessary to ensure congeniality in the presence of clustered data, but these can be computationally intensive and only a limited number of software packages currently implement such methods (Andridge, 2011; Drechsler, 2015; Reiter et al., 2006). Finally, the large amount of categorical data paired with multi-item measures further complicate these procedures (Bernaards et al., 2006; Eekhout et al, 2014; Gottschall et al., 2012; Lee & Carlin, 2010; Yucel, 2011). The extent to which disregarding these aspects of complex surveys during MI may affect analysis results in real-data settings is not clear. The remainder of this report is dedicated to investigating this problem through comparative application of MI methods within the HBSC.

5 Application: Multiple Imputation within the Health Behaviour in School-aged Children Study

Most variables within the HBSC data set have some proportion of missing data, and for analyses which involve multiple variables, this can add up to a considerable portion of data being ignored during a complete-case analysis. Therefore, missing data within the HBSC must be handled appropriately, and MI is a highly advantageous solution (Little & Rubin,

2014; Reiter & Raghunathan, 2007; Schafer & Graham, 2002; White, Royston, & Wood, 2011). Three research questions of interest within the HBSC were identified as being of special interest. Missingness patterns the variables involved in the selected research questions (outlined in the following section) are described below, and can be seen in Table 1.

5.1 Research Questions

Three substantive research questions of interest within the HBSC were chosen to illustrate the potential effects of different missing data methodology. Two of these research questions related to the association between hunger during childhood and specific health outcomes, while the third examined the relationship between the health outcomes themselves.

Hunger during childhood is a prognostic factor for many negative health outcomes including obesity, lower health related quality of life, poorer emotional health, and many physical symptoms (Nackers & Appelhan 2013; Niclasen et al., 2013). While several studies have examined the various health consequences of hunger in adults, only recently was this topic investigated within the Canadian HBSC, providing one of the only large-scale studies on the topic in adolescents (Pickett, Michaelson, & Davison, 2015). The HBSC study collects data on self-reported hunger due to inadequate food supply, and many health outcomes both psychological and physical in nature. The study by Pickett, Michaelson, and Davidson (2015) found relationships between this type of hunger and a number of negative emotional, physical, and social outcomes. These findings motivated the further examination of these same topics during the present investigation. The health outcomes selected for the present investigation were chosen, in part, based on the occurrence of missing data in the corresponding variables. These included measures of psychosomatic symptoms and body mass index (BMI), which acts as a proxy measure of adiposity. Investigation of the relationships between BMI and psychological health outcomes in adolescent populations at the scale of

the HBSC is limited, further motivating examination of the relationship between these two selected health outcomes (Erickson et al., 2000; Ford et al., 2001).

In summary, the following research questions were investigated while employing a variety of MI methods to treat the missing data:

1. *Analysis one*: How is hunger due to inadequate food supply related to the outcome of psychosomatic symptoms?

2. *Analysis two*: How is hunger due to inadequate food supply related to the outcome of adiposity?

3. *Analysis three*: How is adiposity related to the outcome of psychosomatic symptoms?

5.2 Variables and Missingness

Hunger

The HBSC measures students' perception of hunger using the question: "Some young people go to school or bed hungry because there is not enough food a home. How often does this happen to you?" to which students may respond with one of four options including: "always", "often", "sometimes" and "never". This measure has been used as an indicator of child hunger, and as a proxy for socioeconomic status and food availability (Pickett, Michaelson & Davison, 2015). As suggested by previous research involving this hunger variable, response categories "always" and "often" were grouped together to form one category called "frequent", since there is a small number of respondents in each of these categories (Pickett, Michaelson, & Davidson, 2015). This hunger variable has very few missing values, only approximately 0.8% (as reported in Table 1).

Table 1: Item non-response in variables that are part of substantive analyses

Variable	% Missing
Hunger	0.8
Psychosomatic symptoms[a]	8.4
Headache	2.8
Stomache Ache	3.1
Backache	3.9
Feeling Low (Depressed)	3.8
Irritability	3.6
Nervousness	3.7
Difficulty Sleeping	3.3
Diziness	2.9
BMI	23.0
Height	16.5
Weight	15.7
Family affluence	9.0
Family structure	3.7
Immigration status	1.1
Age	0.9
Total[b] for *Analysis 1*: Association between hunger and psychosomatic complaints	17.7
Total for *Analysis 2*: Association between hunger and adiposity	29.6
Total for *Analysis 3*: Association between adiposity and psychosomatic complaints	33.3

[a]Percentage of individuals who failed to respond to at least one of the 8 psychosomatic symptoms items

[b]Total percentages for each analysis indicate the percent of individuals with at least one missing value for a required variable; this is the percent of individuals which were removed from the respective complete-case analysis.

Psychosomatic Symptoms

The HBSC asks students to report the frequency (on a 5 point Likert-like scale ranging from "about every day" to "never") of experiencing 8 psychosomatic symptoms: headache, stomach ache, backache, feeling low (depressed), irritability, nervousness, difficulty sleeping, and dizziness. The 8 items are typically totaled to form a composite score with strong

psychometric properties (Hetland et al., 2002). This score was then dichotomized to form an outcome variable which indicated either frequent (on average, daily or weekly), or not frequent reporting of symptoms. Approximately 8.4% of survey participants failed to respond to at least one psychosomatic symptoms (PS) score item, so PS score was missing 8.4% of the time; non-response for the 8 individual PS score items ranged from 2.7% to 3.9% (as reported in Table 1).

Adiposity

Self-reported height and weight is collected in the HBSC, and BMI was calculated from these values. Although BMI measured in this fashion is subject to measurement error, it considered an acceptable measure of adiposity (Booth et al., 2000). Students were classified as "normal" "overweight" or "obese" based on sex and age-specific percentiles established by the WHO. For the analysis in which adiposity was the outcome of interest, "overweight" and "obese" individuals were grouped together to form a single category (and therefore a dichotomous outcome). Approximately 22.6% of students reported neither their height nor weight, while 16.5% reported only their height and 15.7% reported only their weight. In addition, approximately 0.45% of students reported heights and weights that combined to give extreme BMI outliers (outside of a realistic range of 11 to 37), so these unrealistic values were deleted. Therefore, BMI was unknown for approximately 23.0% of the sample.

Covariates

In accordance with past research, a standard set of individual-level confounders was controlled for during these analyses which included, family affluence (as a proxy measure for socioeconomic status), family structure, immigration status, age, and grade level (Pickett, Michaelson & Davison, 2015).

Family affluence is a validated measure of socioeconomic status (Currie et al., 1992). It was

measured by asking students four questions about the material conditions of their household. These included whether students had their own bedroom, family vehicle ownership, family computer ownership, and family holidays. Following past research practices, responses from these four items were then totaled to create a nine point Family Affluence Scale (FAS) which was categorized into low, medium, and high family affluence (Pickett, Michaelson, & Davidson, 2015). Approximately 9.0% of the survey participants failed to respond to at least one family affluence item, so total FAS was missing 9.0% of the time. The four family affluence items each had between 7.5% and 7.9% non-response.

Following past research, family structure was based on the number and type of adults living in the participant's primary home (Pickett, Michaelson, & Davidson, 2015). Constructed categories included: both mother and father at home, one of mother or father at home, and "other" family structures. Family structure was missing for the 3.7% of individuals who did not provide any information regarding the caregivers living in their primary home.

Immigration status of participants was determined by the length of time they had been in Canada as collected by the question, "How many years have you lived in Canada?" (to which 1.1% individuals failed to respond). Because of the small numbers in certain groups, the responses were collapsed to form three categories: born in Canada, immigrated recently (within 1-5 years), and immigrated not recently (greater than 5 years ago).

Age was calculated from participant's birthdate, which 0.9% of students failed to report. There are no missing values for grade level since this information is known during sampling. Gender (missing 0.12% of the time) was not controlled for as a confounder, but involved in the classification of participants into BMI categories (see above) and is considered relevant as a descriptive covariate.

Distributions of the three main variables of interest in the present investigation (hunger,

psychosomatic symptoms, and adiposity) by the covariates described in this section can be seen in Table 2. Some of the trends that appear to exist within in Table 2 highlight the importance of missing data treatment. For example, upon first examination it may appear as though younger grades have higher percentages of obese students. In actuality, this pattern is likely attributable to the higher occurrence of missing BMI values in these groups.

5.3 Complete-case Analysis

Performing a complete-case analysis (CCA) prior to implementing MI is necessary to fit and evaluate the substantive models of interest using the complete data. The information gathered through this process is used during development of a congenial imputation model.

Methods

Multi-level logistic regression was used to model the dichotomous health outcomes described in Section 5.2 while accounting for the clustered nature of the data. Since there was often only one class sampled from each school, the differences between classes are likely largely indistinguishable from differences between schools (Nezlek, 2011). As such, the multi-level model involved two levels, where students are level-one and schools are level-two. Based on model fit, random intercept models were considered most suitable. The set of confounders described above were controlled for during each analysis, which included family affluence, family structure, immigration status, age, and grade. The total percentages of missing values for each analysis can be seen in Table 1; these percentages of the data are ignored during this CCA.

Table 2: Key variables and missing values as distributed by relevant covariates

Subgroup	Sample (n)	% Who report going to school/bed hungry because there is not enough food at home				% In each of the age-sex specific World Health Organization BMI groups				% Experiencing frequent psychosomatic symptoms		
		Never	Sometimes	Frequent	Missing	Normal	Overweight	Obese	Missing	No	Yes	Missing
Total sample	26078	73.47	21.73	4.01	0.79	57.55	12.24	7.217	23.00	77.33	14.28	8.39
By gender												
Boys	12878	72.07	22.68	4.33	0.92	54.54	14.57	9.28	21.61	81.55	9.58	8.87
Girls	13169	74.86	20.81	3.71	0.62	60.60	9.98	5.19	24.23	73.26	18.89	7.85
Missing	31	64.51	19.36	0	16.13	16.13	6.45	0	77.42	54.84	6.45	38.71
By grade level												
Grade 6	5165	68.25	26.31	4.28	1.16	44.92	10.67	8.03	36.38	77.46	10.40	12.14
Grade 7	5205	72.83	22.96	3.38	0.83	54.93	11.49	6.88	26.71	78.83	11.24	9.93
Grade 8	5266	75.47	19.94	3.84	0.75	58.93	12.34	7.24	21.50	78.73	13.58	7.69
Grade 9	5395	75.89	19.59	3.84	0.68	62.21	13.42	7.06	17.31	76.03	17.59	6.38
Grade 10	5047	74.80	19.93	4.78	0.50	66.79	13.26	6.82	13.14	75.59	18.57	5.85
By immigration status												
Born in Canada	18326	74.81	20.75	3.79	0.64	58.70	12.60	7.32	21.38	77.40	14.89	7.71
Immigrant: recent	2069	71.73	23.05	4.40	0.82	56.16	9.42	4.88	29.53	78.20	13.05	8.75
Immigrant: not recent	5407	69.84	24.63	4.59	0.94	55.28	12.26	7.86	24.60	77.14	12.87	9.99
Missing	276	68.48	19.93	4.71	6.88	36.23	9.42	3.99	50.36	69.93	10.51	19.57
By family structure												
Both parents	16504	76.46	19.82	3.25	0.47	60.87	11.77	7.05	20.30	81.27	12.24	6.50
One parent	7148	70.05	24.51	4.94	0.50	56.56	13.81	7.79	21.83	74.16	18.79	7.05
Other	1470	65.78	26.05	6.73	1.43	44.63	12.65	7.69	35.03	68.30	19.05	12.65
Missing	956	59.21	27.30	6.07	7.43	27.51	7.95	4.71	59.83	44.07	8.37	44.56
By family affluence												
High	15704	77.45	19.13	3.00	0.42	62.11	12.46	6.99	18.44	80.30	13.48	6.22
Medium	6221	70.82	23.74	4.79	0.64	55.39	12.54	7.88	24.19	77.29	15.56	7.15
Low	1797	60.04	31.61	7.35	1.00	46.47	12.35	8.51	32.67	71.79	19.14	9.07
Missing	2356	64.13	26.23	6.20	3.44	41.4	9.89	5.86	42.87	61.93	12.48	25.59
By age												
9-12 years	8457	70.62	24.64	3.87	0.88	49.56	11.08	7.70	31.67	78.51	10.49	11.00
13-15 years	10452	75.35	20.33	3.65	0.67	59.92	12.75	7.31	20.02	77.89	14.72	7.39
15+ years	6944	74.42	20.23	4.74	0.60	65.60	13.28	6.68	14.44	75.36	18.33	6.31
Missing	225	63.56	23.56	4.44	8.44	0	0	0	100	68.0	10.67	21.33

The post-stratification weights available in the HBSC data set are overall unconditional weights, not appropriate for use in the MPML method for weighted multi-level models (Asparouhov, 2006; Rabe-Hesketh and Skrondal, 2006). Furthermore, there is no clear way to suitably approximate the weights necessary for use in multi-level analyses from the weights provided in the data set (see Section 3.4; Rabe-Hesketh & Skrondal, 2006; Stapleton 2012). Therefore, the weights were left out of analyses and a fully model-based approach was used. Dummy variables for grade and province strata status were included as covariates to account for the disproportionate sampling between provinces (Little, 1993,2004).

To briefly evaluate this decision, hybrid (weighted) multi-level regressions were attempted (with the MPML method). These were performed with weights included either in the form they are provided in the data set (standardized to sum to the total sample size), or approximated with a naive method (Zhou, 2014). In order to implement Zhou's naive approximation, the cluster-level weights were approximated as the average of the within cluster weights $\hat{w}_j = \sum_i w_{ij}/n_j$, (where n_j is the sample size of cluster j) and the within-cluster conditional weights were scaled to sum to the cluster sample size $\hat{w}_{i|j} = w_{ij}/\hat{w}_j$. Although there is no evidence that this approximation should perform adequately, it represents a simple and seemingly reasonable approximation that can be implemented with the information available. As discussed earlier, the inclusion of the weights when implementing a model-based strategy should be based on the informativeness of the weights (Asparouhov, 2006). Therefore, these weighted analyses also served to offer guidance in this respect. Since the weighted analyses do not include fixed parameters for grade province strata (as with as the fully model-based approach) estimates of random intercept variance won't be directly comparable. Therefore, the fully model-based analysis was repeated failing to include the grade and province strata dummy variables, in order to examine the effect the weighted approaches have on estimated random intercept variance.

PROC GLIMMIX in SAS 9.4 was used to fit the logistic multi-level models with the inclusion of random intercepts for each school. Fixed parameter variance estimates during the weighted analyses were computed using empirical variance estimators. This specification of the "empirical" computation option is necessary in SAS 9.4 to ensure weights are treated as sampling weights, rather than precision weights, which would lead to a completely different variance estimator.

Results

Results (odds ratios and 95% CIs) from the complete-case analyses are presented in the first columns of Table 3,4, and 5. After controlling for confounders, hunger due to inadequate food supply is significantly associated with the outcome of psychosomatic symptoms: as the frequency of experiencing hunger increases, so do the odds of experiencing psychosomatic symptoms (see Table 3). The results of the analysis examining the association between hunger and the outcome adiposity are not as clear (see Table 4). The intermediate category of hunger experience ("sometimes") is significantly associated with increased odds of adiposity, however, the experience of frequent hunger does not show the same association. Finally, the analysis examining the association between adiposity and the outcome of psychosomatic symptoms shows a significant relationship between increasing adiposity and increased odds of experiencing psychosomatic symptoms (see Table 5).

The results of the explorative analyses which used the available post-stratification weights can be seen in the remaining columns of Table 3, 4, and 5. There are only slight differences between the "model-based" analyses and both "hybrid" methods in terms of fixed parameter estimates (relative to the increased width of confidence intervals), indicating that the weights are not informative past the extent of what is already controlled for in the model (compare column 1 to columns 3-4 of Tables 3-5). Furthermore, including the inappropriate weights did not appear to have notable repercussions.

Notably, all four implemented methods generated different random intercept variance estimates. Inclusion of province and grade strata indicators resulted in much lower random intercept variance estimates since these covariates accounted for a portion of the variability that would otherwise be explained at the school-level. Therefore, as expected, it is more suitable to compare the random intercept variance estimates from the "hybrid" methods to the model-based method which failed to include grade and strata indicators. It can be seen in Tables 3-5 that application of the unconditional single-level weights appears to inflate random intercept variance estimates, as opposed to the approximated multi-level weights in which variance estimates more closely agree with the un-weighted approach. This finding is consistent with the warnings of Rabe-Hesketh and Skrondal (2006) discussed in Section 3.4: use of inappropriate weights can result in biased estimators for random intercept variance. Moreover, fixed parameter estimates following use of the unconditional single-level weights tend to be slightly higher than use of the approximated multi-level weights. That is, the upward bias in random intercept variance that appeared to occur following use of the single-level weights seemed to propagate to fixed parameter estimates, likely due to the ties between fixed parameter estimates and random components variances in logistic multi-level models (as discussed in Section 3.3; Breslow & Clayton, 1993; Demidenko, 2013; Rabe-Hesketh & Skrondal, 2001, 2006). Therefore, the weight approximation method may improve upon the un-approximated single-level weights, but remains unjustified based on current literature (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2012). Overall, the explorative analyses involving the post-stratification weights found them to be largely uninformative and, further, capable of biasing random intercept variance estimation..

Table 3: Odds ratios and 95% confidence intervals (CIs) from complete-case analysis 1: How is hunger related to the outcome psychosomatic complaints?

Variable	Model-based[a]	Model-based ignoring strata[b]	Hybrid with single-level weights[c]	Hybrid with multi-level weights[d]
Hunger				
Often	5.062 (4.346, 5.897)	5.219 (4.48, 6.076)	4.873 (3.963, 5.991)	4.826 (3.928, 5.930)
Sometimes	1.830 (1.674, 2.001)	1.792 (1.640, 1.959)	1.825 (1.622, 2.054)	1.814 (1.613, 2.040)
Never	ref	ref	ref	ref
Random intercept variance (SE)	0.037 (0.0122)	0.064 (0.016)	0.123 (0.022)	0.077 (0.021)

[a]Included province and grade stratum indicators
[b]Failed to included province and grade stratum indicators
[c]Used overall unconditional single-level weights are they are available in the data set
[d]Used naively approximated multi-level weights

Table 4: Odds ratios and 95% confidence intervals (CIs) from complete-case analysis 2: How is hunger related to the outcome of adiposity?

Variable	Model-based[a]	Model-based ignoring strata[b]	Hybrid with single-level weights[c]	Hybrid with multi-level weights[d]
Hunger				
Often	1.155 (0.967, 1.380)	1.157 (0.969, 1.382)	1.192 (0.967, 1.486)	1.186 (0.955, 1.476)
Sometimes	1.179 (1.083, 1.283)	1.181 (1.086, 1.285)	1.249 (1.101, 1.417)	1.242 (1.093, 1.411)
Never	ref	ref	ref	ref
Random intercept variance (SE)	0.065 (0.015)	0.103 (0.017)	0.123 (0.019)	0.091 (0.019)

[a]Included province and grade stratum indicators
[b]Failed to included province and grade stratum indicators
[c]Used overall unconditional single-level weights are they are available in the data set
[d]Used naively approximated multi-level weights

Table 5: Odds ratios and 95% confidence intervals (CIs) from complete-case analysis 3: How is adiposity related to the outcome psychosomatic complaints?

Variable	Model-based[a]	Model-based ignoring strata[b]	Hybrid with single-level weights[c]	Hybrid with multi-level weights[d]
Adiposity				
Overweight	1.391 (1.212, 1.596)	1.392 (1.214, 1.595))	1.403 (1.157, 1.702)	1.393 (1.151, 1.685)
Obese	1.247 (1.114, 1.395)	1.240 (1.115, 1.379)	1.334 (1.139, 1.562)	1.333 (1.138, 1.563)
Normal	ref	ref	ref	ref
Random intercept variance(SE)	0.032 (0.015)	0.054 (0.015)	0.147 (0.031)	0.073 (0.024)

[a]Included province and grade stratum indicators
[b]Failed to included province and grade stratum indicators
[c]Used overall unconditional single-level weights are they are available in the data set
[d]Used naively approximated multi-level weights

In brief, the substantive analyses of interest in the present investigation are based on multilevel logistic regression models with the inclusion of random intercepts for each school. There were significant findings across all three substantive analyses in the complete-case setting. The next step taken was to investigate the changes in these findings following application of MI.

5.4 Multiple Imputation Methods

The results from the above CCA may suffer from bias and a loss of efficiency due to the deletion of individuals with incomplete information (Little & Rubin, 2014). Therefore, efforts were made to remedy these problems by employing a set of MI procedures. MI methods were selected to represent a progressing range of complexity in order to achieve the goals of the present investigation as stated in Section 1. Specifically, the imputation methods differed according to three main characteristics:

i. The extent to which clustered structure of the data was incorporated;

ii. The methods used to impute the composite psychosomatic symptoms (PS) score; and

iii. The parametric modelling assumptions involved.

For the parametric approaches, the imputation models progressively incorporated the clustered structure of the data to ensure congeniality. Since the substantive analyses involve multi-level logistic models with random intercepts, truly congenial MI procedures must be complex enough to also incorporate this multi-level structure in the imputation model. Therefore, to vary complexity three imputation models were employed including: single-level imputation over the whole country, single-level imputation within province, and multi-level imputation (the most congenial approach). For each of these methods, three different approaches were used to impute the composite PS score: imputing each of the 8 individual items, imputing the total composite score with the use of rejection sampling, or imputing the total composite score without rejection sampling. A non-parametric imputation methods was performed in order avoid the restrictions of the strong parametric assumptions in the other models. This involved a k nearest neighbour (KNN) approach based on the distance metric *Gower's distance* (Gower, 1971). See Table 6 for a summary of the implemented methods. For simplicity, the following sections discuss details of the MI methods according to each of the key methodological characteristics.

Table 6: Summary of multiple imputation methods

Method	Description
SLI: Single-level imputation using latent normal models for categorical variables and imputation of 8 PS items	The method does not account for clustering within schools. Here any missing PS score items are imputed individually and the total score is calculated by summing over these (potentially imputed) items. Latent normal models are used for the imputation of categorical variables.
SLI-P: Single-level imputation within provinces using latent normal models for categorical variables and imputation of 8 PS items	The method performs imputation separately within each province. Here any missing PS score items are imputed individually and the total score is calculated by summing over these (potentially imputed) items. Latent normal models are used for the imputation of categorical variables
MLI: Multi-level imputation using latent normal models for categorical variables and imputation of 8 PS items	This method uses a multi-level imputation models which includes random intercepts in order to appropriately account for the clustered structure of the data. Here any missing PS score items are imputed individually and the total score is calculated by summing over these (potentially imputed) items. Latent normal models are used for the imputation of categorical variables
SLI-C: Single-level imputation using multinomial logistic models for cateogorical variables and imputation of 8 PS items	This method does not account for clustering within schools. Here any missing PS score items are imputed individually and the total score is calculated by summing over these (potentially imputed) items. Multinomial logistic regression models are used for the imputation of categorical variables
SLI-PC: Single-level imputation within provinces using multinomial logistic models for categorical variables and imputation of of 8 PS items	This method performs imputation separately within each province. Here any missing PS score items are imputed individually and the total score is calculated by summing over these (potentially imputed) items. Multinomial logistic regression models are used for the imputation of categorical variables
TSLI: Single-level imputation using latent normal models for categorical variables and imputation of total PS score with rejection sampling	This method does not account for clustering within schools. Here the total PS score is imputed (rather than individual items) with the use of rejection sampling. Latent normal models are used for the imputation of categorical variables.
TSLI-P: Single-level imputation within provinces using latent normal models for categorical variables and imputation of total PS score with rejection sampling	This method performs imputation separately within each province. Here the total PS score is imputed (rather than individual items) with the use of rejection sampling. Latent normal models are used for the imputation of categorical variables
TMLI: Multi-level imputation using latent normal models for categorical variables and imputation of total PS score with rejection sampling	This method uses multi-level imputation models which include random intercepts in order to appropriately account for the clustered structure of the data. Here the total PS score is imputed (rather than individual items) with the use of rejection sampling. Latent normal models are used for the imputation of categorical variables
TMLI-NR: Multi-level imputation using latent normal models for categorical variables and imputation of total PS score *without* rejection sampling	This method uses multi-level imputation models which include random intercepts in order to appropriately account for the clustered structure of the data. Here the total PS score is imputed (rather than individual items) but no rejection sampling is implemented. Latent normal models are used for the imputation of categorical variables
KNN: k nearest neighbour imputation	This method utilizes a non-parametric approach which identifies k nearest neighbours based on a distance function, and randomly selects one from which to impute missing values of the recipient. Donors are restricted to belong to the same province and grade.

Parametric versus Non-parametric:

Although the underlying theory of MI is based on parametric modelling, non-parametric (implicit) MI methods have important benefits since they do not require the same distributional assumptions and are less sensitive to model misspecification. Therefore, both parametric and non-parametric MI methods were implemented in the present investigation.

Non-parametric MI could be considered advantageous in the present setting since a large number of mixed-type (categorical and continuous) substantive and auxiliary variables were involved in the imputation, and higher-order and non-linear relationships were not known a priori (Carpita & Manisera, 2011; Liao et al., 2014; Shah et al., 2014).

Implementing non-parametric MI allowed the restrictions of parametric assumptions to be avoided. The non-parametric method employed was a k nearest neighbour procedure with an Approximate Bayesian Bootstrap (as implemented by Koller-Meinfelder, 2009) using a distance function that allows for continuous and categorical variables (Gower, 1971). Donors were restricted to those who had observed values for all the missing values of the recipient, and observed values from selected donor were used to impute the missing values for the partially-observed individual

Fully conditional specification

Joint-multivariate normal modelling and FCS (as described in Section 4.1) are the dominate approaches used for parametric MI. Overall, these methods tend to perform equivalently (Kropko et al., 2013; Lee & Carlin, 2010; van Buuren, 2007), so the FCS approach was selected due to its flexibility, which allowed for implementation of the rejection sampling methods (further described below). Due to software limitations, multi-level imputation was performed with each conditional model specified as a normal linear multi-level regression (implementation of multi-level FCS in publicly available software is limited to conditional

specification of linear models). This required the use of the latent normal model approach for categorical variables (which included the 8 Likert variables of the PS score, family affluence, family structure, hunger, immigration status, and some auxiliary variables). That is, they were imputed with the assumption that they were discretized realizations of some underlying latent normally-distributed continuous variable (following the methods discussed in Section 4.2.1). For comparability, single-level imputation methods were also performed with latent normal model approach for categorical variables (note: since this involves specifying a normal linear regression for each conditional model, these methods are equivalent to the joint-multivariate normal imputation approach; Hughes et al., 2014). However, to take full advantage of the flexibility of the FCS approach, single-level imputation methods were also performed with conditionally specified categorical regression models where appropriate (e.g. multinomial logistic regression models were used for imputation of the 8 Likert variables of the PS score, family structure, hunger, and immigration status).

Multi-level versus single-level versus within province imputation

The goals of the present investigation necessitated a range of MI methodological complexity. The parametric imputation methods were performed using: single-level imputation over the whole country, single-level imputation within province, and multi-level imputation with the incorporation of random intercepts for schools. These represent a progression from disregarding clustering during imputation, to incorporating clustering in the most theoretically appropriate (and most congenial) way. As mentioned above, single-level imputation methods (across the whole country and within province) were performed using both latent normal models and polytomous/multinomial regression models for categorical variables within the FCS algorithm.

Non-parametric imputation methods do not offer a clear ideal approach to incorporating the clustered structure of the data. During the non-parametric k nearest neighbour method

donors were restricted to belong to the same province and grade as the recipient individual and cluster indicators were a component of distance function calculation.

Imputing the multi-item psychosomatic symptoms score

The parametric MI methods imputed the PS score variable based on either each individual score item or the total summed measure. When imputing the total PS score, rejection sampling was implemented to make use of the information available from the partially-observed individual score items (Carpenter & Kenward, 2012). In order to implement the rejection sampling, maximum and minimum achievable values were calculated based on the number of items which were missing. During imputation, any drawn value which fell outside of these maximum or minimum values was rejected and redrawn until meeting these requirements. For example, an individual may have responded to 6 of 8 PS items, for which their responses sum to 25. Since the range of the PS score Likert-type variables is 1 to 5, any values imputed outside of a range of 27-35 would be deleted and resampled. To examine the impact of failing to exploit the additional information present in the partially-observed PS score items, one implemented method imputed the total PS score without rejection sampling (since this approach was only attempted once, it was performed with multi-level imputation).

The latent normal model approach was used for methods imputing the total PS score; the main advantage of the conditionally specified categorical regression models was less relevant when no longer imputing the 8 categorical Likert-type variables. The non-parametric KNN approach only involved imputation of the 8 individual PS score items; the main advantage of imputing the total score was not as relevant when the challenges of parametric modelling were removed.

Imputation model variables

The first step in developing a congenial imputation model is to include all the variables that are present in the substantive analysis (Meng, 1994). Since all three substantive models controlled for the same confounders and otherwise involved three main variables of interest, imputation models were fit to include all variables required for the three substantive analyses (listed in Table 1). In this way, it was only necessary to implement each MI method once, following which, the same imputed data could be utilized for each of the three analyses. Following the recommendations of Morris et al., (2014) BMI was imputed as log transformed height and weight variables during parametric MI (as discussed in Section 4.2). As with the substantive analysis models, fixed parameters were included for grade and province stratum indicators.

Beyond the variables which make up the substantive analyses, it was necessary to identify a set of auxiliary predictors for the imputation models (based on the two criteria mentioned in Section 4.2; Carpenter & Kenward, 2012). Since the HBSC data set involves a large number of potential auxiliary variables, a candidate set was first ascertained through investigation of the literature (Berntsson & Gustafsson, 2000; Cohen & Duffy, 2002; Kalsbeek et al., 2002; Patrick et al., 2004; Paxton et al., 1991 Poikolainen et al., 2000; Swallen et al., 2005). This set of potential predictors was then reduced to a parsimonious set through model selection techniques (forwards and backwards selection into models which already included the variables necessary for the substantive analysis), trial and error, and common sense. Since many of the candidate auxiliary variables were also incomplete, variables which had a large number of simultaneous missing values with the variable to be imputed were considered unsuitable for use an auxiliary variable (van Buuren, 2012). See Table 7 for a list of the auxiliary variables selected for each variable in the substantive model requiring imputation. Although this table displays auxiliary predictors individually for each variable in the

Table 7: Imputation model auxiliary variables

Imputed variable	Auxiliary variables
Hunger	Gender Grade Ethnicity Breakfast habits "How often do you usually have breakfast..." Self-reported quality of life Home life dissatisfaction "There are times I would like to leave home."
Psychosomatic Symptoms	Gender Grade Ethnicity School work `How pressured do you feel by the schoolwork you have to do?..." Self-reported health "Would you say your health is...?" Experiencing bullying "How often have you been bullied at school in the past couple of months?..." Experiencing injury "During the past 12 months, how many times were you injured and had to be treated by a doctor or nurse?..." Self-reported quality of life Home life dissatisfaction "There are times I would like to leave home..." Loneliness " " often feel lonely..." Self-reported wellbeing "Thinking about the last week...have you felt fit and well...felt sad...felt full of energy..."
Height	Gender Grade Age Body Dissatisfaction "Do you think your body is...?... much too thin...much too fat..."
Weight	Gender Grade Ethnicity Self-reported quality of life Body Dissatisfaction Dieting Breakfast Habits Dietary Habits and unhealthy food consumption Physical activity
Family structure, family affluence, immigration status	Grade Ethnicity Immigration status Family affluence Family structure
Age	Grade

substantive analyses, each imputation model included all variables (auxiliary and otherwise). That is, each incomplete variable was imputed based on all other variables. Although this is not strictly necessary when implementing the FCS algorithm, it improves the likelihood that the specified conditional models adhere to an existing joint distribution. As described in Section 4.1, this is a theoretical concern when employing FCS, and although the practical importance of this concern appears limited, the more theoretically valid approach was selected for this application.

Since the FCS algorithm involves fitting univariate imputation models, assessment of imputation model fit was accomplished by fitting each conditional imputation model individually. Residual diagnostic plots, and the reasonability of parameter estimates were examined.

Details of implementation

Each of the imputation methods was specified to produce 25 imputed data sets, which has been shown to be an adequate number to remove noise from estimations (Schafer & Graham, 2002). Although arguments have been made for more imputations (Graham, Olchowski, & Gilreath, 2007), 25 was considered adequate given the computational time required for the more complex imputation procedures. All imputations were completed using R statistical software. The k nearest neighbour method was adapted from the methods implemented in the *StatMatch* package (D'Orazio, 2011). The *mice* package (van Buuren & Groothuis-Oudshoorn, 2011) was used to perform all parametric MI methods (joint modelling and FCS). The *mice* software allows for customization of user specified imputation procedures for use in the FCS algorithm so the rejection sampling procedures could be readily incorporated into the already existing FCS framework. Multi-level imputation was performed by interfacing *mice* with the, *PAN* package (Schafer, 2012). *PAN* implements a Gibbs sampler in order to draw imputed values from a specified joint normal distribution, for which 1000 iterations was specified (Schafer, 2012). Ten iterations between imputations was specified for FCS algorithm

(for all parametric MI methods), which is a recommended number to allow imputed values to be stabilized and independent from each other (Royston & White, 2011). The FCS algorithm was specified to impute variables in order of increasing proportions of missing values to improve the speed of convergence (van Burren, 2012). Trace plots of the mean and variance of imputations for each iteration of the FCS procedure were examined to ensure the algorithm converged and imputations were independent. Ideally, by the point of imputation (the last several iterations) these plots should show no distinct trends and each iteration should be uncorrelated. Examples of such plots can be seen in Figure 1.

Following imputation, data sets were analysed using the GLIMMIX procedure in SAS 9.4 for generalized linear multi-level models, and combined according to Rubin's rules (Rubin, 1987) with the SAS procedure, MIANALYZE. Density plots (or histograms for categorical variables) of imputed versus observed values were generated following imputation to assess the feasibility of imputed values. Several of these plots, based on the first 5 out of m=25 imputation data sets, are presented in Figure 2,3, and 4; the remainder of the plots were very similar to Figures 2-4 and are therefore not presented.

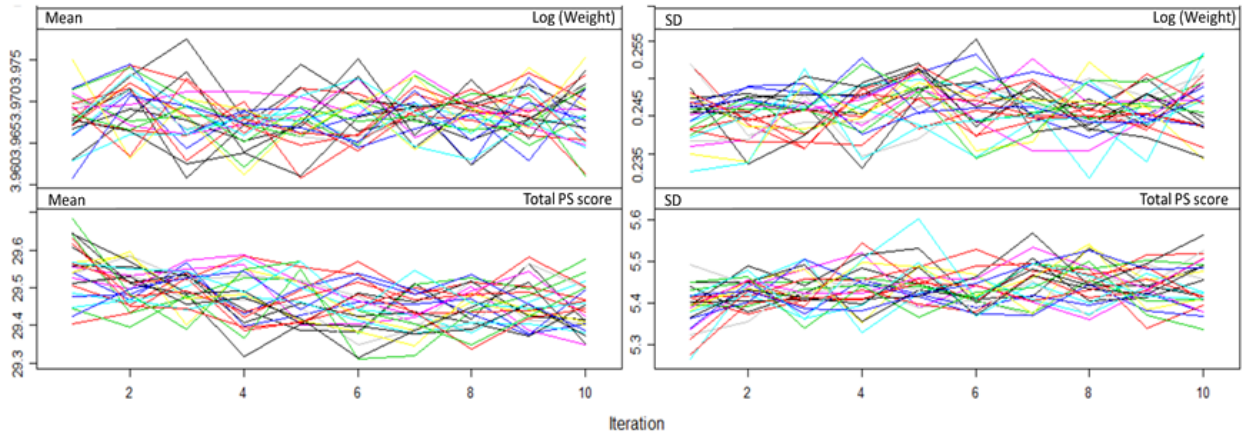
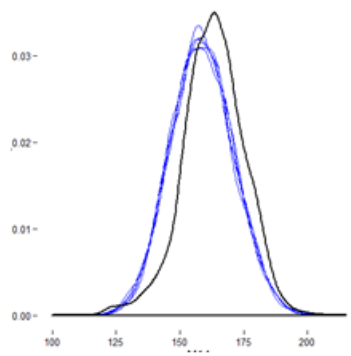


Figure 1: Example of trace plots used to assess convergence of the FCS algorithm. Each line represents a separate imputed data set. The mean and standard deviation of imputed values at each of the 10 iterations are plotted for the log of the weight variable and the total psychosomatic complaints score variable. Plots are shown from the multi-level imputation method imputing the Total PS score.

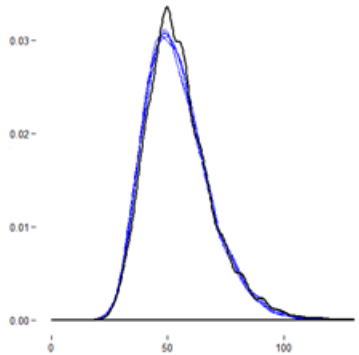
5.5 Multiple Imputation Results

Table 8, 9, and 10 present the results of the substantive analyses following implementation of each multiple imputation method (note: results from methods imputing the total composite PS score are not presented in Table 9 as the PS score was not part of this analysis). Examination of plots for imputed versus observed values can offer some preliminary information (see Figures 2-4). Figure 2 compares the imputed values to the observed values for height and weight variables, as well as the subsequently calculated BMI. From Figure 2, it is clear that imputed values for the height variable systematically deviate from the observed values. This is expected, as younger individuals more frequently fail to report their height, and the inclusion of age and grade in the imputation model has accounted for this. Although younger individuals also tend to more frequently not report their weight, the same systematic difference between imputed and observed values is not seen for this variable. This may reflect the imputation model generating values for missing weight values that are systematically higher than the observed values, which is reasonable based on past literature (Cohen & Duffy, 2002). Finally, the calculated BMI for imputed values show slightly higher proportions of low-range and high-range BMI values compared to observed values. This may be indicative of two co-occurring missingness mechanisms: younger (with a lower BMI) individuals and overweight (as predicted by weight related auxiliary variables) individuals may both have a higher likelihood of failing to report height or weight information.

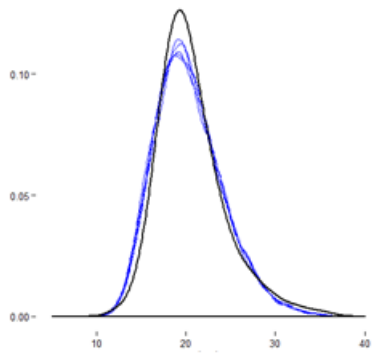
Figure 3 compares the imputed values to the observed values for various methods used to impute the PS score. It can be seen that the PS score was imputed somewhat lower than the observed values. It is not unreasonable, based on past literature, that this is reflective of a systematic difference between respondents and non-respondents in terms of psychological or overall health (Cohen & Duffy, 2002). Further details regarding what can be gathered from this figure are discussed in Section 6.



(a) Height (cm)

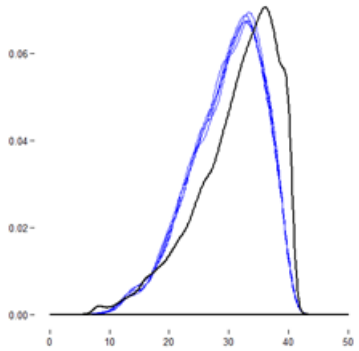


(b) Weight (kg)

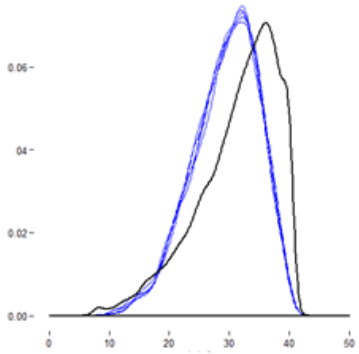


(c) Calculated BMI

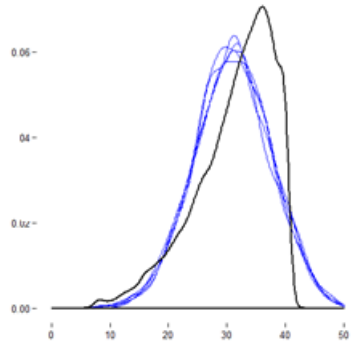
Figure 2: Density plots of observed values (black lines) versus imputed values (blue lines) for height, weight and BMI from single-level imputation.



(a) PS score items imputed and subsequently totaled

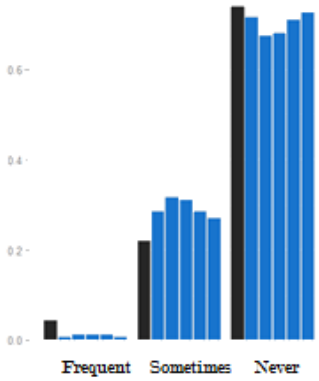


(b) PS total score imputed with rejection sampling

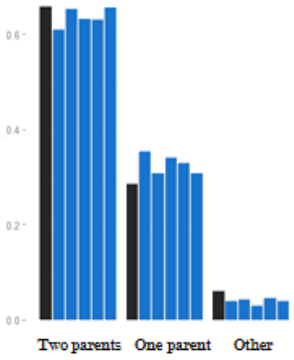


(c) PS total score imputed without rejection sampling

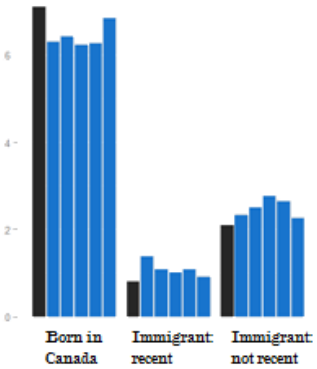
Figure 3: Density plots of observed values (black lines) versus imputed values (blue lines) for the psychosomatic symptoms score from each multi-level imputation method.



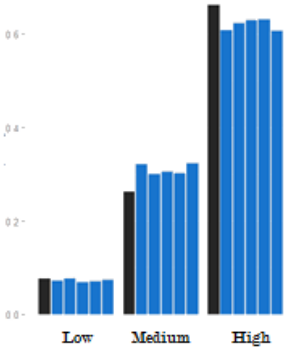
(a) Hunger



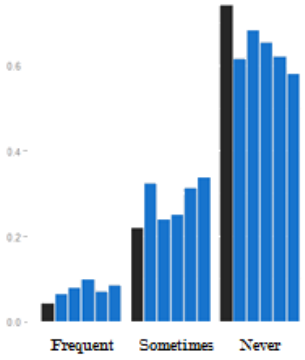
(b) Family structure



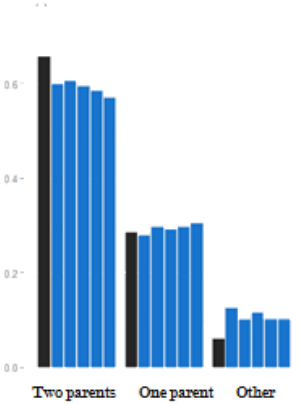
(c) Immigration status



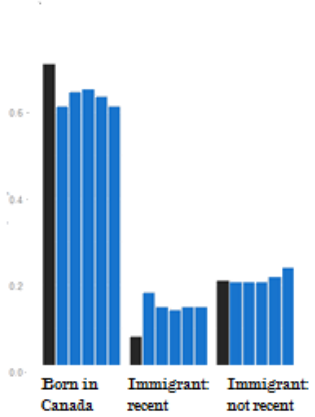
(d) Family affluence



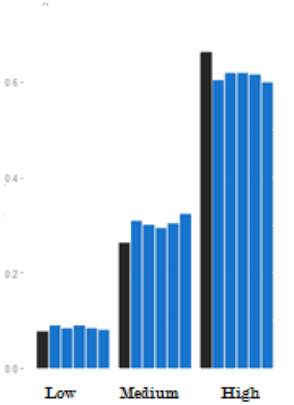
(e) Hunger



(f) Family structure



(g) Immigration status



(h) Family affluence

Figure 4: Histograms of frequencies of observed values (black lines) versus imputed values (blue lines) for categorical variables from single-level latent normal model (top line) and multinomial logistic model (bottom line) imputation methods.

Table 8: Odds ratios and 95% confidence intervals following multiple imputation for analysis 1: How is hunger related to the outcome psychosomatic symptoms?

Variable	Imputation method										
	CCA	SLI	SLI-P	MLI	SLI-C	SLI-PC	TSLI	TSLI-P	TMLI	TMLI-NR	KNN
Hunger											
Frequent	5.062 (4.346,5.897)	4.777 (4.160, 5.486)	4.766 (4.149,5.476)	4.790 (4.177,5.494)	4.783 (4.165,5.492)	4.777 (4.160,5.486)	4.756 (4.141,5.463)	4.721 (4.113,5.419)	4.782 (4.170,5.485)	4.734 (4.105,5.459)	4.649 (4.054,5.332)
Sometimes	1.830 (1.674,2.001)	1.775 (1.636,1.927)	1.775 (1.636,1.926)	1.781 (1.642,1.931)	1.791 (1.652,1.942)	1.787 (1.647,1.938)	1.770 (1.631,1.921)	1.772 (1.633,1.924)	1.784 (1.644,1.935)	1.776 (1.633,1.933)	1.800 (1.660,1.952)
Never	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref
Random intercept variance (SE)	0.032 (0.013)	0.037 (0.011)	0.038 (0.011)	0.039 (0.012)	0.037 (0.011)	0.038 (0.011)	0.038 (0.011)	0.038 (0.012)	0.041 (0.012)	0.040 (0.012)	0.047 (0.012)

Table 9: Odds ratios and 95% confidence intervals following multiple imputation for analysis 2: How is hunger related to the outcome of adiposity?

Variable	Imputation Method						
	CCA	SLI	SLI-P	MLI	SLI-C	SLI-PC	KNN
Hunger							
Frequent	1.155 (0.967, 1.380)	1.189 (1.025, 1.378)	1.196 (1.028, 1.392)	1.202 (1.039, 1.390)	1.183 (1.019, 1.374)	1.191 (1.021, 1.389)	1.176 (1.002, 1.381)
Sometimes	1.179 (1.084, 1.283)	1.170 (1.083, 1.264)	1.170 (1.086, 1.260)	1.171 (1.084, 1.264	1.167 (1.082, 1.259)	1.168 (1.086, 1.255)	1.139 (1.059, 1.226)
Never	ref	ref	ref	ref	ref	ref	ref
Random intercept variance (SE)	0.065 (0.015)	0.047 (0.011)	0.048 (0.012)	0.074 (0.014)	0.047 (0.011)	0.048 (0.011)	0.056 (0.013)

Table 10: Odds ratios and 95% confidence intervals following multiple imputation for analysis 3: How are psychsomatic symptoms related to the outcome of adiposity?

Variable	Imputation method										
	CCA	SLI	SLI-P	MLI	SLI-C	SLI-PC	TSLI	TSLI-P	TMLI	TMLI-NR	KNN
Adiposity											
Obese	1.391 (1.212, 1.596)	1.363 (1.208, 1.539)	1.362 (1.213, 1.528)	1.368 (1.215, 1.540)	1.383 (1.229, 1.558)	1.382 (1.231, 1.551)	1.368 (1.207, 1.552)	1.378 (1.222, 1.554)	1.371 (1.220, 1.540)	1.369 (1.208, 1.552)	1.384 (1.225, 1.563)
Overweight	1.247 (1.114, 1.395)	1.239 (1.125, 1.364)	1.238 (1.124, 1.363)	1.242 (1.121, 1.376)	1.232 (1.115, 1.362)	1.225 (1.114, 1.348)	1.247 (1.131, 1.376)	1.245 (1.124, 1.378)	1.250 (1.134, 1.379)	1.246 (1.123, 1.382)	1.262 (1.141, 1.395)
Never	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref
Random intercept variance (SE)	0.031 (0.015)	0.035 (0.011)	0.036 (0.011)	0.037 (0.011)	0.036 (0.011)	0.036 (0.011)	0.036 (0.011)	0.036 (0.011)	0.039 (0.012)	0.038 (0.012)	0.044 (0.012)

Figure 4 shows the imputed values versus the observed values for the categorical variables imputed from the both the latent normal model and the multinomial logistic model imputation methods. These plots are useful to assess the impact of imputing categorical variables under the assumption of normality. For the most part, the two methods impute categorical variables similarly. Differences can be seen primarily in the categories with small marginal probabilities, particularly in the case of the hunger variable. Since the categorical variables that involve small marginal probabilities for some categories (hunger and, to a lesser extent, family structure) have relatively small amounts of missingness (see Table 1), the impact of these differences was expectedly minor as is apparent in the results of the substantive analyses following MI (see Tables 8-10). Therefore, for simplicity, the latent normal model methods will be focused on during discussion in Section 6.

Now turning to the results presented in Tables 8 9 and 10. Across all three analysis MI was effective in regaining the efficiency lost during the CCA, as apparent through lower SEs. This is evident across all imputation methods. The results of the CCAs showed a significant association between increasing hunger frequency and increased odds of psychosomatic symptoms (see Table 8). Following MI, a slight decrease in the magnitude of these effects occurred, but this was not substantial enough to change the interpretations of the analysis. In the second analysis, the CCA demonstrated a significant association only between hunger "sometimes" and increased odds of being overweight or obese; "frequent" hunger was not significantly associated with adiposity (see Table 9). Following MI the magnitude of these effect estimates were reversed, "frequent" hunger now demonstrating larger effect estimates than hunger "sometimes", and becoming significant at the 5% level. In the final analysis, the CCA showed a significant association between increasing adiposity and increased odds of psychosomatic symptoms (see Table 10). In this analysis, changes in point estimates following MI were minimal and did not impact the findings or conclusions. Notably, across

all three analyses no one particular MI method led to practically important differences in results and, therefore, application of any one of the MI methods would have led to the same findings and conclusions being drawn from these analyses. This has implications for the main goals (stated in Section 1) of this investigation as discussed in Section 6 below.

Beyond the conclusions of the substantive analyses, the results following MI methods show more subtle differences across covariates which are presented Table 11, 12, and 13 (for simplicity, imputation methods not of particular relevance in following discussion were left off these tables). Across covariates the most notable impact of MI can be seen for the covariate immigration status: following MI, the effect of being a "recent" immigrant is no longer significantly associated with decreased odds of psychosomatic symptoms. Following MI, all three analyses show an increase in estimated random intercept variance as the models increasingly account for the clustered data structure (see Tables 11-13). All implemented MI methods generated higher random intercept variance estimates compared to the CCA, except in the analysis with adiposity as the outcome (the analysis with the most missingness in the outcome; Table 12) in which all MI methods resulted in decreased RIV with the exception of the multi-level method. The multi-level imputation methods also resulted in slightly different point and SE estimates. For the most part, the point estimates for the multi-level imputation method were slightly larger in magnitude then those from the single-level methods. Sources for these patterns of results are addressed in Section 6. No substantial differences in point estimates were seen between methods imputing PS score as a composite measure with rejection sampling and methods imputing PS score as set of 8 individual items. Furthermore, only a minimal loss of efficiency was seen when the rejection sampling was implemented. However, imputing the total score without rejection sampling was accompanied with slight differences across covariate estimates and a more notable loss of efficiency. These differences between imputation methods are discussed in detail in the following section.

Table 11: Parameter estimates and associated Standard Errors (SEs) following multiple imputation methods for analysis 1: How is hunger related to the outcome psychosomatic symptoms?

Variable	Imputation method						
	CCA	SLI	SLI-P	MLI	TMLI	TMLI-NR	KNN
Intercept	-3.814 (0.654)	-3.439 (0.582)	-3.449 (0.584)	-3.449 (0.585)	-3.355 (0.587)	-3.191 (0.599)	-3.489 (0.583)
Hunger							
Frequent	1.622 (0.078)	1.564 (0.071)	1.562 (0.071)	1.567 (0.070)	1.565 (0.070)	1.555 (0.073)	1.537 (0.070)
Sometimes	0.605 (0.045)	0.574 (0.042)	0.574 (0.042)	0.577 (0.041)	0.579 (0.042)	0.575 (0.043)	0.588 (0.041)
Never	ref	ref	ref	ref	ref	ref	ref
Immigration Status							
Immigrant: Recent	-0.167 (0.077)	-0.120 (0.070)	-0.123 (0.070)	-0.119 (0.071)	-0.117 (0.071)	-0.144 (0.073)	-0.124 (0.070)
Immigrant: Not Recent	-0.153 (0.051)	-0.123 (0.045)	-0.125 (0.046)	-0.127 (0.046)	-0.133 (0.046)	-0.146 (0.048)	-0.127 (0.046)
Born in Canada	ref	ref	ref	ref	ref	ref	ref
Family structure							
Other	0.500 (0.082)	0.536 (0.073)	0.534 (0.071)	0.530 (0.072)	0.533 (0.072)	0.502 (0.074)	0.512 (0.072)
1 Parent	0.438 (0.043)	0.414 (0.040)	0.412 (0.040)	0.413 (0.039)	0.410 (0.040)	0.414 (0.040)	0.420 (0.040)
2 Parents	ref	ref	ref	ref	ref	ref	ref
Family affluence							
Low	0.249 (0.072)	0.232 (0.067)	0.235 (0.067)	0.232 (0.068)	0.212 (0.066)	0.186 (0.070)	0.242 (0.069)
Medium	0.081 (0.045)	0.068 (0.043)	0.068 (0.043)	0.068 (0.043)	0.063 (0.042)	0.061 (0.042)	0.074 (0.042)
High	ref	ref	ref	ref	ref	ref	ref
Age	0.111 (0.040)	0.093 (0.036)	0.094 (0.036)	0.095 (0.036)	0.090 (0.036)	0.079 (0.037)	0.097 (0.036)
Random intercept variance (SE)	0.032 (0.013)	0.037 (0.011)	0.038 (0.011)	0.039 (0.012)	0.041 (0.012)	0.040 (0.012)	0.047 (0.012)

Table 12: Parameter estimates and associated Standard Errors (SEs) following multiple imputation for analysis 2: How is hunger related to the outcome of adiposity?

Variable	Imputation method				
	CCA	SLI	SLI-P	MLI	KNN
Intercept	0.462	0.758	0.878	0.944	0.815
	(0.630)	(0.544)	(0.544)	(0.558)	(0.564)
Hunger					
Frequent	0.144	0.173	0.179	0.184	0.162
	(0.091)	(0.075)	(0.077)	(0.074)	(0.082)
Sometimes	0.165	0.157	0.157	0.158	0.131
	(0.043)	(0.039)	(0.038)	(0.039)	(0.037)
Never	ref	ref	ref	ref	ref
Immigration Status					
Immigrant: recent	-0.233	-0.195	-0.197	-0.197	-0.186
	(0.074)	(0.065)	(0.064)	(0.067)	(0.066)
Immigrant: not recent	0.038	0.036	0.040	0.043	0.039
	(0.043)	(0.038)	(0.038)	(0.038)	(0.040)
Born in Canada	ref	ref	ref	ref	ref
Family structure					
Other	0.336	0.257	0.250	0.241	0.222
	(0.080)	(0.068)	(0.070)	(0.069)	(0.065)
1 Parent	0.163	0.168	0.170	0.159	0.173
	(0.039)	(0.037)	(0.036)	(0.036)	(0.036)
2 Parents	ref	ref	ref	ref	ref
Family affluence					
Low	0.278	0.223	0.245	0.231	0.176
	(0.070)	(0.061)	(0.064)	(0.062)	(0.063)
Medium	0.131	0.124	0.123	0.114	0.112
	(0.041)	(0.038)	(0.037)	(0.036)	(0.039)
High	ref	ref	ref	ref	ref
Age	-0.126	-0.145	-0.139	-0.156	-0.107
	(0.039)	(0.034)	(0.033)	(0.034)	(0.034)
Random intercept variance (SE)	0.065	0.047	0.048	0.074	0.056
	(0.015)	(0.011)	(0.012)	(0.014)	(0.013)

Table 13: Parameter estimates and associated Standard Errors (SEs) following multiple imputation for analysis 3: How is adiposity related to the outcome psychosomatic symptoms?

Variable	Imputation method						
	CCA	SLI	SLI-P	MLI	TMLI	TMLI-NR	KNN
Intercept	-4.184 (0.732)	-3.596 (0.577)	-3.612 (0.568)	-3.617 (0.578)	-3.516 (0.583)	-3.367 (0.595)	-3.604 (0.579)
Adiposity							
Obese	0.330 (0.070)	0.310 (0.062)	0.309 (0.059)	0.313 (0.060)	0.315 (0.059)	0.314 (0.064)	0.325 (0.062)
Overweight	0.221 (0.057)	0.214 (0.049)	0.213 (0.049)	0.217 (0.052)	0.223 (0.050)	0.220 (0.053)	0.233 (0.051)
Normal	ref	ref	ref	ref	ref	ref	ref
Immigration Status							
Immigrant: Recent	-0.201 (0.089)	-0.091 (0.069)	-0.097 (0.069)	-0.089 (0.070)	-0.087 (0.070)	-0.114 (0.071)	-0.093 (0.069)
Immigrant: Not Recent	-0.136 (0.056)	-0.092 (0.045)	-0.094 (0.045)	-0.096 (0.045)	-0.102 (0.046)	-0.115 (0.047)	-0.095 (0.046)
Born in Canada	ref	ref	ref	ref	ref	ref	ref
Family structure							
Other	0.504 (0.095)	0.570 (0.071)	0.570 (0.070)	0.568 (0.071)	0.570 (0.071)	0.539 (0.072)	0.550 (0.071)
1 Parent	0.486 (0.047)	0.436 (0.039)	0.434 (0.039)	0.438 (0.039)	0.433 (0.039)	0.437 (0.040)	0.442 (0.039)
2 Parents	ref	ref	ref	ref	ref	ref	ref
Family affluence							
Low	0.339 (0.082)	0.341 (0.065)	0.344 (0.065)	0.335 (0.066)	0.319 (0.065)	0.297 (0.068)	0.345 (0.067)
Medium	0.161 (0.050)	0.120 (0.043)	0.119 (0.042)	0.116 (0.043)	0.112 (0.041)	0.112 (0.041)	0.119 (0.041)
High	ref	ref	ref	ref	ref	ref	ref
Age	0.144 (0.045)	0.115 (0.036)	0.114 (0.035)	0.117 (0.036)	0.112 (0.036)	0.101 (0.037)	0.116 (0.036)
Random intercept variance (SE)	0.031 (0.015)	0.035 (0.011)	0.036 (0.011)	0.037 (0.011)	0.039 (0.012)	0.038 (0.012)	0.044 (0.012)

6 Discussion: Comparison of Multiple Imputation Methods

It is important to consider the value of implementing MI for the current analyses. Following MI, there were improvements in efficiency across all analyses and all covariates (see tables 11-13). Some changes in point estimates were seen between CCA results and the results following MI, most apparently in the analysis examining the relationship between hunger and the outcome adiposity. Following MI, the trend between increasing hunger frequency and increasing odds of adiposity became more clear. Prior to MI, only hunger "sometimes" was significantly associated with adiposity. Although these differences were minor, it does represent a change in substantive results interpretation following MI. Changes were also seen for the effect of immigration status in analyses involving psychosomatic complaints as an outcome. Following MI, being a "recent" immigrant was no longer significantly associated with a lower likelihood of experiencing psychosomatic symptoms (an effect which was present in the CCA). Based on past literature, this may be reasonably explained by cultural factors influencing the likelihood of responding to height/weight or psychosomatic symptom variables (Lee et al., 2002). Overall, the differences between results of CCA and those following MI are small, but they do reduce concerns about potential biases present in the complete data. Therefore, MI was considered beneficial in the present application.

The results of the present investigation also serve to highlight an important aspect of multiple imputation: the complexity of the imputation procedure must be guided by the goals of the analysis. (Andridge, 2011; Carpenter & Kenward, 2012; Kim et al., 2006; Meng, 1994; Reiter et al., 2006; Seaman et al., 2012). What is considered a sufficient level of complexity in the context of complex survey data can be challenging to determine (Drechsler, 2015), and the current application of MI was no exception. The implemented MI methods were selected to represent a progressive range of complexity, and varied primarily based on the extent to

which the clustered nature of the HBSC data was accounted for, how the composite PS score was imputed, and the parametric assumptions involved in the imputation procedure. For simplicity, each of these aspects will be discussed independently in the following sections.

Imputation of clustered data

In the present investigation, it was seen that the random intercept variance will be underestimated if a multi-level imputation is not used, especially when imputation was carried out across the whole country rather than within province (see Table 11, 12 and 13). This is accordance with findings from past literature, and occurs due to the inappropriate variance structure of data imputed using a single-level model (Andridge, 2011; Diaz-Ordaz et al., 2014; Drechsler, 2015; Reiter et al., 2006; Taljaard et al., 2008; Van Buuren, 2011). Imputations for a given cluster generated from this model will be skewed towards the mean in the overall data, thereby reducing between-cluster variability (and consequently random intercept variance). Interestingly, as missingness in the outcome variable increases this effect becomes more apparent (as with the analysis involving adiposity as the outcome in Table 11).

Notably, the random intercept variance estimates following KNN imputation were sometimes higher (see Tables 11 and 13) than those achieved during multi-level imputation. The KNN method included cluster status as a component of the distance function computation, which could possibly be considered a non-parametric analogue to the fixed-effects approach for clusters discussed in Section 4.3.4. When employing the fixed-effects approach with a parametric imputation procedure, imputations for any particular cluster are skewed towards the cluster specific means, thereby inflating between cluster variability and inflating random intercept variance (Andridge, 2011; Drechsler, 2015; Van Buuren, 2011). A similar effect may be occurring here in a non-parametric setting, therefore this approach cannot be recommended to appropriately account for clustering.

Based on these findings, when random intercept variance, or the related measure of ICC, is of particular interest in an analysis, a multi-level imputation must be implemented to ensure congeniality. When interest lies, instead, in the estimation of a fixed parameters, the consequences of disregarding the clustered structure during imputation are less obvious. The repercussions of imputed data with an inappropriate variance structure may propagate to estimates of fixed parameters estimates and their variances (Carpenter & Kenward, 2012; Gelman & Hill, 2006; Reiter et al., 2006).

When imputing clustered data, fixed parameter variance estimates may be impacted by use of a single-level imputation model depending on the degree to which this effect varies within schools relative to between schools (Carpenter & Kenward, 2012; Reiter et al., 2006). A single-level model will result in variances being underestimated for any effects that largely occur between schools, such as the overall intercept. On the other hand, effects that vary mainly within schools will have over-estimated variances. In the results presented in Tables 11-13, estimates of fixed parameter variances when employing the multi-level imputation were lower, in many cases, than the single-level methods, although there were exceptions. Most clearly, variance estimates for the overall intercept were higher following multi-level imputation. This is expected as this parameter will primarily vary between schools and, therefore, the associated variance was underestimated following single-level imputation.

Fixed parameter estimates themselves may be affected when imputing clustered data under a single-level model, particularly when substantive analyses involve generalized linear multi-level models with certain link functions. As discussed in Section 3.3, certain generalized linear multi-level models (e.g. logistic or multinomial) assume that the individual-level residual variance is a fixed constant and, therefore, not estimated (Breslow & Clayton, 1993; Demidenko, 2013; Rabe-Hesketh & Skrondal, 2001, 2006). As the between-cluster residual variance increases the total residual variance increases as well. To account for this, the

dependent variable is rescaled and model parameter estimates increase in absolute magnitude. Therefore, if random intercept variance is underestimated following imputation of clustered data under a single-level model, the parameter estimates will be accordingly biased toward the null (Breslow & Clayton, 1993; Demidenko, 2013). The results from the present investigation are generally consistent with this point. Parameters estimated following multi-level imputation tended to be greater in absolute magnitude than those estimated following single-level imputation (although not to a great extent; see table 7). As an aside, this effect was also seen during the CCA in Tables 3-5, in which certain weighting methods may have biased random intercept variance estimation. Exceptions to this pattern in Tables 11, 12, and 13 may be attributable to another repercussion of disregarding the multi-level structure of the data. As discussed in Section 3.3, use of a multi-level model can be additionally advantageous in the presence of unbalanced cluster sizes (Gelman & Hill, 2006). Multi-level models allow individual cluster estimates to be more uniquely represented in the final point estimates, while in contrast, the estimates from a single-level model may be unnecessarily skewed towards the relationships that exists within the largest cluster samples. This may be particularly relevant in this data set as the size of the clusters a highly variable ranging from 2 to 315 students (Gelman & Hill, 2006). While the direction of this potential skew was not known a priori, a brief supplementary single-level analysis was used to confirm the occurrence of this effect in the present investigation. Single-level logistic regressions were performed for each of the three substantive analyses. The skew that was apparent between the single-level and multi-level analyses was generally mirrored in the differences between estimates following the single-level and multi-level imputation methods (with multi-level analyses) as expected.

To summarize, the results of the present investigation demonstrate consequences of inadequately capturing the variance structure of clustered data set during imputation. Most

obviously, it was seen that random intercept variance is underestimated if the multi-level data of the structure is not accounted for. This is particularly important if estimates related to between cluster variability (random intercept variance, ICCs, or when random slope is included) are a focus of an analysis; in these circumstances a multi-level imputation model should be used to ensure congeniality. On the other hand, when interest is in estimation of a parameter under the assumption that it is constant across the population (as in the present investigation) disregarding the multi-level structure of the data during imputation had a minimal impact. Subtle repercussions are apparent in point estimates due to three identified causes:

(a) The inappropriate variance structure of imputed values may impact the estimated variance of fixed effects, depending on the degree to which a relationship exists within or between clusters;

(b) There are inherent ties between random intercept variance and point estimates during multi-level logistic regression, so imputing based on an appropriate variance structure during these types of analyses may be particularly important; and

(c) The advantage multi-level analysis provides with unbalanced clusters is lost during single-level imputation, and the point estimates are overly skewed towards the relationships which exist in the clusters which contribute the most individuals to the sample.

In the present investigation, the consequences of the points (a-c) above were minor, and did not result in any practically important differences in results. Therefore, in the present investigation, and likely many others, a single-level imputation method is adequate to achieve the goals of the analyses. In some settings, however, the above points may have more extensive repercussions. Point (a) above is always of concern, and may become more consequential

with increases in the strength of clustering or the rate of missingness in the outcome. When analysis will involve a mutli-level logistic regression model (or certain other generalized linear multi-level models) point (b) must acknowledged. Again, as missingness in the outcome or the strength of clustering increases, the incorrect variance structure of imputed data may result in more substantial underestimation of random intercept variance, in which case, the resulting bias during a multi-level logistic regression would be more substantial. Finally, with unbalanced clusters (without sufficient compensation of a large sample size) point (c) is of concern.

There may be residual bias even with the multi-level imputation model if this model does not perfectly capture the true structure of the data (e.g. if there is insufficient auxiliary information to fully describe inclusion probabilities). However, the parametric assumptions for the multi-level model are more tenable than those necessary for unbiased estimation with the simpler imputation models in the settings discussed above.

Imputation of a composite measurement

Imputation of multi-item scales can make maintaining congeniality during imputation more challenging (Eekhout et al., 2014; Gottschall et al., 2012; Shrive, 2006). Imputing each individual component of the composite measurement is more straightforward to implement, but may require additional considerations for the categorical variables which may make up the measure (Bernaards et al., 2006; Demirtas, 2009; Horton et al., 2003; Jia & Enders, 2015; Rodwell et al., 2014; von Hippel, 2013). On the other hand, imputing the composite measure fails to exploit information available from partially-observed items, but makes assumptions of joint normality more reasonable. A rejection sampling procedure can be used to regain this lost information and limit the efficiency loss associated with imputing the composite measure, however, this requires more complex MI specifications (Carpenter & Kenward, 2012). The results from the present investigation reflect these advantages and disadvantages

of the varying approaches used to impute a multi-item measure.

The rejection sampling method implemented was successful at capturing the information present in the observed PS score items for partially-observed individuals. This is evident through the similarity in parameter estimates between the methods imputing the total PS score and the methods imputing each individual item (see Tables 11-13). Furthermore, the loss of efficiency which occured due to imputing the total PS score variable was negligible when rejection sampling was implemented. On the other hand, the approach which abandoned the rejection sampling during the imputation of the total PS score shows a more notable loss of efficiency (see Tables 11-13). Interestingly, the method which did not utilize rejection sampling also generated noticeably different parameter estimates across covariates compared to both the rejection sampling method and the method imputing each of the 8 PS score items. When rejection sampling was not implemented, values for the total PS score were entirely unrestricted and could be imputed outside of the range of realistic values (as can be seen in Figure 3). The present results suggest that this impacted how the relationships between PS score and covariates were maintained during imputation. Overall, these differences are not substantial, but they do highlight the advantages present in the rejection sampling method.

Figure 3 shows imputed versus observed values of the PS total variable for methods which imputed the individual PS score items, the total PS score with rejection sampling, or the total PS score without rejection sampling. These plots show the ability of the rejection sampling method to re-capture the information lost when imputing the total PS measure.

Non-parametric imputation

Non-parametric MI methods offer an advantage when an imputation model may require specification of higher-order relationships and interactions which are unknown or infeasible

to model (D'Orazio, 2011; Liao et al., 2014; Shah et al., 2014). Furthermore, these methods select values for imputation directly from the observed data, so categorical variables and restricted range variables do not pose additional challenges. Aside from the multi-level data structure, these methods could be considered ideal in the present setting; non-parametric methods may suitably accommodate the potential complex relationships between the large number of mixed-typed variables involved.

The results from the KNN method show only minor differences from the parametric imputation methods. This may indicate that possible misspecified or unmodelled relationships in the parametric approaches did not substantially diminish the performance of these methods. On the other hand, the minimal differences between the KNN and the parametric methods may be due to specific aspects of the implemented KNN imputation methodology. Specifically, only the missing values for a recipient individual were replaced with the observed values of the donor. In contrast, some authors have suggested to replace the entire pattern of the recipient (missing and observed variables) with the values of the selected donor (Allison, 2001). The approach implemented in this investigation was selected to avoid throwing out observed data and maximize the number of potential donors (since donors can have missing values for the observed values of target recipients). However, it is possible that it inadequately maintained the inter-variable relationships present in the data thus limiting the main advantage of this non-parametric MI method.

6.1 Limitations and Areas for Future Research

The present investigation implemented only a subset of all possible imputation procedures which may have been applicable, and in only a specific set of circumstances for which they could be employed. The main goal was to compare MI methodology in a realistic data setting, which brings the limitations present in any real data application of MI: the true

values of estimates are not known. Therefore, interpretations regarding bias and efficiency are only speculative, as informed by past research; it cannot be established which method is truly performing best. Regardless, these types of applications provide important information which cannot be gained through simulation studies alone. Real data does not abide by the restrictive assumptions and parameters involved in simulation studies, and true population values are usually not known for incorporation during simulations. A main objective of the current investigation was to evaluate MI methodology within the context of the HBSC so recommendations could be made for future analyses involving HBSC data. This necessitated the trade-off between real data applications and simulation studies.

The current application focused on three analyses which have certain characteristics that may make generalization of results challenging. Firstly, the strength of the clustering occurring in the analyses was relatively low (as evident from the random intercept variance, see Tables 8-10). Differences between methods used to incorporate the clustered data structure may be more apparent in a cases where the magnitude of clustering is stronger (Drechsler, 2015). Further comparison of these methods in analysis in settings which involve a greater magnitude of clustering would be a useful area of additional research. Secondly, the current investigation focused on substantive analyses which only included random intercepts. Comparison of these methods in settings which involve more complex multi-level analyses may reveal important differences in approaches not seen here. Since the present investigation was limited in these regards, the findings here may not be extendable to investigations involving data with high ICCs, or complex multi-level models with random slopes, or cross level interactions.

It is also important to discuss methodological limitations of the presently applied MI procedures. The FCS algorithm was utilized in the current study based on its flexibility to incorporate the rejection sampling methodology. A common criticism of the FCS approach

is the uncertainty around whether the conditionally specified models adhere to some existing joint distribution. Although studies have shown that the FCS approach performs well despite this, the implication during multi-level imputation are not as clear (van Burren et al, 2006; van Burren, 2007; Lee & Carlin, 2010; White et al., 2011). van Buuren (2011) recommends the use of group-specific residual variances during multi-level FCS imputation, however he does not support this recommendation with evidence from application. The software *PAN* (interfaced with mice) used to implement the multi-level imputation methods in this investigation does not allow for the specification of heterogeneous within group residual variances. There may be a case in which a linear multi-level regression of a variable Y (with a homogenous within-group variance) on a variable X, may lead to a conditional distribution of X given Y with a heterogeneous within-group variance (Snijders & Bosker, 2012). Snijders & Bosker (2012) suggest this may be of concern when cluster sizes are very unbalanced. Therefore, this limitation may be relevant in the current investigation due to the unequal size of samples between schools, although the consequences cannot be determined. As software available to implement multi-level MI procedures expands to make this more feasible, future research should examine the impact of this limitation.

The imputation models employed during the present investigation were not entirely congenial to the models used for analysis. Aside from the uncongeniality introduced by the inclusion of auxiliary variables, the exact relationships investigated in the substantive analysis models were not directly incorporated in the imputation models. For example, BMI was imputed as height and weight variables, despite being categorized into age and gender specific categories during analysis. A more congenial approach may have been to impute these categories directly, and accept the loss of the information present in the partially-observed height and weight variables. Similarly, the psychosomatic complaints variable was imputed in multiple ways, but none of these involved imputing it as the dichotomized variable that

was incorporated during analysis. In general, the consequences of these choices are expected to be minimal. The imputation models still incorporated all the variables of interest, albeit in a more detailed form, leading to an arguably "richer" imputation model (see Section 4.1; Meng 1994; Rubin, 1996; Schafer, 2003).

Lastly, MI methodology in the present investigation involved the assumption that the data being imputed are MAR. This assumption was accepted, however it could be argued that some of the imputed variables were MNAR. The self-reported weight variable is especially likely to be MNAR, as those who are of a heavier weight may be less likely to report it. The HBSC data set contains variables which ask about students experience with dieting, body dissatisfaction, and weight perception, and including them as auxiliary variables in the present investigation may have helped improve the assumption that self-reported weight is MAR. Regardless, it is appropriate in such settings to perform sensitivity analyses to examine the impact of the MAR assumption (Carpenter & Kenward, 2012). Although outside of the scope of this investigation, such sensitivity analyses could strengthen the conclusions drawn from this investigation.

6.2 Conclusions and Practical Recommendations

MI was successful at regaining efficiency lost during the CCA and reducing concerns about biases. Employing MI changed some point estimates and reduced all variances compared to CCA in the three HBSC analyses, although the choice of how to implement MI did not have a substantial impact on the conclusions drawn in any of the substantive questions of interest in this setting. However, more generally, the decision about how to implement MI in complex survey settings may have an important impact in two ways:

i. Neglecting to account for the clustered nature of the data in the MI procedure can result in underestimation of random variances and biases in the SEs surrounding point estimates. Moreover, underestimated random variances will result in biased point estimates when the analysis of interest involves certain generalized linear multi-level regressions (e.g. logistic).

ii. In the presence of unbalanced cluster sizes and disproportionate sampling, use of single-level method MI methods may also cause point estimators to be overly reflective of the relationship among those over-represented in the sample (the largest cluster samples).

Each of these effects was observed within the results of the present investigation, although the consequences were small enough to not significantly affect conclusions. In terms of accommodating missingness within a multi-item measurement, it was found that imputing the items themselves and implementing a rejection sampling method were similarly successful in terms of retaining the partial information that was available. Both methods resulted in very similar point estimates with the rejection sampling approach being only slightly less efficient. Ignoring the partially-observed response information and imputing the composite score as if it were any other incompletely observed variable, however, resulted in a more noticeable loss in efficiency and slightly different point estimates.

Based on examination of the literature and the findings in the present application, recommendations for future MI-based analyses of incomplete complex survey data with composite scores such as the HBSC are threefold:

1. If ICCs or random effects are of particular interest in the analysis, if clusters are largely unbalanced, or if analyses will involve generalized linear multi-level models with a non-identity link function, then one should implement a multi-level MI

method to ensure greater congeniality. If fixed parameter estimates are the only target of interest and if clusters are relatively balanced in size, then a simpler, single-level MI approach will likely suffice to reap the main benefits of MI.

2. The partial information available should always be used when imputing total score of a summary measure. This can either be done either by imputing the individual score components themselves before summing or by incorporating the partial scores to define a valid imputation space that can be incorporated with rejection sampling. The latter approach may have a slight edge in terms of making the underlying normal distribution assumption more tenable, while the former approach may be easier to implement in practice and may be more flexible when considering differing weightings in the creation of the score.

3. When it not necessary to employ multi-level imputation, non-parametric methods are advantageous since these methods do not require assumptions about underlying normality in the presence of many categorical variables, and do not require the often challenging task of specifying a parametric imputation model. Furthermore, in settings with large sample sizes, the potential loss of efficiency accompanying non-parametric imputation is not of practical concern. However, since there is no ideal way to account for a clustered data structure during non-parametric MI, parametric multi-level MI should be used in cases when the clustered structure of the data must be accommodated (see recommendation 1).

In summary, MI is an advantageous technique for handling missing data in complex surveys since it can improve efficiency and reduce bias. Congeniality during MI is important to achieve unbiased results, and requires that an imputation procedure accurately maintain all relationships in the data that will have a role in any subsequent analyses. Therefore, the imputation methodology must necessarily be directed by the goals of the subsequent

analysis. In complex survey settings, uncongeniality can introduce biases in ways that are less obvious than in simpler settings. While simpler MI methods will often suffice for complex surveys, these should not be applied haphazardly. Thorough consideration of analysis goals is essential when determining the methodological approach that will achieve appropriate findings and conclusions when employing MI in complex survey settings

References

Allison, P. D. (2001). *Missing data*, volume 136. Sage publications.

Andridge, R. R. (2009). *Statistical methods for missing data in complex sample surveys*. PhD thesis, The University of Michigan.

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal*, 53(1):57–74.

Andridge, R. R. and Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in StatisticsTheory and Methods*, 35(3):439–460.

Asparouhov, T. (2008). Scaling of sampling weights for two level models in mplus 4.2.

Asparouhov, T. and Muthen, B. (2006). Multilevel modeling of complex survey data. *Proceedings of the American Statistical Association, Seattle, WA: American Statistical Association*.

Bernaards, C. A., Belin, T. R., and Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in medicine*, 26(6):1368–1382.

Berntsson, L. T. and Gustafsson, J.-E. (2000). Determinants of psychosomatic complaints in swedish schoolchildren aged seven to twelve years. *Scandinavian Journal of Public Health*, 28(4):283–293.

Booth, M., Hunter, C., Gore, C., Bauman, A., and Owen, N. (2000). The relationship between body mass index and waist circumference: implications for estimates of the population prevalence of overweight. *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity*, 24(8):1058–1061.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3).

Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*, 9(1):49.

Carpenter, J. and Kenward, M. (2012). *Multiple imputation and its application*. John Wiley & Sons.

Carpenter, J. R., Goldstein, H., and Kenward, M. G. (2011). Realcom-impute software for multilevel multiple imputation with mixed response types. *J Stat Softw*, 45(5):1–14.

Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):571–584.

Carpita, M. and Manisera, M. (2011). On the imputation of missing data in surveys with likert-type scales. *Journal of Classification*, 28(1):93–112.

Cheah, B. C. (2009). Clustering standard errors or modeling multilevel data. *University of Columbia*, pages 2–4.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics-Stockholm*, 16(2):113–132.

Cohen, G. and Duffy, J. C. (2002). Are nonrespondents to health surveys less healthy than respondents? *Journal of Official Statistics*, 18(1):13.

Currie, C. E., Elton, R. A., Todd, J., and Platt, S. (1997). Indicators of socioeconomic status for adolescents: the who health behaviour in school-aged children survey. *Health education research*, 12(3):385–397.

Demidenko, E. (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

Demirtas, H., Freels, S. A., and Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation*, 78(1):69–84.

Díaz-Ordaz, K., Kenward, M. G., Cohen, A., Coleman, C. L., and Eldridge, S. (2014). Are missing data adequately handled in cluster randomised trials? a systematic review and guidelines. *Clinical Trials*.

Dong, Y. (2014). *A simulation study on multilevel multiple imputation methods*. PhD thesis, Indiana University.

D'Orazio, M. (2011). Statistical matching and imputation of survey data with the package statmatch for the r environment. *R package vignette http://www.cros-portal.eu/sites/default/files//Statistical_Matching_with_StatMatch. pdf*.

Drechsler, J. (2015). Multiple imputation of multilevel missing datarigor versus simplicity. *Journal of Educational and Behavioral Statistics*, 40(1):69–95.

Eekhout, I., de Vet, H. C., Twisk, J. W., Brand, J. P., de Boer, M. R., and Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of clinical epidemiology*, 67(3):335–342.

Elliott, M. R. and Little, R. J. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16(3):191.

Erickson, S. J., Robinson, T. N., Haydel, K. F., and Killen, J. D. (2000). Are overweight children unhappy?: Body mass index, depressive symptoms, and overweight concerns in elementary school children. *Archives of Pediatrics & Adolescent Medicine*, 154(9):931–935.

Ford, E. S., Moriarty, D. G., Zack, M. M., Mokdad, A. H., and Chapman, D. P. (2001). Self-reported body mass index and health-related quality of life: Findings from the behavioral risk factor surveillance system. *Obesity Research*, 9(1):21–31.

Freeman, J. G., King, M. A., Pickett, W., Craig, W., Elgar, F., Janssen, I., and Klinger, D. (2011). *The health of Canada's young people: a mental health focus*. Public Health Agency of Canada Ottawa.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, pages 153–164.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Goldstein, H. (2011). *Multilevel statistical models*, volume 922. John Wiley & Sons.

Gottschall, A. C., West, S. G., and Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research*, 47(1):1–25.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213.

Hetland, J., Torsheim, T., and Aarø, L. E. (2002). Subjective health complaints in adolescence: dimensional structure and variation across gender and age. *Scandinavian Journal of Public Health*, 30(3):223–230.

Honaker, J., King, G., Blackwell, M., et al. (2011). Amelia ii: A program for missing data. *Journal of Statistical Software*, 45(7):1–47.

Horton, N. J., Lipsitz, S. R., and Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, 57(4):229–232.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., and Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC medical research methodology*, 14(1):28.

Jia, W. W. F. and Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on likert scale variables.

Jönsson, P. and Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using likert data. In *Software Metrics, 2004. Proceedings. 10th International Symposium on*, pages 108–118.

Kalsbeek, W. D., Yang, J., and Agans, R. P. (2002). Predictors of nonresponse in a longitudinal survey of adolescents. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pages 1740–1745.

Kim, J. K., Michael Brick, J., Fuller, W. A., and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):509–521.

Koller-Meinfelder, F. (2009). *Analysis of Incomplete Survey Data–Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching*. PhD thesis, Dissertation, Otto-Friedrich-Universität Bamberg, 2009.(Zitiert auf den Seiten 135, 140, 141 und 142).

Kovačević, M. S. and Rai, S. N. (2003). A pseudo maximum likelihood approach to multilevel modelling of survey data. *Communications in Statistics-Theory and Methods*, 32(1):103–121.

Kropko, J., Goodrich, B., Gelman, A., and Hill, J. (2013). Multiple imputation for continuous and categorical data: Comparing joint and conditional approaches. *Columbia University, Department of Statistics. New York*.

Lee, J. W., Jones, P. S., Mineyama, Y., and Zhang, X. E. (2002). Cultural differences in responses to a likert scale. *Research in nursing & health*, 25(4):295–306.

Lee, K. J. and Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American journal of epidemiology*, 171(5):624–632.

Lee, K. J., Galati, J. C., Simpson, J. A., and Carlin, J. B. (2012). Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study. *Statistics in medicine*, 31(30):4164–4174.

Liao, S. G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F. C., and Tseng, G. C. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC bioinformatics*, 15(1):346.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296.

Little, R. J. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, 88(423):1001–1012.

Little, R. J. (2004). To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99(466).

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, pages 538–558.

Mistler, S. A. (2015). *Multilevel Multiple Imputation: An Examination of Competing Methods*. PhD thesis, Arizona State University.

Morris, T. P., White, I. R., Royston, P., Seaman, S. R., and Wood, A. M. (2014). Multiple imputation for an incomplete covariate that is a ratio. *Statistics in medicine*, 33(1):88–104.

Nackers, L. M. and Appelhans, B. M. (2013). Food insecurity is linked to a food environment promoting obesity in households with children. *Journal of nutrition education and behavior*, 45(6):780–784.

Nezlek, J. B. (2011). *Multilevel modeling for social and personality psychology*. SAGE Publications Ltd.

Niclasen, B., Molcho, M., Arnfjord, S., and Schnohr, C. (2013). Conceptualizing and contextualizing food insecurity among greenlandic children. *International journal of circumpolar health*, 72.

Patrick, K., Norman, G. J., Calfas, K. J., Sallis, J. F., Zabinski, M. F., Rupp, J., and Cella, J. (2004). Diet, physical activity, and sedentary behaviors as risk factors for overweight in adolescence. *Archives of pediatrics & adolescent medicine*, 158(4):385–390.

Paxton, S. J., Wertheim, E. H., Gibbons, K., Szmukler, G. I., Hillier, L., and Petrovich, J. L. (1991). Body image satisfaction, dieting beliefs, and weight loss behaviors in adolescent girls and boys. *Journal of youth and adolescence*, 20(3):361–379.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, pages 317–337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical methods in medical research*, 5(3):239–261.

Pickett, W., Michaelson, V., and Davison, C. (2015). Beyond nutrition: hunger and its impact on the health of young canadians. *International journal of public health*, pages 1–12.

Poikolainen, K., Aalto-Setälä, T., Marttunen, M., Tuulio-Henriksson, A., and Lönnqvist, J. (2000). Predictors of somatic symptoms: a five year follow up of adolescents. *Archives of disease in childhood*, 83(5):388–392.

Rabe-Hesketh, S. and Skrondal, A. (2001). Parameterization of multivariate random effects models for categorical data. *Biometrics*, 57(4):1256–1263.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):805–827.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96.

Raghunathan, T. E., Solenberger, P. W., and Van Hoewyk, J. (2002). Iveware: Imputation and variance estimation software. *Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan*.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471.

Reiter, J. P., Raghunathan, T. E., and Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32(2):143.

Rodwell, L., Lee, K. J., Romaniuk, H., and Carlin, J. B. (2014). Comparison of methods for imputing limited-range variables: a simulation study. *BMC medical research methodology*, 14(1):57.

Royston, P. (2011). Ice: Stata module for multiple imputation of missing values. *Statistical Software Components*.

Royston, P. and White, I. R. (2011). Multiple imputation by chained equations (mice): implementation in stata. *Journal of Statistical Software*, 45(4):1–20.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489.

Rubin, D. B. and Schafer, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. In *Proceedings of the Statistical Computing Section of the American Statistical Association*, volume 83, page 88.

Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394):366–374.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57(1):19–35.

Schafer, J. L. (2009). Multiple imputation for multivariate panel or clustered data. *R package version 0.2-6*.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4):545–571.

Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of computational and Graphical Statistics*, 11(2):437–457.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295.

Seaman, S. R., White, I. R., Copas, A. J., and Li, L. (2012). Combining multiple imputation and inverse-probability weighting. *Biometrics*, 68(1):129–137.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. *American Journal of Epidemiology*, 179(6):764–774.

Shrive, F. M., Stuart, H., Quan, H., and Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC medical research methodology*, 6(1):57.

Siddique, J. and Belin, T. R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in medicine*, 27(1):83–102.

Skinner, C. J. (2003). Introduction to part d. *Analysis of survey data*, pages 197–204.

Snijders, T. A. (2011). *Multilevel analysis*. Springer.

Snijders, T. A. B. and Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage.

Stapleton, L. (2012). Evaluation of conditional weight approximations for two-level models. *Communications in Statistics-Simulation and Computation*, 41(2):182–204.

Stekhoven, D. J. and Bühlmann, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate behavioral research*, 44(6):711–740.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338:b2393.

Su, Y.-S., Yajima, M., Gelman, A. E., and Hill, J. (2011). Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software*, 45(2):1–31.

Swallen, K. C., Reither, E. N., Haas, S. A., and Meier, A. M. (2005). Overweight, obesity, and health-related quality of life among adolescents: the national longitudinal study of adolescent health. *Pediatrics*, 115(2):340–347.

Taljaard, M., Donner, A., and Klar, N. (2008). Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometrical journal*, 50(3):329–345.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242.

van Buuren, S. (2012). *Flexible imputation of missing data*. CRC press.

van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76(12):1049–1064.

van Buuren, S. et al. (2011). Multiple imputation of multilevel data. *Handbook of advanced multilevel analysis*, pages 173–196.

van Ginkel, J. R., Sijtsma, K., van der Ark, L. A., and Vermunt, J. K. (2015). Incidence of missing item scores in personality measurement, and simple item-score imputation. *Methodology*.

von Hippel, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? *Sociological Methods & Research*, 42(1):105–138.

White, I. R., Daniel, R., and Royston, P. (2010). Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics & Data Analysis*, 54(10):2267–2275.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, 30(4):377–399.

Yu, L.-M., Burton, A., and Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research*, 16(3):243–258.

Yucel, R., Raghunathan, T., and Schenker, N. (2006). Shrimp: sequential hierarchical regression imputations. In *International Conference on Health Policy Research, Boston, USA*.

Yucel, R. M. and Demirtas, H. (2010). Impact of non-normal random effects on inference by multiple imputation: a simulation assessment. *Computational statistics & data analysis*, 54(3):790–801.

Yucel, R. M., He, Y., and Zaslavsky, A. M. (2011). Gaussian-based routines to impute categorical variables in health surveys. *Statistics in medicine*, 30(29):3447–3460.

Zhao, E. and Yucel, R. M. (2009). Performance of sequential imputation method in multilevel applications. In *American Statistical Association Proceedings of the Survey Research Methods Section*, pages 2800–2810.

Zhou, H. (2014). *Accounting for Complex Sample Designs in Multiple Imputation Using the Finite Population Bayesian Bootstrap*. PhD thesis, University of Maryland.

Zhu, M. (2014). Analyzing multilevel models with the glimmix procedure. In *Proceedings of the SAS Global Forum 2014 Conference. http://support. sas. com/resources/papers/proceedings14/SAS026-2014. pdf.*