Why Missing Data Matter in the Longitudinal Study of Adolescent Development: Using the 4-H Study to Understand the Uses of Different Missing Data Methods[1]

Helena Jeličić

National Children's Bureau, UK

Erin Phelps and Richard M. Lerner

Tufts University

[1]This article in based in part on a dissertation submitted by the first author in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Child Development at Tufts University. This research was supported in part by a grant to Richard M. Lerner by the National 4-H Council. The authors would like to thank Aline Sayer and Avron Spiro for their valuable comments. Requests for reprints or for additional information may be obtained from Helena Jeličić, National Children's Bureau, 8 Wakley Street, London, EC1V 7QE, UK or by emailing to helena.jelicic@alumni.tufts.edu .

Abstract

The study of adolescent development rests on methodological appropriate collection and interpretation of longitudinal data. While all longitudinal studies of adolescent development involve missing data, the methods to treat missingness that have been recommended most often focus on missing data from cross-sectional studies. The problems of missing data in longitudinal studies are not described well, there are not many statistical software programs developed for researchers to use, and there are no longitudinal empirical examples involving adolescent development that show the extent to which different missing data procedures can yield different results. Data from the first three waves of the 4-H Study of Positive Youth Development were used to provide such an illustration. The sample contains 2,265 participants (56.7% females) who were in Grade 5 at Wave 1, in Grade 6 at Wave 2, and in Grade 7 at Wave 3, and varied in race, ethnicity, socioeconomic status, family structure, rural-urban location, and geographic region. The results showed that three missing data techniques, i.e., listwise deletion, direct maximum likelihood (DirML), and multiple imputation (MI), did not yield comparable results for research questions assessing different aspects of development (i.e., change over time or prediction effects). The results indicated also that listwise deletion should not be used. Instead, both DirML and MI methods should be used to determine if and how results change when these procedures are employed.

Key words: Missing data; longitudinal research; multiple imputation; maximum likelihood; adolescent development

All of developmental science – and certainly the study of adolescent development – rests on the effective empirical investigation of change. Therefore, advancing theory and application in the study of adolescent development must involve the methodologically sound conduct of longitudinal research and on the drawing of inferences about the nature of adolescent development from longitudinal research. However, among the many methodological challenges involved in longitudinal research (e.g., see von Eye, 1990a, b for a discussion), the problems of missing data, as well as questions about what methods to use to address these problems, are fundamental and among the most frequently discussed in the methodological literature (e.g., Shafer & Graham, 2009). Indeed, missing data can be found in all major longitudinal studies of adolescent development . Sometimes participants do not complete a questionnaire due to fatigue or loss of motivation. They may skip a whole wave of testing, or they may drop out and never return. The question for every developmental researcher, then, is what to do about missing data. Since data analysis procedures were designed for complete data sets, the presence of missing data may create extra steps for researchers, e.g., before they conduct analyses they must find a missing data estimation technique, because just deleting cases with incomplete data biases the data set and compromises the generalizability (Jelicic et al, 2009).

Current missing data procedures can be divided into three groups: procedures based on deletion of cases, imputation-based procedures, and model-based procedures (Little & Rubin, 2002). Procedures based on deletion of cases simply discard participants with incomplete data and analyze only complete data. In imputation-based procedures, missing values are "filled in" with particular values, which are determined by specific procedures. Such steps result in a complete (imputed) data set for data analysis with standard methods. The goal of missing data imputation is to preserve important characteristics, such as mean and variance, of the whole data

set, so that analyses would be most likely to produce efficient and unbiased results. Finally, model-based procedures are those in which a missing data model and a data analysis model are estimated simultaneously and, therefore, direct handling of missing data is avoided.

The literature illustrates that, although various missing data techniques have been used in the past, some options are no longer recommended because of shortcomings in statistical estimation (see, for example, Allison, 2002; Graham, Cumsille, & Elek-Fisk, 2003; Graham & Donaldson, 1993; Jones, 1996; Little & Rubin, 2002; Peugh & Enders, 2004; Schafer & Graham, 2002; Wothke, 2000). These procedures include estimating missing values based on mean substitution or single imputation with regression, or on deleting cases with incomplete data. In regard to such deletion, it seems clear that any researcher should have the goal of making statistically valid inferences about population parameters, even from an incomplete data set. In other words, researchers should do everything possible to make sure that their results are valid and representative of the population that is studied. The amount of missing data may be significant, and therefore, dropping cases can substantially lower the number of cases in the data set. The resulting new, smaller sample size will have significantly less power. In addition, a sample with dropped cases may be a non-random sample from a population (or from the original group that was assessed), and therefore analysis will produce results that are not representative of that population or group. Hence, dropping cases from the data set has been shown to be a poor approach for handling missing data.

Statisticians currently recommend two approaches for approaching missing data analysis, the direct maximum likelihood method (DirML), and the multiple imputation method (MI), both of which are based on statistical theory and have been shown to provide results that are representative of the population of interest (see, for example, Allison, 2002; Arbuckle, 1996;

Enders & Bandalos, 2001; Graham, Cumsille, & Elek-Fisk, 2003; Graham, Hofer, & MacKinnon, 1996; Little & Rubin, 2002; Peugh & Enders, 2004; Schafer & Graham, 2002; Wothke, 2000). With advances in technology and computer software, most of these methods are easily available to researchers. For example, all structural equation modeling (SEM) software programs include the DirML method for missing data. Therefore, the DirML method for missing data is very useful in multivariate longitudinal research studies that use SEM models for testing research questions.

Nevertheless, a review of the literature shows that researchers are mostly using methods that are not recommended by statisticians; e.g., that is, many researchers used deletion of cases or imputation of means. For instance, Peugh and Enders (2004) examined practices of handling missing data in a sample of 23 psychological, educational, and applied journals in 1999 and 2003. They found that, among those studies that were identified as having missing data (30% in 1999 and 77% in 2003), most researchers used deletion of cases as a missing data method. Only a small number of studies (3%) used techniques recommended in the literature over the last 15 years. In addition, Jelicic, et al. (2009) examined longitudinal studies reported in the three most prominent developmental science journals, *Child Development*, *Developmental Psychology*, and *Journal of Research on Adolescence,* that were published from 2000 through 2006; the goal was to see if authors reported having missing data and, if so, how they accounted for it. Twenty-one issues were examined and 100 longitudinal studies were found that reported missing data and missing data strategies. The majority of longitudinal studies (82%) used either listwise deletion or pairwise deletion as a missing data technique. Seven studies (12%) used the DirML method, and one study reported using the MI method. These findings indicated that most longitudinal researchers publishing in top-tier developmental science journals in 2006 were still using a case

deletion method for handling missing data. Few researchers were using the currently recommended techniques. Given this status of the use of missing data methods in contemporary developmental science, the issue of implications for "best practice" in future longitudinal research is raised.

Longitudinal studies: Additional problems with missing data

Longitudinal studies are investigations in which the same group of individuals are observed (i.e., tested) across different time points (i.e., waves; Baltes, Reese, & Nesselrode, 1988). Some longitudinal studies are designed to measure different variables at different waves, whereas others, repeated measures designs, are created to measure the same set of variables with the same group of individuals (i.e., the same sample) across different waves (i.e., at different time points). In both cases, data can be missing for participants within any specific wave of the study as well as across two or more waves (i.e., there may be missing participants). The biggest challenge for developmental researchers in conducting a longitudinal study is maintaining the same sample across different times of testing. Participants may drop out of the study and never return again (i.e., there is attrition). In other cases, participants may be absent from one or more waves of the study but appear at a later wave. In the literature, this type of missing data is usually called wave nonresponse (Graham, Cumsille, & Elek-Fisk, 2003).

In addition to wave nonresponse, item nonresponse may occur for variables within each particular wave. This type of missingness creates an additional concern that needs to be taken into account when addressing the problem of missing data in a longitudinal study. Consequently, one may wonder how both wave and item nonresponse can be taken into account. The literature does not provide an easy solution that is generally applicable. However, one implicit conclusion can be drawn: one must first have an explicit hypothesis or a question to test, which should guide

the selection of variables (either across waves or within one wave only) that will be used for testing the particular question or hypothesis. After selecting the correct variables, exploratory data analysis can be done in order to determine the amount of missing data for each variable in the data set and whether there is a pattern of missing data.

In addition to involving multiple waves of data, each wave of a longitudinal study may be comprised of a large set of variables that are measured across time. The data set may also contain data from different sources (e.g., data about adolescents' habits may be collected from both adolescents themselves and from their parents). In the case where a researcher collects data from different sources at each wave of a study, it is possible to have different wave nonresponse from different sources. For example, in a study examining change over time in adolescents' eating habits, researchers may collect data from both adolescents and their parents. However, there may be parents who are unavailable or unwilling to participate for one wave or even in the whole study, while their adolescents do participate in one, two, or all waves of the study. Trying to define patterns of missing data within the data set, as well as to make assumptions about missing data mechanisms for particular item or wave nonresponse, may be overwhelming in a multivariate longitudinal study. However, we have noted that the current literature does not provide guidelines for dealing with such complex types of missingness in such complex longitudinal studies. We believe, however, that decisions about handling missing data in multivariate longitudinal studies should be made based on an exploratory data analysis of the particular data set, while considering a theory of the phenomena that one wishes to test with that data set and, as well, exploring the reasons for missing data.

As we have noted, there can be many reasons for why data are missing in any given longitudinal study. Using missing data terminology (Little & Rubin, 2002; Rubin, 1976), it is

important to differentiate between three missing data mechanisms: data missing completely at random (MCAR), data missing at random (MAR), and data not missing at random (NMAR). Missing data are considered MCAR when the reason for missingness is not related to underlying values of missing data (i.e., the values that would have been obtained if participants answered the question) and is not related to any measured variable. Missing data that are considered MAR are also unrelated to the underlying values of missing data, but can be related to the outcomes as long as there are other observed variables that capture or control for the cause of missingness. Essentially, when data are thought to be MAR, the missingness is caused by other factors that are measured in the study. For example, say we want to study the timing of puberty for boys and girls. Perhaps fewer boys answer the question, so that the pattern of missing data varies systematically for boys and girls. If, however, within boys and girls, the probability of missing pubertal data is not related to actual pubertal timing, then the data are defined to be MAR. The third category of missing mechanisms, data that are NMAR, is identified when the missingness is related to the underlying missing data values (e.g., participants who are unsatisfied with their school are more likely to not answer questions about school satisfaction).

Defining the missing data mechanism within a data set is important for choosing a method for handling missing data. It is possible to identify whether data are MCAR. By using those observed variables with values that are available for everyone, this determination can be made by comparing whether there are systematic differences for participants who are and who are not missing responses in a particular wave or on a set of variables. However, in real-world social science applications, data that are MCAR are the least likely, and it is not possible to test whether data are NMAR, simply because one does not know the true values of the missing data for a particular variable. Thus, in most cases, a particular missing data mechanism can only be

stated as an assumption, and whether data are MCAR or MAR can be established by exploring the correlations among missing data and other variables in the data set. After the missing data mechanisms, patterns of missing data, and the amount of missing data are identified within a data set, a researcher must decide on the best missing data method to employ.

Missing data methods

The literature has demonstrated unequivocally the serious violations of statistical assumptions that occur when missing data are ignored. Various techniques used in the past, such as deleting cases with missing data, or estimating values with mean substitution or other single value imputation, have been amply shown to be inadequate at best and misleading at worst (Allison, 2002; Graham, Cumsille, & Elek-Fisk, 2003; Peugh & Enders, 2004; Schafer & Graham, 2002; Wothke, 2000). Listwise deletion assumes data are MCAR, and it is a valid technique in this case. However, when considering large multivariate and longitudinal data sets that are more likely to have a large amount of missing data, which are not likely to be MCAR, the listwise deletion method is problematic because it can dramatically reduce the sample size. Large reductions in sample size create two potential problems: Statistical power is significantly reduced, and the reduced sample may no longer be representative of a population. In addition, traditional imputation-based procedures, such as mean substitution, are inadequate since they do not preserve the important characteristics of the whole data set, such as key relationships among the variables and means. The most significant problem with single imputation methods is that variances and standard errors are underestimated, thus producing inefficient and biased results. Newer procedures, such as multiple imputation (MI) and direct maximum likelihood (DirML), are based on theoretical frameworks for missing data estimation as well as on statistical theory, and do preserve the important characteristics of the entire data set.

The DirML method can be used in data analysis, such as in structural equation modeling (SEM) or mixed methods procedures (e.g., multilevel modeling), on incomplete data sets. In DirML methods, inclusion of data from partially completed cases contributes to the estimation of parameters that involve the missing portion of the data as well. The computation process for obtaining these estimates operates "as if" missing data are "replaced" with probable values implied by the observed values of other variables through the use of the linear relationships between the variable with missing data and other variables included in the model (for detailed description see, for example, Allison, 2002; Enders & Bandalos, 2001; Harel & Zhou, 2007; Little & Rubin, 2002; Peugh & Ender, 2004; Schafer & Graham, 2002). However, these implied values for missing data are only used during the DirML method for obtaining final estimates and are not imputed within a data set.

In contrast to DirML, the multiple imputation (MI) method produces more than one imputed data set, each of which contains imputed values that are randomly drawn from a distribution of plausible missing values (for a detailed description of the MI algorithm see Schafer, 1997). Each imputed data set contains slightly different imputed values, since for each imputation random error is added to the predicted value drawn from the regression. In this way the MI method overcomes the problems of the single regression-based imputation method in which imputed values lack error variance (i.e., all imputed values lie on a regression line), and the regression equation that is used for imputation is just one estimate of the regression equation based on the available data (Allison, 2002; Graham, Cumsille, & Elek-Fisk, 2003; Graham & Hofer, 2000). The data analysis procedure is then conducted on multiple created data sets and the results from different data sets are combined using a set of formulae based on Rubin's rules (e.g., see Schafer & Graham, 2002; Schafer & Olsen, 1998). This produces a combined result, which

provides a single parameter estimate (i.e., calculated as a simple average of multiple parameter estimates from imputed data sets) and its standard error (i.e., calculated as a function of two parts: an average of multiple standard errors from imputed data sets and an added term that captures the variability in the parameter estimates across the imputations). Therefore, using the MI method, variability is introduced into the data set and the results from the analysis are more generalizable to the population than results from a single imputation method. Recently, Graham, Olchowski, and Gilreath (2007) increased the recommended number of imputations necessary for estimating small effect sizes, depending on sample sizes, thus increasing the data analyst's workload. However, given the empirical evidence of the performance of the DirML and MI methods for missing data in terms of providing more valid results, these two methods represent the best methods currently available if assumptions of at least MAR data and the multivariate normal distribution are met. The MI method can even yield statistically valid estimates when the assumption of the multivariate normal distribution is not met, which makes the MI method a preferred method over the DirML method.

The third mechanism that can generate missing data is called Not Missing At Random, or NMAR. This produces a more difficult situation for which MI and DirML methods are not appropriate solutions. For detailed discussions on modeling missing data that are NMAR, see Daniels & Hogan, 2000; Foster et al., 2004: Molenberghs, Kenward, & Lesaffre, 1997.

The Current Study

Because many contemporary longitudinal researchers are not following the recommendations of statisticians regarding the use of appropriate missing data techniques (Jelicic et al., 2009), the field of developmental science can be adversely affected. Not only will theoretical understanding be hampered because of inappropriate inferences drawn from

methodologically-biased data sets but, in addition, child-, youth-, and family-centered programs and policies that are based on implications from such studies will be ill-informed. Study results attained through the use of missing data techniques that are not recommended may not be appropriate for informing programs and policies affecting individuals and groups. However, to gauge the severity of bias in the interpretation and extrapolation to applications, it is important to know the extent to which the use of non-recommended approaches to treating missing data actually yields results different from those particularly associated with recommended approaches within actual longitudinal data sets that seek to advance theory and impact program and policy applications.

The present research provides a sample case of the use of an actual longitudinal data set, i.e., the 4-H Study of Positive Youth development (PYD; e.g., Lerner et al., 2005; Phelps et al., 2007, 2009), an investigation aimed at elucidating a developmental systems theoretical model of adolescent development and, as well, at informing youth-services, professionals, parents, and teachers about the ways resources in and out-of-school time programs, families, and schools can promote PYD. Since recommendations for using the DirML and the MI methods over traditional methods come mostly from simulation studies, our intent was to examine the extent to which different missing data methods yield different results using this actual multivariate longitudinal data set. In particular, the focus was on differences in using listwise deletion, the DirML method, and the MI method, since listwise deletion is the most frequently used technique in published research (e.g., Jelicic et al., 2009), and the DirML and MI methods are currently recommended methods for use.

The three missing data techniques were used in order to provide answers to the following two questions. First, how different are the results obtained from an actual longitudinal study

when different missing data approaches are used? Second, to what extent do different results yield different interpretations of the phenomena that are studied? For the purpose of providing illustrations of the sorts of empirical work derived from large-scale, multivariate longitudinal studies, our focus was on two types of analyses: linear growth modeling (e.g., change in social competence across early adolescence); and predicting a specific outcome from a set of variables collected at previous points in time (e.g., to what extent can depressive symptomatology in Grade 7 be predicted with parents' behaviors and youths' self-worth assessed in Grade 5?). These are two types of analyses that contemporary longitudinal researchers are using in their studies. For each type, two different empirical questions are used that address common research themes in the field of social and personality development in early adolescence. Puberty, as well as changing school settings from elementary to middle schools, represents a context in which many physical, cognitive, and social changes usually take place, and such changes have an impact on adolescent self-esteem, self-competence, and depressive symptomatology. Therefore, the development of self-esteem and self-perceived competence, as well as changes in depressive symptomatology, during early adolescence, are frequently studied developmental phenomena (cf. Harter, 1983; Lerner & Steinberg, 2009), and they were selected as the four substantive domains of analysis in the present research. Although these analyses were conducted using longitudinal data, there are important differences in the statistical structure of the two types of questions. When examining prediction effects, the variables included in the analysis are assumed to be independent. However, in the repeated measures, the units of analysis, same variables measured across time, are not independent, and this characteristic is included in the analytic method. Consequently, these two analyses differentiated between repeated measures effects (i.e., testing change over time in a particular variable) and prediction effects (i.e., predicting a specific

outcome with prediction variables from previous points in time) when using different missing data techniques.

Method

Data for this research came from the first three waves of the 4-H Study of Positive Youth Development, a national study of adolescents focused on defining key aspects of positive youth development (PYD), as well as on identifying developmental resources (assets) in homes, schools, and communities thought to promote PYD in adolescents (Lerner et al., 2005).

Participants

At Wave 1, participants came from sites located in 13 states that provided regional, rural-urban, racial/ethnic, and religious diversity. Schools were chosen as the main setting for collecting the sample. Assessment was conducted in 57 schools and in four after school programs. Participants were 1,719 fifth grade adolescents and 1,139 of their parents.

At Wave 2, 976 youth who were in the fifth grade during Wave 1 were retested. In addition, in order to control for the influence of prior testing on the findings, an additional sample of 978 youth and their parents was tested. The additional youth came from both current and new schools and after-school programs in 18 states. At Wave 3, an additional 780 youth and their parents were tested as test-retest controls and in order to increase the overall sample size of the study. Additional youth in Wave 3 came from the schools tested in Wave 1 and Wave 2, but also from 24 new 4-H after school programs. Across all three waves, 3,477 youth and 2,167 of their parents were tested. The sample varied in race, ethnicity, socioeconomic status, family structure, rural-urban location, geographic region, and program participation experiences.

Attrition in the 4-H Study sample is not randomly distributed across schools. In Wave 2 and Wave 3, some principals withdrew consent for their schools to participate, and thus these

students “dropped out” without having the opportunity to remain in the study. Therefore, the attrition in the 4-H Study sample is due to attrition of schools, where an entire school “dropped out” from the study, and attrition of students who *were* allowed to remain in the study. Attrition from Wave 1 to Wave 2, as well as Wave 2 to Wave 3 is about 40%. However, 30% of participants from Wave 1 and 18% of participants from Wave 2 dropped out due to attrition of schools. Therefore, in the present sample, school withdrawals from the study accounted for the largest amount of attrition. In order to exclude the possible effects of school attrition, the sample used in this study did not include those participants who dropped out due to school attrition (N = 674). In addition, participants who were not tested prior to Wave 3, because they joined the study when new 4-H after-school programs were included in Wave 3, were also excluded from the sample (N = 478).

Accordingly, the sample used in this study contains 2,265 participants (56.7% females) who were in Grade 5 at Wave 1 (mean age = 11.0 years, SD = 0.47), in Grade 6 at Wave 2 (mean age = 12.1 years, SD = 0.48), and in Grade 7 at Wave 3 (mean age = 13.1 years, SD = 0.46). Other participants were excluded from this study because they were not in the appropriate grade level at either time of testing (N = 53) or because their sex information was inconsistent across times of testing, thus making it not possible to determine if they were males or females (N = 7).

Of the 2,265 youth participants, 636 participated at all three times of testing, 959 participated in any two times of testing, and 670 participated in only one time of testing. In addition to youth participants, 1,720 (75.9%) of their parents or guardians completed a guardian questionnaire at, at least, one wave of testing.

The youth sample was racially/ethnically diverse, with a majority reporting being European American (60.0%), Latino/a (16.2%), or Asian American (6.1%), and the remaining

reporting an ethnicity of African American (4.0%), Native American (2.6%), multiethnic/multiracial (4.5%), or other (0.6%). In addition, 5% of the youth reported their race/ethnicity inconsistently across three waves and 1% of youth had completely missing information on their race/ethnicity. Although participants in this sample are clustered within schools, for the purposes of this study this information will be ignored.

Procedure

For all three waves of data collection, teachers or program staff gave each youth an envelope to take home to their parent or guardian. The envelope contained a letter explaining the study, a consent form, a parent questionnaire, and a self-addressed envelope for returning the parent questionnaire and consent form. For those youth who received parental consent, data collection was conducted either in the school or youth program (e.g., 4-H), by trained study staff or hired assistants for distant locations. The procedure began with reading the instructions for the student questionnaire (SQ) to the youth. Participants were instructed that they could skip any questions they did not wish to answer. Data collection took approximately two hours, which included one or two short breaks. During Waves 2 and 3, students who were unable to be surveyed at their school or 4-H site, in that they were either absent during the day of testing or the school superintendent did not allow testing to occur in the school, received a survey in the mail.

Measures

Measures used in this research came from the Student Questionnaire (SQ) and the Parent Questionnaire (PQ) of the 4-H Study of Positive Youth Development. Selected variables from the SQ and information on the socioeconomic status from the PQ were used in the analyses.

Socioeconomic status (SES)

Socioeconomic status (SES) was indexed by mother's education and family per capita income. The two SES variables were measured at all three waves with the Parent Questionnaire. Mother's education and family per capita income variables were found to be moderately correlated at each time of testing (correlations range from 0.42 to 0.51). For the purpose of this study, mother's education and family income from Wave 2 were used as indicators of the two SES measures. Wave 1 and Wave 3 data were used for imputing missing information in Wave 2. This step was taken because parents did not fill out the Parent Questionnaire at each wave of study. Wave 2 was chosen as a starting point for obtaining information about two SES measures, since we can expect that differences in mother's education or per capita income will be smaller across one year than across two years. Change in mother's education and per capita income was tested for those participants whose parents returned their completed PQs for any two waves. The analyses showed that mother's education variable did not significantly change over time, whereas there was a small but significant increase ($p<0.001$) in family per capita income between Wave 2 and Wave 3, and therefore, between Wave 1 and Wave 3 as well.

Maternal warmth

The maternal warmth subscale is a portion of the Child's Report of Parenting Behaviors Inventory (CRPBI) (Schludermann & Schludermann, 1970), a widely used self-report measure of children's assessment of parenting practices. Maternal warmth was conceptualized as behaviors that indicate acceptance, nurturance, support, and a feeling of being loved and wanted by the parent (Gray & Steinberg, 1999). Examples of maternal warmth items include "My mother speaks to me in a warm and friendly way" and "My mother cheers me up when I am upset." The response format ranges from 1 (almost never) to 5 (almost always). Higher scores

indicate higher warmth and nurturance. In the present data set, Cronbach's alpha coefficients for this measure ranged from 0.94 to 0.96 across three points in time for maternal warmth.

Parental monitoring

The eight item Parental Monitoring Scale (PMS) (Small, & Kerns, 1993) was used to assess the extent to which youth report that their parents know the whereabouts of their youngster after school and at night, and have knowledge of a youth's friends and their parents. Examples of parental monitoring items include "My parents know where I am after school" and "My parents know how I spend my money." The response format ranges from 1 (never) to 5 (always), with higher scores indicating higher parental monitoring. In the present data set, Cronbach's alpha coefficients for the PMS ranged from 0.88 to 0.96.

School grades

At all three waves of the study, youth were asked about the grades they earn in school. The 8-point response format for this item ranged from 4.0 (Mostly As) to 0.5 (Mostly below Ds). The response on this item was used as a self-reported GPA.

Depressive symptomatology

At all three waves of the study, depressive symptomatology was measured by the 20-item Center for Epidemiological Studies Depression Scale (CES-D; Radloff, 1977). Using a forced choice response format ranging from 0 = *rarely or none of the time (less than 1 day)* to 3 = *most or all of the time (5-7 days)*, participants reported how often they felt a particular way during the past week. An example of an item is "I was bothered by things that usually don't bother me." Higher scores indicate higher levels of depressive symptoms. In the present data set, Cronbach's alpha coefficients for the depressive symptoms scale ranged from 0.81 to 0.87. In addition, the variable was positively skewed at each time of testing, and therefore, it was transformed into the

log-depression for all analyses in order to meet the assumption of multivariate normality in using the missing data methods. We will refer to it as log-depression in the following text for convenience, even though it refers to depressive symptomatology.

Self-perceived Competence and Global Self-worth

The Self-Perception Profile for Children (SPPC) (Harter, 1983) was developed to assess perceived competence in regard to five specific domains of functioning and global self-worth. These subscales are: 1. academic competence (reflecting school performance); 2. social competence (emphasizing peer popularity); 3. physical competence (stressing ability at sports and outdoor games); 4. physical appearance (assessing satisfaction with one's appearance); 5. conduct/behavior adequacy (emphasizing behaving in accordance with rules for conduct); and 6. self-worth (indexing feelings of self-esteem, in general). Harter (1982) developed a structured alternative response format to assess perceived competence or adequacy of functioning. Participants are initially asked to choose between two types of people, for example, "some kids are happy with the way that they look OR other kids are not happy with the way that they look." After a respondent chooses the person he or she is most like, the participant must decide if it is "really true for me" or "sort of true for me." Half of the items begin with a positive sentence, reflecting high competence, and the other half with a negative sentence, reflecting low competence. The items belonging to any one domain are distributed across the scale, and within each scale the items are counterbalanced. Each item is scored from 1 to 4, where a score of 1 indicates low perceived competence and a score of 4 reflects high perceived competence. Overall scores are computed for each of the six scales. For the purpose of this study, only the self-worth, social competence and academic competence scales were used. In the present data set,

Cronbach's alpha coefficients for these three subscales ranged from 0.70 to 0.80 for academic competence, 0.62 to 0.75 for social competence, and 0.69 to 0.79 for global self-worth.

Peer Support

The four items of the Peer Support Scale (PSS) (Armsden & Greenberger, 1987) assess adolescents' relationships with friends. The response format ranges from 1 (always true) to 5 (almost never true). Examples of items include: "I trust my friends" and "My friends care about me." When all items are reverse coded, higher score indicates higher peer support. In the present data set, Cronbach's alpha coefficients for the PSS ranged from 0.89 to 0.93.

Item-level missing data and scale scores

For all multiple item scales used in analyses, scale scores were computed for all participants where 60% or more of the scale items were answered, since all the scales used in this study are well known scales with good reliability and unidimensional factor structure. The literature also suggests that calculating a scale based on 60% of available items will yield more efficient and unbiased results than deletion of cases based on missing data on any scale item(s) if the multiple item scale has relatively high reliability (.50) and represents a unidimensional construct (Roth, Switzer, & Switzer, 1999). Therefore, information from most items was used as valid answer to calculate scale scores, rather than deleting the existing information.

Data analysis procedures

The data analyses followed multiple steps. As noted before, results from the three missing data techniques will be presented: listwise deletion, direct maximum likelihood (DirML), and multiple imputation (MI). These missing data techniques were used in two separate types of analyses: Examining change over time in a specific variable using linear growth curve modeling, and predicting a specific outcome variable from a set of prediction variable

using regression analysis. For each set of analyses, two substantively different empirical questions (using different measurement variables) were conducted in order to evaluate possible variable effects on the results of applying various missing data techniques. Thus, four questions were analyzed with the three missing data techniques.

Empirical questions

In order to examine the effects of change in specific measures, linear growth curve modeling was used to assess two different empirical questions: How does social competence change across Grades 5, 6, and 7 and how do depressive symptoms change across Grades 5, 6, and 7? In addition, the effects of sex on change were tested in both of these analyses. In order to examine how a specific measure can be predicted with sets of measures assessed at previous time points, regression analyses were conducted to assess two additional empirical questions: How well can academic competence at Grade 7 be predicted with parental monitoring and GPA level measured at Grade 5, as well as measures of SES reported by parents and how well is Grade 7 predicted with maternal warmth and global self-worth measured at Grade 5 as well as measures of SES reported by parents? In addition, effects of sex and interactions of sex with the other predictors were also tested in both of these analyses.

The four empirical questions address common research themes in the field of social and personality development in early adolescence. Puberty, as well as changing school settings from elementary to middle schools, represent a context in which many physical, cognitive, and social changes usually take place, and such changes have an impact on adolescent self-esteem, self-competence, as well as depressive symptomatology. Therefore, development of self-esteem and self-perceived competence, with the change in depressive symptomatology during early

adolescence, as frequently studied phenomena in developmental psychology, were selected for the four substantive analyses.

Frequencies of missing data within each wave and attrition

Frequencies of missing data within each wave and across three waves, differentiating among variable and wave non response, are presented before the results of empirical analyses. Attrition analyses, using t-tests and chi-square tests, were conducted with several background and outcome variables, such as, sex, race, social and academic competence, and depressive symptomatology. These analyses provide information on whether attrition can be accounted for by variables measured in the study. In addition to attrition analyses, logistic regression was used to predict both attrition and non-response within each wave on outcome variables using all other variables as predictor variables. Including all available variables simultaneously in the model to predict missingness may provide insights into the reasons why data are missing or why some participants are missing from a particular wave of data collection. Moreover, this analysis can also help with choosing appropriate methods for handling missing data.

Missing data techniques

As noted earlier, results using three different missing data techniques are presented for each substantive analysis: listwise deletion, the DirML method, and the MI method. When using listwise deletion method for linear growth curve analyses, participants who did not participate in all three waves of testing were excluded from analyses. For regression analyses, in which an outcome variable from Wave 3 is predicted with variables from Wave 1, participants who did not participate in both Wave 1 and 3, regardless of Wave 2, were excluded from these particular analyses.

The PROC MIXED procedure in SAS version 9.1 was used to conduct the DirML method with linear growth curve models. The LISREL 8.72 (Jöreskog & Sörbom, 1996) program was used with regression analyses to conduct the DirML method. The MI method was conducted using PROC MI and PROC MIANALYZE components of SAS version 9.1. Ten imputed data sets were created each time the MI method was used. In order to improve the imputation process of the MI method, several auxiliary variables were used. For the linear growth analyses of depressive symptomatology, self-worth and maternal warmth scales from all three time points were used as auxiliary variables. For linear growth analysis of social competence, depressive symptomatology, and peer support scales measured at all three time points were used as auxiliary variables. All auxiliary variables were chosen based on their moderate and high correlations with variables used in further analyses. For regression analyses, information about the same measures assessed at different points in time were used to help the process of imputing missing data. Therefore, for the regression analysis in which academic competence was used as an outcome measure, academic competence variables from Grades 5 and 6, as well as parental monitoring and GPA at Grades 6 and 7, were employed as auxiliary variables. For the regression analysis where depressive symptomatology is used as an outcome measure, depressive symptomatology variables in Grades 5 and 6, as well as maternal warmth and self-worth measured in Grades 6 and 7 were used as auxiliary variables. Table 1 summarizes the empirical questions used for evaluating differences among three missing data procedures, corresponding data analysis methods that were used for answering these empirical questions, as well as variables used in each particular analysis.

Insert Table 1 about here

Results

Before presenting the results of substantive analyses, a preliminary analysis of the frequency of missing data (differentiating among attrition and variable nonresponse), as well as attrition analyses, are presented.

Frequencies of missing data and attrition

Frequencies of missing data for each variable used in analyses are presented in Table 2. Wave nonresponse and variable nonresponse are presented separately within a table in order to differentiate these two possible types of nonresponse. As presented in the Table 2, wave nonresponse for each wave ranged between 23.8% and 45.6%. The amount of variable nonresponse ranged from 6.3% to 22.2% among all three waves of testing.

Differences in the amount of variable nonresponse for the same variable among three waves of testing can be explained in two ways. First, the Student Questionnaire (SQ) of the 4-H Study of PYD changed in each year of testing in such a way that a particular scale was moved from one part of the questionnaire to another. For example, the parental monitoring scale was placed towards the end of the SQ in Waves 1 and 2, but in Wave 3 it was moved towards the beginning of the questionnaire; therefore, more participants were able to answer these questions. Second, the SQ is a large survey instrument comprised of about 400 questions (i.e., it is about 34 pages long at each wave) and participants usually need about an hour and a half to two hours to answer all questions. However, some participants were slower readers and not able to finish answering the whole SQ in the given time. It is possible that from Grade 5 to Grade 7 students became faster readers and therefore were able to answer more questions in Grade 7 than in Grade 5. Therefore, discrepancies in the amount of variable nonresponse for particular variables might

be due to the different position of a particular scale within the SQ at each time of testing, as well as to better reading abilities of participants in Grade 7. When considering the overall amount of missing data for each particular variable, almost all variables had about 55% of missing data at Wave 1, about 35% at Wave 2, and about 40% at Wave 3 (see Table 2). Descriptive statistics on the data available for all variables within each wave are presented in Table 3.

Insert Tables 2 and 3 about here

In addition, missing data for SES variables (i.e., family per capita income and mother's education) were calculated only at the variable nonresponse level, since information on SES variables was gathered from all three waves of the parent questionnaire and used to create one variable for mother's education and one variable for per capita family income. The family per capita income variable (M = 13,675.5, SD = 8569.87) had 30% missing data, and the mother's years of education variable (M = 13.6, SD = 2.48) had 27% missing data.

Considering that missing data in any longitudinal study can be comprised of both item non-response within each wave and wave non-response, the reason for missingness most likely varies depending on these two types of non-response. In order to explore reasons for item or, in this case, variable non-response within each wave, logistic regressions were conducted in order to predict missingness on outcome variables within each wave. Since all participants who had missing data on the social competence scale also had missing data on the academic competence scale (probably because the items for these two scales were placed together on the same page in the questionnaire), only two logistic regressions were conducted for each wave: one predicted missingness on the competence scales and the other predicted missingness on the depressive

symptomatology scale. All other variables available in the dataset for each wave were used as predictor variables.

Results of logistic regressions showed that variable non-response within each wave was predicted with other variables in the dataset. In Wave 1, of all predictor variables, GPA in Grade 5 was the strongest predictor of missingness on the competence scales, and had a negative parameter, such that participants with lower GPA in Grade 5 were less likely to complete questions related to the competence scales. Similarly, missingness on the depressive symptomatology scale was predicted by lower levels of academic competence in Grade 5. However, variable missingness in Wave 2 was predicted by several variables. Participants who were from the Multiracial/Other ethnic group, who were male and who had lower GPA in Grade 6 were less likely to complete questions related to the competence scales. Similarly, participants who were American Indian or Latino(a) and who had lower levels of academic competence in Grade 6 were less likely to complete questions related to the depressive symptomatology scale. In Wave 3, participants who were Latino(a), had lower levels of GPA and lower scores on peer support scale in Grade 7 were less likely to complete competence scale question. Missingness on the depressive symptomatology scale in Wave 3 was predicted by lower levels of academic competence in Grade 7 and sex (i.e., males were less likely to complete questions related to depressive symptomatology scale). These results indicate that non-response on the outcome variables within each wave were related to some variables in the data set which suggest that, within each wave, data are not missing completely at random (MCAR), but might be missing at random (MAR). In addition, the results also show a pattern of non-response across waves, suggesting that participants with lower GPA scores were less likely to respond to questions about

competence and participants with lower academic competence were less likely to respond to questions related to depressive symptomatology scale, across all three waves of data collection.

In addition to variable non-response, as shown in Table 2, wave nonresponse existed for each wave of testing. The wave nonresponse in Wave 1 was due to participants who were recruited later into the study at either Wave 2 or Wave 3. However, wave nonresponse in Wave 2 was due to both attrition and to participants joining the study in Wave 3. Wave 3 nonresponse represents missing data due to only attrition. As expected, attrition appeared in each wave of the study after the initial year of testing. Of 1719 participants from Wave 1, 63% continued to participate in the study. However, for the purpose of this study, participants who dropped out of the study due to school attrition were excluded from the sample. Therefore, of 1,232 participants from Wave 1, 88% continued to participate in the study. Similarly, of 756 new 6th grade participants that joined the study in Wave 2, and that came from schools that continued to be a part of the 4-H Study, 68% continued in Wave 3. Therefore, attrition analyses were conducted with the sample used for this study.

Attrition analyses were conducted on several background variables, such as, sex, race and SES, as well as on outcome variables (i.e., social and academic competence, and depressive symptomatology). Chi-square tests were used to compare attrition vs. non-attrition groups on sex and race distributions, and *t*-tests were used to compare these groups on the outcome and SES variables. Means and variances of these groups were compared separately. Two sets of attrition analyses were conducted using different groups in the following way: 1. Participants from Wave 1 who stayed in the study for the following one or two waves were compared with a group of participants who were present in Wave 1 but afterwards dropped out of the study; and 2. A new group of participants from Wave 2, who participated in Wave 3, were compared with

participants who dropped out of the study at Wave 3 and, therefore, had data from the Wave 2 data collection only.

The comparison of the attrition group from Wave 1 with the group that continued in the study showed that youth who continued in the study were more likely to be European American, less likely to be Asian American, and less likely to be African American. Other analyses did not show systematic differences between the two groups. Both groups had equivalent mean levels, as well as range of scores, on family per capita income, years of mother's education, and all outcome variables.

Comparison of the test-retest control attrition group from Wave 2 with the group that continued in Wave 3 showed that youth who continued in the study were more likely to be European American, less likely to be Asian American, and less likely to be Latino/a. When considering family level variables, youth who stayed in the study came from a family with higher family per capita income and had mothers with slightly more years of education than mothers of youth who dropped out from the study. However, both groups had equivalent mean levels, as well as a comparable range of scores, on all outcome variables (i.e., depressive symptomatology, academic competence, and social competence).

In addition to the attrition analyses described above, we wanted to determine the strongest predictors of attrition when all available variables in the data set were included together in the logistic regression model. As expected, the results were similar to those of the attrition analyses described above. Those participants who dropped out of the study after Wave 1 were more likely to be Asian and African American than European American, whereas those who dropped out of the study after Wave 2 were more likely to be Asian American, Latino(a), and American Indian than European American, and had mothers with fewer years of education.

In sum, both attrition analyses provided information on whether attrition can be accounted for by variables measured in the study. In the 4-H Study sample, differences between the group that continued in the study and the attrition group are not exactly the same for the subsample that started in Wave 1 and the subsample that started in Wave 2. Across both subsamples, participants who continued in the study are more likely to be European American, and less likely to be Asian American. Other variables that showed differences among the continuing versus the attrition group were specific to the Wave 2 attrition subsample.

Empirical Questions

The goal of conducting these empirical analyses was to answer two specific questions. First, how different are the results (i.e., point estimates and standard errors) obtained from an empirical study when different approaches for handling missing data are used? Second, to what extent do different results yield different interpretations of the phenomena that are studied? In order to answer these two questions, several criteria were used for comparing results from the three empirical analyses: Whether the obtained final models that best described the data (i.e., the most parsimonious model) were the same; similarity of individual parameter estimates and standard errors in the final; and comparability of the interpretations of results.

Change in social competence across three points in time

Change in social competence from Grade 5 to Grade 7 and the influence of sex differences on change over time were examined with four nested linear growth models in order to identify the most parsimonious model. Linear growth models were conducted with the PROC MIXED procedure in SAS. The first fitted model, M1, was specified as a model with no change in social competence over time. The second fitted model, M2, presents the linear growth in social competence over time. The sex variable, as a time invariant covariate, was introduced in the last

two models where the effects of sex on the level of social competence was tested in model M3, and the effects of sex on both the level and slope of change in social competence was tested in model M4. All four models were tested using the three different missing data techniques. Table 4 presents the final fitted linear growth model that best described the relationship between sex, time, and social competence using the listwise deletion method, the DirML method and the MI method.

Insert Table 4 about here

Because of the large amount of missing data on the social competence variables at each time of testing (between 16% and 18% of variable nonresponse) and the large amount of wave nonresponse (between 24% and 45%), the final sample size for analysis of social competence with listwise deletion was 422 participants. Therefore, only 19% of participants from the initial sample had data on social competence for all three times of testing. For analyses using the DirML and the MI method, data from 2,057 participants were used. Other participants did not have data on social competence at any point in time.

As Table 4 shows, in the model using the Listwise Deletion method, sex is not related to the level of, or change in, social competence, whereas in the model using the DirML and the MI methods, sex has a significant effect on the level of social competence. This contrast yielded different interpretations of final results: Using the DirML and the MI model, the results showed that social competence significantly increased from Grade 5 to Grade 7 (significant time effect), and the significant sex effect showed that girls have significantly higher levels of social

competence. The effect of sex was not significant in the analyses which used listwise deletion as a missing data method.

When the parameter estimates and their standard errors from the listwise deletion method are compared with the DirML and MI methods, it can be seen that the estimates are quite different for both sex and time effects. For example, the parameter estimate of sex in the listwise deletion method was 63% smaller from the estimate of the sex effect in the MI method, and 50% smaller than the estimate of the sex effect in the DirML method. Similarly, the standard error of the sex effect in the listwise deletion method was 50% larger than standard errors for the DirML and the MI methods.

Change in depressive symptomatology across three points in time

Change in depressive symptomatology from Grade 5 to Grade 7 and the influence of sex differences in change over time were also examined with four nested linear growth models. The analyses for change in depressive symptomatology paralleled those described for change in social competence.

Table 5 presents the final fitted linear growth models using the three methods. Because there was between 7 % and 15 % of variable nonresponse for the depressive symptomatology variable at the three times of testing and a large amount of wave nonresponse (between 24% and 45%), the final sample size for the analysis of depressive symptomatology with listwise deletion was 504 participants. Therefore, only 22% of participants from the initial sample had data on depressive symptomatology for all three times of testing. For analyses using DirML and MI methods, data from 2,121 participants were used. Other participants did not have data on depressive symptomatology at any point in time.

Insert Table 5 about here

As Table 5 shows, the most parsimonious model in describing the relationships between sex, time, and depressive symptomatology was the same using all three missing data methods. For depressive symptomatology, the significant time by sex effect indicated that change over time was different for boys and girls. In order to interpret the results of the final models, prototypical fitted plots were created for females and males, which are presented in Figures 1 (using the Listwise deletion method), Figure 2 (using the DirML method) and Figure 3 (using the MI method). All three prototypical plots afforded the same interpretations for the relationship between sex, time, and log-depression: Boys on average decreased in log-depression scores over time, whereas girls' log-depression scores increased from Grade 5 to Grade 7.

Insert Figures 1, 2, and 3 about here

However, when parameter estimates and their standard errors of the three final models are compared, the estimates are different. The estimates are different for all three techniques. For example, the parameter estimate of the sex variable in the listwise deletion method was 6% larger from the estimate in the DirML method, but the listwise deletion estimate was 62% larger from the MI method. Therefore, the estimate for the sex variable in the DirML method was 60% larger than the estimate from the MI method. In addition, estimates of the interaction effect between the three methods had a similar pattern of difference: The largest variation was between the MI estimate and the DirML and the listwise deletion estimates, where the ML estimate was 51% larger than the MI estimate, and the listwise deletion estimate was 55% larger than the MI

estimate. However, the estimates of the standard errors for the DirML and the MI methods were almost identical, whereas the estimates of the standard errors for the listwise deletion method were 30% larger than standard errors for the DirML and the MI methods.

Predicting academic competence in Grade 7

The relationship between academic competence in Grade 7 and parental monitoring, adolescents' self-reported GPA level at Grade 5, as well as measures of sex and SES, were examined with three nested regression models. The first fitted model, M1, was specified as a model in which self-perceived academic competence in Grade 7 was predicted with sex, self-perceived GPA, and self-perceived parental monitoring variables measured in Grade 5. In the second fitted model, M2, two SES variables were added as predictor variables -- that is family per capita income and mother's level of education were added -- in order to test whether SES variables significantly contributed to the explanation of the outcome variance. Interaction effects among sex and all other predictor variables were added in the last model, M3. All three regression models were tested using the three different missing data techniques, as before. Furthermore, in order to remove nonessential multicollinearity among the variables in the model, two predictor variables, GPA and parental monitoring, were centered, which was done by subtracting the mean of the variable from each score (Cohen, Cohen, West, & Aiken, 2003).

Table 6 presents final regression models that describe the relationship between academic competence in Grade 7, and sex, GPA, and parental monitoring using the listwise deletion method, the DirML and the MI methods. Because of the large amount of variable nonresponse for all variables included in the regression models (e.g., 30% for SES variables) and the large amount of wave nonresponse (45% at Wave 1 and 32% at Wave 3), the final sample size for the analysis of academic competence with listwise deletion was 410 participants. Therefore, only

18% of participants from the initial sample had data on academic competence for all three times of testing. For the analyses using the DirML and the MI methods, data from 1,758 participants were used. Other participants did not have data on outcome and predictor variables.

Insert Table 6 about here

As Table 6 shows, the results of the three missing data techniques show that the final model was the same using any of the three missing data techniques. However, interpretations of the final models were different. The results using the DirML and the MI methods show that a 5th grade student who is a male and had higher scores on self-perceived GPA and parental monitoring had higher scores on self-perceived academic competence in 7th grade. The final model of the listwise deletion method did not have a significant effect of the sex variable, whereas in the final models of the other two missing data techniques, the sex variable had a significant effect.

In addition, when the parameter estimates and their standard errors of the final model from a listwise deletion method were compared with the final models of the recommended techniques, the estimates were quite different. For example, the parameter estimate of the sex variable in the listwise deletion method was 33% larger than the estimates in the MI and the DirML methods, which were almost identical. In addition, estimates of standard errors were much larger in the listwise deletion method, compared to those of the recommended techniques (ranging from 30% to 80%.

Predicting depressive symptomatology in Grade 7

As in the analyses for predicting academic competence, the relationship between depressive symptomatology in Grade 7 and maternal warmth and self-worth in Grade 5, as well as measures of sex and SES, were examined with three nested regression models. The first fitted model, M1, was specified as a model in which depressive symptomatology measured in Grade 7 was predicted by sex, self-perceived self-worth, and self-perceived maternal warmth variables measured in Grade 5. In the second fitted model, M2, two SES variables were added as predictor variables, family per capita income and mother's level of education, in order to test whether SES variables significantly contributed to the explanation of the variance in the outcome variable. Interaction effects between sex and all other predictor variables were added in the last model, M3. All three regression models were tested using the three different missing data techniques. Furthermore, in order to remove nonessential multicollinearity among the variables in the model, two predictor variables, self worth and maternal warmth, were centered, which was done by subtracting the mean of the variable from each score (Cohen, et al., 2003).

Table 7 presents the final regression models that describe the relationship between log-depressive symptomatology and sex, self-worth, and maternal warmth using the listwise deletion method, the DirML and the MI methods. Because of the large amount of variable nonresponse for all variables included in the regression models (e.g., 30% for SES variables), and the large amount of wave nonresponse (45% at Wave 1 and 32% at Wave 3), the final sample size for the analysis of depressive symptomatology with listwise deletion was 403 participants. Therefore, only 18% of participants from the initial sample had data on social competence for all three times of testing. For the analyses using the DirML and the MI methods, data from 1,924 participants were used. Other participants did not have data on the outcome variable, as well as all predictor variables.

Insert Table 7 about here

As Table 7 shows, the final model did not include the SES variables and the interaction effects between sex and other predictor variables since they did not significantly contribute to the explanation of variance in log-depression. Therefore, model M1 represented the most parsimonious model in the explanation of a relationship between log-depression and variables measured at a previous time point. The model showed that sex, self-perceived self-worth, and maternal warmth measured in Grade 5 were significantly related to log-depression measured in Grade 7. The interpretation of results was the same across all three different missing data methods: Higher levels of log-depression in 7th grade were related to lower levels of self-worth and maternal warmth in 5th grade. In addition, girls had higher log-depression levels in 7th grade than did boys.

Even though the results with the three missing data techniques for the final model were the same, differences among these techniques are evident when comparing parameter estimates and standard errors: the estimates of each predictor in the listwise deletion method were larger than the estimates in the recommended techniques; however, estimates in the DirML method were not similar to the MI method either. For example, the estimate of the sex variable in the listwise deletion method was 12% larger than the estimate in the DirML method, and 25% larger than the estimate in the MI method. Similarly, the estimate of self-worth in the listwise deletion method was 18% larger than the estimate in the DirML method, and 31% larger than the estimate in the MI method.

These results suggest differences in parameter estimates between the two recommended techniques. For example, estimates of sex and self-worth variables in the DirML method were 15% larger than the estimates in the MI method. However, the standard errors of predictors in the recommended techniques were almost identical and were between 30% and 50% smaller than the standard errors in the listwise deletion method. Although differences in parameter estimates did not change the interpretation of the results among the different missing data techniques, the differences are evident. Therefore, the results suggest that for the empirical example of predicting depressive symptomatology in Grade 7 of the 4-H Study sample, all three missing data techniques did not yield the same results.

In sum, the results for the four empirical questions suggest that the three missing data techniques produce somewhat different results. These are summarized below.

Discussion

The understanding of adolescent development rests on the methodologically sound collection and interpretation of longitudinal data (e.g., see von Eye, 1990a, 1990b). However, among the key challenges of such research is the treatment of missing data that, inevitably, exists in longitudinal studies of adolescence. The purpose of this paper was to examine the extent to which results of an empirical longitudinal study of adolescent development differ in regard to different missing data techniques: We considered the listwise deletion method, which is the most frequently used method, and the DirML and MI methods, both of which are recommended by statisticians but seldom used in practice. Using the data set from the first three waves (Grades 5, 6, and 7) of the 4-H Study of Positive Youth Development, three different missing data techniques were used in order to provide answers to the following two questions: How different are the results obtained from an empirical study when different missing data approaches are

used; and to what extent do different results yield different interpretations of the phenomena that are studied? In particular, the focus was on two sets of analyses: growth curve modeling (e.g., change in social competence across early adolescence), and predicting a specific outcome from a set of variables at previous points in time (e.g., to what extent can depressive symptomatology in Grade 7 be predicted with parents' behaviors and youths' self-worth assessed in Grade 5). As such, this research explored how various missing data techniques performed depending on the type of data analysis procedure that was used for testing research questions in longitudinal studies. The performance of missing data techniques was evaluated on four substantively different empirical questions.

Results from the first two empirical questions tested with linear growth models showed that the three missing data techniques produced different parameter estimates, but all three methods showed the same patterns of findings. In the example of social competence, sex is not a significant predictor in the listwise deletion results, but it is significant in both the MI and DirML methods, while for the outcome of depressive symptomatology, the pattern of results is the same for all three methods, that is, change over time is different for boys and girls. These results, again, suggest that the listwise deletion method is less appropriate to use, since it produces larger standard errors, which was expected, and different parameter estimates than the recommended techniques. However, in this empirical example it is not clear why there are such large differences in parameter estimates between the two recommended techniques. One possible explanation of the difference between the DirML method and the MI method in this empirical example is associated with the use of auxiliary variables with the MI method. Use of auxiliary variables may have provided better estimates in the MI method, and therefore produced different results from the DirML method (Collins, Shafer, & Kam, 2001).

Results from the prediction analyses also showed that the three missing data techniques produced somewhat different parameter estimates. For depressive symptomatology, the patterns of results were the same, with significant effects for sex, self worth and mother's warmth. For academic competence, the listwise deletion method did not produce a significant effect for sex, but for the MI and DirML procedures all three predictors were significant. Again, the results showed that the three missing data techniques did not always yield identical results for different research questions assessing particular aspect of development (i.e., change or prediction effects); however, the results suggested that the listwise deletion method produces the most disparate results from the recommended missing data techniques, as suggested in the literature. In the cases where listwise deletion produces different results from both the DirML and the MI method, it could be argued that listwise deletion produced biased results, since simulation studies in the current literature come to the same conclusion (e.g., Arbuckle, 1996; Graham, Cumsille, & Elek-Fisk, 2003; Little & Rubin, 2002; Sinharay, Stern, & Russell, 2001; Wothke, 2000).

In summary, the comparison of results from the first two empirical questions shows that the performance of three missing data techniques on two substantively different empirical questions is different, even though the questions tested similar aspects of development (i.e., change) but with a different phenomenon (i.e., adolescents' social competence and adolescents' depressive symptomatology). Discrepancies between results of the two empirical questions using the three missing data techniques suggest that these three techniques might not provide the same results every time they are used in an empirical study. The discrepancies may be due to the amount of missing data for particular variables used in analyses and/or to the pattern of missing data in a data set with different sets of variables, as well as to use of auxiliary variables. In addition, it was evident that smaller sample size and loss of power negatively affected results of

the listwise deletion method. As noted before, these results were expected, since one of the major disadvantages of the listwise deletion method is small sample size and, therefore, lower statistical power.

Implications of results from the empirical examples using different missing data techniques

As the present analyses showed and the literature suggested, the listwise deletion method is not an appropriate method for solving the problem of missing data. With large amounts of missing data it can substantially lower the sample size and, therefore, introduce low statistical power. More importantly, discrepancies in parameter estimates between the listwise deletion method and the recommended techniques, as well as substantially larger standard errors in the listwise deletion method, all contribute to the conclusion that the listwise deletion method should not be used. Since most longitudinal researchers are still using the listwise deletion method as the preferred missing data technique (Jelicic et al., 2009), this conclusion has important implication for research practice. Not only will theoretical understanding be hampered because of inappropriate inferences drawn from methodologically-biased data sets but, in addition, child-, youth-, and family-centered programs and policies that are based on implications from such studies will be ill-informed. Study results obtained through the use of older missing data techniques may not be appropriate for informing programs and policies affecting individuals and groups. As shown from the examples in this paper, the use of a non-recommended approach to treating missing data actually yields results different from those associated with recommended approaches. This finding suggest that the use of the older approaches might have serious implications for theory and programs and policies related to adolescents.

Although the results of the present analyses suggested that the listwise deletion method is less appropriate to use than the recommended techniques, differences in results between the two

recommended techniques were not expected, and it is not clear which of the two techniques provided more “correct” results. Since the “true” data are not available for missing data in any empirical study, it is not possible to conclude which of the two methods provides the “correct” results for a particular sample. However, the present research provided information for evaluating the differences in results and in the interpretations of results that are afforded by the use of different missing data techniques. This information is important for any empirical study, since it can provide insight about the generalizability of study results.

For example, large discrepancies in the interpretation of results analyzed with different missing data techniques suggest that results cannot be generalized to a population or the sample used in the study, and that results can only be applicable to the smaller subsample of the study that has no missing data. Therefore, a suggestion for researchers is to use several different missing data techniques, particularly the DirML and the MI method, in order to see if and how the results of their analysis change. If the results with different missing data techniques are similar, and there are no large discrepancies in parameter estimates and standard errors between the recommended techniques, a researcher could be confident that the results are valid, and can be generalized to a population.

However, before computing any analyses with different missing data techniques, each researcher should examine the amount of missing data for each variable and the pattern of missing data that exists in the data set used for analysis. A small amount of missing data (e.g., 5% or less) does not create a serious problem, and most missing data procedures in such cases will yield similar results (Graham, Cumsille, & Elek-Fisk, 2003). However, if there is a large percentage of missing data, the issue of handling missing data becomes more serious and

complex. Unfortunately, there are still no specific guidelines on how much missing data are considered “too much.”

Decisions about what to do about missing data should be considered on a case-by-case basis. Researchers should explore how much missing data exist in their data set compared to the sample size in their study. For example, 20% of missing data creates more serious problem for studies with a small sample, because of already low statistical power, than it does for studies with a large sample. However, even in the case of studies with a large sample size, missing data still create a problem that needs to be appropriately addressed.

The inspection of missing data accompanied by comparisons between cases with and without missing data should be a part of an exploratory data analysis, and should be reported in research reports in order to provide a valid description of the data used in analyses. In addition, for longitudinal studies, attrition analyses of important demographic variables and of the outcome variables should be conducted in order to obtain information about whether attrition can be accounted for by variables measured in the study. The inspection of missing data and the use of attrition analysis in longitudinal studies should be a part of exploratory data analysis that is conducted before running any analysis to examine research phenomena about which one is interested.

In addition to considering what to do about missing data after the data are collected, researchers should consider how best to avoid having a large amount of missing data in their studies. Even though statisticians are working to improve missing data techniques, these techniques will never replace actual data; therefore, it is still better not to have missing data than to need to use missing data techniques. One way to avoid having large amounts of missing data when conducting survey research is to follow some general research methods guidelines for

creating survey questionnaires, such as creating questions that are appropriate for the age group being studied, making sure that the questionnaire is not too long, and determining that questions are not too complex. It is important to consider the potential participants and their abilities, particularly when survey research is conducted on a group of middle-school students and early adolescents, who may be expected to vary in cognitive abilities, personality, and temperament characteristics.

Moreover, longitudinal researchers have additional challenges with missing data, since they need to consider how to retain their longitudinal sample. It is well known that in longitudinal studies with large samples it is not possible to expect a 100% sample retention rate over the course of the study; attrition is inevitable. However, it is expected that the sample retention rate will be high and attrition will be unbiased in order to obtain valid results, that is, findings representative for that population. Therefore, longitudinal researchers should consider how to decrease attrition and maintain a representative sample. Not surprisingly, in many cases the reasons for attrition are not under the complete control of the researchers and therefore, it is not possible to maintain a 100% retention rate. However, it is possible to explore the reasons for attrition; such explanation can be beneficial for informing the results of the study. For example, as described in the Method section, attrition in the 4-H Study sample was the result of both individual and school attrition. Further analyses of reasons for attrition can determine the nature of the schools that dropped out (if there are systematic differences between schools that stayed in the study and the ones that dropped out), and how these differences can impact the generalization of results.

It is important to consider when planning research studies that a longitudinal researcher should not just depend on missing data techniques to solve the problem of missing data; he or

she should also take actions before data collection, such as developing appropriate questionnaires and developing mechanisms to retain participants at a high rate. Such actions should decrease attrition and missing data. Planning to prevent missing data and handling missing data after they occur should be considered two parts of a process for managing missing data in any research study.

Missing data are present in almost every longitudinal research study, particularly in the study of adolescence. Therefore, it is important to understand the consequences of ignoring missing data, to learn about missing data theory, and to understand the impact of missing data on the generalizability of study results. As such, researchers must become familiar with available missing data techniques in order to conduct valid scientific studies exploring adolescent development.

References

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage Publications.

Arbuckle, J. L. (1996). Full information estimation procedure in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advance structural equation modeling* (pp. 243 – 277). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Armsden, G., & Greenberger, M. (1987). The inventory of parent and peer attachment: Individual differences and their relationship to psychological well-being in adolescence. *Journal of Youth and Adolescence*, *16*, 427-452.

Baltes, P. B., Reese, H. W., & Nesselrode, J. R. (1988). *Introduction to research methods: Life-span developmental psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multivariate regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Collins, L. M., Schafer, J. W., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6(4)*, 330-351.

Daniels, M.J. & Hogan, J.W. (2000). Reparameteirizing the pattern mixture model for sensitivity analysis under informative dropout. *Biometrics, 56,* 1241-1248.

Enders, C. K. & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8(3)*, 430-457.

Foster, E.M., Fang, G.Y., & Conduct Problems Prevention Research Group (2004). Alternative

methods for handling attrition in prevention research: An illustration using data from Fast Track. *Evaluation Review, 28*, 434-464.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.). *Research Methods in Psychology* (pp. 87-114). Volume 2 of *Handbook of Psychology* (I. B. Weiner, Editor-in-Chief). New York: John Wiley & Sons.

Graham, J. W. & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology, 78,* 119-128.

Graham, J. W. & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 201-218). Mahwah, NJ: Lawrence Erlbaum Associates.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*, 197-218.

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

Gray, M. R., & Steinberg, L. (1999). Unpacking authoritative parenting: Reassessing a multidimensional construct. *Journal of Marriage and the Family*, *61*, 574-587.

Harel, O. & Zhou, X.H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine, 26,* 3057-3077

Harter, S. (1982). The Perceived Competence Scale for Children. *Child Development*, *53*, 87-98.

Harter, S. (1983). Supplementary description of the Self-Perception Profile for Children: Revision of the Perceived Competence Scale for Children. Unpublished manuscript, University of Denver.

Jelicic, H., Phelps, E., & Lerner, R.M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, *45(4)*, 1195-1199

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association, 91(433)*, 222-230.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

Lerner, R. M., Lerner, J. V., Almerigi, J., Theokas, C., Phelps, E., Gestsdottir, S. Naudeau, S., Jelicic, H., Alberts, A. E., Ma, L., Smith, L. M., Bobek, D. L., Richman-Raphael, D., Simpson, I., Christiansen, E. D., & von Eye, A. (2005). Positive youth development, participation in community youth development programs, and community contributions of fifth grade adolescents: Findings from the first wave of the 4-H Study of Positive Youth Development. *Journal of Early Adolescence, 25(1),* 17-71.

Lerner, R. M., & Steinberg, L. (Eds.) (2009). *Handbook of Adolescent Psychology, 3rd edition*. New York: John Wiley & Sons.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.) Hoboken, NJ: Wiley-Interscience.

Molenberghs, G., Kenward, M.G., & Lesaffre, E. (1997). The analysis of longitudinal ordinal

data with non-random dropout. *Biometrika, 84,* 33-44.

Peugh, J. L. & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of educational research,. 74(4)*, 525 – 556.

Phelps, E., Balsano, A.B., Fay, K., Peltz, J.S., Zimmerman, S., Lerner, R.M., & Lerner, J.V. (2007). Nuances in early adolescent developmental trajectories of positive and problematic/risk behaviors: Findings from the 4-H study of positive youth development. *Child and Adolescent Psychiatric Clinics of North America, 16(2),* 473-496.

Phelps, E., Zimmerman, S., Warren, A.E.A., Jelicic, H., von Eye, A., & Lerner, R.M. (2009). The structure and developmental course of Positive Youth Development (PYD) in early adolescence: Implications for theory and practice. *Journal of Applied Developmental Psychology, 30(5*), 571-584.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385-401.

Roth, P. L., Switzer, F. S., & Switzer, D. M. (1999). Missing data in multiple item scales: Monte Carlo analysis of missing data techniques. *Organizational Research Methods, 2(3)*, 211-232.

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

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods, 7(2),* 147-177.

Schludermann, E., & Schluderman, S. (1970). Replicability of factors in children's reports of parent behavior (CRPBI). *Journal of Psychology*, *76*, 239-249.

Sinharay, S., Stern, H. S., Russell, D. (2001). The use of multiple imputation for the

analysis of missing data. *Psychological methods, 6(4)*, 317-329.

Small, S. A., & Kerns, D. (1993). Unwanted sexual activity among peers during early and middle adolescence: Incidence and risk factors. *Journal of Marriage and the Family*, *55*, 941-952.

von Eye, A. (Ed.). (1990a). *Statistical methods in longitudinal research. Volume 1. Principles and structuring change*. New York: Academic Press.

von Eye, A. (Ed.). (1990b). *Statistical methods in longitudinal research. Volume 2. Time series and categorical longitudinal data*. New York: Academic Press.

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.): *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 219-240). Mahwah, NJ: Lawrence Erlbaum Associates.

Table 1.

Variables and corresponding data analysis methods used for examining four empirical questions.

	Empirical questions			
	1. Change in social competence over Grades 5, 6, and 7.	2. Change in depressive symptomatology over Grades 5, 6, and 7.	3. Prediction of academic competence in Grade 7	4. Prediction of Depressive symptomatology in Grade 7
Data analysis method	Linear growth modeling	Linear growth modeling	Regression analysis	Regression analysis
Outcome (OV) and predictor (PV) variables used in the analysis	Social competence in Grades 5, 6, & 7 (OV), & Sex (PV)	Depressive symptomatology in Grades 5, 6, & 7 (OV), & Sex (PV)	Academic competence in Grade 7 (OV), Parental monitoring (PV), GPA (PV), SES (PV), & Sex (PV)	Depressive symptomatology in Grade 7 (OV), Maternal warmth (PV), Self-worth (PV), SES (PV), & Sex (PV)

Auxiliary variables used for the MI method	Depressive symptomatology & Peer support from all three waves	Self-worth & Maternal warmth from all three waves	Academic competence in Grades 5 & 6, GPA & Parental monitoring in Grades 6 & 7	Depressive symptomatology in Grades 5 & 6, Maternal warmth & Self-worth in Grades 6 & 7

Table 2.

Percentage of wave and variable (VN) nonresponse for each variable within each wave.

	Wave 1		**Wave 2**		**Wave 3**	
Percentage of wave nonresponse *	45.6		23.8		32.1	
	VN**	Overall*	VN**	Overall*	VN**	Overall*
Maternal warmth	17.0	54.8	12.9	33.6	8.9	38.1
Parental monitoring	19.2	56.0	14.5	34.9	8.3	37.7
Social competence	18.3	55.5	18.1	37.6	16.3	43.1
Academic competence	18.1	55.5	17.9	37.4	16.4	43.2
Self-worth	18.8	55.8	17.7	37.3	16.6	43.4
Depressive symptomatology	15.3	54.0	10.3	31.6	7.3	37.0

Peer support	9.1	50.6	6.3	28.6	18.4	44.6
GPA	22.2	57.7	15.5	35.6	10.5	39.2

*Percentages were calculated based on the overall sample size for the whole data set (N = 2265).

**Percentages were calculated based on the sample size for each particular wave of data: N_{w1} = 1232, N_{w2} = 1726, and N_{w3} = 1538.

Table 3.

Means and standard deviations for each variable within each wave.

	Wave 1			Wave 2			Wave 3			
	M	SD	N	M	SD	N	M	SD	N	Scale range
Maternal warmth	4.04	1.014	1023	3.86	1.059	1504	3.83	1.017	1401	1 – 5
Parental monitoring	3.50	0.622	996	3.42	0.714	1475	3.31	0.753	1411	0 – 4
Social competence	2.99	0.617	1007	3.01	0.611	1414	3.05	0.597	1288	1 – 4
Academic competence	2.94	0.663	1009	2.99	0.622	1417	2.96	0.631	1286	1 – 4
Self-worth	3.14	0.607	1000	3.09	0.602	1420	3.09	0.605	1283	1 – 4
Depressive symptomatology	13.68	8.934	1043	12.77	9.261	1549	13.42	9.695	1426	0 – 60
Peer support	4.27	0.914	1120	4.23	0.864	1617	4.12	0.884	1255	1 – 5
GPA	3.35	0.684	959	3.31	0.710	1458	3.25	0.784	1377	0.5 – 4

Table 4.

Parameter estimates with corresponding standard errors, approximate p-values, residual variance estimates with corresponding standard errors, and the -2LL fit statistics for final growth linear models that describe the relationship between student's social competence over time and the relationship with sex using the Listwise Deletion method, the DirML method and the Multiple Imputation (MI) method.

	Final model using Listwise deletion method (N=422)	**Final model using the DirML method** (N = 2057)	**Final model using the MI method** (N = 2057)
Intercept, γ_{00}	2.963 (0.047)***	2.879 (0.029)***	2.845 (0.026)***
Time, γ_{10}	0.063 (0.015)***	0.034 (0.010)***	0.040 (0.009)***
Sex, γ_{01}	0.033 (0.046)	0.089 (0.024)***	0.066 (0.022)**
σ_{ε}^2 [var(ε_{ij})]	0.171 (0.012)***	0.179 (0.009)***	0.182 (0.008)***
τ_{00} [var(μ_0)]	0.222 (0.051)***	0.255 (0.042)***	0.206 (0.048)***
τ_{11} [var(μ_1)]	0.09 (0.009)	0.011 (0.008)	0.010 (0.008)
τ_{01} [cov(μ_0, μ_1)]	-0.024 (0.019)	-0.029 (0.017)	-0.021 (0.018)
-2LL	1965.6	6335.0	9857.4

Key: *p<.05; **p<.01; ***p<.001

Table 5.

Parameter estimates with corresponding standard errors, approximate p-values, residual variance estimates with corresponding standard errors, and the -2LL fit statistics for final growth linear models that describe the relationship between student's depressive symptomatology over time and the relationship with sex using the Listwise Deletion method, the DirML method and the Multiple Imputation (MI) method.

	Final model using Listwise deletion method (N=504)	**Final model using the DirML method** (N = 2121)	**Final model using the MI method** (N = 2121)
Intercept, γ_{00}	2.519 (0.068)***	2.585 (0.043)***	2.572 (0.045)***
Time, γ_{10}	-0.103 (0.028)***	-0.080 (0.019)***	-0.062 (0.019)***
Sex, γ_{01}	-0.189 (0.090)*	-0.178 (0.058)**	-0.072 (0.061)
T * Sex, γ_{11}	0.120 (0.038)***	0.106 (0.025)***	0.054 (0.024)*
σ_{ε}^2 [var(ε_{ij})]	0.307 (0.019)***	0.286 (0.013)***	0.280 (0.011)***
τ_{00} [var(μ_0)]	0.286 (0.078)***	0.243 (0.055)***	0.237 (0.054)***
τ_{11} [var(μ_1)]	0.021 (0.015)	0.016 (0.010)	0.023 (0.010)*
τ_{01} [cov(μ_0, μ_1)]	-0.030 (0.031)	-0.022 (0.022)	-0.032 (0.023)
-2LL	3189.3	8330.4	12711.9

Key: *p<.05; **p<.01; ***p<.001

Table 6.

Parameter estimates with corresponding standard errors, approximate p-values, and R^2 statistics for the final regression model that describe the relationship between student's self-perceived academic competence in Grade 7 and student's self-perceived GPA and parental monitoring in Grade 5, sex, and SES variables using the Listwise Deletion method, the DirML method and the Multiple Imputation (MI) method.

	Final model using Listwise deletion method (N=410)	**Final model using the DirML method** (N = 1758)	**Final model using the MI method** (N = 1758)
Intercept	3.274 (0.092)***	3.076 (0.017)***	3.050 (0.049)***
Sex	-0.099 (0.056)	-0.077 (0.034)*	-0.078 (0.030)**
GPA	0.376 (0.047)***	0.373 (0.033)***	0.320 (0.030)***
Parental monitoring	0.207 (0.050)***	0.225 (0.036)***	0.136 (0.037)***
R^2	0.20***	0.22***	0.16***

Key: *p<.05; **p<.01; ***p<.001

Table 7.

Parameter estimates with corresponding standard errors, approximate p-values, and R^2 statistics for the final regression model that describe the relationship between student's depressive symptomatology in Grade 7 and student's self-perceived self-worth and maternal warmth in Grade 5, sex, and SES variables using the Listwise Deletion Method, the DirML method and the Multiple Imputation (MI) method.

	Final model using Listwise deletion method (N=403)	**Final model using the DirML method** (N = 1924)	**Final model using the MI method** (N = 1924)
Intercept	2.003 (0.121)***	2.165 (0.019)***	2.221 (0.062)***
Sex	0.193 (0.074)**	0.170 (0.038)***	0.144 (0.036)***
Self worth	-0.303 (0.066)***	-0.248 (0.047)***	-0.208 (0.046)***
Mother's warmth	-0.122 (0.039)**	-0.107 (0.028)***	-0.102 (0.025)***
R^2	0.10***	0.08***	0.07***

Key: *p<.05; **p<.01; ***p<.001

Figure 1

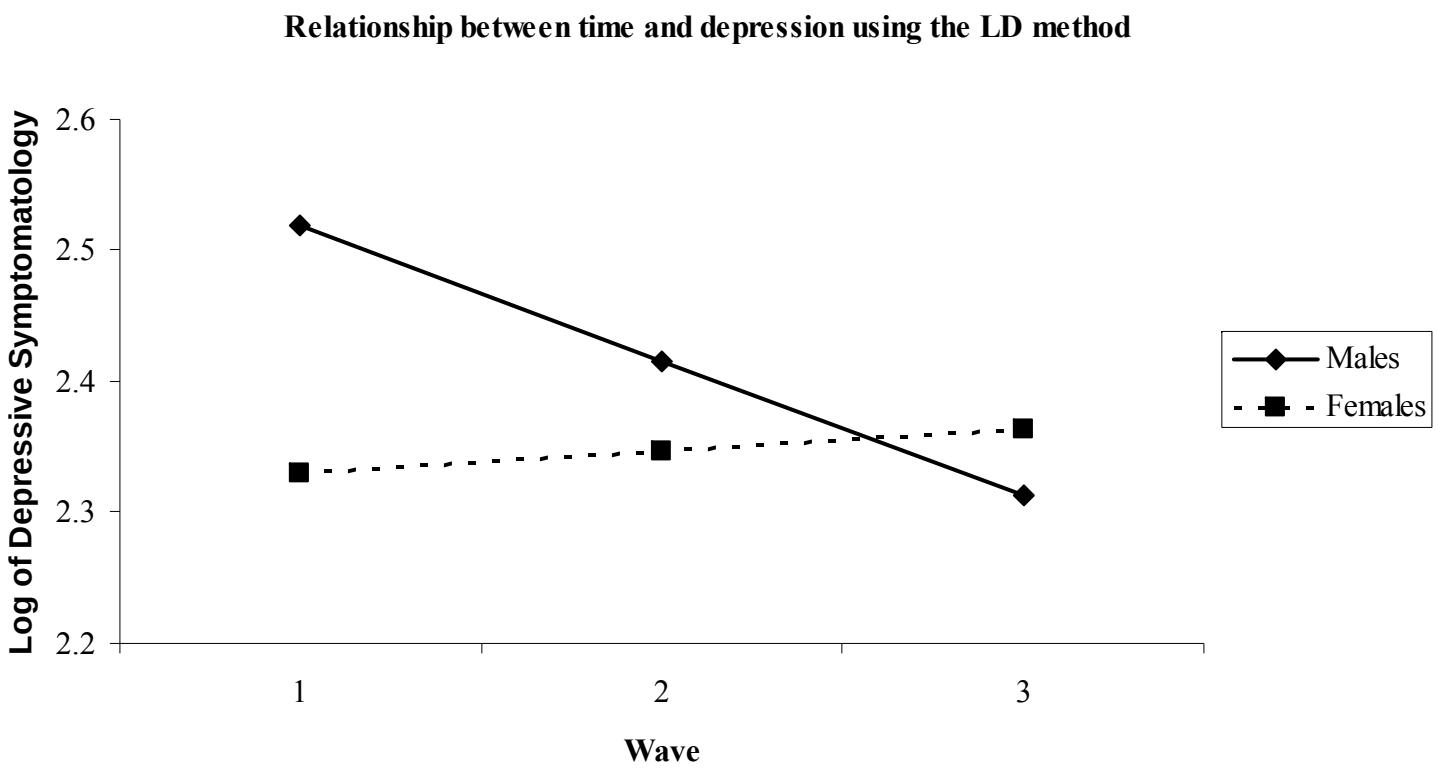

Relationship between time and depression using the LD method
Log of Depressive Symptomatology
2.6
2.5
2.4
2.3
2.2
1
2
3
Wave
Males
Females

Figure 2

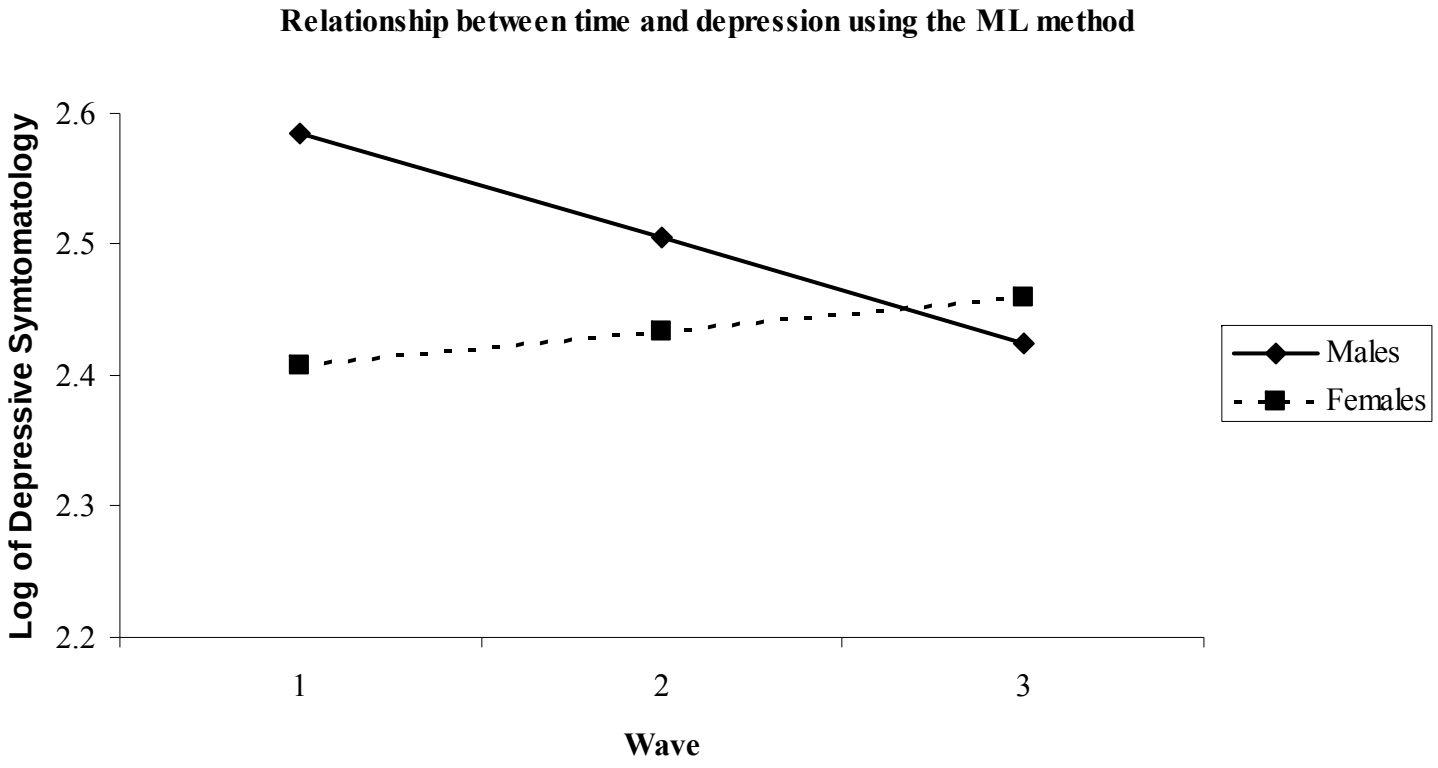
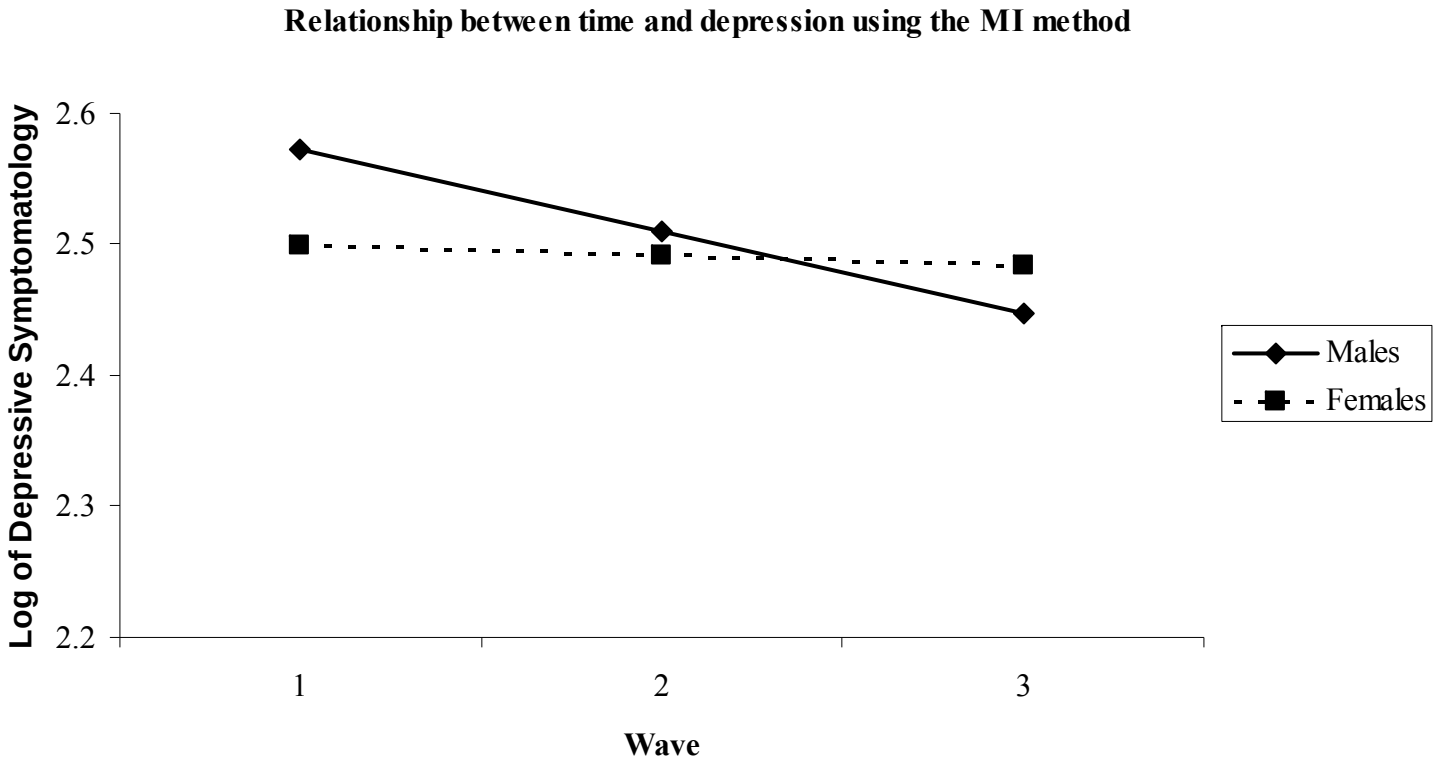

Relationship between time and depression using the ML method
Log of Depressive Symtomatology
2.6
2.5
2.4
2.3
2.2
1
2
3
Wave
Males
Females

Figure 3


Relationship between time and depression using the MI method
Log of Depressive Symptomatology
2.6
2.5
2.4
2.3
2.2
1
2
3
Wave
Males
Females