International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Dealing with Attrition and Missing Data in Longitudinal Studies: A Critique

Meshack Nzesei Mutua[a]*, Shiro Mogeni[b*], Janet Mueni Kilonzi[c*]

[a]*Institute of Education, University College London, 20 Bedford Way, WC1H 0AL, UK*

[b]*Mount Kenya University, P.O Box 13495, Nairobi – 00100, Kenya*

[c]*University of Nairobi, P.O Box 30197, Nairobi – 00100, Kenya*

[a]*Email: m.mutua.16@ucl.ac.uk*

[b]*Email: smogeni@mku.ac.ke*

[c]*Email: janmukil@yahoo.com*

Abstract

Longitudinal studies in teenage pregnancy, like any other topical issue, have been greatly affected by the problem of participant drop out and missing data. Although existing evidence indicate that data from longitudinal studies can provide useful insights regarding individual behaviours; the quality of the data and representativeness of the findings can adversely be compromised by attrition and missing data. Using literature review, this article criticises two papers on teenage pregnancy that have utilised longitudinal data by examining the effectiveness of the measures taken to correct for the attrition and missing data. The article concludes that procedural strategies used to eliminate or reduce attrition and missing data before and during data collection are more effective than employing statistical strategies to deal with the effects afterwards. If statistical procedures have to be used, it is important to first make distinctions of the missing data mechanisms since this has a bearing on whether certain strategies of handling missing data – such as list-wise deletion, pairwise deletion, mean imputation or multiple imputation –will result in biases or not.

Keywords: attrition; missing data; longitudinal studies.

---

* Corresponding author.

1. Introduction

Available literature has shown that adolescent pregnancy has both short-term and long-term consequences to young mothers and their children [7,36,29], and by extension, teenage fathers [19]. Some of these consequences have to do with their education attainment, health status (physical and psychological) and, social and economic wellbeing, although [28] believe that the educational experiences of teenage parents are largely and generally shaped and influenced by macro- and micro -societal proactive/reactionary and conscious or sub-conscious attitudes and response towards adolescent pregnancy and childbearing.

While different research designs have been utilised to understand the phenomenon, a significant number of studies have used longitudinal design – mostly analysing data collected through regional/national surveys and/or census. Each of those designs have their strengths and limitations. Longitudinal study involves continuous or repeated measures with a group of participants over a given period of time. The method is "observational in nature, with quantitative and/or qualitative data being collected on any combination of exposures and outcomes, without any external influenced being applied" [8].

Scientists (for instance, [25,24,27,12,20,22,15,6,8]) have argued that longitudinal studies, though expensive and tiresome to undertake, can yield important data to inform policy and interventions. In fact, Reis *et. al*. (2000) goes on to assert that longitudinal studies have the important advantage of allowing the researcher to directly observe intraindividual change over time. However, some concerns have been raised about attrition rates and missingness of data, and their impacts on longitudinal research which can ultimately compromise the sample and generalizability of the findings.

According to [24], missing data can emanate from a failure by the participants to answer some questions on a survey or due to unavailability of participants (attrition or dropout) in the second and/or subsequent waves in a longitudinal study. Notably, researchers should be concerned about attrition and missing data because they not only reduce sample size substantially effecting the statistical power and subsequently diminishing the efficiency of estimates, but also induce bias especially when nonresponse is selective (non-random). In this essay, we analysed how attrition rates (dropping out of participants during data collection) and missing data have been dealt with, and how effective those measures were. We reviewed two articles which have analysed data collected using a longitudinal design.

One of the articles, which examined how school dropout relate to teenage pregnancy utilised data from the Health and Socio-demographic Surveillance System (HDSS); covering Agnicourt Sub-district in Mpumalanga Province, South Africa. With the exposure being education enrolment status and the outcome being the incident teen pregnancy, the researchers constructed cohort by identifying teenage women aged 12 - 18 years and had resided in the study area for (more than) 12 years, that is between 2000 and 2012. Essentially, the researchers examined how time-varying enrolment in schools affect teenage pregnancy considering other variables such as age, size and socio-economic status (SES) of households, calendar year, and gender, employment status and educational attainment of the household head. Drawing data from the National Longitudinal Study of Adolescent Health, the second article sought to analyse the effects of teenage pregnancy and childbearing on

education. The researchers utilised four different statistical strategies to analyse their data, that is: a) ordinary least squares; b) propensity score matching; c) parametric maximum likelihood estimation and; d) semi-parametric maximum likelihood estimation.

This paper starts with a brief description of the area of interest and the area of focus of the two articles under review, and the context of the paper in terms of motivation and professional background. This is followed by a summary of the two articles, covering literature review, methodology and results, and then delving into the critique section which explores the nature and how participant dropout and missing data are dealt with in the two articles. Lastly, the implication to policy, research and professional practice is examined and a reflection of the review process is covered.

If well-designed, [17] claim that longitudinal studies can yield useful information and real-world data that can be used to explain the relationship between outcomes and their determinants at different macro, meso- and micro-levels. However, attrition and missing data are familiar methodological problems in longitudinal studies possibly due to the protracted nature of the data collection process. These problems have been examined by researchers and scientists and have been confirmed to compromise the external and internal validity of the research ([31,24,13,35,5,16]). According to [18], a longitudinal study not only tends to have a dropout problem, but it also offers possibilities for estimating the effects of the dropout and even for correcting for it. For instance, if a dropout occurs in a later wave of a longitudinal study, Reference [18] argue that, information available for a majority of the dropout from earlier waves can be used to predict how data would have looked like if it were complete. According to them, the effects of dropout on longitudinal study results ought to be judged separately for each specific study with reference to its characteristic features (p.14), and hence the focus of this essay in examining how attrition and subsequent missingness of data have been dealt with in two articles.

In this paper, attrition and missing data in each of the two articles was evaluated in the light of the patterns of missingness of data and how effective each pattern can be corrected for as discussed by [21]; assessing how effective the same has been. We adopted an objectivist view in explaining how and why attrition and missing data occur, how measurements have been carried out and statistical logic in compensating for missing data.

2. Materials and Methods

This article utilised purely document analysis method. Two journal articles exploring teenage pregnancy were selected based on their research design: longitudinal. These articles were analysed to examine how attrition and missing data were addressed by researchers. The analysis was also informed by review of other scholarly articles and books which helped the researchers gain deeper understanding of the effectiveness of methods used to address attrition and missing data in longitudinal studies in different contexts. Materials were searched through Google Scholar, various journal sites and from the University of Nairobi library. Key search words such as ‘longitudinal studies on teenage pregnancy’ were used. The analysis focused on the approaches used to deal with attrition and missing data in the two studies and their effectiveness. This was critically examined in the light of other findings from different literature sources.

3. Summary of the two papers

The first paper "Relationship between school dropout and teen pregnancy among rural South African young women" was authored by Rosenberg, Pettifor, Miller, Thirumurthy, Emch, Afolabi, Kahn, Collinson and Tollman in 2015 and published in the *International Journal of Epidemiology* (IJE) under volume 44, issue 3. Founded in 1972, the IJE is a peer-reviewed and a hybrid open access journal published on bimonthly basis by the Oxford University Press. As at 2015, IJE had an impact factor of 7.522.

The arguments made by [26] are that sexual activity is less likely to occur to adolescents when they are enrolled in school due to the structured and supervised nature of the school environment. Based on empirical studies, they argue that the knowledge/skills/attitudes acquired through education coupled with the peer networks developed in and through schooling provide a safe haven that help keep girls away from unplanned pregnancies.

In their study, the researchers sought to establish if school enrolment could be linked with teenage pregnancy. Using longitudinal demographic surveillance data from the rural Agincourt sub-district in South Africa, they reconstructed the cohort groups by focusing on the population (15,457) of teenage girls who had enrolled in schools between 2000 and 2011, and who were aged between 12 and18 and matched them to the estimated conception date for each pregnancy recorded among the group over the period. Using a Cox proportional hazard model, the researchers examined how time-varying enrolment of teenage girls in schools affected teen pregnancy; adjusting for seven variables, that is, the participant's age, size and socio-economic status of the households, calendar year, and gender, educational attainment and employment of the household heads. In the analysis, school enrolment status was identified as the exposure, and was coded as a time-varying, binary variable, with 1 (one) representing those who had enrolled in school, while 0 (zero) was used to represent those who had not enrolled in school. In 1997, 2002, 2006, 2009 and 2012, data collectors administered an education status module where they recorded the highest level of education attained by each teenage girl and also updated whether or not they were currently enrolled in school.

In their analyses, they found out that school-enrolment-status variable was associated with lower teen pregnancy rates, that is, more cases of teenage pregnancy were recorded during school vacation/holidays compared to school term period [incidence rate ratio (reported at 95% confidence interval)].

The second paper "The Educational Consequences of Teen Childbearing" was authored by Kane, Morgan, Harris and Guilkey in 2013 and published in journal *Demography* under volume 50, issue 6. Established in 1964, the *Demography* is a bimonthly, hybrid open access and peer reviewed academic journal that publishes articles drawn from different disciplines including, but not limited to, geography, statistics, epidemiology, and public health. It is the official journal of the Population Association of America. As at 2016, the journal had an impact factor of 2.802.

In their article, [14] review the debate around teen mothers and the challenges they face in life including truncated educational attainment. They question the relationship between teen motherhood and education attainment, with the cited literature estimating the effects of childbearing among teenage girls on their schooling

to vary widely from "no discernible difference to 2.6 fewer years among teen mothers". The researchers go ahead to conclude that the consequences of teenage pregnancy on education and its magnitude is unpredictable, "despite voluminous policy and prevention efforts that rest on the assumption of a negative and presumably causal effect" (pg.2129).

In a bid to examine the nature of relationship, [14] use the Add Health (National Longitudinal Study of Adolescent Health) data to carry out ordinary least squares regression, propensity score matching, and parametric and semi-parametric maximum likelihood estimation. The Add Health study, according to the researchers, involved a sample of 20,745 students (nationally representative) drawn across 7^{th} and 12^{th} grades in 1994–1995. The study was school-based and involved re-interviewing participants in three subsequent waves, that is, in 1996 (Wave II), 2001–2002 (Wave III), and 2008–2009 (Wave IV).

Based on the Add Health data, [14] focused on a sample of 8,352 female- only participants who had taken part in Waves I and IV of the study. Their choice for only-women participants was supported by past empirical studies that showed that young women bear the weight of pregnancy more than their male counterparts. In order to produce nationally representative estimates, the researchers constrained their sample to those with a valid sampling weight (n = 7,870). Estimating the missing data on analytic variables to be less than 4 per cent, they used the Stata software to carry out a single imputation procedure by replacing the missing data on all the independent variables and verified the same through list-wise deletion which produced substantively and statistically similar results.

Given the broad-ranging causal estimates of teen childbearing on educational attainment, the researchers used four statistical strategies to estimate these effects in their targeted cohort group of young adults, that is, ordinary least squares (OLS), propensity score matching, parametric maximum likelihood estimation, and semi-parametric maximum likelihood estimation. Available literature has reported effect sizes ranging from no difference in years of education, to 2.6 fewer years among teen mothers, while the estimates of the present analysis varied greatly, from 0.7 to 1.9 fewer years of schooling. Results from the Ordinary Least Squares (OLS) analysis showed that, holding all other factors constant, teenage mothers had completed nearly 2 fewer years of schooling than non-teen mothers; this difference was reduced to less than one year when the analyses adjusted for socio-demographic covariates.

In their study, [14] established, through the Propensity Score Matching (PSM), that young women who were living with their two parents exhibited lower odds for teenage pregnancy compared to those who were living in any other family type. Similar results were recorded among teenage mothers whose parents had attained higher education qualifications.

In addition, parametric maximum likelihood results demonstrated that there were more severe consequences of teenage pregnancy on education; with the baseline difference in educational attainment for teen and non-teen mothers being nearly three years, and the difference reducing to below two years with socio-demographic covariates added. Notably, the semi-parametric maximum likelihood strategy illustrated an educational penalty of less than nine months; an estimate that was similar to the OLS and PSM outcomes.

4. Findings and Discussion

4.1 Laying the foundation

Available literature shows that attrition, also referred to as monotone missingness, is the most frequently occurring form of nonresponse (missing data) in longitudinal studies, although other types of nonresponse may occur as well. It is therefore a primary concern for researchers dealing with longitudinal data and requires careful attention when testing longitudinal structural equation models [21]. Attrition occurs when an initially cooperative participant drops out of a study before it ends.

Other types of missingness of data include univariate missingness or item nonresponse which might occur due to poorly worded items or options, sensitive questions or participant burden. For individual item nonresponse in surveys, the pattern of missingness is often idiosyncratic and usually does not involve a large percentage of cases. In such instances, data may be MCAR or MAR or not large enough to seriously bias results. However, univariate missingness can add up to a large percentage of the sample quickly where there are many variables involved. Another pattern of missing data is unit missingness which involves a section or an entire case of a particular observation.

Data missing through attrition or any other form can be deemed to have been missing through three mechanisms: a) missing completely at random (MCAR) where the probability that a measurement is missing does not depend on its own score or on the score on any of the other variables. In this case, the non-respondents form a random subsample of the complete sample; b) missing at random (MAR) where the probability that a measurement is missing may depend on the value of any of the manifest variables but is unrelated to the scores on any of the latent variable; c) missing not at random (MNAR), no correction possible through weighting or imputation. Results tend to be biased.

A wide range of literature has proposed two approaches in dealing with attrition and missing data: minimising/eliminating nonresponse at the data collection stage and adjusting or correcting for it the data analysis stage. At the data analysis stage, a data analyst can ignore missing data which entails complete case analysis, available case method and list-wise deletion, but this technique is likely to induce bias if missing data is not MCAR. Also, an analyst can apply imputation for nonresponse, weighting for unit nonresponse or attrition and maximum likelihood estimation (all based on MAR assumptions). Further, in dealing with NMAR, one can apply joint modelling of nonresponse and substantial analysis but this applies, as [21] argues, only if additional assumptions are met e.g. Heckman selection models. Under MAR, [21] adds, “traditional methods tend to have biased estimates as well as less efficiency estimates when compared to the multiple imputation” (p.347).

Since the two articles reviewed in paper involved already collected data (secondary data), we focused on the second approach of compensating for missing data at the analysis phase.

4.2 Article 1

In their analyses, Reference [26] utilised data derived from the Health and Socio-demographic Surveillance

System (HDSS). As indicated earlier in this paper, data of young women aged between 12 and 18 years who lived in Agincourt Sub-district (a rural area of Bushbuckridge Municipality in Mpumalanga Province, South Africa) between 2002 and 2012 were used to construct a cohort.

In their analyses, the researchers had to deal with several sets of missing data, including 35 per cent of all SES observations which were missing in the baseline data, and 15 per cent of the covariate data observations which were missing for household size, household head employment and household head secondary education among the sample. Existing literature has shown varied vulnerability with some individuals (with unique characteristics) having high probability of dropping out of a study than others, and in such a case external validity is compromised. For instance, in Western society, lower SES students tend to be impermanent and hence tend to move out of the school district and also the study. Obviously, the extrapolation of findings to the lower SES groups is greatly affected by the exit of the participants drawn from that group. In my work experience as a researcher in the African context, participant dropout has been occurring among people without children, young adults, non-owners of homes, low educated, urban areas, lower socio-economic classes, and among old-age groups. With this dropout trend, [21] argues that the list-wise deletion becomes a risky analysis strategy although multiple imputations technique can be more effective in correcting for it.

According to [26] the unadjusted rate of pregnancy among teenage mothers who had enrolled in school was lower compared to those who had dropped out of school. In fact, based on the full sample, the probability of getting pregnant among teenage girls was nearly 50 per cent high for out-of-school individuals compared to those who were enrolled in school. In case, the participants who dropped out of the study substantially varied from those who remained, the outcome would be biased. However, as the researchers indicate, the “adjustment for confounding did not substantially affect the association in the complete case analysis or after correction with multiple imputation.” When the findings from the household fixed effects model were compared with the adjusted estimates from the full sample, it was established that they were not sensitive to uncontrolled household-level confounding variables. In other words, the outcome of the complete (full) sample would have been similar to the current outcome, had there been no dropout. The basic idea of multiple imputations, as in the argument above, is to reproduce values for the missing data over and over again by predicting how the data would look like in its complete form based on its distributions, and this is applicable to both MAR and NMAR missing data patterns, although [21] asserts that standard software routines primarily support MAR imputation schemes.

In the data set, the researchers also had to grapple with missing values in which measurements are available on some but not all variables. In particular, the missing values included 6 socio-economic status (SES) observations, 12 observations on household size, 4 observations on household head employment, 5 observations on household head education and the observations on single-gender-household head, as well as the pregnancy outcome. In this case, the researchers applied multiple imputation for observations with missing covariate information, imputing the missing values using the predictive ability of all observed time points of each covariate. This enabled the team to comparatively assess the distribution n of covariates before and after imputation; the results did not differ significantly between the two data sets and thus the conclusion that the multiple imputation was effective in correcting for the missing values.

According to [18], missing data at a certain occasion cause not only that occasion for that subject to be useless, but also the entire series for that subject in the whole study and, hence the need for appropriate statistical model in order to compensate for them. Presented with the missing data and values, analysts ought to make sound judgments and choices of the technique to use in order to compensate for them, otherwise inappropriate techniques will yield unreasonably huge standard measurement error; indicating ineffectiveness. Based on the outcome of their imputation process, it is agreeable that the researchers utilised an effective missing data technique since – after using multiple imputation- the procedure yielded qualitatively similar results in terms of the independent variable distribution compared to the results yielded by the non-imputed dataset.

Although the final sample included 15,457 out of the overall 22,661 young women covered in the baseline (wave 1), the demographic characteristics of 7,204 young women who were eliminated because their data on education was missing were actually similar to those who had complete data on education. Now that the characteristics of those who were not included in the final sample due to attrition and other forms of missing data were found not to vary significantly from those who remained, I can confidently (95% confidence level) argue that the remaining sample was representative.

In their data, Reference [26] had to deal with 512 young women who dropped out (by moving households) and other missing data which included 136 and 1008 young women who were excluded since they had graduated from high school before follow-up began or experienced a pregnancy before follow-up began, respectively, and also the 1462 and 4086 young women who were excluded from the data set due to logically inconsistent education data and missing education data, respectively. As a way of handling the above missing data, the researchers ignored them (excluded them from the final sample), and this could possibly be justified by the fact that they remained with a huge sample size (15,457) even after the exclusion. Although the final sample is proportionally big (68.2 per cent of the overall sample) the fact that it was derived from a census data could pose some limitations especially because it was not randomised.

Based on the existing literature, the 512 subjects who dropped out of the study after participating in the initial waves gave out their background information which could be used to reconstruct their responses based on the available data. In other words, there are already important information on them (collected at baseline or wave 1): not only as regards to background characteristics, but also with respect to the variables of interest. Thus, the researchers could compare the subjects who remained in the sample with those who dropped out and assess to what extent they differ from one another, thus making the consequences less serious. In fact, Reference [21] argue that if subjects drop out in the course of the study (after a number of occasions/waves), the consequences are not as serious as they are in the case in which subjects do not participate at all.

The outcome of this analysis is consistent with other previous studies conducted to examine the effect of missing data on the findings of longitudinal research, and the choice for missing data techniques. For instance, analyses by [35] illustrated the influence of replacing missing data using different imputation methods. Their illustration showed that when “MANOVA for repeated measurements was used, imputation methods were highly recommendable (because MANOVA as implemented in the software used, uses list-wise deletion of cases with a missing value) and when imputation methods were used in longitudinal data, the point estimates

and standard errors were closer to the estimates derived from the complete dataset." Based on their illustration, it can be concluded that if imputation methods are to be chosen, longitudinal imputation would be more effective since point estimation and standard errors are usually and generally similar to those derived from complete dataset. It is important to note that, although these techniques exist and have been tested and proven effective in correcting for attrition and missing data, the application of the techniques need a careful consideration accompanied with a thorough assessment of data missing mechanisms.

4.3 Article 2

In their article, Reference [14] analysed data collected by the Add Health (National Longitudinal Study of Adolescent Health). According to the researchers, the Add Health involved a school-based, nationally representative sample of 20,745 seventh through twelfth graders in 1994–1995. The study participants were re-interviewed in three subsequent waves, that is, in 1996 (Wave II), 2001–2002 (Wave III), and 2008–2009 (Wave IV). The researchers narrowed the sample down to 8,352 females- only participants who had taken part in Waves I and IV of the study, and further constrained it to a valid sampling weight (n = 7,870). With this constraint on sample, it is difficult to establish the criteria used to exclude the other units, especially because no much detail is provided about the process. It is important for the researchers to give detailed description of how the final sample was arrived at and how exclusion was done in order for the reader to make sound judgment regarding the process. Also, the study does not give details of the characteristics of the final sample (7,870) and explain how they differed from the initial sample (8,352).

Lack of details led us to make an assumption that the final sample excluded units/subjects with missing values in one or more of associated variables, although this is something that could have been dealt with by running tests using the various missing data techniques in order to establish the extent to which the two data sets differed and if the data were NMAR, MAR or MCAR. As [16] assert (based on their testing) that imputation techniques and weighting procedures may not adequate address biases caused by attrition particularly when data are not MAR or MCAR. However, assuming that data were missing completely at random, every participant in the study would have had an equal probability of dropping out, and hence the only problem would have been 'the reduced sample size' which would subsequently lead to reduced or loss of associated statistical power. But unfortunately, Reference [21] believe that a substantial amount of missing data is non-random. According to [33] cumulative nonresponse can greatly reduce the size of the final sample. If the chance that a person will participate in any particular wave of a four-wave study is 0.9, only (0.9^4) 66% of the respondents of the first wave will have remained in the study after four waves, and hence the use of the 66% as the valid/final sample.

This reveals one important lesson that researchers should not be overly optimistic about the response rates in their own research, and they should aim at reasonably large sample size for the first wave of their study although, scholars (for instance, Reference [13,3]) argue that small sample is not the most serious threat, rather the aspect of responders differing systematically from non-responders.

After estimating the missing data on analytic variables to be less than 4 per cent, Reference [14] applied a single imputation technique to correct for the missing data on all the covariates using Stata software and verified the

same through list-wise deletion which produced substantively and statistically similar results. Although the researchers do not expound how they estimated the missing data on analytic variables, the use of single imputation technique and subsequent yield of similar results (between the sample that remained and the complete one) indicates that the technique was effective in correcting for the missing data. This outcome is consistent with an analysis carried out by [9] who found out that "utilizing weights in a complete cases analysis may provide a means of correcting potential biases, but where such weights, as in the case of attrition, are derived from observed data at wave 1 (or later) we show that greater simplicity and greater statistical efficiency can be obtained by an imputation based approach that incorporates all the observed data in a single model" (p.62).

The current and past critiques of dealing with attrition and missing data in longitudinal studies have raised their main concern from a statistical viewpoint; if biased, they lead to the effective samples no longer being representative of the population. It is therefore important for researchers and authors to give more details of how they carried out the sampling and excluded subjects in the final sample in order to enable the reader to make sound judgment.

5. Implications

Although attrition and missing data has been confirmed to affect the external and internal validity of longitudinal studies, studies have been conducted to examine the extent of their impacts and how the same can be addressed (for instance, [23,30,11,34]). While it is greatly advocated for researchers to plan well so as to eliminate missing data, there are extenuating circumstances where missing data become inevitable, for instance due to poor health, death or in case the participant(s) cannot be traced.

According to literature, data – in most cases – are not missing completely at random and not even missing at random and therefore, the most effective way of dealing with missing data is to eliminate them as much as possible during the data collection stage. Procedural strategies (carried out before and during data collection) and statistical steps (at data analysis stage) can be used in combination to help the researcher deal with attrition and missing data. Statistical steps refer to the recent revolution in statistical procedures for handling missing data that has had a big impact on longitudinal research by making it possible to utilise all available data and to eliminate any bias associated with non-random participant drop-out. Using missing data techniques, a data analyst can estimate what the data would have been had there been no dropout, based on the available data.

In the application of statistical procedures, it is important to first make distinctions of the missing data mechanisms since the missing data mechanism has a bearing on whether certain strategies of handling missing data will result in biases or not.

Strategies for handling missing data may include traditional methods such as list-wise deletion, pairwise deletion, or mean imputation or modern methods such as multiple imputation. This however depends whether data are MCAR, MAR or NMAR. Efficient methods of compensating for missing data should not induce bias and should not result in a huge standard measurement error.

6. Conclusion

There is dearth of literature indicating that there is hardly no longitudinal study that is free from attrition and missing data, and if not accounted for, it can greatly compromise the external and internal validity of a study ([1,2,4,37,32,10]). With the traditional and modern techniques available for accounting for missing data, researchers can estimate the variation of the available data from the way it would have been had it been complete. While the argument is based on the application of statistical techniques of reconstructing the data sets, scientists and researchers still believe that the most effective way is to ensure that there are adequate data assurance measures put in place to guarantee quality data at the data collection stage. However, if attrition and subsequently missing data do occur, it is appropriate to employ effective methods of correcting for. As this paper has shown, it is important to first establish the manner in which the attrition and missing data occurred in order to select a more effective method of compensation.

References

[1] S. Arndt. “Data quality in longitudinal research.” American Journal of Psychiatry, vol. 148, issue 9, pp.1257-1267, 1991.

[2] L.R. Bergman. “Measurement and data quality in longitudinal research.” European Child & Adolescent Psychiatry, vol. 5 (suppl 1), pp. 28-32, 1996.

[3] K. Biering, N.H. Hjiollund and M. Frydenberg. “Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated measures of patient-reported outcomes.” Clinical Epidemiology, vol. 7, pp. 91-106.

[4] C.C. Bijleveld, L.J.Kamp, A. Mooijaart, W.A. Kloot, R. Leeden and E. Burg. Longitudinal data analysis: designs, models and methods. London: SAGE Publications Ltd, 1998.

[5] A. Boys, J. Marsden, G. Stillwell, K. Hatchings, P. Griffiths and M. Farrell. “Minimizing respondent attrition in longitudinal research: practical implications from a cohort study of adolescent drinking.” Journal of Adolescence, vol. 26, issue 3, pp. 363-73, 2003.

[6] F.D. Callan and E.M. “Parenting constraints and supports of young low-income mothers in rural United States.” Journal of Comparative Family Studies, vol. 44, issue 2, pp. 157-174, 2013.

[7] J.J. Card and L.L. Wise. “Teenage mothers and teenage fathers: the impact of early childbearing on the parents’ personal and professional lives.” Family Planning Perspectives, vol. 10, issue 4, pp. 199-205, 1978.

[8] E.J. Caruana, M. Roman, J. Hernandez-Sanchez and P. Solli. “Longitudinal studies.” Journal of Thoracic Disease, vol. 7, issue 11, pp. 537-540, 2015.

[9] J. Cumming and H. Goldstein. “Handling attrition and non-response in longitudinal data with an application to a study of Australian youth.” Longitudinal and Life Course Studies, vol. 7, issue 1, pp. 53-63, 2016.

[10] Y. Deng, D.S. Hillygus, J.P Reiter, Y. Si and S. Zheng. “Handling attrition in longitudinal studies: the case for refreshment samples.” Statistical Science, vol. 28, issue 2, pp. 238-256, 2013.

[11] L. Fumagalli, H. Laurie and P. Lynn. “Experiments with methods to reduce attrition in longitudinal surveys.” Journal of the Royal Statistical Society, vol. 176, issue 2, pp. 499-519, 2012.

[12] R.D. Gibbons. “Design and Analysis of Longitudinal Studies.” Psychiatric Annals, vol. 38, issue 12, pp. 758-761, 2008.

[13] K. Gustavson, T. Soest, E. Karevold and E. Roysamb. “Attrition and generalizability in longitudinal studies: findings from a 15-year population-based study and a Monte Carlo simulation study.” BMC Public Health, Vol. 12, pp. 918, 2012.

[14] J.B. Kane, S.P. Morgan, K.M. Harris and D.K. Guilkey. “The Educational Consequences of Teen Childbearing.” Demography, vol. 50, issue 6, pp. 2129-2150, 2013.

[15] D. Kneale, A. Fletcher, R. Wiggins and C. Bonell. “Distribution and determinants of risk of teenage motherhood in three British longitudinal studies: implications for targeted prevention interventions.” Journal of Epidemiology and Community Health, vol. 67, issue 1, pp. 48-55, 2013.

[16] V.L. Kristman, M. Manno and P. Côté. “Methods to Account for Attrition in Longitudinal Data: Do They Work? A Simulation Study.” European Journal of Epidemiology, vol. 20, issue 8, pp. 657-662, 2005.

[17] K.M. MacDonald, S. Vancayzeele, A. Deblander and I.L. Abraham. “Longitudinal observational studies to study the efficacy-effectiveness gap in drug therapy: Application to mild and moderate dementia.” Nursing Clinics of North America, vol. 41, issue 1, pp. 105-117, 2006.

[18] D. Magnusson and L.R. Bergman. (Eds.) Data Quality in Longitudinal Research. New York: Cambridge University Press, 1990.

[19] S. Mollborn. “Exploring variation in teenage mothers’ and fathers’ educational attainment.” Perspectives on Sexual and Reproductive Health, vol. 42, issue 3, pp. 152-159, 2010.

[20] S. Mollburn and E. Morningstar. “Investigating the relationship between teenage childbearing and psychological distress using longitudinal evidence. Journal of Health and Social Behaviour, vol. 50, issue 3, pp. 310-326, 2009.

[21] J.T. Newsom. Longitudinal Structural Equation Modelling: A Comprehensive Introduction. New York:

Routledge, 2015.

[22] R.E. Ployhart and R.J. Vandenberg. "Longitudinal research: The theory, design and analysis of change." Journal of Management, vol. 36, issue 1, pp. 94-120, 2010.

[23] F. Rajulton. "The fundamentals of longitudinal research: an overview." Special Issue on Longitudinal Methodology, Canadian Studies in Population, vol. 28, issue 2, pp.169-185, 2001.

[24] H.T. Reis and C.M. Judd. (Eds). Handbook of Research Methods in Social and Personality Psychology. Cambridge: Cambridge University Press, 2000.

[25] L.M. Rich and S.B. Kim. "Patterns of later life education among teenage mothers." Gender and Society, vol.13, issue 6, pp. 798-817, 1999.

[26] M. Rosenberg, A. Pettifor, W.C. Miller, H. Thirumurthy, M. Emch, S.A. Afolabi, K. Kahn, M. Collinson and S. Tollman. "Relationship between school dropout and teen pregnancy among rural South African young women." International Journal of Epidemiology, vol. 44, issue 3, pp. 928 - 936, 2015.

[27] C. Rosengerd, M.G. Phipps, N.E. Adler and J.M. Ellen. "Adolescent pregnancy intentions and pregnancy outcomes: A longitudinal examination." Journal of Adolescent Health, vol. 35, issue 6, pp. 453-461, 2004.

[28] T. Shefer, D. Bhana and R. Morrell. "Teenage pregnancy and parenting at school in contemporary South African contexts: Deconstructing school narratives and understanding policy implementation." Perspectives in Education, vol. 31, issue 1, pp. 1-10, 2013.

[29] M.A. Silles. "The effect of schooling on teenage childbearing: evidence using changes in compulsory education laws." Journal of Population Economics, vol. 24, issue 2, pp. 761-777, 2011.

[30] J.D. Singer and J.B. Willett. Applied longitudinal data analysis: modelling change and event occurrence. New York: Oxford University Press, 2003.

[31] R. Stratford, J. Mulligan, B. Downie and L. Voss. "Threats to validity in the longitudinal study of psychological effects: the case of short stature." Child: Care, Health and Development, vol. 25, issue 6, pp. 401-409, 1999.

[32] C.T. Street and K.W. Ward. "Improving validity and reliability in longitudinal case study timelines." European Journal of Information Systems, vol. 21, issue 2, pp. 160-175, 2012.

[33] T.W. Tarris. A Primer in Longitudinal Data Analysis. London: SAGE Publications, 2002.

[34] T.W. Tarris and M.A.J. Kompier. "Cause and effect: Optimising the design of longitudinal studies in

occupational health psychology." Work & Stress, vol. 28, issue 1, pp. 1-8, 2014.

[35] J. Twisk and W. de Vente. "Attrition in longitudinal studies. How to deal with missing data." Journal of Clinical Epidemiology, vol. 55, issue 4, pp. 329-337, 2002.

[36] L. Warrick, J.B. Christianson, J. Walruff and P.C. Cook. "Educational outcomes in teenage pregnancy and parenting programs: results from a demonstration." Family Planning Perspectives, vol. 25, issue 4, pp. 148-55, 1993.

[37] C.W. Whitney, B.K. Lind and P.W. Wahl. "Quality assurance and quality control in longitudinal studies." Epidemiologic Reviews, vol. 20, issue 1, pp. 71-80, 1998.