Special issue: Planned missingness

ijbd

International Journal of Behavioral Development
2014, Vol. 38(5) 453–460

DOI: 10.1177/0165025414542713
ijbd.sagepub.com
SAGE

Reducing bias due to systematic attrition in longitudinal studies: The benefits of multiple imputation

Jens B. Asendorpf,[1] Rens van de Schoot,[2,3] Jaap J. A. Denissen,[4] and Roos Hutteman[3]

Abstract

Most longitudinal studies are plagued by drop-out related to variables at earlier assessments (systematic attrition). Although systematic attrition is often analysed in longitudinal studies, surprisingly few researchers attempt to reduce biases due to systematic attrition, even though this is possible and nowadays technically easy. This is particularly true for studies of stability and the long-term prediction of developmental outcomes. We provide guidelines how to reduce biases in such cases particularly with multiple imputation. Following these guidelines does not require advanced statistical knowledge or special software. We illustrate these guidelines and the importance of reducing biases due to selective attrition with a 25-year longitudinal study on the long-term prediction of aggressiveness and delinquency.

Keywords

attrition, longitudinal study, multiple imputation

*This article accepted during Marcel van Aken's term as Editor-in-Chief.

Most multi-wave longitudinal studies are plagued by an increasing drop-out of participants (or knowledgeable judges of them, particularly parents) that is systematically related to variables at earlier assessments (systematic attrition). Systematic attrition causes biases in all results that are influenced by these variables (see, e.g., Baltes, Reese, & Nesselroade, 1988; Little & Rubin, 2002). For example, if one studies the long-term consequences of early aggressiveness and the most aggressive kids drop-out from the study whereas most other kids remain in the study, attrition will restrict the range of aggressiveness and the results of the study will underestimate all long-term effects of aggressiveness. The current article provides guidelines how to reduce biases in such cases, particularly with multiple imputation.

Because of the tendency of participants not to return to the study again once they missed a wave of assessment (Baltes et al., 1988), even small and non-significant selective drop-out effects from wave to wave can accumulate over the course of a multi-wave study such that the results become increasingly biased. In addition, the drop-out tendency can change over the course of a longitudinal study. For example, aggressive adolescents may be unwilling to continue participation in a study into which they were initially placed by their parents. In addition, initially cooperative parents may drop out at later ages of their kids even if these kids remain in the study. For example, the parents may become less personally involved in the study, develop a bad relationship with their kids, or the kids do not agree with being judged by their parents at later ages. Such effects are rarely studied in detail but can contribute to selective attrition and biased predictions.

Although systematic attrition is often analyzed in longitudinal studies, surprisingly few researchers attempt to reduce bias due to systematic attrition. For example, out of the 35 articles reporting longitudinal findings in the *International Journal of Behavioral Development* in 2012 and 2013, 20% did not report on attrition and other kinds of missing data at all.[1] Additionally, 26% reported inappropriate methods such as analyzing only data from participants who continued to participate until the end of the study. Only 51% reported reasonable methods of dealing with attrition although sometimes even these attempts were non-optimal choices (e.g., single imputation when multiple imputation was possible; see later section on imputation).There are no technical reasons for this neglect because many standard statistical software such as *R*, *SPSS,* or *SAS* offers efficient ways of dealing with missing data nowadays. Thus, it seems that not only most authors but also most reviewers and editors of major journals are not sufficiently aware of the importance and possibilities of dealing with systematic attrition.

Systematic attrition is routinely corrected if the longitudinal analyses rely on more advanced statistical techniques such as growth curve analysis through multi-level or structural equation modeling. The standard *Empirical Bayes* (EB) estimations of the individual growth parameters used in multi-level modeling and the standard *Full Information Maximum Likelihood* (FIML) estimations in structural equation modeling correct for selective attrition. But predictions based on less advanced methodology such as correlations (including stability), multiple regressions, or extreme group

[1] Humboldt University Berlin, Germany
[2] North-West University, South Africa
[3] Utrecht University, The Netherlands
[4] Tilburg University, The Netherlands

Corresponding author:
Jens B. Asendorpf, Humboldt University Berlin, Unter den Linden 6, 10099 Berlin, Germany.
Email: asendorpf@gmail.com

analyses require particular methods of controlling selective attrition. However, these methods are still rarely used.

The main reason for this neglect seems to be the misconception of many developmental researchers that controlling for selective attrition in longitudinal data by estimating missing values is akin to "making up data." Even researchers who impute missing values within a wave of assessment are often afraid of imputing missing data at later waves from earlier waves because it seems to be cheating. But this is not the case; instead, it is statistically more accurate to impute with adequate methods both within *and* across waves than not to impute (see e.g. Graham, 2009). The present contribution is meant to encourage longitudinal researchers to go ahead and control for missing values also across waves.

Another reason for the neglect of controlling for selective attrition may be communication problems between methodologists concerned with estimating missing values and longitudinal researchers. Multiple imputation was developed to deal with the more general case of any missing data, not only for the special case of selective attrition. Three cases of missing data can be distinguished by the assumed missing-generating mechanism (see e.g. Graham, 2009). *Missing Completely at Random* (MCAR) means that all missing data occurred independently of all observed and non-observed variables. Systematic attrition violates this assumption because the missing data are related to earlier observed variables. *Missing at Random* (MAR) means that the missing data may depend on observed variables but do not depend on unobserved variables. MAR is a reasonable assumption in longitudinal studies that include a broad range of variables that may be related to attrition and implies that biases due to systematic attrition can be corrected by replacing ("imputing") the missing scores with scores estimated from the observed variables. Continuing the example of selective drop-outs of aggressive kids, MAR is the assumption that drop-out depends only on aggressiveness and other variables included in the longitudinal analyses but not on unobserved variables.[2] Finally, *Missing not at Random* (MNAR) means that attrition is related to unobserved variables. MNAR implies that imputing missing scores may not sufficiently correct biases due to selective attrition.

The term "missing at random" may appear confusing because it describes a case where missing values do *not* occur at random because they depend on observed variables and therefore can be estimated by them. Instead, "random" refers here to the assumption that the missing scores are random once their dependence on the observed variables is controlled.

Although bias due to systematic attrition can be effectively controlled if the missing data can be assumed MAR, researchers may be reluctant to estimate missing data because much of the literature on missing values is rather technical, requires advanced statistical knowledge, and lacks good examples of how important proper reduction of bias due to selective attrition can be (for more easily accessible overviews, see the chapter by Graham, 2009, and the books by Enders, 2010, and van Buren, 2012).

The present article aims at filling this gap by offering non-technical guidelines tailored to selective attrition that are illustrated with a striking example of how important taking account of selective attrition can be when it comes to long-term predictions of aggressiveness. First we present this example and show how biases due to selective attrition can be reduced with multiple imputation. Hopefully we facilitate understanding by presenting details of the multiple imputation methods not in general terms but always with reference to the empirical example. Subsequently, we provide guidelines for bias reduction in more general terms and briefly describe how they can be implemented using *R*, *SPSS,* or *SAS*.

Method

Participants

Participants were part of the sample of the Munich Longitudinal Study on the Genesis of Individual Competencies (LOGIC) which consists of 230 children born in 1980–1981 who started to attended preschools in the Greater Munich area at age 4. The sample was fairly unbiased because the schools were selected from a broad spectrum of neighborhoods and more than 90% of the parents who were asked for permission gave their consent for their child's participation. The sample for the present study consists of the 206 participants who were judged by their teachers in at least two of three yearly Q-sort assessments at ages 4–6 years (see Asendorpf, Denissen, & van Aken, 2008, for more details).

Assessments and measures

The present analyses refer to teacher-reported aggressiveness at ages 4, 5, and 6 years; parent-reported aggressiveness at ages 12, 17, 23, and 29 years; and self-reported delinquency at ages 23 and 29 years.

Teacher-reported aggressiveness. Based on teacher-provided descriptions of a prototypical aggressive preschooler on the 54 items of a German short version of the California Child Q-Set (CCQ; Göttert & Asendorpf, 1989), an 8-item aggressiveness scale was extracted from the CCQ (see Asendorpf et al., 2008, for details). At the end of each school year, the child's main teacher sorted the 54 CCQ-items in terms of their salience for the child ranging from extremely uncharacteristic (1) to extremely characteristic (9); the child's mean salience of the 8 aggressiveness items (e.g., "is aggressive") served as the yearly measure of the child's aggressiveness. It showed a satisfactory internal consistency in all 3 assessments, $\alpha > .78$, and a 2-year stability of $r = .56$. The aggressiveness scores were averaged over the three assessments, allowing for one missing score for each child.[3]

Parent-reported aggressiveness. The main caregiver (nearly always the mother) answered a questionnaire about their child's personality including 4 aggressiveness items at ages 12 and 29 which were rated on 7-point Likert scales. At ages 17 and 23, both parents were asked (if available). At ages 12 and 17, the items referred to aggressiveness with peers (e.g., "is aggressive to peers") that were assessed using paper-and-pencil questionnaires. At ages 23 and 29, the items referred to aggressiveness in general (e.g., "is aggressive"); see Hutteman, Denissen, Asendorpf, and van Aken (2009) for details. The internal consistency of these scales was satisfactory for all ages ($\alpha > .82$); if both parents provided scores, they were averaged.[4] Detailed data on missing observations are reported later in the results section.

Self-reported delinquency. At age 23, participants were asked to complete a Life History Calendar (Caspi et al., 1996) asking for various important events since the 18th birthday using a month-by-month horizontal timeline. The frequency of reported criminal charges for delinquency per year served as the measure of delinquency. At age 29, the participants were asked to report, among

Table 1. Descriptive statistics and intercorrelations with/out multiple imputation.

Variable		*N*	*M*	*SD*	AG4–6	AG12	AG17	AG23	AG29	DEL23	DEL29
Aggressiveness ages 4–6	AG4–6	206	3.91	1.36		.39***	.29***	.21*	.14	.19*	.32***
Aggressiveness age 12	AG12	151	2.52	0.90	.40***		.44***	.27**	.24*	.12	.17
Aggressiveness age 17	AG17	142	2.12	0.68	.31***	.47***		.49***	.42***	.21*	.07
Aggressiveness age 23	AG23	114	2.83	0.83	.24**	.23*	.48***		.56***	.09	.19
Aggressiveness age 29	AG29	92	2.18	0.84	.31**	.37***	.52***	.58***		−.13	.15
Delinquency age 23	DEL23	143	0.04	0.13	.27***	.13	.18	.14	.08		.19*
Delinquency age 29	DEL29	141	0.04	0.13	.27***	.18	.15	.24*	.15	.77***	

Note. Correlations above diagonal uncorrected, below diagonal after multiple imputation (see section on multiple imputation).
* $p < .05$; ** $p < .01$; *** $p < .001$.

Table 2. Systematic attrition for parent-reported aggressiveness and self-reported delinquency.

	Drop-out							
	yes			no		Difference		
	Teacher-reported aggressiveness at ages 4–6							
Age	*n*	*M*	*SD*	*M*	*SD*	*t*(204)	*p*	*d*
Attrition for parent-reported aggressiveness								
12	55	4.15	1.52	3.82	1.29	1.56	.120	0.25
17	64	4.04	1.55	3.85	1.26	0.94	.351	0.14
23	92	4.17	1.39	3.69	1.29	2.53	.012	0.36
29	114	4.16	1.47	3.59	1.13	3.08	.002	0.43
Attrition for self-reported delinquency								
23	63	4.27	1.44	3.75	1.29	2.56	.011	0.39
29	65	4.16	1.50	3.79	1.27	1.87	.063	0.28

Note. $N = 206$. *n* refers to number of drop-outs.

other questions, the frequency of self-reported criminal charges for delinquency since age 23; this frequency was subsequently converted to frequency per year. Detailed data on missing observations are reported later in the results section.

Results

Uncorrected data

The descriptive statistics and the uncorrected intercorrelations (using pairwise deletion) of the aggressiveness and delinquency variables are presented in Table 1. Note that aggressiveness was measured differently at ages 4–6 years versus later such that the *M*s and *SD*s cannot be compared between these assessments. Delinquency showed only a low yearly mean frequency, which was virtually identical for the age intervals 18–23 years and 23–29 years. The stabilities of aggressiveness showed a simplex pattern (monotone decreases within rows and monotone increases within columns), which is expected if instability and drop-out are due to causal mechanisms that are constant across age such that their effects add up over time (Marsh, 1993). The stabilities decreased to a non-significant level of $r = .14$ for the 25-year stability. Together, this stability pattern seems to suggest that the stability of aggressiveness can be described by a transactional model where initial inter-individual differences in aggressiveness fade away over the course of their transaction with environmental or genetic effects, but such a conclusion would be premature because the stability pattern can be influenced also by additional mechanisms including selective drop-out.

One indication for such additional mechanisms is that the predictions of self-reported delinquency from teacher- and parent-reported aggressiveness are not consistent with a simplex model if one assumes that all variables refer to the same underlying construct of externalizing problems (Tremblay, 2000). Early aggressiveness tended to predict delinquency at 29 ($r = .32$) better than at age 23 ($r = .19$). The opposite was true for predicting delinquency at age 29 and 23 from aggressiveness at age 17 (*r*s .07 versus .21).

Although the stability of aggressiveness between the measurement waves increased monotonically from $r = .39$ to $r = .56$, the stability of self-reported delinquency was relatively low ($r = .19$). This inconsistency could be interpreted post hoc by method variance (other- versus self-perception) and the skewed distribution of delinquency (in both assessments, 88% of the participants reported zero delinquency), but caution should be exercised because of the large number of missing values in the later assessments of aggressiveness. Therefore, a closer inspection of the missing value pattern is in order.

Missing pattern

Missing mechanism. If the missing data are MCAR, systematic attrition does not occur. Little's MCAR test (Little & Rubin, 2002) provides an overall statistical test whether MCAR is violated.

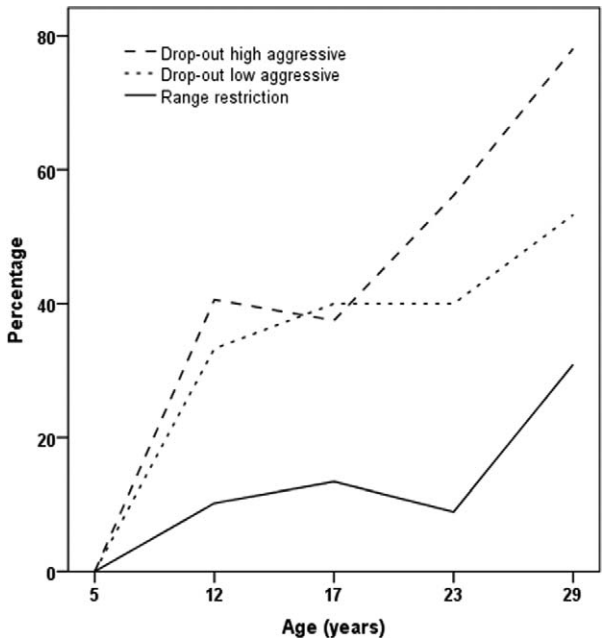

Figure 1. Age-related changes in drop-out rates for the top and bottom 15% of initial aggressiveness and in range restriction of aggressiveness (% reduction of the initial variance for the remaining participants)

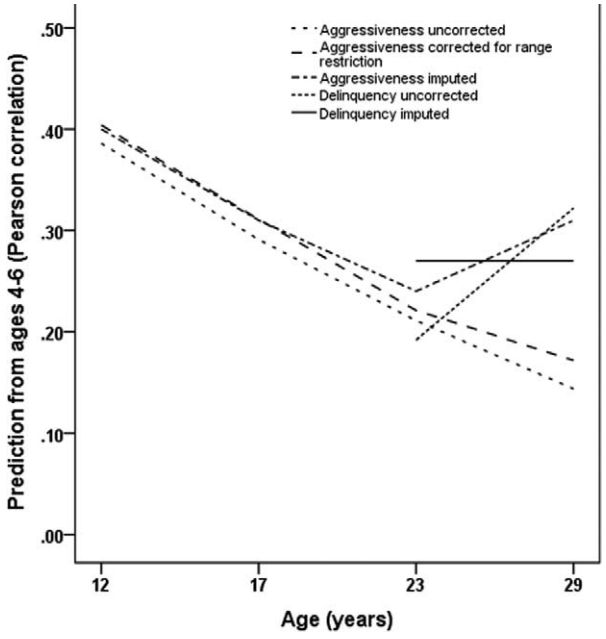

Figure 2. Correlation of teacher-reported aggressiveness at ages 4-6 with later measures of parent-reported aggressiveness and self-reported delinquency with/out controlling for range restriction and with/out multiple imputation.

In the present study, Little's test showed a significant violation of MCAR, $\chi^2(137) = 188.27$, $p = .002$. Table 2 indicates that drop-outs had higher initial aggressiveness scores than the remaining participants at all assessments, and this systematic attrition was at least marginally significant ($p < .10$) for all assessments after age 17. The direction of the effect (high-aggressive rather than low-aggressive kids dropped out) is unsurprising because of the common observation that participants with socially undesirable characteristics drop out more often than those with desirable characteristics (e.g., Weinberger, Tublin, Ford, & Feldman, 1990).

Because of systematic attrition, the variance of the scores of the remaining sample decreased with increasing age, resulting in a reduction of the variance (*range restriction*, see Figure 1).[5] For the parental reports, more participants dropped out at the high end of the initial aggressiveness continuum than at its low end (top 15% versus bottom 15% of the distribution of initial aggressiveness) after age 17. At the last assessment at age 29, 78% of the high-aggressive group but only 53% of the low-aggressive group had missing parental reports (see Figure 1).

This pattern suggests that the parental reports after age 17 potentially underestimate the long-term effects of early aggressiveness. For self-rated delinquency, a reverse bias tended to occur at age 23 (56% high-aggressive vs. 70% low-aggressive drop-out) and no bias at age 29 (56% high-aggressive vs. 60% low-aggressive drop-out). Thus, not the initially highly-aggressive kids but the parents of these kids showed a particularly high drop-out rate after adolescence that may be attributed to frustration of the parents of highly-aggressive kids in the long run that led to a disengagement from the study, or to aggressive participants' unwillingness to be judged by their parents.

The bottom line is that biases were expected for both aggressiveness and delinquency outcomes due to violation of MCAR and that some of these biases might be reduced by predicting missing scores from observed variables.

Monotone missing pattern. Sometimes longitudinal studies show a *monotone missing pattern* where any drop-out is final (drop-outs do not participate at later assessments). It is important to check for this possibility because special imputation procedures are available in this case that use information on the monotone missing pattern. Although the *n*s for aggressiveness reported in Table 1 suggest a monotone pattern, the actual pattern was not monotone (some participants continued participation after a missing wave of data collection), and for the prediction of delinquency from early aggressiveness the pattern was clearly non-monotone (21 participants provided data on delinquency at age 29 but not at age 23).

Reducing bias due to selective attrition

Correction for range restriction. Ever since Pearson (1903) it has been recognized that selection often results in a reduced variance (*range restriction*) that can however be corrected for (e.g., Sackett & Yang, 2000). Figure 2 shows the uncorrected predictions of later aggressiveness from initial aggressiveness (stabilities of aggressiveness) and the stabilities corrected for range restriction.[6] Correction relatively uniformly increased the stabilities by less than .03 across all measurement points (see Figure 2). Thus, correction for range restriction only slightly changed the stability pattern although range restriction was large at the last assessment (see Figure 1).

That correction for range restriction was not effective in this case seems to be due to the fact that correction for range restriction is insensitive to asymmetric attrition effects for high versus low scorers. Corrected is only the effect of overall variance reduction, not specific drop-out for high versus low scorers. For example, if delinquency is driven mainly by the highly aggressive kids who dropped out more frequently than low-aggressive kids, correction

for range restriction will not fully correct for this selective attrition effect on delinquency. Instead, the correction is based on the assumption that the effects of high and low aggressiveness are equally strong. As we will show in the next section, imputation procedures are better suited in the case of asymmetric attrition effects.

Imputation of missing values. All imputation procedures replace missing values by estimated scores (the *imputed values*). Both continuous and categorical variables can be used as predictors or outcomes. A procedure sometimes still offered by standard statistical software is *replacing missing scores by the mean* of the observed variable where the missing scores occur (e.g., in the *SPSS 21* factor analysis procedure). If the data are MCAR, this procedure will bias all (co)variances and thus also all correlations due to a decreased inter-individual variance. If the data are not MCAR, an additional problem is that biases due to systematic dependencies of the missing scores on observed variables are not corrected because the means do not take these dependencies into account. Therefore replacing missing scores by the mean should be generally avoided (see e.g. Graham, 2009).

A better approach is estimating missing scores with *single imputation* based on a regression model or on *EM estimation procedures* that use all relationships between the observed variables in order to maximize the likelihood that the estimated values are correct (see Graham, 2009, for more details). Because each individual missing value is replaced by the best estimation of this value, these imputation methods take asymmetric attrition patterns into account (e.g., they will assign a higher aggressiveness value to participants with higher observed aggressiveness in other assessments than to participants with lower observed aggressiveness values if the aggressiveness assessments are correlated). The parameter estimates based on single imputation procedures are less biased compared to mean imputation. However, with single imputation, one pretends to be sure that the imputed values are correct although single imputation includes a random error associated with each imputed value, and this uncertainty is ignored. Neglecting this uncertainty biases the *SE*s around the above-mentioned parameters and therefore all confidence intervals and statistical tests (see Graham, 2009).

Therefore our recommendation is using *multiple imputation* where the error involved in single imputation is estimated too. This is accomplished by estimating the *SE* of each parameter (e.g., of a correlation) on the basis of the observed variation of this parameter across many different imputation runs. Therefore multiple imputation requires fast computing which is nowadays hardly a problem. Whereas only five imputations have been considered to be sufficient in the older literature (and continue to be the default in standard software such as *SPSS* or *SAS*), more recent simulation studies have led to a revision of this recommendation. Depending on the percentage of missing values, 40 imputations have been shown to be a safe default value (see Graham, Olchowski, & Gilreath, 2007).

Different algorithms are available for multiple imputation, particularly *Markov Chain Monte Carlo* (MCMC) algorithms. MCMC often results in "impossible" imputed values (e.g., negative percentages or scores above 5 on an original 1–5 point scale). This is a consequence of the underlying linear estimation model and should not be "corrected" by hand. For example, replacing such scores by the possible minimum or maximum score would result in biased imputations. If such "impossible" scores should be avoided, *Predictive Mean Matching* (PMM) is a viable option where each missing score is first estimated with linear regression and then replaced by an optimal choice of a non-missing score (see Enders, 2010; van Buren, 2012).

Multiple imputation results in multiple imputed data sets (one set for each imputation run) that should be used for subsequent statistical analyses. Each data set is analyzed separately; subsequently the results are pooled. One should *not* simply use the averaged data from the various runs. The reason is that in such averaged data the error component involved in estimation is lost. Standard software offers procedures that make full use of all imputations (see the Implementation section).

A major decision in using imputation methods is which variables should be included in the analysis. Consider the five aggressiveness variables in the present study (see Table 1). One could use only these "core variables" for imputing. Alternatively, one could add additional *auxiliary variables*. Such variables are correlated with a core variable with missing values *and* can predict some of these missing scores because there are participants with values in the auxiliary variable but not in the core variable (see Collins, Schafer, & Kam, 2001; Hardt, Herke, & Leonhart, 2012). Adding auxiliary variables improves estimation to the extent to which the percentage of missing data that can be predicted increases and the correlations with the core variables increase. It is sufficient if an auxiliary variable correlates with only one core variable but if the correlation is low, the effects of adding this auxiliary variable are often very small (Graham, 2009).

The downside of adding auxiliary variables is that they may add noise to the estimations such that the variance of the estimations between the different imputations increases and therefore the statistical power of the subsequent tests (see Hardt et al., 2012). This becomes a problem if the percentage of missing values is high *and* the number of auxiliary variables becomes large compared to the sample size. Hardt et al. (2012) recommended that the number of auxiliary variables should not exceed 1/3 of the number of participants without missing values. Thus, in a sample of $N = 100$ and 40% drop-out, not more than 20 auxiliary variables should be used.

Application in the present study

We applied the MULTIPLE IMPUTATION procedure of *SPSS 21* to the seven variables listed in Table 1. Because 15.2% of the imputed values using the *SPSS* default *EM* procedure were out of range, we used the PMM option of SPSS; also, we changed the default number of five imputations to 40. For the analysis of the stability of parent-reported aggressiveness, we included self-reported delinquency as auxiliary variables, and vice versa.[7] This inclusive strategy makes sense because delinquency showed some significant correlations with aggressiveness, and although the correlations were low, many missing values could be predicted particularly for the ages 23 and 29 (see Table 1).

As Table 1 shows, the stabilities of aggressiveness and the predictions of delinquency became stronger in most cases, particularly the 24-year stability from ages 4–6 to age 29 and the stability of delinquency, which increased from $r = .19$ to $r = .77$. Figure 2 shows that imputation only slightly increased the estimated stability of aggressiveness from ages 4–6 until age 23. A stronger increase was found for the stability from ages 4–6 until age 29, which was as high as the stability from ages 4–6 until age 17. This is consistent with the increase of the missing rates for the parental reports of aggressiveness (see Figure 1). The increase in the ability of childhood values to predict aggressiveness after age 23 is not an artifact of imputation because the

uncorrected prediction of delinquency from initial aggressiveness also showed a clear increase (see Figure 2).

It seems that the predictions to age 23 were affected by antisocial tendencies limited to adolescence and emerging adulthood whereas the preadolescent and the age 29 assessments reflected more stable, life-course-persistent antisocial tendencies (see Moffitt, 1993, for these two types of antisocial types). This is reflected in the correlates of age 23 and age 29 delinquency. At age 23, delinquency is predicted by both childhood ($r = .19$) and adolescent ($r = .21$) aggressiveness, reflecting the simultaneous externalizing tendencies of both types. At age 29, delinquency is predicted only by childhood aggressiveness ($r = .32$), reflecting the continued expression of the lifecourse-persistent type and the desistence of the adolescence-limited type.

This conclusion is robust concerning the specific method of imputation. The correlational pattern depicted in Table 1 and Figure 2 remained virtually the same if single or multiple imputation with a regression procedure (the SPSS defaults) were used. Dropping the auxiliary delinquency variables for the analyses of aggressiveness slightly but systematically weakened all predictions to the assessments at ages 23 and 29, and therefore, we retained these auxiliary variables. Because the distributions of delinquency were highly skewed, we reran all analyses with delinquency as a categorical variable (delinquent yes or no). The results remained highly similar.

The bottom line is that the major conclusion after imputation is different from the one for the uncorrected data. The stabilities of childhood antisocial tendencies approached a non-zero asymptote of .30 after age 12 and thus do not support a transactional model where long-term stabilities approach zero across extended retest intervals. Instead, they support a transactional model with a stabilizing constant where the stabilities approach a non-zero asymptote (see Fraley & Roberts, 2005). The conclusion based on the uncorrected data would have been mistaken because it rested on biases due to selective attrition.

Guidelines

After this demonstration of how bias due to selective attrition can be reduced, we now present more general guidelines how such a reduction can be achieved:

1. *Analyse the missing pattern* for the core variables of interest (e.g., predictions of outcomes based on earlier assessments). First check whether the data are missing completely at random (MCAR) using Little's test. If MCAR is violated, analyse selective attrition by comparing the drop-outs in each wave of interest with the continuing participants for each variable (see Table 2). Don't rely only on significance, report also effect sizes because in large samples significant effects can be very small. If MCAR is not violated, such a detailed analysis of selective attrition does not make sense and can be skipped. But even in this case, simply using all available data (*pairwise deletion*) results in somewhat reduced statistical power and inconsistent sample sizes (see e.g. Graham, 2009). Therefore, we suggest proceeding with steps 2–4 even if MCAR is not violated. In any case, *listwise deletion* where only participants without any missing data in the core variables are included is even worse because if a substantial percentage of values are missing, all results will be biased if systematic attrition occurred, and the power of statistical tests will be reduced (see e.g. Graham, 2009). Second, if MCAR is violated, check whether the data show a monotone missing pattern such that all missing scores are due to drop-outs that never returned to the study. For this case, special variants of imputation methods are available.
2. *Search for auxiliary variables.* Correlate all other variables in the data set with all core variables. If they correlate with a core variable and have non-missing scores for many cases that have missing scores in this core variable, include them as auxiliary variables. If the number of auxiliary variables exceeds 1/3 of the number of participants without missing values, exclude those with the lowest correlations with the core variables until a rate of 1:3 is reached.
3. *Impute missing values.* Use all core and auxiliary variables as predictors for imputation, and all core and auxiliary variables with missing values as targets for imputation. Impute missing data with multiple imputation based on 20–40 runs; the higher the percentage of missing scores, the more runs are appropriate. If the percentage of missing scores becomes too large (e.g., above 80%), drop the variable from analysis as even the best imputation runs the risk of producing biased results in such cases (see Graham, 2009). Use the standard regression procedure (in case of a monotone missing pattern the specific variant for monotone patterns). If you want to preserve the observed range of scores for the imputed scores, use *Predictive Mean Matching* (PMM).
4. *Use the imputed data sets* for all following statistical analyses. Run the desired analyses (e.g., correlations) separately for each imputed data set and then pool the results of these separate runs using the method offered by your statistical software (see the Implementation section).

If the analyses are based only on structural equation models, steps 3 and 4 can be skipped because the standard estimation procedure *Full Information Maximum Likelihood* (FIML) controls for selective attrition.[8] Similarly, if the longitudinal analyses are based on multi-level models where individual slopes are estimated at Level 1, steps 3 and 4 can be skipped because the standard *Empirical Bayes* (EB) estimation procedure controls for selective attrition (see, e.g. Hox, 2010). However, missing data at Level 2 (constant individual parameters) is not estimated by standard multi-level procedures. Such imputation is not easy because the nested data structure has to be respected (see e.g. van Buren, 2012, Chapter 20).

Implementation

In most instances it is ok using the default values for multiple imputation provided by the statistical software except for the number of imputation runs (should be increased to at least 20).

R *packages*

The freely available *R* software (www.r-project.org) provides many packages for imputation, particularly *mi* and *mice*. Both packages provide many options for describing and plotting missing patterns, and multiple imputation procedures for interval, ordinal, or nominal data based on different models (e.g., linear models, logistic regressions, proportional odds models). PMM is provided by the package *mice*. In both cases, the data are first imputed with multiple runs, then subsequent statistical analyses are performed separately for each run and the output data are stored, and in a last step the results

of these outputs are pooled. This approach takes advantage of the fact that the output of any *R* procedure is stored in an output file that can be read by subsequent procedures.

SPSS *procedures*

The MVA procedure is used for describing missing patterns and testing for MCAR if the EM option is used. The MULTIPLE IMPUTATION procedure is used for multiple imputation. Default is linear regression using MCMC assuming a multivariate normal distribution, *PMM* can be required by option. *SPSS* automatically detects by default whether the missing pattern is monotone and adjusts the procedures accordingly. The output contains the original data along with all imputations; they are identified by the new variable Imputation_ (Imputation_ = 0 identifies the original file). Subsequent analyses are informed that imputed data are used by defining Imputation_ as a split file variable, using the command SPLIT FILE LAYERED BY Imputation_. *SPSS* presents the results first for the original data, then for each imputation run, and last the final results for the pooled imputations.

SAS *procedures*

Both the description of missing patterns and multiple imputation is performed using the procedure PROC MI; the default is linear regression with MCMC assuming a multivariate normal distribution; PMM and methods for monotone missing patterns can be required by option. There is a *SAS* macro available for Little's test for MCAR (www.appliedmissingdata.com/littles_mcar_test.sas, retrieved Dec 13, 2013). After imputation, statistical procedures are run separately for each imputation run by using the PROC MI output file *outmi* and _Imputation_ as a BY variable, and the results are stored in an output file *output*. Then the final pooled result is produced by the procedure PROC MIANALYZE data = *output*.

FIML is available in all major software for structural equation modeling such as *LISREL*, *MPlus*, the *R* package *lavaan*, or *AMOS*. Alternatively Bayesian estimation can be used that is however outside of the scope of this article (see van de Schoot et al., 2014, for a gentle introduction).

Conclusion

Nowadays standard software allows for reducing biases due to selective attrition and other types of missing values in longitudinal studies. *Listwise deletion* should be avoided. *Pairwise deletion* is an option only in the case of MCAR and very few missing data. Therefore *multiple imputation* of missing data with subsequent analyses based on the pooled imputed data should be considered the default procedure unless one-step estimation with FIML is an option or multi-level models are used. Multiple imputation can be done with highest flexibility using *R* and most easily with *SPSS* if one uses *SPSS* for the subsequent analyses.

Notes

1. We thank Sonja Winter for conducting these analyses.
2. This article focuses on methods that deal with the MAR case. MAR is often, but not always, a reasonable assumption. For example, in a longitudinal study on aggressiveness, the dropout tendency of highly aggressive kids may continuously increase such that it is only partly controlled by their earlier aggressiveness. If no concurrently observed variables related to aggressiveness are available, predicting missing aggressiveness scores by earlier variables may still underestimate the stability of aggressiveness. There are approaches to control for such cases of MNAR (see Enders, 2011) but they require complex analysis techniques that are outside the scope of this article.
3. Imputation of the few missing scores (less than 3% within each wave of assessment) from the two remaining scores before aggregation resulted only in minimal changes. In order to be consistent with published results by Asendorpf et al. (2008), here, we prefer reporting the aggregate based on non-imputed missing scores.
4. For the reason mentioned in Note 3, we report here results for aggregates of mother and father reports at ages 17 and 23 based on non-imputed missing values (less than 3% missing item scores for each judge, 16% of the aggregated scores were provided by only one parent). Because of the latter higher percentage, we also ran all analyses with multiple imputation separately for mothers and fathers at ages 17 and 23 before we averaged their scores; the results remained virtually identical.
5. The reduction of variance in Wave i of the study is measured as $100 \times (1\text{-}s_i^2/s_1^2)$ where s_1^2 and s_i^2 are the observed variances in Wave 1 for the full sample and for the subsample participating also in Wave i, respectively.
6. Correction for range restriction used Thorndike's Case 2 formula (see Sackett & Yang, 2000) $r_i' = (s_1/s_i)r_i/(1 + r_i^2(s_1^2/s_i^2 - 1))^{1/2}$ where r_i is the observed correlation between Waves 1 and i ($i > 1$), r_i' is the corrected correlation, and s_1, s_i are the observed standard deviations in Wave 1 for the full sample and for the subsample participating also in Wave i, respectively.
7. Although the number of auxiliary variables was low (30 would be possible according to Hardt et al.'s (2012) suggestion as there were at least 92 participants without missing values, see Table 1), other variables of the LOGIC study did not meet the requirements for auxiliary variables because the correlations with the core variables were non-significant or they shared the missing data of the core variables. For curiosity, we nevertheless included 35 additional parental reports of personality (of shyness, sociability and the Big Five at various ages), violating the 1:3 recommendation. The result was that after imputation most of the correlations reported in Table 1 for the uncorrected variables decreased, 4 of the 12 initially significant correlations became non-significant, and all non-significant correlations remained non-significant. Thus imputation was detrimental in this case due to overuse of auxiliary variables.
8. *MPlus* offers options for including auxiliary variables in FIML estimation.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Asendorpf, J. B., Denissen, J. J. A., & van Aken, M. A. G. (2008). Inhibited and aggressive preschool children at 23 years of age: Personality and social transitions into adulthood. *Developmental Psychology*, *44*, 997–1011.

Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (1988). *Life-span developmental psychology: An introduction to research methods*. Hillsdale, NJ: Erlbaum.

Caspi, A., Moffitt, T. E., Thornton, A., Freedman, D., Amell, J. W., Harrington, H., . . . Silva, P. A. (1996). The Life History Calendar: A research and clinical assessment method for collecting retrospective event-history data. *International Journal of Methods in Psychiatric Research*, *6*, 101–114.

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

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*, 1–16.

Fraley, R. C., & Roberts, B. W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review*, *112*, 60–74.

Göttert, R., & Asendorpf, J. B. (1989). Eine deutsche Version des California Child Q Sort, Kurzform [A German short version of the California Child Q-Set]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *21*, 70–82.

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

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

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC Medical Research Methodology*, *12*, 184–196.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). London, UK: Routledge.

Hutteman, R., Denissen, J. J. A., Asendorpf, J. B., & van Aken, M. A. G. (2009). Changing dynamics in problematic personality: A multiwave longitudinal study of the relationship between shyness and aggressiveness from childhood to early adulthood. *Development and Psychopathology*, *21*, 1083–1094.

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

Marsh, H. W. (1993). Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor models. *Journal of Educational Measurement*, *30*, 157–183.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, *100*, 674–701.

Pearson, K. (1903). Mathematical contributions to the theory of evolution: II. On the influence of natural selection on the variability and correlation of organs. *Royal Society Philosophical Transactions*, *200*(Series A), 1–66.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*, 112–118.

Tremblay, R. E. (2000). The development of aggressive behavior during childhood: What have we learned in the past century? *International Journal of Behavioral Development*, *24*, 129–141.

van Buren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

van de Schoot, R., Kaplan, D., Denissen, J. J. A., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction into Bayesian analysis: Applications to developmental research. *Child Development*, *85*(3), 842–860. doi:10.1111/cdev.12169

Weinberger, D. A., Tublin, S. K., Ford, M. E., & Feldman, S. S. (1990). Preadolescents' social-emotional adjustment and selective attrition in family research. *Child Development*, *61*, 1374–1386.