心理科学进展 2014, Vol. 22, No. 12, 1985–1994
Advances in Psychological Science DOI: 10.3724/SP.J.1042.2014.01985

·研究方法(Research Method)·

# 追踪研究中缺失数据处理方法及应用现状分析*

叶素静[1] 唐文清[1,2] 张敏强[1] 曹魏聪[1]

([1]华南师范大学心理学院, 心理应用研究中心, 广州 510631)
([2]广西大学教育学院, 南宁 530004)

**摘 要** 追踪研究中普遍存在缺失数据, 缺失数据处理方法的选择影响统计推断的精度及研究结果的有效性。首先, 阐述缺失机制及判断方法, 比较追踪研究中主要的缺失数据处理方法的特点、及实际应用中的缺失处理方法的选择和软件实现。其次, 对国内心理学中 92 篇追踪研究文献进行分析, 发现有 59 篇(64.13%)报告不同程度缺失, 其中仅 39 篇报告了处理方法且均为删除法。未来研究应深入探讨现有缺失数据处理方法的有效性, 进一步规范应用研究中缺失数据的处理。

**关键词** 追踪研究; 缺失数据; 缺失机制; 缺失数据处理方法
**分类号** B841

## 1 引言

追踪研究方法作为探讨心理与行为发生、发展和变化的有效途径之一, 已被广泛应用于教育、发展、社会、临床等心理研究领域。但是, 追踪研究需收集一批被试在几个时间点的测量数据, 数据缺失现象普遍存在, 且缺失比横断研究更为严重、复杂。被试作答过程的疏忽、回避等导致某次测量中研究变量数据不完整, 或追踪过程中被试没有时间或不愿参与等都会造成数据缺失。数据缺失比例过大会导致样本不具代表性和估计偏差。一般认为, 缺失率为 5%~10%是可接受的, 最好不超过三分之一(风笑天, 2006), 当缺失达 60%以上时, 数据完全失去利用价值(Barzi & Woodward, 2004)。因此, 当数据缺失不能忽略, 需采用特定方法对其进行处理。

缺失数据处理方法的选择影响处理的精确性和研究结果的有效性。追踪研究中传统的缺失数据处理方法主要有删除法(deletion)、单一插补法(single imputation), 这两种方法简单易行, 但会造成参数估计偏差和统计功效的损失(Enders, 2011a, 2013; Graham, 2009)。近几十年, 特别是 Rubin (1976)提出缺失机制理论框架后, 现代缺失数据处理方法陆续被提出, 其中基于随机缺失的极大似然估计(maximum likelihood estimation, MLE)和多重插补法(multiple imputation, MI)因具有较好的处理精度和较广的应用范围, 最为研究者推崇(Schafer & Graham, 2002)。针对追踪研究中的非随机缺失, 研究者提出了选择模型(selection model)和模式混合模型(pattern-mixture model), 并成为研究热点(Enders, 2011a, 2011b; Muthén, Asparouhov, Hunter, & Leuchter, 2011)。

文章首先介绍缺失机制及其判断方法。其次, 阐述追踪研究中不同缺失机制下缺失数据处理方法的特点, 及实际应用中方法的选择和软件实现。第三, 对国内心理学领域的追踪研究中缺失数据处理方法的应用现状进行文献分析。最后, 针对缺失数据处理方法的研究和应用存在的问题提出建议。

收稿日期: 2014-01-01
* 国家社会科学基金“十二五”规划教育学一般课题(BHA130053)、广东省哲学社会科学规划项目(09sxk1q001)、广州市基础教育学业质量监测系统项目(GZIT2013-ZB0465)、“国家基础科学人才培养基金‘华南师范大学心理学基地’(J1030729)”。
通讯作者: 张敏强, E-mail: zhangmq1117@qq.com

## 2 缺失机制及判断方法

### 2.1 缺失机制

缺失数据处理方法建立在特定的缺失机制基础上，缺失机制描述的是数据缺失概率与研究变量间的关系(Schafer & Graham, 2002)。Rubin (1976)将缺失机制分为三类：完全随机缺失(missing completely at random, MCAR)、随机缺失(missing at random, MAR)和非随机缺失(missing not at random, MNAR)。

假设对 $N$ 个被试进行 $T$ 个测量时间点的追踪，追踪研究变量为 $Y=(y_1,\ldots, y_K)$。最终样本为 $Y_{com}=(y_{ijt})$, $i = 1,..., N$, $j = 1,..., K$, $t = 1,..., T$。缺失数据存在时，该样本包括完整观测部分和缺失部分，分别为 $Y_{obs}$ 和 $Y_{mis}$。另设 $R=(r_{ijt})$为指示变量矩阵，若 $y_{ijt}$ 缺失 $r_{ijt} = 0$，否则为 1。缺失机制本质是描述 $R$ 的分布，若 $y_{ijt}$ 缺失概率只与 $Y_{obs}$ 有关而与 $Y_{mis}$ 无关，即 $P(R|Y_{com},\theta,\xi) = P(R|Y_{obs},\theta,\xi)$ ($\theta$，$\xi$ 分别为 $Y$, $R$ 的分布函数参数)，此类缺失称为 MAR。如果 $y_{ijt}$ 缺失概率与 $Y_{obs}$、$Y_{mis}$ 都无关，即 $P(R|Y_{com},\xi) = P(R|\xi)$，则为 MCAR。MCAR 要求数据缺失是等可能的，是个更强的假设。MNAR 则指数据缺失依赖于未观测到的数据。如一项关于老年人认知功能的追踪研究中，若被试因无关变量(如搬家)未参加第 $t$ 次测量可认为是 MCAR。若被试因前一次已测得的认知功能得分($y_{(t-1)}$)低而不愿参加，$y_t$ 缺失可由已观测的 $y_{(t-1)}$解释，这种情况为 MAR。若被试因当前认知功能($y_t$)衰退不能参加测量，$y_t$ 缺失是由未观测的 $y_t$ 自身导致，该情况为 MNAR。

### 2.2 缺失机制判断方法

缺失数据处理方法的有效性依赖于特定缺失机制的成立。因 $Y_{mis}$ 的存在难以真正判断数据缺失属于哪种类型，研究者提出以下粗略判断方法。

(1) MCAR 机制检验。Dixon (1983)通过多次 $t$ 检验比较在变量 $y_j$ 上有、无缺失的两组样本在无缺失变量上均值的差异性，若差异不显著则为 MCAR。但当变量数较多，检验统计量间会出现相关而影响检验结果。Little (1988)将检验整合成一个服从卡方分布的检验统计量，若卡方检验 $P$ 值大于设定的显著性水平则为 MCAR，该检验可直接在 SPSS 12.0 及以上版本的缺失数据分析模块中实现。对于包括非正态数据的情况，有研究者提出从分布特征入手，若变量 $y_j$ 上有、无缺失的两组样本的分布一致则为 MCAR (孙婕，金勇进，戴明锋, 2013)。

(2) MAR、MNAR 机制检验。目前 MAR 机制检验研究多集中在单调缺失(即未参与第 $t$ 时间点测试的样本也不参与 $t$ 时间点后的测试)情况，如 Diggle (1989)运用 Kolmogorov-Smirnov 检验判断样本 1(未参与 $t$+1,…,$T$ 时间点测试的样本)是否为样本 2(参与 $t$ 时间点测试的样本)的随机样本，若是则为 MAR。正态分布假设下，Listing 和 Schlittgen (1998)检验完全追踪样本与样本 1(未参与 $t$+1,…,$T$ 时间点测试的样本)在 $t$ 时间点的变量均值的差异性，若差异不显著则为 MAR; Listing 和 Schlittgen (2003)通过 Wileoxon 秩和检验将该方法推广到非正态分布的情况。此外，相当部分研究建立以 $R$ 为因变量，以完全观测的变量为自变量的 logistic 回归模型，通过检验回归系数的显著性来粗略判定 MAR 及 MNAR (孙晓松, 2007; Jeličić, Phelps, & Lerner, 2009; 孙婕等, 2013)。

## 3 追踪研究中缺失数据处理方法

### 3.1 传统处理方法

传统的缺失数据处理方法多基于 MCAR，如删除法，它包括列删除法(list-wise deletion)和成对删除法(pair-wise deletion)，前者删除有缺失的样本数据，后者估计不同参数时使用对应变量完整观测的样本数据。删除法简单易行，许多软件提供此功能。若 MCAR 满足且缺失率很小(10%以下)，删除法可获得理想的处理效果(茅群霞，李晓松, 2005; Clarke & Hardy, 2007)。但实际中 MCAR 很难满足，删除法会造成估计偏差，即使 MCAR 成立，也有诸如统计功效损失的问题(刘红云，张雷, 2005; Enders, 2013)。

另一传统处理方法是单一插补法，它用某个“看似”合理的值代替缺失值，主要有均值插补、LOCF (last observation carried forward)及回归插补。均值插补用变量均值代替该变量的缺失值。LOCF 法用前期观测值 $y_{ij(t-1)}$代替 $y_{ijt}$ 的缺失值。回归插补根据变量间的相关关系，利用其他变量信息建立回归方程来预测缺失值。单一插补保留了样本量，但无论用何种方法都可能存在扭曲样本分布的问题(庞新生, 2010)，导致有偏估计(Barzi, Woodward, Marfisi, Tognoni, & Marchioli, 2006;

Enders, 2013)。

### 3.2　基于随机缺失的处理方法

#### 3.2.1　极大似然估计

MLE 通过构造似然函数并求最值来获得参数的估计值。当 $Y_{mis}$ 存在时, 难以求解似然函数的最值, 目前主要通过特殊方法如期望极大(expectation maximization, EM)算法和全息极大似然估计(full information maximum likelihood, FIML)来解决该问题。

**(1) EM 算法。**EM 算法(Dempster, Laird, & Rubin, 1977)的基本思想是 $Y_{mis}$ 含有与估计参数 $\theta$ 有关的信息, 而参数 $\theta$ 反过来有助于寻找最可能的 $Y_{mis}$ 值。基于 $Y_{mis}$ 与参数 $\theta$ 的依赖性, EM 算法以设定参数初始值 $\theta^{(0)}$为起点进行迭代运算：依据 $Y_{obs}$、$\theta^{(k)}$估计 $Y_{mis}^{(k)}$, 然后根据 $Y_{obs}$、$Y_{mis}^{(k)}$ 计算出参数估计值 $\theta^{(k+1)}$。每一次迭代由条件期望步(E 步)、极大化步(M 步)组成。

条件期望步(E 步)：在 $\theta^{(k)}$基础上, 计算对数似然函数对 $Y_{mis}$ 的条件期望, 如公式(1)。其中 $F(Y_{mis}|Y_{obs}, \theta^{(k)})$是用 $Y_{obs}$ 去预测 $Y_{mis}$ 的回归方程, 对 $\theta^{(k)}$进行 SWP 运算(sweep operator)可获得回归系数(Schafer, 1997)。如变量 $Y_1$、$Y_2$ 的样本中, $Y_1$、$Y_2$ 的缺失值可分别由 $F(Y_1|Y_2, \theta^{(k)})$、$F(Y_2|Y_1, \theta^{(k)})$ 的预测值代替。$Y_{mis}$ 被填补后, $Q(\theta|\theta^{(k)})$ 只含有未知数 $\theta$。

$$Q(\theta|\theta^{(k)}) = \int \ln(\theta|Y_{obs}, Y_{mis}) F(Y_{mis}|Y_{obs}, \theta^{(k)}) dY_{mis} \quad (1)$$

极大化步(M 步)：极大化 $Q(\theta|\theta^{(k)})$ 求得估计值 $\theta^{(k+1)}$, 即：

$$Q(\theta^{(k+1)}|\theta^{(k)}) = \max\{Q(\theta|\theta^{(k)})\} \quad (2)$$

EM 算法的每一次迭代均使似然函数值增加。若似然函数有界, 序列$\{\theta^{(k)}\}$将收敛到一个平稳值, 一般情况下该平稳值为 $\theta$ 的极大似然估计值(Little & Rubin, 1987)。通过如上两步, EM 算法将缺失数据问题转化为重复解决完整数据的问题。

EM 算法输出的是样本均值、方差、协方差及相关系数矩阵等的极大似然估计值, 通过一定的转换可计算出模型参数的估计值, 这使得 EM 算法可纳入与缺失有关但非分析模型组成部分的辅助变量(auxiliary variables) (Enders, 2001), 然后选取需要的变量均值、方差、协方差等进行后续分析。Newman (2003)模拟含缺失的追踪数据并使用 EM 算法估计样本相关系数矩阵, 最后将其输入 LISREL 进行结构方程模型参数估计, 研究发现依据该相关系数矩阵获得的参数估计精度优于删除法和回归插补法, 特别是缺失率较大时, 但是标准误的表现略逊于其他基于 MAR 的处理方法如 FIML、MI。

**(2) 全息极大似然估计。**FIML 也称原始极大似然估计(raw maximum likelihood) (Duncan, Duncan, & Li, 1998)、直接似然分析(direct likelihood analysis) (Kadengye, Cools, Ceulemans, & Van den Noortgate, 2012)。与 MLE 步骤一致, FIML 也包括构造对数似然函数并求最值, 但似然函数构造过程有所不同。假设研究变量 $Y$=($Y_1$, $Y_2$, $Y_3$, $Y_4$)服从多元正态分布, 对数似然函数为 $L(\mu,\Sigma) = \sum_{i=1}^{n} \ln L_i$, 其中：

$$\ln L_i = (2\pi)^{p/2} - \frac{1}{2}\ln|W_i\Sigma W_i'| - \frac{1}{2}(Y_i-\mu)'W_i'(W_i\Sigma W_i')^{-1}W_i(Y_i-\mu) \quad (3)$$

$n$ 为样本量, $p$ 为被试 $i$ 有观测值的变量个数。对于 $p$=4 的被试 $i$, $W_i$ 为 4×4 的单位矩阵。若被试在第 $j$ 个变量上有缺失, $W_i$ 则为 4×4 单位矩阵去掉第 $j$ 行后形成的矩阵。

FIML 中缺失数据的处理和参数估计在同一步进行, 研究者只需输入原始数据就能得到参数估计值, 简单易掌握(Duncan et al., 1998)。FIML 可直接得到合理的标准误和置信区间。但若要考虑辅助变量, 就必须将其纳入分析模型中, 这可能会扭曲模型意义。为弥补这方面不足, Graham (2003)在结构方程模型框架下提出两个基于 FIML 的模型：额外因变量模型(extra dependent variable model)和饱和关联模型(saturated correlates model), 两者可在纳入辅助变量的同时不改变模型的意义。Raykov (2005)运用基于 FIML 的结构方程模型分析有缺失的老年人认知追踪数据。相当部分研究表明在处理追踪研究缺失数据时, FIML 在参数估计和标准误方面优于删除法、回归插补法等传统方法 (Duncan et al., 1998; Jeličić, Phelps, & Lerner, 2010; Wothke, 2000; Newman, 2003)。

#### 3.2.2　多重插补

MI 法分三步对缺失数据进行处理。第一步是数据插补, 通过插补模型为每个缺失值产生 $M$

($M$>1)个合理替代值，得到 $M$ 组完整数据集。具体过程是先从参数 $\theta$ 的后验分布 $F(\theta|Y_{obs})$抽样或通过非参数 bootstrap 抽样方法得到 $M$ 个参数值 $\theta^{(1)},\ldots,\theta^{(M)}$，然后从 $F(Y_{mis}|Y_{obs},\theta^{(k)})$ 抽样得到替代值 $Y_{mis}^{(k)}$，$k$=1,..., $M$。根据获得 $\theta^{(k)}$、$Y_{mis}^{(k)}$ 的方法不同，插值过程可采用回归方法、预测均数匹配法(predictive mean matching)、趋势得分法(propensity score)、EM 算法及马尔可夫链蒙特卡罗方法(markov chain monte carlo)等(Yuan, 2010)。关于 $M$ 的取值，早期研究认为 3~5 个插值即可(Rubin, 1987; Schafer & Olsen, 1998)。最近研究指出，$M$ 取值与信息缺失率γ(指由于缺失数据造成的参数估计变异增加程度)有关，γ 越大，$M$ 需相应增加(Graham, Olchowski, & Gilreath, 2007)。一般建议 $M$ 应不少于 20(Little, Card, Preacher, & McConnell, 2009)。第二步是运用标准的统计方法分析每一个完整数据集，得到每个参数的 $M$ 个估计值和对应的标准误。最后整合 $M$ 组结果得到最终参数估计并进行统计推断。

MI 法通过构造不同插补值反映因数据缺失引起的样本变异，能得到精确的参数估计和合理的标准误。Newman (2003)通过模拟研究比较不同缺失率和缺失机制下删除法、回归插补、EM 算法、FIML 及 MI 法对追踪研究中缺失数据的处理效果，发现总体上 MI 法、FIML 及 EM 算法在参数估计方面比删除法、回归插补更精确；MI 法及 FIML 能得到更合理的标准误估计。Jeličić 等(2010)运用列删除法、FIML 和 MI 法处理追踪数据中的缺失，发现 MI 法和 FIML 在参数估计和标准误上比列删除法表现好。其他研究也表明 MI 法比传统方法能更好处理追踪研究中的缺失数据问题，特别是缺失率在 10%以上时优势更明显(Barzi et al., 2006; 易昆南，袁中萸, 2008; 茅群霞，李晓松, 2005)。此外，MI 法可将辅助变量引入插补模型，然后对感兴趣的部分变量进行统计建模分析(Schlomer, Bauman, & Card, 2010)，但相比 EM 算法及 FIML, MI 法的过程更复杂，耗时更长。

### 3.3 基于非随机缺失的处理方法

追踪研究中有时出现 MNAR 情况，基于 MAR 的方法不适用。对有 MNAR 的追踪数据的分析过程需加入一个描述缺失特征的子模型来矫正偏差。子模型的构造不同衍生出选择模型和模式混合模型，前者通过纳入一系列的回归模型来描述在每个观测点数据缺失的概率，后者首先构造缺失模式，然后在每个模式内拟合模型并整合(Enders, 2011a, 2011b; Muthén et al., 2011)。

#### 3.3.1 选择模型

Heckman (1976)提出选择模型，并将其用于矫正回归分析中因数据缺失引起的参数估计偏差，Wu 和 Carroll (1988)用该方法处理追踪研究中 MNAR 问题。选择模型将 $R$ 与 $Y_{com}$ 的联合分布分解为 $Y_{com}$ 的边际分布和给定 $Y_{com}$ 下 $R$ 的条件分布的乘积，如公式(4)所示：

$$F(Y_{com},R|\theta,\xi)=F(Y_{com}|\theta)F(R|Y_{com},\xi) \quad (4)$$

运用选择模型处理追踪研究中的缺失数据时，研究者首先构建 $Y_{com}$ 的统计模型，如潜变量增长模型(latent growth curve model)等。其次通过假定 $R$ 与 $Y_{com}$ 的回归关系来求解 $F(R \mid Y_{com}, \xi)$，常用 Logistic 或 Probit 回归模型。该回归模型以 $R$ 为因变量，用 $Y_{com}$ 直接或间接预测数据缺失概率，如 Diggle 和 Kenward (1994)直接用追踪研究变量 $Y$ 去预测时间点 $t$ 的数据缺失概率，该模型适用于由 $Y_{mis}$ 本身导致的 MNAR (Enders, 2011a)。Wu 和 Carroll (1988)利用个体增长曲线的截距和斜率间接估计 $R$ 的分布，该模型也称共享参数模型(shared parameter model)，它适用于依赖个体发展曲线的 MNAR 情况(Enders, 2011a)。

选择模型基本思想直观，且可直接求解研究模型 $F(Y_{com}|\theta)$。但 $R$ 对 $Y_{com}$ 的回归本质上难以估计，因当 $R$=0 对应的都是 $Y_{mis}$，这时模型可能不可识别。一般的解决方法是加入严格的分布假设，如假定追踪研究变量 $Y$ 或者个体增长曲线的截距、斜率服从正态分布。但研究表明选择模型对该假设较敏感，细微偏离就可能造成实质性的偏差(Schafer & Graham, 2002; Enders, 2013)。Enders (2011a)在多层增长模型(multilevel growth model)框架下，比较 FIML、MI 法、选择模型及共享参数模型在处理 MNAR 时的表现，发现选择模型和共享参数模型由于回归模型定义不正确及正态分布假设不满足而造成较大的偏差，因此有待进一步研究。

#### 3.3.2 模式混合模型

模式混合模型(Little, 1994, 1995; Hedeker & Gibbons, 1997)将 $R$ 与 $Y_{com}$ 的联合分布分解为缺失

数据模式的边际分布和具体缺失模式 $R$ 下 $Y_{com}$ 的条件分布的乘积, 如公式(5)所示：

$$F(Y_{com},R\mid\theta,\xi)=F(R\mid\xi)F(Y_{com}\mid R,\theta) \quad (5)$$

模式混合模型在处理追踪研究中的 MNAR 数据时, 第一步识别不同的数据缺失模式, 缺失模式通常根据缺失数据出现的测量时间点确定。如一个 $T$=3 的追踪研究, 可将三次测量均参加的被试分为第一组, 只有第三次未参加的分为第二组, 以此类推。但当 $T$ 较大时, 缺失模式可能很多, 导致某些模式内被试过少不足以拟合模型。Roy (2003)、Roy 和 Daniels (2008)提出潜在模式混合模型(latent pattern mixture models), 用潜在类别变量代替传统的时间点分组, 既可减少缺失模式个数又能捕捉到被试的本质差别。第二步在每个模式中拟合模型, 得到模型参数估计值 $\hat{\theta}_i$ ($i$ 为缺失模式个数)。假设每个模式内的样本比重为 $\pi_i$, 那么总体的参数估计值是这 $i$ 个估计值的加权平均值： $\hat{\theta}=\sum_i \pi_i\hat{\theta}_i$。

Enders (2011a)比较列删除法、FIML、MI、选择模型与模式混合模型在处理 MNAR 数据的表现, 发现模式混合模型的参数返真性最好。但模式混合模型也存在模型识别问题, 如上面例子中第 2 个缺失模式中第 3 次的观测值全部缺失, 信息不足可能会使 $F(Y_{com}|R_2,\ \theta)$不能识别(Little, 1993)。为达到模型识别, 方法之一是简化模型以减少待估参数数目, 如 Hedeker 和 Gibbons (1997)将 $R$ 以协变量形式引入模型。方法二是将某些模式中不可估计的参数设定为其他可估计模式(如没有缺失数据的模式)中对应参数的函数, 如完整数据模式缺失变量分布假设(complete-case missing variable) (Little, 1993)、可估计模式缺失变量分布假设(available-case missing variable) (Molenberghs, Michiels, Kenward, & Diggle, 1998)和相邻模式缺失变量分布假设(neighboring case missing variable) (Thijs, Molenberghs, Michiels, Verbeke, & Curran, 2002)。Thijs 等(2002)和 Enders (2011b)探讨此类假设在模式混合模型中的表现, 发现参数估计结果因假设不同而有所不同, 具体使用时需仔细分析。

### 3.4　缺失数据处理方法选择及软件实现

#### 3.4.1　处理方法的选择

缺失数据处理方法基于不同的理论基础, 表现也不尽相同。结合 Duncan 等(1998)及刘红云和张雷(2005)的观点, 一种优秀的缺失数据处理方法应该具备：(1) 处理缺失数据过程可考虑造成数据缺失的原因, 特别是辅助变量; (2) 采用与数据分析相同的模型处理缺失数据; (3) 可得到无偏估计和合理的标准误。上述缺失处理方法在三个标准的表现见表 1。

**表 1**　缺失数据处理方法的表现

| 处理方法 | 处理方法评价标准 | | |
|---|---|---|---|
| | 标准 1 | 标准 2 | 标准 3 |
| 传统方法 | 难加入辅助变量 | √ | MCAR 且缺失率≤10%满足 |
| EM | √ | √ | MCAR, MAR 满足 |
| FIML | 难加入辅助变量 | √ | MCAR, MAR 满足 |
| MI | √ | 插值模型一般与最终分析模型不一致 | MCAR, MAR 满足 |
| 选择模型 | √ | √ | 不确定 |
| 模式混合模型 | √ | √/但需整合 | 不确定 |

追踪研究中缺失数据处理不存在一种最佳方法, 研究者在进行方法选择时需考虑缺失机制、缺失率等具体情况。根据各种方法特点(见表 1)及相关研究对追踪研究缺失数据处理方法选择给出初步的建议：(1) 对于 MCAR, 如果缺失率较小(10%以下)且完整观测样本量满足统计分析要求, 传统方法是个简便且效果较好的方法。若缺失率较大, MLE 及 MI 法在参数估计和统计功效方法显著优于传统方法。(2) 对于 MAR, 如果影响数据缺失的变量在统计模型或插值模型中, MLE 和 MI 法一般能得到一致的结果(Enders, 2011a), 但 MLE 法实施过程易于 MI 法。EM 算法及 MI 法在

处理缺失过程更容易纳入辅助变量。

对于 MNAR, 研究者难以知道数据缺失的真正原因, 建立的子模型可能不符合实际情况。已有研究表明 MNAR 处理方法对子模型、正态分布等假设较敏感(Enders, 2011a, 2011b, 2013), 且计算复杂, 较难实施(Muthén et al., 2011; Enders, 2013), 因此有待进一步研究。Schafer 和 Graham (2002)提议在每次测量结束后询问被试“有多大可能不参与下一个时间点的测量”并将该变量引入模型中, 以达到将 MNAR 转变为 MAR。另外, 有研究提议将选择模型和模式混合模型作为敏感性分析的组成部分, 同时用几种不同方法(或假设)拟合数据再比较拟合结果, 然后作出选择(Molenberghs & Kenward, 2007; Molenberghs, Verbeke, & Kenward, 2009; Enders, 2011b)。

**3.4.2** 软件实现

目前, 大部分心理学常用的统计软件可实现对追踪研究中的 MCAR、MAR 缺失数据的处理, 如 SPSS、SAS、R、MPLUS、HLM 和 LISREL 等, 具体见表 2。

(1) 一般软件如 SPSS、SAS 和 R。SPSS 12.0 及以上版本新增的缺失数据分析模块, 提供列删除法、回归插补等传统处理方法。此外, 可实现 EM 算法及基于 EM 算法的单一插补, 但仅限于连续型变量, 且 SPSS 12.0~16.0 版本中基于 EM 算法的插补过程未考虑随机误差项, 获得的插补值会严重偏离真实值。SPSS 17.0 及以上版本纳入随机误差项, 新增了 MI 模块, 可提供回归方法、MCMC 方法的 MI 法。SAS、R 软件除了能实现传统处理方法、EM 算法外, 还提供针对不同数据类型的 MI 法(如 SAS 的 Proc MI, R 中的 MI, MICE), 且易于进行后续分析和结果整合。

(2) 专业软件如 MPLUS、HLM 和 LISREL 等。MPLUS 5.0 及以上版本, LISREL 8.7 及以上版本均可使用 FIML 的缺失数据处理方法。对于 MI 法, MPLUS 6.0 及以上版本可实现 MCMC 方法的 MI。LISREL 8.7 及以上版本中的数据操作和基本分析模块 PRELIS 提供了基于 EM 算法、MCMC 方法的 MI, 并可输出 EM 算法获得的协方差矩阵等估计值, 可将其作为数据输入进行 SEM 分析。HLM 6.0 及以上版本可对 MI 数据进行分析并进行结果整合。

对于追踪研究中 MNAR 数据, MPLUS、SAS 中的 PROC MI, PROC MIANALYZE 可进行选择模型、模式混合模型分析。

**表 2** 缺失数据处理方法的软件实现

| 处理方法 | 统计软件 | | | | | |
|---|---|---|---|---|---|---|
| | SPSS | SAS | R | MPLUS | LISREL | HLM |
| 传统方法 | √ | √ | √ | 列/成对删除法 | 列/成对删除法 | 列/成对删除法 |
| EM | 12.0 及以上版本 | PROC MI | NORM,CAT | × | √ | × |
| FIML | × | PROC CALIS | √ | √ | √ | × |
| MI | 17.0 及以上版本 | PROC MI | MI, MICE | √ | √ | 可分析 MI 数据并整合结果 |
| 选择模型 | × | PROC MI, PROC MIANALYZE | × | √ | × | × |
| 模式混合模型 | × | | × | √ | × | × |

## 4 国内心理学追踪研究中缺失数据处理方法应用现状

为了解国内心理学领域追踪研究中缺失数据处理方法的应用情况, 以中国期刊网全文数据库为数据源, 收集 1980~2013.4 期间追踪研究的相关文献, 去除文献综述、理论研究文献、个案追踪研究、已在心理学期刊上公开发表的学位论文及非心理学期刊发表的学术论文, 最后入选文献 92 篇。其中 10 本心理学期刊文献 71 篇, 博士学位论文 5 篇, 硕士学位论文 16 篇。对 92 篇文献的数据缺失情况、缺失机制检验、缺失数据处理方法进行统计分析。结果如下:

(1) 数据缺失方面, 92 篇追踪研究中有 59 篇(64.13%)报告了不同程度(1.40%~87.30%)的数据缺失(见表 3), 其中缺失 20%以上的占 55.94%。

(2) 缺失机制检验方面, 59 篇有报告缺失的研究中, 仅 7 篇(11.90%)通过卡方检验或独立样本 $t$ 检验分析缺失样本与完全追踪样本间或最终

表 3　59 篇报告缺失的研究中缺失率的分布表

| 缺失率 | 篇数 | 百分比(%) |
|---|---|---|
| (0, 10%] | 13 | 22.03 |
| (10%, 20%] | 13 | 22.03 |
| (20%, 30%] | 11 | 18.64 |
| (30%, 40%] | 3 | 5.08 |
| (40%, 50%] | 10 | 16.95 |
| (50%, 90%] | 8 | 13.56 |
| 其他 | 1 | 1.69 |
| 总和 | 59 | 100.0 |

注：“其他”指的是只汇报最终选取的样本量，无法计算缺失率。

分析样本与初测样本的差异性来粗略判断缺失机制。其中 2007、2008、2013 年各 1 篇, 2009、2011 年各 2 篇。

(3) 缺失数据处理方面，59 篇有报告缺失的研究中，39 篇报告了处理方法且均为删除法。Jeličić 等(2009)分析 2000~2006 年间发表在 3 个重要发展科学期刊上的 100 篇追踪研究，发现 12.28%使用 FIML, 1.75%使用 MI 法。相比之下，国内心理学领域的追踪研究中缺失数据处理的科学性还有待提高。

文献分析结果表明国内心理学领域追踪研究中缺失数据的普遍性及严重性，处理方法以传统方法为主；绝大部分研究未对缺失机制进行检验。出现此情况的原因可能有如下几点。

(1) 忽视缺失数据引起的问题。大部分研究者并不了解具体的缺失数据处理方法，或未意识到传统方法与基于 MAR, MNAR 方法效果的差别。

(2) 缺失数据及其处理报告缺乏规范。虽然国内心理学论文写作规范对被试选取部分有一定的要求，如张侃(2002)提议“如果部分被试没有完成实验，应说明没有完成实验的被试数量，并解释他们没有继续实验的原因”，但对缺失报告、缺失处理要求不够严格、具体。

(3) 大部分统计软件如 SPSS 将删除法、均值插补等设置为默认选项，相比之下，MLE、MI 法的方法原理和实现过程较为复杂，研究者为便于操作一般会选择默认选项。

## 5　问题与展望

追踪研究中普遍存在缺失数据。传统的缺失数据处理方法应用范围狭窄且存在问题。缺失机制的理论框架提出后，现代处理方法被提出和完善。目前追踪研究中缺失数据处理方法的理论研究文献虽较多，但主要集中在数理统计学及生物医学统计领域，且不够全面深入。应用研究方面，文献分析结果显示国内心理学领域追踪研究中缺失数据的处理仍以传统方法为主。鉴于此，结合已有研究及文献分析的结果，对未来的理论和应用研究提出以下建议。

(1) 继续探讨现有缺失数据处理方法的有效性。首先，基于 MAR 的 MLE 及 MI 法在前提假设满足的情况下，能获得很好的处理效果。但实际研究中，理论假设如正态分布不一定成立，特别是在追踪研究中由于人力物力的限制，样本量一般较小，样本分布可能不服从正态分布，此时缺失处理的有效性是否会受影响有待深入研究。进一步的研究可探讨处理方法对模型假设的稳健性。其次，MNAR 处理方法的有效性问题。早期的选择模型和模式混合模型由于对模型假设依赖性强，表现不尽人意。目前已有研究者致力于扩展模型研究(Muthén et al., 2011; 季家超，王刚，张潇雅，刘桂芬，2013)，如何更好发挥它们的作用也需不断探索。

(2) 加强贝叶斯理论框架下的缺失数据处理方法的探讨。由于似然函数与极大化过程复杂，MLE 难以找到全局最大值。随着 MCMC 算法的发展，基于 MCMC 的贝叶斯方法将最大化问题转为抽样问题，避免了求解复杂的似然函数，很大程度简化了求解过程。Daniels 和 Hogan (2008)全面介绍了贝叶斯方法在处理追踪研究中缺失数据的原理和实例。Li (2006)将该方法应用于 MNAR 的情况。国内研究者杨林山和曹亦薇(2012)比较研究了完全贝叶斯(fully bayesian)方法与部分贝叶斯(partially bayesian)方法在处理不同缺失率的追踪数据的表现。未来研究可进一步探讨该方法在不同缺失机制下的表现，及其与 MLE、MI 等方法的比较。

(3) 规范应用研究中缺失数据处理。目前，应用研究者普遍使用删除法等传统方法，为改善这种现状，研究者应重视缺失数据引起的问题并学习相关的知识。除此之外，最重要的是规范学术论文撰写中缺失数据及处理报告的要求。APA 在 1999 年就提出研究者在呈现结论前须报告数据收

集过程中出现的缺失数据问题，包括缺失数据的模式、分布及处理方法(Wilkinson, 1999)。Burton 和 Altman (2004)提议分三步报告缺失数据的处理：报告每个变量的缺失情况；探讨数据缺失原因；汇报处理方法及具体实施过程。借鉴上述观点，具体到追踪研究，若涉及到缺失数据，建议研究者在撰写论文时按照以下步骤阐述：首先报告缺失数据的分布，包括每个测量时间点、每个变量的缺失率；其次通过差异性检验或 logistic 回归等粗略判断缺失机制；最后报告处理方法的选择与原因、具体的实施过程及处理结果。

## 参考文献


风笑天. (2006). 追踪研究: 方法论意义及其实施. *华中师范大学学报(人文社会科学版), 45*(6), 43–47.

季家超, 王刚, 张潇雅, 刘桂芬. (2013). 数据非随机缺失机制的混合效应模式混合模型分析与应用. *中国卫生统计, 30*(2), 221–225.

刘红云, 张雷. (2005). *追踪数据分析方法及其应用*. 北京: 教育科学出版社.

茅群霞, 李晓松. (2005). 多重填补法 Markov Chain Monte Carlo 模型在有缺失值的妇幼卫生纵向数据中的应用. *四川大学学报(医学版), 36*(3), 422– 425.

庞新生. (2010). 缺失数据处理方法的比较. *统计与决策,* (24), 152–155.

孙婕, 金勇进, 戴明锋. (2013). 关于数据缺失机制的检验方法探讨. *数学的实践与认识, 43*(12), 166–173.

孙晓松. (2007). *数据缺失机制及其检验* (硕士学位论文). 苏州大学.

杨林山, 曹亦薇. (2012). 贝叶斯理论框架下的 2 种纵向缺失数据处理方法的比较——以潜在变量增长曲线模型为例. *江西师范大学学报(自然科学版), 36*(5), 461–465.

易昆南, 袁中英. (2008). 对模拟纵向数据集缺失值处理的几种方法比较. *湖南工业大学学报, 22*(2), 48–51.

张侃. (2002). *心理学论文写作规范*. 北京: 科学出版社.

Barzi, F., & Woodward, M. (2004). Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology, 160*(1), 34–45.

Barzi, F., Woodward, M., Marfisi, R. M., Tognoni, G., & Marchioli, R. (2006). Analysis of the benefits of a Mediterranean diet in the GISSI-Prevenzione study: A case study in imputation of missing values from repeated measurements. *European Journal of Epidemiology, 21*(1), 15–24.

Burton, A., & Altman, D. G. (2004). Missing covariate data within cancer prognostic studies: A review of current reporting and proposed guidelines. *British Journal of Cancer, 91*(1), 4–8.

Clarke, P., & Hardy, R. (2007). Methods for handling missing data. In A. Pickles, B. Maughan, & M. Wadsworth (Eds.), *Epidemiological methods in life course research* (Vol. 1, pp. 157–197). New York: Oxford University Press.

Daniels, M. J., & Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for bayesian modeling and sensitivity analysis*. Boca Raton, Florida: CRC Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1–38.

Diggle, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics, 45*(4), 1255–1258.

Diggle, P. J., & Kenward, M. (1994). Informative dropout in longitudinal data analysis. *Journal of the Royal Statistical Society-Series C: Applied Statistics, 43*(1), 49–93.

Dixon, W. J. (Ed.) (1983). *BMDP statistical software*. Berkeley, California: University of California Press.

Duncan, T. E., Duncan, S. C., & Li, F. (1998). A comparison of model-and multiple imputation–based approaches to longitudinal analyses with partial missingness. *Structural Equation Modeling: A Multidisciplinary Journal, 5*(1), 1–21.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8*(1), 128–141.

Enders, C. K. (2011a). Analyzing longitudinal data with missing values. *Rehabilitation Psychology, 56*(4), 267–288.

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

Enders, C. K. (2013). Dealing with missing data in developmental research. *Child Development Perspectives, 7*(1), 27–31.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling, 10*(1), 80–100.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*(1), 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*(3), 206–213.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement,*

*5*(4), 475–492.

Hedeker, D., & Gibbons, R.D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2*(1), 64–78.

Jeličić, 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.

Jeličić, H., Phelps, E., & Lerner, R. M. (2010). 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. *Journal of Youth and Adolescence, 39*(7), 816–835.

Kadengye, D. T., Cools, W., Ceulemans, E., & Van den Noortgate, W. (2012). Simple imputation methods versus direct likelihood analysis for missing item scores in multilevel educational data. *Behavior Research Methods, 44*(2), 516–531.

Lerner & L. Steinberg (Eds.), (2009). *Handbook of adolescent psychology* (Vol. 1, pp. 15–54). Hoboken, New Jersey: John Wiley & Sons.

Li, J. H. (2006). *Analysis of longitudinal data with missing values* (Unpublished doctorial dissertation). University of California.

Listing, J., & Schlittgen, R. (1998). Tests if dropouts are missed at random. *Biometrical Journal, 40*(8), 929–935.

Listing, J., & Schlittgen, R. (2003). A nonparametric test for random dropouts. *Biometrical Journal, 45*(1), 113–127.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198–1202.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association, 88*, 125–134.

Little, R. J. A. (1994). A class of pattern mixture models for normal incomplete data. *Biometrika, 81*(3), 471–483.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90*(431), 1112–1121.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. Hoboken, New Jersey: John Wiley & Sons.

Little, T. D., Card, N. A., Preacher, K. J., & McConnell, E. (2009). Modeling longitudinal data from research on adolescence. In R. M.

Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies.* Hoboken, New Jersey: John Wiley & Sons.

Molenberghs, G., Michiels, B., Kenward, M. G., & Diggle, P. J. (1998). Monotone missing data and pattern mixture models. *Statistica Neerlandica, 52*(2), 153–161.

Molenberghs, G., Verbeke, G., & Kenward, M. G. (2009). Sensitivity analysis for incomplete data. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 501–551). Boca Raton, Florida: CRC Press.

Muthén, B., Asparouhov, T., Hunter, A., & Leuchter, A. (2011). Growth modeling with non-ignorable dropout: Alternative analyses of the STAR*D antidepressant trial. *Psychological Methods, 16*(1), 17–33.

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods, 6*(3), 328–362.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

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

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

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology, 57*(1), 1–10.

Raykov, T. (2005). Analysis of longitudinal studies with missing data using covariance structure modeling with full-information maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal, 12*(3), 493–505.

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics, 59*(4), 829–836.

Roy, J., & Daniels, M. J. (2008). A general class of pattern mixture models for nonignorable dropout with many possible dropout times. *Biometrics, 64*(2), 538–545.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, New Jersey: John Willey & Sons.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., & Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics, 3*(2), 245–265.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.

Wothke, W. (2000). Longitudinal and multi-group 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. 197–216). Mahwah, New Jersey: Lawrence

Erlbaum Associates.

Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics, 44*(1), 175–188.

Yuan, Y. C. (2010). *Multiple imputation for missing data: Concepts and new development (Version 9.0)*. Rockville, MD: SAS Institute Inc.

## Techniques for Missing Data in Longitudinal Studies and Its Application

YE Sujing[1]; TANG Wenqing[1,2]; ZHANG Minqiang[1]; CAO Weicong[1]

([1] *Center for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou* 510631, *China*) ([2] *School of Education, Guangxi University, Nanning* 530004, *China*)

**Abstract:** Missing data are not uncommon in longitudinal studies. Different techniques for handling missing data affect accuracy of the results and validity of statistical inference. Firstly, we will elaborate on missingness mechanism and how to judge them. Then we make a summary of missing data techniques that mainly used in longitudinal study, and how to choose an appropriate missing data technique as well as software for analysis. Secondly, based on a literature review of psychology research in China, among 92 studies, we found that 59 contain a certain degree of missing data. Among these, 39 studies reported using deletion method. The validity of missing data techniques needs further study, and the reporting of missing data in published research also needs to be better established.

**Key words:** longitudinal study; missing data; missingness mechanism; missing data technique