

• 研究方法(Research Method) •

追踪研究中测验信度的估计*

叶宝娟 温忠麟 陈启山

(华南师范大学心理应用研究中心, 广州 510631)

摘要 追踪研究中测验工具的信度是衡量追踪研究质量的重要指标。传统的信度估计方法不适用于估计追踪研究的测验信度。近年来, 研究者提出了四种估计追踪研究的测验信度, 包括估计单个时间点的测验信度系数 ρ_w 和 $\rho(S_w)$, 以及估计整个追踪研究的测验信度系数 R_T 和 R_A 。本文评述了这四种信度估计方法的数学模型、前提假设及其优缺点。 R_T 和 R_A 既可估计追踪研究中单个时间点的测验信度, 也可估计追踪研究中整个追踪研究的测验信度, 所需要的前提假设较少, 推荐同时使用 R_T 和 R_A 来估计追踪研究的测验信度。

关键词 追踪研究; 信度; R_T 系数; R_A 系数

分类号 B841

1 引言

在心理、教育、社会和经济等研究领域, 许多自变量往往需要经过一段时间才能显示出其效应, 要较好地推断变量间的因果关系需进行追踪研究(刘红云, 张雷, 2005; Little, Card, Preacher, & McConnell, 2009)。追踪研究可更有效的揭示心理现象的发展变化, 估计一组变量的跨时间关系(cross-time relations); 能够对某种效应随时间表达的过程进行建模; 在多变量增长曲线模型中, 可对动态(双向)关系进行建模; 可提高假设检验的统计功效(Little et al., 2009; Preacher, Wichman, MacCallum, & Briggs, 2008)。得益于这些优势, 近年来, 追踪研究越来越受到各领域研究者的青睐, 特别是发展心理学领域。比如, 2006年在6种主要的发展心理学期刊上发表的348篇研究论文中, 有41% (141篇)的研究采用了追踪设计(Card & Little, 2007)。

信度反映了测量中的误差, 即使一个完美的研究设计也无法弥补不可靠和不精确测量所带来

的缺陷, 因此, 估计测验的信度就成了数据分析的必须前提和关键性步骤(Bierner, Christ, & Wiesen, 2009; Vangeneugden et al, 2010)。如果测验信度的估计方法不当, 可能会高估或低估变量间的真实关系, 直接影响基于信度基础上所做的统计分析的结果(Lachin, 2004; Laenen, Alonso, Molenberghs, & Vangeneugden, 2009a)。同横断研究一样, 信度估计也是衡量追踪研究结论可靠性的重要信息。

如何估计追踪研究的测验信度呢? 目前, 一部分研究者使用传统的信度估计方法, 用经典测量理论框架下的 α 系数或用概化理论框架下的G系数估计信度; 还有研究者根据追踪研究的特点, 基于体现追踪数据特点的数学模型(包括单纯形模型、概化单纯形模型、线性混合模型), 提出相应的信度估计, 包括 ρ_w , $\rho(S_w)$, R_T , R_A 。问题是, 这些信度估计孰优孰劣? 应用工作者应该报告追踪研究的哪个信度? 本研究试图回答以上问题, 在述评传统的追踪测验信度估计及其局限的基础上, 介绍并比较4种估计追踪测验信度的原理与技术。

2 估计追踪研究测验信度的传统方法

目前, 在学术界就如何准确地估计追踪研究的测验信度还未取得一般共识的情况下, 一部分研究者沿用横断研究中的测验信度估计思路, 使

收稿日期: 2011-05-10

* 国家自然科学基金项目(30870784)、教育部人文社科重点研究基地项目(11JJD190005)和教育部人文社会科学研究项目基金(09YJCXLX007)资助。

通讯作者: 温忠麟, E-mail: wenzl@scnu.edu.cn

用经典测量理论下的 α 系数和概化理论下的G系数来估计追踪研究的信度。

2.1 用 α 系数估计追踪测验的信度

使用基于经典测量理论的 α 系数估计追踪研究的测验信度,通常有两种方法:(1)计算追踪研究中各个时间点测验的 α 系数,用多个 α 系数作为追踪研究测验信度的估计(例如, de Clercq, van Leeuwen, van den Noortgate, de Bolle, & de Fruyt, 2009; Suárez-Orozco et al, 2010; Umaña-Taylor & Guimond, 2010);(2)先计算追踪研究中各个时间点测验的 α 系数,再计算各 α 系数的算术平均值,用一个单一的 α 系数均值作为追踪研究测验信度的估计(例如, Crawford, Cohen, Chen, Anglin, & Ehrensaft, 2009)。

目前虽然研究者广泛使用 α 系数作为测验信度的估计,但是测验学界对 α 系数能否用来估计测验信度有很大的争论。尽管这些争论主要是针对横断研究的,但 α 系数固有的不足在追踪研究中同样存在。比如,当测验是基本 τ 等价且误差不相关时(Novick & Lewis, 1967), α 系数等于测验的信度,否则 α 系数估计的信度可能不准确(例如, Green & Yang, 2009; Raykov, Dimitrov, & Asparouhov, 2010; Revelle & Zinbarg, 2009; 温忠麟, 叶宝娟, 2011; Yang & Green, 2010)。实际研究中,无论横断研究还是追踪研究,测验都不会是基本 τ 等价的。此外, α 系数是一种相关系数,其本身的度量层次是定序的,而不是定距的,不可直接求平均数,因此上面方法(2)用 α 系数的均值作为追踪研究中的测验信度是不妥当的。

更重要的是,信度并非是测验的固有特性,而是对测验结果中测量误差的衡量(Hogan, Benjamin, & Brezinski, 2000; Laenen et al., 2009a),因此,不能认为信度是一个固定的数值,它会随施测情景的变化而变化(Laenen, Vangeneugden, Geys, & Molenberghs, 2006)。经典测验信度是在横断研究的框架下定义的,每个被试只在一个时间点施测,将信度定义拓展到追踪研究中时,应当考虑追踪数据的特定特性:被试随时间会有系统的变化,体现为被试特定的相关效应,经常会有序列相关和异质方差成分(Verbeke & Molenberghs, 2000, 2009),经典测验信度的定义没有考虑追踪数据的这种特性,因此不适合用来估计追踪研究测验的信度(Laenen et al., 2009a;

Laenen, Alonso, Molenberghs, & Vangeneugden, 2009b)。简言之, α 系数是在经典测验理论的框架下针对横断研究定义的,盲目使用 α 系数估计追踪研究测验的信度是不恰当的,可以肯定经常会导致测验信度估计偏差(Bierner et al., 2009)。

2.2 用G系数估计追踪测验的信度

概化理论(generalizability theory)能够同时考虑各种来源的变异,一些研究者用概化理论中的G系数估计追踪研究测验信度,将时间点作为其中的一个测量侧面,进而估计追踪研究测验的信度(例如, Bost, 1995; Vangeneugden, Molenberghs, Laenen, Alonso, & Geys, 2008)。用概化理论估计追踪研究的测验信度需要有很强的假设(Verbeke & Molenberghs, 2009; Molenberghs & Kenward, 2007),如误差不相关,随机效应不相关,方差随时间不变等,如果违背这些前提假设,可能会误估G系数,而这些假设在追踪研究中很难满足,此时用G系数估计信度可能不准确(Laenen et al., 2009a; Laenen, Alonso, Molenberghs, Vangeneugden, & Mallinckrodt, 2009)。Bost (1995)通过模拟研究发现,如果不同时间点的误差相关,则G系数会高估或低估测验的信度。因此,不要用G系数估计追踪测验的信度。

总之,在经典测验理论和概化理论中定义的传统信度估计方法(α 系数,G系数),其前提假设在追踪研究中难以满足,不适合用于估计追踪研究中的测验信度。

3 估计追踪测验信度的新方法

追踪研究中涉及的测试时间点、情境较多,因此追踪研究的测验数据中往往会有多个来源的变异和相关,这给定义和估计其信度带来了挑战(Laenen et al., 2009a, 2009b; Laenen, Alonso, Molenberghs, Vangeneugden, & Mallinckrodt, 2010)。根据追踪研究的特点,近年来研究者逐渐发展了估计追踪研究测验信度的新方法,有些研究者从追踪研究中单个时间点的角度探讨追踪测验的信度估计;有些研究者则从整个追踪研究的角度探讨追踪测验的信度估计。

3.1 单个时间点上的测验信度估计

估计追踪研究单个时间点测验信度,需借助结构方程技术,既可以用单纯形模型(simplex model, Alwin, 2007; Bierner et al., 2009),在该模

型下估计追踪测验的信度得到信度系数 ρ_w ;也可以使用概化单纯形模型(generalized simplex model, Biemer et al., 2009),在该模型下估计追踪测验的信度得到信度系数 $\rho(S_w)$ 。

3.1.1 单纯形模型和 ρ_w 信度系数

Alwin (2007)论述了单纯形模型的数学原理与假设,单纯形模型由测量方程和结构方程两部分组成,对时间点 $w(w=1,2,\dots,W)$ 的测量方程如下:

$$S_w = T_w + E_w \quad (1)$$

其中, S_w, T_w, E_w 分别是时间点 w 的全部题目的观测分数之和、真分数之和、误差之和,简称为(整份测验的)观测分数、真分数、误差。真分数的结构方程为:

$$T_w = \beta_{(w-1)w} T_{w-1} + \zeta_w \quad (2)$$

其中, $T_0=0, \beta_{01}=0, \beta_{(w-1)w}$ 是从时间点 $w-1$ 到时间点 w 的真分数的变化, ζ_1 表示时间点1的真分数, $\zeta_w(w>1)$ 是随机误差,表示真分数随时间点变化的偏差。

模型假设当 $w \neq w' (w, w' = 1, 2, \dots, W)$ 时,有:

$$E(E_w) = 0; \text{cov}(E_w, E_{w'}) = 0;$$

$$\text{cov}(E_w, T_{w'}) = 0; \text{cov}(\zeta_w, T_{w'}) = 0 \quad (3)$$

为使模型可识别,需假设:(1)误差方差或真分数的方差随时间不变;(2)追踪研究至少有3个时间点。

Biemer 等人(2009)基于单纯形模型而提出来的一个信度估计方法,在单纯形模型条件下单个时间点测验的信度 ρ_w 为:

$$\rho_w = \frac{\sigma_{T_w}^2}{\sigma_{T_w}^2 + \sigma_{E_w}^2} \quad (4)$$

其中, ρ_w 表示时间点 w 的测验信度, $\sigma_{T_w}^2$ 和 $\sigma_{E_w}^2$ 分别表示时间点 w 的真分数方差和误差方差。 ρ_w 用被试测验总分估计信度,估计的信度不受题目间误差相关的影响(Biemer et al., 2009)。

ρ_w 信度系数有两个不足(Biemer et al., 2009):

- (1)如果违背前提假设,估计的信度可能不准确,实际施测时可能经常违背其前提假设,如不同时间点施测的间隔较短,不同时间点的误差会相关;
- (2)固定不同时间点的误差方差相等和真分数方差相等得到的信度可能不同,用两种方法得到的信度估计值差别最大达0.20。

3.1.2 概化单纯形模型及 $\rho(S_w)$ 信度系数

如果误差方差和真分数方差都随时间变化,公式(4)不再适合估计追踪研究测验的信度,为克服

服 ρ_w 的局限, Biemer 等人(2009)提出了用概化单纯形模型估计追踪研究中单个时间点测验的信度 $\rho(S_w)$ 。与单纯形模型不同的是,概化单纯形模型在每个时间点将整个测验平分成两个分测验,有关真分数和误差的假设参见公式(3),模型的测量方程为:

$$S_{w1} + S_{w2} = T_w + E_w \quad (w=1, 2, \dots, W) \quad (5)$$

其中, S_{w1}, S_{w2} 表示两个分测验在时间点 w 的观测分数, T_w, E_w 分别是整份测验在时间点 w 的真分数和误差。模型真分数的结构方程同公式(2)。

该模型假设:(1)在每个时间点上,两个分测验的误差的协方差不随时间变化;(2)同一时间点两个测验间的真分数方差相等($\text{var}(T_{w1}) = \text{var}(T_{w2}) = \sigma_{T_w}^2$);(3)结构的测量不随时间变化,即每个时间点的观测分数与真分数之间的线性关系相同。

Biemer 等人(2009)在此模型下推出追踪研究中单个时间点测验的信度估计为:

$$\rho(S_w) = \frac{\sigma_{T_w}^2}{\sigma_{T_w}^2 + \sigma_{12} + \frac{\sigma_{E_w}^2}{2}} \quad (6)$$

其中, $\rho(S_w)$ 表示时间点 w 测验的信度, $\sigma_{T_w}^2$ 表示分测验在时间点 w 的真分数方差, $\sigma_{E_w}^2$ 表示时间点 w 的误差方差, σ_{12} 是两个分测验间误差的协方差(在任一时间点相同)。

$\rho(S_w)$ 信度系数有两点不足:(1)如果违背前提假设,估计的信度可能不准确。实际施测时可能经常违背其前提假设,如结构的测量可能随时间变化,或 σ_{12} 可能随时间变化,此时估计的信度有偏差;(2)如果将测验分成两个分测验的方式不同,则估计的信度可能不同。

3.2 整个追踪研究测验信度的估计

ρ_w 和 $\rho(S_w)$ 都只估计了单个时间点测验的信度,而没有给出整个追踪研究测验的信度。有研究者从估计整个追踪研究测验信度的角度,在线性混合模型(linear mixed model)的基础上发展出两个信度系数: R_A 系数与 R_T 系数。

3.2.1 线性混合模型

线性混合模型可以区分三种变异:被试间变异,序列相关变异和误差变异,被试随机效应解释了被试间的异质性;序列相关解释了相近时间点上测量的测量有更强的相关(Verbeke & Molenberghs, 2000, 2009)。该模型处理追踪数据非常有优势:真分数可以变化,误差可以相关,序列相关可以有不同类型,

可以有异质误差成分,可以有相关的随机效应等(Molenberghs, Laenen, & Vangeneugden, 2007; Verbeke & Molenberghs, 2000, 2009)。

设被试 i ($i=1,2,\dots,N$) 有 p_i 次观测, Y_i 是其观测分数组成的 p_i 维向量, 线性混合模型如下:

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_{(1)i} + \varepsilon_{(2)i} \quad (i=1,2,\dots,N) \quad (7)$$

其中, X_i 和 Z_i 是协变量的 $p_i \times q$ 和 $p_i \times r$ 已知的观测矩阵; β 是 q 维固定效应向量, b_i 是 r 维包含随机效应的向量, $b_i \sim N(0, D)$, D 是 b_i 的 $r \times r$ 的协方差矩阵; $\varepsilon_{(1)i}$ 是 p_i 维残差向量, $\varepsilon_{(1)i} \sim N(0, \Sigma_{Ri})$, Σ_{Ri} 是 $\varepsilon_{(1)i}$ 的 $p_i \times p_i$ 协方差矩阵; $\varepsilon_{(2)i}$ 是 p_i 维的序列相关成份组成的向量, $\varepsilon_{(2)i} \sim N(0, \tau^2 H_i)$, τ^2 是方差参数, H_i 是 $\varepsilon_{(2)i}$ 的 $p_i \times p_i$ 相关矩阵; $b_1, \dots, b_N$, $\varepsilon_{(1)1}, \dots, \varepsilon_{(1)N}$, $\varepsilon_{(2)1}, \dots, \varepsilon_{(2)N}$ 相互独立。而且, Σ_{Ri} 和 H_i 只是维数与 i 有关, 参数值与 i 无关。记 $V_i = \Sigma_{Di} + \Sigma_i$, 其中 $\Sigma_{Di} = Z_i D Z_i'$ 是被试间变异(即真分数变异), $\Sigma_i = \Sigma_{Ri} + \tau^2 H_i$ 包含所有其他来源的变异(即误差变异), 包括残差变异和序列相关。这样, 整个观测分数变异分为两部分: 真分数变异和误差变异。

3.2.2 R_T 系数估计方法

在线性混合模型的基础上, 一个测验重复测量的方差-协方差矩阵的迹(trace)可概括测验重复测量的变异, Laenen, Alonso 和 Molenberghs (2007) 基于公理定义了一个追踪研究测验的信度系数 R_T , 公式如下:

$$R_T = \frac{1}{N} \sum_{i=1}^N \frac{\text{tr}(V_i) - \text{tr}(\Sigma_i)}{\text{tr}(V_i)} \quad (8)$$

其中, $\text{tr}(A)$ 表示方阵 A 的迹, $\text{tr}(V_i)$ 表示被试 i 的总变异, $\text{tr}(\Sigma_i)$ 是被试 i 的误差变异。

在平衡设计中, 假定对所有被试有:

$\Sigma_{Di} = \Sigma_D$ 且 $\Sigma_i = \Sigma$, 则 R_T 为

$$R_T = 1 - \frac{\text{tr}(\Sigma)}{\text{tr}(V)} \quad (9)$$

可将 R_T 理解为测验区分被试的能力, 如果被试间变异解释了大部分总变异, 则差异主要由被试间的差异引起, 这样的测验有高的区分能力, 认为可信(Laenen et al., 2007)。

3.2.3 R_A 系数估计方法

在成长性测量中测量的准确性和信度是时间点数目的单调增函数(Preacher et al., 2008), 对一个被试观测的时间越久, 对其做的相关结论就更可靠, 上述的单个时间点的信度或 R_T 并不支持这

种假设, Laenen 等人(2009b)定义了一个新的信度系数 R_A , 可以支持这种假设。

在线性混合模型基础上, 一个测验重复测量的方差-协方差矩阵的行列式可概括测验重复测量的变异, 在平衡设计中, Laenen 等人(2009b)基于公理定义了另一个追踪研究测验的信度系数 R_A :

$$R_A = 1 - \left| \Sigma V^{-1} \right| \quad (10)$$

其中, $|A|$ 表示矩阵 A 的行列式, V^{-1} 表示矩阵 V 的逆矩阵。

3.2.4 信度系数 R_T 和 R_A 的特性

基于线性混合模型的两个信度系数 R_T 和 R_A 满足以下四个特性: 取值范围为(0, 1); 当且仅当只有测量误差时, 其值为 0; 当且仅当没有测量误差, 其值为 1; 在横断研究中其值等于真分数的方差与观测分数方差的比率。这两个信度系数是基于公理而非相关的角度推导出来的, 利用每个时间点的被试信息, 可估计单个时间点测验的信度, 评价单个时间点的测量质量, 也可估计整个追踪研究测验的信度, 评价整个追踪研究的测量质量。 R_A 和 R_T 基于线性混合模型, 独立于对数据进行后续统计分析的模型, 易于理解和应用(Laenen et al., 2007; Laenen et al., 2009a)。

虽然有许多相同的特性, 两个信度系数有两点不同: (1) R_T 是追踪研究中测验重复测量的平均信度(average reliability), 增加时间点的次数, R_T 不会有规律的变化, 既可增大也可减小, 增加低信度的时间点(测量的真分数变异与误差变异的比率比之前各个时间点的平均比率小), R_T 减小, 反之, R_T 增大(Laenen & Alonso, 2010; Laenen et al., 2009b); R_A 随着测量次数的增多而增大, 增加时间点的次数, 会增加有关真分数的总体信息, 增加对被试有用的信息(即使受到测量误差的影响), R_A 会增大, 当时间点无穷多时, R_A 接近 1, 当测量误差很大时, 如果重复的次数足够多, 仍可得到较高的 R_A (Laenen et al., 2009b)。 (2) R_T 只用到方差-协方差矩阵中的方差, 未用协方差, 没有用整个追踪研究中的所有信息; R_A 用到方差-协方差矩阵中的方差和协方差, 用了整个追踪研究中的所有信息, 可以看作是整个追踪研究的信度(total reliability, Laenen et al., 2009a, 2009b; Laenen et al., 2010; Laenen & Alonso, 2010)。

追踪研究中, 经常会有序列相关, 而 R_A 和 R_T 所基于的线性混合模型有序列相关项 $\varepsilon_{(2)i}$, 保证了 R_A 和 R_T 能准确的估计追踪研究中测验的信度。Laenen 等人(2010)的模拟研究表明, 如果被试内有序列相关, 若在计算 R_A 和 R_T 系数时忽略被试内的序列相关, 此时得到的 R_A 和 R_T 会误估追踪研究测验的信度, 因此, 在估计追踪研究测验的信度时, 考虑序列相关是十分重要的, 如果忽略序列相关, 即使用 R_A 和 R_T 估计的信度可能也不准确。

3.3 追踪研究四种新信度估计方法的比较

最近提出的估计追踪研究测验信度的四种新系数, 即 ρ_w , $\rho(S_w)$, R_T , R_A , 既有相似之处, 也有不同之处。它们的相似点, 都是由经典测量理论的信度估计思想发展而来, 将测验分数的方差进行分解, 用真分数的方差与观测分数的方差之比求得测验信度; 在满足所基于的前提的情况下, 可以较为准确的估计追踪研究中测验的信度。

然而这几个信度估计系数也有不同: (1)前提假设不同。 ρ_w , $\rho(S_w)$ 信度的前提假设较多, 如果违背这些假设, 可能会使信度估计出现偏差, 实际施测时可能经常违背这些假设; 而 R_T , R_A 信度的前提假设少, 比较适用于追踪研究的实际情况, 可以较为准确估计追踪研究的测验信度。(2)信度估计范围不同。 ρ_w , $\rho(S_w)$ 仅能估计单个时间点测验的信度, 信度在不同的时间点可能不同, 一个追踪研究可能得到多个测验信度值, 而且不能估计整个追踪研究测验的信度。 R_T , R_A 系数既可估计追踪研究中单个时间点测验的信度, 也可估计追踪研究中整个追踪研究测验的信度, 也就是说, R_T , R_A 能反映整个追踪测验信度的一致性和稳定性, 提供更多信息; (3)时间顺序效应不同。由于时间点不同, 单个时间点的 ρ_w , $\rho(S_w)$ 可能会不同, 会得到多个信度值, 如果使用平均值的方法, 则可能掩盖时间顺序效应; R_A 的值随着测量次数的增多而增大, 其增加的数值反映了整个追踪研究中增加的真分数的信息, 而 R_T 随着测量次数的增多, 其值不一定增加, R_T 反映的是整个追踪研究的平均信度。因此, 可同时使用信度系数 R_T 和 R_A 来估计追踪研究测验的信度, 能较好的反映追踪研究测验的信度。(4)处理方差的方式不同。 ρ_w , $\rho(S_w)$ 是用方差的比率求得的, 而 R_T , R_A 是用迹和行列式概括方差进而求得比率得来的, 在多变量分析中, 一个随机向量的概化方差可以用相应的方差-

协方差矩阵的迹或行列式来定义(Laenen et al., 2007; Laenen et al., 2009a, 2009b; Alonso, Laenen, Molenberghs, Geys, & Vangeneugden, 2010)。(5)用于计算的模型不同。 ρ_w , $\rho(S_w)$ 是在结构方程模型的基础上推演出来的, 因结构方程模型是基于协方差理论的, 所需的样本容量较大, 并且只有当结构方程模型拟合好, 计算的参数才有意义, 从而计算的 ρ_w , $\rho(S_w)$ 才有意义。而 R_T , R_A 是在线性混合模型的基础上推演出来的, 与结构方程模型相比, 线性混合模型对样本容量的要求相对宽松。

如果研究者要用 ρ_w , $\rho(S_w)$ 系数估计追踪研究的测验信度, 样本要尽量大, 并在计算 ρ_w , $\rho(S_w)$ 前应当检视模型的拟合情况, 只有模型拟合好, 才去计算 ρ_w , $\rho(S_w)$, 否则, 改用其他信度系数估计追踪研究中测验的信度。而计算 R_T , R_A 的关键是确定线性混合模型的随机效应及固定效应。 R_T , R_A 的前提假定少, 并且 R_T 反映整个追踪研究的平均测验信度, R_A 反映整个追踪研究中真分数的信息, 推荐同时使用 R_T , R_A 来评价追踪研究的测验信度。

4 用四种信度系数估计追踪研究测验信度的例子

一项追踪研究考察利哌酮(risperidone)对精神分裂症(schizophrenia)治疗效果, 有 453 名被试参与此项追踪研究, 用简明精神病量表(Brief Psychiatric Rating Scale, BPRS)评定治疗效果, BPRS 有 18 个项目, 采用 1~7 分的 7 级评分法。量表共施测 6 次, 第一次施测后, 分别间隔 1 周, 2 周, 4 周, 6 周和 8 周再次进行施测(参见 Laenen et al., 2009b)。下面分别用 R_T , R_A , ρ_w 和 $\rho(S_w)$ 系数估计这次追踪研究中 BPRS 的信度。

根据 Verbeke 和 Molenberghs (2000, 2009)模型构建准则, 选择与数据拟合最好的协方差结构。有关线性混合模型协方差结构的细节及例子, 读者可参考 Verbeke 和 Molenberghs (2000, 2009)的著作。选取 AIC (Akaike's information criterion) 最小的模型, 并采用 REML (restricted maximum likelihood)估计参数。

在此基础上, 根据公式(9)和(10)求得 BPRS 的 R_T , R_A 分别为 0.821 和 0.991 (计算过程和计算程序可在网址 <http://www.ibiostat.be/software/measurement.asp> 查到)。固定不同时间点的误差方

差相等用公式(4)求得6次施测的BPRS的 ρ_w 分别为0.841、0.836、0.830、0.824、0.817、0.805,固定不同时间点真分数方差相等用公式(4)求得的6次施测的BPRS的 ρ_w 分别为0.802、0.813、0.821、0.827、0.834、0.843。不难看出本例用固定不同时间点的误差方差相等和真分数方差相等得到的 ρ_w 不同,差别最大的为0.039。用固定不同时间点的误差方差相等求得6次施测的BPRS的 ρ_w 呈递减趋势,即越晚施测的BPRS的信度越小;而用固定不同时间点真分数方差相等求得6次施测的BPRS的 ρ_w 呈递增趋势,即越晚施测的BPRS的信度越大。用公式(6)求得的6次施测的BPRS的 $\rho(S_w)$ 分别为0.846、0.835、0.831、0.823、0.814、0.801, $\rho(S_w)$ 呈递减趋势,即越晚施测的BPRS的信度越小。各个时间点的 $\rho(S_w)$ 与固定不同时间点的误差方差相等或真分数方差相等得到的 ρ_w 都不同。

通过本文的示例也可以看出,本次追踪研究中6次施测得到BPRS的6个 $\rho(S_w)$ 系数,12个 ρ_w 系数(固定误差方差和真分数方差各6个 ρ_w),因此,用 $\rho(S_w)$ 和 ρ_w 很难对整个追踪研究的BPRS作出一个整体的评价,而用 R_T 、 R_A 可以对整个追踪研究的BPRS作出一个整体的评价,因此,推荐使用 R_T 、 R_A 估计追踪研究中测验/量表信度。

5 总结

使用 α 系数和概化系数等传统的信度估计方法估计追踪研究测验信度,具有较多的局限性。在追踪研究下, α 系数和概化系数的前提假设很难满足;而且在追踪研究中,随着时间和测验情境变化会产生序列相关,而 α 系数和概化系数无法分离这些序列相关。因此不要用 α 系数和概化系数估计追踪研究测验信度。

四种追踪研究的信度系数 ρ_w 、 $\rho(S_w)$ 、 R_T 、 R_A ,都用到了各个时间点的信息,如果前提假设都可满足,比传统的信度估计方法能更好的估计追踪研究中测验的信度。在追踪研究中,每次测量都会带来额外的信息;总体的信息要比单个时间点上的信息更可靠和丰富全面;在分析、理解和解释时,使用一个整体信度要比使用多个随时间变化的信度更好,有助于比较和解释两个或多个追踪研究结果的可靠性和稳定性,有助于理解其中的心理测量特性(Laenen et al., 2009b)。与 ρ_w 、 $\rho(S_w)$

比较, R_T 、 R_A 是估计追踪研究的测验信度较好的方法, R_T 反映整个追踪研究的平均测验信度, R_A 反映整个追踪研究中真分数的信息,可以用来衡量整个追踪研究的测验信度,推荐同时使用两个信度系数来评价追踪研究的测验信度。

对测验信度的估计都是基于模型的,其范围和适用性局限于其模型的范围和适用性(Laenen et al., 2009a, 2009b; Laenen et al., 2010), R_T 、 R_A 是基于线性模型基础上推导出来的,只适用于追踪数据之间是线性关系的情景。如果追踪数据之间的关系是非线性的,用 R_T 、 R_A 估计追踪研究测验的信度可能会不准确。借鉴推导 R_T 、 R_A 的思路,估计非线性数据的测验的信度时可以参考如下思路:首先,选定能够表达追踪数据特点的非线性模型,然后,用迹(或行列式)表示测验分数的方差-协方差的变异,最后,用真分数方差的迹(或行列式)与观测分数方差的迹(或行列式)的比率推导出测验的信度公式。此外, R_A 是在平衡设计下推导出来的,其在非平衡设计中的表现如何呢?这些均是今后研究中有待解决的问题。

参考文献

- 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.
- 温忠麟, 叶宝娟. (2011). 测验信度的估计: 从 α 系数到内部一致性信度. *心理学报*, 43, 821-829.
- Alonso, A., Laenen, A., Molenberghs, G., Geys, H., & Vangeneugden, T. (2010). A unified approach to multi-item reliability. *Biometrics*, 66, 1061-1068.
- Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. Hoboken, NJ: Wiley.
- Biemer, P. P., Christ, S. L., & Wiesen, C. A. (2009). A general approach for estimating scale score reliability for panel survey data. *Psychological Methods*, 14, 400-412.
- Bost, J. E. (1995). The effects of correlated errors on generalizability and dependability coefficients. *Applied Psychological Measurement*, 19, 191-203.
- Card, N. A., & Little, T. D. (2007). Longitudinal modeling of developmental processes. *International Journal of Behavioral Development*, 31, 297-302.
- Crawford, T. N., Cohen, P. R., Chen, H., Anglin, D. M., & Ehrensaft, M. (2009). Early maternal separation and the trajectory of borderline personality disorder symptoms. *Development and Psychopathology*, 21, 1013-1030.
- de Clercq, B., van Leeuwen, K., van den Noortgate, W., de Bolle, M., & de Fruyt, F. (2009). Childhood personality pathology: Dimensional stability and change.

- Development and Psychopathology*, 21, 853–869.
- Green, S. B., & Yang, Y. Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, 74, 121–135.
- Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, 60, 523–531.
- Lachin, J. M. (2004). The role of measurement reliability in clinical trials. *Clinical Trials*, 1, 553–566.
- Laenen, A., & Alonso, A. (2010). The functional living index-cancer: Estimating its reliability based on clinical trial data. *Quality of Life Research*, 19, 103–109.
- Laenen, A., Alonso, A., & Molenberghs, G. (2007). A measure for the reliability of a rating scale based on longitudinal clinical trial data. *Psychometrika*, 73, 443–448.
- Laenen, A., Alonso, A., Molenberghs, G., & Vangeneugden, T. (2009a). A family of measures to evaluate scale reliability in a longitudinal setting. *Journal of the Royal Statistical Society*, 172, 237–253.
- Laenen, A., Alonso, A., Molenberghs, G., & Vangeneugden, T. (2009b). Reliability of a longitudinal sequence of scale ratings. *Psychometrika*, 74, 49–64.
- Laenen, A., Alonso, A., Molenberghs, G., Vangeneugden, T., & Mallinckrodt, C. (2009). Using longitudinal data from a clinical trial in depression to assess the reliability of its outcome scales. *Journal of Psychiatric Research*, 43, 730–738.
- Laenen, A., Alonso, A., Molenberghs, G., Vangeneugden, T., & Mallinckrodt, C. (2010). Coping with memory effect and serial correlation when estimating reliability in a longitudinal framework. *Applied Psychological Measurement*, 34, 255–266.
- Laenen, A., Vangeneugden, T., Geys, H., & Molenberghs, G. (2006). Generalized reliability estimation using repeated measurements. *British Journal of Mathematical and Statistical Psychology*, 59, 113–131.
- Little, T. D., Card, N. A., Preacher, K. J., & McConnell, E. (2009). Modeling longitudinal data from research on adolescence. In R. M. Lerner & L. D. Steinberg (Eds.), *Handbook of adolescent psychology* (3rd ed., Vol. 1, pp. 15–54). Hoboken, NJ: John Wiley & Sons.
- Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. Chichester: Wiley.
- Molenberghs, G., Laenen, A., & Vangeneugden, T. (2007). Estimating reliability and generalizability from hierarchical biomedical data. *Journal of Biopharmaceutical Statistics*, 17, 595–627.
- Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32, 1–13.
- Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Los Angeles: Sage.
- Raykov, T., Dimitrov, D. M., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling*, 17, 265–289.
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74, 145–154.
- Suárez-Orozco, C., Gaytán, F. X., Bang, H. J., Pakes, J., O'Connor, E., & Rhodes, J. (2008). Academic trajectories of newcomer immigrant youth. *Developmental Psychology*, 46, 602–618.
- Umaña-Taylor, A. J., & Guimond, A. B. (2010). A longitudinal examination of parenting behaviors and perceived discrimination predicting Latino adolescents' ethnic identity. *Developmental Psychology*, 46, 636–650.
- Vangeneugden, T., Molenberghs, G., Laenen, A., Alonso, A., & Geys, H. (2008). Generalizability in nongaussian longitudinal clinical trial data based on generalized linear mixed models. *Journal of Biopharmaceutical Statistics*, 18, 691–712.
- Vangeneugden, T., Molenberghs, G., Laenen, A., Geys, H., Beunckens, C., & Sotto, C. (2010). Marginal correlation in longitudinal binary data based on generalized linear mixed models. *Communications in Statistics-Theory and Methods*, 39, 3540–3557.
- Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.
- Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data* (2nd ed.). New York: Springer.
- Yang, Y., & Green, S. B. (2010). A Note on structural equation modeling estimates of reliability. *Structural Equation Modeling*, 17, 66–81.

Estimating Test Reliability of a Longitudinal Study

YE Bao-Juan; WEN Zhong-Lin; CHEN Qi-Shan

(Center for Studies of Psychological Application, South China Normal University, Guangzhou 510631, China)

Abstract: Test reliability of a longitudinal study is an important index in evaluating the quality of the longitudinal study. Conventional methods that estimate test reliability are not appropriate to estimate the test reliability of a longitudinal study. In recent years, four methods for estimating the test reliability of a longitudinal study were proposed. ρ_w and $\rho(S_w)$ coefficients were designed to estimate the test reliability of a longitudinal study at each wave, whereas R_T and R_A were designed to estimate the test reliability of the whole longitudinal study. The mathematical models and assumptions of the four reliability coefficients were briefed and their merits and demerits were pointed out in this article. R_T and R_A were recommended because they can be used to estimate the test reliability of a longitudinal study both at each wave and as the whole under a few assumptions.

Key words: longitudinal study; reliability; R_T coefficient; R_A coefficient