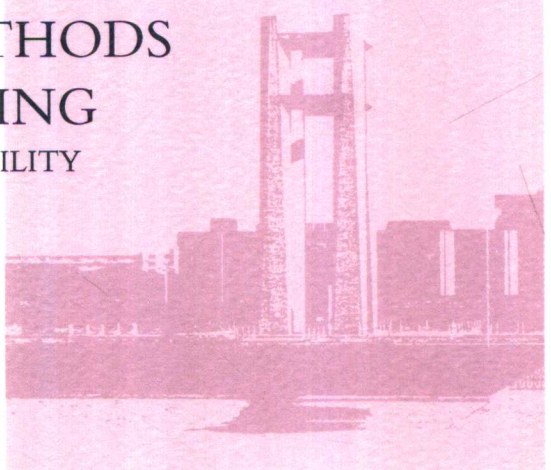
JIANGXI NORMAL UNIVERSITY

# 经典测验信度的现代估计方法

MODERN METHODS
FOR ESTIMATING
CLASSICAL TEST RELIABILITY



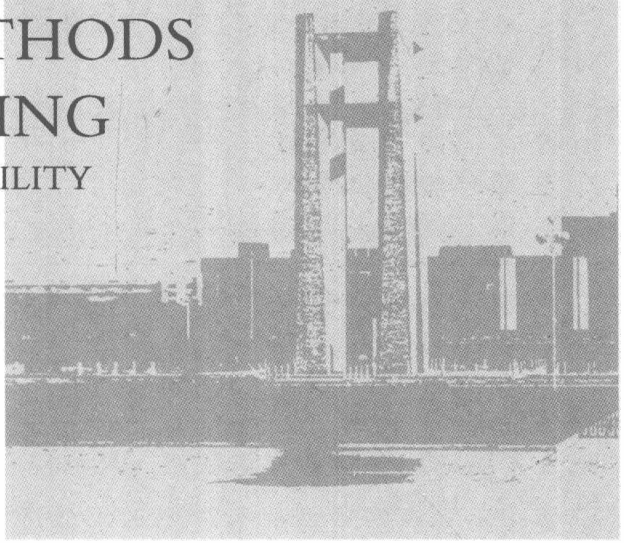
叶宝娟 著

中国社会科学出版社

JIANGXI NORMAL UNIVERSITY

# 经典测验信度的现代估计方法

# MODERN METHODS FOR ESTIMATING CLASSICAL TEST RELIABILITY



叶宝娟 著

中国社会科学出版社

**图书在版编目(CIP)数据**

经典测验信度的现代估计方法／叶宝娟著．—北京：中国社会科学出版社，2014.11

ISBN 978-7-5161-5054-2

Ⅰ.①经… Ⅱ.①叶… Ⅲ.①信度-估计 Ⅳ.①B841.7

中国版本图书馆CIP数据核字(2014)第262046号

出 版 人 赵剑英
责任编辑 宫京蕾
责任校对 张依婧
责任印制 何 艳

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲158号（邮编100720）
网 址 http：//www.csspw.cn
中文域名：中国社科网 010-64070619
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印 刷 北京市兴怀印刷厂
版 次 2014年11月第1版
印 次 2014年11月第1次印刷

开 本 710×1000 1/16
印 张 14.25
插 页 2
字 数 235千字
定 价 45.00元

凡购买中国社会科学出版社图书，如有质量问题请与本社联系调换
电话：010-64009791

# α系数：晃而不倒的信度标杆[①]（代序）

自1951年Cronbach（1916—2001）在*Psychometrika*上发表了那篇讨论α系数的论文以后，α系数逐渐成为心理和教育测验工作者的一根标杆，用来估计测验信度。世界上心理测验的编制、修订和使用，几乎都会报告α系数，这已经成为一种常识（也可以说是一种仪式）。在近代科学中，至少在心理学科中，恐怕不容易找到其他专业论文及其推荐的公式，有如此长久且广泛的影响力。

## α系数的身世

1945年，以Guttman量表闻名于世的心理测量学家Louis Guttman在*Psychometrika*上发表了研究信度的文章，推出6个λ系数（记为$\lambda_1$—$\lambda_6$），它们都是信度的下界。其中的$\lambda_3$，被Cronbach相中，从此走红。Cronbach将其命名为α系数，应当是希望有更好的信度系数出现，可以接着用后面的希腊字母命名。后来还真的有研究者提出$\beta$系数、$\gamma$系数等，它们在某些方面的表现优于α系数，但没有一个受到重视，连昙花一现也只是小范围的事，更别说流行了。其实，Guttman的$\lambda_2$，就比α系数（即$\lambda_3$）更接近信度（即$\lambda_3 \leqslant \lambda_2$），因而是信

① 原文发表在《中国社会科学报》（2011年10月13日），略微修改并增加“补记”后作为本书代序。

度更好的估计。然而，历史选择了 α 系数。α 系数有什么过人之处，成就其标杆地位呢？

## α 系数的优点

α 系数的优点可以简单地概括为公式简单、计算方便、理解容易。

设一份测验由 $k$ 个题目组成，α 系数的计算公式为：

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\text{各题方差之和}}{\text{总分方差}}\right)$$

公式简单，要记住不难，而且其中的方差是很基本的统计量。

由样本计算方差只涉及简单的四则运算和平方，所以在计算机（器）出现之前，α 系数不难用手工计算得到；只需要一次施测，单个样本就可以完成计算；1995 年以后，流行的社科统计软件 SPSS 中用窗口操作很容易计算 α 系数。

在测验题目为偶数时，α 系数是所有可能分半信度的均值，这个简洁的意涵是 α 系数的一个诱人之处。分半信度的背景很直观，但一份试题分半的可能性很多，所以分半信度也很多，摇摆不定，因而不是估计信度的好指标。而 α 系数用来评价信度比任何一个分半信度都更合理。这样，α 系数就有了一个简单的统计背景，容易理解。

## 对 α 系数的美丽误会

信度，是评价一个心理或教育测验质量最重要的指标之一，衡量了测验的一致性和稳定性。在经典测验理论中，信度定义为真分数方差与观测分数方差之比。在有了观测分数以后，不仅真分数是未知的，真分数方差也是未知的，所以无法根据信度的定义给出信度的估计，这看似简单的信度却引出诸多问题的原因。

既然信度不能直接通过样本数据计算，人们就采用别的办法。例如，对同一组被试使用同一份试题在一定间隔时间内施测两次，用前

后两次测验分数计算相关系数，得到所谓重测信度，从测验稳定性角度反映信度；用两份“等值”（内容、题型、题数、难度等都相同或接近）但具体题目又不同的试题，在短期内对同一组被试进行施测，用两份试题所得分数计算相关系数，得到所谓的复本信度，从测验一致性角度反映信度。

α 系数有一个与生俱来的名称——内部一致性系数，给人的感觉是它衡量了测验内部题目之间的一致性，即从题目一致性的角度反映信度。α 系数还有另一个名称——同质性系数，给人的感觉是它衡量了测验题目测量了相同特质的程度。

然而，后来的研究发现，上述关于 α 系数的感觉是不对的。但这些误会以各种方式被写进国内外教科书，客观上巩固了 α 系数的地位。

## α 系数是非不断

相信 Cronbach 本人也想不到 α 系数会有后来的标杆地位，因为他一开始就预留了其他信度系数的位置。在 α 系数流行以后，对它的质疑和评论不断。例如，权威期刊 *Psychometrika* 在 2009 年第 1 期上就有 6 篇文章研究信度，其中 5 篇的标题都包含有 α 系数。

Novick 和 Lewis（1967）研究发现，如果（1）各题的误差不相关（这个条件容易满足）；（2）测验是基本 $\tau$ 等价（这是一个非常强的条件，现实中的测验没有一个能满足），则测验信度等于 α 系数。后来许多研究发现，如果没有对测验附加条件，α 系数既可以小于信度，也可以大于信度。α 系数甚至还会出现负值，此时 α 系数一点信度的影子都没有了。

许多应用工作者都有这样的经验，如果一个测验有多个分测验（多个维度），整个测验的 α 系数通常比各个分测验的 α 系数高，所以 α 系数不能用来评价测验的内部一致性和同质性。关于这一点，早在 20 世纪 70 年代，已经有人撰文讨论。后来有许多模拟研究证明无论一个测验是单维还是多维，都可能有非常低的或非常高的 α 值。这

说明，α 系数高不代表测验是同质的，同质测验的 α 系数不一定高。

## 标杆为何晃而不倒

虽然 α 系数一再受到质疑甚至否定，也有人提出一些别的统计量，试图代替 α 系数。其中，20 世纪 70 年代伴随验证性因子模型而出现、最近十多年被 Tenko Raykov 等人发扬光大的合成信度，对 α 系数构成最大的威胁。

但 α 系数至今只晃不倒。首先，期刊编辑和审稿人都会要求报告测验信度，但没有具体推荐，报告 α 系数还是管用。其次，虽然已经有了更好的信度指标（如合成信度），但这些信度指标比较难理解、更缺乏应用工作者可以接近和容易使用的计算程序，用家只好继续使用 α 系数。

## 新的内部一致性信度和同质性信度

前面说过，α 系数不能用来评价内部一致性和同质性，那应当用什么指标来评价呢？Bentler（2009）认为，合成信度 $\omega_t$ 可以衡量内部一致性；而本书作者和笔者发现，McDonald（1985）的信度 $\omega_h$ 适合用来衡量同质性。

举例来说，设有 10 个题目测量自信，其中有 5 个题目是正向题，5 个题目是反向题。以前，这样的量表是当作单维来分析的，但我们其实可以建立一个三因子模型，其中有一个是全局因子（即自信，影响全部 10 个题目）；两个局部因子（分别影响正向题和反向题，可以看作是两个方法因子）。

如果将全部三个因子引起的变异都当作真分数变异，计算信度就得到内部一致性信度。如果将全局因子引起的变异当作真分数变异，计算信度就得到同质性信度。

## 新的信度双标杆：α 系数是主杆

笔者和本书作者根据最新研究，提出测验信度双标杆，其中 α 系数是主杆，做法如下：

（1）确定要做信度分析的测验（整份测验或者分测验均可），前提是测验的总分有意义。

（2）判断误差是否相关。如果不相关，进行步骤（3a）；否则进行步骤（3b）。

（3a）计算 α 系数，如果 α 系数高到可以接受，报告 α 系数并说明：因为误差不相关，所以测验信度不低于 α 系数；如果 α 系数过低，转到（3b）。

（3b）计算合成信度，如果合成信度高到可以接受，报告合成信度并说明：测验信度不低于合成信度；如果合成信度过低，虽然不能说测验信度也一样低，但在没有其他方法更准确地评价测验信度之前，只能认为测验的信度不能接受，停止进一步的统计分析。

上述做法，是在验证性因子分析视角下得到的，不仅重塑了 α 系数的标杆地位，也让用家清晰地知道，什么时候 α 系数不再适用，应当使用合成信度。对于绝大多数的测验来说，计算并报告 α 系数已经足够。这样，年届花甲的 α 系数，在经久的质疑声中，部分地得到了新的合法性。

## 补　记

对于不同的数据类型，需要不同的信度估计方法。例如，对于多水平数据和追踪数据，不仅 α 系数不合适用来评价信度，连合成信度也不合适了。所以，无论从理论研究的角度，还是从实际应用的角度，本书都很有意义。

本书是国内第一部专门研究测验信度的专著。重点是 α 系数、合成信度和同质性信度的估计方法，特别是区间估计方法，以及这些信

度之间的关系。还讨论了多水平数据、追踪数据的信度估计方法，以及信度元分析方法。通过本书，读者可以更好地理解和估计测验信度，也可以学习到验证性因子模型、双因子模型的应用，还可以学习用 Delta 法估计标准误的做法以及相应的 LISREL 和 MPLUS 程序。

温忠麟

2014 年 3 月于华南师范大学

# 前　言

测验在心理、教育、市场、管理等社科领域被广泛使用。信度是评价测验质量的最重要指标之一。即使完美的研究设计也无法弥补信度低的测量带来的缺陷，因此，估计测验的信度是数据分析的必须前提和关键性步骤。本书以验证性因子分析为主要工具，介绍各种测量条件下合适的信度系数估计方法，并将向测验使用者推荐好的信度估计方法，会极大丰富信度研究，有利于提高社科领域基于测验的量化研究质量。

本书不仅包括了笔者博士论文的研究成果，还介绍了其他研究者的有关成果。第一章是信度概述，简单介绍了信度含义、信度作用、影响信度的因素、信度种类。第二章是传统的 α 系数及置信区间估计方法。第三章是现代的合成信度及置信区间估计方法。第四章是同质性信度及置信区间估计方法。第五章是各种信度之间的关系，揭示了内部一致性信度、合成信度、同质性信度、α 系数等之间的关系。第六章是两水平研究的测验信度及置信区间估计方法。第七章是追踪研究的测验信度估计方法。第八章是信度元分析，介绍了 α 系数和单维测验合成信度元分析的方法。

本书主要有以下特色：

第一，注重理论与实践相结合。本书不仅详细介绍了各种估计信度的方法，而且提供了各种估计信度方法的示例，并提供了简单的 LISREL 和 MPLUS 程序，应用工作者很容易掌握新的信度估计方法。

第二，注重学术前沿性。本书紧跟国际研究前沿，介绍了笔者和合作者的最新研究成果。

第三，适合不同读者群。笔者在撰写过程中，充分考虑了不同的读者群，尽量做到通俗易懂。本书既可以作为一般应用工作者的参考书，也可以作为从事心理测量学专业读者的参考用书。

叶宝娟

2014 年 5 月于江西师范大学

# 目　录

**第一章　信度概述** …… (1)
第一节　信度含义 …… (1)
第二节　信度作用 …… (4)
第三节　影响信度的因素 …… (6)
第四节　信度种类 …… (9)
本章小结 …… (12)

**第二章　α系数** …… (13)
第一节　α系数的点估计 …… (13)
第二节　α系数区间估计方法 …… (16)
第三节　单维测验α系数的区间估计方法比较 …… (22)
一　研究设计 …… (23)
二　研究结果 …… (25)
三　计算α系数置信区间示例 …… (35)
第四节　多维测验α系数的区间估计方法比较 …… (36)
一　研究设计 …… (36)
二　研究结果 …… (37)
本章小结 …… (43)
附录2－1　估计3个题目测验α系数置信区间的SAS程序 …… (44)

**第三章　基于因子模型的合成信度** …… (68)
第一节　单维测验合成信度点估计 …… (68)

第二节 单维测验合成信度区间估计 …… (70)
一 研究设计 …… (72)
二 不同估计方法的置信区间 …… (73)
三 研究结果 …… (74)
第三节 用 SPSS 软件计算单维测验的合成信度 …… (87)
一 用 SPSS 软件计算合成信度 …… (87)
二 应用例子 …… (88)
三 小结 …… (90)
第四节 误差相关的单维测验合成信度区间估计 …… (90)
一 用 Delta 法计算误差相关单维测验合成信度置信区间 …… (91)
二 用 Delta 法计算误差相关单维测验合成信度置信区间示例 …… (93)
第五节 多维测验合成信度点估计 …… (94)
第六节 多维测验合成信度区间估计 …… (95)
一 用 Delta 法估计多维测验合成信度的置信区间 …… (95)
二 用 Delta 法估计多维测验合成信度置信区间示例 …… (96)
三 多维测验合成信度区间估计方法的比较 …… (97)
本章小结 …… (103)
附录 3-1 Bootstrap 取样的 PRELIS 程序 …… (104)
附录 3-2 计算 6 个题目单维测验的合成信度的 LISREL 程序 …… (104)
附录 3-3 计算 6 个题目单维测验的合成信度置信区间 SPSS 程序 …… (105)
附录 3-4 计算 6 个题目单维测验合成信度的 MPLUS 程序 …… (106)
附录 3-5 用 SPSS 软件计算单维测验合成信度 …… (107)
附录 3-6 计算一般单维测验合成信度的 LISREL 程序 …… (107)
附录 3-7 计算一般单维测验合成信度的 MPLUS 程序 …… (108)
附录 3-8 计算 2 因子 3 个题目多维测验合成信度 LISREL 程序 …… (109)

附录3-9　计算2因子3个题目多维测验合成信度的MPLUS程序 …… (110)

**第四章　同质性信度** …… (130)
第一节　同质性信度的点估计 …… (130)
第二节　同质性信度的区间估计 …… (134)
本章小结 …… (138)
附录4-1　求测验同质性信度置信区间的LISREL程序 …… (139)
附录4-2　求测验同质性信度置信区间的MPLUS程序 …… (140)

**第五章　信度之间关系** …… (142)
第一节　内部一致性信度 …… (142)
第二节　信度之间关系 …… (143)
一　内部一致性信度、信度与同质性信度关系 …… (143)
二　内部一致性信度与α系数 …… (145)
第三节　信度分析流程 …… (147)
本章小结 …… (148)
附录5-1　证明：非负定矩阵对角元素均值大于或等于非对角元素均值 …… (149)

**第六章　两水平研究测验信度** …… (151)
第一节　现有的两水平研究中测验信度估计公式的局限 …… (151)
一　以单水平研究的信度估计方法来估计两水平研究的测验信度 …… (152)
二　加上限定条件的两水平研究中的信度公式 …… (152)
第二节　一般的两水平研究中单维测验信度的点估计 …… (155)
第三节　两水平研究中单维测验信度的区间估计 …… (156)
第四节　两水平研究中多维测验信度的点估计 …… (159)
第五节　两水平研究中多维测验信度的区间估计 …… (161)
第六节　两水平研究中组水平上的测验信度的点估计 …… (163)
第七节　两水平研究中组水平上的测验信度的区间估计 …… (165)
本章小结 …… (165)

附录6-1 计算两水平研究中单维测验的信度及其置信区间 …… (166)
附录6-2 计算两水平研究中多维测验的信度及其置信区间 …… (168)

**第七章 追踪研究中信度估计** …… (170)
第一节 估计追踪研究测验信度的传统方法 …… (171)
一 用α系数估计追踪测验的信度 …… (171)
二 用G系数估计追踪测验的信度 …… (172)
第二节 估计追踪测验信度的新方法 …… (173)
一 单个时间点上的测验信度估计 …… (174)
二 整个追踪研究测验信度的估计 …… (176)
三 追踪研究四种新信度估计方法的比较 …… (180)
四 用四种信度系数估计追踪研究测验信度的例子 …… (181)
本章小结 …… (183)

**第八章 信度元分析** …… (184)
第一节 α系数元分析 …… (184)
第二节 合成信度元分析 …… (188)
一 合成信度元分析 …… (188)
二 合成信度元分析区间估计模拟研究 …… (189)
三 对合成信度元分析的示例 …… (192)
本章小结 …… (194)

**参考文献** …… (195)

**后记** …… (212)

# 第一章

# 信 度 概 述

测验在心理、教育、市场、管理等社科领域被广泛使用。信度（reliability）是衡量测验质量的一个重要指标。没有信度的测验毫无用处。[①] 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。[②]

本章介绍信度含义、信度作用、影响信度的因素、信度种类。

## 第一节 信度含义

信度是指测验结果的稳定性（stability）或一致性（consistency）程度，就是说，若能用同一测量工具反复测量某人同一种心理特质，则其多次测量的结果间的一致性程度就叫信度，有时也叫测量

---

① Aiken, R. L., *Psychological Testing and Assessment* (11th ed), Allyn and Bacon Press, 2003.

② Biemer, P. P., Christ, S. L., & Wiesen, C. A., "A General Approach for Estimating Scale Score Reliability for Panel Survey Data", *Psychological Methods*, Vol. 14, No. 4, December 2009, pp. 400 - 412; Vangeneugden, T., Molenberghs, G., Laenen, A. et al., "Marginal Correlation in Longitudinal Binary Data Based on Generalized Linear Mixed Models", *Communications in Statistics - Theory and Methods*, Vol. 39, No. 19, 2010, pp. 3540 - 3557.

的可靠性。[①] 一个具有良好信度的测验，测验结果不会随施测情境的变化而变化，也就是说，在不同的主试、时间、地点等情境下，测验分数应该接近或一致。一个好的测验必须具有较高的信度，也就是说，只要遵守测验施测规则，测验结果就不应随测验使用者或施测时间等方面的变化而发生较大的变化。一个好的测验，就如同一把好的尺子，对同一被试反复测量多次，其结果要一致。一个测验对同一组被试进行多次施测，如果被试的分数忽高忽低，则说明该测验缺乏信度；如果被试分数变化不大，则说明该测验的信度高。如果不是因为成长、学习、疾病等引起的永久性变化，测验分数随着情境的变化而变化，则测验是不可靠的，不能用于描述或评价人以及预测其相关行为。

设一个测验（或量表）共有 $p$ 个题目，第 $j$ 题的（观测）分数 $x_j$ 是一个随机变量，在真分数（true score）理论中，题目分数 $x_j$ 可以分解为题目真分数 $t_j$ 和题目误差 $e_j$ 的和，即 $x_j = t_j + e_j$，其中 $e_j$ 与 $t_i$ 不相关（$i,j = 1,2,\cdots,p$），即一个题目的误差跟任何题目的真分数不相关。整个测验的测验分数（量表分数）记为 $x = \sum_{j=1}^{p} x_j$，相应地，记 $t = \sum_{j=1}^{p} t_j$ 为测验真分数，$e = \sum_{j=1}^{p} e_j$ 为测验误差，则 $x = t + e$。注意到误差与真分数不相关，所以测验分数方差等于真分数方差与误差方差之和，即 $\sigma_x^2 = \sigma_t^2 + \sigma_e^2$，其中 $\sigma_x^2 = \mathrm{var}(x)$，其余符号类推。

在经典测验理论中，测验信度定义为真分数方差与测验分数方差之比，即测验分数方差中真分数方差所占的比例：[②]

$$\rho_x = \frac{\sigma_t^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2} \tag{1-1}$$

由式（1-1）可以看出信度是真分数的变异在测验分数变异中所占的比重，或者说测验分数的变异在多大程度上是由真分数的变异引

① 戴海崎、张峰、陈雪枫：《心理测量》，暨南大学出版社 1999 年版，第 68—87 页；金瑜：《心理测量》，华东师范大学出版社 2001 年版，第 139—168 页。

② Lord, F. M., & Novick, M. R., *Statistical Theories of Mental Test Scores*, Addison-Wesley Press, 1968.