资格认证测验的信度估计及其特征分析

赵世明

(中国浦东干部学院 领导研究院,上海 201204)

摘 要:资格认证测验属于典型的标准参照测验,在国内已得到普遍应用,但在报告心理测量学指标时很少提到标准参照模式的信度估计指标。该文归纳了标准参照测验信度估计的指标体系,分析讨论了适用于资格认证测验的信度估计及其与测验长度、分界标准分布、样本同质性的关系与特征。

关键词:标准参照测验;资格认证测验;信度估计;概化理论

中图分类号:B841.7 文献标识码:A 文章编号:1003 - 5184(2006)03 - 0084 - 04

在标准参照测验的研究文献中,信度估计是研究最为深入和广泛的课题之一。以资格认证测验为代表的标准参照测验在国内已得到普遍应用,但在报告心理测量学指标时很少提到标准参照模式的信度估计,多数以相关法报告内部一致性信度或分半信度,在美国八十年代初期也是这种状况[1]。资格认证测验属于典型的标准参照测验,由于被试变异的减小,以常模参照模式估计信度指标时容易低估,相关法不再适用。因此在探讨各类信度指标的适用性基础上,应尽早将标准参照模式的信度估计指标体系引入各类资格认证测验或能力水平测验。

1 标准参照测验信度估计的指标体系

Hambleton 等人(1978)归纳了标准参照测验的三类信度估计指标:基于分类一致性信度、测验分数信度和领域分数估计值信度[2]。在一般意义的信度概念中,人们习惯将后两者信度估计指标看作是标准参照测验的信度。研究主要讨论这两类标准参照测验的信度估计。在此领域中,测验信度估计主要是由来自被试、题目、测量误差的方差分量之间的关系所决定。方差分析是信度估计的基础。实际上,标准参照模式的信度估计更能体现信度概念的基本假设:由分数变异之间的关系界定测验分数的可靠性。借助方差分析技术,概化理论将被试变异与总体变异的比例界定为测验的信度。

鉴于概化理论于标准参照测验的应用,平行测验的复本假设问题对测验分数的信度估计产生了影响。依据不同方法编制的平行复本对应不同的信度估计程序,要求采用不同的信度指标[3]。以经典方法编制的平行测验复本,属于同一测量目标下"题目集合"的同质样本,这些题目样本或测验复本的测量内容、平均分数、方差、难度及题目间相关(同质性)均是相同的,这类复本称为"经典的平行测验复本";以随机方式编制的测验复本并不要求各题目样本或测验复本同质,复本之间允许有不同的均数与方差,这种假设相对较弱一些,题目样本是从"所有可能的题目领域"中以随机或分层随机方式抽取的,此类复本称为"随机的平行测验复本"。在估计信度时,如果只考虑被试差异带来的组内效应及被试与题目的交互作用,即利用经典复本的信度估计方法;在此基础上,如果还考虑题目差异带来的组间效应(实际上属于一种随机误差),就要利用随机复本的信度估计方法。

可以利用概化理论中题目 × 被试设计的"D 研究"。将题目视为"面",形成一个单面交叉设计;测验看作是随机平行复本。利用代表领域分数变异与观测分数变异之比的"依存性指数(Index of dependability)"估计标准参照测验的分数信度,其大小反映估计被试领域分数的可靠性。在概化理论中的 D 研究中,复本是依照随机原则编制的。也就是说,编制复本并不要求按照严格的平行原则,从总体中抽取题目样本组成的试卷,不必具有相同的平均分和方差,允许存在题目间的难度差异。这样题目和被试的交互作用被引入了估计领域分数信度的过程。

以下是标准参照测验基于方差分析的信度估计指标体系。

表 1　基于方差分析的标准参照测验信度估计指标体系

效应（变异源）	自由度 *df*	平方和 *SS*	均方 *MS*	G 研究方差分量	D 研究样本数	D 研究方差分量
被试(p)	$n_p - 1$	SS(p)	MS(p)	$\sigma^2(p)$	1	$\sigma^2(P)$
题目(i)	$n_i - 1$	SS(i)	MS(i)	$\sigma^2(I)$	n'_I	$\sigma^2(I)$
交互作用(pi)	$(n_p - 1)(n_I - 1)$	SS(pi)	MS(pi)	$\sigma^2(pi)$	n'_i	$\sigma^2(PI)$
方差分量的计算	$\sigma^2(p) = [MS(p) - MS(pi)]/n_i$ $\sigma^2(I) = [MS(I) - MS(pi)]/n_p$ $\sigma^2(pi) = MS(pi)$			$\sigma^2(P) = \sigma^2(p)$ $\sigma^2(I) = \sigma^2(i)/n'_I$ $\sigma^2(PI) = \sigma^2(pI) = \sigma^2(pi)/n'_i$		
信度指标体系	经典复本方式			随机复本方式		
领域分数估计	估计的相对测量误差			估计的绝对测量误差		
置信区间	$\sigma^2(\delta) = \sigma^2(pI)$			$\sigma^2(\Delta) = \sigma^2(I) + \sigma^2(pI)$		
领域分数	概化系数			依存性指数		
估计信度	$\varepsilon\rho^2 = \sigma^2(p)/[\sigma^2(p) + \sigma^2(\delta)]$			$\Phi = \sigma^2(p)/[\sigma^2(p) + \sigma^2(\Delta)]$		
测验分数信度估计	$K^2(x, T) = [\sigma^2(p) + (X_{PI} - \lambda)^2 - \sigma^2(X_{PI})]/[\sigma^2(p) + (X_{PI} - \lambda)^2 - \sigma^2(X_{PI}) + \sigma^2(\delta)]$			$\Phi(\lambda) = [\sigma^2(p) + (X_{PI} - \lambda)^2 - \sigma^2(X_{PI})]/[\sigma^2(p) + (X_{PI} - \lambda)^2 - \sigma^2(X_{PI}) + \sigma^2(\Delta)]$		
相关统计量的计算	样本平均分：$\bar{X}_{PI} = \Sigma\Sigma X_{pi}/n_p n_I$					
	样本平均分的估计方差：$\sigma^2(\bar{X}_{PI}) = \sigma^2(p)/n'_p + \sigma^2(i)/n'_i + \sigma^2(pi)/n'_p n'_i$					

2　实证研究

以某项资格认证测验为样本，根据标准参照模式的信度估计模式估计资格认证测验的分数信度估计指标，讨论并分析适用于此类测验的标准参照测验信度指标与测验长度、分界标准分布、样本同质性的相互关系及信度指标的技术特征。

2.1　被试样本来自该项资格认证测验

采用分层随机抽样，从全国主要行政区划（华中、华东、华南、西北等地区）选出五个被试样本组，共计 8701 人。抽取上述样本时合并考虑了不同地区被试的同质性。考虑因素为被试的学历构成（具备专业学历的被试占被试总人数的百分比）和测验分数的离散程度。

2.2　测验样本来自该项资格认证测验试卷，测验长度为 230 题，均为 0、1 计分。

2.3　研究方法

1）将研究对象看作是一个 P * I 的单面交叉设计，将题目视为面；

2）利用 GENOVA 进行 G 研究，计算各种方差分量；

3）利用 GENOVA 进行 D 研究，计算各种相关统计量。

3　结果与讨论

3.1　信度估计的基本模式

选择华东地区被试（样本一）作为信度估计的研究样本。该样本在被试同质性和分数分布方面均比较符合标准参照测验的被试与分数特征。表 2 列出了 G 研究和 D 研究在测验长度分别为 230、150、100 和 50 题时，不同分界标准下的方差分量与相关信度指标的统计结果。

表 2　样本一的 G 研究与 D 研究

效应	G 研究	D 研究的方差分量			
变异源	方差分量	$n'_i = 50$	$n'_I = 100$	$n'_i = 150$	$n'_i = 230$
被试(p)	0.0080	0.0080	0.0080	0.0080	0.0080
题目(i)	0.0490	0.0010	0.0005	0.0003	0.0002
交互作用(pi)	0.1616	0.0032	0.0016	0.0011	0.0007
	$\sigma^2(\delta)$	0.0032	0.0016	0.0011	0.0007
$\sigma^2(\Delta)$		0.0042	0.0021	0.0014	0.0009
$\sigma^2(X_{PI})$		0.0010	0.0005	0.0004	0.0003
	$\varepsilon\rho^2$	0.7123	0.8319	0.8813	0.9193
Φ		0.6551	0.7916	0.8507	0.8973

表 2 样本一的 G 研究与 D 研究

效应	G 研究	D 研究的方差分量			
变异源	方差分量	$n'_i = 50$	$n'_I = 100$	$n'_i = 150$	$n'_i = 230$
$\Phi(\lambda = X_{PI} = .68)$ *		0.6234	0.7800	0.8447	0.8943
	$K^2(x, T)$	0.7123	0.8319	0.8813	0.9193
$\Phi(\lambda = 0.40)$		0.9530	0.9761	0.9839	0.9895
	$K^2(x, T)$	0.9644	0.9817	0.9864	0.9920
$\Phi(\lambda = 0.50)$		0.9034	0.9498	0.9661	0.9777
	$K^2(x, T)$	0.9276	0.9620	0.9715	0.9830
$\Phi(\lambda = 0.60)$		0.7605	0.8682	0.9091	0.9392
	$K^2(x, T)$	0.8268	0.6364	0.7447	0.9023
$\Phi(\lambda = 0.70)$		0.6364	0.7889	0.8512	0.8990
	$K^2(x, T)$	0.7447	0.8468	0.8788	0.9267
$\Phi(\lambda = 0.80)$		0.8354	0.9122	0.9401	0.9603
	$K^2(x, T)$	0.8788	0.9342	0.9500	0.9699
$\Phi(\lambda = 0.90)$		0.9293	0.9637	0.9756	0.9840
	$K^2(x, T)$	0.9467	0.9724	0.9784	0.9877

注：$\lambda = X_{PI} = 0.68$ 为该样本的平均通过率

3.2 信度指标的基本特征

表 2 中的统计结果表明，标准参照模式的信度指标具备二个基本特征：

随着题目样本的增加，题目变异和误差变异（交互作用）减小，相对测量误差 $\sigma^2(\delta)$ 和绝对测量误差 $\sigma^2(\Delta)$ 减小，领域分数信度估计和测验分数信度估计均提高。

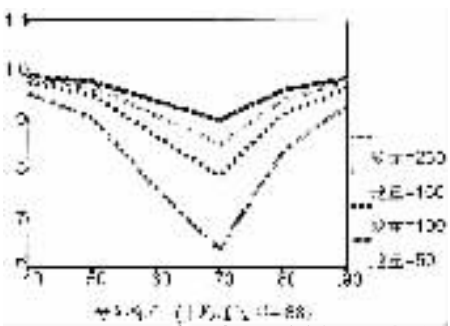

图 1 不同测验长度下 $\Phi(\lambda)$ 与分界标准的关系

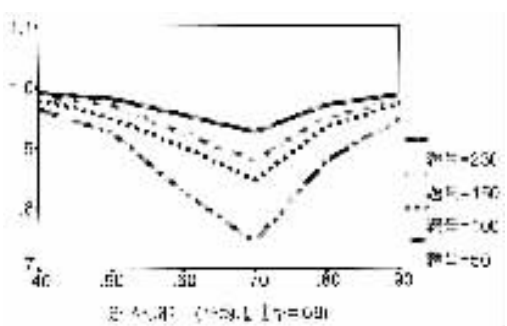

图 2 不同测验长度下 $K^2(x, T)$ 与分界标准的关系

随着分界标准离均差的增大，两个测验分数信度估计值 $\Phi(\lambda)$ 和 $K^2(x, T)$ 提高；当分界标准接近样本平均分时，这两个信度估计值达到最低点。说明标准参照测验的分数信度指标具有趋中性：分界标准距离样本平均分越近，区分误差越大，分类的一致性和可靠性下降。见图 1 和图 2。

3.3 不同复本方式下的测验分数信度估计

由于经典复本方式下的测验与领域分数信度估计在计算测量误差时忽略了题目变异 $\sigma^2(I)$ 的影响，没有考虑题目难度差异造成的分数差异，因此，经典复本方式的测验信度估计值均大于随机复本方式下的测验信度估计值。即：

由于 $\sigma^2(\Delta) > \sigma^2(\delta)$，总有 $\varepsilon\rho^2 > \Phi$ 和 $K^2(x, T) > \Phi(\lambda)$

表 2 的统计结果也验证了这一点。提示人们在估计资格认证测验分数信度时，应考虑不同复本模式的影响。经典复本方式下的信度估计高于随机复本方式下的信度估计。如果在标准参照测验条件下，仅考虑经典复本模式而未采用随机复本模式，测验信度可能会被高估。

3.4 被试样本同质性对测验分数信度指标的影响

表 3 给出了同质性递减的五个被试样本的相关统计量及信度估计结果。结果表明，各种信度估计指标均与样本分数方差成正比关系：随着样本同质性的降低，样本分数变异增加，被试变异也随之增加，信度估计值也相应提高。说明资格认证测验的信度估计与被试样本的同质性存在密切关系。被试样本越同质，信度估计值越小，这与传统信度理论是一致的。需要说明的是，由于题目样本比较大的原因，尽管测量误差也随分数方差增大而增加，但是变化幅度不大。

表 3　不同被试同质性条件下的信度估计(n = 230)

	样本一	样本二	样本三	样本四	样本五
样本分数方差 S^2	21.40	25.95	27.65	30.05	33.52
被试变异 $\sigma^2(p)$	0.0080	0.0120	0.0136	0.0163	0.0199
测量误差 $\sigma^2(\delta)$	0.0007	0.0008	0.0008	0.0008	0.0008
测量误差 $\sigma^2(\Delta)$	0.0009	0.0009	0.0010	0.0010	0.0010
概化系数 $\varepsilon\rho^2$	0.9193	0.9406	0.9420	0.9525	0.9606
依存性指数 Φ	0.8973	0.9289	0.9305	0.9431	0.9523
$\Phi(\lambda = X_{PI})$	0.8943	0.9276	0.9293	0.9423	0.9516
$K^2(x,T;\lambda = X_{PI})$	0.9193	0.9406	0.9420	0.9525	0.9606

3.5　标准参照与常模参照模式信度估计的关系

尽管经典复本方式的信度估计可用于常模参照测验,但是这些信度估计都是基于概化理论的,与传统测验理论的信度估计仍有不同。以上统计结果也可以证明:

$\varepsilon\rho^2 > \Phi > \Phi(\lambda = X_{PI}) = KR21$,即随机复本依存性指数的最小取值为 KR21;

$K^2(x, T) > K^2(x, T;\lambda = X_{PI}) = KR20 = \varepsilon\rho^2$,即经典复本依存性指数的最小取值为 KR21;对于 0、1 计分的测验来说,概化系数等于 KR20。

在两种复本方式下,标准参照模式的测验分数信度估计均以常模参照模式的测验信度估计为最小值,说明标准参照测验分数的信度估计一般高于常模参照测验分数的信度估计。在资格认证测验中,由于被试的同质性增强,测验分数变异减小,使得以常模参照模式估计标准参照测验的信度指标时容易低估。因此,相关系数方法用于资格认证测验信度估计的适用性就降低了。

4　结论

在资格认证测验中,如果以经典方式编制测验复本,并假设错误区分的性质和程度所导致的损失并不是同样严重的,就应考虑使用平均误差损失函数方法中的 $K^2(x, T)$ 指标;如果是以随机方式编制测验复本,认为题目差异对于区分一致性有影响,就可以选择 $\Phi(\lambda)$ 指标;如果关心测验分数对领域分数的估计程度,可以选用依存性指数 Φ 和概化系数 $\varepsilon\rho^2$。一般情况下,既看重测验分数的一致性和可靠性,也关心以测验分数估计领域分数的精确性,因此既应报告测验分数信度,也应报告领域分数估计信度。在实际应用中,标准参照模式的信度估计过程可能会较为繁杂,但是可以利用多种方式计算方差分量,并通过编写专门的计算机程序来解决其中的技术问题。

在标准参照测验的研究文献中,尚未发现对标准参照测验信度可接受标准的专门研究。可以认为,在可接受标准的问题上,标准参照的信度估计与常模参照的信度估计不应存在双重标准。因此资格认证测验的信度系数应达到 0.90 以上;短测验或分测验的信度系数应达到 0.80 – 0.85 以上。

参考文献

1　Berk R A. Criterion – referenced measurement: the state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980: 231 – 232.

2　Hambleton R K, Swaminathan H, Algina J, et al.. Criterion – referenced testing and measurement: A review of technical issues and development. Review of Educational Research: 1978, 48: 15 – 23.

3　张厚粲, 刘昕. 考试改革与标准参照测验. 沈阳: 辽宁教育出版社, 1992: 166.

Credentialing Test Reliabilities and Their Characteristic Analysis

Zhao Shiming

(Institute of Leadership, China Executive Leadership Academy Pudong, Shanghai 201204)

Abstract: The professional credentialing tests as the typical example of criterion – referenced test have been used universally in China, but reliability indexes of criterion – referenced format was rarely reported. Taking a national professional credentialing examination as research sample, this article induces the reliability index system for credentialing tests and discusses the characteristics of reliability estimating in relation with test length, standard setting distribution and subject .

Key words: Criterion – referenced test; credentialing tests; reliability; generalization theory