质性研究中编码者信度的多种方法考察

徐建平[1,2*] 张厚粲[3]

([1]陕西师范大学心理系,西安,710062) ([2]中国科学院心理研究所,北京,100101)

([3]北京师范大学心理学院,北京,100875)

摘 要 质性研究中检验编码者信度的方法有归类一致性指数、编码信度系数、相关系数、中位数检验、概化系数等。基于教师胜任力访谈数据集,对编码者信度考察结果表明,归类一致性指数和编码信度系数受相同编码数影响而不稳定,相关系数受数据类型制约,中位数检验受研究设计影响,概化系数则受编码者和编码项目的数量影响。研究中须合理选用。

关键词:编码者信度 归类一致性指数 编码信度系数 相关系数 中位数检验 概化系数

1 引言

由于质性研究中文本资料编码不一致现象的存在,编码者信度历来受研究者关注[1]。为确保编码者在内容分析、文本理解、主题提取、编码及解释中一致,通常对编码数据要做必要考察。本文结合教师胜任力研究数据集,探讨质性研究中多种检验编码者信度方法的效用。

2 方法和步骤

预访谈。采用行为事件访谈法(BEI)对 7 名教师做个别访谈,要求叙述"教学工作中三件成功和三件失败的事件",涉及发生的情境、任务、对话、行动、思考、感受和影响。访谈全程录音。基于录音文本,运用主题和内容分析法,练习编码,形成《教师胜任力编码词典》。

筛选编码者。从预访谈文本编码者中,筛选出识别胜任特征行为指标准确率高、编码一致性高的编码者 2 名。他们均接受过严格的 BEI 编码方法训练。

正式访谈。程序与预访谈相同。受访中小学教师 24 名,高绩效组和普通绩效组各 12 名。高绩效组由获全国和省级优秀教师、特级教师、优秀教育工作者、模范教师、教育系统先进工作者、教学能手、骨干教师等荣誉称号,且近三年来教学业绩考核评价为优秀的在岗教师组成。普通组由一般教师组成,随机选自高绩效教师所在的学校。

编码。2 名编码者仔细阅读正式访谈文本,依照《教师胜任力编码词典》编码并登录胜任特征出处(文本编号、页码、行号)、胜任特征代码、等级水平。

用多种方法考察编码者信度。统计 2 名编码者对每个受访者叙述的所有行为事件中各胜任特征在不同等级上出现的频次。依此计算各个胜任特征发生的总频次、等级分数、平均等级分数和最高等级分数。最后,采用多种方法对这些数据的编码者信度进行检验。

3 结果和分析

3.1 归类一致性指数及编码信度系数

归类一致性指数指对编码归类相同数占归类总数的百分比,即 $CA = 2 \times S/(T_1 + T_2)$。式中 S 表示两名编码者归类一致数,$T_1$、$T_2$ 表示每人的编码总数[2]。编码信度系数公式为 $R = (n \times \text{平均相互同意度})/(1 + n \times \text{平均相互同意度})$[3]。其中平均相互同意度 $= 2 \times M/(N_1 + N_2)$,式中 M、N_1、N_2 与前式中 S、T_1、T_2 意义相同。显然,后者基于归类一致性指数。结果见表 1。

表 1 两名编码者归类一致性及编码信度系数

被试编号	*CA*	*R*	被试编号	*CA*	*R*	被试编号	*CA*	*R*
01	0.322	0.487	09	0.365	0.535	17	0.752	0.859
02	0.577	0.732	10	0.442	0.613	18	0.534	0.697
03	0.402	0.574	11	0.339	0.507	19	0.500	0.667
04	0.571	0.727	12	0.455	0.625	20	0.556	0.714
05	0.436	0.608	13	0.587	0.740	21	0.654	0.791
06	0.447	0.618	14	0.575	0.730	22	0.674	0.805
07	0.423	0.595	15	0.438	0.609	23	0.704	0.826
08	0.424	0.596	16	0.444	0.615	24	0.659	0.795
全体被试				*CA* = 0.500			*R* = 0.667	

3.2 相关系数

计算两名编码者编码频次的等级相关,平均等级分和最高等级分的积差相关。结果如表 2。

3.3 中位数检验

将两名编码者的平均等级分转化为标准分,计算每个被试 28 个胜任特征平均等级分总和,再求其联合中位数。结果发现优秀组高于此数 9 个,普通组 3 个;优秀组低于此数 3 个,普通组 9 个。即编码者能够正确识别出 75 %的优秀组被试和 75 %的普通组被试。采用精确概率法计算 $\chi^2 = 6$, $p = 0.039$; *Kappa* = 0.5, $p = 0.014$。其差异在 0.05 水平上具有统计学意义,反映一名编码者和另一名参与抽样的编码者编码一致性较好。结合 *Kappa* 值,确认一致性中等。

* 通讯作者:徐建平,男。E-mail:xping@snnu.edu.cn

表 2 两名编码者编码的频次、平均等级分、最高等级分的相关

胜任特征	频次	平均等级分	最高等级分
挑战与支持	0.641**	0.538**	0.370
自信心	0.863**	0.604**	0.777**
创建信任感	0.746**	0.304	0.275
尊敬他人	0.673**	0.705**	0.410*
分析性思考	0.271	0.322	0.290
概念性思考	0.225	0.465*	0.311
提升的动力	0.689**	0.547**	0.735**
重视次序、品质与精确	0.625**	0.766**	0.711**
信息搜寻	0.471*	0.729**	0.713**
主动性	0.511**	0.389	0.460*
灵活性	0.653**	0.233	0.519**
倡导责任感	0.732**	0.777**	0.893**
创新性	0.724**	0.586**	0.612**
建立关系	0.796**	0.779**	0.770**
管理学生	0.594*	0.623**	0.755**
学习的热情	0.787**	0.614**	0.857**
冲击与影响	0.553*	0.185	0.482*
发展他人	0.806**	0.773**	0.894**
团队精神与协作	0.636**	0.675**	0.772**
理解他人	0.454*	0.051	0.413*
诚实正直	0.752**	0.678**	0.895**
自我控制	0.780**	0.726**	0.781**
专业知识与技能	0.791**	0.776**	0.846**
关注师生的需求	0.668**	0.831**	0.799**
自我评估	0.825**	0.634**	0.778**
情绪觉察力	0.573**	0.845**	0.608**
应变能力	0.560**	0.490*	0.619**
职业偏好	0.776**	0.073	0.602**

** $p = 0.01$ * $p = 0.05$

3.4 G系数与 ϕ指数

根据概化理论,用平均等级分采用随机双面交叉设计(P×I×R)考察编码者信度。其中,I和 R 都随机。表 3 是运行 GENOVA3.1 软件得到的 G研究结果。

表 3 结果显示编码者(R)变异分量接近零,说明两名编码者评分客观独立。受访者与编码者交互效应(P×R)变异分量值很小,说明编码者做到了盲评。胜任特征侧面变异较大,说明编码词典中部分胜任特征存在高相关,需要合并以改善区分结构。胜任特征与编码者交互效应(I×R)值较小,说明两名编码者对胜任特征的理解比较一致。三个侧面的交互效应(P×I×R)占总变异分量的 29.564%,表明编码者对胜任特征的理解、编码、胜任特征在不同受访者之间的差异对最后的分数影响最大。

进一步实施 D 研究,以了解胜任特征、编码者样本容量对编码信度的影响。表 4 是编码者侧面样本容量分别为 1、2,胜任特征侧面样本容量随机时的 D 研究结果。

表 4 显示在初始情况下 G值和 ϕ值最大,编码者信度最高,说明编码一致性好。由此推论,两个侧面样本容量增大,G系数和 ϕ指数都会提高。图 1 显示了这种渐变提高的趋势。

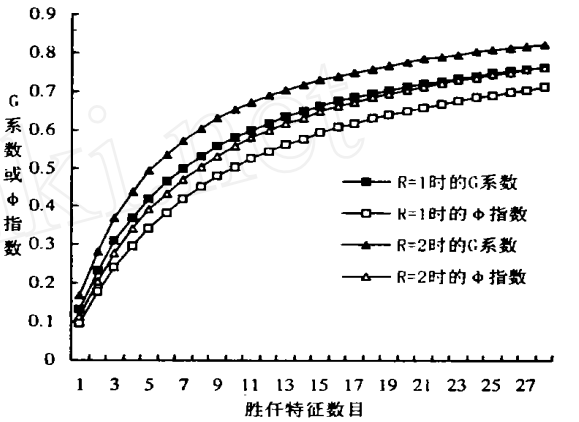

图 1 G系数、ϕ指数与编码者、胜任特征容量间的关系

表 3 P×I×R 设计胜任特征编码的 G研究变异分量估计

变异来源	*df*	*SS*	*MS*	变异分量估计值	占总变异分量百分比
P(受访者 = 24)	23	535.303	23.274	0.342	9.816
I(胜任特征 = 28)	27	1254.672	46.459	0.885	25.402
R(编码者 = 2)	1	0.001	0.001	(0.0)	0
P×I	621	2097.689	3.378	1.174	33.697
P×R	23	41.418	1.801	0.028	0.804
I×R	27	44.289	1.640	0.025	0.718
P×I×R	621	639.630	1.030	1.030	29.564

表 4 P×I×R 设计 D 研究的概化系数和可靠性指数

胜任特	评分者侧面样本容量(R = 1)		评分者侧面样本容量(R = 2)	
征项目	G系数	ϕ指数	G系数	ϕ指数
1	0.133	0.098	0.167	0.116
3	0.309	0.243	0.372	0.281
5	0.422	0.344	0.493	0.391
10	0.579	0.502	0.652	0.556
15	0.662	0.592	0.730	0.647
20	0.713	0.651	0.777	0.705
25	0.747	0.692	0.808	0.744
28	0.763	0.711	0.822	0.763

表 5 P×I×R 设计不同固定侧面时的 D 研究结果

	样本容量	G系数	ϕ指数
胜任特征侧面固定,编码者侧面随机	1	0.856	0.855
	2	0.923	0.922
编码者侧面固定,胜任特征侧面随机	1	0.174	0.121
	5	0.513	0.407
	10	0.678	0.579
	15	0.759	0.673
	20	0.808	0.733
	25	0.840	0.774
	28	0.855	0.794

受编码项目数量、编码者人数等条件约束,研究推广面会受限。因此,当胜任特征项目或编码者侧面固定,而另一侧面随机时,D 研究结果会因此不同。如表 5 所示。

结合完全随机和某一侧面固定时的分析结果,发现随着各个侧面样本容量的增大,G系数和 ϕ指数都会增加,但在编码项目数量较小时,增幅较大。因此,编码项目并不是越多越

好,只要 G值满足心理测量学要求即可。

4 讨论

就 BEI 方法在胜任力建模中的信度、效度而言,Boyatzis 和 Burrus(1977),Klemp(1977)等研究证实,受过训练的不同编码者采用最高分和频次编码,其一致性介于 0.74 - 0.80 之间[4]。Motowidlo 等(1992)研究也表明,对同一组被试两次访谈的胜任特征评价结果具有较高稳定性[5]。本研究中归类一致性在 0.322 - 0.752 之间,总的归类一致性为 0.500,编码信度系数值在 0.487 - 0.859 之间,总体编码信度系数为 0.667。这些值不稳定,部分较低,随着深入讨论,编码一致性逐渐在提高。一方面由于编码时除了记录胜任特征频次,还要评定其等级水平,归类编码难度较大;另一方面由于文本内容非常广泛,编码时需要反复核查修正和确认。

用相关系数法考察编码一致性的结果显示,绝大多数胜任特征频次、平均分、最高等级分相关显著,说明编码一致性较高。表 2 结果也显示,用相关法考察编码者信度要注意数据类型。如果把频次转化为百分比,也可用皮尔逊积差相关法。

中位数检验法能从整体上评估编码者准确区分两组被试的能力。在单盲或双盲设计中,编码结果如果能够完全区分不同水平的被试,就表明编码客观一致。编码一致性高,编码项目对不同被试的区分会更准确,也越能反映抽样的准确性。本研究为单盲设计,一名编码者不了解被试信息,文本编号随机,参与抽样的编码者也不知文本的隶属,需要评定编码者与抽样者之间的编码一致性。若针对每一等级编码频次做检验,会得到更严格的分析结果。

编码项目数量是影响归类一致性的另一重要因素。在本研究中,编码主题可以在讨论基础上修改。结果分析表明,编码一致性比较满意,但编码者等级评分差异较大,这与编码词典设计、编码项目数量有关。当信度较低时,减少编码项目是提高编码信度的重要方法。概化分析有助于确定编码项目数。在 G研究中,编码者侧面、以及与编码者相关的交互效应变异分量越小,编码者一致性就越高,这与方差分析原理相同。因而,也可用 t 检验和 ANOVA(R > 2)考察编码一致性[6]。也有人提出用 SEM 方法估计变异分量研究评分者信度[7]。D 研究是用不同条件下的 G系数确定不同情况下编码者一致性。G系数高,说明编码者信度较高。图 1 显示胜任特征大约为 15 时,G值会达到心理测量学允许接受的范围。

因此,通过合并或缩减编码项目、提供体现编码项目特征的典型例证、训练编码者、参照其他行为指标等方法,可以提高编码的一致性和可靠性,从而提高编码者信度和研究效度。

5 结论

不同编码者对相同文本独立编码的一致性是检验编码者信度的重要指标。用多种方法考察编码者信度的结果表明,归类一致性要求最严格,其稳定性更多的受相同编码数目的影响;相关系数受数据类型制约;中位数检验受实验设计的影响;G系数则受编码者和编码项目数量的影响。就本研究数据集而言,编码一致性符合研究基本要求,编码结果可靠。

6 参考文献

1 Kvale, S. Validity in the qualitative research interview. Methods, 1987,1(2): 37 - 42

2 Schultheiss, O. C., Brunstein, J. C. Assessment of implicit motives with a research version of the TAT: picture profiles, gender differences, and relations to other personality measures. Journal of Personality Assessment, 2001,77(1):71 - 86

3 董奇. 心理与教育研究方法. 广州:广东教育出版社,1992:398

4 Boyatzis, R. E. The Competent Manager: A Model for Effective Performance. New York: Wiley, 1982:52

5 Motowidlo, S. J., et al. Studies of the structured behavioral interview. Journal of Applied Psychology, 1992, 77(5): 571 - 587

6 (美)安托尼特·D·露西亚,(美)理查兹·莱普辛格著,郭玉广译. 胜任:员工胜任能力模型应用手册. 北京:北京大学出版社,2004:134

7 严芳,李伟明. 用结构方程建模估计概化理论中的评分者信度. 心理学报,2002,34(5):534 - 539

Testing Intercoder Reliability by Multi-approaches in Qualitative Research

Xu Jianping[1,2]

([1] Department of Psychology, Shaanxi Normal University, Xi'an, 710062)
([2] Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101)

Zhang Houcan[3]

([3] School of Psychology, Beijing Normal University, Beijing, 100875)

Abstract This paper reported several different statistical procedures for assessing intercoder (or interrater) reliability in qualitative research. The procedures included the index of category agreement, intercoder reliability coefficient, correlation coefficient, median test and G coefficient. Based on the data set of the interview study on teachers' competency, the results of analysis indicated that reliability indices obtained with different methods for coding in qualitative research had different efficiency. The index of category agreement and intercoder reliability coefficient was shown to be less stable and more affected by consistency among coders. Correlation coefficient was affeced by data type. Median test was limited by the research design. G coefficient was influenced by the number of items and coders. The authors suggested that researcher should choose the appropriate methods from these procedures and use them rationally.

Key words: intercoder reliability, index of category agreement, intercoder reliability coefficient, correlation coefficient, median test, G coefficient