领导干部结构化面试信度的多元概括化理论分析

洪自强[1]，涂冬波[2]

（1. 中国领导干部考试与测评中心，北京 100036；2. 江西师范大学 教育学院，南昌 330027）

摘 要：本研究尝试运用多元概括化理论对北京市某区副处级干部准入资格结构化面试测评数据进行测量信度分析，为提高领导干部考试与测评工作科学化水平提供了有益的实证依据。主要结论有：(1)本次结构化面试难度适中，区分度较高；(2)各测评要素及合成分数的类信度系数均较高，合成分数的测量信度高于单个测评要素的测量信度；(3)各测评要素及合成分数的类信度系数随着考官数量的增加而增加，且从确保信度和降低成本考虑，考官数量以 5－9 位为宜；(4)在这次面试测评中，各项测评要素间的相关系数较高，这为目前在选拔面试中将各项测评要素得分进行合成提供了依据，说明用合成分数计算总分具有一定的合理性。

关键词：结构化面试；多元概括化理论；信度

中图分类号：B841.2 文献标识码：A 文章编号：1003－5184(2006)01－0085－06

1 概括化理论简介及研究目的

概括化理论（generalizability theory）是经典真分数理论特别是经典信度理论的重大发展，是一种具有广阔前景的现代测量信度理论。在经典真分数理论中，测量分数是真分数和测量误差之和，其中，测量误差是一种笼统的随机误差，无法对造成测量误差的具体来源做出分析。概括化理论则能按照不同的误差来源（在概括化理论中，对拟研究的误差来源，称为测量侧面），对测量分数进行具体分解，并采用方差分析统计方法，深入考察误差来源对测量信度的影响程度，在此基础上，可以帮助人们针对具体误差来源，提出有效控制和改善测量精度的措施和办法[1～2]。

运用概括化理论对测量分数进行分析，简称概括化理论分析，具体包括 G 研究（概化研究）和 D 研究（决策研究）两个阶段。G 研究的主要任务是，针对拟研究的误差来源或测量侧面，提出研究设计，并实施测试，收集数据，分析与各种误差来源有关的方差分量。D 研究的主要任务是，在 G 研究的基础上，用类似经典真分数中的信度指标（在概括化理论中，称为类信度系数或 G 系数）对测量精度做出评价，并针对误差来源，通过改变取值区间或固定某些测量侧面等方法，考察减少测量误差、提高测量信度的具体策略。从这两个阶段的内在联系看，G 研究是 D 研究的基础，D 研究是对 G 研究所得结论的进一步推广，使提高测量信度的某些结论具有普遍意义。

结构化面试是领导干部考试与测评方法体系的重要组成部分，是在笔试基础上进一步测试被试在领导能力素质和个性特征等方面与选拔职位的匹配程度[3]。结构化面试作为一种测评方法，能否选准选好干部，始终要受到试题质量、实施程序和考官等因素的影响。近年来，经过各地各部门的积极探索和实践，领导干部结构化面试在制度化建设、职位针对性以及考务规范化等方面均有明显改进，实际效果比较好，得到各方面肯定。但是，从实际情况看，考官因素比较复杂，诸如如何优化考官构成、研究制定考官资格、合理确定考官数量等问题，目前主要靠经验把握，未得到科学解释。因此，考官因素将是进一步提高结构化面试质量的一个重要突破口。本文尝试运用概括化理论，主要分析结构化面试的信度尤其是考官小组成员数量对结构化面试信度的影响，并探讨考官小组的合理规模。

2 数据来源

2005 年上半年，北京市某区委组织部开展了副处级领导干部资格准入结构化面试。这次面试试题限时 15 分钟，共 3 道题，评价 4 种测评要素，总分为 100 分，其中 3 道题依次主要测试自我认知与综合分析能力、组织协调能力、决策应变能力，权重都是

28 分,语言表达与气质则在整个面试过程中考察,权重为 16 分。

考官小组有 7 位成员,由领导干部和测评专家构成。在第一位被试回答后,考官小组作了简要讨论交流,进一步统一评判标准,从第二位被试开始,每位被试回答完毕后,各位考官坚持独立评分,互不讨论。共有 82 位被试参加了这次面试。男被试 65 位,占 79.3%,女被试 17 位,占 20.7%,被试年龄在 26~43 岁之间,平均年龄为 36.8(标准差 3.30)。不管是从面试实务来看,还是从理论研究来看,这是一个非常难得的样本。

3 多元概括化理论分析方案

从结构化面试的特点看,这次测评工作的的测量目标是 82 个被试(P),测量目标被细分为 4 项测评要素(自我认知与综合分析能力、组织协调能力、决策应变能力、语言表达与气质),7 位考官(R)则是面试测评的测量侧面。因此,对这次面试测评的信度进行概括化理论分析,应采用多元概括化理论分析,即有 4 个目标测评要素的单侧面完全交叉设计(P×R)。

本次面试测评的数据格式构成一个 82 位被试×(4 项测评要素×7 位考官)矩阵,共 2296 个元素(数据)。每一个数据即为每位考官对每一位被试在某项测评要素上的打分。

4 试题的难度和区分度

根据每位被试的最后得分,采用得分率和积差相关(各项测评要素上的得分与总分的相关系数),分别计算面试的难度和区分度。自我认知与综合分析能力、组织协调能力、决策应变能力、语言表达与气质的难度分别为 0.73、0.69、0.70、0.76,区分度分别为 0.92、0.91、0.87、0.92。

一般来说,项目难度应在 0.20~0.80 之间,以 0.50 为最佳。这次面试四项测评要素的难度在 0.69~0.76 之间,平均难度为 0.72,从面试实践的角度看,这个难度比较合适。项目区分度应在 0.20 以上。从表 2 可以看到,这次面试各项测评要素的区分度在 0.87~0.92 之间,平均为 0.90,比较理想。

5 多元概括化理论分析的 G 研究

用布伦南的 mGENOVA[4] 分析软件进行数据处理,计算测量目标的被试、测量侧面的考官以及被试和考官的交叉项等随机效应在 4 项测评要素上的方差和协方差分量矩阵,结果如表 1 所示。

表 1 G 研究方差与协方差分量的估计

效应	测评要素 1	测评要素 2	测评要素 3	测评要素 4
P	4.91	0.8	0.72	0.96
(被试)	3.53	3.91	0.78	0.83
	2.97	2.86	3.46	0.77
	2.29	1.76	1.55	1.16
R	4.49			
(考官)	4.37	5.13		
	4.26	4.65	4.42	
	1.3	1.56	1.44	1.25
P×R	4.16			
(被试×考官)	1.86	4.86		
	1.21	1.93	3.86	
	1.03	1.04	1.09	1.49

注:①测评要素 1:自我认知与综合分析;测评要素 2:组织协调;测评要素 3:决策应变;测评要素 4:语言表达与气质。(下同)

②上表矩阵中对角线上元素为 4 个测评要素效应的方差分量,对角线以上元素为相关系数,对角线以下元素为协方差分量。

由表 1 可知，4 项测评要素之间的相关系数较大，说明用这 4 项指标的得分来反映被试的素质水平，其结果会比较一致。这样，不但可以分别从自我认知与综合分析能力、组织协调能力、决策应变能力以及语言表达与气质等方面做信度评估，还可以将各方面分数合成起来作整体评估。

6　多元概括化理论分析的 D 研究

D 研究分两步走，第一步分析当前实施方案（7 位考官）的类信度系数（概化系数或 G 系数），第二步探索考官数量对面试测评类信度系数的影响，并确定合适的考官数量。

6.1　D 研究当前实施方案的方差、协方差分量及类信度系数

6.1.1　D 研究当前实施方案的方差与协方差分量

D 研究中方差与协方差分量的估计是建立在 G 研究中方差与协方差分量估计的基础之上。对有 7 位考官测量侧面的当前实施方案进行 D 研究，应用 mGENOVA 软件估计方差和协方差，结果见表 2。

表 2　D 研究当前实施方案的方差与协方差分量估计

效应	除数	测评要素 1	测评要素 2	测评要素 3	测评要素 4
P	4.91	0.8	0.72	0.96	
（被试）	3.53	3.91	0.78	0.83	
	2.97	2.86	3.46	0.77	
	2.29	1.76	1.55	1.16	
R	7	0.64			
（考官）	7	0.62	0.73		
	7	0.61	0.66	0.63	
	7	0.19	0.22	0.21	0.18
P × R	7	0.59			
（考 × 生）	7	0.27	0.69		
	7	0.17	0.28	0.55	
	7	0.15	0.15	0.16	0.21

注：上表矩阵中对角线上的元素为方差分量，对角线以上元素为相关系数，对角线以下元素为协方差分量。

由上表可知，考官效应（R）及考官与被试的交互效应（P × R）的方差分量远小于被试效应（P）的方差分量。这样的结果说明，这次面试测评的误差得到了较好的控制。

6.1.2　D 研究当前实施方案的各测评要素误差以及各效应在 4 项测评要素上的类信度系数

继续使用 mGENOVA 软件，计算当前实施方案各指标全域分数（Universe Score）的误差估计及类信度系数等指标，结果分列于表 3 和表 4。

表 3　当前实施方案的各测评要素误差估计值

误差项目	测评要素 1	测评要素 2	测评要素 3	测评要素 4
Universe Score	4.91	0.8	0.72	0.96
（全域分数）	3.53	3.91	0.78	0.83
	2.97	2.87	3.46	0.77
	2.29	1.76	1.55	1.16
Relative Error	0.59	0.41	0.3	0.41

误差项目	测评要素 1	测评要素 2	测评要素 3	测评要素 4
（相对误差）	0.27	0.69	0.45	0.39
	0.17	0.28	0.55	0.45
	0.15	0.15	0.16	0.21
Absolute Error	1.24	0.67	0.65	0.48
（绝对误差）	0.89	1.43	0.72	0.5
	0.78	0.94	1.18	0.53
	0.33	0.37	0.36	0.39
Errorfor Mean	0.71	0.9	0.93	0.58
（均值误差）	0.67	0.79	0.96	0.63
	0.65	0.7	0.68	0.62
	0.22	0.25	0.23	0.19

注：全域分数矩阵中对角线上的元素为方差分量，对角线以上元素为相关系数，对角线以下元素为协方差分量。

由表 3 可知，4 项测评要素的全域分数的相关及其协方差都相对较大，表明 4 项测评要素的相关程度很高，可将 4 项测评要素上的得分合成，可对面试总分进行整体分析。

表 4 当前实施方案各测评要素全域分数的 G 系数

项目	测评要素 1	测评要素 2	测评要素 3	测评要素 4
全域分数方差分量	4.91	3.91	3.46	1.16
相对误差方差分量	0.59	0.69	0.55	0.21
绝对误差方差分量	1.24	1.43	1.18	0.39
均值误差方差分量	0.71	0.79	0.68	0.19
G 系数	0.89	0.85	0.86	0.84
可靠性指数（PHI 系数）	0.80	0.73	0.75	0.75

从表 4 可知，4 项目测评要素的类信度系数都在 0.8 以上（可靠性指数也都在 0.7 以上），均值为 0.86。这说明每项测评要素的评估误差均很小，各项测评要素的评估信度较好。

6.1.3 当前实施方案合成分数的 G 系数

根据表 1 和表 3 的结果，各项测评要素之间的相关系数较高，说明将各项测评要素上的得分合成，可对总分进行分析。按各项测评要素上的原始得分合成，并计算其全域合成分数（Composite Universe Score）的类信度系数。结果显示，本次面试测评的全域合成分数的类信度系数高达到 0.91。这表明，从总体上看此次面试测评的评估误差小，信度很高。同时，这一合成分数的类信度系数高于未合成前的 4 项测评要素的类信度系数，说明核分时将 4 项分数合成在统计上具有一定合理性。

6.2 考官数量对各项测评要素及全域合成分数类信度系数的影响

6.2.1 考官数量对各测评要素全域分数的类信度系数的影响

为了研究考官侧面中考官数量对各项测评要素测量信度的影响，采用 6.1.2 中相同的方法，计算考官数量在从 1 位到 12 位中取值时，各测评要素相应的类信度系数，结果见表 5。

表 5　各测评要素全域分数的类信度系数(G 系数)及增量(△)

考官数量	测评要素 1		测评要素 2		测评要素 3		测评要素 4	
	G 系数	△	G 系数	△	G 系数	△	G 系数	△
1	0.54	–	0.45	–	0.47	–	0.44	–
2	0.70	0.16	0.62	0.17	0.64	0.17	0.61	0.17
3	0.78	0.08	0.71	0.09	0.73	0.09	0.70	0.09
4	0.83	0.05	0.76	0.05	0.78	0.05	0.76	0.06
5	0.86	0.03	0.80	0.04	0.82	0.04	0.80	0.04
6	0.88	0.02	0.83	0.03	0.84	0.02	0.82	0.02
7	0.89	0.01	0.85	0.02	0.86	0.02	0.84	0.02
8	0.90	0.01	0.87	0.02	0.88	0.02	0.86	0.02
9	0.91	0.01	0.88	0.01	0.89	0.01	0.88	0.02
10	0.92	0.01	0.89	0.01	0.90	0.01	0.89	0.01
11	0.93	0.01	0.90	0.01	0.91	0.01	0.90	0.01
12	0.93	0.00	0.91	0.01	0.91	0.00	0.90	0.00

从表 5 可知,随着考官数量的增加,4 项测评要素的类信度系数均增大,也即 4 项测评要素的测量误差均逐步减小。考官数量从 1 增至 5 时,4 项测评要素的类信度系数均有大幅度提高,且当考官数量为 5 时,各项测评要素的类信度系数均在 0.80 以上。当考官数量从 5 增加到 9 时,各项测评要素的类信度系数仍有一定程度的增加,且当考官数量为 9 时,各项测评要素的类信度系数接近 0.90。但是,当考官数量从 9 开始增加时,各项测评要素的类信度系数增加很小。以上结果表明,结构化面试考官数量控制在 5 ~ 9 位之间比较合适,不能少于 5 位,多于 9 位则对提高信度的作用不大,且增加了考官成本,在 5 ~ 9 位之间,考官数量越多,测量误差越小,测量信度越高。

从表 5 还可看出,在考官数量相同时,4 项测评要素类信度系数的大小依次为:自我认知与综合分析、决策应变、组织协调、语言表达与气质。这就是说,在面试测评中,对被试自我认知与综合分析能力的评价误差较小,而对语言表达与气质的评估误差较大。前三项测评要素均有对应的试题及评分参考,而语言表达与气质没有对应的试题,是依据被试在整个面试过程中的综合表现做出评判的,考官之间要在评判标准上把握统一口径的难度更大一些。这可能是造成语言表达与气质类信度系数略小的主要原因。

6.2.2　考官数量对全域合成分数类信度系数的影响

继续分析全域合成分数类信度系数随考官数量变化的规律,结果见表 6。

表 6　全域合成分数 G 系数及增量(△)

考官数量	G 系数	△
1	0.59	–
2	0.74	0.15
3	0.81	0.07
4	0.85	0.04
5	0.88	0.03
6	0.89	0.01
7	0.91	0.02
8	0.92	0.01
9	0.93	0.01
10	0.93	0.00
11	0.94	0.01
12	0.94	0.00

从表 6 可知，随着考官数量从 1 增至 12 时，全域合成分数的类信度系数不断增大，这说明考官数量越多，对被试测评结果的误差越小。从类信度系数的增量上看，当考官数量从 1 增加到 5 时，类信度系数的增量为 0.30，且当考官数量为 5 时，类信度系数达到 0.88，说明信度已经较高。当考官数量从 5 增加到 9 时，信度还有一定程度的增加，增量趋于稳定，当考官数量为 9 时，类信度系数高达 0.93，这时评估的误差已很小。当考官数量从 9 增加到 12 时，信度增量极小，说明考官数量超过 9 时，对提高测量信度意义不大。这些结果与各项测评要素类信度系数随考官数量变化的规律类似。

从表 5 及表 6 可知，在考官数量不变时，4 项测评要素的全域分数合成后的类信度系数均大于未合成的任一单个测评要素的类信度系数。这一结果说明，在面试评判过程中进行分项打分，有利于提高最后合成分数的信度，评价效果更好。

7 结论

研究采用概括化理论，从难度、区分度、信度特别是考官数量对信度的影响等方面，对北京市某区副处级干部准入资格结构化面试测评数据进行了较为全面的分析。主要结论有：

7.1 本次结构化面试的难度适中，区分度较高。

7.2 本次结构化面试中各测评要素和合成分数的测量信度均较高，合成分数的测量信度高于单个测评要素的测量信度，其中，各测评要素的平均类信度系数为 0.86，合成分数的类信度系数为 0.91。这说明在结构化面试中测量误差得到了较好控制。

7.3 结构化面试中各测评要素或合成分数的类信度系数会随着考官数量的增加而提高。从本研究看，考官数量应在 5～9 之间。考官数量少于 5，不能保证较高的信度。考官数量多于 9，信度没有明显提高，且因考官数量增加提高了面试成本。从目前的面试实践来看，考官数量通常采用奇数，这样，考官数量一般为 5、7、9 为宜。当考官数量为 9 时，面试测评信度一般就能达到比较理想的水平。

7.4 本次面试中，各项测评要素的相关系数较高，这为目前在选拔面试中将各项测评要素得分进行合成提供了依据，说明分数合成计算总分具有一定的合理性。如果各测评要素间的相关系数小，则不宜强行合成。同时，测评要素间高相关的结果也表明了各测评要素间的辨别效度较低，各项测评要素实际的评价结果可能代表了同一能力素质，而不是不同的能力素质。这有背于面试设计中努力测试不同能力要素的初衷。实际上，在很多面试中都发现了类似情况[5]，需要进一步探讨研究。

7.5 在结构化面试中，各测评要素的测量信度略有差异。其中，自我认知与综合分析能力的测量信度最高，决策应变能力和组织协调能力的测量信度次之，语言表达与气质的测量信度相对略低。

以上结论对更好地组织实施结构化面试尤其是确定考官数量，提供了一定的实证依据。但是，也应注意到这次结构化面试的一些局限性：(1)7 位考官均来自考试测评机构，有较强的专业素质和面试评判技能，而这一点在其它领导干部面试中未必具备；(2)82 位被试均是副处级后备干部，职位层面比较集中，样本量有限。这些结论仅来自对这次结构化面试的分析结果，是探索性的，在推广结论时应持谨慎态度。这些结论是否具有普遍意义，需要进一步验证。

参考文献

1 漆书青. 现代测量理论在考试中的应用. 武汉：华中师范大学出版社，2003.143－184.

2 漆书青，戴海崎，丁树良. 现代教育与心理测量原理. 南昌：江西教育出版社，1998.60－91.

3 中共中央组织部. 领导干部公开选拔和竞争上岗考试大纲. 北京：党建读物出版社，2004.

4 Brennan R L. MGENOVA. Lowa Testing Programs. University of lowa，1999.

5 Schmitt N，Chan D. Personnel Selection：A Theoretical Approach. SAGE Publications，Inc，1998.

(下转第 95 页)

5　Cooper C L，Williams S. Occupational Stress Indicator. Version 2. England. North Yorkshire：RAD Ltd.
6　Siu O L，Cooper C L. Managerial Stress in HongKong and Taiwan：A comparative study. Journal of Managerial Psychology，2002，14(1) 6－25.
7　周跃萍，等.不同职业人员工作压力源及压力反应的比较研究. 心理学探新，2004，23(1)：63－65.

A Study of Occupational Stressor，Coping Style and Mental & Physical Well－being in IT Firm

Zhang Xichao　Lian Xu　Che Hongsheng
(Beijing Normal University，Beijing 100875)

Abstract：The purports of this research is to explore predictors of mental & physical well－being among many job stressors，and to find the impact of coping style on mental & physical well－being. The results were as follow：all job stressors and controlling coping style had significant correlation with mental & physical well－being. The predictors of male's mantlewell－being were relationships，home & work balance，role ambiguity and career development，and controlling coping style acted as an mediator between role ambiguity and mantle well－being. The predictors of male's physical well－being were workload，role conflict and controlling coping style. The predictor of female's mantle well－being was home & work balance. The predictor of female's physical well－being was workload.
Key words：occupational stressor；coping style；mental health；physical well－being

(上接第 90 页)

Reliability Analysis of Structured Interview：A Multivariate Generalizability Theory Approach

Hong Ziqiang[1]，Tu Dongbo[2]
(1. China Center for Leadership Assessment and Selection，Beijing 100036；
2. Education College，Jiangxi Normal University，Nanchang 330027)

Abstract：This paper attempted to apply the MGT in analyzing the reliability of structured interview. The results showed that，(1) the structured interview is moderately difficult and highly discriminated，(2) the reliabilities of both single and composite measurement aspects are high，(3) the reliability of structured interview increases with the number of assessors (five to nine assessors are appropriate when balancing reliability and cost)，and (4) measurement aspects in structured interview are highly interrelated，which makes composite score reasonable. The implications for improving the present practice of selection interview are discussed.
Key words：Structured interview；MGT；Reliability