China Family Panel Studies

CFPS

中国家庭追踪调查

技术报告系列：CFPS-31

系列编辑：谢宇　责任编辑：张聪

中国家庭追踪调查

2012 年数列测试题

徐宏伟　谢宇

2015.1.7

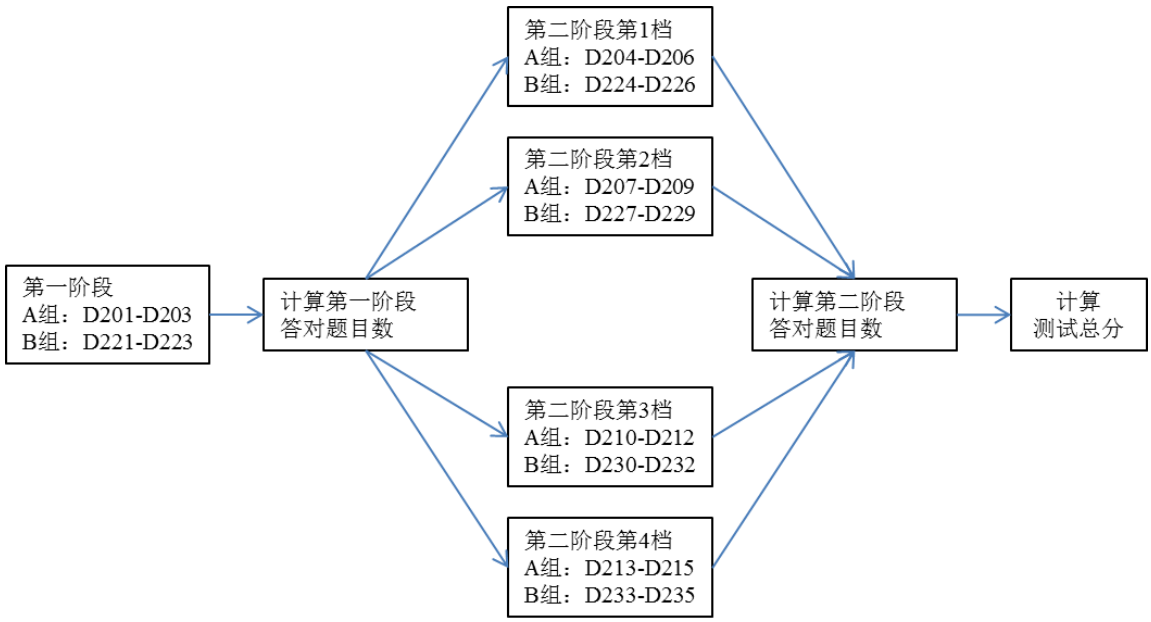
一、数列测试题简介

根据调查设计，CFPS2012 的数列测试以所有 10 岁及 10 岁以上的被访者为测试对象。在进行测试前，调查员会为所有满足年龄条件的被访者提供两道介绍性质的题目：介绍题目 Q1 和介绍题目 Q2。在做完这两道题目之后，由被访者自己决定是否参加接下来的数列测试。决定参加数列测试的被访者会被随机分成两组：A 组和 B 组。CFPS2012 为 A、B 两组被访者准备了两套内容不同、难度相当的题目。A 组对应的一组数列测试题目为 D201-D215，B 组对应的一组数列测试题目为 D221-D235，每组题目的总数都是 15 道, 但如下介绍的两阶段测试设计让每位受访者实际回答的题数为 6 道。

所有决定接受数列测试的被访者都会接受一个分为两阶段的“适应性测试（adaptive test）”：在第一个阶段，被测试者需要作答 3 道题目（A 组为 D201-D203；B 组为 D221-D223）。在第一个阶段回答正确的题目数将被用来确定第二阶段被测试者需要作答的题目。对应于在第一个阶段所有可能的回答正确的题目数：0 题、1 题、2 题和 3 题，被测试者在第二个阶段会被相应地分配到由易到难排列的第 1 档题目、第 2 档题目、第 3 档题目和第 4 档题目，每档题目均为 3 道。例如，A 组的某位被测试者在第一个阶段回答正确的题目数为 0，那么他在第二阶段要作答难度最低的第 1 档题目：D204-206；A 组的另一位被测试者在第一个阶段正确作答了全部 3 道题目，那么他在第二阶段的测试要作答难度最高的第 4 档题目：D213-D215。同样，若是 B 组中有两位被测试者在第一个阶段分别答对了 0 道和 3 道题目，那么他们在第二个阶段将对应作答第 1 档题目 D224-D226 和第 4 档题目 D233-D235。

被测试者在第一阶段答对的题目数或在第二阶段作答哪一个难度档的题目，以及在第二个阶段正确作答的题目数，共同决定了被测试者最终在该项测试上的得分。图 1（摘引自 Fisher et al. 2014）给出了详细的测试流程，并列出了两个被测试组分别在各阶段要回答的题目。

图 1 CFPS2012 数列测试流程图



二、对一道出错的测试题目（D226）的作答填补

由于拼写错误，在 B 组所对应的测试题目中，位于第二阶段第 1 档的题目 D226 没有正确答案。在假设 A 组和 B 组各阶段题目难度相当的前提之下，我们利用 A 组测试者在第二阶段第 1 档相应题目（D206）上的回答情况来对 B 组测试者回答 D226 的回答情况进行填补。具体过程如下：首先，在 A 组测试者回答第二阶段第 1 档题目（D204-D206）的样本中，我们对 D206 的回答情况（0=错误；1=正确）做二分 Logit 回归模型（模型所包含的协变量参见表 1）；然后，根据该模型得出 A 组测试者中相应人群对题目 D206 回答正确的概率，概率大于或等于 0.5 的测试者被预测为答对了该题目，概率小于 0.5 的则被预测为答错了该题目。这样，回答了 D206 的 A 组测试者同时会得到一组模型预测出来的 D206 回答情况。而同时，在满足 A 组和 B 组各阶段题目难度相当的假设前提之下，我们用模型预测出来的 A 组测试者对 D206 的回答情况来填补 B 组测试者在相应题目 D226 上的回答情况。

表 1 对 A 组测试者中回答第二阶段第 1 档题目（D204-D206）D206 的二分 Logit 模型 (N = 618)

	系数		标准误
第二阶段第 1 档题目中前两题的回答情况（参照组：两题都错）			
D204 错；D205 对	2.786	**	0.936

D204 对；D205 错	2.479	**	0.910
两题都对	3.400	***	0.751
介绍题目 Q1（参照组：回答错误）	0.132		0.508
介绍题目 Q2（参照组：回答错误）	1.046	**	0.326
年龄（均值对中）	-0.437	***	0.110
年龄平方	0.107	*	0.046
男性	0.228		0.310
2010 年字词测试得分	-0.022		0.024
2010 年数学测试得分	0.101	*	0.040
居住地（参照组：城市）			
乡镇	1.709	**	0.644
农村	1.027	†	0.599
常数项	-7.301	***	1.107

注：† P < .1; *P < .05; **P < .01; ***P < .001

表 1 给出了对 A 组测试者中回答 D206 对错情况的二分 Logit 回归模型。总的来看，我们选择的协变量大部分可以显著影响到测试者对 D206 的回答情况。值得注意的是，A 组测试者对第二阶段第 1 档题目前两题，即 D204 和 D205 的回答情况，可以显著影响到对随后第三题 D206 的回答情况。

表 2 A 组测试者中回答第二阶段第 1 档题目（D204-D206）D206 的实际观测值和模型预测值分布对比

实际观测答案	模型预测答案		总合
	错误	正确	
错误	543	9	552
正确	56	10	66
总合	599	19	618

表 3 A 组测试者和 B 组测试者中各回答第 1 档题目（分别为 D206 和 D226）的模型预测情况对比

	A 组		B 组	
	N	%	N	%
回答错误	599	96.9	873	94.2
回答正确	19	3.1	54	5.8
总合	618	100.0	927	100.0

表 4 第二阶段测试题回答情况汇总

		A 组		B 组	
	答对题目数	N	%	N	%
第 1 档 [a]	0	270	36.4	359	32.5
	1	115	15.5	171	15.5
	2	286	38.6	521	47.1
	3	70	9.5	54	4.9
第 2 档	0	179	6.1	543	17.5
	1	878	30.1	1,154	37.2
	2	1,213	41.6	1,108	35.7
	3	649	22.2	297	9.6
第 3 档	0	562	31.7	1,248	51.1
	1	763	43.1	840	34.4
	2	370	20.9	280	11.5
	3	77	4.3	73	3.0
第 4 档	0	1,097	34.7	634	34.3
	1	922	29.1	541	29.3
	2	824	26.0	380	20.6
	3	321	10.2	292	15.8

注：[a]此处 B 组测试者第二阶段第 1 档题目的回答情况是经过模型预测填补后的结果。

表 2 对比了 A 组测试者中回答 D206 的实际观测答案分布和用上述模型预测得出的模型预测答案分布。整体来讲，模型预测答案和实际观测答案的匹配程度达到 89.5%（(543+10)/618），测试者实际回答错误而模型预测为回答正确的占 1.5%（9/618），测试者实际回答为正确而模型预测为回答错误的占 9.1%（56/618）。也就是说，我们的模型倾向于低估 A 组测试者答对 D206 的可能性，这是因为在实际观测到的数据中 D206 的正确率比较低[1]：在 741 位理应回答 D206 的被访者中（即 A 组第 1 档），只有 83 人给出了正确的答案，折合约 11.2%的答对比例。

在用模型预测对 D226 的回答情况进行填补之后，表 3 进一步对比了基于模型预测的 A 组测试者回答 D206 的对错情况和 B 组测试者回答 D226 的对错情况。模型预测结果为，回答 D226 的 B 组测试者答对比例为 5.8%，高于回答 D206 的 A 组测试者 3.1%的答对比例。

[1] 我们还尝试了基于最大似然估计的多重填补方法（Multiple Imputation），对于 A 组测试者答对 D206 的可能性依然是低估的。此处我们选择报告了相对较简单的 Logit 回归模型结果。

表 4 汇总了两组测试者在第二阶段对各档题目的回答情况。卡方检验结果显示，A 组测试者和 B 组测试者在相应档次题目上的答对题目数分布，存在着显著的差别。这意味着 A 组和 B 组两套题目在难易程度上可能存在差别。

三、对数列测试题无应答情况的调整

在符合参加数列测试条件的样本中，有大约一半的人没有选择参加测试。我们用二分 Logit 模型拟合样本对数列测试题的参与情况，然后预测出样本中每个人参与数列测试题的概率或倾向分（propensity），最后用该倾向分的倒数对数据进行加权。

表 5 的 Logit 模型显示，对数列测试题目的应答情况确实具有选择性。例如，男性比女性参与数列测试的可能性更高；教育水平越高、家庭经济水平越高的人，参与数列测试题的可能性也倾向于越高。

表 5 对数列测试题参与情况（0=未参与；1=参与）的 Logit 回归模型

	系数		标准误
年龄 （均值对中[a]）	-0.291	***	0.010
年龄平方	0.028	***	0.004
男性	0.197	***	0.027
教育水平（参照组：小学以下）			
小学	0.945	***	0.047
初中	1.553	***	0.046
高中	1.860	***	0.054
大学及以上	2.141	***	0.069
家庭人均收入分位数（参照组：最低 25%）			
中下 25%	0.184	***	0.038
中上 25%	0.273	***	0.038
最高 25%	0.395	***	0.041
记忆测试得分	0.209	***	0.008
城镇	-0.015		0.028
常数项	-2.306	***	0.056
N	31911		

注：*** p<.001

[a]均值对中（mean-centering），即把每个人的年龄减去整个样本的平均年龄后，得到的转换后的“年龄”；该“年龄”的 0 值对应的是转换前的样本年龄均值。

表 6 对于数列测试题最终得分给出了未经加权和经过加权两套描述性结果。不同于 CFPS 的问卷设计，我们在此采用常规的 18 岁作为儿童与成年人子样本的年龄分界。数列测试题得分的计算采用了两种方法：（1）哥特曼量表法（Guttman Scale），基于该方法得出的数列测试题得分范围为 0-15 分；（2）W-测量法（W-score 或 W-scale），基于该方法得出的最终得分是定距的，例如，得分为 390 分和 400 分的差距，被认为是与 500 分和 510 分之间的差距完全相同的。关于两种测量方法更详细的介绍，可参见 Fisher et al. (2014)。如表 6 所示，对数列测试题目无应答做出调整会显著影响到 18 岁及 18 岁以上的成年人样本最终的数列测试得分。例如，未调整无应答时，成年人样本的 W-得分均值为 515.8，这一数字在概率倒数加权后降为 504。相较于成年人，由于 10-17 岁的儿童样本本身对数列测试的参与率较高，他们的数列测试得分在加权后几乎不会发生变化。

表 6 未加权和以概率倒数加权后数列测试题得分对比

	未加权		概率倒数加权	
	儿童	成人	儿童	成人
W-得分 [a]				
均值	528.6	515.8	527.6	504.0
标准差	26.5	34.9	27.1	41.4
W-得分标准误				
均值	10.9	11.4	10.9	12.0
标准差	2.3	2.7	2.3	3.1
哥特曼量表得分 (0-15)				
均值	9.5	7.8	9.4	6.7
标准误	3.5	4.0	3.6	4.2
N	2827	14264	3466	28292

注：[a]计算方法参照了 Fisher et al. (2014)；此处儿童为 10-17 岁，成人为 18 岁及以上。