

已经制作了以下书籍，可在 <http://www.douban.com/group/topic/14833042/> 或 <http://blog.sina.com.cn/u/1408815407> 查阅更新和下载，或加入科学心理学与书籍交流群，群号 42506864

1.进化心理学 by 巴斯 2. 社会生物学:新的综合 by wilson 3.谁会认错？我们为什么会有如此愚蠢的信念错误的决定和让别人受伤的行为 4.社会心理学-阿伦森版 5. 布鲁克林有棵树 6. 学习的艺术 维茨金 著 7. 疯狂实验史 8. 哪来的天才 9.为什么选错的总是我 10.思维改变生活 11. 积极心理学：关于人类幸福和力量的科学 12.欲望之源 13. 吃的真相 14. 直觉——你所不知的潜力与危害 by 迈尔斯 15.进化的大脑 16. 态度改变与社会影响 by 津巴多 17.害羞心理学 by 津巴多 18. 简捷启发式 19. 学习乐观 by 塞利格曼 (现有的版本是超星的，不太清晰) 20. 认知心理学 第五版 艾森克著 21. 实验心理学（原书第八版）索尔索 22. 生命的心流 23.幸福的真意 24. 别太苛求：摆脱完美主义的束缚 25. 批判性思维:思维、写作、沟通、应变、解决问题的根本技巧 26. 为什么大猩猩比专家高明(How we decide) 27. 心理和脑 脑与心智历程 100 项 28. 心理和脑与生活 训练脑与心智的 75 项窍门 29. 结构方程模型: AMOS 的操作与应用 30.颠倒思维 31.动机心理学 皮特里著 32.决策的艺术 33. 改变: 问题形成和解决的原则 34.象与骑象人 35.心理学研究方法 by 舒华 36.美丽圣经 by 宝拉·培冈 37.撬动幸福 by 奚恺元 38.津巴多普通心理学 39.思维与决策(第四版) 40. 内向者优势 如何在外向的世界中获得成功 41. 行为科学统计 第七版 42. 不确定状况下的判断: 启发式和偏差

图片来自读秀网，我用的是简单的办法——咨询下载，但每次只能咨询 50 页，咨询时间选在晚上 22 点以后可以免受咨询不得超过 80%的限制。具体咨询下载的方法我会公布在我的博客上。

快捷的办法是直接从掌握读秀、超星漏洞的那些人手里购买，价格基本在 1 元/100 页左右（taobao 上搜索读秀或超星 咨询 下载即可，可以讲价的），或者在读书宝库、诺贝尔学术网等论坛上通过论坛币来求助。

以上书籍之外，再推荐些绝对值得一看的书籍：

心理学类:社会性动物 少有人走的路 社会心理学(迈尔斯版) 决策与判断 与众不同的心理学 亲密关系 影响力 20 世纪最伟大的心理学实验 人类的性存在

其他类:如何阅读一本书 学会提问:批判性思维指南 把时间当做朋友 神奇的睡眠 中国健康调查报告，别让不懂营养学的医生害了你 营养圣经-最佳营养学指南 外语学习的真实方法及误区 异类

社会学译丛
学术经典系列

不确定 状况下的 判断

启发式和偏差

Judgment under uncertainty:
Heuristics and biases

丹尼尔·卡尼曼
Daniel Kahneman

[美] 保罗·斯洛维奇 编
Paul Slovic

阿莫斯·特沃斯基
Amos Tversky

方文 吴新利 张肇 等译

方文 校

中国人民大学出版社



图书在版编目 (CIP) 数据

不确定状况下的判断：启发式和偏差/

(美) 卡尼曼等编；方文等译.

北京：中国人民大学出版社，2007

(社会学译丛·学术经典系列)

ISBN 978-7-300-08575-3

I. 不…

II. ①卡…②方…

III. 决策-判断-文集

IV. C934-53

中国版本图书馆 CIP 数据核字 (2007) 第 152613 号

社会学译丛·学术经典系列

不确定状况下的判断：启发式和偏差

丹尼尔·卡尼曼

[美] 保罗·斯洛维奇 编

阿莫斯·特沃斯基

方文 吴新利 张擘 等译

方文 校

出版发行 中国人民大学出版社

社 址 北京中关村大街 31 号

邮政编码 100080

电 话 010-62511242 (总编室)

010-62511398 (质管部)

010-82501766 (邮购部)

010-62514148 (门市部)

010-62515195 (发行公司)

010-62515275 (盗版举报)

网 址 <http://www.crup.com.cn>

<http://www.ttrnet.com> (人大教研网)

经 销 新华书店

印 刷 北京新丰印刷厂

规 格 150 mm × 230 mm 16 开本

版 次 2008 年 9 月第 1 版

印 张 40.5 插页 2

印 次 2008 年 9 月第 1 次印刷

字 数 598 000

定 价 79.80 元

版权所有 侵权必究 印装差错 负责调换

目 录

序言	1
第一部分 引论	1
第1章 不确定状况下的判断：启发式 和偏差	2
第二部分 代表性	21
第2章 小数律的信念	22
第3章 主观概率：代表性判断	32
第4章 论预测心理学	49
第5章 对代表性的研究	71
第6章 有关代表性的判断和依据代表性 的判断	88
第三部分 因果性和归因	105
第7章 普通的归纳：信息不必然使人	

不确定状况下的判断

知情	106
第 8 章 不确定状况下的判断中的因果图式	123
第 9 章 归因过程中的缺陷：错误的社会评估的起源和维持	137
第 10 章 基率的证据式影响	162
 第四部分 便利性	173
第 11 章 便利性：用于频率和概率判断的启发式	174
第 12 章 便利性和归因中的自我中心偏差	193
第 13 章 社会知觉和互动中的便利性偏差	204
第 14 章 模拟启发式	216
 第五部分 共变和控制	225
第 15 章 非正式的共变评估：基于数据的判断与基于理论的 判断的比较	226
第 16 章 控制的错觉	247
第 17 章 测验结果的随意性	257
第 18 章 临床医学中的概率推理：问题及机遇	268
第 19 章 从经验中学习及决策中的次优规则	290
 第六部分 过分的信心	309
第 20 章 在个案研究的判断中的过分信心	310
第 21 章 关于概率评估训练的进展报告	318
第 22 章 概率的校准：从当前的状况（1977）到 1980	333
第 23 章 致那些注定要研究过去的人：事后诸葛亮的启发式 和偏差	365
 第七部分 多阶段评估	385
第 24 章 序列选择中复合概率的估计	386
第 25 章 人类信息处理中的保守主义	391

第 26 章 多阶段推理中的最佳推测假设	404
第 27 章 根据回忆信息推论个人特征	414
第八部分 校正程序	425
第 28 章 决策中非常规线性模型的优美之处	426
第 29 章 虚构数字的生命力	445
第 30 章 直觉预测：偏差和校正程序	453
第 31 章 去偏差过程	462
第 32 章 改善归纳推理	488
第九部分 风险知觉	505
第 33 章 事实对恐惧：理解可觉知的风险	506
第十部分 附录	537
第 34 章 论统计直觉的研究	538
第 35 章 不确定性的变式	557
参考文献	571
索 引	627
译后记	638



序 言

本书所代表的有关判断研究的路径，其源头是在20世纪50年代和60年代所发展的三条研究线索：由米尔（Paul Meehl）所开创的有关临床预测和统计预测的比较；由爱德华兹（Ward Edwards）导入心理学而体现在贝叶斯范式（Bayesian paradigm）中的有关主观概率的研究；由西蒙（Herbert Simon）提供规划并由布鲁纳（Jerome Bruner）提供例证的有关推理启发式和策略的研究。我们的文集也代表了判断研究和另一种心理研究在近期的汇合：由海德（Fritz Heider）作为先驱的有关因果归因和常人心理解释的研究。

米尔的经典著作出版于1954年。他在总结大量证据的基础上得出结论：在预测显著的行为标准时，有关线索的简单线性组合，优于专家的直觉判断。这项研究持久的理智遗产，以及与之相随的激烈论争，也许并不是临床医生在临床预测的工作中表现不佳的明证，正如米尔所指出的，这些工作并非他们所长。

相反，它是在预测任务中人们成功的客观记录和他们对其表现性质的主观信念之间存在实质性偏离的明证。这个结论，并不仅仅局限于临床医生或临床预测：人对自己如何推理以及推理的成效的印象，并不是显而易见的。也许因为临床判断的学生经常用自己和朋友作为研究对象或被试，他们对错误和偏差的解释，往往是认知性的，而不是心理动力学的：是错觉而不是幻觉，作为错误和偏差的模板。

xvi 由于爱德华兹和其助手把贝叶斯观念导入心理学的研究中，心理学家第一次被提供了在不确定状况下有关最优绩效的完全清晰的模型，援用这个模型，可以比较人的判断。在对不确定状况下的判断研究中，人的判断和规范模型之间的匹配，成为一个主要范式。毋庸置疑，它导致对归纳推理易犯的偏差以及可校正的方法的关注。这种关注，反映在本书所选择的大部分文章中。然而，早期的大部分研究采用规范模型来解释人的绩效，但又引用不同的过程来解释与最优性（optimality）的偏离。与之相反，有关判断启发式的研究，则依据同样的心理过程，致力于解释正确和错误的判断。

认知心理学新范式的突现，对判断研究有深远的影响。认知心理学关注内在过程和心理限度，以及内在过程受心理限度所形塑的方式。在这一脉络中，概念研究和经验研究的早期例证，是布鲁纳和其助手对思维策略的研究以及西蒙对推理启发式和有限理性（bounded rationality）的研究。布鲁纳和西蒙都关注能降低判断任务复杂性的简单化策略（strategies of simplification），从而使人所拥有的心智能驾驭这些任务。本书所包含的大部分研究，受同样的关注所激发。

近年来，大部分研究已致力于揭示判断启发式，探索其后果。本书提供了这一研究路径的完备样本。它汇集了为本文集专门撰写的新评论，以及已出版的有关判断和推理的论文。尽管判断和决策之间的界限并不总是很清晰，我们这里专注于判断而不是选择。决策的主题足够重要，从而需要由另一本书来处理。

本书分成十个部分。第一部分，包含有关直觉判断的启发式和偏差的早期评论；第二部分，则具体关注代表性启发式（the representativeness

heuristic);而在第三部分中,它被拓展至因果归因的问题;第四部分,描述便利性启发式(availability heuristic)及其在社会判断中的角色;第五部分,则覆盖有关协方差(covariation)的知觉和学习的问题,并且详尽说明在常人和专家的判断中存在伪相关;第六部分,讨论概率评估的度量,并且证明在预测和解释中普遍存在的信心过分(overconfidence)的现象;第七部分,讨论与多阶段推理有关的偏差;第八部分,评论校正和改善直觉推理的正式和非正式的程序;第九部分,在特定的关注领域即风险知觉领域中,总结有关判断偏差的效应的研究;而最后的部分,则包括与启发式和偏差研究有关的概念和方法论难题的当前思考。

为方便起见,所有参考文献置于书后。文中的黑体数字,是指本书所引的材料在本书中的章节序号。我们用省略号(……)来指代所选文章的删节部分。

xiii

在准备本书的过程中,我们的研究工作得到如下资助:海军研究办公室给斯坦福大学的资助(N00014-79-C-0077)以及海军研究办公室给决策研究(Decision Research)的资助(N0014-80-C-0150)。

我们感谢下列人士在本书准备过程中所给予的帮助:Peggy Roecker、Nancy Collins、Gerry Hanson 和 Don MacGregor。

卡尼曼 (Daniel Kahneman)

斯洛维奇 (Paul Slovic)

特沃斯基 (Amos Tversky)



第一部分
PART ONE

引 论



第 1 章

不确定状况下的判断：启发式和偏差

特沃斯基和卡尼曼 (Amos Tversky and Daniel Kahneman)

许多决策基于不确定事件的可能性的信念，这些事件包括诸如一次选举的结果、被告的罪行或美元的未来价值。这些信念，常常以这样的陈述表现出来：“我认为……”“机会是……”或“它是不可能的……”诸如此类。偶尔，有关不确定事件的信念，以数字形式如几率 (odds) 或主观概率 (subjective probabilities) 来表达。这些信念的决定因素是什么？人们如何估计不确定事件的概率或者不确定量值的数值？本章将表明人们依赖数目有限的启发式原则，以把概率估计和数值预测的复杂任务降低为较简单的判断操作。通常，这些启发式原则很有用，但有时，它们也会导致严重的和系统的错误。

概率的主观估计，与物理量值如距离或尺寸的主观估计类似。这些判断，都基于效度有限的数据或资料，其加工依据启发式规则。例如，物品的表面距离部分受制于其明晰度。物品看起来越清晰，似乎越

近。这个规则有一定效度，因为在任何给定的视域中，较远的物品比之较近的物品，看起来更不清晰。然而，在对距离的估计中，依赖这个规则，会导致系统的错误。尤其在能见度很差时，距离会被高估，因为物品的轮廓很模糊。而另一方面，在能见度很好时，距离会被低估，因为物品的轮廓很清晰。因此，依靠明晰度作为距离的指标，会导致惯常的偏差（biases）。而这些偏差，也会在直觉性的概率判断中被发现。本章将描述被用来估计概率和预测数值的3种启发式。我们将清查这些启发式所导致的偏差，并且讨论这些研究报告的应用和理论内涵。

一、代表性

人们所关注的许多概率问题，都可归于下列类型中的一种：物品A属于类别B的概率是多少？事件A源于过程B的概率是多少？而过程B产生事件A的概率又是多少？在回答这些问题时，人们典型地依赖代表性启发式，即A代表B的程度或者A类似B的程度，来评估概率。例如，当A高度代表B时，A源于B的概率，就被判断或认为是高的。另一方面，如果A和B并不类似，A源于B的概率，就被认为是低的。

为了详细说明依据代表性的判断，请考虑这样的一个人，他以前的邻居对他的描述如下：“斯迪夫，害羞而畏缩，乐于助人，但对人或事没有多少兴趣。一个温顺而清洁的灵魂。他渴望秩序和结构，对细节充满激情。”人们如何从一系列可能的职业名单（如农民、商人、驾驶员、图书管理员或物理学家）中推测他的特定职业？按照可能性的高低，如何排列这些职业？依照代表性启发式，斯迪夫是图书管理员的概率，通过他代表或类似图书管理员的刻板观念（stereotype）的程度，而被评估。事实上，这一类问题的研究已经证明，人们用同样的方式通过概率和相似度（similarity）来排列职业（Kahneman & Tversky, 1973, 4）。这种概率判断的路径，会导致严重的错误，因为相似度或代表性，并不受影响概率判断的几个因素的制约。

1. 对结果的先验概率不敏感

对代表性没有影响但对概率有主要影响的一个因素，就是结果的先验概率（prior probability），或者是结果的基率频率（base-rate frequency）。就

斯迪夫的例子而言，在总体人口中，农民比图书管理员要多得多，而这个事实，应该在推断斯迪夫是个图书管理员而不是农民的概率时，被合理地考虑到。然而，基率频率的考虑，并不影响斯迪夫和图书管理员或农民的刻板观念之间的相似性。因此，如果人们通过代表性来估计概率，先验概率会被忽视。我们在一个实验中检验了这个假设，在其中，先验概率被操纵（Kahneman & Tversky, 1973, 4）。给被试呈现几个人的概要人格描述，他们被宣称是从由工程师和律师组成的 100 人的群体中随机抽取的。要求被试基于每个描述来推测他是工程师而不是律师的概率。在一种实验条件下，告知被试群体中有 70 个工程师和 30 个律师；在另一种实验条件下，告知被试群体中有 30 个工程师和 70 个律师。在第一种条件下，任何特定描述属于工程师而不是律师的几率都要高，因为其中工程师是多数。应用贝叶斯规则，每个描述的几率比率是 $(0.7/0.3)^2$ 即 5.44。但结果严重违背了贝叶斯规则，两种条件下的被试的概率判断，本质上相同。很明显，被试评估特定描述是属于工程师而不是律师的可能性，是依靠这个描述对两种刻板观念的代表性状况来进行的，而没有关心范畴或类别的先验概率。

而当被试没有其他信息时，他们会正确地利用先验概率。没有概要的人格描述，他们在两种不同的基率条件下，依次判断特定对象是工程师的概率为 0.7 和 0.3。而当描述甚至是完全无用的描述被引入后，先验概率则被严重地忽视。对下列描述的反应，证明了这种现象：

迪克 30 岁，已婚无子女。他能力强，有干劲，想在自己的领域中成就一番事业。他受同事喜爱。

这个描述，本意上不包含与迪克是工程师还是律师有关的任何信息，结果应该是迪克作为工程师的概率和群体中工程师的比率相等，正如没有人格描述的那样。但被试不管群体中工程师的比率是 0.7 还是 0.3，都把迪克是工程师的概率判断为 0.5。显然，在面对无证据或无用证据的情形时，人们的反应不同。没有特定证据时，人们会恰当地利用先验概率；而有无用证据时，人们会忽视先验概率（Kahneman & Tversky, 1973, 4）。

2. 对样本规模不敏感

在从特定整体抽出的样本中，为了评价特定结果所获得的概率，人们典型地应用代表性启发式。或者说，他们通过这个结果和对应参数（即男人整体中的平均身高）的相似性，来评价样本结果的可能性，如 10 个男人的随机样本的平均身高是 6 英尺（180 厘米）的可能性。样本指数（statistic）和整体参数的相似性，并不依赖于样本规模。因此，如果概率由代表性来评价，那么，所判断的样本指数的概率，本质上就独立于样本规模。事实上，当被试评价不同规模的样本的平均身高的分布时，他们得出同样的分布。例如，无论样本规模是 1 000、100 还是 10，要获得男人平均身高高于 6 英尺的概率，对他们而言是一样的（Kahneman & Tversky, 1972b, 3）。并且，甚至在强调样本规模在问题中的重要性时，被试也不能体会到它的作用。请思考下面的例子：

某镇有两个医院，大医院每天约 45 名婴儿出生，小医院每天约 15 名婴儿出生。众所周知，约 50% 的婴儿是男孩。但每天男孩出生的实际比例都不一样，有时高于 50%，有时低于 50%。

在一年内，每家医院都记录新生儿中男孩比例大于 60% 的天数。请问你认为哪家医院会记录更多这样的天数？

大医院 (21)

小医院 (21)

大致相同（也就是在 5% 的范围内）(53)

括号中的数值，是选择该答案的大学生被试的数目。

大部分被试认为，60% 以上的新生儿是男孩的概率，在大医院和小医院都是一样的。这似乎是因为这些事件是用同样的指数来描述的，因此也同样代表一般的整体。相反，抽样理论，则要求 60% 以上的新生儿是男孩的天数，更可能是在小医院中发生的，因为大样本更不可能偏离 50%。这种基本的统计观念，很显然，并不是人的直觉全域（repertoire of intuitions）的一部分。

在后验概率 (posterior probability) —— 从一个整体而不是另一个整体中抽取样本的概率——的判断中, 也发现了类似的漠视样本规模的现象。请思考下面的例子:

想象一个容器装满小球, 其中 $\frac{2}{3}$ 是一种颜色, $\frac{1}{3}$ 是另一种颜色。一个人从容器中拿出 5 个球, 结果发现 4 个是红的, 1 个是白的。另一个人从容器中拿出 20 个球, 结果发现 12 个是红的, 8 个是白的。谁会对“容器里的球 $\frac{2}{3}$ 是红的而 $\frac{1}{3}$ 是白的”更有信心, 而不是相反? 每个人给出的几率应该是多少?

在这个问题中, 如果假设同样的先验概率, 那么正确的后验几率对 4: 1 的样本而言是 8 对 1, 而对 12: 8 的样本而言则是 16 对 1。然而, 大多数人则认为第一个样本提供了更强的证据, 支持容器中有更多红球的假设, 因为红球的比例在第一个样本中比在第二个样本中更大。直觉判断在这里再次受样本比例的主宰, 且本质上不受样本规模的影响, 而样本规模在决定实际的后验几率中, 扮演决定性的角色 (Kahneman & Tversky, 1972b, 3)。并且, 后验几率的直觉估计, 并不比正确值极端。在这一类的问题中, 不断观察到低估证据影响的现象 (Edwards, 1968, 25; Slovic & Lichtenstein, 1971) 被称为“保守主义”。

3. 误解机会

人们期望由随机过程而产生的事件序列, 甚至是很短的序列, 也能代表这个过程的本质特征。如考虑硬币抛掷正面或反面出现的情况, 人们会认为序列正一反一正一反一反一正比正一正一正一反一反一反更可能发生, 因为后者似乎不随机; 它也比正一正一正一正一反一正更可能发生, 因为后者并没有代表硬币的公正性 (fairness)^① (Kahneman & Tversky, 1972b, 3)。因此, 人们期望一个过程的本质特征, 不仅表现在其整体序列中, 而且也表现在其局部的部分序列中。然而, 局部的代表性序列会系统地偏离

^① fairness 在这里有双关之义。一指没有磨损的好硬币; 二指好硬币在抛掷过程中正面或反面出现的机会, 应公平。——译者注

机会预期：它包含太多的变化和太少的簇（runs）。而局部代表性的信念的另一后果，就是著名的赌徒谬误（gambler's fallacy）。例如，在轮盘赌中，看到一长串红色之后，大多数人错误地相信现在应该是黑色了，似乎因为较之于再出现一个红色，黑色的出现，将会导致一个更有代表性的序列的出现。机会，通常被看成是自我校正的过程，其中一个方向的偏离，将诱发相反方向的偏离以维持平衡。事实上，正如机会过程所显示的，偏离不会被“校正”，仅仅被稀释。

对机会的误解，并不局限于普通的被试。一项由有经验的研究心理学家从事的关于统计直觉的研究（Tversky & Kahneman, 1971, 2），揭示出可称之为“小数律”（law of small numbers）的持久的信念。依照小数律的信念，即使是小样本，它也高度代表由此而产生的整体。这些研究者的回答，反映了这样的一种预期：有关整体的有效假设，无论在多大规模的样本中，都会通过统计上有显著意义的结果而被表征。其结果是，研究者过分关注小样本的结果，而全然高估这些结果的可复制性或可重复性（replicability）。在实际研究中，这种偏差导致对不充分规模的样本的选择，以及对结果的过分解释。

4. 对可预测性（predictability）不敏感

人们经常要进行数值预测，如一支股票的走势、一件商品的需求量或者一场足球赛的结果。这些预测，也常常依靠代表性。如给你一份公司的描述，要求你预测其未来的利润。如果公司描述很诱人，那么高利润似乎最有代表性；如果描述很差，那么很差的业绩将显得最有代表性。描述诱人的程度，并不受描述信度的影响，也不受预测精确度的影响。因此，如果仅仅依据精彩的描述来进行预测，其预测将对证据的信度不敏感，也对预测的预期精确性不敏感。

这种判断模式，违背了规范的统计理论，根据规范的理论，极值和预测区间，可以根据预测值推断。当可预测性为零时，应该在所有情形下进行同样的预测。例如，如果公司描述没有提供与利润有关的任何信息，那么，对公司的公司应该预测同样的数值如平均利润。当然，如果可预测性是完备的，那么，所预测的价值会匹配其实际价值，而预测的范围，也会

等同于实际结果的范围。通常而言，可预测性越高，所预测的价值的范围越大。

一些数字预测的研究已经表明靠直觉预测违背上述原则，被试很少或没有考虑可预测性（Kahneman & Tversky, 1973, 4）。在其中的一项研究中，给被试呈现几段文字，每段文字描述一个实习老师（a student teacher）在一实习课程中的表现。一些被试被要求用百分制来评价所描述的课程相对于特定总体的质量；而其他被试被要求也用百分制分数来预测每个实习老师在5年后的表现。两种条件下的评价是一样的。也就是说，对遥远尺度的预测（5年后一个老师的成就），和对当前预测所基于的信息（即实习课程的质量）的评价是一样的。而进行预测的这些被试，毫无疑问地意识到基于5年前的一次实习课程来对教学能力进行预测，其可预测性是有限的；尽管如此，其预测和其评价一样极端。

5. 效度错觉

我们已经指出，人们经常通过选择输入信息（如一个人的描述）中最有代表性的特征（如职业）来进行预测。他们在预测时所拥有的信心主要依赖于代表性的程度，亦即所选择的特征和输入信息之间匹配的质量，而很少或没有关注限制预测精确性的因素。因此，在面对与图书管理员的刻板印象相符的人格描述时，人们很有信心地将其预测为图书管理员，即使这种描述很单薄、不可信或者已过时。由预测特征和输入信息之间的良好吻合而产生的无保证的信心，可称为效度错觉。这种错觉，即使在判断者意识到存在限制其预测精确性的因素时，仍会继续。进行选择访谈（selection interview）的心理学家，常常在其预测中有足够的信心，即使他们确知有大量的文献表明选择访谈是高度不可靠的，而这已成为公开的秘密。尽管持续有证据证明选择访谈的不恰当性，但临床上对它的依赖充分地证实了这种效应的强度。

输入信息的内在一致性，是人们预测有信心主要决定因素。例如，有两个大学生，第一个学生一年级的绩点都是B，而第二个学生有许多A和C，如要预测学生在大学期间整体的平均绩点，人们对第一个学生会更有信心。当输入变量是冗余的或高度相关的时候，最易观察到高度一致的

输入信息。因此，在基于冗余的输入变量进行预测时，人们似乎信心更大。然而，相关统计学的基本结果指出，如果给定输入变量的确切效度，基于这些变量进行的预测，在它们相互独立而不是冗余或相关时，能获得更高的精确性。因此，变量间的冗余性（redundancy），虽然提高了预测信心，但它降低了预测精确性；而人们的预测信心，常常超出人的能力（Kahneman & Tversky, 1973, 4）。

6. 误解回归

假设用能力测验的两个等价版本对一大群儿童进行能力测试。如果你从中挑出10个儿童，他们在其中一套测验中表现优秀，那么你会发现这些儿童在另一套测验中的表现有些令人失望。相反，如果从中挑出的10个儿童在其中一套测验中表现很差，那么你会发现这些儿童在另一套测验中的表现就平均而言会有进步。普遍而言，设想两个变量 X 和 Y ，它们有同样的分布。如果所选择的一些个体，在 X 的平均分数上偏离 X 的均值 k 个单位，那么，它们在 Y 的平均分数上通常会反方向地偏离 Y 的均值 k 个单位。这些观察表明向均值回归的普遍现象，它首先是由高尔顿（Galton）在100多年前证明的。

在正常的生命历程中，人们会碰到许多向均值回归的例证，如父亲和儿子身高的比较、丈夫和妻子智力的比较，或个体在连续的测验中成绩间的比较。然而，人们对这种现象并没有发展出正确的直觉。首先，人们在许多肯定会发生回归的语境中，并没有预期回归的发生；其次，在他们确认回归的发生时，常常对之创造虚假的因果解释（Kahneman & Tversky, 1973, 4）。我们主张回归现象不易把握，是因为它和这种信念并不相容，而这种信念就是：所预测的结果，应该最大限度地代表输入信息。因此，结果变量的数值，也应该和输入变量的数值一样极端。

正如下面的研究（Kahneman & Tversky, 1973, 4）所证明的，没有意识到回归的意义，会产生有害的后果。在有关飞行训练的讨论中，有经验的教练发现赞扬一次平滑的着陆，通常会伴随下一次糟糕的着陆；而对糟糕着陆的严厉批评，则常常会伴随下一次着陆的改善。这个教练得出这样的结论：与广为接受的心理学定律相对立，言语奖赏对学习是有害的，而言

语惩罚则是有益的。但这个结论是没有根据的，因为存在向均值的回归。正如在重复测验的案例中那样，好的成绩经常会伴随差的表现，而差的表现却常常伴随好的成绩，即使教练对学员第一次的成绩并没有作出任何反应。因为教练在其学员良好着陆后给予赞扬而在糟糕着陆后进行批评，他们会得出错误而有害的结论，即惩罚比奖赏更有效。

因此，理解回归（regression）效应的失败，导致人们高估惩罚的有效性和低估奖赏的有效性。在社会互动和训练中，表现好时，通常会给予奖赏，而表现差时，通常会给予惩罚。因此仅仅通过回归，惩罚之后行为更可能改善，而在奖赏之后行为可能更糟。其结果是人们碰巧在惩罚他人时经常得到奖赏，而在奖赏他人时经常得到惩罚。而人们则常常意识不到这种偶然关系（contingency）。事实上，回归在决定奖赏和惩罚这一明显后果中难以辨别的作用，似乎逃脱了这个领域的学者的注意。

二、便利性

有许多这样的情景，人们通过能想到的例证或事件发生的容易性（ease）来评估这类事件发生的频率或概率。如人们通过回忆其熟人中心脏病的发生情况来评估中年人患心脏病的风险。类似地，通过想象可能遭遇的种种困难，评估一个给定的企业会失败的概率。这种判断启发式，被称为便利性。在评价频率或概率时，便利性是一条有用的线索，因为与较小可能的类别的例证相比，我们能更好和更快地接近或获得较大可能的类别的例证。然而，便利性受频率和概率之外的其他因素的影响，因此，对便利性的依赖会导致预测的偏差，其中一些，在下面予以详细说明。

1. 由例证的可提取性而引发的偏差

当通过其例证的便利性来判断类别的规模时，其例证容易提取的类别，比之频率相同但其例证更难提取的类别，数量上显得更多。在证明这个效应的一项基本研究中，让被试听到一串著名男女的名单，然后要求他们判断这一串名字中是否男人比女人多。不同的名单呈现给不同的被试组。在一些名单中，男人比女人更有名，而在其他名单中，女人比男人更有名。结果表明，在每个名单中，被试都错误地进行判断：名人更多的类别（性

别)，数目更多 (Tversky & Kahneman, 1973, 11)。

除熟悉性 (familiarity) 之外，还有其他因素如显著性 (salience)，会影响例证的可提取性 (retrievability)。例如，目睹房子着火对这类事件主观概率的影响，比之在当地报纸上读到此类事件可能要大。并且，近期的事件比早期的事件，可能会更为便利地提取。众所周知，目睹翻车事故后，主观上认为发生交通事故多的概率会暂时地升高。

2. 由搜索集的有效性而引发的偏差

12

设想你从英文文本中随机抽出一个词（它有3个字母或更多）。第1个字母是 r 的词还是第3个字母是 r 的词哪种更有可能？在处理这个问题时，人们得通过回忆以 r 打头的词（如 road）和第3个字母是 r 的词（如 car），并且通过能回忆起这两类词的容易性来评估它们的相对频度。因为用第1个字母比用第3个字母来搜索单词要容易得多，大多数人会判断以某个辅音打头的词比第3个字母是这个辅音的词要多。而实际上，第3个字母是 r 或 k 的英文单词，要多于第1个字母是 r 或 k 的英文单词 (Tversky & Kahneman, 1973, 11)。

不同的任务激发不同的搜索集 (search sets)。如果要你来评定书面英文中抽象词（如思想或爱）与具体词（如门或水）的频度，回答这个问题的自然方法，是去搜索这些词会出现的语境。想到提及抽象词如爱情故事中的爱的语境，似乎比想到提及具体词如门的语境，要容易些。如果单词的频度是通过它们所出现的语境的便利性来判断的，那么，抽象词会被认为比具体词要多。近期的一项研究 (Galbraith et al., 1973) 发现了这种偏差。这项研究证明，尽管客观频度相同，抽象词的频度，被认为比具体词要多得多。

3. 想象力的偏差

有时，人们要评估一类事件的频率，这类事件在脑海中无法找到实例，却可以依照一定的规则来推断。在这些情况中，人们通常通过相关的容易被建构的例证来形成想要评估的例证并评估其频率或概率。然而，例证建构的容易性并不总是反映实际的频率，这种评价模式易导致偏差。例如，一个10人的群体，他们要组成有 k 个成员的委员会，其中 k 在2和8之间。

可组成多少个有 k 个成员的委员会？这个问题的正确答案，是由二项系数 $\binom{10}{k}$ 给定的，当 k 为 5 时，其极大值为 252。很明显，有 k 个成员的委员会的数目，等于有 $(10 - k)$ 个成员的委员会的数目，因为任何有 k 个成员的委员会，也界定了 $(10 - k)$ 个非成员的独特群体。

如果不通过计算来回答这个问题，就要通过心理估算来建构有 k 个成员的委员会，并且通过他们能想到的简单的办法来估计其数目。成员较少如 2 个人的委员会，比成员较多如 8 个人的委员会更可行。

最简单的委员会建构的图式，就是把群体分割为不同的集合 (disjoint sets)。你会看到，很容易建构有 2 个成员的 5 个不同的委员会，但几乎不可能产生有 8 个成员的即使是 2 个不同的委员会。结果，如果频率通过想象力或者建构的便利性来估计，小委员会比大委员会似乎更多，而与正确的钟形函数相对立。实际上，要求普通被试来估计不同规模的委员会的数目时，其估计是委员会规模的单调减函数 (decreasing monotonic function) (Tversky & Kahneman, 1973, 11)，如 2 个成员的委员会的中位值为 70，而 8 个成员的中位值为 20，但两者的正确答案都是 45。

而在现实生活中，想象力在概率评估中扮演重要的角色。例如，探险的风险，是通过能想象到的但无法应对的偶然事件来评估的。如果对这些困难的描绘栩栩如生，那么这次探险将显得格外艰难。尽管很容易想象这些灾难，但这并不反映其实际可能性。相反，如果一些可能的困难很难被意识到，或没有想到，那么风险性会被完全低估。

4. 伪相关

在判断同时发生的两个事件的频率时，查普曼等 (Chapman et al., 1969) 发现了一种有意思的偏差。他们给普通被试呈现与几种假想的心理病人有关的信息。每个病人的资料，包括一份临床诊断和这个病人所画的人像图。然后，被试要估计每种诊断（如妄想狂或疑症）和人像图中不同特征（如奇怪的眼睛）的频率。被试明显地高估自然联想物（如疑症和奇怪的眼睛）之间共同发生的频率。这种效应被称为伪相关 (illusory correlation)。在对自己所接触的资料的错误判断中，普通被试“重新发现”许多普遍而没有根据的临床信条 (lore)，而这些信条，与“画人测验” (the

draw-a-person test) 的解释有关。伪相关效应, 极端地抗拒相互矛盾的资料。即使症状和诊断之间实际存在负相关时, 它还会维持; 并且, 它使判断者无法觉察所存在的真实关系。

便利性也为伪相关提供了自然的解释。对两个事件同时发生的频率判断, 基于两者之间相关联的强度。当关联很强时, 人们可能会确信两个事件常常会同时出现。因此, 强关系常常被判断为同时发生。依照这种观点, 如疑症和奇怪的眼睛之间的伪相关是因为疑症更多地与眼睛而不是身体的其他部分有关联。

人生的经验告诉我们, 通常, 发生频率高的事件比频率低的事件能回忆得更好和更快; 可能的事件比之不可能的事件, 更容易想象; 而当事件经常同时发生时, 事件之间的关联性得以增强。其结果是人们在推测类别的量值、事件的可能性和事件共同发生的频率时, 基于容易实施的心智操作如提取、建构和联想, 随意使用便利性启发式的程序。然而, 正如前面的研究所证明的, 这种有价值的推测程序, 会导致系统错误。

三、调适和初始化

在许多情景中, 人们的推测, 是以初始值为参照而获得最终答案的。初始值, 或出发点, 可以理解为确定问题的明确表达, 也可作为部分计算的结果。但在任何情形下, 调适 (adjustments) 显然都是不充分的 (Slovic et al., 1971)。也就是说, 不同的出发点产生不同的推测, 它偏向初始值。我们称之为初始化 (anchoring)。

1. 不充分的调适

在证明初始化效应的一项研究中, 被试被要求用百分制来估计种种量值, 如非洲国家在联合国中所占席位的百分值。当被试在场时, 通过转动幸运轮盘来确定一个初始数字, 它在 0 和 100 之间。让被试首先指出这个初始数字是大于还是小于所估计的量值, 然后从这个初始数字开始向上或向下移动转盘来估计量值。不同的被试群体, 被给予不同的初始数字, 而这些武断的数字对估计有显著的影响。例如, 对得到的初始数字为 10 和 65 的两个群体而言, 非洲国家在联合国中所占席位的中位估值, 依次为 25 和

45. 对于估计尽量精确的要求，也没有降低初始化效应。

15 初始化效应，不仅发生在被试被给予出发点的时候，而且发生在被试的推测基于不完备计算的结果的时候。对直觉性的数值估计的研究，证明了这种效应。两组高中生要在 5 秒钟以内估计写在黑板上的数值。一组估计的数值如下：

$$8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$$

而另一组估计的数值如下：

$$1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$$

要迅速地回答这样的问题，人们得计算几步，然后通过外推（extrapolation）或调适来推测结果。因为调适明显不充分，这种程序会导致低估。并且，就最初几步乘积（从左到右）的结果而言，第 1 个下降的数字序列高于第 2 个上升的数字序列，因此，对第 1 个序列的数值判断，应大于对第 2 个序列的数值判断。这两种预测，都得以证实。对上升序列的中位估值为 512，而对下降序列的中位估值为 2 250。但正确值为 40 320。

2. 在评估连续事件和非连续事件时的偏差

巴-希勒尔在近期的研究中（Bar-Hillel, 1973），给被试机会在两个事件中打赌。研究采用了 3 类事件：（1）简单事件，如从装有一半红色一半白色弹珠的包中拿出一个红色弹珠；（2）连续事件，如连续 7 次从装有 90% 红色 10% 白色弹珠的包中拿出一个红色弹珠；（3）非连续事件，如在连续 7 次的尝试中至少有一次从装有 10% 红色 90% 白色弹珠的包中拿出一个红色弹珠。就这个问题而言，绝大部分被试偏好压赌连续事件（其概率为 0.48）而不是简单事件（其概率为 0.50）；被试也偏好压赌简单事件而不是非连续事件（其概率为 0.52）。因此，在两种比较中，被试倾向于压赌较小可能性的事件。这种选择模式证明了一个普遍地发现。对赌博和概率判断中的机会的研究表明，人们倾向于高估连续事件的概率（Cohen et al., 1972, 24），而低估非连续事件的概率。基本事件的既定概率（即任一阶段的成功），为连续和非连续事件的概率推测提供了基本的出发点。因为从出发点的调适明显不充分，因此在两种情形下，最后的推测，也都接近基本事件的概率。请注意连续事件的整体概率（the overall possibility），低

于每个基本事件的概率；而非连续事件的整体概率，则高于每个基本事件的概率。而作为初始化的结果，连续问题中的整体概率将被高估，而非连续问题中的整体概率将被低估。

在计划 (planning) 的语境中，评估复合事件时的偏差，尤其显著。成功完成一项艰巨任务如开发新产品，明显地具有连续 (事件) 的特征：为了任务的成功，每个事件必须发生。即使每个单一事件很可能成功，但如果事件数目大时，整体的成功概率仍然很小。因此，在评估计划成功的可能性或者项目能否按期完成的可能性时，高估复合事件的概率的普遍倾向，会导致无谓的乐观主义。相反，非连续结构，通常会遭遇风险评价。复杂系统如原子反应堆或人体，如果其任一基本部分功能失调，其整体功能也会失调。即使每一部分的失败可能性很小，如果包含很多部分，其整体失败的概率仍然很大。由于初始化，人们会倾向于低估复杂系统失败的概率。因此，初始化偏差的方向，有时可从事件的结构中推演出来。连续事件的链式结构，导致高估；而非连续事件的漏斗式结构 (the funnel-like structure) 则导致低估。

3. 在评估主观概率分布时的初始化

在决策分析中，专家常常被要求表达对某种量值的信念，如某一天以概率分布的形式呈现出的道-琼斯平均指数。这种分布的建构，常常是通过专家所选择对象的不同数值，而这些数值，对应于其主观概率分布的百分位 (percentiles)。例如，判断者也许被要求选择一个数字如 X_{90} ，这意味着他认为这个数字高于道-琼斯平均指数的主观概率是 0.90。也就是说，他选择 X_{90} ，所以他愿意接受 9 对 1 的几率，认为道-琼斯平均指数不能超过这个数值。根据几个对应于不同百分位数的这种判断，我们能够构造出道-琼斯平均指数的主观概率分布。

通过收集许多不同量值的主观概率分布，就有可能检验判断者的度量或校准 (calibration) 是否合适。在一组问题中，待估量值 (assessed quantity) 的真实值，恰好有 $\Pi\%$ 落在他给出的 X_{Π} 值之下，那么该判断者被正确 (或者外部) 校准。例如，1% 的量值，真实值应该落在 X_{01} 之下，并且 1% 的量值，真实值在 X_{99} 之上。因而，有 98% 的问题，真实值应该落在置

信区间 X_{01} 到 X_{99} 之间。

几位研究人员 (Alpert & Raiffa, 1969, 21; Staël von Holstein, 1971b; Winkler, 1967) 根据大量的判断者已经得到许多量值的概率分布。这些分布表明, 它们与合适的校准之间存在巨大而系统性的偏差。在大多数研究中, 对于大约 30% 的问题, 待估量值的实际值, 或者小于 X_{01} , 或者大于 X_{99} 。也就是说, 被试给出的置信区间过小, 与他们关于待估量值的知识所能证明的相比, 置信区间反映出更大的确定性。经验较少的被试和老练的被试, 同样存在这种偏差, 并且合适的记分规则的引入, 也不能够消除这种偏差, 而记分规则为外在校准提供了刺激 (incentive)。这种效应至少部分可归因于初始化。

例如, 为了选择 X_{90} 为道-琼斯平均指数的数值, 通常一开始就考虑它对道-琼斯的最佳估计值, 然后向上调节这个数值。如果这种调节——和大多数其他调节一样——不充分, 那么 X_{90} 就不会足够大。一种类似的初始化效应, 出现在 X_{10} 的选择上, 据推测向下调节它的最佳估计值能够得到 X_{10} 。结果, 置信区间 X_{10} 到 X_{90} 将会太小, 而待估量值的分布, 将会过于高耸 (tight)。为了支持这种解释, 有一种程序能够系统地改变主观分布, 在这种程序中, 最佳估计值并不作为初始值。

有两种不同的方法, 能够得到给定量值 (如道-琼斯平均指数) 的主观概率分布: (i) 要求被试选择与他的主观概率分布的指定百分位数相对应的道-琼斯数值, (ii) 要求被试估计真实值将会超过指定数值的概率。这两种程序在形式上是等同的, 因此应该产生相同的分布。然而, 它们暗示的是对于不同初始值的不同调节模式。在程序 (i) 中, 通常的起始点, 是对变量的最佳估计值。在程序 (ii) 中, 被试的注意力, 可能一开始就被固定在问题中给出的数值上; 或者, 被试可能一开始就被固定在半分比率 (even odds) 或者 50-50 概率上, 它是可能性估计中通常的起始点。在任何一种情况下, 程序 (ii) 得到的几率与程序 (i) 相比, 都更不极端。

为了对比这两种程序, 向一组被试出示一组 24 个量值 (例如新德里到北京的空中距离), 每一个问题让被试估计 X_{10} 或者 X_{90} 。另外一组被试, 得到的是第一组关于 24 个量值中每一个量值判断的中位数。他们被要求估计

每一个给定数值超过相应量值的真实值的几率。在没有任何偏差的情况下，第二组应该检索出指定给第一组的几率，也就是 9:1。然而，如果半分比率或者给出的数值充当初始值，那么第二组的几率应该更不极端，也就是说，接近于 1:1。实际上，所有问题，第二组给出的几率中位数都是 3:1。当测试两个组的判断确定外部校准时发现，与早期的研究一致，第一组中的被试非常极端。实际上，他们规定的概率为 0.10 的可能性事件在 24% 的个案中实现了。与此相对，第二组的被试过于保守，他们指定的平均概率为 0.34 的可能性事件，实际上在 26% 的个案中实现了。这些结果，说明了校准的程度依赖于启发程序 (the procedure of elicitation) 的方式。

四、讨论

本章关注认知偏差，而这些偏差源于对判断启发式的依赖。不能把这些偏差归结为动机性的影响，如有意的思考或者由收益和惩罚而引发的判断的歪曲。事实上，尽管鼓励被试要尽量准确，并且其正确回答会有奖赏，前面所报告的严重的判断错误，仍然存在 (Kahneman & Tversky, 1972b, 3; Tversky & Kahneman, 1973, 11)。

对启发式的依赖和偏差的普遍存在，并不仅仅局限于普通人。有经验的研究者，也容易在其直觉思考时犯下同样的错误。例如，在统计学上有良好训练的个体，在其直觉判断时，也存在这种倾向：在预测最能代表资料的结果时，没有充分地关注先验概率 (Kahneman & Tversky, 1973, 4; Tversky & Kahneman, 1971, 2)。虽然有丰富的统计知识的个体，能避免基本错误如赌徒谬误，但在更为复杂、透明度更小的问题上，他们的直觉判断，也会犯同样的谬误。

毫不奇怪，有用的启发式如代表性和便利性，可以保持，即使它们在预测和估计时会偶尔导致错误。而令人惊异的是，人们不能从其人生经验中推论出如此基本的统计规则，如向均值回归，或者样本规模对抽样变异的影响。虽然在正常的生命历程中，每个人都碰到无数的例子，从中应该能归纳出这些抽样和回归原则，但很少有人能独自发现这些原则。例如，人们不能发现英文课文中相连行之间的字词的平均字母数，比之相连页之

间的字词的平均字母数，变异更大，仅仅因为他们没有注意单行或单页的字词的平均字母数。因此，人们并没有知悉样本规模和抽样变异性之间的关系，尽管为这种学习而提供的资料是丰富的。

缺乏合适的编码可以解释人们为什么常常不能在其概率判断中觉察到偏差。人们应该确切了解其判断是不是外在度量的，通过在要分配同样概率的事件中保持实际发生的事件的比例。然而，通过他们所判断的概率来分类组合事件，并不合适。在缺乏这种分类组合的时候，个体不可能发现他所分配的概率为 0.9 或更高的预测，一半会真的发生。

认知偏差的经验分析，对所判断的概率的理论和应用角色，有特定的内涵。现代决策理论（de Finetti, 1968; Savage, 1954）把主观概率看成是理想化个人的量化观点。具体说，给定事件的主观概率，被界定为特定个人愿意接受的有关这个事件的一组赌注（bets）。如果他在不同赌注之间的选择满足某种原则如概率论的公理，那么，内在一致而连贯的主观概率测量，就得以引申出来。所引申的概率是主观的，是因为不同个体对同一事件允许有不同的概率。这种研究路径的重要贡献在于，它为概率提供了有力的主观解释，而这种解释可应用于独特事件，并且嵌入在理性决策的一般理论中。

也许应该注意，主观概率有时可从赌注之间的偏好中推演出来，那它们就和上述的形成方式不同。一个人在 A 队上而不是 B 队上加赌注，是因为他相信 A 队更可能会赢。他并没有从他的打赌偏好中推论这种信念。因此，事实上正如理性决策的公理论（axiomatic theory）所言（Savage, 1954），主观概率决定赌注之间的偏好，而不是从中引申出来。

概率的内在一致的主观本质，导致许多学者形成这样的信念：连贯性或内在一致性，应该是评价所判断的概率唯一有效的标准。从主观概率的这种形式理论的观点来看，任何内在一致的概率判断集，都优于其他。这种标准并不完全令人满意，因为内在一致的主观概率集，可能与个体所持有的其他信念并不相容。请考虑这样的一个人，他有关硬币投掷游戏所有可能结果的主观概率反映了赌徒谬误。或者说，随着硬币正面连续地出现，在后面的投掷中，他所估计的反面出现的概率会增大。而依照这种形式理

论的标准，其判断就是内在一致的，因此也是合适的主观概率。但这些概率，与普遍持有的信念并不相容；而这种信念就是，硬币并没有记忆，因此也不可能生发序列依存性（sequential dependencies）。对所判断的概率而言，充分、合理或内在一致，都不完备。判断必须与个体所持有的整体信念网相容。不幸的是，并没有简单的形式程序，来评价概率判断集和判断者的整体信念系统的相容性（compatibility）。然而，合理的判断要追求相容性，即使内在一致性更易达到、更易评估。尤其是，他会尽力使其概率判断和其有关主题的知识、概率规律以及自身的判断启发式和偏差相容。

五、概要总结

本章描述了 3 种启发式，它们被用于在不确定状况下进行判断：（1）代表性，它常常被用于这样的情景，即人们被要求判断一个物品或一个事件 A 隶属于类别或过程 B 的概率的时候。（2）例证或场景的便利性，它常常被用于这样的情景，即人们被要求评估一个类别的频率或特定发展的拟真性（the plausibility of a particular development）的时候。（3）从出发点开始的调适，当有相关数字可资利用的时候，它常常被用于数值预测。这些启发式，是高度经济的，也常常是有效的，但它们也会导致系统的和可预测的错误。更好地理解这些启发式和它们所导致的偏差，会改善不确定状况下的判断和决策。

本章原载于《科学》（*Science*, 1974, 185, 1124-1131）。允许重印。



第二 部分
PART TWO

代表性



第 2 章

小数律的信念

特沃斯基和卡尼曼 (Amos Tversky and Daniel Kahneman)

23 “假设你用 20 位被试做实验，并已获得显著结果来证实你的理论 ($z = 2.23$, $p < 0.05$, 双侧检验)。现在你有理由要对另外的 10 人组被试再做实验，你认为通过单侧检验其结果显著的概率是多少?”

如果你认为概率在 0.85 左右，你将高兴地得知自己属于多数人群体。事实上，这是在数学心理学组 (the Mathematical Psychology Group) 和美国心理学会 (the American Psychological Association) 的会议上，两个小组对我们所发问卷做出慷慨回答的中位值。

如果你认为概率在 0.48 左右，你则属于少数人群体：84 位被试中仅有 9 人给出的答案在 0.40 与 0.60 之间。然而，0.48 恰恰是一个比 0.85 更为合理的估计。^[1]

显然，大多数心理学家对于成功复制已获发现的可能性，怀有夸大的信念。而这种信念的来源及其对
24 科研行为的影响，正是本章想探讨的。我们的论点

是，人们对随机抽样有着强烈的直觉；这些直觉本质上是错误的；而普通被试及训练有素的科学家，也都有这种直觉；当它们被应用于科学探寻的过程中时，会产生不适当后果。

我们认为，人们相信一个从总体中随机抽取的样本具有高度代表性，或者说，这个样本在所有本质特征上都与总体相似。相应的结果是，他们预期从特定总体中抽出的任意两个样本之间以及它们与总体之间，都有着比抽样理论所预测的更多的相似性，至少对小样本而言是这样的。

这种把样本作为整体代表的倾向，明显表现在种种情景中。比如，当要求被试给出投掷一完整的硬币（fair coin）的随机序列时，他们给出的序列，在任意部分，头像或正面的比例都远远比所预测的几率，都要接近0.50（Tune, 1964）。因此，反应序列中的每一部分，都高度代表着硬币的“公正性”。类似的效应，在诸如概率学习实验（probability learning experiments）（Estes, 1964）或者其他机会的序列博弈（sequential games of chance）中，当被试在随机产生的序列中连续预测事件的情况时，均能发现。被试的行为，似乎依照随机序列中的每一部分都应反映真实比例的观念。如果序列偏离了总体比例，一种在反方向上的校正性偏差（corrective bias）就会预期出现。这被称为赌徒谬误（gambler's fallacy）。

赌徒谬误的实质，是对机会规律之公正性的误解。赌徒认为，硬币的公正性有理由让他预期任意方向上的偏差，将被另一方向上对应的偏差所抵消。然而，即使是最公正的硬币，考虑到它的记忆和道德感的限度，也并不总是如赌徒所预期的那样公正。这一谬误并不为赌徒所独有。请思考下面这个例子：

已知某一城市八年级学生总体的平均智商是100。你已经选出50人的随机样本来做教育成就的研究。第一个被试儿童的智商是150，你认为整个样本的平均智商将是多少？

正确答案是101。但有相当多的人相信这一样本的平均智商仍是100。这一预期，只有在认为随机过程是自我校正（self-correcting）的情况下，

才有合理性。通常所说的“错误会相互抵消”，正反映了这种积极的自我校正过程的意象。在大自然中，我们熟悉的一些过程，就符合此定律：与均衡的偏离，会产生重新回复到均衡的动力。但与之相反，机会规律并不按此行事：偏差并不会随抽样过程而消失，它们仅仅会被稀释。

25 至此，我们已尝试描述有关机会的两种相互关联的直觉。我们已经提出一种代表性假设（representation hypothesis）。人们据此认为，从同一总体中抽出的样本相互间十分相似，且与总体相似。我们也主张，人们相信抽样是一个自我校正的过程。这两种信念导致同样的结果。它们都引发对样本特征的预期，且这种预期的变异性要小于真实的变异性，至少对小样本而言是这样的。

大数定律（law of large numbers），保证非常大的样本确实能高度代表它从中抽出的总体。另外，如果自我校正的过程在起作用，那么，小样本也具有高度的代表性并且彼此相似。人们对随机过程的直觉，似乎正满足小数律。小数律声称，大数定律同样适用于小样本。

试考虑一位假想的科学家，他信守小数律，他的信念将怎样影响他的科学工作？假设我们的科学家在研究这种现象，它与非控变异性有关的量度很小，也就是说，他从自然中所接收的信息的信-噪比很小。我们的科学家，可能是一位气象学家、一位药理学家或者是一位心理学家。

如果他信守小数律，他就会对在小样本基础上得出结论的有效性抱有夸大的自信。为表明这一点，我们试想他正研究婴儿对于两个玩具的偏好选择。在所研究的前五个婴儿中，四个显示出对同一玩具的偏爱。许多心理学家此时已经有一定把握认为没有偏好的虚无假设或零假设是错误的。幸运的是，这样的裁决，并不足以用来作论文发表，尽管或许能用于书中。通过快速运算，我们的心理学家将发现虚无假设成立的极值将高达 $3/8$ 。

显然，将统计假设检验运用于科学推论中，面临重重困难。但是，显著性水平〔或如贝叶斯统计所偏好的似然比率（likelihood ratio）〕的计算，迫使科学家依据有效的抽样方差估计而不是他主观的偏差估计，来评价所得结果。因此，统计检验，通过规诫那些宁愿信守小数律的科学共同体的许多成员，而使科学界免遭过于仓促地拒绝虚无假设（即第一类错误）的

厄运。但另一方面，我们也没有对应的保障，来消除无法证实有效研究假设的风险（即第二类错误）。

设想一位心理学家要研究成就需求（need for achievements）与成绩的相关性。在确定样本规模时，他可能这样推理：“我期望的相关是多少？ $r=0.35$ 。我需要多大的样本规模来使我的结果显著？（查一下表格） $N=33$ 。不错，这就是我的样本。”这一推理过程中的唯一缺陷，在于我们的心理学家忘记了抽样方差，也许他认为任何样本都应该高度代表整体。然而，如果他对于总体中的相关的猜测是正确的话，样本中的相关系数，就有可能低于或高于0.35。因此，当 $N=33$ 时，获得显著结果的可能性（即检验力）大约为0.50。

在一个对统计检验力的详细研究中，科恩（J. Cohen, 1962; 1969）为我们提供了大效应、中效应和小效应（effects）的合理定义，以及一套广泛的用来估计各种统计检验力的计算辅助方法。比如，在对两个均值差异进行正态分布检验时，按照给出的定义，0.25Q的差异是小效应，0.50Q是中等效应，而1Q是大效应；普通员工与半技术化员工的智商差异，是中等效应。在一个精妙的研究中，科恩（1962）研究了刊登在一卷《变态和社会心理学杂志》（*Journal of Abnormal and Social Psychology*）上的所有统计分析，并计算出发现上述三种效应的可能性。他发现小效应的平均检验力是0.18，中等效应是0.48，大效应是0.83。如果心理学家普遍地预期中等效应并依此来选定样本规模，他们所做研究的检验力，实际仅为0.50左右。

科恩的分析表明，许多心理学研究的统计检验力出奇的低。这是一个自拆墙脚的举动：它正是冲着那些沮丧的科学家和无效的研究来的。一位正检验一个有效的假设却不能获得显著结果的研究者，不禁怨叹大自然是如此地不值得信赖，甚至充满敌意。进一步，正如欧沃劳（Overall, 1969）所揭示的，欠缺统计检验力的研究如此普遍，它不仅虚耗人力，而且十分有害：在出版的结果中，它常常导致对虚无假设的无效拒绝。

考虑到在重复研究（replication studies）设计中，统计检验力有特别重要的意义，我们在问卷中考察了心理学家对重复研究的态度：

假设你的一位博士生刚完成 40 个动物的实验研究，困难且耗时。他计量和分析了大量变量。他的结果是非结论性的，但是实验的前后比较（before-after comparison）是高度显著性的， $t=2.70$ 。这个结果令人惊异，并且在理论上可能有重大意义。

综合考虑到这项结果的重要性、它的惊人价值以及你的学生所做分析的动物数量，你会建议他在发表前重做这项实验吗？如果是，你认为他的重复实验还需要多少动物？

27 在被调查的心理学家中，压倒性的情感是偏好重做实验：在 75 位被试中，66 位建议重做，也许因为他们怀疑单一的显著结果是出于偶然。他们推荐重复实验所需动物数目的中位数是 20。思考这项建议的可能性结果，对我们很有裨益。如果在第二个样本中，均值与方差与第一个样本中的恰恰一样，那么， t 值则为 1.88。根据注释 [1] 中的推理，这位学生在重复实验中取得显著结果的概率仅略高于 0.50 ($p=0.05$ ，单侧检验)。因为预料到 20 个动物作为重复实验的样本对我们的被试来说是合理的，我们加上了下面这个问题：

假定你那位不高兴的研究生事实上已用 20 个动物重复了原初的实验，并且在同一方向上获得了不显著的结果， $t=1.24$ 。现在你会有什么建议？请选择一个：（括号中的数字，是选择该答案的被试人数。）

- (a) 他应该汇集他的实验结果并客观地出版其结论。(0)
- (b) 他应该将结果作为有待验证的发现予以报告。(26)
- (c) 他应该对另一组动物再做实验（中位数值为 20）。(21)
- (d) 他应该对两组结果的差异尽力做出解释。(30)

请注意，不管对原初发现的信心如何，研究的可信度，通过重复实验显然得以加强。在两个样本的实验中，不仅实验结果是同方向的，而且重复实验的结果的量度，也完全是原结果的 $2/3$ 。就被试所建议的样本规模 (20) 而言，重复实验正如所预期的，基本上是成功的。但是，回答的分

布，却反映出在重复实验之后对其结果的持续怀疑。这一令人不快的情形，正是不充分的统计检验力的典型结果。

反应 b 和 c ，在某些方面有其合理性。与之对照，最为普遍的反应 d ，则是站不住脚的。我们怀疑，即使被试知道两项结果的差异甚至没有接近显著性（如果两样本方差相等， $t = 0.53$ ），他们的答案也会是一样的。在缺乏统计检验的情况下，我们的被试追随代表性假设：当两样本差异超出他们的预期时，他们认为这值得解释。然而，这种“对两样本差异寻找解释”的尝试，是无谓的徒劳。

总体而言，我们的被试对重复实验的评估是相当苛刻的。这符合代表性假设：如果我们认为所有样本相互间都非常相似，那么几乎所有针对有效假设的重复实验在统计上都应该是显著的。这一判断重复实验成功与否的标准之苛刻，在对下面这个问题的回答中显现无遗：

一位研究者报道了一个你认为不合理的结果。他对 15 位被试进行研究，且有显著结果， $t = 2.46$ 。另一位研究者重做了他的实验，然而对同样数目的被试进行研究得出的结果是不显著的。但两个数据的方向一样。

你正在检阅该文献，当第二组数据中的 t 的最高值为多少时，你会认为重复实验失败？

大部分被试认为 $t = 1.70$ 时，重复实验是失败的。如果将这两次研究中的数据（ $t = 2.46$ 和 $t = 1.70$ ）汇集，则合并后的数据的 t 值大约为 3.00（假定方差相等）。因此，我们面临一个自相矛盾的情况，即同样的数据，如果作为原来研究的一部分，会增加我们对结果的信心，但作为独立研究，则会动摇我们的信心。这一双重标准十分令人不安，因为重复实验，出于多种原因的考虑，通常被认为是独立研究，而假设的有效性，则是通过列举证实和否证的报告来衡量的。

与普遍的信念相反，做重复实验时，样本所需数目通常应大于原来样本。实施重复研究的决定，常表现出对于这一发现的极大喜爱和希望看到

它被怀疑者共同体所接受的渴望。因为怀疑者共同体无端地要求重复实验应是独立显著的，或至少接近显著性，实验者因此必须用更大的样本进行研究。为说明这一点，我们重提前面所提的那位不幸的研究生。假定他认定他原初的结果（ $t = 2.70$, $N = 40$ ）是有效的，而且仅愿意接受 0.10 的获得低于 1.70 的 t 值的风险，则他必须在重复实验中对差不多 50 个动物进行研究。如果原初结果更弱一点的话（ $t = 2.20$, $N = 40$ ），要获得同样的检验力，重复实验的样本规模将增加到 75。

迄今为止所讨论的这种效应，并不局限于与均值有关的假设。而与方差有关的假设的类似效应，在对下面的问题的回答中得以体现：

你在做一项相关性的研究，对 100 位被试进行 20 个变量的评分。190 个相关系数中，有 27 个在 0.05 的显著性水平上有显著性；9 个在超过 0.01 的显著性水平上有显著性。显著相关系数的绝对平均值是 0.31，而且结果模式（pattern）在理论上也是非常合理的。你认为在对 40 个被试严格进行的重复实验中，这 27 个显著相关系数将有多少个还是显著的？

当 $N = 40$ 时，在 0.05 的显著水平上，显著性所需相关系数的大小大约为 0.31，这是原来研究中显著相关系数的平均值。因此，当 $N = 40$ 时，原来显著的相关系数中，仅有一半（也就是 13 或者 14 个）仍然显著。当然，重复实验中的相关系数，必然不同于原初研究中的相关系数。而由于回归效应（regression effects），原初有显著性的相关系数，极有可能减少。因此，原初 27 个显著相关系数中，大约 8 个至 10 个将仍然显著，这理应是一个比较乐观的估计。但我们被试的平均估计值是 18。这比从原来的 100 位被试中随机抽取 40 位的情况下重新计算出来的相关系数要大。显然，在重复实验样本中，人们预期的不仅仅是复制原来的统计推论；他们预期的是复制原来的结果显著性，而丝毫不考虑样本规模的变化。这样的预期，要求对代表性假设进行可笑的拓展；因为即使是小数律，也无法产生这样一个结果。

结果模式几乎能被完全复制的期望,为那些普遍的却十分可悲的研究实践提供了准则。一位计算三种焦虑指数(indexes of anxiety)及三种依赖性指数之间的所有相关的研究者,常常会以极大的信心来报告和解释所获得的单个显著相关。而他对其虚弱发现的信心,则来源于他的信念,即所获得的相关矩阵具有高度代表性并且能够被复制。

综上所述,我们已发现小数律的信徒是这样从事科学研究的:

(1) 在检验研究假设时,他把赌注放在小样本上,而未意识到他的失败机会非常之高。他高估了检验力。

(2) 他对于初期的趋势(如最早的几个被试的数据)以及观察到的模式的稳定性(如显著结果的数量和属性),有过分的自信。他高估了结果的显著性。

(3) 在评价自身或别人的重复实验的时候,他对显著结果的可重复性,抱有非分的高预期。他低估了置信区间的范围。

(4) 他很少将实际结果与预期间的偏离归结为样本的变异性,因为对于任意的偏差,他都能发现因果“解释”。因此,在实践中,他鲜有机会认识到抽样的变异性。他对于小数律的信念,因而始终坚不可摧。

我们的问卷调查,引发相当的证据,证实小数律的信念的广泛存在。^[2]我们多数的被试,都是小数律的信徒,无论他隶属于哪个群体。实际上,数学心理学会议上的听众和美国心理学会分会上的与会者,给出答案的中位数并无区别,尽管我们并不宣称这两个样本具有代表性。显而易见,对形式逻辑以及概率理论的了解,并不能消除错误的直觉。那我们能做些什么?对小数律的信念能被消除或者至少被控制吗?

从事研究的经验,似乎帮助不大,因为抽样差异太容易被解释。能与之校正的经验仅在于:既没有动机也没有机会进行虚假解释。因此,统计课程上的学生,应从总体中重复抽取固定规模的样本,并通过个人观察来学习到样本规模对抽样差异性的影响。但我们还远不能确定,人们的预期可以通过这种方式来纠正。因为相关的偏差,如赌徒谬误,在面对足够多的矛盾证据时,依然存在。

即使这种偏差不能被消除,学生仍然能学会意识到它的存在,并采取

必要的防备。因为统计的教学，从来都不乏于规诫，有关统计直觉偏差的警告，也不会显得不合时宜。最明显的预防就是计算。小数律的信徒对显著性水平、检验力和置信区间，都有不正确的直觉。人们通常会计算和报告显著性水平，却没有计算和报告检验力和置信限度（confidence limits）。也许应该计算和报告它们。

与有些合理假设相关的检验力的详细计算，如科恩（1962，1969）的小、大、中等效应，无疑是必须在任何研究之前进行的。此类计算，常常能让人意识到没有理由去进行那项研究，除非，比如说，样本规模再乘以4。我们拒绝相信一位严谨的研究者，会有意接受存在0.50风险可能被否证的有效假设。另外，检验力的计算，对于否定性结果——也就是不能拒绝虚无假设的结果——的解释，也是基本的。因为，读者对检验力的直觉性估计，极有可能是错误的，所以出版计算出来的结果，既不是浪费读者的时间，也不是浪费杂志的版面。

在早期的心理学文献的报告中，有一种普遍的做法，比如说，将样本均值表达为 $M \pm PE$ ，这里 PE 是可能的误差（即以均值为中心的置信度为50%的置信区间）。这一做法，后来由于人们对假设检验程式的青睐而被抛弃。然而，置信区间，能为抽样差异性提供有用的指标，并且正是由于这种变异性，我们倾向于低估。而对显著性水平的强调，倾向于遮蔽效应值（size of an effect）和它的统计显著性之间的基本区别。如果不考虑样本规模，在一项研究中，效应值也是重复实验中的效应值的合理估值。相反，重复实验中所估计的显著性水平，则严格依赖于样本规模。如果效应值与其显著性水平之间的区别，能被辨明，并且效应值的计算得以常规性地披露，那么，有关显著性水平的可重复性的不真切预期，就有可能得以校正。至少，就此来说，接受假设—检验模型，对心理学而言并不是一个纯粹的福音。

小数律的虔诚信徒，真诚地犯下了违背统计推论逻辑的重重罪过。代表性假设，描述了一种不受制于动机因素的认知或感知偏差。因此，如果说匆忙拒绝虚无假设是痛快淋漓的，那么，拒绝珍贵的假设则是令人扼腕的，但小数律的虔诚信徒，则屈从于这两种境况。他的直觉预期，受制于

对这个世界的一贯的误解，而并非一时的异想天开。如果给他一些警告和刺激，他或许会愿意恰当地怀疑其统计直觉，并且在任何可能的情况下，以计算来取代那些印象形成（impression formation）。

本章原载《心理学通报》（*Psychological Bulletin*, 1971, 2: 105-110）。
允许重印。

注释

[1] 所要求的估计，可以用多种方式来解释。一种可能的路径，就是遵循常规的研究实践，即用从一个研究中获得的值来定义虚无假设之外的替换假设（alternative hypothesis）。问题中所要求的概率，因此可理解为针对由第一个样本所获结果所定义的替换假设的检验力（即在第二个样本中获得显著结果的概率）。在用已知方差进行均值检验的特定情形下，可以计算出总体均值等于第一个样本均值这一假设的检验力。因为第二个样本的规模是第一个样本的一半，计算出的使 z 大于或等于1.645的概率，只有0.473。而理论上更合理的路径，是在贝叶斯框架中来解释所要求的概率，并计算出与合理选择的先验分布（prior distribution）有关的概率。假设同一的先验概率，所需的后验概率是0.478。显然，如按通常情况，先验分布支持虚无假设，则后验概率会更小。

[2] W. Edwards (1968, 25) 主张人们不能从概率数据中提取充分的信息或者是确定性，他称此失败为保守主义。而我们的被试，却根本无法用保守主义来描述。相反，根据代表性假设，他们倾向于去获取比数据实际包含的更多的确定性。



第 3 章

主观概率：代表性判断

卡尼曼和特沃斯基 (Daniel Kahneman and Ames Tversky)

32 主观概率在我们生活中扮演着重要的角色。我们做出的决定、得出的结论及给出的解释，通常都是基于对不确定事件的可能性判断，如新工作的成功、选举结果或市场行情。事实上，在概率学习、直觉统计 (intuitive statistics) 和风险决策的语境中，已有大量的实验文献致力于研究人们如何感知、加工和评估不确定事件的概率问题。虽然从这些文献中并未产生有关不确定心理学的系统理论，但一些经验性的概括已经确立。从大量的研究中所获得的最普遍的结论也许就是，人们并不按照概率理论的原则来判断不确定事件发生的可能性。这一结论一点也不奇怪，因为许多机会定律 (law of chances) 既不在直觉上显而易见，也难以运用。然而，难为人知的一个事实是，主观概率与客观概率 (objective probability) 的偏离，似乎是稳固而系统性的，且难以消除。^[1] 显然，人们用那些偶尔合理但大多数时候无效的启发式，代替了机会

定律。

在这篇论文里，我们将详细探讨这种启发式——代表性启发式。代表性启发式的使用者，根据下面两点来评估不确定事件或者样本发生的概率：
(1) 它与总体在基本特征上的相似性程度；(2) 它反映由此生发的那个过程的显著特征的程度。我们的论点是，在很多情景中，只要事件 A 似乎比事件 B 更具代表性，则事件 A 比事件 B 就被认为更可能发生。换言之，主观概率的事件序列，与其代表性序列相一致。

与知觉相似性 (perceptual similarity) 一样，代表性易于评估而难以特征化。这两个词语都没有公认的定义，但在很多情景中，人们都同意两个刺激中某一个与标准刺激更相似，或者两事件中某一个更能代表给定的过程。在这篇论文中，我们并不准备去度量代表性，尽管这是一个可行的办法。而我们将要考虑的是，什么时候依据代表性的事件序列是明显的，并说明人们是否一致性地认为越具代表性的事件越可能发生。虽然代表性可能在如政治预测、临床诊断等诸多概率判断中都扮演重要角色，但本章仅限于本质上可重复的情景，在其中，可方便地计算客观概率。

本章所报告的大部分数据，是通过问卷而收集的，被试为总共大约 1 500 名以色列的 10、11、12 年级的大学预科高中生 (15~18 岁)。我们采用了特别的措施来维持被试 (Ss) 的注意力和动机。问卷是以小测验的形式在教室里进行的，被试的姓名记在答题纸上。每位被试，在最多 2 分钟的时间内，回答少量 (通常是 2~4 个) 所要求的问题。这些问题被介绍为有关人们机会直觉的研究。实验前，研究者用标准的口头指导语来对问题进行详细解释。实验设计，通过处理以防止与年龄、学校因素有关的含混变量。大多数问题，都预先对大学本科生 (20~25 岁) 进行了测试，且两总体所得结果没有显著差别。

一、代表性的决定因素

在这一节，我们将讨论使之具有代表性的样本或事件的特征，并说明这些特征对主观概率的影响。首先，我们将描述决定样本与其总体相似的一些特性。随后，我们将转而分析表面随机性 (apparent randomness) 的决

定因素。

34

1. 样本与总体的相似性

代表性的观念，最好用实例来详细说明。请思考下面这个问题：

研究者对某市有 6 个孩子的所有家庭进行了调查。在 72 个家庭中，孩子确切的出生顺序是：女男女男男女。请你估计在所有被调查的家庭中有多少家庭的出生顺序是：男男女男男男。

这两种出生顺序有大致等同的可能性，但大多数人会肯定地同意它们不具有相同的代表性。有 5 个男孩和 1 个女孩的出生顺序，没有反映出总体中男孩与女孩的比例。事实上，在 92 位被试中，有 75 位认为这一出生顺序比标准顺序出现的可能性要小 [$p < 0.01$ ，符号检验 (sign test)]。估计的中位数是 30。科恩等 (1956) 也报告了相似的结果。

有人也许会问，被试是否并不是简单地忽视了出生顺序的信息，而只是通过估计有 5 个男孩和 1 个女孩的家庭与有 3 个男孩和 3 个女孩的家庭的频率来回答这个问题。但是，当我们要求同样的被试估计“男男男女女女”这一出生顺序的频率时，他们认为它比“女男男女男女”出现的可能性明显要低 ($p < 0.01$)，这似乎是因为前者显得不随机。由此看来，出生顺序的信息，并没有被简单地忽视。

代表性的一个有关的决定因素在于，样本是否保持了总体中的多数—少数的对比关系。我们预期保持这一对比关系的样本，将被认为比违背了这一关系的样本更可能发生，尽管两者（客观）概率等同。这一效应，在下面的问题中得以体现：

在一所高中有两个项目，男生在 A 项目中占多数 (65%)，而在 B 项目中则占少数 (45%)。每个项目都有相等的班级数。

你随机进入一个班级，发现 55% 的学生是男生。你认为该班级属于 A 项目还是 B 项目？

因为这个班上大部分同学是男生，因此该班更能代表 A 项目而不是 B 项目。其结果，在 89 位被试中，有 67 位猜测这一班级属于 A 项目 ($p < 0.01$ ，符号检验)。事实上，该班级属于 B 项目的可能性，要略高于属于 A 项目的可能性（因为， $p = 0.45$ 的方差大于 $p = 0.65$ 的方差）。

通常来说，一个包含各种可能结果的样本，会比一个不包含某些可能结果的样本更具有代表性。例如，给定一个 $p = 4/5$ 的二项过程 (binomial process)，绝大多数被试认为 10 次成功 0 次失败的样本比 6 次成功和 4 次失败的样本可能性更小，但事实上前一样本更可能出现。

代表性的这种偏差效应 (biasing effects)，并不局限于普通被试。它们同样出现于审慎的心理学家的直觉判断中 (Tversky & Kahneman, 1971, 2)。统计显著性，通常被认为是科学真理的表征。因此，(总体中的) 真实效应被认为理应 (在样本中) 以显著结果出现，而不论样本容量如何。结果，当研究者坚信零假设是错误的时候，他们就倾向于高估显著结果的可能性。

例如，给参加数学心理学会议和美国心理学会会议的与会者呈现下面这个问题：

假设你用 20 位被试做了一项实验，并已获得显著结果，它能证实你的理论 ($z = 2.23$, $p < 0.05$ ，双侧检验)。现在你有理由还要对另一 10 人组的被试进行实验。就这组被试而言，你认为单侧检验结果显著的概率是多少？

所求概率的现实估计，略低于 0.50。而被试给出估计的中位数高达 0.85。对重复研究的显著性的过分自信，对研究实践有严重影响：它导致对显著性的不现实的预期，并产生缺乏统计检验力 (statistical power) 的研究计划。请参见科恩 (1962)。

2. 随机性的反映

不确定事件与总体的相似性，并不足以使它具有代表性。这一事件还需反映产生该事件的不确定过程的特质，也就是说，它应该是随机的。与

样本和总体的相似性一样，决定表面随机性的特定特征，随情境而变。不过，有两个普遍特征，不规则性和局部代表性（local representativeness），看来抓住了随机性的直觉内涵。下面我们依次探讨这些特征。

表面随机性的主要特征是系统模式的缺乏。比如说，硬币投掷^[2]的结果序列，若存在显然的规律就不具代表性。因此，正面与反面交替出现的序列，如“正反正反正反正反”或“反反正正反反正正”，就没有体现过程的随机性。事实上，被试认为此类序列更不太可能发生，而他们在构造模拟的随机序列（simulated random sequences）时，也会放弃它们（Tune, 1964; Wagenaar, 1970）。

正如下面这个问题所展现的，有些不规则性不仅被预期发生在结果的顺序中，也发生于结果的分配上：

36

在每轮游戏中，要将 20 个弹珠随机分给 5 位儿童：艾伦、本、卡尔、丹和阿德。请思考下面两个结果：

	(1)	(2)
艾伦	4	4
本	4	4
卡尔	5	4
丹	4	4
阿德	3	4

在多轮游戏的结果中，第一种结果与第二种结果哪种会更多呢？

第二种结果中的平均分配，客观上比第一种非平均分配更有可能，但它如此规则以至于不像是随机过程的结果。而第一种分配与平均分摊有点偏离，但更能代表随机分配。明显多数的被试（52 人中有 36 个， $p < 0.01$ ，符号检验）认为第一种分配比第二种分配更有可能。扰动的出现增加了代表性，也因此增加了不确定事件发生的表面可能性（apparent likelihood）。

被试的回答，似乎忽略了这两种分配的单一性，转而比较这两类分配，而无视上面的特定分配。这并不意味着被试没有意识到类别与其实例的区

分，他们未意识到的是这一区分对相关频率判断的合理影响。

人们认为机会不可预知但本质上是公正的。因此，在一次纯粹的随机分配弹珠的过程中，他们期望每一位儿童应当获得大致（但不是真正）相同的弹珠。同理，他们认为即使很短的一个硬币投掷的序列也应包含相同数目的正面和反面。更一般地，有代表性的样本，不但在整体上体现了总体的本质特征，且在每一个局部都是如此。然而，一个具有局部代表性的样本，会系统地偏离机会预期：它包含太多的变换和太少的聚类（cluster）。

大数定律保证极大的样本能高度代表它所从中抽出的总体。在其他地方（Tversky & Kahneman, 1971, 2），我们根据“大数定律同样适用于小样本”的断言，将局部代表性的预期作为小数律信念的一个特征。我们认为，这一信念支撑着有关随机性的错误直觉，它明显地体现在多种情境中。

对随机性感知的研究（Tune, 1964; Wagenaar, 1970）表明，当人们被要求模拟随机过程时，如硬币投掷的序列，他们会给出具有局部代表性和有许多短簇^①的序列。另外，人们倾向于认为那些短簇的长度有正确分布的序列不太可能发生，或者是不随机而予以拒绝。这也许是因为长簇^②不具有局部代表性。 37

在大量有关概率学习和二元预测（binary prediction）（Estes, 1964; M. R. Jones, 1971）的研究中，也发现了相似结果。赌徒谬误（gambler's fallacy）或者说否近效应（negative-recency effect），就是信守局部代表性的一个明证。因为如果两种结果的比例关系必须在较短区间内予以保持的话，则一种结果的长序列必然紧跟另一结果的长序列以维持平衡。事实上，在局部代表性的世界中，赌徒谬误不再荒谬。

在费乐的《概率理论导论》（*Introduction to Probability Theory*）一书中，他（Feller, 1968, p. 160）曾举例说明局部代表性的错误信念。第二次世界大战期间，在对伦敦的密集轰炸中，人们普遍认为轰炸模式不可能是随机的，因为市区中有的地方被轰炸了数次，而其他地方却丝毫未损。因此，轰炸模式违背了局部代表性，随机轰炸的假设看来不成立。为检验这一假

① 即长的连续出现的正面或反面。——译者注

② 即短暂的连续出现的正面或反面。——译者注

设, 整个伦敦南部地区被分割为相同面积的小块, 然后将每小块轰炸次数的实际分布与随机轰炸假设下的分布 (泊松分布) 相对比。与这种普遍信念相反, 这两种分布非常吻合。“对未受过训练的人来说”, 费乐评论道: “随机似乎有规则性或者有聚类倾向 (tendency to cluster)。”

大部分学生惊讶地发现在一个只有 23 人的小组里, 至少有两人生日相同 (也就是同月同日) 的概率超过 0.5。显然, 23 人的预期生日天数与一年中的天数之比要小于 $1/15$ 。因此, 一天有两个生日, 而 343 天则是“空缺”的情况, 极不具有代表性, 因此, 问题中所述事件也就显得不太可能。更一般地, 我们推测正是概率理论中许多结果的反直觉的本质促成了它们对代表性的违背 [有关随机游动理论中的一个引人注目的例子, 可参见费乐 (Feller, 1968, pp. 84-88)]。

如此说来, 一个有代表性的样本, 在本质特征上与总体相似, 并且如人们认为的那样体现随机性; 也就是说, 它的所有部分都具代表性, 且无一部分过于规则。在所有可能的样本中, 只有少部分符合所有这些约束条件。大部分样本并不符合, 也就显得不随机。比如说, 在投掷 6 次硬币结果的 20 种可能的序列中 (不考虑方向与标号), 我们敢说, 只有正反反正反正似乎是真正随机的。而在投掷 4 次硬币的序列中, 可能没有一个。

认为某些二元序列 (binary sequence) 比其他二元序列更为随机的倾向, 在天空广播实验中有着重要的影响。^[3]在这项实验中, 听众被要求猜测由一个游戏小组编制的五个二元符号 (binary symbols) 的具体含义。通过对超过百万个回答的分析 (Goodfellow, 1938), 研究者发现有些序列被猜中的概率远远超过随机概率, 而有些则远远低于随机概率, 这主要取决于目标序列的表面随机性。这项发现对实验社会心理学 (ESP) 研究的意义是显而易见的。

对表面随机序列的语言描述是最冗长的。试想你在口述一个长序列的二元符号, 比如说正面与反面。你毫无疑问会使用简便的表达, 如“四个反面”或者是“正面一反面, 连续三次”。一个有许多长簇 (long run) 的序列, 可以使用第一种简便表达, 而有许多短簇的序列, 则要求第二种简便表达。一个表面随机序列的呈现结构, 最大地限制了简便表达的便利性,

也因此拒绝了经济的描述。因此，表面随机性，是一种结构复杂的形式。而结构复杂性的决定因素，如可编码性（Garner, 1970; Glanzer & Clark, 1963; Vitz & Todd, 1969），也会影响表面随机性。

二、抽样分布

我们已提出，被试总是给更具代表性的事件分配更高的概率，而给同等代表性的事件分配同等的概率。在这一节，我们将考察这一假设对主观抽样分布研究的意义（subjective sampling distribution）。而主观抽样分布，也就是被试分配给从一特定总体中抽出的既定容量的样本的概率。

当用诸如比例或均值等单一统计量来描述样本时，该样本能在多大程度上代表总体，取决于该统计量的数值与相应的总体参数的相似性。因为样本容量并不能反映出总体的任何特征，它不影响代表性。因此，比如说，在1 000个婴儿的样本中发现600个男孩的事件与在100个婴儿的样本中发现60个男孩的事件，具有同等的代表性。这两个事件也就被认为有同等的可能性，尽管后者事实上更有可能。同理，按照现在的分析，一个平均身高在6英尺与6英尺2英寸之间的男性样本的主观概率，与样本容量无关。

为了检验这一预测，9组被试被要求给出以下3种总体中3种样本容量（ N 为10或100或1 000）下的主观抽样分布：

（1）性别分布（二项分布， $p = 0.50$ ） 被试被告知某一地区每天大约有 N 个婴儿出生。当 $N = 1\,000$ 时，问题如下：

39

在1 000个婴儿中有下列数目男孩的天数与总天数的比例各是多少：

0 ~ 50 个男孩

50 ~ 100 个男孩

150 ~ 250 个男孩

.....

850 ~ 950 个男孩

多于 950 个男孩

请注意这些类别包括所有的概率，您回答的总和应为 100%。

PDG

当 $N=100$ 时，有 11 个类别：0~5，5~15 等。当 $N=10$ 时，每一结果就是一类别，如 6 个男孩。

(2) 心跳类型的分布（二项分布， $p=0.80$ ） 被试被告知某地区每天大约有 N 个婴儿降生，且 80% 的新生儿心跳类型是 a ，剩下的 20% 的心跳类型是 b 。被试要对 3 种容量的样本按照如图 3—1 和图 3—2 中 11 种类别给出每天新生儿心跳类型为 a 的抽样分布。

(3) 身高分布 被试被告知某地区的就业中心记录每天来访男性的平均身高。被试了解到男性总体的平均身高在 170~175 厘米之间（在以色列身高是按厘米来测量），且身高离均值越远频率越小。对每一类样本容量，被试要给出如图 3—3 所示的 7 类平均身高的抽样分布：低于 160，160~

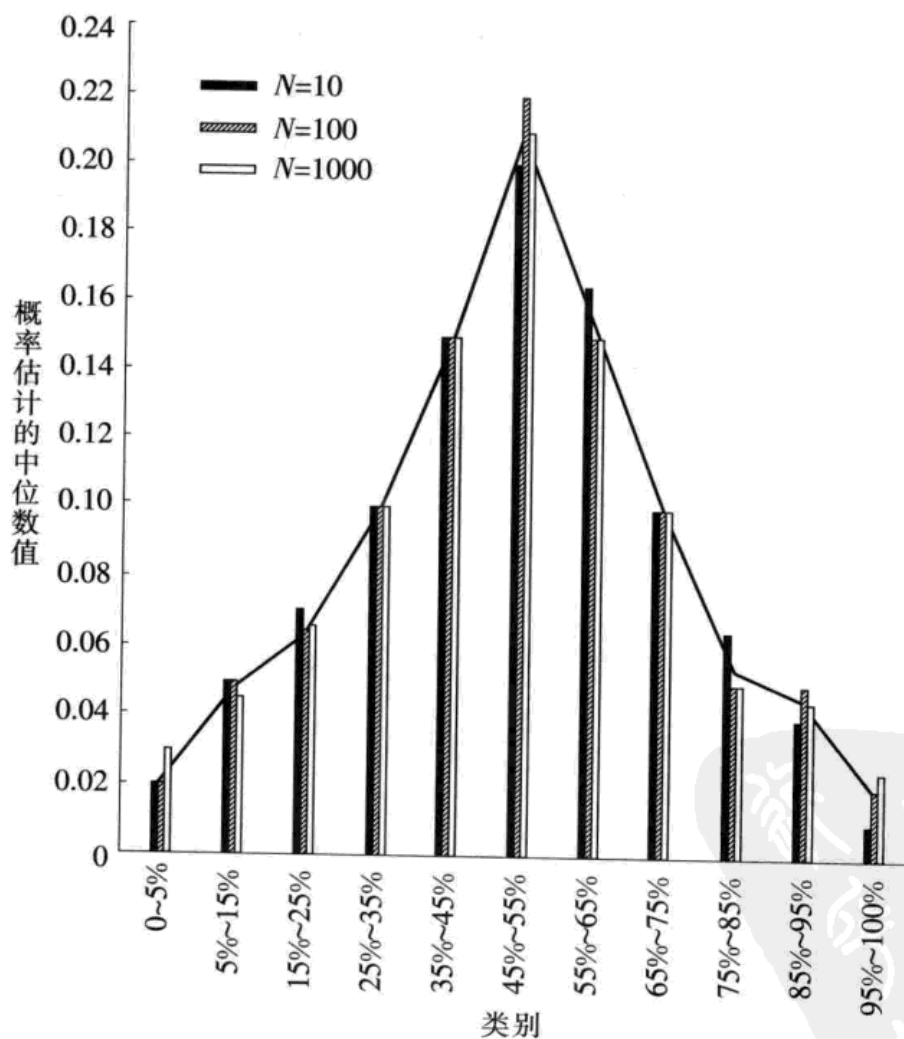


图 3—1 性别分布

165, ……高于185。

在图3—1、图3—2、图3—3中，我们给出了3种总体在不同取值 N 下估计的中位数值（被试的数量在45与84之间，均值为62）。显然，样本容量对主观抽样分布几乎无任何影响。各自独立的小组，面对仅仅是样本容量不同的相同问题，给出相同的分布。这一结果对于抽象的总体，如二项式以及日常生活中所熟知的总体，如男性身高，同样成立。

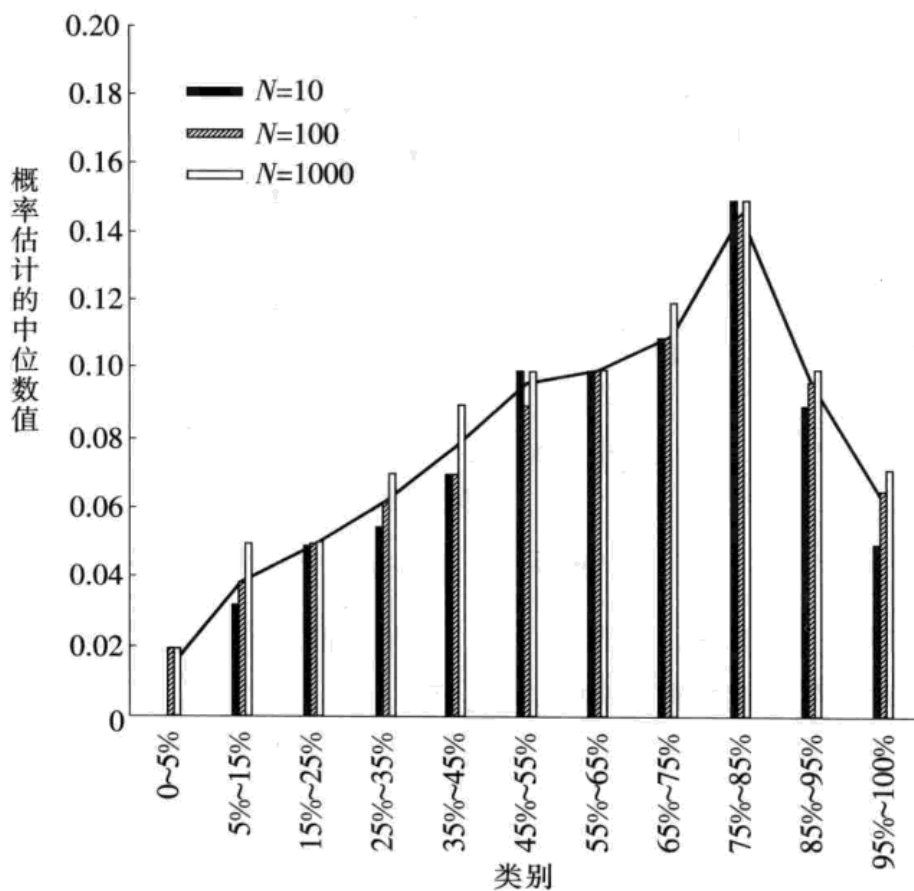


图3—2 心跳类型分布

因为主观抽样分布与 N 无关，在每个图中连接估计中位数的均值的实线可被认为对各自总体是“通用”的抽样分布。为说明被试完全忽略的样本容量真实影响的大小，我们分别在图3—4和图3—5中给出了 $p=0.50$ 和 $p=0.80$ 时的正确抽样分布以及相应的“通用”抽样分布。

可以看出，那条“通用”的曲线比 $N=10$ 时的正确曲线还要平坦。 $p=0.50$ 时，“通用”抽样分布的方差（0.048）差不多等于 $N=5$ 时正确样

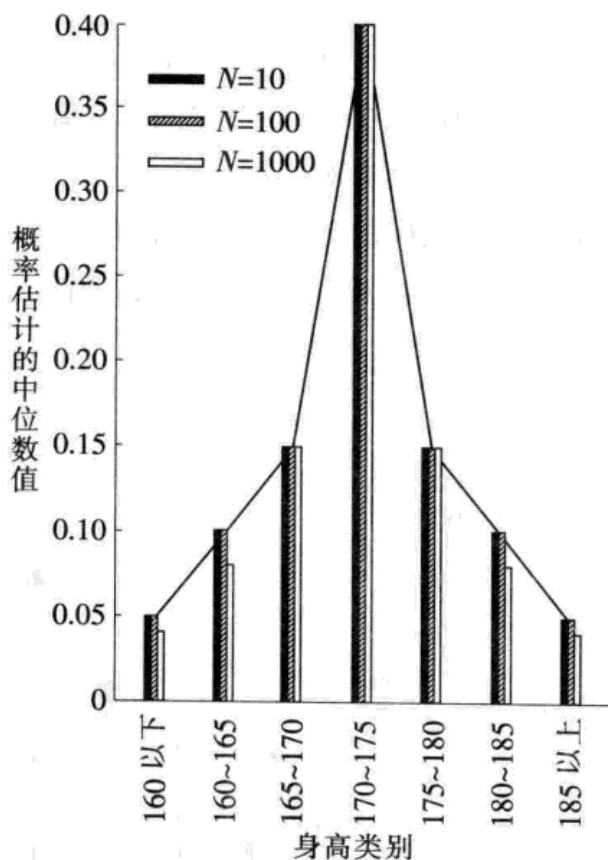
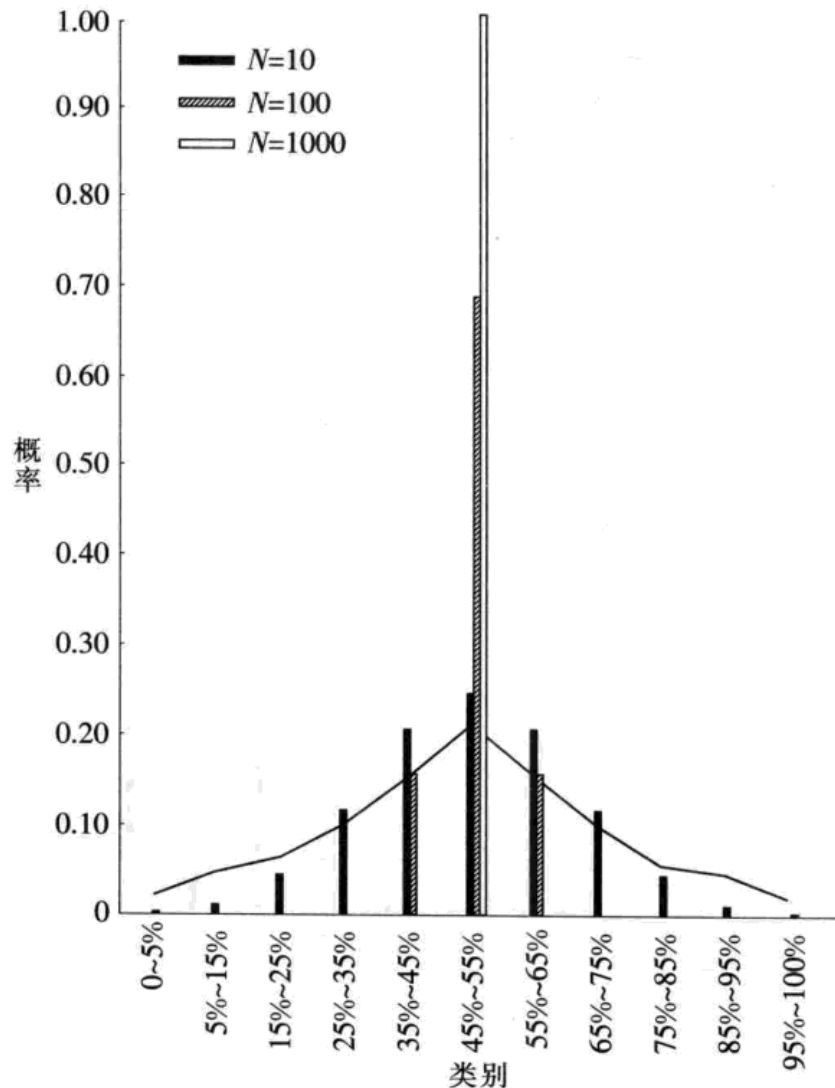


图 3—3 身高分布

本分布的方差 (0.05)。 $p=0.80$ 时，“通用”曲线的方差 (0.068) 在 $N=2$ 及 $N=3$ 时的正确抽样分布的方差之间。

在二项分布中，均值通常与众数相等。因此，当 $p \neq 0.50$ 时，较短的尾端应比较长的尾端高；比如，请看图 3—5，当 $N=10$ 时的正确分布。图 3—5 说明 $p=0.80$ 时的“通用”曲线违背这一特点，它的均值只有 0.63。因此，尽管主观抽样分布的众数合理地出现在最具代表性的值上，均值则出现在较长尾端的方向上。同样的结果在其他研究，如科恩等的研究中，也得到同样的结果。因此， $p=0.80$ 时比例的“通用”样本分布根本不是二项分布。

这一实验在两方面区别于以往的主观二项分布的研究 (Peterson, DuCharme & Edwards, 1968; Wheeler & Beach, 1968)。第一，早期研究关注的样本的容量远小于这一研究。第二，更重要的，对不同的样本容量被试要给出概率分布的事件数目不同：样本容量为 N 的样本，被试需评估 $N+1$ 个结果。与之不同，在这项研究中，被试对所有样本容量的样本都评估相

图 3—4 $p = 0.50$ 时的抽样分布

同数量的类别。当类别的数量变化或者是样本足够小以致可以心算出概率时，图 3—1、图 3—2、图 3—3 中所展示的主观抽样分布对样本容量 N 不变的特点将不存在。对大样本，心算是不可能的，自然的解决方法是直接运用代表性启发式，而这主要由样本均值或样本比例掌握着。

为进一步探讨样本容量在代表性启发式的预测中的影响，我们另外做了一项实验，被试是 97 名斯坦福大学的没有概率或统计背景的本科生，他们被分成由 5 至 12 名成员组成的小组。按照固定顺序，被试看到三个问题，每一问题都定义了一个抽样过程并给出特定的均值和一个待评估的高于均值的数字，然后要求判断一个特殊的抽样结果更可能在小样本还是大

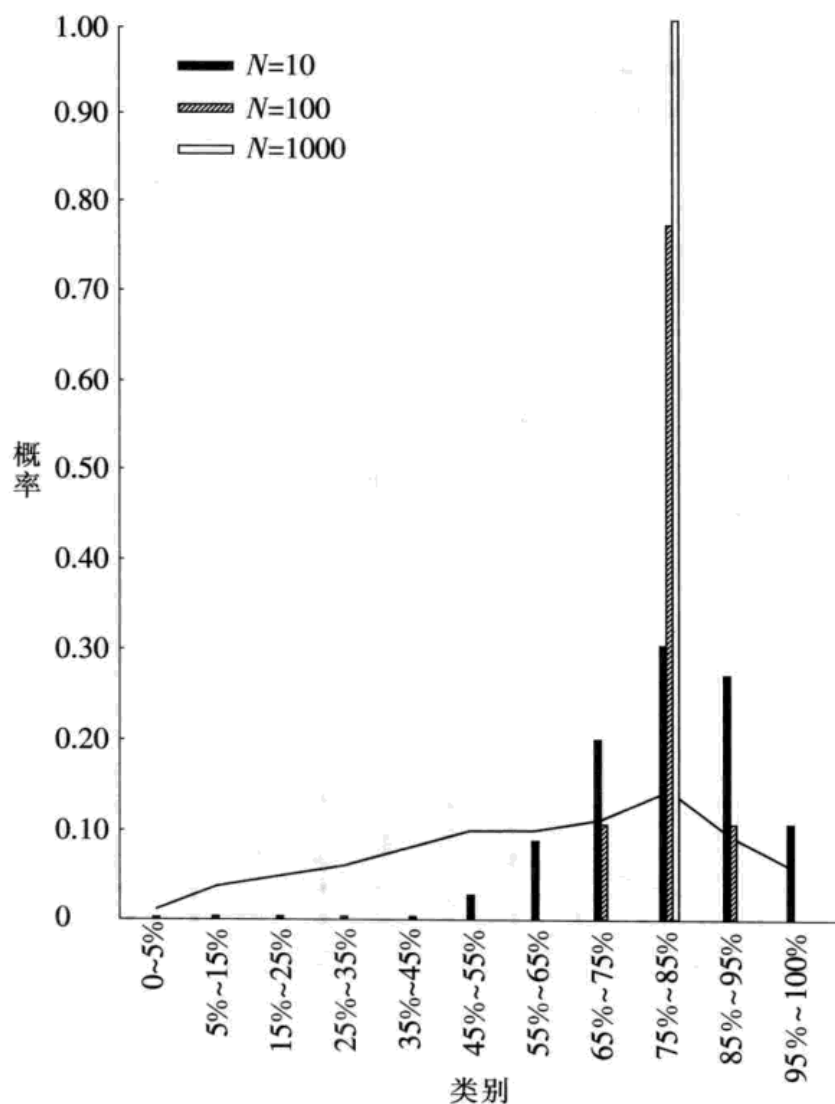


图 3—5 $p = 0.80$ 时的抽样分布

样本中发生。每位被试都获得 1 美元的实验报酬；如果他答对其中一个问题（完成任务后从中随机抽取一个），将获得额外 1 美元的奖励。

为控制反应偏差（response bias），每个问题都以两种方式出现。一半被试判断，在 3 个问题中，与给定数值相比，更极端的值更可能在小样本还是大样本中出现。正确答案当然是极端值更可能在小样本中出现。剩下的被试判断，与给定数值相比，不极端的值更可能在小样本还是大样本中出现。这里正确答案是这种数值更可能在大样本中出现。这 3 个问题如下。其中，表格中的数据是两种问题形式中选择每一回答类别的人数。正确的答案用星号表示。

(1) 某镇有两个医院，大医院每天约 45 名婴儿出生，小医院每天约 15 名婴儿出生。众所周知，约 50% 的婴儿是男孩。但每天男孩出生的实际比例都不一样，有时高于 50%，有时低于 50%。在一年的期间内，每家医院都记录（多于/少于）60% 的新生儿是男孩的天数。请问你认为哪家医院会记录更多这样的天数？

	多于 60%	少于 60%
大医院	12	9*
小医院	10*	11
大致一样（即在 5% 的差异之内）	28	25

(2) 一位研究语言特性的研究者选出一本平装书，并计算该书每页单词的平均长度（即用该页字母数除以单词数）。另一位研究者挑每页的第一行，并计算该行单词的平均长度。整本书单词的平均长度是 4，但并不是每一页或每一行都如此。有些平均长度高些，有些则低些。第一位研究者记录每页单词平均长度为 6 或多于/少于 6 的数量。第二位研究者记录每行单词平均长度为 6 或多于/少于 6 的行数。你认为哪位研究者将获得更大的数值（其中前者单位是页，后者是行）？

	多于 6	少于 6
页数研究者	8	10*
行数研究者	21*	15
大致一样（即在 5% 的差异之内）	20	23

(3) 一项医学调查旨在研究冠状动脉疾病的相关因素。两个小组正收集数据。其中一组每天检查三个人，另一组每天检查一人。被检查者是从总体中随机抽取的。在检查中每个人身高都要予以测量，成年男性的平均身高是 5 英尺 10 英寸，高于平均值与低于平均值的人一样多。每天检查 3 人的小组对他们的身高排序，并记录下中等身高的人身高高于/低于 5 英尺 11 英寸的天数。另一组仅记录被检查者高于/低于 5 英尺 11 英寸的天数。你认为哪一小组记录的天数更多？

	高于 5 英尺 11 英寸	低于 5 英尺 11 英寸
检查 3 人的小组	7	14*
检查 1 人的小组	18*	17
大致一样（即在 5% 的差异之内）	23	17

如果被试了解样本容量所扮演的角色，他们会轻而易举地挑出这几个简单序数问题的答案。相反，如果他们认为同等代表性的结果发生的可能性等同，他们将不会表现出对正确答案系统一致的偏好。事实正是如此。在几乎所有的比较中，最多的回答都是“一样”；而且，没有哪个问题表现出对正确答案的明显偏好。

尽管有一些程序上的差异，但这项实验确认了我们前面研究的结论。这里，每位被试，都在追求准确性的实验设计中，对两个不同样本容量下结果发生可能性做一个直接的排序判断。这一程序可以突显样本容量的显著性。并且，最后一个问题，将单个观测与三个观测组成的样本的均值进行比较。显然，被试没有注意到一个最明显的事实，即中位数值比单个观测值更稳定。

抽样变异（sampling variance）与样本容量成比例降低的观念，显然不是人们直觉中的常备曲目。事实上，有关样本容量的错误观念，在我们日常生活中经常发生。一方面，人们愿意对用百分比表示的结果认真看待，却忽略了观测的数值可能过分的小。另一方面，人们面对从大样本处获得的确凿证据依然持怀疑态度，正如一位著名的政治家措辞激烈地抱怨道，生活成本指数（cost of living index）不是建立在整个人口总体的基础上，而只是个大样本，他还补充道：“更糟糕的——还是个随机样本。”

我们当然不是说人们无法意识到样本容量对抽样变异存在影响。人们可以学会正确的规则，甚至没有多大难度。问题在于，当让他们独自判断时，他们并不遵守正确的规则。另外，对研究心理学家行为的研究（J. Cohen, 1962; Tversky & Kahneman, 1971, 2）表明，他们有一种很强的低估样本容量影响的倾向，尽管他们拥有有关正确规则的知识和广泛的统计训练。对任何一位认为人们是合理的直觉统计学家的人来说，这些结果

都是令人沮丧的……

三、规范模型和描述性启发式

有些学者 (W. Edwards, 1968, 25) 认为人们大体上遵守正确的贝叶斯规则, 只是不能看清证据的全部影响, 因此是保守主义的。有人 (Peterson & Beach, 1967) 总结说, 规范模型提供了一个很好的对被试行为的近似描述, 认为“被试在合适的方向上受到合适变量的影响” (第 43 页)。这一观点并不为所有人所认可。较近的一篇文献评论 (Slovic & Lichtenstein, 1971) 认为, 以上对人们的表现作为一个直觉统计学家的评价“过于慷慨大方”; 而皮茨等 (Pitz et al., 1967) 根据他们的数据总结说, 人们在贝叶斯任务中的表现, “比保守主义说法所暗示的更差” (第 392 页)。

用规范的贝叶斯路径来对主观概率进行分析和建模的有效性, 很大程度上并不取决于主观估计的准确性, 而是这一模型是否抓住了判断过程的主要决定因素。这篇论文中的研究表明它并没有做到。特别地, 我们注意到样本容量对样本分布没有任何影响, (至少在总体上) 后验二项估计 (posterior binomial estimates) 是由样本的比例而不是样本的差异来决定的, 还有, 样本分布并不取决于总体比例。在他的证据评估过程中, 人们显然不是一个保守的贝叶斯主义者: 他根本就不是贝叶斯主义者。

有人也许会说, 规范模型对人们行为描述的失败, 仅局限于面对不熟悉的随机过程的普通被试, 它对人们对日常生活中碰到的熟悉的随机过程的评估, 仍然能提供充分的解释。然而, 几乎没有证据支持这一观点。第一, 研究已表明 (Tversky & Kahneman, 1971, 2), 在严谨科学家的直觉判断中, 同样能发现援用代表性启发式所引起的相同类型的系统性错误。显然, 对概率理论的了解, 并不能消除所有的有关机会定律的错误直觉。第二, 在日常生活中, 我们会碰到许多以很高的近似度符合二项式定理 (binomial law) 的随机过程 (如生男孩还是生女孩, 在一个既定交叉路口撞上红灯, 抓了一手没有红心的牌), 但人们并没从这些经验中获取对二项式过程 (binomial process) 的充分理解。显然, 广泛的对众多实例的接触, 并不足以产生理想的行为。

在日常生活中，人们会问自己或他人这类问题：这个 12 岁的男孩长大后成为一名科学家的可能性有多大？这位候选人当选的可能性有多大？这个公司关门倒闭的可能性有多大？这些问题与我们论文中早先讨论的问题的不同，在于它们的特质，即人们既不能用过去发生的频率也无法用定义完备的抽样过程来给出现成答案。

在本章中，我们详细地考察了一种启发式，人们依此通过评估一事件多大程度上代表产生该事件的过程或总体的主要特征来判断事件发生的概率。尽管我们的实验局限于定义完备的抽样过程（客观概率可以计算），我们推测，这一启发式，在那些没有“正确”答案的独特情况下对不确定性的评估中，也扮演重要的角色。比如说，那位 12 岁的男孩成为科学家的可能性，可以通过判断这位男孩的形象在多大程度上代表了科学家的角色来评估。同理，当思考某公司关门或某位政治家当选的可能性时，我们的大脑中也需存有公司或政治局势的模型，那些最能代表相应模型的核心特征的结果将被评估为最有可能发生的……

本章是《认知心理学》（*Cognitive Psychology*, 1972, 3, 430-454）上的一篇论文的缩写。允许重印。

注释

[1] 我们用“主观概率”来指被试给出的或者由他的行为推断出的对一事情发生概率的估计。这些估计并不需要满足任何公理或者一致性要求。我们用“客观概率”来指基于既定假设，并根据概率运算的法则算出的数值。显然，这一术语与任何概率哲学观无关。

[2] 本章只讨论完整硬币（fair coin，均质硬币）的情况。

[3] 感谢 R. P. Abelson 提请我们注意这项研究。

第4章

论预测心理学

卡尼曼和特沃斯基 (Daniel Kahneman and Amos Tversky)

在这篇论文中，我们将探讨决定直觉预测和信心判断的规则，并将其与标准的统计预测原则进行比较。我们讨论两类预测：类别预测（category prediction）和数值预测（numerical prediction）。在类别预测中，预测以定类（nominal）形式出现，比如，竞选中的胜出者、对病人的诊断或一个人将来的职业。在数值预测中，预测以数值形式出现，比如，某一股票的将来市值，或某一学生的平均绩点（grade point average, GPA）。

在不确定状况下进行预测与判断时，人们并不根据概率的计算或者预测的统计理论。相反，他们依赖于数目有限的几种启发式，他们有时能获得合理判断，有时却导致严重的系统性错误（Kahneman & Tversky, 1972b, 3; Tversky & Kahneman, 1971, 2; 1973, 11）。本章关注其中一种启发式即代表性启发式在直觉预测中的作用。

在给定证据下（如人格描述），所考虑的结果（如职业或成就水平），可以按它们在多大程度上代表了证据来排序。本章的论点就是，人们用代表性启发式来预测，也就是说，他们按照结果体现证据的本质特征的程度来选择或者排列结果。在很多情景中，具代表性的结果，确实比其他结果更可能发生。然而，事实并不总是如此，因为有些因素（如结果的先验概率以及证据的可信度），而不是其代表性，会影响结果发生的可能性。由于49 忽视了这些因素，直觉预测，系统地和本地地违背了统计预测的规则。为证实这一假说，我们将说明按可觉知可能性对结果的排序，与按代表性的排序是一致的，并且直觉预测基本上不受先验概率和预期的预测准确性的影响。

在第一节，我们将考察类别预测，并说明它符合一种独立的代表性的评估，且它基本上也独立于结果的先验概率。在第二节，我们将考察数值预测，并说明它并不是合理的回归性的，且基本上不受证据可信度的影响。接下来的三节依次讨论有关预测研究中的方法论问题、预测中过分信心的来源，以及有关回归效应（regression effects）的错误直觉。

一、类别预测

1. 基率、相似度和发生可能性

下面这个实验的例子展现了用代表性启发式做出的预测，以及与这种直觉预测模式相关的谬见。一个由 69 位被试组成的小组^[1] [基率组（the base-rate group）] 被要求回答下面这个问题：“请考虑今天在美国的所有的一年级研究生。请写下你对以下 9 个专业领域的注册学生所占比例的最佳估计。”这 9 个领域见表 4—1。表格中的第一列是各个领域基率估计的均值。

表 4—1 对 9 个专业领域基率的估计，和有关汤姆的相似度及预测数据的总汇

专业领域	判断出的基率的均值 (%)	相似度排名的 均值	可能性排名的 均值
工商管理	15	3.9	4.3
计算机科学	7	2.1	2.5
工程学	9	2.9	2.6

续前表

专业领域	判断出的基率的均值 (%)	相似度排名的 均值	可能性排名的 均值
人文学与教育学	20	7.2	7.6
法律	9	5.9	5.2
图书馆管理学	3	4.2	4.7
医学	8	5.9	5.8
物理与生命科学	12	4.5	4.3
社会科学与社会工作	17	8.2	8.0

第二个小组的 65 位被试（相似度组）被给出下面这个人格描述：

汤姆有很高的智商，尽管他缺乏真正的创造力。他有着对秩序、清楚、整洁和所有部分各司其职的整齐系统的渴望。他的文章相当枯燥和机械，只是偶尔的陈腐的双关语和科幻小说式的想象迸出的火花才令其生色。他有强大的动力去胜任工作。他看起来对他人毫不在意，缺乏同情心，也不愿与人交流。尽管以自我为中心，但他有很强的道德感。

被试被要求按照“汤姆与上述 9 个领域中的典型的研究生有多像”来对这 9 个领域排序。表 4—1 中的第二列就是给不同领域的相似度排名的均值。

最后，预测组由 114 名美国三所主要大学的心理学研究生组成，他们被给出了汤姆的人格描述，但附加了下面的额外信息：

前面的人格描述是一位心理学家在汤姆读高中高年级时，根据投射测验写下的。汤姆现在是一位研究生。请根据汤姆现在分别是这 9 个领域中的研究生的可能性，对这 9 个领域排序。

表 4—1 中的第三列是预测组的被试赋予结果的排名的均值。

我们计算了表 4—1 中各列之间的积矩相关系数（product-moment correlation）。判断的可能性与相似度之间的相关系数是 0.97，而可能性与估计的基率^[2]之间的相关系数是 -0.65。显然，对可能性的判断与相似度的判断基本一致，而与基率的估计截然不同。这一结果，对人们按照代表性或

者相似度来预测的假说，是一个直接的证明。

由心理学研究生给出的对可能性的判断，严重违反了标准的预测规则。多于 95% 的回答者认为汤姆更可能学计算机科学而不是人文学或者教育学，但同时他们也确知在后一领域有更多的研究生的事实。根据表 4—1 列出的基率估计，人文学或教育学与计算机科学的先验几率之比（prior odds）
51 大约为 3:1（实际比率要高很多）。

根据贝叶斯规则，仅当对汤姆的人格描述既准确又切中要害时，才有可能逆转汤姆更可能是人文学或教育学的研究生而非计算机科学研究生的倾向。然而，我们研究中的研究生并不认为这一条件得到了满足。在预测任务之后，被试被要求估计通过数种不同信息实现一选就中（在 9 个领域中第一选择就正确）的比例。当预测是基于投射测验时，估计的一选就中的比例的中位数是 23%，而基于高中生对自己的兴趣与计划的报告的预测，这一数值是 53%。显然，投射测验并不被看重，不过，研究生们确实是依赖于由此种测验获得的描述，而忽视了基率。

总体上，有三类信息与统计预测相关：（1）先验信息或者是背景信息（如研究生专业的基率）；（2）关于个例的特定证据（如对汤姆的描述）；（3）预测的预期准确度（如估计的一选就中的概率）。统计预测中的一个基本规则是，预测的预期准确度，控制着赋予特定证据及先验信息的相对权重。当预测的预期准确度降低时，预测应该更为回归，也就是更接近按先验信息做出的预测。在汤姆这个例子中，预期准确度较低，先验概率理应被赋予更高的权重。但相反，我们的被试按照与特定证据的相似度来对答案排序，而无视先验概率。

在他们完全依赖于人格描述时，预测组中的被试，显然忽视了以下几点。其一，基于人格投射测验无效性这一臭名昭著的事实，汤姆极有可能并不像人格描述暗示的那样冲动和孤僻。其二，即使对汤姆读高中时的描述是真实的，在汤姆读研究生时可能就不再有效了。最后，即使描述依然是真实有效的，与读计算机科学的学生比较，极有可能在读人文学或教育学的学生中，有更多人符合这一描述，道理很简单，因为有更多的学生就读于前一领域。

2. 对预期准确度的操纵

另外一个研究被用来检验这一假说, 它认为与统计模型不同, 对预测的预期准确度的操纵, 并不影响预测模式。实验材料是5份对9年级男孩的简略的人格描述, 据称这是由一位咨询员在纵向研究(longitudinal study)的访谈基础上写成的。实验设计与汤姆研究中的设计是一样的。同一组中的被试($N=69$), 按照被描述的男孩与“那一研究领域典型的一年级研究生的形象”的相似度, 对这9个研究生专业排序。在相似度判断之后, 他们估计这9个研究领域基准频率(base-rate frequency)。这些估计均见表4—1。剩下的被试被告知, 这5个案例是从现在是研究生一年级学生的早先研究的参加者中随机抽取的。其中一组是高准确度组($N=55$), 他们得知“在此类描述基础上, 与你相似的同学所做出的准确预测约占所有案例的55%”。低准确度组($N=50$)则得知在一个任务中能做出的准确预测约占所有案例的27%。被试对每一描述, 按“被描述的人现在是该研究领域研究生的可能性”对这9个领域排序。他们同样要对每一描述估计他们的第一选择就正确的概率。

预期准确度的操纵, 对概率判断有重大影响。高准确度与低准确度组的平均估计分别为0.70和0.56 ($t=3.72, p<0.001$)。然而, 在低准确度的条件下, 对9个领域排序的结果与高准确度条件下的排序结果相比, 并不显著地更接近基率分布。对每一判断, 研究者计算出被试赋予(五项描述)9个研究领域的平均排名与基率间的积矩相关系数。这一相关系数, 是对被试的预测多大程度上符合基率分布的总体衡量。高准确度组和低准确度组的单个相关系数的均值分别是0.13和0.16。差异并不显著($t=0.42, df=103$)。这一判断模式违背了预测的标准理论, 后者认为任何预期准确度的下降, 都将伴随着预测对基率的移动。

由于预期准确度的操纵对预测没有影响, 这两个预测组被汇总在一起。随后的分析与汤姆研究是一样的。对每一描述, 两个相关系数被算出: (1) 可能性的平均排名与相似度的平均排名之间的相关系数; (2) 可能性的平均排名与平均基率之间的相关系数。对每一描述判断出的最可能的结果和这些相关系数都列示于表4—2。预测与相似度之间的相关系数始终很

高。与之相反，预测与基率之间没有系统的关系：相关系数波动很大，它取决于每一描述中最具代表性的结果是否经常或极少发生。

表 4—2 可能性平均排名与相似度平均排名、基率之间的积矩相关系数

	统计数据中的第一选择				
	法律	计算机科学	医学	图书馆管理学	工商管理
与相似度平均排名	0.93	0.96	0.92	0.88	0.88
与基率	0.33	-0.35	0.27	-0.03	0.62

59 这里对基率的考虑再一次被忽略。在统计理论中，只有当个体认为自己的预测万无一失时，他才可以忽略基率信息。在所有其他情况下，在按照描述所做出的排序与按基率的排序之间必须取得合理的折中。难以相信，基于一次访谈中的对 14 岁少年的粗略描述，就可以成为由被试的预测所暗示的其预测绝对可靠的理由。

在 5 份人格描述之后，让被试回答另一问题：

对于丹，除了他参加了原来的研究而现在是一年级研究生之外，你没被告知任何其他信息，请记住你的排序，并回答你对这一案例的信心。

丹的可能性平均排名与估计的基率之间的相关系数是 0.74。因此，在具体描述给定的情况下没被利用的基率信息，在没有具体证据可资利用时，会被利用。

3. 先验证据对个体证据

下面的一项研究，为直觉预测受制于代表性启发式且对先验概率相对不敏感的假设，提供了更严格的检验。在这项研究中，先验概率将十分显著，且与回答模式谐调一致。被试看到下面的封面故事：

一组心理学家访问 30 位工程师和 70 位律师，并进行人格测验，他们在各自的领域都很成功。在这些信息的基础上，他们写出 30 位工程师与 70 位律师的简略描述。你将在表格上发现 5 份描述，它们是从

100 份可利用的描述中随机抽取的。对每一份描述，请在 0 ~ 100 的量表上记下你认为被描述者是工程师的概率。

同样的任务由一组专家完成，他们所给出的对不同描述的概率是高度准确的。如果你的估计接近专家组的估计，你将得到奖金。

这些说明，提供给 85 名被试组成的小组（低比例工程师组，或 L 组）。而另一被试组（高比例工程师组，H 组； $N = 86$ ），也给出同样的说明，只是先验概率不同：他们被告知，抽取这些描述的集合由 70 位工程师和 30 位律师组成。所有被试都得到 5 份同样的描述，其中一份描述如下：

杰克今年 45 岁，他已婚并有 4 个小孩。总体上，他保守、小心、雄心勃勃。他对政治和社会话题没有丝毫兴趣，大部分业余时间都花在他众多的嗜好上，比如家庭木工、航船、数学难题。

在 100 人中有 30 位工程师的样本中，杰克是工程师的概率是_____%。

在五份描述之后，被试碰到一份无内容的描述：

假设现在你没有获得有关从样本中随机抽取的人的任何信息。

此人是由 100 人组成的样本中的 30 位工程师之一的概率是_____%。

对每份描述，在高比例工程师组和低比例工程师组，一半被试要求给出被描述者是工程师（如上面的例子）的概率，而另一半被试给出被描述者是律师的概率。这一操纵没有任何影响。对每份描述，在两份不同表格中对“工程师”和“律师”给出的概率相加约为 100%。因此，两种表格的数据被汇总，结果以工程师的形式列示。

这项实验设计，使研究者能够算出标准的合理判断模式，这项推导是依据贝叶斯公式，以几率（odds）的形式出现。用 O 来表示某一描述属于

工程师而不是律师的几率，按照贝叶斯规则， $O = Q \cdot R$ ，其中 Q 表示随机抽取的表述是工程师而不是律师的先验几率； R 是对某一描述的可能性比率，也就是，从工程师总体中随机抽取的个体与从律师总体中随机抽取的个体被如此描述的概率之比率。

对得知样本由 70 名工程师和 30 名律师组成的高比例工程师组，先验几率 Q_H 等于 70/30。而对低比例工程师组，先验几率 Q_L 等于 30/70。因此，对每份描述，两组的后验几率的比值是

$$\frac{Q_H}{Q_L} = \frac{Q_H \cdot R}{Q_L \cdot R} = \frac{Q_H}{Q_L} = \frac{7/3}{3/7} = 5.44$$

55 由于可能性比率在公式中被消除，对所有描述都能得到同样的 O_H/O_L 。因此，在现有实验设计下，对操纵先验几率的准确影响，可以在不了解可能性比率的情况下算出。

图 4—1 给出了在两种先验几率条件下，对每份描述的概率估计的中位数。对每份描述，先验几率高时（ $Q_H = 70/30$ ）的估计概率的中位数对应先验几率低时（ $Q_L = 30/70$ ）的估计概率的中位数的点，被绘制出来。按照上一段得出的标准方程，所有的点都应落在曲线（贝叶斯曲线）上，而事实上，只有对应无内容的描述的空白方框落在这条线上：当没有给出描述时，被试在 Q_H 下判断概率为 70%， Q_L 下概率为 30%。在其他 5 个案例中，所有点都落在直线附近。

先验概率的影响虽然细小，但在统计上是显著的。除了无内容描述外，对每位被试，都计算出概率估计的均值。在低比例工程师组中，所有这些值的平均是 50%，而高比例工程师组是 55%（ $t = 3.23$ ， $df = 169$ ， $p <$
56 0.01）。然而，由图 4—1 可以看出，相比贝叶斯曲线，所有点更靠近直线，得出对先验分布的明确操纵对主观概率有极小的影响的结论是合理的。与前一实验一样，被试仅在没有给出明确证据时才使用先验信息。正如代表性假设所必然推出的，当个体证据可资利用时，先验概率大体上都被忽略。

这一效应的强度，在对下面的描述的回答中得以体现：

迪克 30 岁，已婚无孩子。他能力强，有干劲，要在自己的领域中成就一番事业。他也受同事喜爱。

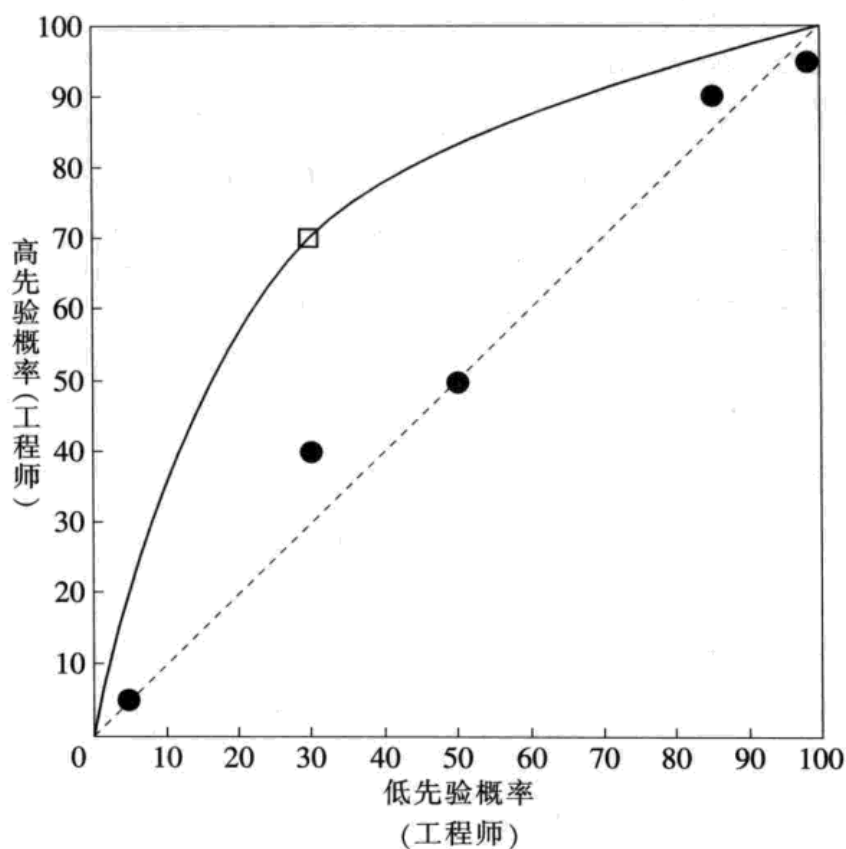


图4—1 在高先验概率和低先验概率下，对五份描述和无内容描述（方框符号表示）的判断概率的中位数（工程师）。（曲线表示的是根据贝叶斯规则得出的正确关系。）

这一描述的组织不包含任何有关迪克职业的信息。我们的被试取得一致：在低比例工程师组和高比例工程师组，估计值的中位数都是50%（见图4—1）。对这一描述与对无内容描述的反应之间的反差是明显的。显然，人们在没给出明确证据和给出无价值的证据时的反应是不同的。在没给出明确证据时，先验概率被合理利用；在给出无价值证据时，先验概率被忽略。^[3]

在有些情况下，先验概率扮演更具实质意义的角色。在迄今讨论的所有例子中，替代结果总是与独特的刻板观念联系在一起，而判断——我们已说明，是由描述代表那些刻板观念的程度所控制。在其他问题中，结果更自然地被视为一个区域内割离的不同部分。比如说，某人被要求判断数位学生中每一位获得奖学金的概率。在这个问题中，没有泾渭分明的奖学

金获得者与非获得者的刻板观念。相反，视结果（即获得奖学金）为由学术成就与能力区域中的分割点决定的，是更自然的事情。先验概率，也就是相关群体获奖学金的比例，可以定位这些分割点，从而可被用来定义结果。因此，它们不太可能被忽略。而且，我们可预期，即使存在有关结果的清晰的刻板观念，极端的先验概率也将产生一些影响。对在什么条件下先验信息会被使用或被丢弃的明确阐释，还有待进一步的研究。

统计预测的一个基本原则是，先验概率总结了在获得独立的明确信息之前我们对问题的了解，即使在获得此类信息之后，它也依然相关。贝叶斯规则，将这一定性原则转化为先验几率与可能性比率之间的相乘关系。然而，我们的被试，没能将先验概率与明确信息加以综合。在面对一项描述时，无论它多么空洞或者可疑，无论是有关汤姆的还是迪克的（工程师/律师），他们都明显感到描述对象所在组中的职业分布不再相关，不能意识到在有明确证据下先验概率依然相关，这也许是直觉与标准预测理论最重大的差别。

二、数值预测

标准预测理论中的一个基本规则是，在一组案例中，预测的变异性，应该反映预测准确度。当预测准确度极佳时，个体将预测标准值（criterion value）。而当不确定性极大时，在所有案例中，个体都将预测固定值（在类别预测中，个体将预测最高频率的类别。在数值预测中，个体会依据损失函数，预测均值、众数、中位数或其他数值）。因此，预测准确度极佳时，预测的变动将等同于标准值的变动。预测准确度为零时，预测变异性也为零。预测准确度为中等时，预测的变异性也取中等的值，也就是，预测依标准而回归。因此，不确定性越大，预测变异性越小。而使用代表性启发式的预测并不符合这一规则。在前一节，我们已知，预测的预期准确度下降时，人们并不会回归于高频率的类别。本节将在数值预测情境下显示类似的错误。

1. 结果预测对输入信息的评估

假设你被告知，一位大学新生被咨询人员形容为聪明、自信、渊博、

勤奋，还富有好奇心。请考虑下面对该描述可能提出的两类问题：

(1) 评估：关于学习能力，这一描述给你留下多深的印象？这种描述适合多少比例的新生，会给你留下更深印象？

(2) 预测：你估计该生会获得多高的平均绩点？获得更高的平均绩点的新生的比例是多少？

这两个问题之间有一重要区别。在第一个问题中，你评估的是输入信息；在第二个问题中，你预测结果。由于第二个问题比第一个问题肯定有更大的不确定性，你的预测应该比你的评估更趋向回归。也就是说，你给出的作为预测的比例将比作为评估的比例更接近 50%。为凸现这两个问题之间的差异，请考虑这项描述不准确的可能性。它应该不会对你的评估产生任何影响：对这些描述给你印象的排序与它们的准确度无关。另一方面，就预测而言，只有当你怀疑描述不准确或者你的预测不准确时，你才会回归。

然而，代表性假设，使预测和评估一致成为必然。在评估一份给定描述时，人们很可能选取最能代表这一描述的分值。如果人们用代表性预测，他们会选取最具代表性的分值作为预测。因此，评估与预测将基本上等同。有多项研究被用来检验这个假设。在每项研究中，被试获得有关一组案例的描述性信息。评价组评价与特定总体相关的每一描述的质量（quality），预测组则预测将来的表现。这两组的判断将被比较，以检验预测是否比评估更趋向回归。

在这两项研究中，被试得到对大学新生的描述，据称是由咨询员在对新生班级的访谈基础上写成的。在第一个研究中，每一描述由五个形容词组成，如我们引用的例子一样，它们涉及智力品质与性格。在第二个研究中，描述是大段的报告，包括这些学生的详细背景和目前对大学生活的适应情况。在两项研究中，评估组都被要求，以评估“在整个班级中，由描述显示的比该生更具学习能力的同学的比例”的方式，评估每一份描述。预测组得到同样的描述，并被要求预测每位学生在大一学期末取得的平均

绩点与百分比方式的班级排名。

这两项研究的结果都在图 4—2 中有显示，它为每一描述描出百分比平均绩点的预测均值对应着平均评估的点。观察到的预测与评估之间的唯一的系统的偏差发生在形容词研究中，在该研究中，预测始终高于对应的评估。用每位被试的数据算出预测或评估的标准差（standard deviation）。在我们研究中的数值范围内，这些数值的对比没有显示出评估组与预测组在变异性上的显著差异。在形容词研究中，评估组（ $N=38$ ）的平均标准差是 25.7，预测组（ $N=36$ ）是 24.0（ $t=1.25$ ， $df=72$ ，没有显著差异，简写为 *ns*）。在报告研究中，评估组（ $N=37$ ）的平均标准差是 22.2，预测组（ $N=63$ ）是 21.4（ $t=0.75$ ， $df=98$ ，没有显著差异，简写为 *ns*）。在两项研究中，预测组与评估组都产生同样极端的判断，尽管前者基于粗略的访谈信息预测出稍具客观性的标准值，而后者仅仅是评估从每一描述那里获得的印象。在预测的统计理论中，只有在预测准确度极佳时，我们观察到的预测与评估之间的相等才具合理性，而在这些研究中，难以想象这一条件能得到满足。

对评估与预测相等的进一步的证据，源于贝丝（Beyth, 1972）的硕士论文。她呈现给三组被试 7 段话，每段话都描述一位实习老师在某一特定试讲课程上的表现。被试是希伯来大学统计课程上的学生。他们被告知，这些描述是从 100 位小学教师的档案中抽取出来的，这些老师在 5 年前完成了他们的教师培训计划。评估组的被试，要求用相对于给出总体的百分比分数的方式，评估这段话描述的课程的质量。预测组的被试，要求用百分比分值来预测每位教师目前的名次，也就是在写出这些描述的 5 年后他的总体胜任能力。评估—预测组要完成这两项任务。与上述研究一样，评估与预测的差异不明显。这一结果对被试之间与被试内（within-subject）的比较同样成立。尽管判断者都毫无疑问知道在一次试讲课程与五年后的教学能力之间有多重因素起作用，这一知识却并未导致他们的预测比评估更趋向回归。

2. 预测对转换

前面的研究说明，对变量的预测相比于对变量输入信息的评估，并不

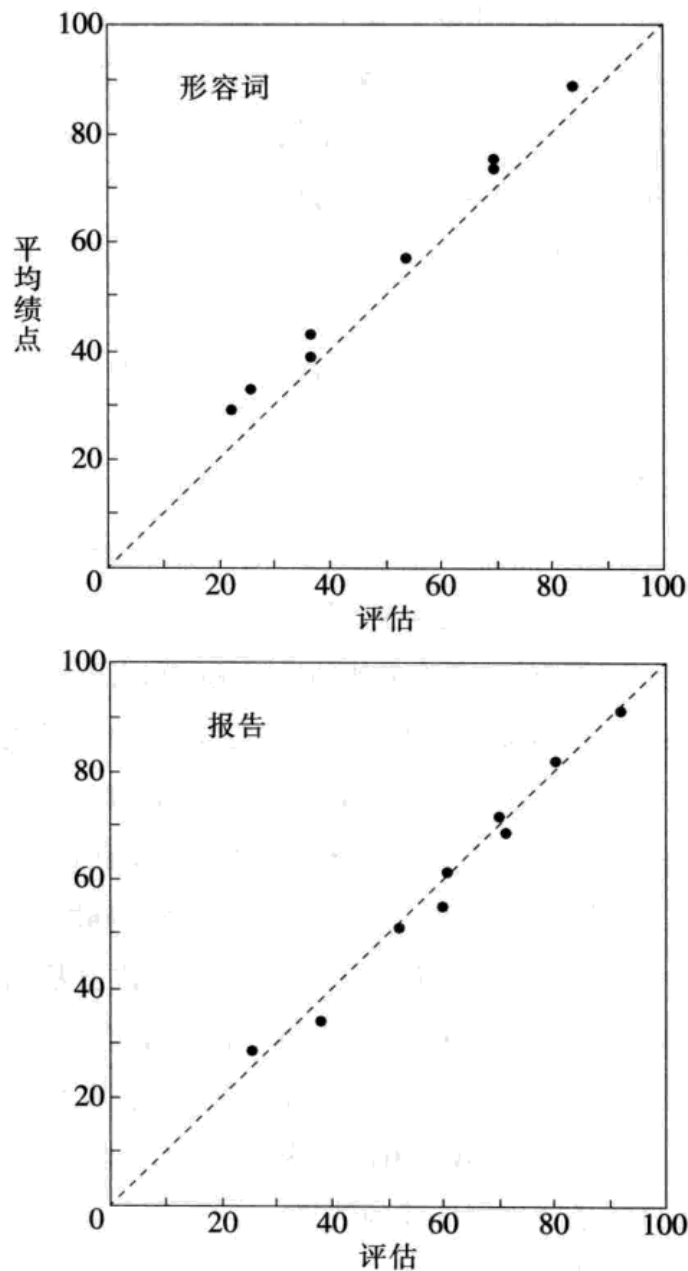


图 4—2 预测百分比平均绩点是对形容词和报告描述的百分比评估的函数

会更趋向回归。在下面这个研究中，我们将说明在有些情况下，对某一变量（学习成就）的预测，并不比变量在不同尺度之间的转换，更趋向回归。平均绩点，被选作输出变量，是因为它具有相关性，且它的分布特性对被试是非常熟悉的。

三组被试参加了实验。所有组的被试，都要在单一每位学生获得的百分比分值基础上，预测 10 位假想的学生的平均绩点。所有小组都拿到同样

的一组百分比分值，只是三组获得如下的对输入变量的不同表述：

(1) 百分比平均绩点：第一组被试 ($N=32$) 被告知，“对这数名学生的每一位，你都会得到一个代表他大学第一年的学习成就的百分比分值，你将需要对他当年的平均绩点给出你的最佳的猜测。”研究者也向被试解释，“比如，百分比分值是 65，表示该生获得的平均绩点好于他所在班级的 65% 的同学，以此类推。”

(2) 心理集中度 (mental concentration)：第二组被试 ($N=37$) 被告知，“对心理集中度的测试是测量集中注意力和提取由复杂信息传达的所有信息的能力。研究发现，获得较高平均绩点的学生，往往在心理集中度测试中取得高分，反之亦然。然而，研究发现，心理集中度测试中的表现取决于测试者参加测试时的情绪和心理状况。因此，同一人重复测试会取得差别很大的分数，这取决于测试前一天晚上他的睡眠状况和当天他的心境。”

(3) 幽默感：第三组被试 ($N=35$) 被告知，“幽默感的测试可测量一个人为卡通配上诙谐的文字说明和欣赏不同形式的幽默的能力。研究发现，在此测试中取得高分的学生比取得低分的总体上趋于获得更高的平均绩点。不过，用幽默感去预测平均绩点，不可能有较高的准确度。”

在这个实验设计中，所有被试都在同一组百分比值基础上预测平均绩点。第一组仅仅将百分比平均绩点的数值转换为平均绩点尺度上的数值。另一方面，第二组和第三组依据关系甚远的证据来预测平均绩点。标准的推理将要求这两组的预测比第一组的判断更回归，也就是有更少的波动。然而，代表性假设暗示了不同模式的结果。

第二组用可能有效但不可靠的、被看作学习能力的一种衡量方式的心理集中度来预测。我们假设这一组的预测与第一组预测相比，并不会更回归。总之，我们推测，最能代表在某一能力测试（如心理集中度）中取得的百分比数值的成就分值（如平均绩点），是对应在成就量度上的同一百分比数值。因为代表性并不受不可靠性的影响，我们预期用不可靠的心理集中度的测试做出的平均绩点预测，将基本上与用百分比平均绩点做出的预测相同。另一方面，我们预期第三组的预测将会回归，因为幽默感的测试，并不被普遍认为是学习能力的衡量方式。

由这三组赋予10个百分比数值的预测显示于图4—3。在该图中，显而易见的是，第二组的预测并不比第一组的预测更回归，而第三组的预测显得更回归。

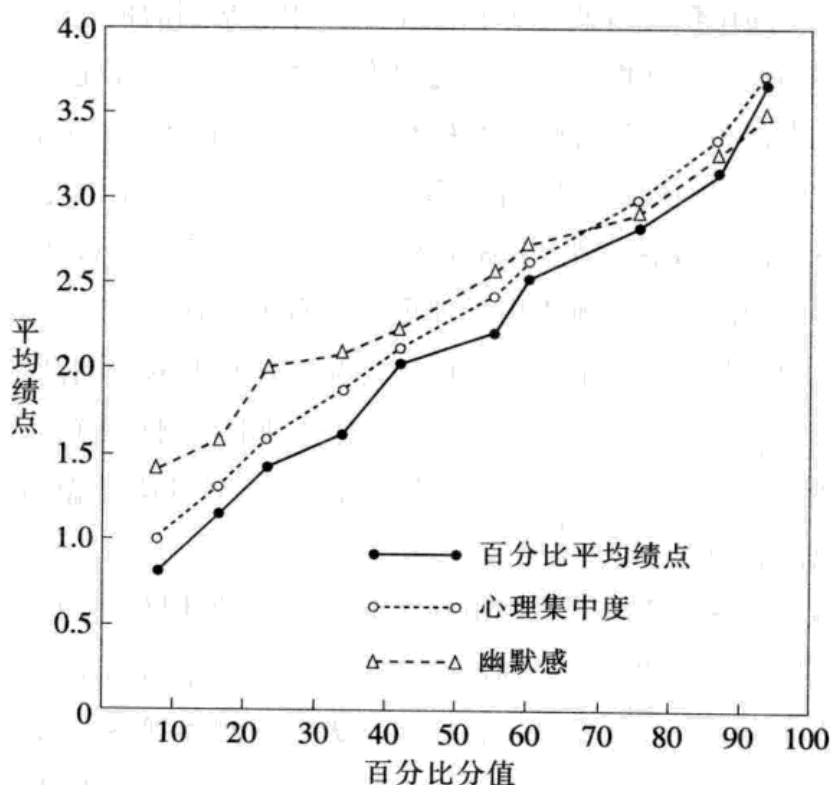


图4—3 用三个变量上的百分比分值做出的平均绩点预测

在每位被试数据的范围内，计算出4个指标：他的预测的均值，他的预测的标准差，输入数值基础上做出的预测平均绩点的回归的斜率，以及它们之间的积矩相关系数。三组中的这些数值的均值均列示于表4—3。

表4—3 三组的单个预测数据的均值以及预先计划的第一组和第二组之间、第二组和第三组之间比较的结果

统计量	小 组				
	1. 百分比平均绩点	1 对 2	2. 心理集中度	2 对 3	3. 幽默感
预测的平均绩点的均值	2.27	ns	2.35	0.05	2.46
预测的标准差	0.91	ns	0.87	0.01	0.69
回归线的斜率	0.030	ns	0.029	0.01	0.022
r	0.97	ns	0.95	ns	0.94

在表格中,三组都产生有顺序的数据是显而易见的,这由输入变量与预测之间的高度相关证明[平均的相关系数,是通过将个人的数据转换为费希尔系数(Fisher's Z)而获得的]。预先计划的第一组和第二组、第二组和第三组之间比较的结果确证了在图4—3中观察到的模式。在用百分比平均绩点和心理集中度做出的预测之间没有显著差异。因此,人们在用对能力的衡量来预测对成就的衡量时并不回归,无论它多么不可靠。

另一方面,用幽默感做出的预测,是回归的,尽管不够充分。由回归曲线的比较而推定的平均绩点和幽默感之间的相关系数是0.70。另外,用幽默感做出的预测,显著地比用心理集中度做出的预测高。同时,由心理集中度做出的预测,倾向于比百分比平均绩点基础上的预测高。我们在很多研究中都获得这一发现。当在不完备信息基础上预测某人的学习成就时,被试表现得宽大仁慈(Guiford, 1954)。他们对有效性降低做出的反应是,提高预测出的表现水平。

只要输入变量与输出变量都被视为同一品质的外显,就可预期,预测基本上是不回归的。此类预测的例子,在一个真实生活场景——以色列军队军官选拔委员会中被观察到。在完成几天的检测和观察之后,那些参加评估小组的富有经验的军官,通常会在7分量表上对候选人进行评估。为了研究的目的,他们被额外要求对每位入选的候选人,在军官培训学校最后能取得的分数做出预测。在由大量不同的判断者做出评估的个案中,预测分数的分布与军官培训学校最后获得的实际分数的分布实际上是一致的,只有一个明显的例外:预测的失败比实际失败的频率要少。特别地,在最高的两个类别上的预测频率,与实际的频率完全一致。所有判断者都深知,研究证明他们的预测有效性仅仅是一般水准(介于0.20与0.40之间)。然而,他们的预测是非回归的。

三、方法论的考虑

代表性假设认为,预测与按相似度进行的评估或评定没有差别,尽管标准的统计理论认定预测应该比这些判断更不那么极端。因此,对代表性假设的检验,需要一种可以将预测与其他类别的判断相比较的实验设计。

本章报告的实验中,就使用了两种比较实验设计的变异形式。

在其中一个设计中,我们标记为 $A - XY$,不同组的被试在同一输入信息 (A) 的基础上判断两个变量 (X 和 Y)。比如在汤姆的例子中,两个不同的小组被给予同样的输入信息 (A),也就是,一份人格描述。其中一组按相似度 (X) 对结果排序,而另一组按可能性 (Y) 排序。相似地,在数值预测的几个研究中,不同组的被试被给予同样的信息 (A),比如说,对一位学生的五个形容词的描述,其中一组提供评估 (X),另一组则提供预测 (Y)。

在另一设计中,我们标记为 $AB - X$,两组不同被试在不同输入信息 (A 和 B) 的基础上判断同一输出变量 (X)。比如,在工程师/律师研究中,两个不同小组做同一判断 (X),即一个特定的人是工程师的可能性。他们得到此人的简要人格描述和有关工程师与律师的基率频率 (base-rate frequencies) 的不同信息 (A 和 B)。而在数值预测的情境中,不同组根据不同变量——百分比平均绩点 (A) 和心理集中度 (B)——的分数对平均绩点 (X) 做出预测。

这些比较实验设计表明,与标准模型相反,预测与相似度的评估或判断一样不回归,从而支持了代表性假设。我们同样可以问,与实际结果或者当输入信息与结果在同一量表上衡量时的输入信息相比,直觉预测是否回归。由于著名的趋中错误 (central-tendency error) (Johnson, 1972; Woodworth, 1938),即使当预测与转换后的形式一样不回归时,我们也预期他们将比实际结果更回归一些。在各种广泛的判断任务中,包括纯粹的将输入变量从一个量度转换到另一个量度上,被试都倾向于避免极端的回答且限制他们判断的变异性 (Stevens & Greenbaum, 1966)。由于这一反应偏差,相比输入信息或结果,判断将是回归的。本章所使用的设计通过两种判断的比较中和了这一效应,因为二者都受同一偏差的影响。

这里的系列研究,关注的是人们基于在实验前就可使用的信息来做预测的情形,这些信息是以刻板观念(如对某一工程师的刻板观念)和对变量之间关系的预期的形式出现的。它们没有提供结果的回馈,对每一被试要求的判断的次数也很少。与之相反,大部分以前的预测研究,都是处理

64

被试并不了解的对变量之间的函数或者统计关系的学习，这些研究通常有大量次数的试验和不同形式的结果反馈。在斯洛维奇等（Slovic et al., 1971）的研究中，有对部分这类文献的回顾。在有回馈的反复预测的研究中，被试通常选择那些使整个预测序列或模式高度代表实际结果的分布的结果。比如，在概率学习的研究中，被试构造出与结果序列的统计特征大体一致的预测序列。同样地，在数值预测中，被试相似地构造出散点图，也就是输入与结果的联合分布（参见 Gray, 1968）。为了这样做，被试使用一种混合策略：对任一给定输入信息，他们构造出一个有不同预测的分布。这些预测反映的事实是，任一输入信息，在不同的试验下产生不同的结果。显然，在这两种范式中，其预测规则是不同的，尽管都涉及了代表性启发式。在有回馈的范式中，被试构造出代表输入信息与结果关系的整体范式的回答序列。在本章考察的情形中，被试选择最能代表每一个体给他们的印象的预测。这两种方法导致对标准规则不同的背离：在回馈范式中用混合策略代表不确定性；但在本章涉及的范式下，则根据评估的预测而摒弃不确定性。

四、信心与效度错觉

如前面章节所示，人们通过选取最能代表输入信息的结果来预测。我们提出，人们对预测的信心程度反映出被选结果在多大程度上比其他结果更能代表输入信息。在输入信息（如分数描述）具有多重属性的数值预测的情境中，代表性的一个主要决定因素，是输入信息的一致性或者连贯性。输入信息越一致，预测分数将显得越具代表性，从而对预测越有信心。比如，在两门基础课程成绩都是 B 的基础上做出总体成绩是 B 的预测，相比基于获得 A 和 C 的成绩做出的同样的预测，更具信心。事实上，输入信息内部的变异性或者不一致，已被发现会降低预测的信心（Slovic, 1966）。

那种认为一致的描述（profile）相比不一致的将给予更多的可预见性的直觉，是不由自主的。然而，值得注意的是，这一观念与被普遍应用的多元预测模型（即正态线性模型）不相符合，后者认为预测的预期准确度独立于描述内部的变异性。

当判断者依据高度相关的分数来预测时，通常将碰上一致的描述；相对地，在相关性较低时，不一致的描述会更常见。因为信心随着一致性而增长，当输入变量高度相关时，信心通常也会比较高。然而，就给定效度的输入变量而言，它们与标准值（criterion）的多重相关性和它们之间的相关性是反向相关的。因此，一个两难的情况出现了：输入变量间的高度相关增强了信心却降低了效度。

为说明这一效应，我们要求被试基于两组能力测试预测平均绩点。被试被告知，其中一组测试（创造性思维和符号能力）是高度相关的，而另一组测试（思维灵活性和系统推理能力）是不相关的。他们看到的分数都符合这些预期（对其中一半的被试，测试贴上的相关和不相关的标签被调转过来）。被试被告知，“所有测试在预测大学成绩上，被发现是同样成功的。”在这种情况下，当然地，不相关组的测试将比相关组测试实现更高的预测准确度。但正如预期的，在总体的预测分数范围内，被试对依据相关组测试做出的预测更具信心。也就是，他们在一个较低的预测准确度的情境中更具信心。

在许多预测研究中，包括我们自己的，都观察到另一个发现：信心是表现的预测水平的J状函数（参见Johnson, 1972）。被试对杰出表现的预测，具有相当的信心，并且他们对预测彻底的失败比预测平庸的表现，也有更多的信心。正如我们前面所看到的，直觉预测通常都是不充分地回归。因此，预测与结果之间的偏差，在极端值时是最大的。J状信心函数，使被试在预测最有可能偏离常态时最具信心，成为必然。

上述分析表明，增强信心的因素，比如一致性和极端性，通常与预测准确度是负相关的。因而，人们在极易犯错误的判断中倾向于体验相当的信心，这一现象我们可称之为效度错觉（illusion of validity）。与其他感知和判断上的错误一样，即使在其虚幻的特性被意识到时，效度错觉通常仍会存留。比如，当采访一位候选人时，大多数人都会对我们对他将来的表现的预测具很高的自信，尽管我们都知道采访的极不可信是臭名昭著的。

五、有关回归的直觉

回归效应与我们所有人都息息相关。在我们的经验中，大部分杰出的

父亲都有稍稍令人失望的儿子，光彩照人的妻子有暗淡的丈夫，适应不良的开始适应，幸运儿最终遭遇飞来横祸。尽管有这样的经历，人们并没有获取回归的合理概念。首先，在许多注定会发生回归的情况下，他们没有预测回归。其次，正如任何一位统计学老师会证实的，对回归的合理概念是十分难获得的。最后，当人们观察到回归时，他们通常会对其臆造出动态的解释。

是什么让回归的概念违反直觉，且难以获得和应用呢？我们认为这一困难的一个主要来源在于，回归效应，通常违背了所预测的结果应该最大限度地代表输入信息的直觉。^[4]

67 为显示尽管与统计学有相当的接触但非回归的直觉仍然持续的事实，我们为—组心理学研究生被试提供了下面这个问题：

一个测试的问题。一位随机抽取的个体，在一个标准的 IQ 测试中获得 140 分。假定 IQ 分数，是“真实”分数和服从正态分布的随机测量误差的总和。请给出你对此人真实 IQ 的置信水平为 95% 的置信区间的上限和下限的最佳估计。也就是说，给出一个高估计值，这样你有 95% 的把握使真实 IQ 事实上低于这一估计，以及一个低的估计值，这样你有 95% 的把握使真实 IQ 高于此估计。

在这个问题中，被试被告知将观测到的分数视为“真实”分数和误差成分之和。由于观测到的分数显著地高于总体均值，误差成分更有可能是正值，因而此人在随后的测试中也将获得一个略低的分数。然而，大部分被试（108 位中有 73 位），报告了以 140 为中心的对称的置信区间，而未能传达出任何的回归预期。在剩下的 35 位被试中，其中 24 位报告了回归的置信区间，11 位报告反回归的区间。因此，大部分被试忽略了输入信息中的不可靠因素的效应，而视 140 为真实分数一样来预测。这种在预测时把输入信息看成是毫无错误的倾向，在本章中已被反复观察到。

回归现象的发生，有时也能被认识到，这可能是因为我们在自己的观

察中发现了回归效应，也可能是因为我们被明白无误地告知回归现象已发生。当意识到时，回归效应通常会被视为需要进行大量解释的系统的变化。事实上，在社会科学中，提供了大量的对回归效应的虚妄的解释。^[5]已有人运用动态原则来解释为何有些企业在某一时段生意兴隆但随后濒于垮台，为何面部表情的训练有益于在前测中分值很低的培训者，而有害于表现极佳的人。如果那些作者能够意识到给定两个有相同方差的变量，下面两个命题在逻辑上是等价的，则部分这样的解释就不会被提出：(1) Y 有关于 X 而回归；(2) Y 和 X 之间的相关系数小于 1 (unity)。因此，解释回归效应与解释为何相关系数小于 1 是等价的。

作为对认识到和合理表达回归效应是多么困难的最后一个说明，请考虑下面这个提供给研究生样本的问题。这个问题所描述的，是发生在作者之一身上的真实经历。

一个培训的问题。一所飞行学校的老师采纳了心理学家推荐的一致正面强化的政策。他们对每一个飞行操作的成功实施都进行口头强化。在使用这个方法一段时间后，老师们声称，与心理学原理相反，对正确实施复杂操作的高度赞赏通常会导致下一次尝试中表现的下降。心理学家应该如何应答？ 68

在飞行操作中，回归是不可避免的，因为飞行表现并不是十分的稳定，且成功操作之间的进步也是缓慢的。因此，飞行员在一次尝试中表现极佳，在下一次有可能弄糟，无论老师对最初的成功做出什么样的反应。富有经验的飞行教员事实上已发现回归现象，但归结为是正面强化的负面效果。这一真实的故事表现了我们人类生活中令人悲哀的一面。我们一般在他人表现良好时予以强化，在他人表现糟糕时予以惩罚。仅仅通过回归效应，他们就极有可能在惩罚后提高成绩，而在奖赏后成绩下滑。于是，我们被置于一个终生的计划中，我们常常为惩罚他人而受到奖励，为奖励他人而受到惩罚。

回答这一问题的研究生中，没有一位提出回归效应可能导致这一问

题。相反，他们提出口头强化对飞行员可能是无效的，或者会导致过分自信。有些学生甚至怀疑教员的印象的正确性，并探讨他们对情境的感知偏差的可能来源。这些被试都毫无疑问接触过彻底的对统计回归的处理。然而，当它不是用父亲和儿子的身高这一熟悉的方式表达时，他们没能认识到这一回归事件。显然，单凭统计学训练，并不能改变对不确定性的基本直觉。

本章原载《心理学评论》(*Psychological Review*, 1973, 80, 237-251)。
允许重印。

注释

[1] 除非另加说明，本章中实验的被试，均为通过俄勒冈大学的一份学生报纸招聘的有偿志愿者。数据是在小组的情境下收集的。

[2] 在计算这一相关系数时，我们把排名反过来。这样一个更高的判断可能性被赋予了更大的数值。

[3] 参见本书第 159 页。

[4] 每份有意义的行为片段，应高度代表行动者的人格。这种期望，可解释为什么普通人和心理学家一样总是对这些因素之间可忽略的相关表示惊奇，而这些因素，就是对诚实、风险承担、攻击行为和依从的明显可交换的测量 (Mischel, 1968)。

[5] 对研究中的回归谬误 (regression fallacies) 的富有洞见的讨论，可参见 Campbell (1969)、Wallis & Roberts (1956)。

第5章

对代表性的研究

巴-希勒尔 (Maya Bar-Hillel)

卡尼曼和特沃斯基提出，人们在判断一些不确定事件发生的概率时，通常会求助于启发式或者拇指法则，它们与真正决定事件发生的概率的变量是不完全相关的（即便有相关关系，也是不完全的）。其中一种启发式是代表性启发式，它被定义为对问题中的事件在多大程度上“与母体在核心特征上相似”或者“反映了产生该事件的过程的显著特征”的主观判断（Kahneman & Tversky, 1972b, p. 431, 3）。尽管在有些情况下，更可能发生的事件显得更具代表性，反之亦然。但将事件的代表性作为判断其发生概率的指示器，会将两种系统的错误引入判断之中。第一，它赋予那些影响事件的代表性但不影响其发生概率的变量以过高的影响力。第二，它会降低那些对事件发生概率起关键作用但与代表性无关的变量的重要性。

代表性这一概念时常被批评为过于模糊和隐晦，这可能是因为缺乏一个一般性的可操作定义。不

过，这并不等于认同了批评者常常暗示的一个结论，即独立于概率判断来评估代表性是不可能的。例如，在“汤姆”研究中，卡尼曼和特沃斯基就将代表性定义为某些个体（汤姆）与“某些研究生专业领域中的典型研究生”的相似度（1973，p. 238），并在汤姆就读这些研究领域的可能性之外，对这些领域进行排名。在其他研究中，按照代表性进行的独立排名能够被规避（sidestep），仅仅是因为读者可以如此轻易地通过思想实验（thought experiment）将其予以运用。

一、相似度和代表性

70 在代表性的概念被引进后不久，我进行了一项小研究，被试在其中根据一项先验标准（在本章后面会予以明确说明）和实验中的判断对代表性排序。我随后将这些排名与独立收集的的概率判断进行比较（Bar-Hillel, 1974）。此研究中的任务的优点在于，与之相关的概率是唯一的且定义完备，而且它展示方式简便，可以不费力地帮助被试进行感知相似度的判断。

本研究中的刺激是三条形图（triples of bar graphs），我们分别用 *L*、*M*、*R* 来表示左边、中间和右边的图。其中一例如图 5—1 所示。每一个图都由三个竖条组成，它们从左到右被涂成绿色、黄色和红色。每一竖条的长度都是 5 的倍数，而且每个图中的三个竖条的长度之和都是 100。

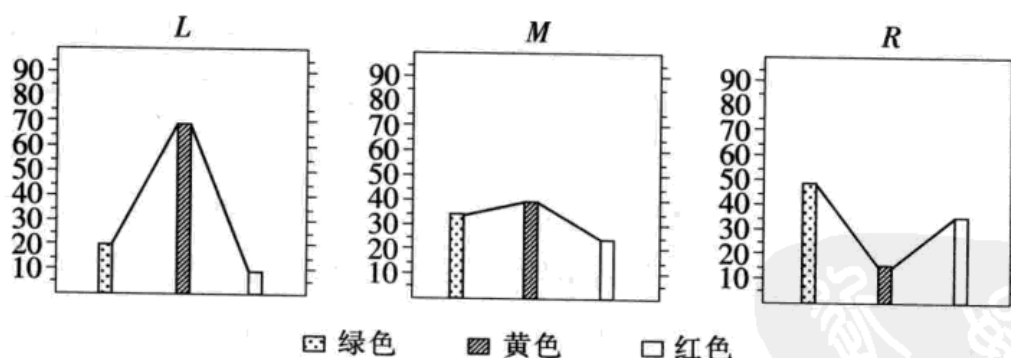


图 5—1 在相似度和代表性的研究中使用的特征刺激（characteristic stimulus）

资料来源：Bar-Hillel, 1974。

本研究中的所有三条形图（*L*、*M*、*R*）都具有下面一些特征：

- (1) 中间条形图中的每一竖条在长度上都是其相应的左边和右边图中

的竖条长度的中点（不过，如果左边和右边的竖条长度的均值不是5的倍数，将被近似为与其最接近的5的倍数）。比如，在图5—1中，在M中的绿色竖条的长度是35，它是20和50之间的中点，也就是L和R中相应竖条长度的中点。这对黄色和红色竖条同样成立（不过，请注意这些竖条的长度都是被近似处理过的：42.5近似为40，22.5近似为25）。

(2) M中竖条长度的排布顺序与L或者R一致，但不是都一致，因为L和R的排布顺序总是相反的。比如，在图5—1中M和L的排布顺序是一样的，都是红的最短而黄的最长，而R的不一样，它最短的竖条是黄色的。依次连接每个图中所有竖条顶端的细线使这一特点变得明显，不过X轴是定类变量。

(3) 假设M被表述为是描述一个三项式总体 (trinomial population)，而L和R是可能从中抽取的容量为20的样本，则排布顺序与M一致的样本将总是概率较小的样本。例如，在图5—1中，M描述的是参数为0.35、0.40和0.25的三项式总体，则抽取出样本R（10个绿色、3个黄色和7个红色珠子）而不是样本L（4个绿色、14个黄色和2个红色珠子）的几率将大约为8/5，因为

$$\frac{P(L|M)}{P(R|M)} = \frac{(20! / 4! 14! 2!)}{(20! / 10! 3! 7!)} \times \frac{(0.35)^4 (0.40)^{14} (0.25)^2}{(0.35)^{10} (0.40)^3 (0.25)^7} = 0.6127$$

因此，R是更可能的样本。

(4) 假设反过来，L和R被表述为总体，M是一个样本，则M从与其排布顺序一致的总体中抽取出来的可能性将总是较小。比如，在图5—1中，M描述的是一个7个绿色、8个黄色和5个红色珠子的样本，则M是从总体R（参数是0.50、0.15和0.35）而不是从总体L（参数是0.20、0.70和0.10）中抽取的几率大约为7/5，因为

$$\frac{P(M|L)}{P(M|R)} = \frac{(20! / 7! 8! 5!)}{(20! / 7! 8! 5!)} \times \frac{(0.20)^7 (0.70)^8 (0.10)^5}{(0.50)^7 (0.15)^8 (0.35)^5} = 0.7017$$

因此，R是更可能的总体。

在三组不同的说明条件下，28种此类三条形图被展示给三组不同被试。其中一组得知，每一条形图都是描述三项式分布，其中L和R描述的是两个总体，而M描述的是从其中一个总体中抽取的样本。该组25位被试

需判断 M 描述的样本是更可能从 L 还是从 R 中抽取的。另一组得知 M 描述的是三项式总体，而 L 和 R 是两个三项式样本。该组的 26 位被试需判断两个样本中哪个更有可能是从给定总体中抽取出来的。实验者首先向这些组呈现一个装满绿色、黄色和红色珠子的大罐子，然后向他们展示如何可能用合适的条形图来描述罐子（总体）的组成成分。实验者当场展示随机抽取 20 个珠子的样本，同时展示了它的条形图。被试随后被要求将 L 和 R 视为罐子总体的代表， M 代表 20 个珠子的样本，或者视 L 和 R 为样本， M 为总体，这要视被试所在组而定。

29 第三组，也就是最后一组，由 25 位被试组成，他们未得到对条形图的任何解释，仅仅被要求判断 L 和 R 中哪个与中间的 M 更相似。

按照每一个标准（1. 更可能获得样本 M ；2. 更可能从总体 M 中抽取；3. 与 M 更相似），根据大部分被试选取的图的情况，刺激被归类为 L 或者 M 。

本研究两个主要发现是：

（1）以被试给出的经验性的相似度排名与体现于特征 2 中的相似度的先验标准完全一致。换句话说，在所有 28 个刺激中，多于 80% 的被试判断， M 与这种图形更相似，它的各种颜色的竖条高度的排布与 M 是一致的。

（2）无论是作为样本还是总体，按照可能性对 L 和 R 的排名，与按照相似度对它们的排名是高度相关的。在 R 被判断为与 M 更相似的 17 个三条形图中，作为样本时，仅仅 2 个被判断为更不可能，作为总体时，仅仅 4 个被判断为更不可能。在 L 被判断为与 M 更相似的 11 个样本中，仅仅只有 1 个排名低于 R ，且它是作为样本的。因此，只有少于 13% 的刺激的相似度排名与可能性排名不一致。相似度判断和可能性判断的联合分类在图 5—2 中给出；在 A 中， L 和 R 是样本，在 B 中，它们是总体。

作为样本的可能性排名与作为总体的可能性排名，它们之间的相关系数 ϕ 是 0.75，这与每个可能性排名与相似度排名之间的相关系数 ϕ 基本上是一样高。这支持了在两个概率判断组中被试的排名基本上都是基于相似度的立场，其中对相似度的判断是由第三组进行的。

		(A)				(B)			
		L		R		L		R	
相似度	R	2	15			R	4	13	
	L	10	1			L	11	0	
样本可能性				总体可能性					
$\phi=0.79$				$\phi=0.75$					

图 5—2 被试按相似度对 L 组和 R 组的排名，分别与它们的样本可能性 (A) 或总体可能性 (B) 的联合分布。

资料来源：Bar-Hillel, 1974。

二、使样本显得具有代表性的特征是什么？

奥尔森 (Olson, 1976) 指出，尽管“在广泛的判断情境中，基于对代表性的评估做出判断的观念在实验和思辨上都有相当的支持”，但只有当我们能够确定“哪些因素使某些特定任务和问题的特征，成为与进行代表性判断相关的显著特征”时，这一观念才是完整的 (p. 608)。在某些情境下，比如在相似度和概率研究中，这些都被识别。举另一例子来说：决定从贝努利分布中抽取的各种容量的无序样本的样本代表性的线索 (cue)，看来是样本中“连续”的比例与相应的总体参数之间的差异 (Kahneman & Tversky, 1972b, 3)。我接下来要描述的研究 (Bar-Hillel, 1980b) 提示我们一种一般性的方法，它可识别出对从任何总体中抽取的样本的代表性起控制作用的那些特征。这体现在对从钟形分布 (如正态分布) 中抽取的 (三个观察的) 样本的代表性起决定作用的线索进行识别的尝试中。

基本原则很简单：如果我们假定对代表性和可能性的判断是由同样的基本因素决定的，则这些因素可以通过询问对合理选出的样本集的概率给出判断的人而获得。这一程序倒转了概率和代表性的角色。换句话说，概率判断不是被用来确认代表性，而是去推定代表性。然而，直接要求被试按照代表性对样本排序很有理由被认为是在乞求答案。他们也许会反驳：“那你说的代表性是什么意思呢？”我们建议的策略是，用在标准假设下的有标准答案的清晰明确的问题 (“下面的样本中，哪个更有可能是从这一总体中随机抽取的？”) 来取而代之。

研究人员给被试如下的问题：

美国大学男生的平均身高是 175 厘米。从大学注册办公室中随机抽取的三份档案，它们分别是 John、Mike 和 Bob 的。在三人身高的结果中，哪个更有可能？

John - 178 厘米		John - 177 厘米
Mike - 170 厘米	或	Mike - 177 厘米
Bob - 176 厘米		Bob - 177 厘米

实际给出的数字在不同的表格上都不一样，它们都能在表 5—1（上述例子是表中第一行）中的刺激列中看到。通过系统地改变被使用的数字，我们可以查明究竟是什么特征，又是以什么顺序在决定被试的回答。在对浮现出的图景给出总结前，我会依次给出一些解释性的评论来帮助读者理解该表。

（1）标号为 A 的样本始终是被试认为比与之配对的样本更有可能的样本。在原来的问卷表上，A 出现在左边和出现在右边的频率是一样的。被试中选择 A 和选择 B 的确切比例标于回答列中。

76 （2）从标准观点看更有可能的样本，可以在统计测量列中看到。当且仅当 $P(A)/P(B) > 1$ 时，也就是当且仅当

$$\frac{f(X_1) \cdot f(X_2) \cdot f(X_3)}{f(Y_1) \cdot f(Y_2) \cdot f(Y_3)} > 1 \text{ 时,}$$

这里 $A = [X_1, X_2, X_3]$, $B = [Y_1, Y_2, Y_3]$ ，样本 A 比样本 B 更有可能。这一系数是根据 $\mu = 175$ 厘米， $\sigma = 6$ 厘米的正态密度函数（normal density function）算出的。因为统计意义上的正确答案出现于 A 或者 B 列中的频率是一样的，被试与这些标号的关系显然仅仅是随机的。

也许表 5—1 的结果中最令人震惊的，是不存在单一的为被试对样本排序所依赖的线索。比如说，有时他们选择均值与总体均值比较接近的样本（如第 1 行），而有时选择相反的（如第 2 行）。有时他们选择有更多极端数据的样本（如第 13 行），而有时选择相反的（如第 14 行）。事实上，似乎有一大堆的线索是相关的——相同数值的个数，总体均值两侧的数据是否都得到了体现，样本均值和全距（range），且如图 5—3 显示的那样，以

表 5-1 对一对样本的可能性的统计衡量和判断

		样本 A 的刺激				样本 B 的刺激				回 答		$\frac{P(B)}{P(A)}$	正确 答案	KS 统计量	
		John		Mike		Bob		Mike		Bob				A	B
		John	Mike	Bob	John	Mike	Bob	A (%)	B (%)	n					
三个不同数据 V. S. 三个相同数据	1.	178	170	176	177	177	177	95	5	188	1.39	B	0.31	0.63	
	2.	181	165	177	175	175	175	93	7	29	6.42	B	0.31	0.50	
	3.	178	170	176	175	175	175	97	3	30	1.59	B	0.31	0.50	
	4.	178	180	176	175	175	175	85	15	26	1.59	B	0.57	0.50#	
	5.	178	180	176	177	177	177	80	20	25	1.39	B	0.57	0.63	
	6.	191	183	189	175	175	175	72	28	25	1300	B	0.91	0.50#	
三个不同数据 V. S. 两个相同数据	7.	178	170	176	176	170	176	72	28	25	1.12	B	0.31	0.43	
	8.	178	170	176	178	170	178	76	24	25	0.68	A	0.31	0.36	
	9.	176	174	175	178	170	178	57	43	29	0.56	A	0.43	0.36#	
较大全距者被优先考 虑	10.	178	170	176	176	174	175	62	38	29	1.58	B	0.31	0.43	
	11.	178	170	176	177	173	176	71	29	31	1.43	B	0.31	0.37	
	12.	178	170	176	178	172	177	54	46	28	1.14	B	0.31	0.31	
	13.	178	170	176	177	171	175	54	46	28	1.14	B	0.31	0.37	
较小全距者被优先考 虑	14.	178	170	176	179	169	177	54	46	26	0.75	A	0.31	0.30#	
	15.	178	170	176	181	165	177	77	23	30	0.23	A	0.31	0.30#	
	16.	178	170	176	180	170	175	64	36	25	0.81	A	0.31	0.20#	
	17.	178	170	176	181	169	175	59	41	27	0.60	A	0.31	0.17#	

续前表

	样本 A 的刺激			样本 B 的刺激			回 答		$\frac{P(B)}{P(A)}$	正确 答案	KS 统计量		
	John	Mike	Bob	John	Mike	Bob	A(%)	B(%)			A	B	
双侧样本 V. S. 单侧 样本	18.	178	170	176	172	170	174	78	22	1.00	皆可	0.31	0.57
	19.	178	170	176	178	180	176	64	36	1.00	皆可	0.31	0.57
	20.	178	170	176	184	176	182	76	24	0.26	A	0.31	0.57
	21.	181	165	177	178	180	176	63	37	4.29	B	0.57	0.57
	22.	176	174	175	184	176	182	73	27	0.87	A	0.37	0.57
	23.	177	171	175	178	180	176	93	7	0.17	A	0.43	0.57
	24.	176	174	175	178	180	176	79	21	1.58	B	0.43	0.57
	25.	183	175	177	184	176	182	87	13	0.42	A	0.50	0.57
两样本均为单侧	26.	185	175	177	184	176	182	85	15	0.68	A	0.50	0.57
	27.	178	180	176	184	176	182	91	9	0.26	A	0.57	0.57
	28.	178	180	176	181	179	183	81	19	0.32	A	0.57	0.75

资料来源: Bar-Hillel (1980b)。



某种序列的形式依次被考虑到。这更强化这一观点，即是什么决定样本被判定的可能性本身就构成了一个概念。如果被试仅仅使用单一的线索，那就没有必要称其为代表性。事实是，被试的判断看来是基于一个复杂的对样本格式塔结构性的考虑，它促成了对被试的启发式的术语化——还有概念化——按照代表性进行判断。

我们能将图 5—3 中的流程图视为不仅仅是对表 5—1 中的结果的方便简洁的总结吗？在多大程度上，我们可以将其视为被试在上述任务中的真实推理的过程模型？一方面，显然，并不是模型中提到的所有特征，都可以推广到其他模型、其他的样本容量和其他的分布。比如说，我不敢说任

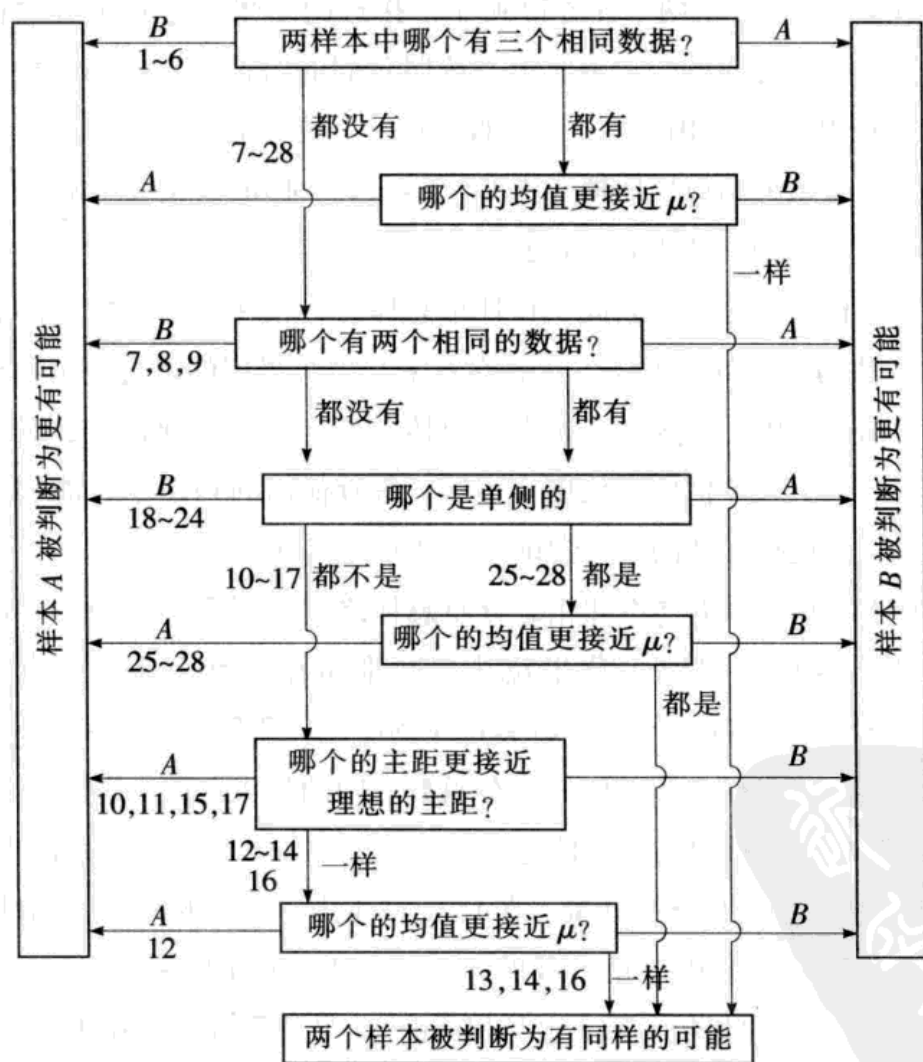


图 5—3 对按照可能性对样本排序时使用到的线索和过程进行汇总的流程图

何存在变异的样本总是被认为比没有变异的样本更有可能出现, 尽管这对本实验中使用的样本是成立的。因此, 这一序列模型仅仅是对人们真实的认知策略的一种近似模拟。尽管如此, 它具有可信这一诱人的心理学上的优势, 因为它仅仅纳入了少量的需要计算的简单线索。

这里, 一个有趣的发现给出了暗示, 并将前面描述的研究和现在的研究联系在一起。在条形图研究中, 两种概率排名被用来与相似度排名进行比较: (1) 对两个样本中哪个更有可能是从某一总体中抽取出来的进行的排名, 我们将其标为 $P(s/p)$; (2) 对两个总体中哪个更有可能是某一样本来源进行的排名, 我们将其标记为 $P(p/s)$ (有时也被称为逆概率 inverse probabilities)。这两个任务得到非常接近的结果, 这暗示着它们是用非常相关的方式完成的——事实上, 按照我们的假设, 它们都是根据代表性 (这里是视觉上的相似度) 完成的。有一个统计概念与心理学上的代表性的概念具有某种相似性即拟合优度 (goodness of fit)。对拟合优度的测量通常是在假设检验中被用到, 在其中, 没有任何先验概率, 而必须找到产生几个已知样本的总体。因此, 在评估总体时, 相比于已知样本的总体的条件概率 $P(p/s)$ 这一概念, 代表性看来与拟合优度的概念更为贴近。在 27 身高分布的研究中, 作为样本来源的总体是给定的。因此, 被试既不必评估逆概率, 也不必评估拟合优度。不过, 看看他们的判断是否能够被设想为对拟合优度进行评估的直觉尝试, 也是非常有意思的。

对样本容量为 3 的样本和正态分布总体之间的拟合优度的一种可能衡量, 是单样本双侧的柯尔莫果洛夫—斯米尔诺夫统计量 [Kolmogorov-Smirnov (KS) statistic] (参见 Siegel, 1956, pp. 47-52)。这一统计量被表述为, 值 (它是指样本和总体的分布函数之间的最大差距) 越大, 此样本是从该总体中随机抽取的可能性越小。表 5—1 中的最后两列列出的就是每一个被研究的样本的 KS 统计量。在 18 对样本中, 按照 KS 统计量对样本的排序与大多数被试的排序是一致的; 有三对 KS 统计量相同 (其中一对, 第 12 行中, 被试的判断实际上也是相同的), 另外 7 对中 (在表 5—1 中用 #号做了标示) 的排序是不一致的。

将 KS 统计量排序和被试的排序之间的一致程度, 与标准排序和被试的

排序之间的一致程度进行比较是没有多大意义的（对于后者，28个刺激中有13个是一致的），因为后一数据仅仅是被使用的刺激的副产品，而这些刺激不是从所有可能的刺激中随机抽取的。考虑特征层级模型（hierarchical-features model）与拟合优度模型在什么地方存在差异，会更具有建设性的意义。KS排序相同而被试的排序不同的情况（第21、27行，还有程度较轻的12行）可以归结为是KS值造成的，它是用6厘米的标准差计算出来的。这导致全距为12厘米的样本（181、169、175）（样本B，第17行）有最小的KS值，而被试判断中的“理想”（也就是最具代表性的）的全距要稍小——8到10厘米。由于在目前的数据基础上，我们没有办法确定人们主观的身高分布中的标准差，因此我们难以获知是什么导致了这一差异。

第二种差异对第4、6、9和14~17行能够做出解释，它的产生是源于KS模型而不是特征层级模型，且它提供了某种补充。换句话说，当某一样本存在一个重要的“瑕疵”（比如由三个相同数据组成，或者仅代表了钟形总体的一侧）时，被试将判断它比任何一个没有那种瑕疵的样本更不具代表性，但KS模型会考虑特征层级模型之外的因素。正如我前面所说的，我们建议的过程模型的层级性质应该被视为一种近似，因此这种差异更多的是程度上的差异而不是性质上的不同。

其他的拟合优度的测量由统计学家进行。他们未达成一致的事实表明，他们声称要予以测量的属性是不清晰的和没有唯一定义的。尽管训练有素的统计学家比起一般被试来有更复杂的直觉，但他们都极有理由去获取同样的潜在概念。迄今为止，即使是统计学家在这个方向上的努力也依然是 29 不尽如人意的。

三、样本容量在样本评估中的角色

某些样本特征与一些总体参数相对应，这些特征被称为样本统计量。人们普遍地将代表性启发式运用其上，其中代表性等同于样本统计量和相应的总体参数之间的相似度或者说贴近度（closeness）。我称之为代表性的首要含义（primary sense of representativeness）。但样本同样可以借助于产生该样本的程序来描述。比如说，样本的抽取有放回和没有放回两种方式；

从已知总体或未知总体中抽取等等。代表性的概念，也可能被运用其上。因此，人们可能（已有逸事证据支持）判断一个随机抽取的样本比一个分层样本更不具代表性，或者一个大样本比小样本更具代表性（Bar-Hillel, 1980b）。我称之为代表性的次要含义（secondary sense of representativeness）。因为它涉及的是抽样方法而不是样本统计量，代表性的次要含义反映的是对代表性的首要含义的预期。换句话说，判断一个大样本比一个小样本更具代表性，是在被告知样本的确切属性之前，就判断它的“显著特征”或者“基本属性”能更好地反映总体的相应特征。

我在这最后一节要描述的实验问题，研究的就是代表性的次要含义，它是被运用于样本容量的特征上的。

问题 1：两家民意调查公司进行一项调查，以估计在某一全民公决中投票者将投赞成票的比例。公司 A 调查的是一个 400 人的样本，公司 B 调查的是 1 000 人的样本。你将更有信心接受哪家公司的估计？
公司 A 的 _____ 公司 B 的 _____ 基本一样 _____

研究者将这一问题及其两种变异形式的问题（它们分别加上被调查社区的人口数为 8 000 和 50 000 的信息）呈现给 72 位被试。多于 80% 的被试，对大样本更有信心，相比之下只有 4% 的人倾向于小样本。

如果从中我们仅仅总结出，样本越大，则它将被判断为越具有代表性（即被预期为更可靠或更准确），这一结果将过于平常而不值得予以报告。并不是样本本身决定了样本的代表性，而是由某种类似于样本和总体容量之比的东西决定的。当被考虑的样本是从同一总体中抽取的时，这两者是线性相关的，因而被试给出同样的排序。然而，当样本容量和总体容量都变化时，差别就变得明显。请思考下面这个问题：

问题 2：两家民意测验公司进行调查，以估计在它们分别所属的城市中投票者将在某一全民公决中投赞成票的比例。

公司 A 是在一个有 100 万投票者的城市中进行的

公司 B 是在一个有 5 万投票者的城市中进行的

两家公司都是从每 1 000 人中抽取 1 人。

你将更有信心去接受哪家公司的估计? _____

这里, 尽管公司 A 的样本有 1 000 人而公司 B 只有 50 人, 被试中对大样本更有信心的比例却降到了 50%, 而有 29% 的被试表示对两者有同样的信心。在另一组被试中, 他们被告知两家公司都是抽取 1 000 人 (而不是每 1 000 人中抽取 1 人), 9% 的人表示对两样本有同样的信心, 而相比之下有 62% 的人对在小城市中抽取的样本更有信心。被试给出的解释性评论, 也确认了他们除考虑绝对样本容量以外, 也考虑比例性的样本容量的想法。

有时, 容量和总体之比对评估一个样本确实很重要——比如说, 当采用无放回抽样时, 还有, 尤其当相当大的比例被抽取的时候。然而, 当抽样是可放回的 (总体容量是无限的), 或者一个较大总体中仅仅有小部分被抽取到时 (这实际上使总体容量为无限), 在绝对容量之外考虑相对容量则是苍白无力的。

问题 3: 在你的面前放着两个被掩盖住的瓶子。它们里面都放着混合的红色和绿色的珠子。两个瓶子中珠子的数目不一样: 小瓶子中有 10 个珠子, 而大瓶子中有 100 个珠子。不过, 红色和绿色珠子的比例在两个瓶子中都是一样的。抽样过程如下: 你从瓶子中随机抽出一个珠子, 将它的颜色记下, 然后将其放回。将珠子混合, 再次随机抽取, 然后再记下它的颜色。在小瓶中抽取次数为 9 次, 在大瓶中抽取次数为 15 次。你认为在哪种情况中, 你猜中多数珠子的颜色的机会更大?

因为问题 3 中描述的抽样程序是有放回的, 因此, 从标准视角来看, 这两个瓶子中珠子的数目是完全不重要的。被试的选择应该压倒性地倾向于有 15 个珠子的大样本。但相反, 110 位被试中有 72 位选择了有 9 个珠子的小样本。这只能由后者中的样本容量和总体容量之比为 90% 而前者为

15% 来解释。由巴-希勒尔 (Bar-Hillel, 1979) 报告的其他结果, 也支持了同样的结论。

87 因此, 如果在样本统计量中谁是“基本属性”这一角色未知时, 样本容量就替代了它的位置。人们都很清楚地意识到样本容量是有优势的样本属性, 那为何这一意识没能上升为代表性的首要含义呢? 换句话说, 既然人们相信大样本相比小样本更可能成为总体的好的代表, 那为何他们没能判断出某些不正常的结果在大样本中相比小样本出现的可能性更小呢?

判断特定样本结果在大样本中出现的可能性较小, 可能是取决于它被感觉为不具代表性。从标准视角来看, 只有当某一样本的均值与总体均值相同时, 它才是“准确”的。其他样本都是不准确的, 虽然某些样本比其他样本更不准确。从心理学角度来看, 即使样本结果有些偏离总体参数, 它们也可能被判断为具有代表性。也许 60% 并不被感觉为不具代表性, 因此它没有启动被试的代表性的次要含义。

为检验这一可能性, 我给被试几个有所变动的卡尼曼和特沃斯基 (1972b, 3) 的孕妇产房 (maternity-ward) 的问题:

某镇有两家医院。大医院中每天约有 45 个婴儿出生, 小医院每天有 15 个婴儿出生。众所周知, 在所有婴儿中大约有 50% 是男孩。不过, 男婴的确切比例每天都不一样。有时它高于 50%, 有时又低于 50%。在一年的期间内, 两家医院都记下出生婴儿中超过/少于 60% 的是男婴的天数。你认为哪家医院会记下更多的此类天数? (1972b, p. 443)

在对不同组给出的不同问题中, 男婴的比例从 60% 到 70%、80% 和 100% 不等。表 5—2 显示的就是这一变化对被试回答的分布的效应。

第二列复制了卡尼曼和特沃斯基版本中的问题和答案。但请注意, 一旦男婴的比例超过 70%, 最常见的回答 (modal response) 就变成了“小医院”, 而这是正确的答案。当被告知所有的婴儿都是男婴时, “小医院”甚至是多数人的回答。另一组问题给出的比例少于 60%、70% 或者 80%。对

于这些问题，“一样”从始至终都是最常见的回答。

这些结果表明，一旦某样本结果被认为不具代表性，样本容量更有可能在人们对样本的评估中扮演一个适当的角色，这可能是因为较小的样本容量相比较大的样本容量，与不具代表性的结果更一致。然而，如果给定的样本结果没有激发起“不具代表性”的标签，人们大体上都忽略了样本容量。高于70%的比例看来被编码为不具代表性的，而包括50%在内的其他比例则没有。

表 5—2 对有所变动的孕妇产房问题做出不同回答的被试比例

	超过 60% ¹	超过 60% ²	超过 70% ²	超过 80% ²	100% ³
大医院	24%	20%	25%	26%	19%
小医院	20%	20%	43%	42%	54%
一样	56%	60%	32%	32%	27%
N	50	40	28	27	41
	少于 60% ¹	少于 70% ²	少于 80% ²		
大医院	20%	31%	29%		
小医院	24%	28%	25%		
一样	56%	41%	46%		
N	45	29	28		

1. 源自 Kahneman 和 Tversky (1972b, 3)。

2. Bar-Hillel 的未发表数据。

3. 源自 Bar-Hillel (1979)。这里，大医院和小医院每天出生的婴儿数分别为 15 和 5。

四、讨论

本章呈现的是对主观概率的判断的三项研究的总结，它们与代表性概念是相关的。每个研究都是用不同的方式做的。

第一项研究是对相似度判断和可能性判断具有相关性这一主题的一个直接的呈现。与许多其他检验这个同样假设的任务不同（参见 Kahneman & Tversky, 1972b, 3; 1973, 4），被试既不会因为考虑其他因素（或者，至少其他显而易见的或者说简单的因素）而被挑剔，也不会因为没有审慎思考错误的或者不相关的因素而被指责。事实上，在缺乏理应进行的概率比较的情况下，被试除了使用他们实际采纳了的策略外，别无他法。这在某种程度上削弱了这项研究的效力。不过，与其他研究一起，它指出在对

计算或概念上的复杂事件做出判断时，对相似度的考虑——即便是如本研究中所引发的对简单的视觉相似度的考虑——是多么的不可抗拒。

在样本特征 (sample-features) 研究之前，对代表性的检验，通常都集中于单一参数的样本或者单一维度的线索。这使迅捷地按相似度对刺激的先验排名成为可能，因而使得通过实验获取的独立排名变得没有必要。不过，即使是一些“显而易见”的线索随后也遭到质疑（参见 Olson, 1976），但当刺激变得复杂时，它们之间的相似关系就变得隐晦，这是因为许多线索都会影响总体的判断。本研究就意图通过挖掘已建立的概率和代表性之间的关系，以求对后者如何按照多重线索被样本特征决定给予一定的阐释。

通常，查明两个刺激有多么相似的最好的办法，是直接考察相似度。但要查明样本 S 多大程度上代表了总体 P 时，这一方法就有点问题了。因此，询问样本 S 多大可能是从总体 P 中抽取出来的，以求间接地对它做出评估，这是有裨益的。本章中的第二个研究在描述了什么样本特征使样本在某一特定情境下显得具有代表性时，显示了这一方法的可能性。

最后，第三个研究指出了样本的代表性和样本结果的代表性之间一个重要的区别。代表性的首要含义，是应用于样本结果的判断。换句话说，为了评估在这个意义上某一样本对其总体具有多大的代表性，我们必须知道两者的基本特征。因此，参照一般的新生婴儿的性别分布（男婴的比例已知），对有 9 个男婴的由 15 个婴儿组成的样本的代表性做出评估，是有意义的。如果样本或者总体中的男婴的比例未知时，对该样本的代表性做出评估是没有意义的。但是我们还有代表性的次要含义，它将使第二种评估也成为有意义的。就其次要含义而言，如果样本被判断为在首要的含义上具有代表性，则它更具有代表性。换句话说，无论总体均值是多少，在样本均值未知之前，一个 45 名婴儿组成的样本（在次要含义上）比 15 名婴儿组成的样本更具代表性，因为一旦得知它的均值，它更有可能与总体相似。因此，当然地，45 人的样本能被证实有 60% 的男婴，而相比之下小样本只有 53% 的男婴。但如果这两层含义区分很清楚，这不会造成任何

混淆。

第三个研究表明，当人们做出次要含义上的代表性判断时，即便他们未能意识到首要和标准的代表性判断必然包含的含义，他们也会对样本容量做出反应。



第6章

有关代表性的判断和依据代表性的判断

特沃斯基和卡尼曼 (Amos Tversky and Daniel Kahneman)

84 几年前，我们分析了不确定性状况下的判断，它将主观概率和关于代表性的预期及印象的直觉预测联系在一起。有两个不同的假设组成了这个概念：(1) 人们预期样本与它的母体高度相似，而且能体现抽样过程的随机性 (Tversky & Kahneman, 1971, 2; 1974, 1); (2) 代表性作为一种启发式，人们常常依赖它进行判断和预测 (Kahneman & Tversky, 1972b, 3; 1973, 4)。

第一个假设，用来解释机会是自我校正的这一普遍信念，以及对在小样本中观测到的结果的稳定性持有的夸大信念即赌徒谬误，还有对随机性判断中的相关偏差。我们指出，在普通人有关几率的概念中，包含了对小数律的信念，根据小数律，即使是小样本，也能够高度代表它们的母体 (Tversky & Kahneman, 1971, 2)。一个类似的假设也能解释下面这一普遍的倾向，即夸大人格特质的一致性和可预测性 (Misch-

el, 1979), 以及高估类似的变量 (见第 15 章) 和行为之间的相关关系 (Shweder & D'Andrade, 1980)。人们似乎相信一个全息摄影式的 (hologram-like) 人格模型, 在这个模型中, 任何行为片段, 都被认为代表了行动者的真实品格 (Kahneman & Tversky, 1973, 4)。

人们预期样本会高度代表它们的母体, 这一假设在概念上独立于第二个假设, 即人们通常使用代表性启发式做出预测和判断概率。也就是说, 人们评估不确定事件或样本的概率, 是按照“(1) 样本在多大程度上与母体的核心特征相似; (2) 体现了产生它的过程的显著特征”进行的 (Kahneman & Tversky, 1972b, p. 431, 3)。我们在多种情境下对该假设进行研究, 这包括直觉统计判断和对职业选择的预测 (Kahneman & Tversky, 1972b, 3; 1973, 4)。

这两个代表性的假设, 被用来解释许多观察到的现象, 诸如共识信息的相对无效性, 以及在解释投射测验时对相似度的使用等 (Nisbett & Ross, 1980)。这些假设也为一项回报甚丰的、在直觉判断中对标准规则的严重违背的研究指明了方向。这些研究大部分都关注的是依据代表性的判断, 也就是, 代表性在预测和推论中的角色。相对而言, 较少的研究, 指向关于代表性的判断, 也就是, 在随机抽样情境之外的这一关系的本质和决定要素 (Bar-Hillel, 1980b)。本章的第一部分, 将关注代表性关系的本质, 以及在什么条件下代表性的概念被用来解释直觉预测和概率判断。在本章的第二部分, 我们将展示, 在对复合事件的可能性判断中, 代表性逻辑和概率逻辑之间的差别。

一、代表性关系

代表性, 是某一过程或模型 M 和与该模型相关的某一实例或事件 X 之间的关系。代表性和相似度一样, 可以通过经验研究来评估, 比如说, 我们可以要求人们对两个事件 X_1 和 X_2 哪个更能代表某一模型 M 作出判断, 或者是判断事件 X 更能代表 M_1 还是 M_2 。所讨论的模型, 可以是一个人、一块理想硬币, 也可以是世界经济, 而与此相应的结果, 则是一个评价、一个由正面和背面组成的序列, 或者是黄金的当前价格。

代表性是一个单向关系：我们说，一个样本对某一总体更具或更不具代表性，或者是，这一行为更能代表某个人。我们通常不会说，这个总体对该样本具代表性，或者是，这个人能够代表这一行为。但是，在有的问题中，我们可以逆转模型和结果的角色。比如说，我们可以评估一个人是否代表了图书管理员的刻板印象，也可以评估图书管理员的职业是否能代表这个人。

我们区分出代表性概念通常被激发的四种基本情形：

(1) M 是一类别， X 是定义在该类别中的一个变量的数值。我们是在这个意义上提及（更具或更不具）代表性的大学教授的收入，或者是某一文化中的结婚年龄。自然的、最具代表性的数值，应该要接近于类别 M 的相关变量分布的均值、中值和众数。在这一情况下，代表性关系，主要是由判断者所了解的相关变量的频率分布决定的。

(2) M 是一类别， X 是这一类别中的一个实例。大多数读者都可能会同意，厄普代克（John Updike）是比梅勒（Norman Mailer）更具代表性的美国作家。显然，这一判断不是基于频率，它反映的是这些作家的风格、主题和思想谁更接近当代美国文学的主流。相似的考虑，也决定了那些本身就是类别而非个体的实例的代表性。比如说，知更鸟被认为是比鸡更为典型的鸟类，尽管它比较少见（Rosch, 1978; Smith, Shoben & Rips, 1974）。因此，如果一个实例拥有该类别的成员都共有的特征，而没有太多其他类别成员不具备的独特特征，则它可以成为该类别的代表性的实例（Rosch, 1975; Tversky, 1977）。

当前对概念形成（Rosch & Mervis, 1975; Mervis & Rosch, 1981）、语义记忆（Bransford & Franks, 1971）和类型识别（Posner & Keele, 1968）的研究已说明，一个类别中最具代表性或原型性的元素，比起更常见但缺乏代表性的元素，要更易被学习、记忆和识别。而且，人们经常犯“再认”一个实际从未显示过的原型刺激的错误。因此，代表性对频率判断和识别记忆都会带来偏差。

大家也许会注意到，元素可以有两种方式高度代表它所属的类别。代表性的这两层含义，与典型性关系和原型性关系非常一致。如果一个元

素是典型的或者普遍的,则它对所属类别具高度的代表性;如果它是体现了所属类别的核心特征的理想型,则它也是具代表性的。比如说,纽约是美国城市的原型,但辛辛纳提更可能被选作典型的城市。相似地,我们观念中典型的法国女性和法国女性的原型也可能是非常不一样的。前者可能是一位年轻优雅的巴黎女郎,而后者更可能是外省的肥胖中年妇女。

(3) M 是一类别,而 X 是 M 的子集。大部分人可能会同意,佛罗里达州的人相比伊利诺伊州的人更不能代表美国人,而天文学的学生比心理学的学生更不能代表整个的学生群体。代表性的标准对一个子集和单个的实例是不一样的,这是因为,一个实例仅仅能体现属性的趋中度 (central tendency),而子集还可以体现全距和变异性。一个身高、体重、年龄、收入和美国人的平均值一致的人,显然能够代表这一总体。由 100 个这样的人组成的群体,却无法体现总体属性的变异性。 87

如果类别 M 由这样一些不同的聚类 (cluster) 构成,在这些聚类中,每一聚类内的变异性相比聚类间的变异性都十分小,我们将倾向于把每一聚类视为类别的一个实例而不是子集。因此,尽管知更鸟是鸟类的一个子集,但我们自然地将“知更鸟”视为一种鸟或者说“鸟”这一类别中的一个实例。更一般地,我们可以将 (2) 视为 (3) 的一个特例,这里子集 X 只有一个成员。相同地, (1) 可以被视为 (2) 的单维版本。这三种代表性因为 X 的复杂性而区分开来,其中 (1) 是单个元素、单个属性的情况, (2) 是单个元素、多属性的情况,而 (3) 是多元素的情况——有一个或多个属性。

子集代表性中的一个特别重要的例子,是这样的情形,即 X 是从特定总体中随机抽取的样本。人们预期,一个随机样本要体现抽取过程的随机性,而不仅仅是体现它所从中抽取的总体的核心特征。譬如,当 100 个人被随机抽取时,由 53 位男性和 47 位女性构成的样本比由 50 位男性和 50 位女性构成的样本更具代表性,这是因为前者体现了随机抽样的不规则性,而后者没有 (Kahneman & Tversky, 1972b, 3)。代表性样本的统计概念,有些学者已讨论过 (Kruskal & Mosteller, 1979a; 1979b)。

(4) M 是 (原因) 系统, X 是一个 (可能) 结果。这一情形区别于前

面几个的地方在于， M 不再是一类物体或实例，而是产生不同结果的系统。比如说， M 可以是美国经济而 X 是通货膨胀率，或者 M 是一个人而 X 是 M 采取的一个行为，比如，离婚、自杀和选择职业等。这里， X 要代表 M ，可能是因为它通常和 M 联系在一起（比如，肺炎一般都伴随着高烧），也可能是因为人们相信——无论正确的还是错误的——是 M 造成了 X （比如，死刑防止了绑架的发生）。在条件概率的判断中，因果图式（causal schemas）的进入，已在特沃斯基和卡尼曼（1980，8）一文中有所说明和讨论。

综上所述，代表性关系可以在下面四类情形中被定义：（1）一个数值和一个分布，（2）一个实例和一个类别，（3）一个样本和一个总体，（4）一个结果和一个原因。在所有这四种情形中，代表性都表达的是 X 和 M 的一致程度，但其决定要素在四种情形中不都是一样的。在情形（1）中，代表性是由感知的相对频率或统计关联（statistical association）所操纵。在情形（2）和（3）中，代表性主要由相似度决定，比如说，一个实例与其他实例的相似度，或样本统计量与相应的总体参数的相似度。最后，在情形（4）中，代表性主要受（正确的或错误的）因果信念控制。

代表性和概率

88 使用代表性来解释概率判断和直觉预测是基于下面的假设：

（1）“ X 对 M （非常/……/全然不）具有代表性”这一关系，可以由判断者进行有意义的评估。

（2）这些评估不是基于概率或者频率的印象，而是由代表性来解释的。

（3）代表性有其自己的逻辑，它系统地区别于概率逻辑。

当这些条件满足时，来检验概率判断是否受代表性的评估的调节，是有重要意义的。

对不确定事件发生概率的评估或者未知数值的预测，是一个复杂的过程，它包括对问题的解释、对相关信息的搜索和合理回答的选择。它可以和一个灵活的计算机程序的运行相比较，这种计算机程序包容了多种潜在有用的子程序。依照这个类比，代表性启发式是其中的一个程序，它可以

被用来提取、解释和评估信息。当然，这一启发式的使用，并不排除其他程序的使用，这也正如形象性（imagery）启发式在回忆中的使用，并不排除其他策略的使用。但是，对启发式的依赖会导致特有的偏差。当形象性启发式被用来回忆在某一会议中在场的人时，与会者中可以被清楚看到，相比那些不能清楚看到的更能被记住。同样地，用代表性启发式评估主观概率会造成对某些概率的高估和对其他概率的低估。

早先的研究导致极端的假设，即认为某些概率判断完全是基于代表性的。比如说，主观样本分布基本独立于样本容量这一观察（Kahneman & Tversky, 1972b, 3）表明，人们根据样本统计量和相应的总体参数的相似度来评估样本的概率。但是，大部分可资使用的数据都支持一个更温和的假设，即直觉预测和概率判断，对代表性是高度敏感的，但不是完全受其控制。因此，主观概率受到那些（标准上）不相关但影响代表性的因素的强烈影响，而对（标准上）相关但不影响代表性的因素相对不敏感。代表性偏差的大小以及样本容量、可靠性和基率等变量的影响，取决于问题的性质、实验设计的特点、回答者的严谨程度和在场的暗示性的线索或者其他需求特征（demand characteristics）。我们将在第34章对这些因素在有关判断的研究中的角色进行讨论。

如果对代表性的依赖会造成系统的错误，那为何人们还使用这一关系作为预测和判断的基础？对这一问题的回答有三点。第一，代表性显得容易获取和易于评估。当代对范畴化的研究（Mervis & Rosch, 1981; Rosch, 1978）表明，概念性的知识，通常是用原型或者代表性的实例的方式进行组织和处理的。因而，我们会发现，评估某一实例对所属类别的代表性比评估它的条件概率更容易。第二，更可能发生的事件，通常比更不可能发生的事件更具代表性。比如说，一个与总体相似的样本，通常比同样样本容量的高度非典型的样本更可能发生。第三，样本通常对母体具有代表性，这一信念致使人们高估频率和代表性之间或者统计关联和隐含的相似度之间的相关关系。因此，代表性之所以被使用，是因为：（1）它易获取，（2）它通常与概率相关联，以及（3）人们高估了这一相关关系。但是，对代表性的依赖会导致可预见的判断错误，因为代表性有它自己的逻辑，

而这一逻辑与概率逻辑不同。

代表性和概率间的反差，在下面的情况中最为显著：（1）当证据是错误的，或者（2）当目标事件是高度特殊的。在第一种情况中，对我们的模型具有高度代表性的结果，也许是不可能发生的——如果我们观念中的模型，是建立在不具有有效性的证据上。比如说，请思考在一次面试中给人留下突出印象的候选人完成一项非常困难的任務的可能性。因为基于面试的印象，是臭名昭著地不可靠，且工作成功与否是由多种因素控制的，而这些因素，并不是从一次简单的交谈中就可以预测到的。因此，尽管成功非常能体现候选人给我们的印象，但成功还是很有可能不会发生。

就第二种情况而言，代表性的结果也许会非常不可能发生，因为它太特殊或者说太具体。一般而言，一个事件不可能发生，或者是因为它是非典型的，或者因为它太特殊。低于 135 磅的体重对中年男子来说是非典型的；157.625 磅的体重是典型的，但过于特殊。事实上，尽管前者更有可能，但后者更能代表中年男子。如本例所示，特殊性的增加，通常不会导致代表性的减少。因此，具有不同特殊性的事件之间的比较，通常会导致代表性和概率间的冲突。比如说，随机抽取的四张牌的样本由红心 K、黑桃 A、方块 9 和梅花 4 组成，它比四张一样的牌组成的样本显得更具代表性，尽管后者往往更可能发生。因此，概率判断中的代表性偏差，在对具

90 代表性但高度特殊的事件的评估中最为显著。这类偏差，在我们下一节要描述的对复合事件的概率判断的研究中有所展示。

二、对复合事件的评估

概率和代表性之间的最大差异，存在于对复合事件（compound events）的评估中。假设我们被提供某人的一些信息（比如人格描述），然后我们推测此人拥有的各种属性或者属性的组合，比如说职业、业余爱好或政治倾向等。概率的一个基本定律是，详细的说明（specification）只会降低概率。某一特定的人既是民主党人又是艺术家的概率，必然小于此人是艺术家的概率。这一条件不仅在标准的概率计算中成立，也在非标准的模型中成立（如 Shafer, 1976; Zadeh, 1978）。

但是, $P(A \& B) \leq P(B)$ 的要求——也被称为联合定律 (conjunction rule), 对相似度或者代表性并不适用。比如说, 一个蓝色的正方形比圆形更相似于蓝色的圆形; 相比于民主党人给我们的印象, 某个人给我们的印象可能与民主党艺术家更相似。因为某客体与目标物的相似度, 可以通过给目标物添加客体所共有的特性的方式而增加 (见 Tversky, 1977), 因此, 相似度或代表性, 可以通过对目标物的详细说明而增加。如果概率判断受到代表性或相似度的调节, 则我们可以构造出这样的问题, 在这些问题上, 结果的联合显得更具代表性, 因而显得比它单个的组成成分更有可能发生。

1. 联合效应 (conjunction effect): 研究 1

这一预测, 首先是在 1974 年于耶路撒冷进行的一项实验中被检验的。我们给 184 位被试四份人格描述。每一描述, 都与某一职业 (如出租车司机) 的刻板印象相符合, 但与某一政治党派 (如工党) 的刻板印象截然不同, 或者相反。因此, 每份描述 (X) 都对某一目标物——我们标记为 A ——具代表性, 而对另一目标物——我们标记为 B ——不具代表性。每份描述的后面, 是一组五件或六件目标物事件, 它们是由从事某份职业的人、某党派的人, 或者它们的组合——比如, 一位加入了工党的出租车司机——讲述的。对每份描述, 一半被试收到的列表上包括目标物 A 和目标物 B , 而另一半被试收到的列表包括混合目标物 ($A \& B$)。剩下的四个目标物, 在两份列表上是一样的。一半被试, 被要求根据“ X 对该类别具代表性的程度”对目标物排序, 另一半则根据“ X 是该类别的成员的概率”对目标物排序。 91

研究设计可以让我们对目标物 B 和混合目标物 ($A \& B$) 相比于四个不变的替代选择的代表性和概率, 进行间接的比较。结果如下。第一, 相比于单个目标物 B , 所有这四份描述都被判断为对混合目标物 ($A \& B$) 更具代表性。第二, 在所有情况中, 对每组代表性和可能性的排序都是几乎一致的; 平均排名之间的平均积距相关系数是 0.96。特别要注意的是, 混合目标物 ($A \& B$), 在概率排序中, 被赋予的平均排名, 要显著高于单个的目标物 B 。显然, 对代表性启发式的依赖, 导致回答者认为混合事件比

它单个的组成部分更可能发生，而这与概率理论中的联合定律是不一致的。这种判断的模式，我们称为联合效应。

2. 研究2：比尔和琳达

因为早先研究中所使用的刺激材料 (stimulus material)，是置于以色列文化中的具有高度特殊性的材料，所以我们对其进行了几项重大修改，设计了一份英文版的问题并重复了这项研究。首先，我们对被试间设计的结果和被试者内设计的结果进行比较，前者要求每位被试将混合目标物 ($A \& B$) 或者单个目标物 B 与相同一组替代选择进行比较，后者则是每位被试对两个关键目标物进行直接比较。我们假设联合定律在前一设计中不会生效——正如我们前面的研究一样，但我们预期在后面的设计中，违背联合定律的频率会显著降低，因为后者中的被试，在本质上被要求将 $P(A)$ 和 $P(A \& B)$ 进行比较。其次，我们预期，即使被试不具有多少统计上的严密性，大部分对联合定律的违背也会被消除，至少在被试者内设计中是这样的。

为检验这些假设，我们用相同的刺激材料进行被试内（直接）和被试间（间接）研究。重复研究中的三组被试的统计严密性各不一样。统计知识空白组，是由英属哥伦比亚大学和斯坦福大学的本科生组成，他们没有任何概率或统计的背景。中等组由斯坦福大学心理学、教育学和医学的研究生组成，他们都修过数门统计课程，且对概率的基本概念非常熟悉。统计知识严密组，由斯坦福商学院决策科学项目中的研究生组成，他们都修过多门概率和统计的高级课程。

22 我们撰写了两份简要的人格描述。每位参加者都在被试内研究中看到其中一份描述，而在被试间研究中看到另一份描述。在前者中，人格描述后面是八个可能的结果，包括一个代表性的结果、一个不具代表性的结果和这二者的组合。在被试间研究中，结果的清单包括两个关键的单个结果或者它们的组合。被试内研究中的问题如下所示。括号中的数字是收到这份表格的被试赋予不同结果的排名的均值。

比尔 34 岁，他聪明但缺乏想象力，冲动，通常都是毫无生气的。
在学校时，他擅长数学，但在社会科学和人文科学方面比较弱。

第 6 章 有关代表性的判断和依据代表性的判断

请根据下面陈述的可能性对其排序，用 1 表示最有可能的，用 8 表示最不可能的。

- (4.1) 比尔是位内科医生，打扑克是他的一项业余爱好。
- (4.8) 比尔是位建筑师。
- (1.1) 比尔是位会计师。(A)
- (6.2) 弹爵士乐是比尔的一项业余爱好。(J)
- (5.7) 冲浪是比尔的一项业余爱好。
- (5.3) 比尔是位记者。
- (3.6) 比尔是位会计师，弹爵士乐是他的一项业余爱好。(A & J)
- (5.4) 登山是比尔的一项业余爱好。

琳达 31 岁，单身，坦率直言，且非常伶俐。她主修哲学。做学生时，她对歧视和社会正义的问题非常关注，也曾参加反核游行。

请根据下面陈述的可能性对其排序，用 1 表示最有可能的，用 8 表示最不可能的。

- (5.2) 琳达是小学教师。
- (3.3) 琳达在书店工作，她参加瑜伽功的练习。
- (2.1) 琳达是女权运动中的积极分子。(F)
- (3.1) 琳达是精神治疗的社会工作者。
- (5.4) 琳达是女选民联盟 (League of Women Voters) 的成员。
- (6.2) 琳达是银行出纳。(T)
- (6.4) 琳达是保险销售员。
- (4.1) 琳达是银行出纳，且是女权运动中的积极分子。(T & F)

读者也许已猜到，对比尔的描述被设计为对会计师 (A) 具代表性，而对爱好弹奏爵士乐的人不具代表性 (J)。对琳达的描述被设计为对女权运动的积极分子 (F) 具代表性，而对银行出纳 (T) 不具代表性。根据心理学的相似度原则 (Tversky, 1977)，混合目标物——爱好弹奏爵士乐的

会计师 ($A \& J$) 和积极参加女权运动的银行出纳 ($T \& F$), 将介于相应的单个目标物之间。为验证这一预测, 我们要求 88 位统计知识空白的被试, 根据“比尔 (琳达) 与该类别的典型成员相似的程度”对八个目标物排序。相似度的排名验证了我们对这些描述的假设。回答者中表现出我们对比尔的预测排序 ($A > A \& J > J$) 的比例是 87%; 表现出我们对琳达的预测排序 ($F > T \& F > T$) 的比例是 85%。

所有被试都收到被试内研究样式的对比比尔的描述或者对琳达的描述, 并根据八个目标物的发生可能性对其排序。这些数据汇总在表 6—1 中的上半部分, 其中标记了“联合效应 (%)”的一行表示的是, 被试中将混合目标物排在单个的不具代表性的目标物之前的比例。标记为“ $A \& B$ ”和“ B ”的行分别表示的是, 赋予混合目标物和单个的不具代表性的目标物的排名的均值。我们在图 6—1 中描出了被试内设计的相似度排名对应概率的总体平均排名的点。

表 6—1 联合效应

	空白		中等		严密	
	琳达	比尔	琳达	比尔	琳达	比尔
被试者内设计						
联合效应 (%)	89	92	90	86	85	83
平均排名: $A \& B$	4.2	3.6	3.9	3.5	4.0	3.4
平均排名: B	6.3	6.4	6.2	6.4	6.1	5.6
N	88	94	53	56	32	32
被试者间设计						
平均排名: $A \& B$	3.3	2.3	2.9	2.4	3.1	2.5
平均排名: B	4.4	4.5	3.9	4.2	4.3	4.6
N	86	88	55	56	32	32

在被试间设计的条件下, 两个问题的版本, 是通过从目标物清单中删去混合目标物或者两个单个的目标物而得到的。人格描述、说明和其余的五个目标物, 与被试内设计的版本是一样的。在被试间设计中的所有被试的结果, 呈现在表 6—1 的下半部分。

表 6—1 中汇总的结果表明, 混合目标物在被试内设计和被试者间设计中, 都被判断为比关键的单个目标物更有可能发生。这一结果对两个描述

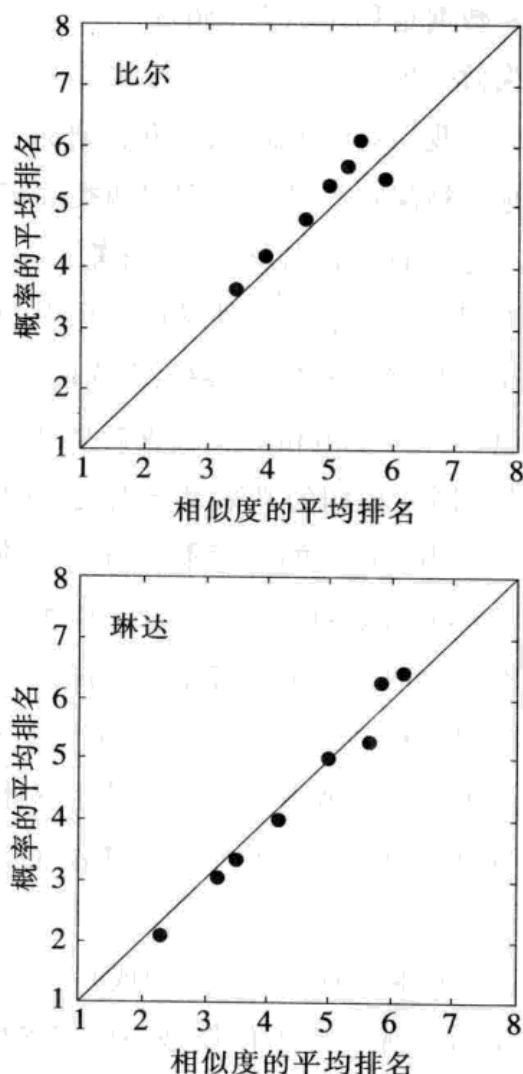


图6—1 在两个描述中，按照概率和相似度对八个结果的平均排名的散点图

和所有组都成立。让我们吃惊的是，统计知识的严密性对联合效应的影响甚微，这体现在三组中都有多于80%的被试受联合效应的影响。

在前面的研究中，关键目标物都是放置在大量的可能结果中，这可能会掩盖它们之间相互包容的关系。因此，对目标物之间的逻辑关系非常明显时，人们是否会违背联合定律进行考察，是有意义的。为检验这一假设，我们将比尔和琳达的描述呈现给一组新的（统计知识空白）被试。每位被试都得到两份描述中的其中一份，并被要求回答两个关键目标物中〔即 J 和 $(A \& J)$ ，或 T 和 $(T \& F)$ 〕哪个更有可能发生。这一程序并没有降低联合效应：在比尔的情形中，92% 的被试 ($N = 88$) 选择了混合目标物，

在琳达的情形中，这一数据是 87% ($N=86$)。

联合定律的严重失效，提出了一个饶有趣味的关于它作为标准诉求的问题。为考察这个问题，我们对中等组中的 36 位研究生进行了访谈，他们都参加了这项实验。他们被问到：(1) 你们对两类关键的目标物是如何排序的，(2) 为什么你们这么排序，以及 (3) 请考虑这一论述，“比尔同时是会计师和爵士乐手的概率不可能超过他是爵士乐手的概率，因为前一类别中的所有成员同时也是后者的成员。”超过 2/3 的被试表现为：(1) 说他们选择了混合目标物，(2) 给出相似度或者典型性的论述作为理由，以及 (3) 经过一段考虑，同意他们的答案是错误的，因为它不符合联合定律。只有两位被试坚持认为，概率的排序不必与类别的包容保持一致，只有一位被试声称自己误解了问题的含义。尽管访谈可能有诱导被试认可联合定律之嫌，但结果至少表明，有统计背景的被试，愿意将对联合定律的违背视为令人遗憾的错误。有关该问题的进一步讨论，请参见第 34 章。

在解释联合定律的失效时，考虑这一结果是否整体地或部分地归结于语言惯例或者交谈规则是很重要的。譬如，在早先的一个研究中，我们给人们下面这个描述，“约翰 27 岁，性格开朗。在大学时，他是一位出色的运动员，但在学识问题上没有表现出多大能力或兴趣。”我们发现，约翰被判断为更像是一位“体育老师”而不仅仅是一位“老师”。虽然每位体育老师在某种意义上都是老师，但我们可以认为，老师这个词语在这里的含义排除了体育老师或者驾驶学校的教员。在本章的实验设计中，我们在外延上将关键的结果限定为两个集合的交集，比如说，业余爵士乐手的会计师，这样就避免了这个问题。

对联合定律的违背，在序列问题中——目标物由一个序列的事件组成——也被观察到。斯洛维奇等 (Slovic et al., 1976) 呈现给被试某人的 人格描述，此人与工程师的刻板印象相类似，但与新闻记者不类似。被试赋予“汤姆将选择新闻学作为他的大学专业”这一事件的概率，要低于赋予“汤姆将选择新闻学作为他的大学专业，但很快不乐意于他的选择而改学工程学”这一事件的概率。严格来讲，前一事件包括了后者，因而上面的判断违背了联合定律。但是，这个例子存在可商榷之处，根据谈话的标

准规则，汤姆选择新闻学作为他的大学专业，这一表述意味着他仍然是新闻学专业的。否则，这一表述就是有歧义的。

相同的异议也可以对比尔和琳达的例子提出。人们可以争辩说，被试将“一位银行出纳”这一类别理解为“一位不积极参加女权运动的银行出纳”，这与给出的类别“一位积极参加女权运动的银行出纳”是不一样的。⁹⁶但是，在被试间设计中——其中，关键目标物之间不是进行直接的比较——所呈现的联合效应表明，该效应，不能由根据标准的谈话蕴意（implicature）对目标物类别进行的重新表述给予充分的解释。相反地，观察到的判断，表现出一种普遍的倾向，即根据琳达代表各自类别中的典型或者原型性的成员的程度，评估相关事件发生的概率。

另外，我们也注意到在有些任务中，联合效应显得不受谈话蕴意的影响。下面的问题关注的是对未来事件的预期，这里将 B 解释为 $(B \& \text{非} A)$ 看来是不合理的。

3. 研究3：对1981年的预测

这里描述的问题，被设计来检验在预测真实世界中的事件时的联合定律，其中被试可以依赖他们的常识进行预测。这些问题是在1980年的12月，由93名没有统计背景的被试组成的小组回答的。说明如下所示：

在这份问卷中，你被要求评估各种事件在1981年发生的可能性。每个问题都包括四个可能的事件。你的任务是对这些事件按照概率进行排序，用1表示最可能发生的事件，2表示第二可能的事件，3表示第三可能的事件，4表示最不可能发生的事件。

这份问卷包括六个问题。我们下面给出了其中的两个问题。其他问题的结果非常相似。括号中的数字是每个事件的平均排名，我们也给出了将复合目标物排在单个目标物之前的被试的比例。

1981年网球赛（联合效应：72%）

假设博格（Bjorn Borg）进入了1981年温布尔登网球赛决赛。请

对下面几种结果按照从最可能到最不可能的顺序排序。

(1.7) 博格将赢得这场比赛。

(2.7) 博格将输掉第一局。

(3.5) 博格将赢得第一局但输掉整场比赛。

(2.2) 博格将输掉第一局但赢得整场比赛。

1981 年的美国政治 (联合效应: 68%)

请根据下面事件在 1981 年发生的概率对其排序。

(1.5) 里根将削减对地方政府的联邦支持。

(3.3) 里根将为未婚妈妈提供联邦支持。

(2.7) 里根将以低于 5% 的幅度提高国防预算。

(2.9) 里根将为未婚妈妈提供联邦支持, 而削减对地方政府的联邦支持。

97 如前面的研究一样, 复合类别, 被判断为比单个的组成部分更可能发生。该结果与代表性的概念是一致的, 代表性在这是指原因系统和其结果之间的关系, 而不是某项描述与刻板观念之间的相似度。比如说, 在第二个问题中, 为未婚妈妈提供支持, 对总统里根不具代表性, 而削减对地方政府的联邦支持非常具有代表性。这些行为的复合显得具有中等的代表性, 对概率的评估显然也遵循同样的模式。

在第一个问题中, 大部分回答者评估博格赢得比赛是最有可能的, 而视博格输掉第一局不太可能。这两者的复合, 也就是博格输掉第一局但赢得整场比赛, 再次被判断为比第一种可能性的发生概率要低而比第二种可能性要高。显然, 被试对事件的组合是根据代表性原则, 或者因果效应, 而不是根据概率的法则。

4. 讨论

前面研究的结果为我们的假设——人们根据这些事件代表相关模型或过程的程度, 评估事件发生的概率——提供了直接的支持。因为事件的代表性, 可以通过详细化而增加, 混合目标物, 会被判断为比它的单个组成

部分更可能发生。这一预测，由使用被试内设计和被试间设计的研究给予了支持，被试总体，覆盖了广泛的有各种统计背景的群体。

不同于其他的概率规则，如向均值的回归——普通被试发现它们难以理解和接受，联合律既简单又引人注目。大部分被试，都愿意以抽象的形式去接受它，但当联合律与代表性的直觉发生冲突时，他们当中几乎所有人都在实际中违背了它。现有结果与有的学者（Johnson-Laird & Wason 1977）关于对“如果……就……”表述证实的发现恰恰相反（参见 Johnson-Laird et al., 1972）。这些研究者发现，在确认的任务中，大部分被试，成功完成使用抽象材料的证实任务，但在使用具体例子的任务中则会失败。相反地，我们的被试，在抽象形式中认可联合定律，而在具体例子中则违背了联合定律（见第34章）。

混合目标物比其单个组成部分显得更有可能发生，这一发现可能会有深远影响。我们没有理由相信，那些政治分析员、陪审员、法官和内科医师的判断会不受联合效应的影响。在以评估某一特定剧情或场景（scenarios）发生可能性的方式来预测未来的尝试中，这一效应可能是尤为有害的。当他们向水晶球里注视时，政治家、未来学家和普通人，都一样地在寻找一个未来的形象，使其最能代表他们心中的当前的动态前景的模型。这种寻找，将导致对具体剧情的构建，这些剧情将是内部一致的，且高度代表着我们心中世界的模型。此类剧情，往往显得比不具体的预测更有可能发生，但事实上，后者更有可能发生。当剧情中的细节数量增多时，它的概率只会稳定地下降，但它的代表性以及相应的表面可能性则会增加。我们相信，对代表性的依赖，是人们对此类构建所经常提供的具体剧情和虚幻洞察力无根据青睐的主要原因。

对概率的考虑和对相似度的考虑之间的混淆，不仅发生在对不确定未来的预测上，也出现在对不确定的过去的重构中，比如说在历史和刑事案件中。这里，对过去事件的描述，通常被纳入到一个代表性的剧情中，它包括对未知事件进行的似是而非的猜测。这些猜测的纳入，只会降低整个事件是真实的可能性，但它给人具有代表性和一致性的感觉，因而可能会增加感觉中的剧情的发生可能性。比如说，“被告离开了犯罪现场”的假

设，相比于“被告因为害怕被指控犯谋杀罪而离开了犯罪现场”的假设，似乎合理性更小，但后一描述比前一描述的可能性要小。一个好的故事，比一个不怎么动听的故事，更不可能发生。

最后，认识到联合效应只是一个更基本问题的表征是很重要的。它仅仅暴露出了概率逻辑和代表性逻辑之间的不一致，后者往往操控着人们对不确定事件的信念。因为人们的判断与许多对我们生活有重大意义的问题密不可分，所以直觉上的概率概念和逻辑架构上的概率概念之间的冲突，是非常令人忧虑的。一方面，我们无法轻易地放弃我们用来评估不确定性的启发式，因为我们关于这个世界的知识，有很大部分和它的运作联系在一起；另一方面，我们又不能违抗概率法则，因为它们抓住了我们所处世界中的真实。无论你喜欢与否， A 不可能比 $(A \& B)$ 的概率要小，相反的观点将是有害的。我们的问题将是，在纠正直觉判断所倾向造成的错误和偏差的同时，保留直觉判断中有益的和有效的部分。

本章是按照与斯坦福大学的合同 N00014-79-C-0077，由海军研究办公室提供资助的。



第三部分
PART THREE

因果性和归因



第 7 章

普通的归纳：信息不必然使人知情

尼斯比特等 (Richard E. Nisbett,
Eugene Borgida, Rick Crandall, and Harvey Reed)

目前 (1976) 对社会心理学家影响最大的认知理论是归因理论。其形式化的版本，由凯利 (Harold Kelley) 于 1967 年提出。这个理论把人看成非专业的科学家，力图对所观察到的结果进行原因推论。他的原因归因，决定其社会世界观，而这种世界观，又会决定其行为。从阿希 (Asch) 的从众研究到沙赫特 (Schachter) 的情绪研究，有极其广泛的现象，可以有效地看作是因果归因过程在发生作用的例证。实际上，凯利最为重要的贡献，最终极有可能被认为是他创造了一种语言，或者说是一个路线图，用以描述并连接不同的社会心理现象。

除了其组织性贡献，凯利还提出了因果归因过程中三个正式的影响源。在试图对“行动者在 A 情境下以 X 进行反应”这一类事件进行归因时，他认为平常的归因者要面对三种信息来源：特异性 (distinctiveness) 信息 (行动者对同类的所有情境还是仅仅在 A

情境下，以 X 进行反应?); 一致性 (consistency) 信息 (行动者是一直有这种反应，还是纯属偶然?); 共识性 (consensus) 信息 (是大多数人还是只有相对的少数人有这种反应?)。原因归因便依赖于对上述每个问题的回答。如果行动者在同类的所有情境中都如此反应，而其他人在此情境下并不这样，那么行动者本人就是这个行为的首要原因。而如果行动者只在 A 情境中有如此反应，且大多数人在此情境中也是如此，那么情境就是行动者行为反应的原因。 102

凯利的归因过程分析，由于其极端的形式化而受到赞同，但也受到批评。不管你对它是欣赏还是厌恶，你都会惊讶地发现其中的一个基本公理，在后续的研究中基本得不到任何支持：这就是人们在分配原因时会对共识性信息作出反应。尽管有理论和常识的支持，仍有越来越多的证据显示，人们在因果归因时基本上不受有关他人行为的知识影响。共识性反应的知识，似乎并不会促进情景归因而不是行动者归因。与之相反，低共识性反应的知识，也不会促进行动者归因而不是情景归因。

下面我们要评论相关证据，它们很少支持这种观点，即人在归因时会利用共识性信息。这些证据涉及两类例证：行动者是他人，以及行动者是自我；这后一类的例证，主要源于我们自己的研究。然后，我们将揭示共识性信息对归因影响的失败与卡尼曼和特沃斯基 (1973, 4) 所证明的基率信息对预测影响的失败之间的相似性。我们将依据抽象信息 (共识和基率) 与具体信息 (和行动者或目标相关的信息) 的相对影响，来对上述两种失败进行解释。最后，我们将把抽象信息和具体信息之间的区分应用于沟通与劝说的研究中。

一、共识性信息与对他人的感知

有两个研究，考察了有关他人行为的共识性信息对归因的影响。它们都显示出共识性信息的影响极其微弱。其中的第一个实验，是由麦克阿瑟 (L. Z. McArthur) 在 1972 年做的。她的研究，直接检验了凯利关于特异性、一致性和共识性信息对因果归因影响的命题。被试拿到一个只有一行的描述：“行动者对刺激作出 X 反应”；除此之外，他们还得到一些凯利的

特异性、一致性和共识性维度上的信息。例如，被试可能被告知：“在跳舞的时候，拉尔夫踩了琼的脚”，此外还被告知拉尔夫差不多会踩所有女孩子的脚（或几乎没踩其他女孩子的脚）；或者拉尔夫几乎总是（或者从来没有）踩琼的脚，而差不多所有其他人（或者没有他人）都踩琼的脚。然后，研究者让被试回答踩脚事件是拉尔夫的错、琼的错，还是仅仅舞场环境的错？他们还要求被试预测反应普遍性（拉尔夫建议琼报名参加一个交际舞培训的可能性有多大？）以及刺激普遍性（拉尔夫有多大可能会在一个结冰的人行道上摔跤？）。

特异性信息，占有原因归因（将所有原因加总）的10%，和所有刺激普遍性预期的63%；一致性信息，占有原因归因的20%，和所有反应普遍性预期的14%。相反，共识性信息，在上述三种推论的任一种中所占的比重都不超过3%。这个结果，不仅违背了归因理论的常识，也根本违背了所有常识。虽然被试似乎相信这样的信息是重要的，即拉尔夫是否踩大多数女孩子的脚以及他是否经常踩琼的脚，但对他们而言，其他人是否也踩琼的脚，则是无足轻重的。

对他人的感知的其他研究中，归因的或任何其他类型的常识，也被违背。1973年，米勒（Miller）等让大学生阅读米尔格拉姆（Milgram）在1963年所作的经典的服从研究的程序。其中一半被试被提供的是米尔格拉姆研究的真实数据，它们显示出实际上所有被试都对实验同谋施以高量度的电击，而且大多数被试的电击都一致达到了最高量度。而其余的被试，则让他们保留其质朴的期望，即那样的行为应该是很少发生的。然后，所有被试被要求对两个个体进行人格特质评定，这两个个体实施了最高量度的电击，而特质评定是在11项负荷强评价意义的人格特征维度上进行的，如吸引力、热心、可爱、攻击性等。在这11个维度的评定中，仅有一个受到了共识性信息的显著影响。因此，施行最大量度的电击是普遍行为的信息，实际上，对评价已实施了最大量度电击的人，没有影响。

二、共识性信息与自我感知

共识性信息，似乎对自我行为的归因也没有多大影响。贝姆（Bem，

1967) 提出, 人们对自己的行为进行因果分析所采用的方式与对他人行为进行归因的方式完全类似, 凯利则在同年把此观点结合到他的归因理论中。人们观察自己的行为反应, 注意这些反应所发生的情境, 然后推论其感受和动机状态。例如, 在沙赫特和辛格 (Schachter & Singer) 的经典实验中, 被试知道自己被注射了一种能导致自发唤醒的药物, 在被置于能引发强烈情绪的环境中, 他们就会采取特定的因果分析。他察觉到了自身唤醒的症状, 在通常情况下, 他会将此归因于情景的影响, 而现在他则归因于他所注射的药物。其结果是, 比之那些不知道自己被注射了唤醒药剂的被试, 甚至是没有注射唤醒药剂的控制组被试, 他更少报告也更少在行为上表现出其情绪症状。他因此将自身的自发反应归于“外在的”的原因, 并依此进行感受和行动。

104

几年前, 以人能受引导对自己行为反应进行外在归因的观点为基础, 我们开始了治疗干预的研究项目。似乎看来, 无论个体的反应是适应不良的、精神崩溃的, 或者是病态的, 劝说他将这些反应归结为某种外在的原因, 也许会有一定的成效。第一个, 也是唯一成功的一个研究, 是由斯多姆斯和尼斯比特 (Storms & Nisbett) 在 1970 年做的。所研究的病状是失眠症。我们要求入睡困难的大学生在就寝前 15 分钟吃一片药 (实际上是安慰剂), 并告诉他们这药物会引起心跳加速、呼吸短促而无规律、全身发热, 以及一种平常的清醒状态。显然这些都是失眠症的症状。吃了药的被试报告说, 他们在吃了药的夜晚要比没有吃药的夜晚入睡更快, 而且比那些没有吃药的控制组被试也要入睡快。斯多姆斯和尼斯比特分析认为, 有一个或两个不同的归因过程能解释此结果。失眠症可能主要是由就寝时的唤醒所造成, 而唤醒有种种原因, 这些原因包括对个人问题的焦虑, 白日里不适当的生活节奏, 或者慢性的神经官能症。当人躺在床上却处于唤醒状态时, 他沉醉在富含情感内容的想法中。这种唤醒, 能和情感认知相联结, 并促进情感认知。而由此增强了的情绪状态, 又会强化唤醒状态, 于是形成恶性循环。然而, 如果认识到唤醒本质上是外源性的, 恶性循环就可以打破。他就不会从对自身唤醒状态的观察中, 推论出他多么地担心考试, 或者对室友多么地气愤。在服用过药物的夜晚, 唤醒状态被视作是不可避免

免的，与他脑中的任何想法都没有关系。不断强化唤醒状态的恶性循环圈，因此被打破，睡眠接踵而至。

作为替代，或者作为补充，一个略微不同但更有普遍适用性的过程，应该也在发生作用。我们的失眠被试报告说，他们对自己是失眠症患者的事实非常担忧。他们把此看成是更为普通的病理学的证据，并反映了自身不良的心理调节状态。对一个具有如此忧虑的被试来说，意识到唤醒是不可避免的，并且是外在引发的，这将使他们安心。至少在服药的当晚，被试可以推断，失眠症不会是普通心理病理学的证据。即使这种担忧本身对失眠负有部分责任，但在服药当晚，睡眠还是会来得更快。

1. 抑郁操纵

有了成功干预失眠症的经验，我们开始了一系列的尝试来减轻抑郁。在所有研究中，所采用的技术都是共识性操纵，通过让被试确信抑郁是普遍共享的，以求将抑郁原因外在化。既然这种状态为类似的其他人所共有，它的存在就更少消极地反映自我。它就不会是根源于被试对特定情境和环境所作的独特的并且可能是病态的反应。随着对自己处理生活情景能力的担忧的减弱，抑郁会部分减轻。

2. 研究1：星期天忧郁症

许多大学生，在星期天会体会到一种普遍的低落情绪。虽然这一天开局很好，有早午餐、咖啡、周末报纸，但在下午通常会萌生厌倦感。要多多地筹划下周的工作，筹划太多而无法决定周末的约会，尽管可能不是多到马上就得开始进行。临近黄昏，没有外出，也没有做任何事，周末报纸甚至包括财经和旅行版面，也嚼了一遍，而漫长枯燥的苦修之夜就要开始。到了晚上，如果所有事情还像通常的一样糟糕，那么学习就会痛苦不堪，或者毫无感觉，而黯淡的心境以及自我怀疑，就会涌现。

这让我们想到，如果这种星期天抑郁现象非常普遍，而且人们也意识到这一点，那么它在强度上就能得以缓解。如果一个学生知道他周围的室友都处于同样的厌倦状态，他自己的消极情绪也就会略微减轻。他也许意识到人们都会在星期天心情沮丧而顺其自然，而不再整天担忧自己的学业，或因自己不再可能与某个女孩子维持良好关系而耿耿于怀。

为了检验这个观点，我们在耶鲁大学请了大批的男生在星期天的下午4点及晚上10点填写心境量表。心境量表由魏斯曼等（Wessman et al., 1966）所设计的量表中的项目所构成，但主要专注于其中的欣快感—忧郁感（euphoria-disphoria）因素。此外，我们还要求他们在晚上10点填写一份问卷，报告他们当天的学习和社交活动，以及他们借以发泄不快情绪事件的次数，如大叫或者哭泣。最后，还要求被试从文件夹中取出一叠卡通画并评价它们的有趣程度。

经过这些最初的星期天预先测试，把被试分成三组，每组18人。第一个组是控制组，被试仅被告知调查者正在研究星期天的心境模式，需要他们在下个星期天填写与上个星期天相同的材料。对于第二组被试，星期天抑郁综合症被详细地加以讲解，并给他们看一些伪造的统计资料，以说明这种抑郁症在大学生中的广泛存在。被试被告知耶鲁学生中有92%都报告说至少偶尔有过这样的情绪，而65%的学生则在大多数星期天都有同样的体验。而第三组被试，被给予和第二组相同的共识性信息，但附加对此现象的理论解释。研究者告诉他们星期天抑郁是由一种在星期天发生的“唤醒冲突”（arousal clash）造成的：平日里的唤醒状态在周六达到高潮，而到了星期天，则坠入谷底；而这种唤醒缺乏则经常就被解释为，或者转化为抑郁。

我们预期后两个实验组的被试，将重新解释他们在星期天的苦恼情绪，将它们更少地归于自身原因，顺其自然，不再因抑郁情绪而沮丧。果真如此，他们应该在心境量表上显示出，从操纵前的星期天到操纵后的星期天，他们的不快情绪在减少，吼叫或哭泣等不悦行为也在减少；而学习或社会活动则在增加，对卡通画有趣程度的评价也会普遍提高。但上述这些现象，都没有在他们身上发生。实验组被试的任何心境指标，与控制组被试相比都没有提高。

3. 研究2：慢性抑郁

我们没有因此气馁，而试图对男大学生进行类似的治疗干预，这些学生认为自己处于慢性抑郁状态。用在密歇根大学的学生报纸上登广告的方式，我们招募了20名被试，广告词为招募“抑郁的男性高年级学生参加一

个由抑郁社会研究所实施的研究”。

被试来到实验室后，就被随机分成两组。我们告诉控制组被试已设计出一个新的心境量表，以及使用这个量表获取那些自认为是抑郁者的日常心境报告的重要性。而它会有助于评定量表的有效性及其探测心境转变的能力。

107 给予实验组被试的是同样的故事，此外，他们还被告知实验者正处在一个理论检验的最后阶段，而这个理论和年轻成年男子的抑郁有关，至少基于一些特异的事实。这个理论如下所述：被试被告知我们已知道成年人心境的保持部分依赖于性激素的存在，而男性则依赖于睾丸激素。然而，直到现在还存在着一个悖论，即儿童几乎从不沮丧。由于儿童体内的性激素非常少，这显然就与荷尔蒙能促进良好心境这一普遍规则相抵触。不过最近这个矛盾得到了解决，因为有人发现大脑的情绪中枢——边缘系统（limbic lobe）在青春期时会依据性激素进行转换以保持心境。在大多数男性中，这个转换的时间期限与体内激素的增长相当吻合，而体内激素水平大概在 25 岁时达到顶点。但是，对许多年轻男性而言，这个转换过程，在“燃料”的数量并不足以保持心境之前，就结束了。

对这些被试而言，这个理论的强烈含义在于他们就是这样的年轻男子。我们预期这样的操作化，由于下列三个复杂原因而使心境改善：

（1）它暗含着抑郁情绪有时间界限。

（2）被试现在可以把消极情绪归于一个不幸的生理事件，而不是自己的生活圈以及无法控制它们的任何病理学的无能。从这个意义上讲，他们把不良情绪“外因化”了。

（3）这种消极情绪及其原因为多数他人所共有，它意味着被试的苦恼及反应并不是唯一的。

所有的被试都被要求在两个星期的实验期内，每天晚上填写前述的魏斯曼心境量表。问卷还包括被试前一晚睡眠情况的报告，因为睡眠障碍是抑郁的常见症状。最后，在研究进行的学期期末，统计被试学习成绩的平均绩点。

在两个星期的实验期间，实验组和控制组的心境报告没有任何差异，睡眠质量的报告也是如此。实际上，这些变量没有向假设的方向发展的任

何迹象。虽然正如我们所预期的，实验组被试平均绩点较高，但它缺乏统计的显著性 ($0.05 < p < 0.10$) ……

三、归因与预测心理学

卡尼曼和特沃斯基所揭示的推论失败，似乎与共识性信息影响研究中所观察到的推论失败高度相关。在《论预测心理学》(1973, 4)的论文中，卡尼曼等人表明，在预测目标成员的范畴成员身份 (the category membership) 时，人会忽略整体基率 (population base rates) ……

如果被试在预测目标个体的范畴成员身份时不受基率信息的影响，那么他们的归因似乎也不会受到共识性信息的过多影响。而共识性信息，恰恰就是基率信息。它是关于行为反应而非范畴成员身份的基率信息。而且归因，是比预测更复杂也更间接的推论。卡尼曼和特沃斯基让他们的被试组织一个相当直接也并不复杂的推论链：“如果整体中的大多数成员都属于特定范畴，那么目标个体也有同样的几率属于该范畴。”可他们的被试，都没能够作这样的推论。而在我们已讨论的归因研究中，所要求的是一个更为精致的推论链：“如果整体中的大多数成员都表现出特定行动，那么情境的强大压力必然趋向这种行为；因此，如果目标个体的行为也是如此，那么利用该个体的特质去解释其行为就是多余的。”

当然，被试是否不愿意利用行为基率去预测目标个体，这仍有待检验。如果他们确实不愿意，那么我们一直探求的问题，就必须从“为什么人们不能依照共识性信息改变他们的归因？”，转向一个更为基础的问题“为什么人们如此处理基率信息，就好像它并没有提供任何信息一样？”

研究5：行为基率、预测和归因

为了研究人在面对行为基率信息时是否愿意改变其预测，我们 (Nisbett & Borgida, 1975) 给被试提供了两个心理实验。被试是来自密歇根大学的学生。他们要仔细阅读两个实验的描述：(1) 尼斯比特等 (1966) 的实验。在实验中，被试被要求尽可能地承受最大量的电击。(2) 达利和拉塔尼 (Darley & Latané, 1968) 的助人行为实验。在实验中，当几名学生在彼此隔离的小房间里通过耳机讨论大学生活适应问题时，一名“被试”突

然发出类似疾病发作的声音。我们之所以挑选这两个实验，是因为我们在教学经验中发现，大学生对行为基率的猜测是不准确的。虽然学生倾向于假定很少有被试能承受高量度电击，但实际的普遍行为，是被试可承受仪器所能提供的最强电击，它足以引起被试整个手臂不由自主地痉挛。而就达利和拉塔尼的实验而言，在六人一组的条件下，其典型被试，从来不帮助受害者，但学生则假定，多数人都会迅速离开他们的小房间去帮助受害者。

10.9 鉴于被试并不知道真实的行为比率，就有可能只告诉一部分被试两个实验的真实比率，以创造分化的比率信息条件情景。获得比率信息（共识性信息条件）的被试，观看了原初实验中对被试进行采访的简短录像带（或者，采用一种共识性信息条件的变体形式，让他们读一些对那些学生的背景和人格的简单描写）。共识性信息条件下的被试，被要求预测他们所看到或读到的目标个体在实验中的行为反应。这样就有可能将获得共识性信息的被试的预测，与他们所了解的实际比率信息进行比较，以及与那些缺乏共识性信息的被试对比率的猜测进行比较。

图 7—1 所显示的，是对有关电击实验的结果。最上面的柱形图所显示的，是提供给共识性信息被试组的实际比率数据。第二排的柱形图所显示的，是缺乏比率信息的被试组所作的比率预测。可以看出这些无共识性信息的被试所作的预测与实际数据相差很大。他们认为承受中等强度的电击应该是普遍的行为。第三排图所显示的，是具有共识性信息的被试对他们所看到或读到的目标个体的行为的猜测。虽然这些被试完全清楚比率信息，但仍可以看到它的分布与实际比率甚至没有一丁点儿类似。相反，它的分布与缺乏比率信息的被试所作的对比率的猜测相当接近。关于助人行为的实验结果也完全一样。

在一篇题为《小数律的信念》的论文中，特沃斯基和卡尼曼（1971，2）提出一个假设：即使是科学家对样本规模也相当迟钝，他们往往基于极少数案例的知识就草率地对整体进行强推论。我们这个实验对之也提供了一个检验机会。为了在现在的语境中验证这个观点，我们让一些被试对两个实验的比率一无所知，并让他们观看每个实验中的两名被试被采访时的简短录像。这些目标信息条件下的被试，被告知两个实验中的那两个被试

都采取了最极端可能的行为方式，也就是说，在电击实验中的两名被试都承受了可能达到的最强电击，而助人实验中的两名被试则都没有去帮助受害者。在告知这些信息之后，我们要求这些目标信息条件下的被试预测这两个实验中整个被试群体的行为分布。就这两个实验而言，被试都倾向于推断整个被试群体的行为模式与他们所观看的两个人的行为模式是一致的。从图7—1底端的柱形图可以看出，对电击实验的基率猜测，跟实际的基率

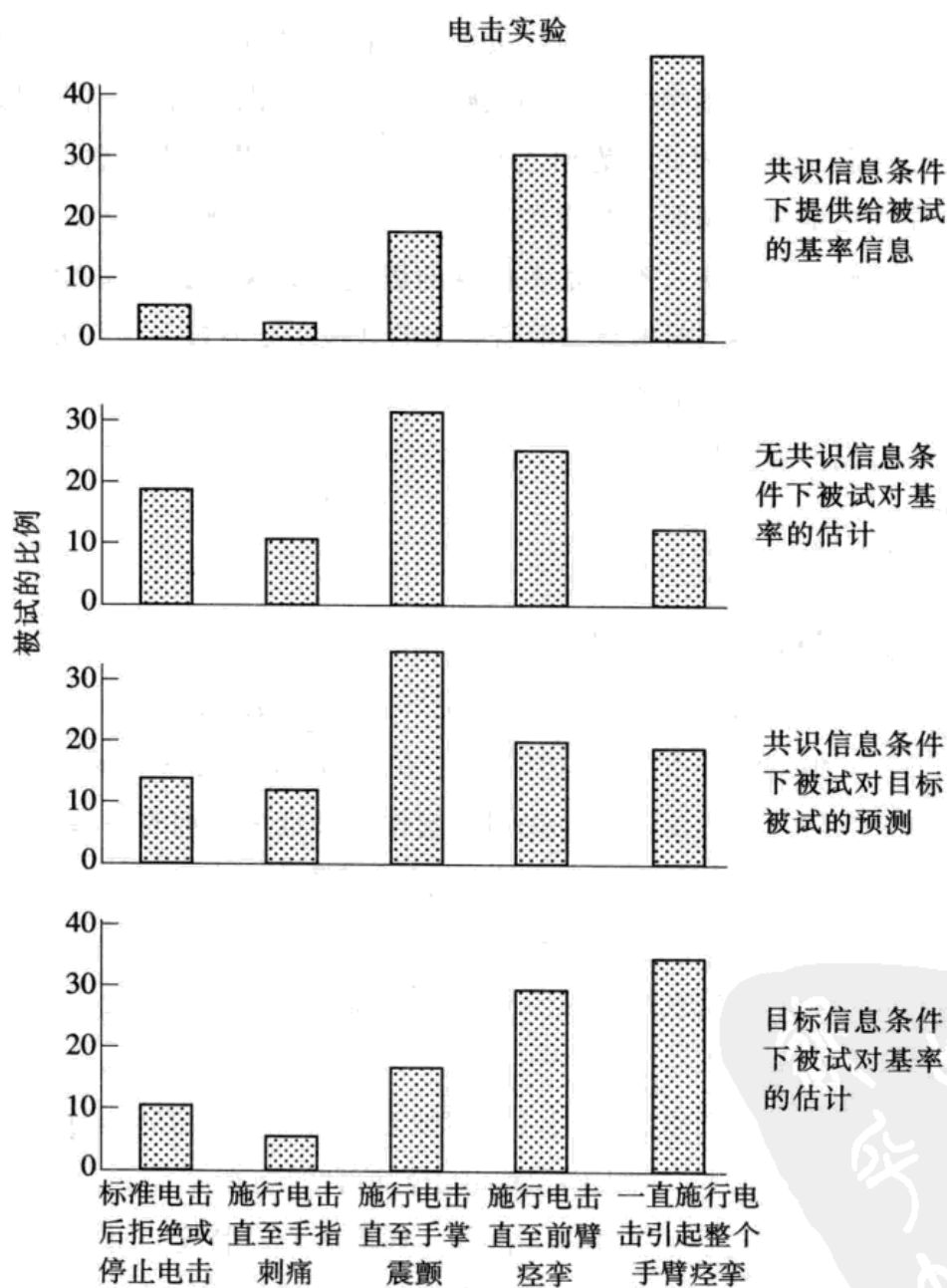


图7—1 电击实验中的基率、基率估计以及对目标被试的预测

110 惊人地相似。而对助人实验的预测虽然不像电击实验那般近似，但与实际基率的J型曲线也相当接近。如果样本挑选过程反复而鲜明地使他们注意到那两个被试是随机挑选的，而不是指定的，我们的被试就愿意推论所有被试的行为模式与他们所看到的两名被试的行为相类似。

111 总之，当对目标个体的行为进行预测时，他们并不利用基率信息。重要的是，在预测问题之外，我们也提及了一些归因问题，例如是情景因素还是个人偏好对目标个体的行为负责？共识性信息对这些归因问题没有实质影响。由于前者（基率信息）的失败，后者（共识性信息）的失败，实质上是不可避免的。因此，为何人在归因时会忽略共识性信息这个问题，就应还原为更基本的问题：即使在像预测这样简单的推论中，基率信息为什么会被抛弃？人为什么不受基率信息的影响，对这个更为基本问题的任何回答，都理当能够同时解释当前研究中的其他主要发现。这个发现就是，被试在被告知只有两名那样少数的被试实施了最极端可能的行为后，愿意假定这种极端行为就是普遍的行为模式，他们实际上是“过多地”受到目标个体信息的影响。

四、抽象信息对具体信息

卡尼曼和特沃斯基（1973，4），并没有详尽地思考人们不受基率信息影响的原因。他们的基本解释，似乎集中在人们不擅长于处理概率性数据。即使在赌博领域，人们知道机会法则在起作用，且至少拥有一些处理或然性的基本图式，但他们仍显示出明显的盲目和偏差。而在这种情景之外，人们则完全看不到像基率那样的概率性信息的相关意义。或者，因为不知道如何正确结合基率信息和目标个体信息，他们就只好简单地完全忽略基率信息。

当然，也有很多的事实证实，人们缺乏处理概率性信息的良好图式。实际上，在归因文献中，就人们忽略共识性信息的规则而言，有一个例外，良好图式的缺乏则能很好解释这一例外。如当人们对特定个体的能力进行归因时，就显然利用了成功与失败的基率信息（Weiner et al., 1971）。如果多数人在特定任务上都失败了，而目标个体成功了，那么目标个体就被

认为具有很强的能力；如果大多数人都成功了，而目标个体失败了，目标个体就会被视作能力较低。显然，我们每个人一生中都在依照他人的表现评价能力。能力，实际上，就其本质是一个导源于基率的推断。

然而，在我们看来还有另一个原则也可能在起作用。基率或者共识性信息，从本质上，似乎就是冷漠的、苍白的和抽象的。相比而言，目标案例信息，则是逼真的、醒目的和具体的。在抑郁研究中，我们试图让对枯燥的统计性信息的记忆与对现实世界刺激的逼真的紧迫的反应相对抗。在饼干品尝研究中，从瓶子中液体高度的证据中抽象出来的共识性信息，被用来与感官印象相比较。在描述电击和助人实验的研究中，表格式的频率数据，则对决于一个具体的人的录像或书面描述，而他有父母、职业计划、爱好以及个人癖好。尽管基率信息具有逻辑上的关联性，但它们也许就是缺乏激发更进一步认知活动的动力。 112

正如所见，上述假设并不新颖。在1927年，罗素就提出“普通的归纳所依据的，是事例的情感兴趣，而不是其数目”（p. 269）。在卡尼曼和特沃斯基的实验中，以及在我们自己和其他人关于共识性信息的实验中，纯粹的事例数目，与具有情绪兴趣的事例进行较量。与罗素的假设相一致，情绪兴趣在每个实验中都取胜。

我们可以推断，具体的、有情绪兴趣的信息会有更大的效力促生推论，因为这样的信息更有可能唤起包含了类似信息的具体图式。推论的进行，因此会依照预先存在的具体图式的习惯路线。比较而言，抽象信息与具体图式能够达到的连接网络之间，可能只有更少的潜在联系。与这些推测相一致，在电击实验和助人实验中，尼斯比特等（1975）发现，有关他人行为的共识性信息，不仅未能影响被试对他们如果参加这个实验会如何反应的预测，而且在实验后关于他们为什么会作如此预测的采访中，也没有一个被试提到此共识性信息。相反，被试抓住了实验场景中一些特别具体的细节信息，并将它们与自己过去经历中的类似情境相联系。“我相信我会去帮助发病者，因为我一个朋友的姐姐是癫痫病患者。”

罗素的假设，对于日常行动也有重要的内涵。一个简单的例子可以为例证。我们假定你想买一辆新车，而且基于经济和使用寿命方面的考虑，

118 你认为你需要一辆结实而坚固的中产阶级的车——沃尔沃（Volvo）或者萨博（Saab）。作为一个谨慎和明智的购买者，你求助于《消费者报道》，它告知你他们专家的一致意见是沃尔沃的机械性能更好，而且读者也一致认为沃尔沃有更好的维修记录。拥有这些信息后，你决定在下星期之前去跟沃尔沃的经销商商讨价钱。然而，在这段间隔时间里，你参加了一个鸡尾酒会，并跟你的一个熟人讲了这个决定。他对此极不相信，并劝告你说：“买沃尔沃？你在开玩笑！我妹夫曾买过一辆沃尔沃。起先，那个奇特的加油计数器坏了，花了250块。接着汽车的尾灯也开始出问题了，只好换掉。然后是变速器和离合器。不到三年就只好当垃圾卖掉了。”这个信息的逻辑地位是，给几百名拥有沃尔沃的《消费者报道》的读者的总人数增加了一位，让维修记录的平均频率在三个或四个方面提高了一点点。但是，任何人若坚持他将不受这种信息的影响，那他要么是不坦诚，要么就是缺乏最基本的自知之明。

研究6：抽象信息对具体信息的影响力

看起来将《消费者报道》操作化为思想实验是很有价值的（Borgida & Nisbett, 1977）。由于我们可资利用的大多数被试，都是密歇根大学心理系的学生，我们就选择密歇根大学的心理学课程作为消费商品。于是十门较高水准的讲座课程被挑选出来，它们在所称的授课质量上各不相同。一群准备加入心理学专业的低年级学生，由一个本身就是系里老师的实验者召集到一间教室里。实验者告诉他们，他现在为系里一个委员会工作，该委员会要为系里制定长期发展规划。而关于发展规划的一个问题，就是要决定将来会有多少人选择其中的哪些课程。他们被告知，为了得到有关计划招生人数的一些启发，需要他们为心理学本科的学习填写一张尝试性的课程表。

于是，我们要求控制组被试先浏览一份课程目录（实际上是一个包括了27门课程的真实课程表的模拟，只是排除了实验课、统计课以及通选课），然后要求他们在愿意选的5~10门课程旁边作上记号，并给他们认为自己一定会选的课程旁的记号画圈。

而两个实验组的被试，被告知为了帮助他们作决定，将提供给他们一

些额外的关于具有很高选课率的讲座课程的信息。对这两个实验组来说，这些额外信息中的一部分都是比目录式的介绍更为全面的对10门课的内容和形式的详细介绍。之后，让其中的一组被试（面对面条件组）加入一个由高年级的心理系学生组成的座谈小组。这些高年级学生依次对表上的10门课程进行简短评论。有1到4个人，通常是2个或3个人，对每门课进行评论。每个评论，都是先在下面五个词中挑一个，这五个词是：“优秀”、“非常好”、“好”、“一般”、“糟糕”，然后有一个概要的评价。下面是所有评论中的一个例子：

在这门课上尽管要读许多资料，但它们都会被清楚地介绍给你们。你们始终都会知道自己的进展，这对你们通过这门课程是非常有帮助的。这门课会向你们介绍心理学中一个非常广阔且重要的领域。但我为何不评定它为“优秀”而是“非常好”呢？原因在于那些资料并不能明显地启发学生思考。

在另一个实验条件下（基率条件组），我们让被试看一份对这门课的平均评价，它们是在所有选这门课的学生在上学期期末所填写的量表基础上得出来的。每门课程底下，有一个从“优秀”到“糟糕”的五个等级量表。在每一量表上，有个标记表示这门课程的平均评价，并注明参与评价的学生人数。这些人数从26人到142人不等。每门课程的平均评价，是由实验者设计的，以使它们与在面对面条件下的实验同谋所提供的平均评价价值相同。

这样的设计使得两种信息的有效性的比较成为可能。第1种信息，是被试亲耳所闻的推荐，亦即选修过这门课程的2或3名学生的简短评论；第2种信息，是他们得到的更稳定、更有广泛基础的信息。表7—1显示的，是这三组被试对推荐课程（平均评价价值是2.50或更高）、未推荐课程（平均评价价值是3.75或更低）以及未提及课程的平均选课人数。在每一类的平均人数旁边的，是加权的选择倾向。这是一个指数，没被选的课程加权值是0，被选课程的加权值是1，而画圈必选的课程加权值是2。

表 7—1 平均被选课程数以及加权的选择倾向

条件	推荐的课程		未被推荐的课程		未提及的课程	
	选择数目	加权的 选择倾向	选择数目	加权的 选择倾向	选择数目	加权的 选择倾向
面对面组 ($N=22$)	4.73 ^a	8.31 ^{a,b}	0.50 ^a	0.77 ^{a,b}	3.09 ^{a,b}	4.32 ^{a,b}
基率条件组 ($N=18$)	4.11	6.33 ^b	0.94	1.56 ^b	4.17 ^b	5.89 ^b
控制组 ($N=18$)	3.33 ^a	5.22 ^a	1.39 ^a	2.17 ^a	5.39 ^{a,b}	7.17 ^a
$F(2, 55)$	6.14 [*]	10.34 ^{**}	6.59 [*]	6.65 [*]	13.24 ^{**}	8.19 ^{**}

^a 标有 ^a 上标的栏在 Newman-Keuls 检验中在 0.01 显著水平上区别于其他栏。

^b 标有 ^b 上标的栏在 Newman-Keuls 检验中在 0.05 显著水平上区别于其他栏。

^{*} $p < 0.005$, ^{**} $p < 0.001$ 。

从表 7—1 中可以看出, 面对面的方式, 对课程选择具有更大的影响。面对面条件组的被试, 跟控制组的被试相比, 更倾向于选择被推荐的课程, 而更少选择未被推荐或未提及的课程。相比而言, 基率的方式, 仅仅影响到对未提及课程的选择。

也许有人会主张, 面对面条件组的被试, 比基率条件组的被试有更多的信息。这种主张的一种版本, 正是我们想提出的观点。学生们在作选择时, 就好像他们已经从由两个人所作的现实的评价中, 提取了比基率信息更多的信息。然而, 这个主张的另一版本, 则是面对面条件下的学生所作的评价, 包含了基率条件下所没有的真正有价值的信息, 如关于课程组织、评分程序或者老师亲和力等方面的信息。

为了应对后一种反对意见, 我们在对原实验进行了一个重要的改动后重复了该实验。我们给基率条件组被试同样提供了面对面条件下的实验同谋所作评论的书面副本。而且, 我们明确地指出, 这些评论是那些选了该门课的学生的代表性意见, 是从上学期末的所有评价中精选出来的。这种条件下的被试, 既了解整个学生群体稳定的平均评价, 也看到了实验同谋的逐字报告, 并且也知道这些评价具有高度代表性, 但他们在作选择时, 仍比面对面条件下的被试更少受到前述信息的影响。

五、和具体性动物沟通

不难看出罗素关于普通归纳的格言以及上面提及的例证, 对沟通和劝

说的一般问题所具有的意义。如果人们对这种让科学家及政策制定者所钟爱的呆板的统计数据无动于衷的话，社会及技术的进步就会受到阻碍，除非发展出有效的、具体的、具有情绪影响力的沟通议定方式。我们收集了几个有关劝说的“个案研究”，我们坚信依据抽象信息和具体信息的区分，对之能够更好地理解。我们在下面将它们罗列出来，希望它们能作为现实世界的灵感之源，并引导有关信息的本质及其劝说性影响的研究。

(1) 20 世纪 30 年代初期，由于农业技术的进步，早期的绿色革命成为可能。通过由县农业部门介绍统计数据以及发放政府传单的方式，政府及时地向全国的农场主介绍这些新技术。然后他们便停下来，期待传来粮食产量大量增长的喜讯。可是接下来，并没有出现这样的报道，而且马上清楚农场主们并没有转向这些新技术。于是，一些明智的政府官员着手一项计划，由政府农业组织直接进入到了被挑选出来的一些农场中，与农场主们一起利用这些新技术进行粮食耕作。邻近农场的农场主，看到使用新技术的产量效果后，立即转向新技术的应用中。 116

(2) 在癌症检测诊所，候诊者名单如同本文一样冗长，排到了几个月之后，而且从 1974 年秋季开始就一直是这种状况。这并不是由于美国医生协会 (Surgeon General) 或美国医学学会 (AMA) 或其他组织发布的最新统计数据造成的。这种冗长的候诊者名单，始于福特夫人 (Mrs. Ford) 和洛克菲勒夫人 (Mrs. Rockefeller) 的乳房切除之后。

(3) 克娄瑟 (Timothy Crouse) 1974 年写了一本关于 1972 年总统竞选的新闻报道的书，题目是《巴士上的男孩》。在这本书中，他提到在竞选日前一天，一大群追踪报道麦克戈文竞选活动的记者，都很明智地同意麦克戈文的失利不会超过 10 个百分点。这些人包括有线电视记者、网络电视记者以及一些主要的报纸和新闻杂志的记者。而他们也了解到所有的主要民意测验都认为麦克戈文会落后 20 个百分点，而且也清楚在 24 年里没有哪一个主要民意测验结果的偏差会超过 3%。然而，他们已亲眼看到成千上万的狂热的群众都支持麦克戈文。

(4) 纽约时报 (Kaufman, 1973) 最近对一个纽约地铁里的胡乱涂鸦者进行了采访。一个电火花点着了这个人作喷涂画的喷涂器，而他则在这

场大火中被严重烧伤。这个笔名叫“Ali”的男孩，承认在此事故的两星期前，曾经读过关于一个名叫 Bernard Brown 的男孩在火车上作喷涂画时被火车辗死的报道。他说：“也许如果我们知道他所用的名字‘Joe 146’，就会留下深刻印象。但是我记得我当时在嘲笑他，认为他肯定是一个不知道自己在做什么的家伙……”

我们相信，现在的研究以及这些日常生活的事例，揭示了某类信息通常被人们所忽略，而这些信息，在科学家看来，则是高度关联的，且具有逻辑必然性。而其他一些更少逻辑性的信息，却能够激发很强的推断和行为倾向。对那些研究信息加工的心理学家来说，最有用的研究，莫过于去发现他们的被试认为哪些信息是值得加工的。

本章是一篇文章的节选，它原刊于 J. S. Carroll 和 J. W. Payne 编的 *Cognition and Social Behavior*. Hillsdale, N. J. : Lawrence Erlbaum. 1976。允许重印。



不确定状况下的判断中的因果图式

特沃斯基和卡尼曼 (Amos Tversky and Daniel Kahneman)

我们所作的许多决策，无论是琐细的还是极其重要的，均取决于事件的显著可能性，如遵守承诺、一个企业的成功或对一项行动的反应。但由于我们通常缺乏适当的模型去计算此类事件的概率，对它们的估计就必然是主观的和直觉性的。近年来，人们以估计概率的方式评估证据，已经激发了许多的研究兴趣，如爱德华兹 (1968, 25)，卡尼曼和特沃斯基 (1979a, 30)，斯洛维奇 (1972a)，斯洛维奇等 (1977)，特沃斯基和卡尼曼 (1974, 1)。这类研究已经识别出几种判断启发式，而它们总与一些特征性的错误和偏差相关联。本章关注的是因果推理在不确定状况下的判断中的角色，以及与这种思维模式相关的一些偏差。

人们总是力图对身边的事件给出一个连贯的解释，而利用因果关系模式来组织事件有助于达到这个目的，这是心理学的一个常识。人们根据因果关系来

感知事件序列，即使感知者完全意识到事件间的联系是偶然的，导入其中的因果关系是虚假的。这种不可抗拒的趋势，由米休特（Michotte, 1963）的经典研究提供了令人关注的解释。而海德（Heider, 1958）的原创性研究，也突出了在对基本社会关系的感知中因果模式的普遍性。而因果归因研究，也是当代社会心理学的焦点之一（Jones et al., 1971; Ross, 1977）。

118 本章主要关注在不确定状况下的判断中因果模式的角色。尤其，我们要在一些证据或数据 D 的基础上，来研究对目标事件 X 的条件概率 $P(X/D)$ 的判断。要对证据的影响进行心理学的分析，有必要来区分判断者所感知的 D 和 X 间的几种不同的关系类型。如果 D 被认为是 X 发生或不发生的原因，我们就称 D 为原因资料。另一方面，如果 X 被视为 D 的一个可能原因，我们称 D 为诊断资料。例如，对 A 人格的描述，通常被认为是提供了预测他行为的原因资料，而对 A 的行为的描述，则为其人格提供了诊断信息。如果 D 既不是 X 的原因也不是它的结果，但它们都被视为另一因素的结果，我们称 D 为指征性资料。因此，如果两种行为都被认为是同一特质的表现形式，那么，一种情景中的行为，就为另一情景中的行为提供指征性资料。最后，如果 D 和 X 看来并没有直接或间接的因果联系，我们就称 D 为偶然性资料。

在对条件概率的规范处理中， D 和 X 的各种关系类型的区分并不重要，数据资料的影响，仅仅依赖于其信息量（informativeness）。与此相反，我们主张信息的心理影响，关键取决于它们在因果模式中的角色。我们还特别假定，原因资料比其他具有同等信息的资料，有更大影响；而且如果存在一个能引发因果图式的资料，不符合该图式的偶然资料，就只有较弱影响或根本无足轻重。

本论文的第一部分，我们比较了原因资料和诊断资料的效果，揭示出相对于具有同等信息的诊断资料而言，人们赋予原因资料以更大的影响力。我们还研究了一类问题，其中总有一个独特的资料同时具有原因和诊断意义，并且证明了在对条件概率 $P(X/D)$ 的直觉性估计中， D 对 X 直接的原因影响占有支配地位，而其诊断影响没有被给予足够的考虑。

原因和诊断推理

1. 推论的不对称性

因果图式，具有天然的演化进程：从原因到结果。因此，我们认为，依照此正常序列从原因推论结果，比颠倒此顺序由结果推论原因，更加自然，也更加容易。如果原因推论实际上比诊断推论更容易也更自然，那么我们可以预期，人们从原因推演结果比从结果推演原因会信心更强——即使结果和原因实际上提供了同等的关于对方的信息。我们用了两种不同的测量来检验这个假设：对条件概率的判断和对预测准确性的信心。

119

在一组问题中，我们要求被试在两种条件下比较一对事件 X 和 Y 的两个条件概率 $P(Y/X)$ 和 $P(X/Y)$ ，这两个条件是：(1) X 被自然地认为是 Y 的原因，(2) $P(X) = P(Y)$ ，即两事件的边际概率等同。后一个条件意味着 $P(Y/X) = P(X/Y)$ 。大多数被试会认为我们的预测是因果关系强于诊断关系，因而错误地坚持 $P(Y/X) > P(X/Y)$ 。

在另一组问题中，我们要求被试来比较他们在预测两个连续变量时的信心差异，而预测则依据哪个变量是给定的，哪个是需要预测的。同样地，问题的设计，是使其中一个变量自然地被视为另一个变量的原因。如果这两个变量有着近似的边际分布，就没有有效的统计理由来预期由一个变量预测另一个变量的准确性上存在着差异。然而，我们假定，多数被试将认为由原因推测结果比由结果推测原因具有更大的信心。

在这两类问题中，因果推论和诊断推论之间被预测的不对称性，都被观察到。其效果可用下面的问题来说明，括号中的数值表示选择每个答案的大学生（来自俄勒冈大学）人数。^[1]

问题 1：下列哪个事件更具有可能性？

- (1) 如果母亲是蓝眼睛，那么她女儿也是蓝眼睛。(N=69)
- (2) 如果女儿是蓝眼睛，那么她母亲也是蓝眼睛。(N=21)
- (3) 这两个事件有同样的可能性。(N=75)

问题 2：在对一个城市高年级中学生的调查中，比较男孩的身高

与他们父亲的身高。你对下面的哪个推测更确信？

(1) 从儿子的身高推测父亲的身高。(N=23)

(2) 从父亲的身高推测儿子的身高。(N=68)

(3) 同等信心。(N=76)

显然，在连续的代际中，身高或者眼睛颜色的分布基本上是相同的。为了证实被试对这个事实的觉知，我们询问了另一组 91 个被试，其问题是：在一个由母亲和女儿组成的更大的样本中，蓝眼睛的母亲所占比例相对于蓝眼睛的女儿所占比例是更多、相等、还是更少。结果选择更多的是 15 人，选择相等的 64 人，选择更少的 12 人。因此，尽管被试意识到这两者的概率是相等的，他们仍然判断“原因的”条件概率高于“诊断的”条件概率。

当然，严格来讲，父亲的身高并不是儿子身高的原因。然而，在惯常的用语中，说一个男孩高是因为他父亲有 6 英尺 4 英寸，是可以被接受的，而说父亲有 6 英尺 4 英寸是因为儿子高，则显然说不过去。更一般地讲，我们料想在两个变量间存在推论的不对称性，只要看起来用其中第一个变量推论第二个似乎比用第二个变量推论第一个更好。为了举例说明，请看下面我们提供给不同的两组被试的问题：

问题 3：以下哪一个描述更好理解？

(1) 汤姆很重是因为他很高。(N=63)

(2) 汤姆很高是因为他很重。(N=7)

问题 4：你对以下哪个推测更有信心？

(1) 由一个男人的体重推测他的身高。(N=16)

(2) 由一个男人的身高推测他的体重。(N=78)

虽然体重和身高不能被视为彼此的原因，但大多数被试，仍觉得用身高解释体重比用体重解释身高有更好的解释力。这也许是因为典型的高个

子男人会很重，而典型的重的男人却未必高。因此，绝大多数的被试，对由男人的身高推测体重给予了相比于由其体重推测身高更大的信心。当然，这个不对称性，无法得到统计学意义上的解释。

问题3和问题4，还暗示了即使在两个变量之间并不存在直接的因果关系，推论的不对称仍然会发生——倘若其中的一个变量（如身高）更自然地被认为能解释另一个变量（如体重）的话。下面的两个问题，涉及的都是两个变量都被认为是某个潜在特质的指征或表现的情况。但其中的一个变量，提供了对潜在特质更直接的体现或者更有效的测量。我们预计从强指征到弱指征的推断，将比其相反方向的推断更有信心。

问题5：下面哪个事件更具可能性？

(1) 如果一个运动员赢了十项全能中的第一项，他就能赢得比赛。
($N=21$)

(2) 如果一个运动员赢了十项全能，他就能赢得十项全能中的第一项。
($N=75$)

(3) 两事件有同等的可能性。
($N=70$)

问题6：对一大群学生进行了两项智力测验，一项是一个小时的综合测试，一项是十分钟的简短测试。你比较相信下面哪一个推测？

(1) 以学生在综合测试中的分数推测他在短测试中的分数。
($N=80$)

(2) 以学生在短测试中的分数推测他在综合测试中的分数。
($N=47$)

(3) 同样确信。
($N=39$)

同样，这里两个问题的正确答案仍是“相等”。在问题5中，一个运动员（非特指的）在十项全能比赛中获胜的先验概率是 $1/N$ ， N 代表参赛者数目。这也是一个非特指的运动员能够在第一项运动中获胜的先验概率。因此，这两个条件概率应该相等。在问题6中，线性回归的标准假定使从

一个测验推测另一个测验具有一定的正确性。但对这两个问题的回答，都显示出了明显的预测方向的偏好。

在问题5和问题6中，所包含的都是对同一潜在特质的两种指征，只是强度有别。在十项全能中的胜利和在一个单项运动中的胜利，都是杰出运动才能的表现，但前者比后者对杰出性有更强的指示力。同样地，在智力测验中的表现反映了其智力特质，而更全面的测验比缩减后的测验，能提供更好的测量。这些结果证实了前面的假设，即从强指征到弱指征的预测，比反向的推测会更有信心。

在前述问题中所观察到的推断的不对称性，可与特沃斯基（1977）研究的接近关系的不对称性联系在一起。经验研究显示，在相似性判断中，一个显著物体或原型与不太显著物体或变体之间的相似性，要小于变体与原型之间的相似性。例如，深红对浅红不如浅红对深红近似（Rosch, 1975）；“好”与“坏”之间的相似性，不如“坏”与“好”之间的相似性；一个地位突出的国家同一个不那么突出的国家之间的相似性，小于其反向的相似性（Tversky, 1977）。预测的不对称性，似乎遵守了同样的规则。因此，我们通常认为儿子更像父亲而不是相反，并以很大的信心把儿子的品性归于父亲而不是相反。因此，同样的过程，可能潜藏在相似性的不对称性和推论的不对称性背后。

2. 证据的因果意义和诊断意义

前一节的讨论，显示了因果资料对结果概率判断的影响大于诊断资料对原因概率判断的影响。本节要研究的问题是，在考虑到目标事件的时候，某证据同时具有原因和诊断两种意义。我们研究的假设是，人们倾向于关注资料对将来的因果影响，而倾向于忽略了它们对于过去的诊断性意义。
122 我们首先讨论一类问题，在其中，对原因的考虑相对于诊断性的考虑占据了优势，而导致不一致甚至矛盾的概率估计。下面这一对问题，是图罗夫（Turoff, 1972）在关于预测的交叉影响方法的讨论中提到的。

问题7a：下面哪一个的概率更高？

(1) 如果在以后的五年里由于汞污染导致的死亡人数超过500人

的话，议会将在这五年内通过一项控制汞污染的法令的概率。

(2) 如果在以后的五年里由于汞污染导致的死亡人数没有超过 500 的话，议会将在这五年内通过一项控制汞污染的法令的概率。

问题 7b：下面哪一个的概率更高？

(1) 如果议会在五年内通过一项控制汞污染的法令，这五年内由于汞污染导致的死亡人数将会超过 500 的概率。

(2) 如果议会在五年内没有通过一项控制汞污染的法令，这五年内由于汞污染导致的死亡人数将会超过 500 的概率。

我们用 C 表示在五年内议会将通过一项控制汞污染的法令，用 D 表示在五年内汞污染导致的死亡人数超过 500。然后用 $\bar{C}$ 和 $\bar{D}$ 分别表示 C 和 D 的反面。

大多数被试认为死亡人数超过 500 将比不超过 500 更有可能使议会通过限制汞污染的法令，也就是 $P(C/D) > P(C/\bar{D})$ 。多数人也认为这项法令在五年内通过，相对于不通过而言，死亡人数更不可能达到 500，也就是 $P(D/C) < P(D/\bar{C})$ 。这些判断反映了因果信念，即高的死亡人数将增加实施一项反污染措施的压力，而这样的措施又将会有效控制汞污染。在一个由 166 名学生构成的样本中，有 140 人对两个问题都选择了上述的回答。这种看起来似乎合理的判断方式，违反了最基本的条件概率原则。

很显然， $P(C/D) > P(C/\bar{D})$ ，意味着 $P(C/D) > P(C)$ 。

而且，若不等式 $P(C/D) = \frac{P(C \& D)}{P(D)} > P(C)$ 成立，

那么 $P(C \& D) > P(C) P(D)$ ，

由此进一步得出必须 $P(D) < \frac{P(C \& D)}{P(C)} = P(D/C)$ 。

这样反过来意味着只要 $P(C)$ 和 $P(D)$ 不为 0，那么 $P(D/C) > P(D/\bar{C})$ 。

因此， $P(C/D) > P(C/\bar{D})$ 意味着 $P(D/C) > P(D/\bar{C})$ ，这与通常的判断模式相矛盾。

很容易找出其他的同类例子，来证明人们的直觉违反了概率运算。这样的例证，通常由一对事件 A 和 B 构成，B 事件的发生增加了其后 A 事件出现的可能性，而事件 A 的出现又降低了随后 B 事件发生的可能性。为了举例说明，请看下面的问题。

问题 8：用 A 表示在明年年底之前，彼得将会在他家里安装一个防盗警报系统。用 B 表示在明年年底前彼得家将会遇窃。用 $\bar{A}$ 和 $\bar{B}$ 分别表示 A 和 B 的反面事件。

问题： $P(A/B)$ 和 $P(A/\bar{B})$ 中，哪一个条件概率更高？

问题： $P(B/A)$ 和 $P(B/\bar{A})$ 中，哪一个条件概率更高？

大多数被试（161 人中的 132 人）选择了 $P(A/B) > P(A/\bar{B})$ 和 $P(B/A) < P(B/\bar{A})$ ，有悖于概率规律。我们将此判断模式解释为因果分析比诊断分析占优势的另一种表现。为了认识这种效应的实质，让我们分析一下问题 8 的结构。

首先，考虑 $P(A/B)$ ：如果彼得家在明年年底前的某个时候会遇窃，他将在此期间给家里安装一个防盗警报系统的条件概率。防盗警报系统的安装，或在受盗之前，或在之后。条件中所携带的信息，也就是对遇窃的假设，对于将来有原因的意义，而对于过去又有诊断的意义。具体而言，盗窃的发生，提供了之后安装防盗警报系统的原因，同时又提供了在遇窃之时房子没有安装警报系统的诊断性提示。因此，盗窃的影响增加了安装警报系统的可能性，而它的诊断影响又降低了其可能性。几乎是一致的 $P(A/B) > P(A/\bar{B})$ 的判断，显示出 B 的原因影响比其诊断性影响占优势。

同样的分析也适用于 $P(B/A)$ ：如果在明年年底前某时彼得给家里安装一个警报系统，他家在此期间会遇窃的条件概率。警报系统的存在，具有原因性的影响，以减少之后遇窃的可能性；它也会提供一个诊断性的暗示，即盗窃的发生，可能促使了彼得安装警报系统。警报系统的影响减少了遇窃的可能性；而警报系统的诊断性影响增加了这个可能性。这里，绝

大多数对 $P(B/A) < P(B/\bar{A})$ 的判断, 又一次证实了 A 的原因影响相对于其诊断性影响的优势地位。人们显然首先是依据条件的直接原因影响来估计条件概率 $P(A/B)$ 和 $P(B/A)$ 的, 而不对此事件的原因和诊断性影响都加以权衡, 这样必然导致在这类问题中出现矛盾。

图罗夫的问题中的一个明显特征, 是条件事件和目标事件之间时间关系的不确定性。然而, 即使不存在时间的不确定性, 也经常会有条件事件同时具有原因和诊断性意义的情况。现在要分析的假设是: 即使在两事件的时间关系确定的情况下, 主导条件概率估计的仍然是原因分析。

问题 9: 下面的哪一个概率更高?

$P(R/H)$ 假如 20 世纪 80 年代利用太阳能进行室内供暖有显著增长, 90 年代美国将对个人消费者实施燃料配给的概率。

$P(R/\bar{H})$ 假如 20 世纪 80 年代利用太阳能进行室内供暖没有显著增长, 90 年代美国将对个人消费者实施燃料配给的概率。

按照我们对原因和诊断性考虑的区分, 用贝叶斯的标准方式来考虑这个问题将会有所启发。20 世纪 80 年代利用太阳能进行室内供暖将有显著增长这一事件 H, 既有原因的也有诊断性的意义。它对事件 R 的直接原因影响显然是否定性的。在其他情况不变的条件下, 使用太阳能的明显增长, 只会减轻以后年代里的燃料危机。但是, 80 年代太阳能使用的显著增加, 也提供了强有力的一个对即将发生的能源危机的暗示。尤其是, 它暗示着 80 年代的燃料价格已非常高, 致使对许多消费者而言投资太阳能进行室内供暖会更为经济。而 80 年代高的燃料价格, 反过来又揭示了一个矿物燃料短缺的状态, 这个状态增加了在之后的年代里进行燃料配给的可能性。因此, H 对 R 的原因影响, 减少了 R 的可能性, 而它的诊断性含义又间接增加了 R 的可能性。

虽然原因的和诊断性意义的相对强度问题, 还没有正式解决, 我们仍相信事件 H 的诊断性意义有可能比它的原因影响更重要。由于使用太阳能供暖的增长而节省下来的燃料数量, 不可能大到足以制止即将发生的危机。

125 另一方面, H 所暗含的燃料不足, 对即将到来的能源危机是一个强有力的提示。按照这个推理路线, $P(R/H) > P(R/\bar{H})$, $\bar{H}$ 是 H 的对立面。

然而本节的假设是, 人们通常在条件概率估计中, 将条件事件的直接原因影响看得过重, 而没有给予条件事件的诊断意义足够的重视。这个假设在问题 9 中的体现, 就是 80 年代使用太阳能供暖增长这一条件减小了对 90 年代进行燃料配给的概率判断。实际上, 83 名回答者中的 68 名都认为, $P(R/H) < P(R/\bar{H})$ 。同样的判断模式, 还可以见诸同类的其他问题中, 其中条件的直接诊断意义与它的直接原因意义相冲突。虽然这个判断模式没有像图罗夫的问题一样违背概率原则, 但我们认为它反映了在条件概率判断中忽略条件事件的诊断意义的惯常趋势。

3. 预测、解释和修正

在前面的几节里, 我们列举了一些证据来支持原因推论比诊断推论具有更大效力的假设。首先, 我们揭示了从原因推论结果比从结果推论原因更有把握。其次, 我们指出当一个同样的信息同时具有原因和诊断性意义的时候, 前者通常比后者在条件概率判断中占有更大优势。

现在我们转向一个更普遍的问题, 即一个系统的意象、模型或者图式 (比如能源情景或某个人的人格) 与该体系的结果或表现 (比如太阳能使用的增长或敌意的表现) 之间的关系。模型或者图式, 通常被用于预测和解释结果, 这个结果反过来又被用于修正或者更新原模型。因此, 一个人可以用这个模型去预测事件的结果或估计其概率; 也可以用它解释某特定事件的发生或其结果。最后, 他还可以用某特定事件的发生所提供的信息, 去改正或者修正他原有的模型。

预测和解释代表了两种不同的原因推论, 而模型修正则是诊断性推论的例证。在作预测时, 判断者力图选择与他所采用的系统模型最一致的结果; 在进行解释时, 判断者从模型中辨识出最有可能引起某特定结果的特征。另一方面, 当需要对模型进行修正时, 判断者更改或者完善模型里与资料最不一致的元素。

126 日常生活中的大多数推论, 都基于不准确、不完备甚至有时候是不正确的模型或图式。不过, 人们已经意识到这一点, 而且经常欣然地承认他

们的系统模型（如某人的意图或者能源情景）可能是错误的。这种有关模型精确性的不确定性的存在，对于合适的预测、解释和修正，有特定意义。如果一个模型可能出错，由此模型导出的推测，将是适度的或者保守的，也就是说它们不应太偏离基率预测。例如，如果某人对另一个人的了解，出自一个并不可靠的信息来源，相比于同样信息来自一个比较可信的渠道的情况，他会更不愿意预测这个人是否会做出少有或不寻常的行为。

基于不确定模型的解释，也应被谨慎地加以调整，因为用于解释的因果因素，可能在现实中并不存在。而且，在不确定状况下的解释，经常与模型的修正结合在一起。例如，若一个人的某种行为与我们对他人格的印象明显不符，我们就会认真地考虑是不是我们对他的印象不正确，原有印象就会按照这个新信息所暗示的方向进行修改。模型的不确定性越大，其行为越不寻常，对模型的改动就会越多。一个恰当的解释，应该将要解释的事件所暗含或揭示的模型的变化考虑进去。因此，就规范的观点而言，不确定情况下的解释，包括了诊断性的和原因的这两种推论。

早先的研究发现，人们通常由高度不确定的模型进行过分预测。比如被试会依据对某人人格的一个简短描述，非常确信地推测他的职业选择或学业表现，即使这个人格概述的来源并不可靠（Kahneman & Tversky, 1973, 4）。从人格概述中推导出的某人的意向和特质，被自然地视为他的职业选择或学业上的成功等结果的原因。在这类问题中所发现的过度预测，与前面所例证的原因性资料的强影响是一致的。

在解释和修正的语境中，原因推论的强势地位和诊断性推论的劣势地位，体现在人们更易于对他们难以预测的结果构建原因性说明，而不容易对不确定模型进行修正以调适新的信息。将一个新的事实同化到已有的原因模型中，似乎比依照此事件修改原模型，要容易得多。而且，用以调适新事实的修正，通常只限于在极小的范围和局部的特征上。

为了举例说明这个观念，我们转向以前所作的有关直觉性预测研究中的一些尚未发表的结果（Kahneman & Tversky, 1973, 4）。在那个研究中，给 114 名心理学专业的学生看了一段对研究生汤姆的描述，并声称这段描

127

还是在他念中学高年级时由一个临床心理学家在投射测验基础上所写的。下面就是这段描述：

汤姆有很高的智商，尽管他缺乏真正的创造力。他有着对秩序、清楚、整洁和所有部分各司其职的整齐系统的渴望。他的文章相当枯燥和机械，只是偶尔的陈腐的双关语和科幻小说式的想象迸出的火花才令其生色。他有强大的动力去胜任工作。他看起来对他人毫不在意，缺乏同情心，也不愿与人交流。尽管以自我为中心，但他有很强的道德感。

首先要求被试推测的是汤姆的研究生专业，需要他们对九个备选专业按照其可能性排序。结果被试中的绝大部分人认为汤姆最有可能学的是计算机科学或工程技术，最不可能的是社会科学和社会工作或者人文科学和教育。对一个附加问题的回答，也显示出一个基本的共识：投射测验为职业选择的预测，并不能提供一个有效的信息来源。在完成预测任务后，被试被要求回答下面的问题。

实际上，汤姆是一个教育学院的研究生，他已经加入了一个为残疾儿童的教育而开设的专门培训计划。请简要勾勒你认为最能解释汤姆的人格和他的职业选择之间关系的理论。

回答这个问题的正确路径是什么呢？被试面临一个明显的冲突，在一个确定的事实即汤姆的职业选择和对他的一个详细但并不可靠的人格描述之间的冲突。人们从人格描述来预测他的职业选择时的十足把握，意味着他们相信在人格和职业选择之间存在着高相关。这种信念，依次使职业选择对于人格有很强的诊断性意义。在上述例子中，汤姆的职业选择，从对他的人格描述来看是不可能的，而且这个描述的来源信度很低。因此，一个合理的诊断式推论，将导致对汤姆性格印象的实质性的修正，以使它与汤姆所选职业的刻板观念相容。如果一个人相信特殊教育专业的学生通常

都是富于同情心的，那么汤姆的职业选择，会导致对心理学家的描述中所说的他“缺乏同情心，也不愿与人交流”产生怀疑。对此问题的一个恰当的反应，就是至少应提高汤姆的人格并不像描述的那样的可能性，并且提高实际上他要比描述中介绍的更亲切、更仁慈的可能性。

但我们的被试并没有采用这种解决路径。一小部分人（占21%）甚至完全承认原描述的可靠性。绝大多数被试（包括那些怀疑者）为了解决冲突，要么选择汤姆人格描述中合适的片段进行推论（如他有很强的道德观念），要么重新解释他的职业选择的心理意义（如这是他渴望支配权的表现）。128

可以论断被试没有修改他们对汤姆的印象，仅仅是反映了他们所要完成任务的要求特征，也就是“解释汤姆的人格和他的职业选择之间的关系”。按照这种说法，这个任务被自然地理解为要求他们尝试着将汤姆的职业选择与他的人格描述联系起来，而不去质疑该人格描述的可靠性。但是，尽管对他的人格描述深表怀疑，被试却似乎把它视作非常可靠的信息以处理对汤姆的印象。我们认为这种趋势，为一个更普遍的现象提供了例证。这个现象就是：即使用以解释的模型高度不确定，人们也倾向于不修改模型。

在我们看来，被试的回答，显示了不愿意和舒适两方面的特征：无论模型多么不确定，人们都不愿意修改一个内容丰富内在连贯的模型；以及他们非常舒适地用这样的模型去解释新事实，而不管该事实是多么的出乎意料。我们惊讶于被试在发展对汤姆意料之外的职业选择的因果解释时所表现的流畅自然，而且没有理由不认为他们在解释汤姆令人惊讶的其他行为时，也会同样地轻而易举。

高度发展的解释技巧，也许要归功于人们在面临不相容的证据时，他们脑中的印象、模型、概念以及范式的牢固和稳定（Abelson, 1959; Hovland, 1959; Janis, 1972; Jervis, 1975; Kuhn, 1962）。修改一个模型的推动力，只能来自他们对此模型和新事件的不一致性的承认。如果只对已有观念作少量而且局部的修改就能令人们满意地解释大多数事情的发生，那他们几乎不会觉得有需要对那些观念做彻底的修改。这样讲的话，原因的

思考，阻碍了诊断性修正的过程……

本章是以下书中某篇论文的第一部分，M. Fishbein 编： *Progress in Social Psychology*. Hillsdale, N. J. : Lawrence Erlbaum. 1980。允许重印。

注释

[1] 基于以色列的样本，本章初稿报告了不真实的低比例的“同样”反应。我们感谢冈萨雷斯（Gonzalez, A. M. B）和库伯威（Kubovy, M. ），他们使我们注意到这个问题。现在的数据，是在俄勒冈大学收集的。它和冈萨雷斯在耶鲁大学收集到的数据一致。



归因过程中的缺陷： 错误的社会评估的起源和维持

L. 罗斯和安德森 (Lee Ross and Craig A. Anderson)

一、归因理论和归因错误引论

1. 归因理论和直觉心理学

从最广泛的意义上讲，归因理论所关注的是普通人企图理解他们亲眼目睹的事件的原因及其隐含意义的各种尝试。它处理的是人们在对自己的行为及他人的行动进行解释时的“朴素心理学”。由于归因理论正在社会心理学中处于优势地位，这就使长期以来关于提升本学科对于人的观念的斗争达到一个顶峰。心理学上的人，不再是极端行为主义者所认为的刺激—反应 (S—R) 式的机器人，它被提升到了信息加工者和认知协调寻求者的层次之上，而最终被赋予与研究他们的科学家同等的地位。从归因理论的角度来看，人是直觉心理学家，他们力求解释行为并对行动者和他们的社会环境进行推论。

为了更好地理解这些直觉心理学家的感知和行为，我们必须探求他们的方式。首先，如同专业的心理学家一样，他会受一些关于人的本性和行为的隐含假设的引导——如追求快乐和逃避痛苦是普遍存在的且强有力的人类动机，或者遵从其同伴的心愿和期望相对于不遵从这些期望是更正常的，并且更不需要作进一步的解释。这些非专业的心理学家和专业心理学家一样严重依赖于数据，尽管这些数据从随机性和代表性上考虑很难满足正规的需求。有时候这些数据来源于第一手的实验；但更为通常的是，它们来自非正规性的社会交流、大众传媒或其他间接来源。非专业的心理学家，必须进一步采用或者发展出编码、存储以及提取这些信息的技术。最后，他还必须采用各种总结、分析和解释信息的策略——也就是说，他必须应用能使他建立新推论的一些规则、启发式或图式。相应地，这个直觉心理学家控制他的社会环境的能力，将依赖于他的假设、证据和分析的精确性和充分性。反过来，任何现存理论中的系统性错误都可利用数据中的偏差或者分析方法的不足产生严重的后果。这个后果既影响到非专业心理学家个人，也将影响到他所构建和维系的社会。从当代归因理论的优势地位角度来探究的这些缺陷，就是本章的焦点所在。

归因理论的主要轮廓由海德（1944，1958）最早勾画出来，并由琼斯和戴维斯（Jones and Davis, 1965）、凯利（Kelley, 1967, 1971, 1973），以及他们的同事（Jones 等, 1971；Weiner, 1974）将其发展得更为精致、具体。这些理论家主要关注的是社会观察者所面临的两个紧密相连的任务。一个任务是关于因果判断的：观察者力图辨别出能最合理地将一些特殊的结果（如某些行动或者结果）归结于其上的原因，或者是一组原因。另一个任务是关于社会推论的：某事件的观察者会对相关行动者的品质（attributes）或者品性，以及行动者所应对的情境的特征或者属性（properties）进行推断。

因果判断和社会推论这两个任务，一直都是大量的理论和经验调查的主题，而且直至今日，它们实际上占据了归因理论的整个领域。然而近来直觉心理学家的第三个任务，开始受到某种关注，这个任务是对结果及行为的预测或者估计。直觉心理学家不仅必然会寻求解释，对品性进行推论，

他还必定会对目前未知或将来可能发生的行动和结果产生期盼，并进行猜测。例如，当一个总统竞选者承诺将“减轻普通纳税人的负担”，我们就可能会考虑他作这个声明的原因，以及它对于竞选者个人品性的某种暗示（这个承诺是否仅仅反映了某种政治上的权宜之计？我们能由此推断出竞选者的任何真实信念吗？）。但是我们也可能会推测他接下来的行为以及他对于那些没有被考察到的相关问题的看法（如果选举成功，他会削减财产所得税吗？他会支持削弱社会福利计划吗？）。简而言之，关于直觉预测的心理学，是归因理论领域的一个自然的扩展。

当然，无论从逻辑上还是心理学上讲，这三个归因任务都是相互依存的。对一个事件的解释，与对在该事件中扮演角色的行动者和组织实体的推论是密切联系的。而它们又共同提供了对目前未知或在将来有可能发生的事件本质进行推测的基础。然而，每一个任务又提供了揭示直觉科学家的表现背后隐含的假设和策略的独特可能性（以及一些关于解释和方法论的独特问题；见 Ross, 1977, pp. 175 - 179）。值得一提的是，近几年来，在对普通人的推论所进行的研究中，将估计和预测用作因变量已经越来越流行。造成其日益流行的其中一个原因是尤其重要的。与感知者对一事件进行分析后所作的因果判断和品性推论不同，对新的或未知事件的估计或预测，通常能依据它们的准确性进行评价。也就是说，一个人能将他对事件的预测和估计与实际观察或测量结果进行比较。这就使得有可能评估直觉科学家采用的归因策略的适当性，以及他产生特定错误或偏差的方向。

2. 逻辑归因原则对自助（self-serving）偏差

当代归因理论追求的是两个截然不同但互相补充的目标。一个目标，是要揭示出社会感知者进行估计和推论时通常会遵从一些逻辑的或理性模式的指引。另一个目标，则是要对会歪曲那些通常是真实的估计和推论的偏差或错误的来源进行阐明和解释。我们首先将简短地关注为直觉心理学家所采用的所谓的逻辑的或理性的图式，然后在本章的余下部分集中考察他在试图理解、预测和控制即将发生事件时所犯错误的来源。

(1) “共变”原则和“折损”原则

在很大程度上，人们必须具有一个共同的、对影响他们的社会行动和结

果的理解，因为如果没有这种共识，社会的互动将会混乱而不可预测，并且脱离参与者的控制。通过归因理论家们内省的方式和一些实验证据的支持，已经得出对在行为和结果解释中经常采用的一套“规则”的假设。这些“常识”性的规则或者图式在某些方面，非常类似于社会科学家和统计学家在对数据进行分析和解释时所遵从的更为正规的规则和程序。

182 凯利和琼斯，以及他们的同事，已经区分出逻辑规则和图式会被应用到的两种情况。在多重观察情形中，归因者将有机会得到以行或列的方式表示的关于行为的信息，它是一个行动者 $\times$ 对象 $\times$ 环境（或场合）的反应矩阵。特别的是，在这个研究领域中，参与者所能够得到的是概括性的陈述而不是实际的反应。因此潜在的归因者得到的是如下这样的信息：“大多数戏剧爱好者都喜欢这部新的品特派戏剧”，或“玛丽无法拒绝迷失的动物”，或“安看的唯一的一个电视节目是名著剧场（Masterpiece Theatre）”。而在单一观察情形中，归因者所面对的是单独的一个行动者在个别场合下的行为。例如，他可能看到山姆顺从一个实验者的要求，对同伴施以痛苦的电击；或者他得知“路易用他所有的钱在赛马上进行了大风险的投注”。

在这两种情形中，对归因进行调节的逻辑规则或原则是截然不同的（Kelley, 1967, 1971, 1973）。在多重观察情形中，归因者应用的是“共变原则”；也就是，他会估计已观察到的行为或结果，在所考虑的每一个候选原因出现时发生而在其缺失条件下不发生的不同程度。按照这个原则，若新品特派戏剧在很大程度上受到众多戏剧爱好者的喜爱，并得到那些很少赞赏戏剧的人的称赞，以及它在其上映的第九十天与第九天一样受到热烈欢迎的话，这个归因者就会得出新品特派戏剧是优秀戏剧的结论。

在单一观察情形中，归因者所采用的估计策略包括“折损原则”。据此，归因者在解释一事件时，若能发现其他可能的原因或决定因素的话，就会“折损”任何已有的候选原因。这个归因原则，用社会推论的术语重述，会比用因果归因描述更好：若环境或外在的因素能对一个事件进行“充分”的解释，那么这个事件将被归因于该环境，且不会再作关于行动者品性的某种逻辑推论（而且也可能不会作任何经验上的推论）。相反，如果一个行动或结果看起来在该环境因素不存在的情况下也会发生，因此

不是由于现场的环境影响，那么相应的事件就会归因于行动者，并进行“对应推论”（correspondent inference）（Jones & Davis, 1965），即归因者会推断是行动者的某些性格、能力、意图、情感或其他品性的存在和影响导致了行动者的行动或某种结果。因此，若一些诸如令人激动的小道消息（a hot tip）、让人绝望的金融危机或者下注前的七瓶马提尼酒等因素被提及的话，我们就不会认为路易在赛马上的高风险性下注反映了他稳定的个人特征。反之，若我们知道他在面对如果再一次在跑马上输掉薪水那么妻子就会离开他的威胁时，在他知道如果输掉就无法偿付租金的情况下，以及偷听到赛马行家关于赛场上可望获胜的马“并不为获胜几率所能预测”的评论之后，仍继续下注，那么我们会判断路易是一个积习成癖的高风险性赌徒。

值得一提的是，这两种不同原则的应用，对直觉科学家们提出了迥然不同的要求。共变原则，需要归因者采用在本质上基本是逻辑性的或统计性的规则，并且不需要对所考察实体的特征作进一步深入的了解。与之相对应，折价原则的应用，则需要对此人的本性以及如经济需求、喝酒、妻子的遗弃威胁等环境因素的影响进行相当深入的了解。从某种意义上讲，共变原则能被一个纯粹的“统计学家”所应用，而折价原则则需要一个能够估计各种社会压力和环境因素，甚至能区分有意行为和非有意行为结果（比较 Jones and Davis, 1965）的“心理学家”。

关于常识性归因原则的系统运用的证据，主要来源于一些问卷调查，在这些调查中被试会阅读关于一个或多个行动者在特定情境下对一些特定的对象或者“实体”作出反应的简短情节描述，并对之进行解释（如 L. Z. McArthur, 1972, 1976）。较小范围的特殊场合的研究，使得归因者所面对的似乎是可信的一些反应、遭遇和结果（如 Jones et al., 1961; Jones & DeCharms, 1957; Jones & Harris, 1967; Strickland, 1958; Thibaut & Riechen, 1955）。这类研究显示归因者们能够，而且通常会适量应用到一些假定的原则或拇指法则（rules of thumb）。现在所应用的研究方法论，仍然无法解释清楚的是普通人归因准确性的程度以及他们的归因错误的幅度和方向。

(2) 归因中的自助动机偏差

在考察一个不同的逻辑归因体系中可能存在的歪曲时，理论家们很快假定出“自我保护”偏差，通过这种偏差，归因者得以保持或者提升他们惯有的自尊或者关于他们自己特有的品性和能力的肯定性评价（Heider, 1958; Jones & Davis, 1965; Kelley, 1967）。证明这样一个动机性偏差存在的尝试，通常也揭示出在对积极和消极结果的归因中存在着不对称性——更确切地讲，就是行动者趋向于将“成功”归因于他们自己的努力、能力或者品性，却将“失败”归因于运气、任务的难度或者其他外在的因素。成就任务（如 Davis & Davis, 1972; Feather, 1969; Fitch, 1970; Wolosin et al., 1973）和教学表现（如 Beckman, 1970; Freize & Weiner, 1971; Johnson, Feigenbaum & Weiby, 1964）实验，已经提供了关于不对称性的海量证据。而且还显示出相对于某行动的旁观者而言，其行动者可能给予自己的成功以更多的称赞，而更少对失败加以责备（Beckman, 1970; Gross, 1966; Polefka, 1965）。

134 然而，对动机性偏差的广泛存在持怀疑态度的批评者，很容易就对这些研究提出了挑战（见 Miller & Ross, 1975; Nisbett & Ross, 1980, 第10章，对此进行了详细讨论）。对意欲揭示动机性偏差的研究所作的一个最有力的反驳，是在被试个人内在的感知与他们的公开评判之间进行了明显的区分。我们很容易设想一个情境，在此情境中，一个人会公开否认（或承认）他对于某事件的责任，而在私下里他则会承认（或者否认）这是他的责任。虽然从保持某人公众形象的意义上讲，这些公开的评判是自助性的，但若从维护个人自我形象的意义考虑，它们并不意味着自我保护偏差在起作用（Mileer, 1978）。

而且，对成功与失败进行自我归因（它们是否能用于研究者身上？）中的不对称性，以及行动者与旁观者的判断上的差别，有可能反映了其他非动机性的偏差来源。正如几个研究者所指出的，至少在考试的情境下，成功可能是被期待的，并且与行动者过去的经历相符，而失败则在意料之外，也是不常见的。同样地，成功的结果是人们所期待的，也是行动者的计划和行动的目标，而失败则是发生在行动者的计划和努力之外的未预料

事件。而旁观者，则很少真正了解他们所看到的结果的行动者过去的经历以及他们目前的种种期待和意图。

对存在普遍的自我保护偏差的观点的挑战，既有经验上的，也有概念上的。例如，一些研究发现被试显示出的是“反自卫性”或者自尊衰减性偏差。比如，罗斯等（1974）通过采用一个非常可信的教员—学生范式，发现教员将自己的表现和能力评定为失败而不是成功的更为重要的决定性因素。反之，他们认为其学生的努力和能力则是成功而非失败的更为关键的决定因素。同样的研究还证明与那些缺乏经验的大学生相比，这些看起来是反自卫性的归因趋向，在职业教师中甚至更为显著。这个结果与自我防卫理论的一个明显推论相抵触，即那些受到失败经历最直接威胁的人将最具有自卫性。

当然，坚持自我保护动机性偏差存在的研究者，能对那些似乎显示出不存在动机性偏差或者存在反自卫性偏差的研究提出替代性的解释（比较 Bradley, 1978）。实际上，动机性偏差的支持者和怀疑者之间争论的许多方面，都让我们回想起一些较早的在学习理论和基础知觉中间进行的更广泛的争论。在这些争论中，寻求一个关于动机性影响问题的“决定性”实验（即一个无法从“另一方面”作出解释的实验）的努力，在资料收集得越来越多，而概念分析也日益深入的时候，却越发显出其毫无益处。

对这样一种事态的一种反应，就是暂时放弃动机性概念，而把注意力集中于那些能影响并可能歪曲因果判断的非动机性因素（如信息的、知觉的和认知的因素）。除开已经存在的概念上的困难、混杂的经验研究结果以及使寻求动机性偏差存在的研究者气馁的历史教训外，还有其他两个促成当前这种转换的原因。第一，我们越来越确信对非动机影响的充分认识，将有助于我们理解并预期在何种情境下对责任的归因可能提升归因者的自尊，而在何种情境下这样的归因又可能削弱他的自尊（比较 Miller & Ross, 1975）。第二，我们也逐渐认识到准确的归因比不准确的归因更倾向于“自助”——也就是不管某种不准确的感知及影响的直接后果是多么令人愉快，归因判断的失真，更易于使有机体难以准备好应对长期的生存任务。

本文接下来的部分，将关注有限的几个非动机性偏差（见 Nisbett &

Ross, 1980, 有一个更为全面的回顾)。而且我们也会讨论一个将增加这些偏差“代价”的普遍现象——错误的印象、判断甚至更广泛意义上的理论,在面对与它们相抵触的逻辑严密的数据资料时仍会继续存在的趋势。不过,让我们在刚开始时就认识到,本文所探讨的这些错误和偏差,在直觉科学家的角度来看,并不是难以理解的反常现象。它们典型地反映出了一些机制和策略的作用,而且这些策略和机制,在很多情况下对个体而言,都是相当有效的;否则,它们也绝不可能在个人的认识历程或者人类的进化历程中保留下来。这些错误和偏差,应当被公平地看作是推论策略和技巧在“特定领域”中的不足,而这些推论策略和技巧,在个体整个经历中,是最为经济有效的(并且可能通常是相当正确的)。

二、非动机性归因偏差

1. 基本归因错误

最先被发现(Heider, 1958),也是最被常被引用的非动机性偏差,是指归因者在对控制行为的归因中趋向于低估环境因素的影响,而高估品性因素(dispositional factors)的作用,我们称之为基本归因错误。作为“直觉”心理学家,我们似乎经常都是本土主义者,或者比较赞同个体性差异,很少是刺激—反应式的行为主义者。我们非常乐于对各种个人品性进行推断,并预期在广泛而各异的情境和背景下个人的行为或者其结果都具有一致性。若我们亲眼目睹了同伴的行为,便会匆忙地据此得出结论,而忽视相关的环境压力和环境限制的影响。

基本归因错误的普遍证据。除开一些趣闻逸事以及求助于个人体验,196 最常被用来说明这种基本归因错误的证据(如 Jones & Nisbett, 1971; Kelley, 1971),包括归因者明显地喜欢对处于非常显著的环境压力之下的行动者进行“一致性”个人推论的倾向。例如,琼斯等(1967)发现,即使听众知道谈话者仅仅是在遵从实验者的直接要求,因而处于毫无选择的境地,他们仍然会假定谈话者的批评卡斯特罗的评论与他们私人的观点之间存在着某种一致性。

观察者会忽略或者低估环境压力的另一个更直接的证据,来自另一学

者的研究（Bierbrauer, 1973）。她主要研究的，是被试对经典的米尔格拉姆实验环境所产生的压力的印象。在她的研究中，参与者亲眼目睹了一个丝毫不差的被试“服从”的重新演示，该被试对假定的受害者施加了最大量级的电击。不管被试作判断前有怎样的时间间隔以及间隔的长短，也不管参与者在这个重演中是否实际扮演了被试的角色或者仅仅是旁观，其研究中的参与者，都犯了基本归因错误：即他们一致地，并且是显著地低估了普通被试对米尔格拉姆实验环境中要求服从的情境压力的屈服程度（见图9—1）。换句话讲，他们假定这个被试的服从反映了他独特的个人品性而不是作用于所有被试身上的潜在的环境压力和控制。

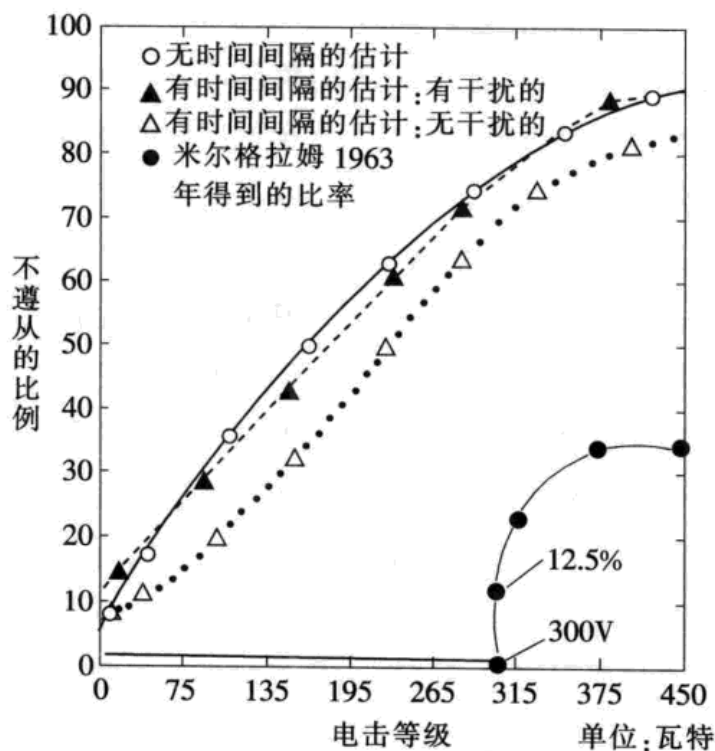


图9—1 估计的和实际的不遵从比率的比较

自我展示中授予性角色优势的一个特例

社会观察者低估环境压力和限制的影响力并高估个人品性的作用的趋向，在策略、概念性分析，甚至当代社会心理学的学术争论中，已被广泛地讨论到（见 Nisbett & Ross, 1980; Ross, 1977）。关于这个基本归因错误的一些特例，将帮助我们注意力集中到直觉心理学家的判断过程以及他

们的一些更细微的错误之上。罗斯等（1977）的一个实验，是一个非常恰当的例子。这个实验所研究的，是被试对具有优势角色或不利角色的行动者（通过随机分配产生）所进行的评价。罗斯等人设计的特定角色，是在一个常识测验游戏中的发问者以及竞争者。发问者的角色，要求被试向竞争者提出一系列具有挑战性的常识问题，并且在竞争者每次回答问题后给予其准确的反馈。而竞争者的角色仅限于回答，或者尝试回答相关问题。之后，这两组参加者（在后来重新操作的实验中，还包括旁观者）都被要求对发问者以及竞争者的常识水平进行评判。

137

很显然，这种武断的角色分配，以及对这些角色的执行，将迫使参与者和旁观者面对明显不具有代表性的，或者说具有偏差性的关于发问者和竞争者的知识水平的“样本”。发问者的角色，激励他们展示出高深的知识水平，并且保证他们能够避免涉及那些他们无知的领域；竞争者在自我展示中则没有这样的有利条件。实际上，这里的参与者及旁观者对这种角色分配的专断性或者与各个角色相联系的不同特权并不是不清楚。它不像现实世界中的许多情境，其社会角色同样赋予了个人在自我展示中的优势或者劣势，却对于个人而言是不明确的。尽管如此，发问者与竞争者的这种尽管不平等的竞争，仍然导致了一致的偏差和错误的印象。从某种意义上讲，参与者仅仅没有对由情境赋予的与各个角色相关的优势和劣势给予足够的考虑。因此，竞争者认为他们的发问者比他们自己优秀许多，而未参与其中的旁观者也显然同意这样的判断（见图9—2）。旁观者明白他们自己不会比竞争者更能回答发问者提出的那些深奥问题，因此会认为竞争者的常识水平并不差。但是他们仍然会得出发问者确实在知识水平上高出一筹的结论。有意思的是，发问者自己并不会被他们的遭遇所误导。对这样一个事实的关注，将我们的注意力从基本归因错误的普遍存在以及社会角色的特定影响，转移到各种参与者进行推论所依赖的特定“资料样本”之上。与竞争者和旁观者不同，发问者并不一定要专门依赖于常识水平的“偏差性”样本。他们显然拥有许多其他的证据，表明他们自己的常识水平及其局限，以及在他们所提出的问题中展示出的深奥信息的非代表性；结果，他们将自己以及竞争者都评定为“一般”。

138

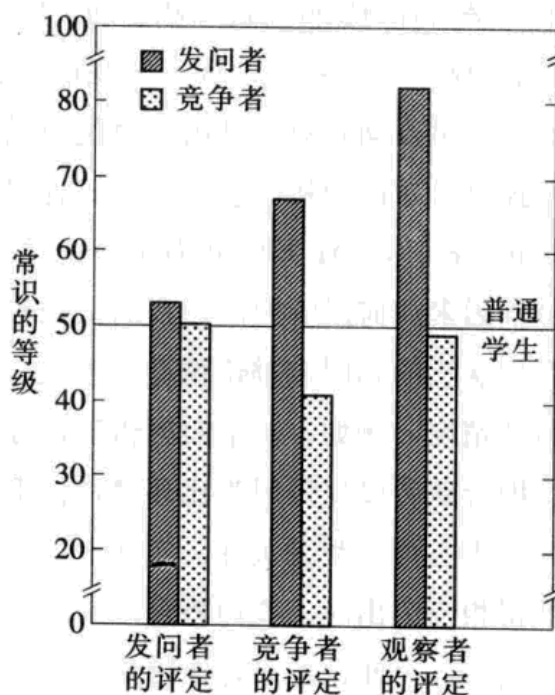


图9—2 对发问者和竞争者的常识的评定

罗斯等人的研究的社会和理论意义都是很清楚的。它促使我们去考虑无数的社会场景，其中正式或者非正式的角色限制着人际境遇，并且因此使参与者的印象发生偏差——甚至达到试图将由优势角色和劣势角色所造成的特权和限制予以合理化的程度。它也促使我们加强对直觉科学家的一个特别失误的关注——他们似乎没有意识到偏差性的资料样本只具有有限的推论价值（见 Hamill et al., 1980; Nisbett & Ross, 1980, 第四章）。

2. 显著性或便利性偏差

也许归因偏差中注意力以及调节注意力的知觉和认知因素的影响得到最有力地研究。简而言之，就是似乎无论何时，若环境中的某些因素对于感知者来说变得相当突出或“便利”（比较 Tversky & Kahneman, 1973, 11），这些因素就会对因果归因具有更大的影响力。因此，若一个行动者在 139 一个较大的群体中因为某种独特的种族或性别身份（Taylor 等, 1976），或因为一些外表或衣着的显著特征（McArthur & Post, 1977; McArthur & Solomon, 1978），或因为处于某种指导性的位置（Regan & Totten, 1975），甚至是因为座位的顺序或其他一些视觉感知上的决定因素（如 Storms, 1973; Taylor & Fiske, 1975），而使他在视觉上显著的话，这个行动者就会被认定

对她或他所做事情的结果负有与实际情况不相称的责任（见 Taylor & Fiske, 1978, 可得到更详尽的回顾）。实际上, 从“客观的自我知觉”理论导出的一系列研究, 已经揭示出行动者对自己因果角色的感知同样可以为某种简单操纵影响, 这些操纵引导他们将注意力集中或者偏离作为社会客体的自我（如, Duval & Hensley, 1976; Ellis & Holmes, 1979）。

对显著性或便利性因素如何影响归因过程的认识和理解, 可以帮助我们更好地理解许多我们熟悉的归因和推论偏差的基础, 甚至有可能将它们包括进来作为一些特殊情况。例如, 基本归因错误, 可能在很大程度上反映了这样一个事实, 即行动者仅仅是因为比环境特征更显著而使他在归因者最初寻找候选原因时更有可能受到关注。实际上, 当环境因素和限制因素对于归因者来说变得相当突出时, 我们将可以预期会出现与所谓的基本错误完全相反的归因错误。因此, 若某外在因素高度显著, 一个主管就可能将某员工的忠诚表现错误地归因于该因素——该主管的监视作用（见 Strickland, 1958）。出于同样的原因, 某行动者对一指定任务的内在兴趣将会被减弱, 如果他被引导将注意力集中于某些外在的看起来似乎激励了他在该任务中的表现的刺激或限制因素的话, 虽然这些因素实际上并不必然具有激励作用。

我们再来看一下琼斯和尼斯比特（1971）提出的经验性归纳。他们发现行动者在对自己的行为进行归因时, 相比于旁观者对同样行为的归因, 相对更乐于提及那些环境因素, 而不愿归因于自己的品性因素。从一定程度上讲, 行动者和旁观者显示出了相应的关注点的不同——行动者注意的是环境中的相关特征, 而旁观者把注意力集中于行动者本身——因此他们的归纳成为注意力/归因原则的一个特例。实际上, 实验证据显示通过操纵调节行动者和旁观者的关注点或视角, 他们引证环境或者品性原因的倾向也同样会得到调整（如 Storms, 1973; Taylor & Fiske, 1975）。

最后, 让我们考虑一下由精明的侦探小说家, 同时也是心理学家的霍姆兹（Sherlock Holmes）提出的一个推论缺陷, 即人们倾向于忽略未发生事件的推论价值。这种能提供信息的未发生事件, 指的是在某种背景下没有出现的事件或行为, 也正因为没有出现才潜在地包含了重要的信息（比

较 Ross, 1977)。很显然，认为这种未发生事件具有相对低的认知便利性，是特别恰当的。当某人寻求或考虑候选原因时，未发生事件不可能得到高度凸显或者受到适当的注意；相应地，他也不易在解释所观察行为或事件结果时赋予未发生事件足够的影响力。吉尔更有可能将杰克的愤怒归因于一些她已经“做过”的事情，而不是某些还未做的事，这仅仅是因为对于她来说前者比后者更为显著。实际上，若假定违法罪相对于疏忽罪更为显著的话，杰克就可能在对自己的愤怒作解释时犯同样的错误。

3. 虚假共识或自我中心的归因偏差

本章要考虑的最后一个非动机性的或者“信息性的”偏差，是关于人们对社会共识的估计——人们判断自己所目睹的各类行为反应是普遍的还是奇异的。普通人不能像专业的心理学家一样，依赖于精心设计的抽样技术和统计程序去做这样的估计，他们只拥有一些获得相关资料的有限渠道，因而只能依据以此为基础的直觉以及主观印象进行判断。因此，这种估计中的偏差，以及反映类似估计的各种社会推论或归因中的偏差，数不胜数。这里我们要考虑的特殊归因偏差，是关于人们感知“虚假共识”的倾向——人们通常将自己的行为选择和判断视作相对普遍的，并且对现有情境而言是适当的，却将其他替代性的反应视为少见的、异常的并且不恰当的。

在关于社会感知和归因的研究文献（比较 Katz & Allport, 1931; Kelley & Stahelski, 1970）中，偶尔会提到“自我中心的归因”（Heider, 1958; Jones & Nisbett, 1971）、“归因投射”（Holmes, 1968）以及与虚假共识偏差相关的一些特别的发现和现象。然而，最具启发性的此类证据，或许是由罗斯等（1977）所做的一系列研究提供的。

在公布的第一项实验中，被试阅读了对一种假设的冲突情境的描述，这种情境是他们可能会亲身面对的。之后他们要完成三个任务：一是估计两类可能的反应选择的普遍性；二是指出他们个人将选择其中哪一类；三是对每一类反应选择的“典型”个体的性格进行评定。

这些估计和评定显示了“虚假共识”效应；被试认为他们所选择的反应类型比他们没选的一类更具有普遍性。从虚假共识命题出发，可以得出

141

的一个显而易见的推论，即直觉心理学家会将那些与自己的反应不同的行为认定为比与自己类似的行为更能揭示行动者稳定的品性特征。罗斯等人（1977）得出的资料支持了这个推断：相对于对依照被试所选的反应类型进行反应的特定个体的评价而言，被试对依照自己未选择的行为类型行事的特定个体的评价，更确定而且更极端。

相对的这个词，对于虚假共识偏差的明确表述非常关键，因此需要作一些解释。显然，一个会在两栋摩天大楼之间走绳索的人、会发起一场革命的人或者选择牧师独身生活的人，会认识到他的选择仅为少数的同伴所认同。然而，虚假共识偏差，会导致这些个人的选择在他们自己眼中，比在那些不会走绳索、不会发动革命以及不会成为独身牧师的人眼中，更正常。同样地，当前的命题也会承认，对于某些反应类型来说，所有评判者的估计实际上都会在同一方向上发生偏差。比如，虐待以及从不虐待婴儿的父母，都会低估虐待婴儿事件。但虚假共识假设中相对的这个概念，仅仅要导出这样一个推论，即有虐待行为的父母相对于无虐待行为的父母，会将婴儿虐待事件估计为更普遍，更少揭示个人的品性特征。

罗斯等人（1977）在所做的最后一项实验中，放弃了假定性的问卷方法，而让被试面对一个真实的并且有意义的冲突情境。被试被要求挂着一个印有“EAT AT JOE'S”标志的广告牌绕着校园走30分钟。实验者让被试清楚地知道他们可以很轻易地拒绝参与这项关于广告牌的研究，但是他建议他们参加，并因此“在帮助实现研究目的的同时学到一些有趣的东西”。然后，被试被要求进行是否加入研究的个人选择，估计其他人可能的决定，并对那些同意或者拒绝加入研究的同伴进行性格推论。

采用“真实的”冲突情境的结果（见表9—1），证实了先前只涉及假定行为反应的问卷研究的发现。总的来看，同意佩戴广告牌标志的被试估计62%的同伴会作出同样的选择。而拒绝佩戴此标志的被试认为仅有33%的同伴会遵从实验者的要求。而且，如我们所预计，遵从和不遵从的被试，对同意佩戴广告牌的同伴以及拒绝佩戴的同伴所作的推论的相对强度上，存在着严重的差异。遵从研究要求的被试，对不遵从的同伴的个人性格特征进行了更确定和更极端的推论；不遵从的被试，则对遵从要求的同伴作

了更强的推论。

表 9—1 虚假共识效应：评定者对两种行为选择的普遍性的估计以及性格推断

	估计的同意 比例 (%)	估计的拒绝 比例 (%)	性格推断的强度	
			对同意佩戴的 被试的推断	对拒绝佩戴的 被试的推断
同意佩戴广告牌的被试 ($n=48$)	62	38	120.1	125.3
拒绝佩戴广告牌的被试	33	67	139.7	106.8

注：对四种性格的评定的总和；数值越高代表着评定者的推断越确定以及越极端。

资料来源：摘自 Ross, Greene, & House (1977)。

罗斯等人的实验 (1977)，显然对我们关于直觉心理学家的概念，具有某种普遍意义。普通人对事件普遍性和奇特性的估计，以及与此估计相伴的许多社会推论和人际反应，都存在系统性的而且是自我中心的偏差，偏差的方向与他自己的行为选择相一致。更普遍的是，归因分析遭受的歪曲，显然不仅来自于直觉心理学家在对社会信息的最终分析中存在的错误，而且还来自于早先对这些信息进行选择或者估计时就已产生的偏差。

几个非动机性因素，显然对于虚假共识现象的出现有一定的作用。其中的几个主要因素是：(1) 选择性暴露以及便利性因素，(2) 与解决不明情境相关的因素。

在虚假共识中隐含着选择性暴露因素，是非常明显的。显然，我们会与那些和我们拥有共同的背景、经历、兴趣、价值观及发展前景的人结识并且交往。这样的一些人中的一小部分，在面对许多情境时，确实会与我们作出相同的反应。实际上，我们与他人的密切交往，部分决定于一种普遍共识的感觉，我们会倾向于避免与那些我们认为不可能跟我们拥有同样的判断和反应的人交往。这种对有偏差的人和行为的样本的暴露，并不必然使我们在对相关人数的估计中犯错，但它确实可能导致这样的错误。更微妙并且在特性上更具认知性的因素，是那些能够提高我们回忆、形象化或者想象典型行为事例能力的因素。在一个既定情境中，我们已经选择或者将要选择的特定行为，很可能更易于从记忆中提取，并且比与之相反的行为更容易想象。用卡尼曼和特沃斯基 (1973, 4) 的术语来讲，我们所偏爱的行为选择，会在认知上更具“便利性”，我们也容易在估计相关行为

142

143

选项的可能性时，受这种获取信息的难易度的误导。

第二个非动机性的虚假共识的影响源，来自直觉心理学家们对不明情境的反应——既包括对环境压力的性质和强度的不明确，也包括对各种不同反应选择的意思及含义的不明晰。解决这种不明确的努力，通常包括解释、估计以及猜测。所有这些方式，都会对归因者自己的行为选择，以及他对其他人的选择的预测和推论，产生类似的效果。因此，预计同伴会因他佩戴“EAT AT JOE'S”的广告牌而嘲笑他，并且害怕这种嘲笑的被试，以及那些认为实验者的请求和期望无足轻重的被试，很可能会拒绝佩戴广告标志，并假定他们的同伴也会同样拒绝，同时对选择佩戴此广告牌的被试的性格，进行了很确定的推论。当然，对上述问题的轻重缓急的相反看法，将会导致相反的个人选择以及相反的社会估计和推论。

144 简而言之，虚假共识偏差，反映了同时也创造了归因过程中存在的歪曲现象。它来源于对证据的非随机的抽样和提取，以及对不明情境因素和压力的特异性处理。反过来，它又使得对异常性和偏离的判断发生偏差；而且，更普遍的是，它促使了社会现象解释中的变化和错误。

三、面对经验挑战时信念的坚定性

在本章以及其他地方（见 Nisbett & Ross, 1980），所介绍的直觉心理学家的各种缺陷，能使他对自己、对他人甚至对社会世界的本质的信念坚定不移，尽管这些信念是不成熟的，而且在很多情况下是错误的。只要这些信念只保持私下的状态，并且未导致相应的行为，似乎就无关紧要——本质上仅仅是暂时的，当输入新的信息时，可以随之进行调整。但是，一系列逐渐积累的理论和研究，现已证明其对立面的成立。

各种信念——从相对比较细微的个人印象到更宽泛的社会理论——在面对看起来具有严整逻辑性的经验挑战时，似乎都相当具有伸缩性。两个范例可以说明这种伸缩性。一是信念在遇到新信息后能够继续保留甚至被强化。而这些新信息，从一个标准的立场来看，应该是能够导致这些信念减弱的。另一个范例是在信念得以确立的最初证据基础都被否定之后，信念仍然能被保留。

1. 面对新信息时信念的坚定性及极化

对于一些紧迫的社会或政治问题，不同个人、社会派别、利益群体甚至国家，通常会持不同的信念。这种观点上的分歧是不足为奇的。由于这些观点赖以形成的基础是不规范的且通常是纯粹直觉性的，以及社会沟通（通常是高度偏差性的）在我们的信念形成过程中的作用，产生真正意见不合的情况是不可避免的。但是，当这些不同观点的持有者，有机会接触相关的证据——尤其这些证据本质上是相对规范的，并且对于所有相关群体而言都是一致的时候会如何呢？

一种比较乐观的预测，是这些彼此争斗的派系之间的信念差别会减小。这种减小包括两种情况，一种是如果那些相关证据具有逻辑一致性和说服力，不同信念就会都朝着这种为证据所支持的方向变化，由此而减小差别；另一种是如果相关证据是混杂的或者不是决定性的，不同信念就会更加抑制自身或者互相宽容。还有一个不那么乐观的预测，则是这些竞争派别将保持立场不变，也就是说他们不会理会这些新信息，而牢固坚持原有立场。最近由劳德等（Lord et al., 1979）所做的一个实验，得出一个甚至更加令人沮丧的结果（至少对那些希望或者期盼社会科学家的客观资料能够消除社会争斗的人来讲是令人沮丧的）。145

劳德等人（1979）首先选出一批被试，他们要么支持死刑并认为这是一种有效的威慑力量（支持者），要么反对死刑并认为它不具有威慑力（反对者）。接着采用一种具有相互抵消作用的设计（counterbalanced design），实验者提供给他们两份据称是非常可信的研究。其中的一份似乎从经验上支持他们的观点；而另一份似乎反对他们的立场。在他们阅读这两份研究过程中的一个关键时间点上，实验者要求两组被试评估他们对这两个研究的评价以及他们自己的态度和信念的改变。这次评估，显著地反映出了理论持有者按可以强化和支持其原理论的方式对新证据进行解释的能力。首先，死刑的支持者和反对者们，都一致地将支持他们信念的研究评估为相比于反对他们信念的研究“更令人信服”和“得到了更好的实施”。其次，与任何可以想见的在合并与某人信念相关的新证据时所采用的标准策略形成对照的是，阅读这两个研究的最终效果，是导致死刑支持者和反

对者的信念都进一步极化。这种极化发生的方式，尤其具有启发性的（见图9—3）。通过阅读支持他们观点的结果的简单描述，被试的信念变得相当极端，而且在被试进一步考察关于研究过程和资料的细节之后，其信念的极化，得以保持或者更加扩大。相比而言，通过阅读反对他们自己观点的结果的简单描述，被试仅仅变得稍微不那么极端，但在考察研究过程和资料的细节之后，被试往往返回到他们原有的信念，甚至是在知道该研究存在之前所持有的信念。实际上，许多既阅读了反对他们信念的研究的结果简介，又看了其进程细节的单个被试，最终都变得更加确信其原有信念的正确性。而对那些其信念在最初就得到支持的被试来说，即使读了同样的结果简介和进程细节，上述效果也不会发生。

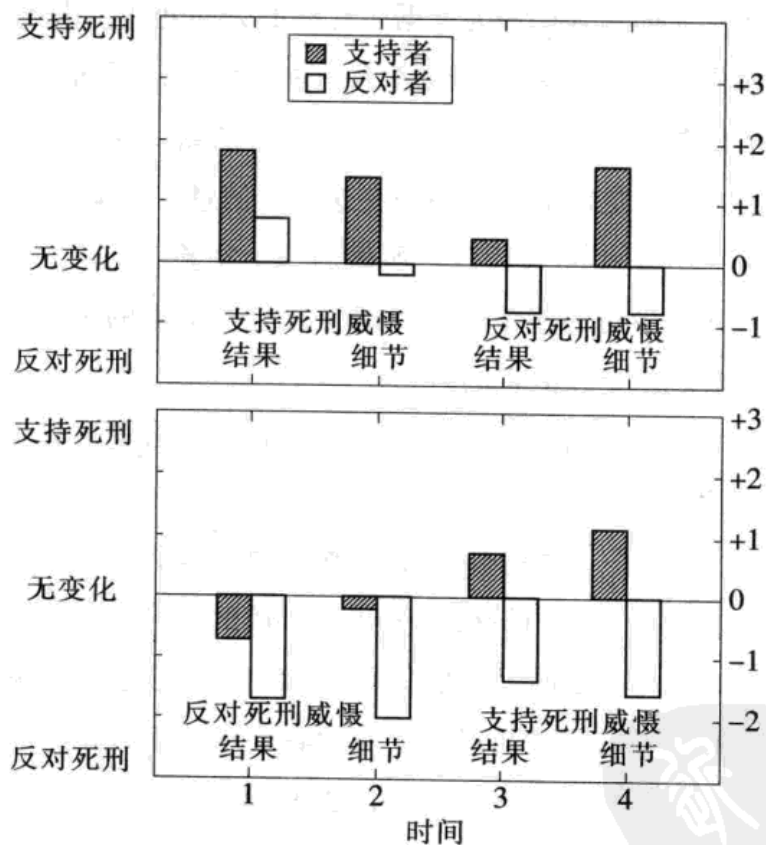


图9—3 上面的图反映的是先阅读了支持死刑的研究的被试在实验过程中随时间推移而报告的对死刑的态度改变。而下面的图反映的是首先阅读了反对死刑的研究的被试在实验过程中报告的随时间推移而产生的对死刑态度的变化。这些变化都是相对于实验刚开始时的态度而言的。

显然，专业的科学家也经常会犯和直觉心理学家们同样的错误。我们能一次次地看到存在于学术争论中的竞争性派别——不管这些争论是关于宇宙的起源、原始人类的进化路线或是自我防御性归因偏差是否存在——从一系列相同的研究发现中获取对于各自不同观点的支持。在本章的后面我们将更详细地考察此种现象背后隐藏的过程，并将专门评析科学家们努力按照他们现有理论和期望处理证据的标准情形。这里有必要首先考虑第二类普遍的信念坚定现象。

2. 证据遭怀疑后信念的坚定性

在很多情况下，个人信念受到的威胁，不是来自于新信息，而是来自对支持此信念的“正式的”证据的挑战——对导致此人在最初形成该信念的那些信息或者分析的挑战。我们很容易在逸事趣闻的层次上举出一个这样的例子。萨丽第一次学习滑冰的尝试非常失败，结果她发现是因为她借的溜冰鞋太大了，以至于不能给她的脚踝以所需的力量支持。玛丽认为约翰送她一束花反映的是他的体贴、浪漫情怀，以及某种传统礼节，结果却得知仅仅是因为约翰的父亲开着一间花店。更大范围的关于世界的理论或信念，同样也会遭到挑战。一个婴儿照看者依据她对一个会整夜号啕大哭的婴儿的经验，认定用奶瓶代替母乳喂养会导致婴儿害疝痛病，结果却发现这个婴儿整夜哭只是由于发高烧。或者一个科学家发现对某项特定理论的形成具有很大作用的一项经典实验，由于一些严重的人为操作或彻底的失误而失去了意义。

按照信念坚定假设的通常表述，它认为上面提到的社会感知者对他们原有看法的坚持会达到一种毫无根据并且毫不恰当的程度。但是这个论点中的概念，显然太具概括性，也太模糊，而无法检验。我们何时才能精确地推断出一个社会感知者对其基础已被削弱了的印象或信念“不恰当地”坚持？为了用实验证明信念坚定性假设，我们显然需要一个范式，能让我们精确地确认出到底存在多大程度的坚定性以及多大程度的信念改变。

这样的一个范式，可以由社会心理学家通过在实验中先欺骗被试，然后对作为此欺骗目标的被试进行解密的过程所产生的困境来提供。这些心理学家，在揭示了先前所提供的信息的纯粹人为的、不真实的本质后，假

146

147

定这样一个解密过程，将消除先前信息对被试的感情或者信念所可能产生的任何影响。然而，许多专业人士表达了他们公开的担忧，即这种实验性欺骗，有可能产生严重的不良后果，而它们是不能由传统的解密过程完全消除的（如 Kelman, 1972; A. G. Miller, 1972; Orne, 1972; Silverman, 1965）。

罗斯等人所做的一系列研究（早期也有类似的研究，如 Walster 等, 1967; Valins, 1974），就运用了完整的折损或者解密范式，来考察当人们面对证据遭怀疑的情况时仍保持信念坚定的现象。我们可以通过略述罗斯（1975）的一组实验来开始对此研究的讨论。他们的这组实验处理的是被试经历解密过程之后对在某特定任务中自己以及同伴的能力的印象。

（1）个人印象解密后的坚定性

罗斯等人（1975）所采用的程序相当简明。首先，在被试完成一项小说鉴别任务（如从虚构的自杀描写中辨别出真实的片段）时，实验者不断地给予他们错误的反馈。在所报道的第一个实验中，该程序是用于控制被试对自己的表现及能力的感知。在第二个实验中，旁观者被引入，他们在观看错误反馈操作时将形成自己的社会印象。在这两个实验中，当对第一印象的操作化完成之后，实验者将行动者及/或旁观者的印象得以形成的基础“证据”完成推翻。尤其是行动者（在第二个实验中被旁观者旁听）经历了一个标准的解密过程，在此过程中，他得知他的推断结果是被预先确定的，而且其反馈与他实际的表现毫不相关。在因变量被引入之前，实际上每一个被试都被引导着明确地承认了他对于实验欺骗的性质及目的的了解。

在完全推翻原有的信息之后，被试被要求填写一个关于行动者表现和能力的因变量调查量表。结果显示印象在解密后仍具有坚定性的证据，对于行动者以及旁观者都是同样明显的。实际上，在每一项测量中（如客观评价行动者刚才的表现，估计被试在将来的一系列分辨问题中的表现，主观评价行动者的能力）被彻底推翻掉的先前的结果操纵过程对行动者和旁观者的评估，都产生了显著的“剩余”效应（见表9—2）。

表 9—2 解密后对行动者的业绩和能力的感知

	行动者自身的感知			观察者对行动者的感知		
	成功	失败	<i>t</i>	成功	失败	<i>t</i>
所估计的最初正确数目	18.33	12.83	5.91***	19.00	12.42	4.43***
所预测的未来正确数目	18.33	14.25	4.23***	19.08	14.50	2.68*
所评定的能力	5.00	3.83	2.65*	5.33	4.00	3.36**

注：* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$ 。

资料来源：摘自罗斯等（1975）实验2。

一些后续的实验，都显示出许多毫无根据的个人印象一旦通过实验程序形成，就能在各种彻底的证据失效过程之后仍然保留。例如，杰宁斯等（1980）就揭示出被试对于自己的人际劝说能力的印象（他们在劝说一实验同谋献血的努力中是成功或是失败了）能够一直持续，即使在他们得知最初的结果完全不可靠之后。类似地，其他学者（Lepper et al., 1979）的两个相关实验，也显示出学生对他们的“逻辑问题解决能力”的错误印象（以及两星期后的一个后续测量中的学院选择）可以一直保持，即使在他们知道作为这些印象基础的对先前问题成功或者失败的解决，完全是由教学过程的好坏所造成的之后。

（2）已失效理论解密后的坚定性

最近由安德森（Anderson et al., 1980）所做的一系列实验，将坚定性的实证领域由个人印象扩展到更广泛的关于世界的信念。他们的研究，首先操纵了继而试图破坏被试关于两个测量变量间的函数关系的理论：消防队员职业表现的好坏与他们在之前的一项关于冒险偏好的书面测验中的分数。在一个与之相关的实验变式中，提供给被试用以形成理论的是一组仅包括一对特殊个案的证据——一个成功的和一个不成功的消防队员，在他们各自关于冒险承担偏好的测验中获得的是相应不同的分数。有趣的是，这样简单的资料，就足使被试形成坚定的关于这两个相关变量的可能关系的理论。然而，更重要的发现，是在对被试揭示前面所考察的个案完全是虚构的，其他被试实际上获得的是相反的关于冒险测试分数与工作表现的数据组合之后，这些理论仍然会被保留。实际上，当我们将经历了解密过程的被试与那些没有经过这个过程的被试进行比较时，会发现“个案记录”信息的最初效应的50%，会在解密之后仍然保持。

149

简而言之，信念显然能够在逻辑性强或经验性强的挑战中幸存。它们不仅能够幸存，甚至还能由这些证据而得到巩固，而这些证据，在局外的旁观者看来，在逻辑上是应该能对那些信念有削弱作用的。这些信念，甚至还能在它们最初的证据基础被完全破坏之后继续存活。尽管还需要做更多的工作来确认此现象的精确界限以及考察那些不可避免的例外情况，但显而易见的是，普通人的归因偏差与其他推论缺陷造成的损失不容易被校正，相反还会因随后的经历和考虑而被固化。因此，我们最后必须强调的问题，是如此的信念坚定是如何以及为何产生的？也就是说，在我们对于印象、信念和广泛的社会理论的毫无根据的坚持背后，隐藏着怎样的认知机制？

3. 信念坚定性中可能的潜在机制

(1) 偏差性的搜寻、回想以及信息同化

毫无疑问，我们的信念会影响我们搜寻、储存和解释相关信息的过程。实际上，如果没有先前存在的知识以及相应的先见，我们对于日常经历的理解将需要更多的时间、更多的努力，而且这种理解十有八九是严重不充分的。但是，我们依据先前信念处理证据的偏好所带来的一个不可避免的后果就是，我们倾向于从手头的证据中提取比它实际存在的更多的对于前信念的支持。

这种“证实偏差”（见 Einhorn & Hogarth, 1978; Hamilton, 1979; Hastie & Kumar, 1979; Wason & Johnson-Laird, 1972）长期以来一直为科学哲学家（如 Bacon, 1620/1960）所关注。也许最值得一提的是这些理论持有者对于模棱两可的或者不明确的资料的反应。如劳德等人（1979）所证实的，那些潜在可证实的证据，在第一印象上就易于被视作有价值的，而那些不具有潜在证实性的证据，就将受到高度批判性的和怀疑性的审查。因此，这就会引出下面两种结果：一是按照这种方式处理的任何类型的证据，即使是那些本质上随机的证据，也往往会支持最初的信念。二是证据一旦依据此方式经过处理，当先前信念遇到新的经验性反驳，或者其最初的证据基础遭到攻击时，它就具有了维持先前信念的能力。

我们认为当理论持有者面临新资料的情况时，偏差性同化的作用，已经得到了相当令人信服的展现（如 Lord et al., 1979）。但是，这种机制，

在折损或者解密范式中的作用，也许不是很明显，我们必须依赖的是推测而不是坚实的资料。我们假设，那些力图形成关于自己、他人或者一些函数关系的最初印象的被试，倾向于从他们的记忆以及当前的环境中寻求与此印象相关的其他资料。而这些资料若支持即将形成的印象，就更易于被回想起，并被视为相关的或者能够提供证据的。因此一个在给定任务中成功或者失败的被试，就会依据相应结果的一致性而回想起在相关任务中的同样的成功或者失败经历——并判断它们与目前情形的相关性。类似地，一个逐渐相信变量X与变量Y存在函数关系的被试，就会回想那些证实而不是挑战该假定关系的例子，并给予它们更多的信任。而且，这种偏差性的搜寻、回想和同化，不仅能支持某人先前的信念，还能创造出一种偏差性证据，它们在其先前信念基础受到攻击甚至完全被摧毁时，能始终高度有效地给予支持。这里存在的一个关键的假设，是人们不会经常对与他们信念相关的证据进行更新或者重新评估。他们通常不会判定“我先前的假设现在已经略微被削弱了，我必须回过头去重新评估我曾依据此假设加以考虑的所有证据”。

(2) 因果解释的形成

人们所做的不仅仅是注意那些与他们的印象或者信念相关的证据。他们也致力于进行因果分析或者解释 (Heider, 1958)。或者说，人们力图解释他们自己或他人的性格特征，或者解释那些他们已逐渐相信其存在的函数关系。因此，在罗斯等人 (1975) 的研究中，认为自己是一个优秀的或者较差的自杀描写分辨者的被试，就会在他们的背景中寻找一些能说明此种才华或者缺陷的情况。同样地，被引导着相信在消防队员的能力和冒险偏好之间存在某种正相关或者负相关的被试，就会毫不困难地为其相关关系假定一个逻辑基础。而且，这个过程，不仅能够支撑最初的印象或信念，还能在这些印象或信念面临之后的挑战或者攻击的时候，给予支持。

这种坚定性机制发生作用的证据，主要来自于两个解密范式的研究。它们揭示出若被试在经历解密过程之前被明确要求对其信念作出解释，那么坚定性效应的强度就能增大。在1980年安德森等人的研究中，就明确要求一组被试对由两个消防队员的个案所暗示的正相关或负相关关系进行解

释。如我们所料，这项操纵极大地增强了相应的坚定性效应。实际上，在经历解密前就对这种正相关或负相关关系的基础进行解释的被试，与那些没有经历解密的被试相比，只是对此关系有些不确定。罗斯等人（1977）的研究，也得出了同样的结果，他们发现若被试被要求对临床病人（他们读了这些病人先前的病历）经历中的某些结果进行了解释，就会一直认为这些结果是相对可靠的，即使他们了解到所解释的事件，仅仅是实验者杜撰的，毫不真实。

（3）行为证实或者“自我实现”假设

罗斯和其同事在研究坚定性现象中所采用的两种研究范式，缺少了一个要素，而该要素对于许多日常情境是非常重要的。确切地讲，就是这些研究中的被试，缺少依照他们信念行动的机会。这种行动之所以非常重要，部分原因在于它们会导致某人信念改变时的心理损失或者“失调性”（Festinger, 1957）增加（比较 Ashmore & Collins, 1968; Collins & Hoyt, 1972; Hovland et al., 1957）。而且，这样的行动，能够产生与那些信念相关的新的资料。这些新资料不仅会以一种偏差性的方式加以处理，而且它们自身也可能会朝着支持相关假设的方向而偏差化。

自我证实，或者自我实现假设的概念，对于社会科学家来讲并不新鲜。由罗森塔尔等人（Rosenthal & Jacobson, 1968）所做的著名的但极具争议的“皮格马利翁效应”（Pygmalion）研究，就是一个典型的例子。它处理的是教师的期望对于其学生的能力和表现的“大幅度提升”的影响。然而，最近由斯奈德（Snyder）及其同事所做的一系列研究，揭示了被试的期望或者他们所意图证实的假设的形式，就能产生对这些期望或假设的“客观支持”（如 Snyder and Swann, 1978a, 1978b; Snyder et al., 1977），由此很大程度上扩展了我们对于此种现象的评判或者理解。

4. 结束语：信念总在改变！

我们之前对现象及机制的讨论，不应使读者忽略掉这样一个事实，即有关我们自己，我们的政治领导人，甚至我们的科学理论的信念，都在发生着改变。在部分情况下，这种改变仅仅是毫无理性的影响力的结果。逻辑的或经验的挑战，尽管不如由正规标准所赋予它们的那样具有影响力，

但它们的影响确实存在（见 Ross & Lepper, 1980）。这些信念改变，部分反映了这样一个事实，即那些正规的假设验证方式，经常被特意地用于保护我们的信念免受非正规验证的威胁。但是我们怀疑这里面还有更多的东西，因为已有证据表明，先前的理论有时候会遭到替代，但却不是因为存在大量驳斥性证据或者决定性的操纵完美的实验。因此，由生动的、具体的、第一手的经验所导致的，以及由那些实现了引人注目的政治或宗教变革的群体或领导者的影响力所导致的观点和信念的改变（见 Nisbett & Ross, 1980），就构成了将来研究中的极具吸引力的研究对象。

本章的内容和结构，主要源于第一作者在 *Advances in Experimental Social Psychology* (1977) 中的论文。衷心感谢他们允许使用这些材料。



第 10 章

基率的证据式影响

特沃斯基和卡尼曼 (Amos Tversky and Daniel Kahneman)

158

在很多语境下，人们需要对一些目标事件的发生概率进行估计，如对某病人的诊断，或者某教科书的销售状况。而这些估计往往以两方面的证据作为基础，其一是某些相关参照整体中目标事件发生的基率性频率（如不同诊断的发生频次或教科书销售的频率分布），其二是与现有事件有关的一些具体证据（如病人在一个诊断性测试中的反应或所考察的课本的目录表）。

米尔等人 (Meehl & Rosen, 1955) 注意到基率 (base-rate) 数据在对个别事件进行的直觉预测中所扮演的角色。他们应用贝叶斯规则，认为在不准确数据基础上所进行的对稀有事件的预测，是临床预测中发生错误的主要原因。他们虽然没有进行实验研究，但他们从临床诊断的文献中引用了许多例证，在其中，并没有考虑基率信息。

为了对基率数据的影响进行实验检验，我们提供

给被试有关某个研究生或专业人员的描述，然后让他们各自预测此人的研究领域或他的职业（Kahneman & Tversky, 1973, 4）。这些研究，显示后验概率判断主要取决于那些描述在多大程度上近似于或者代表了各自的职业的刻板观念（如关于图书馆馆长或律师的刻板观念）。被试从他们的日常经验里，或者该研究的问题里，都无法得知有关这些职业类型的基率性频率，因此基率信息在很大程度上被忽视了（我们用忽视“neglect”这个词来描述基率信息被忽略或者严重考虑不足的情况）。

根据代表性或者相似性进行的预测，对基率性频率通常是不敏感的。然而，忽视基率的现象远比此更为普遍，因为它们同样出现于那些不能轻易地用代表性来解释的判断中（Hammerton, 1973）。例如，有学者（Casscells et al., 1978）向 60 名哈佛医学院的学生和教职工提出下列问题：

如果一种疾病的流行程度是 1/1 000，而检测该疾病的一项测试，出现错误的阳性结果的概率是 5%，那么你判断一个有阳性检测结果的人实际患有该疾病的概率是多少？若你一点儿也不知道这个人的症状或者迹象的话（p. 999）。

最普遍的答案是 95%，几乎有一半的被试作此回答。答案的平均值为 56%，并且仅仅只有 11 名被试回答正确，即 2%，他们假定该疾病测试正确地诊断出了每一个患有该疾病的患者。显然，即使是受过高等教育的被试，通常也未关注事件基率在相对简单的正规问题中的价值（如 Bar-Hillel, 1980a; Lyon & Slovic, 1976）。米尔等人对基率信息未得到关注的批评，不仅仅局限于临床心理学家，也同样适用于一般的医生以及其他人员。

判断和社会心理学的学者，详尽地研究了基率数据能得到采用或者会被忽视的一些条件（可参阅 Borgida & Brekke, 1981; Kassin, 1979b。他们对此类文献有一定回顾）。在这些研究中，考察的因变量可以分成两种类型：程序性的和证据性的。程序性变量，指程序设计的特征、任务以及其对被试的展示，而证据性变量，指的是信息来源的性质以及对这些证据的解释。

例如，一个非常重要的程序性变量，是判断者将各个问题作为特殊任务处理，还是将它们作为一个总的复合性预测任务处理。对概率学习以及相关任务的一些研究，提供了大量的证据。它们显示人们在进行复合预测时，倾向于匹配其标准的分布，尤其是在具有结果反馈的情况下。由于人们总是力图形成一种对结果分布具有代表性的预测类型，因此那些使用相同基率进行重复判断的实验，就能比那些每个判断被当作一个特殊问题进行处理实验，产生更大的基率效应（见 Bar-Hillel & Fischhoff, 1981; Manis et al., 1980）。

155 另一个有意思的程序性变量，是关于被试内（within-subjects）与被试间（between-subjects）设计的区分。比如，费希荷夫等人（1979）揭示出如果基率的差别是存在于提供给每个被试的所有问题中，那么相比于这种差别存在于不同被试的问题之间的情况，基率数据能够具有更大的影响力。这种被试内程序，能够导致一种普遍的对不同的分布赋予更大重要性的趋向，即使这些分布根本与问题没有关系（Fischhoff & Bar-Hillel, 1980）。要更深入地讨论比较被试内设计与非比较被试间设计的差异，以及他们对于常人统计性直觉检验的意义，请参阅本书第 34 章。

虽然程序性变量能产生相当大的效应，本章在这里却只关注对证据性变量的讨论，它们控制着对基率数据的解释和其影响。我们将关注点特别集中于我们称为原因基率和偶然基率的两种基率类型的区分。

一、原因基率和偶然基率

一个基率，如果暗示着一种因果因素，它能解释为什么某特定情况比其他情况更容易导致某种结果，那么这个基率就被称为原因基率。而一个被称为偶然基率的基率，则无法导致这样的因果推断。

有学者阿杰岑（Ajzen, 1977）提供了一个引人注目的关于因果基率和偶然基率比较的例证。在他所做的一项实验中，实验者向被试提供了对一名学生的学术能力的简要描述，然后要求他们估计这名学生通过一特定考试的概率。以下是其原因基率：

两年前，耶鲁大学一门课程的期终考试中，有大约 75% 的学生未能通过（已通过）。

这个基率是原因基率，因为它暗示着这门考试相当难（如果 75% 的学生都未通过）或者相对简单（如果 75% 的学生都通过）。这个推出的原因（如考试很难）“解释”了此基率，并使每个单个的学生更不容易（或更容易）通过这个考试。

如下是其偶然基率：

两年前，耶鲁大学的一门课程进行了期终考试。一个对教育成绩感兴趣的教育心理学家采访了许多参加这门考试的学生。由于他主要关注的是对成功（失败）的反应，因此他选择的大多是通过（未通过）考试的学生。具体而言，在他抽取的样本中，有 75% 的学生通过（未通过）了这门考试。

这个基率是偶然的，或者说是非原因的，因为样本中的成功与不成功学生的比例是由调查者主观选定的。因此它不能像原因基率那样，允许对考试的难易度进行推断。

阿杰岑的这个研究，显示出原因基率比偶然基率更具影响力，尽管两种类型的基率的变化都能产生明显的效力。对于原因基率来讲，被试在成功基率高的情况下所做的该学生通过考试的平均概率预测，比在成功基率低的情况下所做的预测，要高出 34%。对于偶然基率而言，相应的这种差别只有 12%。依据以上的分析，考试的难易，是影响学生成绩的一个有效原因，而且它与其他一些起作用的原因相结合，比如问题中涉及的那个学生的智力水平和动机。

在上述研究中，成功的基率，被用于确定考试的难易。在下面的研究中，偏好或优选基率，则被用于确定选项是否具有吸引力（Ajzen, 1977）。在此实验中，要求被试评估学生选择历史或经济学作为一门通选课的概率，同时实验者给出了对这些学生人格的简单描述。这里的原因基率，包括选

修这两门课的学生比例（分别为 70% 和 30%），它被用于确定这两种选择的相对吸引力大小。而偶然基率，则以下面的方式给出：

为了了解学生的反应，历史学（经济学）教授最近采访了 70 名选择历史学（经济学）为通选课程的学生。为了能够进行比较，他也采访了 30 名选择经济学（历史学）的学生。

我们可以注意到，偶然基率不同于原因基率，它没有提供关于这两门课程的受欢迎程度的信息。在这个研究中，偶然基率的影响是不显著的，虽然对概率的预测，在偶然基率暗示的方向上存在 2.5% 的差异。相反，原因基率产生了明显的效应：对受欢迎课程（有高的基率）的选择概率的预测平均值为 65%，而对不受欢迎课程（基率较低）的选择概率的预测平均值为 36%。显然，课程的吸引力，是从选择基率中推断出来的，并在判断特定的学生将选择某门课程而不是另一门的概率时，与该学生的个人性格特征结合在一起被考虑。然而，从一个标准的立场来看，在这些例子中的原因基率和偶然基率，应该具有大致相同的作用。

下面我们将举例说明另一类不同的原因基率；它也允许在一些合理的假定下计算正确的后验概率。让我们来看关于出租车问题（cab problem）的一个修正版本，该问题最早是由卡尼曼和特沃斯基（1972a）提出来的，后经巴-希勒尔（Bar-Hillel, 1980a）以及特沃斯基和卡尼曼（1980, 8）作进一步研究。

某天晚上一辆出租车撞人并逃逸。该城市只有两家出租车公司，绿的以及蓝的。提供给你下面的数据：

(a) 该城市中 85% 的出租车是绿的，15% 是蓝的。

(b) 一个目击者确认那辆出租车是蓝的。法院在与事故发生当晚同样的环境下检验该目击者的可信度，结果目击者在 80% 的时间里正确辨认出了每一种颜色，而余下的 20% 的时间里却没有。

157

那么在此事故中牵涉到的出租车是蓝的而不是绿的概率是多少？

为了得到正确的答案，我们用 B 和 G 分别指代事故中的出租车是蓝的或绿的假定，用 W 指代目击者的报告。按照贝叶斯概率公式，在先验概率为 $15/85$ ，似然比 (likelihood ratio) 为 $80/20$ 的情况下，

$$\begin{aligned} P(B/W) / P(G/W) &= P(W/B) P(B) / P(W/G) P(G) \\ &= (0.8) (0.15) / (0.2) (0.85) = 12/17 \end{aligned}$$

因此：

$$P(B/W) = 12 / (12 + 17) = 0.41$$

因此，尽管目击者证实车是蓝的，但撞人并逃逸的出租车更有可能是绿的，因为基率比目击者的可信度更强。

提供给大量被试的是一些关于此问题的经过细微改动的版本，但保留了相当一致的结果。被试得出典型的答案以及答案的中位值均为 80%，它与目击者的可信度一致，显然没有受到蓝的和绿的出租车的相对频率的影响。

不过，在缺少个案数据的情况下，基率信息就会被采用。在本研究中，若 (b) 条件项从问题中删除，几乎所有的被试都会将基率 15% 作为他们的答案。而且，基率信息限制着被试对于证据的预期。我们在提供给另外一组被试的上述问题中，将“一个目击者确认那辆出租车是蓝的”一句改为“一个目击者看到了那辆出租车的颜色”。然后要求被试回答“该目击者确认那辆出租车是蓝的概率是多少？”回答者给出的典型答案及答案中位值为 15%。而正确的答案应该是 $0.2 \times 0.85 + 0.8 \times 0.15 = 0.29$ 。因此，在缺少其他数据的情况下，基率被正确地用于目标结果的预测，同时被不恰当地用于目击者的报告的预测。

如果将偶然基率（关于出租车）替换为原因基率（关于事故），我们就可以发现一种不同的判断方式。这可以通过将前面的 (a) 替换为下面的 (a') 来完成

(a') 虽然这两家公司在规模上基本相当，但该城市中 85% 的出租车事故都是绿车公司的出租车造成的，15% 的事故出自蓝车公司的出租车。

虽然对此问题的回答差异很大，但基率不再遭到忽视。答案的中位值为60%，居于目击者的可信度（80%）和正确答案（41%）之间。（a'）中的基率为原因基率，是由于两个同等规模的公司事故发生率上的差别能很容易地导出这样的推论：绿车司机比蓝车司机更鲁莽，而且/或者能力更差。这个推论，可以解释事故基率的不同，并且暗示着绿车公司任何一辆出租车都会比蓝的出租车更可能牵涉进一场交通事故中。相比而言，（a）中的基率，之所以是偶然的，在于该城市中蓝的和绿的出租车数量的差别并不能证明一个因果推论，使任何一辆绿的出租车都比任何蓝的出租车更可能牵涉到一场交通事故中。

应注意到根据现有分析，关于犯错的是蓝的出租车而不是绿的出租车的后验概率在（a）和（a'）两种条件下都是相等的。然而，出租车颜色与事故之间的相关，在偶然基率下为0，而在原因基率下则为0.7！这个统计上的事实，反映了两种基率间的差别，并有助于解释为什么原因基率被采用而偶然基率则遭到忽视。

二、其他的证据性变量

基率数据的原因或偶然性质，并不是唯一影响直觉判断的证据性变量。即使在缺乏因果解释的情况下，基率数据也不会被不明确的、无力的或不一致的个案数据所取代。例如，巴-希勒尔（1980a）研究了早期出租车问题的另一个版本，其中对关于目击者的信息（即问题中的b项）被替换为：撞人并逃逸的出租车配备了对讲系统，而80%的绿的出租车和20%的蓝的出租车配备了该系统。在这个问题的回答中，（偶然）基率并未被被试所抛弃，答案的中位值为48%。巴-希勒尔认为，关于对讲系统的证据没有替代基率，是因为它不如由目击者提供的证实明确。因此，在基率数据能提供因果解释，或者其他一些证据并不比基率更明确的情况下，基率会与其他证据结合在一起（Bar-Hillel, 1980a），

确切性和因果性，都有助于解释卡尼曼和特沃斯基（1973，4）与麦克考利等人（McCauley & Stitt, 1978）的研究结果之间的差别。差别在于，卡尼曼和特沃斯基发现被试在获得对某学生的性格刻画基础上预测该学生

的研究领域时会严重忽视基率信息，而麦克考利等人发现被试判断的关于某特征的基率与他所判断的在一特定国籍下该特征出现的概率（例如，如果一个人是德国人，那么他就会很能干的概率）之间存在着本质的关联。除开几个程序上的差别，后一个研究，在三个重要的方面不同于前一研究。其一，实验者要求被试预测的是相对频次（如能干的德国人的比例），而不是预测关于某个体的概率。其二，证据中涉及的是群体成员，如德国人，而不是对一特定个体的详细描述。其三，关于某特征的基率性频率，与关于某专业的基率信息相比，更易于导出因果性的解释。常人的人格理论，提供了一些为何大多数人是乐观的，而仅有少数人具有自虐性的原因。这些原因适用于普遍意义上的人，也适用于特定的德国人，由此对特征的基率提供了一个因果解释。 159

一个非常有意思的情形，是关于具体的但非诊断性的证据（例如，对一个人的描述既适合于此人是工程师的情况也适合于他是律师的情况）。此种情况下的实验结果并不完全一致。卡尼曼和特沃斯基（1973，4）的被试忽视了基率，而另一些学者（Ginosar & Trope, 1980）在明显类似的实验条件下所发现的则是被试对基率的完全依赖。然而，大多数研究得到的是一种中间结果，即基率并没有被完全抛弃，但由于有关手头案例的非诊断性证据的存在，其影响被削弱（参见 Manis et al., 1980; Wells & Harvey, 1977）。

三、内在归因与外在归因

当具体证据与基率信息分别代表的是影响结果发生的内在性情因素和外在环境因素时，让社会心理学家们特别感兴趣的一系列基率问题就产生了。例如，某考试的难度和某学生的才干共同决定了该学生在该考试中的成功。类似地，某人是否同意向某特定目标捐款的请求，既取决于他本人的慷慨程度，也决定于这个请求的性质。外在因素，如上述的考试的难度或请求的效力等，自然是由相应的基率来表示（如：75% 的学生未通过考试；大多数人为该目标捐款）。因此，社会归因中关于环境因素和性情因素的相对影响力问题，就可以转化为相应基率的重要性问题。

尼斯比特等最先开始探究在判断研究中基率信息的使用与行为归因研究中环境因素的相对重要性之间的联系。他们发现在达利等人(1968)的助人研究中,被试虽然知道助人行为发生的低频次,但并没有影响他们对该研究中个别参与者的行为的预测,而对这些个别参与者的观察是通过一个简短的电视访问进行的。尼斯比特等人的研究(Nisbett & Borgida, 1975)有助于认知的方法与社会心理学的方法在判断研究中的会合。它也引发了一些争论(Borgida, 1978; Wells & Harvey, 1977, 1978),并激起了一股研究共识性信息在行为预测中的作用的热潮(Borgida & Brekke, 1981; Kassin, 1979b; Nisbett & Ross, 1980; Ross, 1977)。

在上述关于考试和出租车的例子中,原因的和偶然的基率都具有明确的区分,与之相对照的是,在许多共识研究中,基率通常会被给予其他的解释。为了说明这一观点,让我们比较一下尼斯比特(1975)的研究和阿杰岑(1977)的实验中的原因基率情形,其中被试要估计某特定学生通过一项有75%的人未通过的考试的概率。这两个问题的标准结构完全一样,但是前一研究中基率被大多数被试忽略掉了,而后一实验中的被试则利用了基率信息。显然,在阿杰岑的研究中,被试对令人惊讶的基率信息进行的是情境解释,而尼斯比特等人的研究中,被试将之解释为一次抽样失误。

阿杰岑的被试们的判断,显示出他们由考试成功的低基率推断出了该考试很难,尽管他们也可以用这同样的证据推出参加考试的学生能力差的结论。与之相对照,尼斯比特的被试所作的推断,显然是助人研究中的参与者大多是冷酷无情的(Wells & Harvey, 1977)。他们没能得出正确的结论,即达利等人的研究设计的情境,并不易于助长助人行为。

把极端的基率归因于一次抽样失误还是归结于情境因素,依赖于问题的前后关系:一次考试结果的不寻常分布似乎更有可能是由于此考试很难(或很容易),而不是因为该班级学生构成的特殊性。另一方面,相比于假定助人研究中的参与者非常不乐于助人,被试更难改变对于那些使人们帮助一个遇到麻烦的陌生人的条件的认识。

在预测个体案例时,被试明显忽视了基率数据,而这也同时伴随着他们对群体成员的不寻常特征的推断。如果这种推断受到阻碍,那么被试对

基率进行因果解释就变得更为可能。这个假设已经得到了许多研究的支持，在这些研究中，通过强调观察到惊人行为的样本的代表性，使基率效应得到保留（Hansen & Donoghue, 1977; Hansen & Lowe, 1976; Wells & Harvey, 1978）。在其中一个研究中，被试被告知提供基率信息的样本非常大，因而具有可靠性，基率信息对他们的预测的影响甚至由此得到了加强（Kassin, 1979a）。这个研究所得出的主要结论是，在个人预测中，对共识性信息是利用还是忽略，关键在于人们对该共识性信息的解释。

本研究受海军办公室的资助，合同号为斯坦福大学 N00014-79-C-0077。



1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53. 54. 55. 56. 57. 58. 59. 60. 61. 62. 63. 64. 65. 66. 67. 68. 69. 70. 71. 72. 73. 74. 75. 76. 77. 78. 79. 80. 81. 82. 83. 84. 85. 86. 87. 88. 89. 90. 91. 92. 93. 94. 95. 96. 97. 98. 99. 100.

11.1

11.2

11.3

11.4

11.5

11.6

11.7

11.8

11.9

11.10

11.11

11.12

11.13

11.14

11.15

11.16

11.17

11.18

11.19

11.20

11.21

11.22

11.23

11.24

11.25

11.26

11.27

11.28

11.29

11.30

11.31

11.32

11.33

11.34

11.35

11.36

11.37

11.38

11.39

第**四**部分
PART FOUR

便利性



第 11 章

便利性：用于频率和概率判断的启发式

特沃斯基和卡尼曼 (Amos Tversky and Daniel Kahneman)

一、引论

163 近来的许多研究都在关注频率判断和概率判断的有效性以及一致性。但是对于人们采用哪些心理机制去估计某些类别的出现频率或某些事件的发生概率，我们却知道得很少。

我们假定人们在面临概率判断或频率判断这样比较难的任务时，会采用数量有限的几种启发式以简化这些判断。在其他地方，我们已经详细分析过其中的一种启发式——代表性。如果某事件体现了它的整体或原发过程的某些基本特征，人们就会利用这个启发式而判断该事件是可能的……

在通过代表性判断某事件的概率的时候，人们会将这个事件的基本特征与导致它发生的整体结构的基本特征进行比较。按照这种方式，人们评估其相似性或隐含距离 (connotative distance)，再对概率进行估

计。另外，人们也会通过评估便利性或联结距离（associative distance）来估计概率。长时期的经验已经教会我们，回忆大数量类别的例证，通常比回忆那些出现较少的类别的例证，更快也更清晰；想象可能出现的事例比想象不太可能发生的事件也更容易，而且若两个事件经常同时发生的话，其联结关系会得到进一步加强。因此，人们在估计某类别的数量、某事件的发生概率或事件同时发生的频率时，会通过对与此相关的回忆、建构或联想等智力活动发生的容易程度来完成。 164

例如，某人会通过回忆他熟人圈子中的离婚事件来估计某特定群体中的离婚率；他会考虑一个政治家失去支持的多种途径，以估计该政治家在选举中失败的概率；他还会评估残暴与食肉兽之间的关系强度，以估计一个残暴的人会在罗夏（Rorschach）卡片中“注意”食肉兽的概率。在所有这些例子中，便利性评估，都在调节着对某类别频率或某事件概率的估计。^[1]无论什么时候，只要人通过回想实例或者相关性的难易度来进行频率或概率估计，我们就称他是采用了便利性启发式。评估便利性，并不一定需要进行实际的回想或建构活动。我们只需要评估这些活动发生的容易程度就足够了，就好像我们并不需要考虑详细的解答，就能够评估某个谜语或数学问题的难度一样。

联结关系能够通过重复得到强化，或许是人们所知道的最古老的记忆法则。便利性启发式所考察的，是这个法则的相反形式，即它将联结关系的强度作为频率判断的基础。在这个理论中，便利性是一个中间变量，而不像在记忆研究中是一个典型的因变量。便利性，在生态学上，是一个对频率判断很有效的线索，因为频率高的事件通常比频率低的事件，更容易被回忆或者想象。然而，便利性也会受到一些与实际频率无关的因素的影响。如果便利性启发式得到采用，那么这些因素就会影响人们所感知的类型的频率以及对事件概率的主观判断。结果，便利性启发式的采用，将导致系统性的偏差。

本章将考察 10 个研究中的便利性启发式。^[2]我们首先证明人们能以较好的速度和准确性估计便利性。然后，我们将揭示人们所判断的某类别的频率，会因他们建构和回忆实例的便利性而产生偏差。本章所介绍的实验

165 研究,关注的都是频率判断或概率判断,而后者很容易被转换为相对频率。我们将在第五部分及最后一部分,讨论便利性对基本上是唯一的事件(它们不能被转换为相对频率)的概率判断的影响。

二、便利性评估

1. 研究 1: 建构

我们给被试($N=42$)出示了一系列词语建构问题。每个问题,由一个包含了 9 个字母的 3×3 的矩阵构成,在这中间我们要建构由 3 个及其以上的字母组成的词语。在这个研究的训练阶段,我们给每个被试提出了 6 个问题。对每一个问题,他们有 7 秒钟的时间,来估计他们认为自己能在 2 分钟内建构出来的词语数目。在这每一个估计之后,我们给他们 2 分钟来写下他们能从这个矩阵的字母中建构出的词语(写在一些已经编号的横线上)。我们并没有采用训练阶段的数据。在实验阶段,建构和估计这两个任务被分开了。每一个被试要在 8 个问题中,估计他们认为自己能够在 2 分钟内建构出来的词语数。而对另外 8 个问题,他们不是事先进行估计,而是直接建构词语。估计和建构问题交替进行。我们采用了两本相同的小册子,因此对每一个问题,有一半的被试在作估计,而另一半被试在建构词语。

结果

平均建构的词语数,最低为 1.3 (XUZONLCJM 这一问题),最高为 22.4 (TAPCERHOB 这一问题),总平均数为 11.9。而估计的平均数,最低为 4.9,最高为 16.0 (与前面的问题相同),总平均数为 10.3。这 16 个问题中,估计数与实际建构数之间的积矩相关 (product-moment correlation) 为 0.96。

2. 研究 2: 提取

除了任务的性质有差别外,本实验的设计和程序都与研究 1 相同。这里每一个问题包含着一个目录,如花或俄罗斯小说家,它们的实例将由被试来回忆。被试($N=28$)被限定在 7 秒钟内估计他们能够在 2 分钟内能回忆到的实例数,或他们在 2 分钟内实际回忆到的实例数。跟实验 1 —

样，建构和估计任务，在训练阶段是结合在一起的，而在实验阶段则交替进行。

结果

被试回忆到的平均实例数，最小为 4.1（以 F 开头的城市名），最大为 23.7（四条腿的动物），总平均值为 11.7。估计的平均数，最小为 6.7，最大为 18.7（在上两个目录中），总平均值为 10.8。这 16 个目录的回忆值与估计值之间的积矩相关系数，为 0.93。

3. 讨论

在前面的两个研究中，通过被试在任一给定问题中回忆或建构出的实例的总数量，我们可以测量出获得实例的便利性。^[3]这两个研究揭示出人们能迅速并且准确地估计便利性。这种估计是如何进行的呢？有学者（Bousfield & Sedgewick, 1944）的研究，提出了一个可能的机制，即能回忆起的实例数目是随着时间的长短呈负向指数增长的。因此，被试能用他们在短时间内回忆起的实例数，来估计那些仅能在更长时间内回想起的实例数。另外，被试根本不需要明确地回想或者建构任何实例就能够估计便利性。例如哈特（Hart）在 1967 年的研究中，就已经发现人们能够准确地估计他们认识某些事物的能力，而他们在单词配对联想（paired-associate）记忆实验中，却不能回想起这些事物。

166

三、建构的便利性

现在我们转向另一系列问题，在其中，我们提供给被试一些建构实例的规则，并让他们估计这些问题的总频率（或相对频率）。与大多数估计问题一样，在这些问题中，被试无法建构和列举所有的例子。相反，我们假设他们会尝试建构一些实例，并通过建构的便利性，来判断整体的频率而这里所谓的建构的便利性，即指对回忆实例的容易度的评判。因此，其实例很容易建构或想象的类别，相对于数量相等但很难想到实例的类别，将被视为出现频率更高的。我们用词语频率判断以及对一些组合表达式的估计研究，来验证上面这个推测。

1. 研究 3：词语频率判断

假定你从一篇英语文章中随机地选出一个单词。这个单词，更有可能

以 *K* 开头呢，还是以 *K* 作为它的第三个字母？根据我们的理论，人们在回答这个问题的时候，是通过比较这两类任务的便利性来完成的，即他们会估计哪一类的例子更容易回想起。我们当然更容易想出那些以 *K* 开头的单词而不是把 *K* 放在第三个位置的单词。如果这种频率判断确实受到便利性估计的影响，那些以 *K* 开头的单词将被判断为更常出现。实际上，某特定文本中所包含的 *K* 为第三个字母的单词，会是以 *K* 开头的单词的两倍。

根据学者 (Mayzner & Tresselt, 1965) 所进行的广泛的单词计数，一共有 8 个辅音字母会经常出现在单词的第三个位置而不是第一个位置。在这 8 个字母中，有两个是相对比较少见的 (*X* 和 *Z*)，另一个 (*D*)，仅仅在由 3 个字母构成的单词中，才会更多出现在第三个位置。剩下的 5 个辅音字母 (*K*, *L*, *N*, *R*, *V*) 被选用来进行研究。

我们给被试出示下面的说明：

我们已经研究过字母在英语中出现的频率。选出某篇文章，然后记录下字母表中不同字母出现在整个词的第一个位置和第三个位置的频率。少于 3 个字母的词不在我们的计数范围内。

现在我们将给你几个从字母表中选出的字母，要求你判断这些字母出现在第一个位置的次数多，还是出现在第三个位置的次数多，并且估计它们出现在这两个位置的频率比。

给他们的問題是这样的：

考虑一下 *R* 这个字母。

R 更可能出现在

____ 第一个位置？

____ 第三个位置？（请选择一个）

我估计这两个位置的频率比为 ____ : 1。

我们要求被试估计频率高的一类与频率低的一类的比率。对其中的一半被试，问题中这两个位置的顺序是颠倒过来的。而且，对这 5 个字母我们采用了 3 种不同的排序。

结果

在所有 152 名被试中，有 105 人认为 5 个中的多数字母更可能出现在第一个位置，另外 47 人则认为大多数字母会出现在第三个位置。这种对第一个位置的偏差，是非常明显的（在显著性检验中， $p < 0.001$ ）。而且，这 5 个字母中的每一个，都被大多数被试判断为经常出现在第一个位置而不是第三个位置。对每一个字母，他们估计的比率平均值为 2:1。尽管事实上，这里所有的字母，都常出现在第三个位置，但被试还是得出了它们更多出现于第一位置的判断。

在其他研究中，我们发现了同样的对于第一位置的偏好，且不管其研究采取的是每一个被试只判断一个字母的被试内（within-subject）设计，还是字母出现在第一及第三个位置的频率由不同的被试来判断的被试间（between-subject）设计。我们还观察到，在被试内设计中，引入正确判断的回报机制（payoffs for accuracy）根本不会产生任何作用。由于所有这些实验方法都得出的是大致相同的结果模式，我们在这里就只介绍用最简单的程序得出的结果。

菲利普斯（Phillips, 1966）在一个关于贝叶斯推论的研究中得到了相同的结果。他们让一个学生出版物的 6 名编辑估计从他们自己的文章中抽取出的各种双字母组合（bigrams）出现在词语的开头或是结尾的概率。在这个研究中观察到的一个偶然效应，是所有的编辑都具有同样的偏差，即赞同双字母组合出现在词语开头的假设。比如，这些编辑错误地判断以 re 开头的词比以 re 结尾的词更多。而前者显然比后者更易于被想到。

2. 研究 4：排列（permutations）

请看 A 和 B 两种排列，如下所示。

(A)

X X X X X X X X
X X X X X X X X

(B)

X X
X X

X X X X X X X X

X X
X X
X X
X X
X X
X X
X X

一个排列中的路径指的是一条线，它能够将第一排的某个元素与最后一排的某个元素连接起来，而且在每一排它通过一个而且仅仅是一个元素。

这两种排列中的哪一个拥有的路径更多？

你认为在每一个排列中共有多少种路径？

大多数读者都可能跟我们拥有相同的直观印象，认为在 A 中的路径要比 B 更多。我们的被试也同意这种意见：54 名被试中有 46 人说自己在 A 中看到的路径要多于 B（显著性检验中， $p < 0.001$ ）。他们对 A 的平均估计值是 40 条路径，而 B 是 18 条。实际上，在两个排列中的路径数是一样的，都是 $8^3 = 2^9 = 512$ 。

为什么人们在 A 中会看到比 B 更多的路径呢？我们假定这种结果所反映的，是两种结构中的路径在便利性上的差异。有几种因素使得 A 中的路径比 B 中的路径更易获取。首先，最直接易获的路径，是各个排列中的纵列。在 A 中有 8 条纵列，而在 B 中只有这样的两条。其次，在跨越这些纵列的路径中，属于 A 的比属于 B 的更特别且更不易被混淆。平均而言，A 中的任两条路径共同通过的元素为总元素的 $1/8$ ，而 B 中的任两条路径共同通过的元素为总元素的一半。最后，A 中的路径比 B 的路径更短，因此也更易于想象。

3. 研究 5：组合

1629

让我们来想一下下面这个问题。我们需要从一个 10 人群体中选出一个由 r 名成员构成的委员会， r 值位于 2 和 8 之间。那么这 10 个人能够构成多少种不同的包含 r 名成员的委员会？这个问题的正确答案由一个二项式系数 (r^{10}) 给出，它在 $r=5$ 的时候达到最大值 252。显然， r 名成员组成的委员会的数量，

与 $10-r$ 名成员组成的委员会的数量是相等的，因为任何一个选出来的群体，比如两名成员构成的群体都能对应一个特定的未入选的八人群体。

然而，根据我们对直觉估计的分析，两个人的委员会比 8 个人的委员会更容易被想到。首先，建构一个委员会的最简单的策略，就是将整个群体分成一些分散的子群体。那么，人们很容易看出由两个人构成的独立的委员会有 5 个，而由 8 个人构成的独立的委员会甚至不足两个。其次，由 8 人构成的委员会由于成员之间的重叠度较高而显得不那么易于区分；任两个 8 人委员会之间都会共有至少 6 名成员。以上的分析暗示着小委员会比大委员会在想象时更具有便利性。因此，依据这个便利性假设，小委员会将显得数量更多。

有 4 组被试（总人数 $N=118$ ）估计了一个 10 人群体所可能组成的包含 r 名成员的委员会数量。不同的被试组在估计时所用的 r 值分别为 2 和 6, 3 和 8, 4 和 7, 5。

图 11—1 显示了对委员会数量的平均估计值以及正确值。与我们的预测一致，被试所判断的委员会的数目多少随着其构成成员的增加而减少。

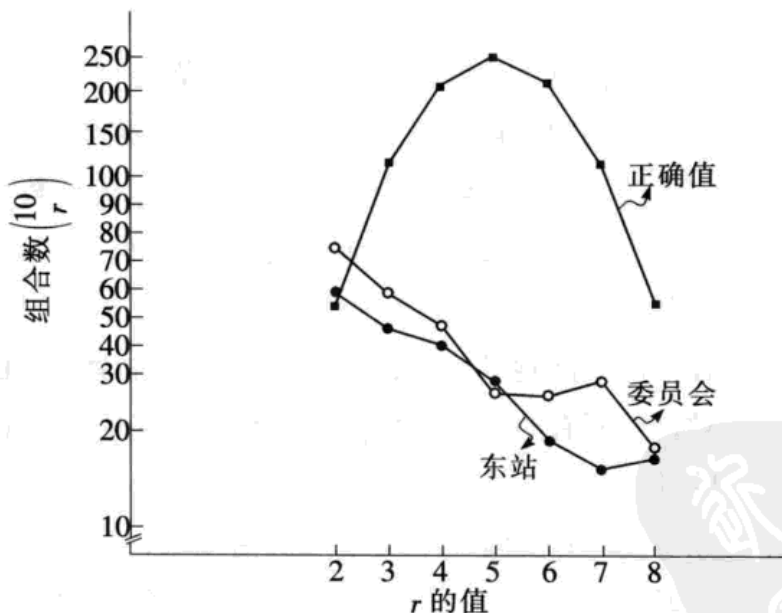


图 11—1 委员会问题以及车站问题的正确值和在对数尺度上的估计中值

我们还设计了对这同一个问题的其他表述方式，以检验之前的结果的

普遍性：

下图表示在始发站和终点站这条路线上共有 10 个车站。一辆在这条路线上运营的巴士会在 r 个车站靠站。



这辆巴士在这条线上停 r 个车站的不同方式有多少种？

停靠 r 个车站的不同方式的总数，仍然是由二项式系数 $\binom{10}{r}$ 给出的。

而且显然这里停两个车站的方式，与停靠 8 个车站的方式总数是相等的，因为任一种停靠类型都会对应着一个特定的与之互补的未停靠车站类型。然而，似乎人们在建构停靠两个车站的方式时，会比建构停靠 8 个车站的方式有更大的自由度，因为若只需在两个车站停靠，会有很多车站供你从中选择，而若要在 8 个车站停靠，则意味着几乎每一个车站都需要停。我们之前的分析认为停靠两站的方式在想象时更具便利性；乍一看上去，这样的组合方式更多，它们更容易区分，而且易于想象。

我们让与之前不同的四组被试（总人数 $N = 178$ ）回答了 $r = 2, \dots, 8$ 时的停靠站组合数量，程序都依照上一个设计。停靠类型数目的平均估计值也在图 11—1 给出。与委员会问题研究的结果一样，表面上组合数随着 r 值的增大而减小，这与从便利性假设推出的结论一致，但与正确值差别极大。而且，在这两个问题中，对组合数的估计是非常相似的。在另一个组合问题中，被试对所有的正确值都存在着明显的低估，单单 $r = 2$ 这个最易想象的类型除外。

我们假定，在实验 4 和实验 5 中发生的低估现象，是由于人们对组合值的估计来自他们通过最初印象得出的推测。人们在第一眼获得的印象以及通过简单的几步计算所得出的认识，并不能够给予他们足够的关于很多组合表达式急剧的增长率的信息。在这种情况下，从最初印象导出的推测，就会导致显著的低估。在这里需要作出一种区分，即推测的基础是来自于如前两个研究所实例最初的便利性，还是来自像下面这个研究中所呈

示的最初计算的结果。

4. 研究 6：外推 (extrapolation)

我们要求被试在 5 秒钟内估计写在黑板上的一个数学表达式的值。其中一组被试 ($N=87$) 需要估计的数值是 $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ ，而另一组 ($N=114$) 估计的数值是 $1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$ 。对于下降序列的表达式，被试的中位估值为 2 250。而对于上升序列的表达式的中位估值为 512。这两个估值的差别是高度显著的 (中位值检验, $p < 0.001$)。而且这两个估计值，都大大低于正确值 40 320。

不管是对正确值的低估还是这两个估计值之间的差别，都支持了同一个假设，即人们在估计 $8!$ 的时候，是通过局部计算得出推断。与其他组合表达式一样，阶乘也具有增长率不断上升的特点。结果，那些只通过局部的计算进行推断的被试，就会严重低估阶乘积。因为在下降序列的组合表达式的前几步乘积 (计算顺序从左至右)，要大于上升序列的组合表达式的前几步乘积，被试就判断前一个表达式的结果要大于后者。对下降序列的表达式的评价，可能是像下面这样：“8 乘以 7 为 56，再乘以 6 就已经超过 300 了，因此我们是在计算出一个相当大的数值。”而另一方面，在评估上升序列的表达式时，被试可能如此推理：“1 乘以 2 为 2，再乘以 3 等于 6，再乘以 4 也只等于 24，这个表达式的结果显然并不会太大……”

5. 研究 7：二项式——便利性对代表性

本章讨论的最后一个研究，是要探索便利性在二项式分布评估中的作用，并阐明问题的表达方式会影响人们在直觉估计中对所采用的启发式的选择。

我们给被试 ($N=73$) 提供了下面的说明：

请看下面这个图表：

X X O X X X

X X X X O X

X O X X X X

X X X O X X

X X X X X O

O X X X X X

该图表中的路径指的是任一条递减的直线，它起始于第一排，结束于最后一排，并且在每一行只通过一个符号（X 或 O）。

你认为下面这些类型的路径各占的比例为多少？

通过 6 个 X，0 个 O _____ %

通过 5 个 X，1 个 O _____ %

.....

.....

.....

通过 0 个 X，6 个 O _____ %

请注意这些包括了所有可能的路径类型，因此你的所有估计值加起来应该等于 100%。

172

路径类型的实际分布是一个 $p = 5/6$ ， $n = 6$ 的二项式。当然，人们既不能靠直觉得出正确答案，也无法列举出所有相关的例子。相反，我们假定他们会扫一眼图表，然后依据某类型的个别路径能够在脑海中建构出来的容易度，估计每一种路径类型的相对频率。由于在建构一条路径的每一个步骤（即在图表的每一行）中，X 的数量都要多于 O 的数量，因此建构包含 6 个 X 的路径就要比建构包含 5 个 X 和 1 个 O 的路径容易得多，尽管后者实际上数量更多。相应地，我们预计被试会错误地判断包含 6 个 X、0 个 O 的路径将是数量最多的一类。

图 11—2 显示的，是所有路径类型的相对频率的中位估值，以及正确的二项式值。这个结果证实了我们的假设。在 73 个被试中，有 54 人错误地判断包含 6 个 X、0 个 O 的路径要多于包含 5 个 X 和 1 个 O 的路径，而只有 13 人认为后者比前者数量更多（符号检验， $p < 0.001$ ）。被试主观认定的路径类型分布的单调性，显然是一个普遍现象。我们在其他条件下也得到了同样的结果，如改变 p 和 n 的赋值（ $p = 4/5$ 和 $5/6$ ， $n = 5$ 、6 和 10），或改变总体的分布（如在路径图的第一行安排 4 个 X、1 个 O，或者 8 个

173

X、2 个 O。)

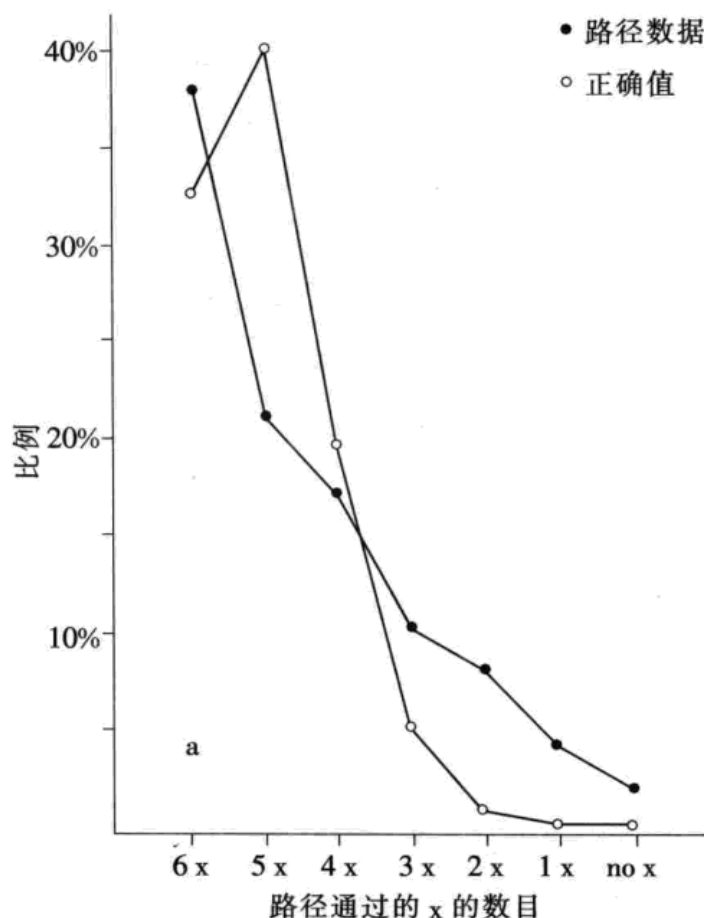


图 11—2 路径问题中的正确估值与估计中值

为了进一步研究这个结果的稳定性，我们进行了下面这个补充性实验。我们向 50 名对于二项式一无所知的斯坦福大学的学生，提出了上面的路径问题。这里，我们并不要求被试估计相对频率，而是仅仅要求他们判断“包含 6 个 X 和 0 个 O 的路径数量多呢，还是包含 5 个 X 和 1 个 O 的路径多？”这些被试单独进行判断，而且我们向他们许诺如果他们判断正确，将得到 1 美元的奖金。结果绝大部分的被试（50 人中的 38 人，符号检验 $p < 0.001$ ）再次认为前者的数量更多。显然，错误的直觉，不可能通过引入金钱回报机制，来轻易地加以改变。

我们曾假定当二项分布以一个路径图表来表现时，人们会通过评估在想象每一种路径类型的单个路径时的容易度来判断它们的相对频率。采用这种评估方式，是受到了路径定义中的有序特征以及问题的图像式表达的

提示。下面我们来看对同一问题的另一种表达方式。

在一个纸牌游戏中有 6 个参与者。一轮游戏中，每一个参与者都要从一副洗好的纸牌中随便抽出一张牌。在这副牌中， $5/6$ 的纸牌印的是 X，而剩下的 $1/6$ 印的是 O。在多次游戏中，出现下面情况的轮数各占多少比例

6 个参与者抽到 X，没有参与者抽到 O _____ %

5 个参与者抽到 X，1 个参与者抽到 O _____ %

.....

.....

.....

没有参与者抽到 X，6 个参与者抽到 O _____ %

请注意这些已经包含了所有可能的抽取类型，因此你的估计数加起来应该等于 100%。

这个纸牌问题在形式上是与路径问题相同的。但是它意欲引出一个不同的评估模式。在路径问题中，个别例子是通过直接的展示来加以强调的，且总体比例（即 X 在每一行的比例）并没有明确地表达出来。而在纸牌问题中，明确地告诉了总体比例，但没有提及个别例子。因此，我们假定，对纸牌问题的结果的评估，所依据的是其中的每一类抽取方式对于整副牌的构成的代表性，而不是个别例子的获取便利性。在纸牌问题中，“5 个 X 和 1 个 O”的结果是最具有代表性的，因为它与总体完全吻合（见 Kahneman & Tversky, 1972b, 3）。因此，依据代表性启发式，这个结果将被判断为比“6 个 X 和 0 个 O”的结果出现次数更多，这与路径问题中所观察到的判断模式是完全相反的。回答纸牌问题的 82 个被试中，有 71 人的判断证实了我们的预测。在路径问题中，73 名被试中，只有 13 人，以同样的方式得出了与此相同的判断；这两类问题表述所造成的差别，是非常显著的（在卡方检验中， $p < 0.001$ ）。

图 11—3 显示的是纸牌问题中被试的中位估值。图 11—2 和图 11—3

之间的差异，证实了我们的假设，即同一问题的不同表达方式会引出不同的启发式。具体而言，如果某类型事物的个别例子在题目中得到强调的话，对该类型事物的频率判断就有可能依据便利性，而如果该类型事物的整个类型特征在题目中很显著的话，人们的判断的就更可能依据代表性。

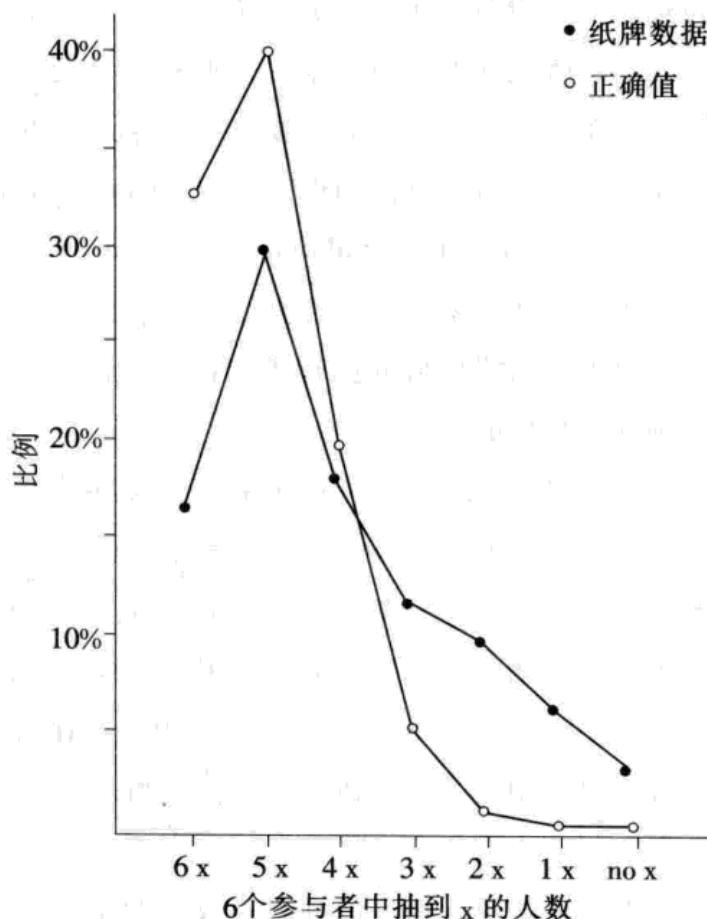


图 11—3 纸牌问题中的正确值和估计中值

四、提取的便利性

在本节我们要讨论的几个研究中，研究者先让被试看一些相关信息（如一张姓名列表），然后要求他们判断信息中提到过的某给定类型的具体项目出现的频率。跟我们在前一节中讨论过的问题一样，被试无法回忆和计算所有的例子。相反，我们假定他们会尝试回忆一些实例，然后依据便利性判断整体的频率，即依据实例浮现在脑海中的容易度来作判断。由此，

175 那些实例很容易被回想的类型，就会被判定为比实例难以回想的类型数量更多，尽管它们的数量其实是相等的。这个预测，首先通过一个关于范畴判断的研究来加以证实……

研究 8：声望、频率和回忆

我们给被试播放了一份包含了许多知名人士的姓名列表的录音，其中男女性别都有涉及。在听完所有的姓名后，一些被试被要求判断此姓名表中包含的男士多，还是女士多，另一部分被试则要尝试回忆列表中的名字。表中所列的一些名字是相当有名的（如 Richard Nixon、Elizabeth Taylor），而另一些却不那么有名（如 William Fulbright、Lana Turner）。通常比较著名的名字会更容易回想。因此，如果频率判断受到被试所估计的便利性的影响，那么与由不太出名的名字构成的比较类相比，包含了出名人士的姓名的一类，就会被判断为数量更多。

我们准备了 4 份姓名列表，其中两份由演艺人员构成，另两份列出的是其他公众人物。每一份列表包括 39 个名字，它们的播放频率是每 2 秒钟一个名字。有两份列表（一份是公众人物，一份是演艺人员）包含 19 个著名女士的名字，20 个不太著名的男士的名字。另两份列表则包括了 19 个著名男士的名字以及 20 个不太著名的女士的名字。因此，在所有的列表中声望和频率都存在着反向的相关关系。依据所有名人姓名中的名字，都能够明确地辨认其性别。

我们要求被试用心听录音。每一份姓名表放给两个被试组听。在听完录音之后，我们要求其中一组被试写下所有他们能够回想到的列表中的姓名。要求另一组被试判断列表中包含的男士姓名多还是女士姓名多。

结果

(1) 回忆。对 19 个著名人士的姓名，被试能平均回忆出 12.3 个，对另 20 个不太著名的姓名，他们能平均回忆出 8.4 个。4 个回忆组的 86 名被试中，有 57 人回想的著名人士名字多于非著名人士的名字，只有 13 个回想出的非著名人士名字多于著名人士的名字（符号检验， $p < 0.001$ ）。

(2) 频率。在比较列表中男士和女士的姓名频率的 99 名被试中，有 80 人错误地判定包含了更多著名人士名字的一组频率更高（符号检验，

$p < 0.001$)。

五、事件的提取和情节的建构

在本章我们所讨论过的所有经验研究中，都存在着一个列举实例的客观步骤（如以 K 开头的词或图表中的路径），因此每一个问题，都有一个客观的正确答案。但在许多需要进行概率判断的现实情境中，却往往没有一个客观正确答案。每一次经济衰退、一次成功的医学治疗，或者一桩离婚案，基本上都是独特的。对这些事件的概率不能只通过事例间简单的吻合来判断。然而，便利性启发式，也可能被用于对这些事件的发生可能性进行估计。

176

比如，在判断某对夫妇将会离婚的可能性时，人们会在记忆中搜寻因这个问题而浮现于脑海中的类似的夫妇。如果在他以此方式想到的所有例子中离婚非常的普遍，这对夫妇离婚的可能性就显得很大。或者，人们会通过尝试着建构一个能导致离婚的故事，或是一个情节来估计其可能性。这个建构出的情节的合理性或者其被虚构时的容易度，提供了一个可能性判断的基础。在本节中，我们将讨论便利性在这种判断中的作用，推测一些有所预计的偏差来源，并勾画出在进一步的研究中应该努力的方向。

我们来看下一个假想出来的临床情境，以解释便利性偏差。^[4] 一个医生，他听到某个病人抱怨说他对生活已经非常厌倦。他在猜测这个病人会不会自杀的时候，会仔细地回想他所知道的类似的病人。有时候可能只有一个相关的病例会进入脑海中，这大概是因为这个病例最让人难忘。这时候，对该病人自杀可能性的主观判断的主要依据，就是该病人与所想到的案例之间的相似性。如果这两者非常类似，那么这个医生就会预计过去发生的事情会重演。如果该医生回想到许多的案例，他就可能会衡量这些案例在一些关键特征上与目前病人的吻合程度。

相关的例子是如何被选出来的呢？在浏览他过去的经历的时候，这个临床医生想到的是与当前病人相似的病例呢，还是那些试图自杀的病例，或是既与当前病人相似又试图自杀的病例？当然，从一个保险统计的观点来看，相关的类型应该是与当前病人在某些方面相似的病例，而相关的统

计量是这些病例中试图自杀的频率。

记忆搜寻也可能遵循其他的规则。由于试图自杀是一个非常引人注目的和显著的事件，自杀的病人，就可能比那些并未企图自杀的病人更难以忘记，也更易于回想。因此，这个临床医生会回想起他所遇到的自杀病人，然后依据这些病人与当前病人的相似程度来判断当前病人试图自杀的可能性。这种方法会导致严重的偏差。注意到自己能想起的所有自杀病人几乎都相当沮丧的医生会得出结论说，如果一个病人显示出严重的沮丧情绪的话，他就可能自杀。另一方面，如果该病人看起来与这个医生之前遇到的自杀病例并不类似，他就会断定该病人不可能自杀。这样的推理，忽略掉了在精神沮丧的病人中只有一小部分人会企图自杀的事实，以及当前病人与该临床医生遇到过的其他病例完全不同的可能性。

最后，一个临床医生可能会想到一个既沮丧又企图自杀的病人。于是，他会通过这个病例进入脑海中的容易度或者当前病人对此类病例的代表性，来评估该病人自杀的可能性。这种推理仍然会导致严重的错误。有许多沮丧的病人企图自杀的事实，并不足以支持一个沮丧的病人必然企图自杀的可能性。然而，这种评估模式并不少见。许多研究（Jenkins & Ward, 1963; Smedslund, 1963; Ward & Jenkins, 1965）显示出像症状和疾病这样的两个二元变量之间的偶然关系（contingency），通常是根据它们同时发生的频率来判断的，对其中任一方不出现的情况，很少考虑或根本不考虑。

有些事件感觉非常独特，以至于过去的历史看起来对估计它们的发生可能性并没有太大作用。在考虑这些事件的时候，我们通常会建构情节，即能将当前情境导引至目标事件的一段故事。想象出来的这些情节的合理性，或者虚构这些故事的容易度，会成为我们估计当前事件发生可能性的一个线索。如果无法想出一个合理的情节，这个事件就会被认为不可能或者非常不可能。如果我们能想出许多的情节，或者我们想象出来的某情节相当具有说服力，那么这个事件在我们看来就非常可能发生。

人们想要估计其可能性的许多事件，都建立于一些相互关联的因素上。然而对人的智力而言，要弄清楚这些相互关联的因素变化的次序，是相当困难的。我们假定在估计复杂事件的概率时，人们只能想到最简单也是最

便于想象的情节。具体来讲，人们倾向于虚构这样一些情节，其中许多因素根本就不会有改变，即使有发生也只是最显而易见的变化，而且因素之间相互影响的变化极少。由于这种想象出的情节受到了极度简化，用计算机模拟相互影响过程而得到的结果，往往是反直觉的（Forrester, 1971）。这种只考虑相对简单的情节的倾向，对于冲突情境具有非常显著的影响。这里，某人获得自己的情绪和想法，当然比获得其对手的情绪和想法要容易得多。参赛者很难采用对手关于整个棋局或战场的看法，这也正是为什么当普通的竞争者在比赛中与对手转换一下位置之后，会发现更多新的制胜可能性的原因。因此，参赛者倾向于认为他的对手采用的策略是相对固定的，且与他自己的战术或招数（moves）毫无关系。这样的考虑意味着参赛者很容易受到主动性谬误（fallacy of initiative）的影响，这种主动性谬误，指的是参赛者认为自己比对手具有更多的主动性和想象力的倾向。这个假设，与琼斯等人（Jones & Nisbett, 1971）所做的一项归因研究的结果，非常吻合。在那个研究中，被试倾向于认定他们自己的行为反映了环境不断变化的要求，而他人的行为仅反映了他们的特质的影响。 178

虚构出一个极具说服力的情节，很可能限制人们未来的思考。有很多的证据显示，一旦某个不确定情境以某种特定方式被理解或者解释之后，就很难以另一种方式再来看它（可参考 Bruner & Potter, 1969）。因此，对某特定情节的虚构，将会限制其他情节的出现，尤其是那些会引出不同结果的情节……

偶然的事例或者情节所导致的便利性的影响，也许是现实生活中最明显的对于便利性的证明。很多读者都有过这样的经历，即亲眼看到一辆轿车在路旁翻车之后，其主观判断的此类事件的发生可能性就会瞬间上升。同样地，许多人应该留意到人们在看了某部生动地展现一次事故或故障引发一场热核战争的电影之后，对此类事件发生率的主观判断也会增长不少。连续地遇到同一事件就会提高该事件的便利性，从而增强对它的发生可能性的感知。人们总会专注于那些非常令人满意的事件或者非常令人不快的事件，比如赢得彩票大奖或者一次飞机坠毁。因此，便利性就提供了一种机制，使得那些非常有利的事件（或者非常不利的事件）看起来比实际情

况更可能发生……

本章是《认知心理学》(*Cognitive Psychology*, 1973, 4, 207-232) 上一篇论文的缩写。允许重印。

注释

[1] 现在所采用的“便利性”，与这个术语在言语学习文献中的一些用法，并不吻合（如 Horowitz et al., 1966; Tulving & Pearlstone, 1966）。

[2] 大约有 1 500 名被试，参与了这些研究。除非特别指明，他们在研究中都分成 20~40 人一组。研究 1、2、3、9 和 10 中的被试的招聘，是通过在俄勒冈大学的学生报纸上登广告的方式来进行的。研究 8 中的被试，是以类似的方式在斯坦福大学招聘的。而研究 5、6 和 7 中的被试，是以色列 10~11 年级的高中学生。

[3] 单词建构的问题，也可看成是提取问题，因为反应单词储存在记忆中。在这篇论文中，我们所说的提取，是指被试从自然类别或范畴中回忆实例，如研究 2 和 8 一样。而我们所说的建构，是指被试依照特定规则来生发范例，如研究 1 和 4 一样。

[4] 这个实例的选择，也是因为便利性。我们没有理由相信股民、运动员、政治分析家或研究心理学家在直觉预测中，会少有偏差。



第 12 章

便利性和归因中的自我中心偏差

罗斯和斯考利 (Michael Ross and Fiore Sicoly)

几乎所有从事过合作研究的人，都很熟悉我们现在的实验所研究的现象。试想，你和另一个人共同进行一个研究项目，涉及谁是“第一作者”（即谁对最终的成果贡献更多），问题就来了。通常，你们两人都认为自己完全有理由享此荣耀。尤其当你确信同事就事实真相（只有一种真相）也肯定认可你的观点的时候，你认为对方正试图占你便宜。有时候，采用随意决定的原则，可以解决或是避免这一忧虑，如字母优先原则，这是姓氏排在字母表开始部分的人最喜欢的一种方式。

由此我们认为，对于一个合作性成果，跟其他合作者所认定的自己的贡献相比较，个体倾向于认为自己承担更多的责任。进一步，我们主张当合作探索的责任被参与者分担时，这种现象尤其普遍。然而，在许多共同的努力中，因为没有必要去分配“著述权”，参与者并未意识到他们意见的分歧。其结果，这种现

象并不明显。我们现在研究的目的，就是要评估在种种情景中这种自我中心的感知是否存在，并且考察与之相关的心理过程。

在探索这些分化感知的基础时，我们不会天真地认为有意的自我夸耀从不存在。然而，在不存在有意识的欺骗时，这些感知，可能会有差别。就是基于这种视角，我们来研究这个难题。

180 为了给合作探索分配责任，善意的参与者大概都会试图回忆每一名参与者对最终成果所做的贡献。然而，互动的某些方面，比起其他方面，可能更容易或是更便利地被回忆。而且，更易回忆的特征，并非整体的随机子集。尤其是，同对其他参与者的贡献相比，个人可能记起其自身贡献的更大部分。

记忆中信息便利性的自我中心偏差，依次会在合作任务中产生有偏差的责任归因。正如特沃斯基和卡尼曼（1973，11）所阐明的，人们使用便利性——相关实例被回忆的容易性（1973，p. 209）——作为估算频率的基础。因此，如果自身的输入的确更为便利，跟其他合作者所认定的自己的贡献相比较，个体更可能在合作任务中声称自己承担更多的责任。

为了增加个人自身贡献的便利性，至少有四种过程在起作用：（1）信息的选择性编码及储存，（2）分化提取，（3）信息的不一致，（4）动机影响。

1. 选择性编码及储存

自生反应中的分化编码及储存，因为种种原因，会促进个人自身输入的便利性。首先，个体自身的思想（关于他们接下来说什么，白日梦等）或行动，可能会分散对其他人贡献的注意力。第二，个体会复述或重复自己的想法和行动，如在言说和辩护之前权衡自身的观点。其结果，他们自身的输入能够赢得更多的“学习时间”，而保持程度和学习时间是紧密相连的（Carver, 1972）。第三，个体的贡献，可能更容易地契合其自身的认知图式，也就是基于其先前的经验、价值等对特定问题的独特概念。而与这种先存的图式相契合的贡献，更可能被保存（Bartlett, 1932；Bruner, 1961）。

2. 分化提取

对记忆中的信息的选择性提取，也会产生便利性偏差。在合作性成果

的责任分配过程中,从每一名参与者的观点来看,基本问题是,“我做了多大的贡献?”因此,参与者会主要地尝试回忆自己的贡献,以及不适当地使用所提取的信息去评估自己的相对贡献。而这种判断,如果没有考虑他人的输入,是不可能合适的。

181

3. 信息的不一致

合作者可资利用的信息,可能存在差异。而这会助长自我中心的回忆。较之观察者,个体更易进入其自身的内在状态、思想和策略。而且,合作的参与者对于彼此独立贡献的比例及意义,也有不同的注解。例如,导师同学生相比,可能较少注意到学生在操纵被试、数据分析以及撰写论文初稿上所投入的时间、努力及独创性。另一方面,导师会更为看重学生参与之前自己所投入的思想、阅读等量值及其重要性。

4. 动机影响

动机因素也会调节便利性中的自我中心偏差。对个人输入(inputs)的注重和强调,会提高个人的自尊感。类似地,对于个人效能或控制的关注(参见 deCharms, 1968; White, 1959),能够使个体在合作成果中看重自身的贡献。

前面的讨论,概要描述了几种发生作用的心理过程,它们使得个体自身输入的呈现较其他人的贡献,更为便利(或更容易被回忆)。因此,想来否认记忆与归因是自我中心的假设,会很困难。然而,正如有人(Greenwald, 1978)所观察到的,记忆的自我中心特征“并非必然的真理。比如,对过去经验组织的构想,可能更类似于参考书,如历史教材或辞典的索引”(p. 4)。而且,我们还不能找到已发表的数据,来直接支持便利性中的偏差假设。最后,在有关行动者—观察者的归因偏差文献中,其最近的发展,似乎与记忆与归因的自我中心假设,并不一致。琼斯和尼斯比特(1971)推测,行动者倾向于将其行为原因归于环境,而观察者则将相同的行为归因为行动者所具有的稳定特质。尽管有种种方案来解释这种效应(Jones & Nisbett, 1971),但最近的重点集中于感知的信息加工(Storms, 1973; Taylor & Fiske, 1975)。行动者的视觉感受器指向环境,而观察者则直接集中于行动者。因此,行动者与观察者,会关注情境的不同

182

方面的显著性，反应在他们不同的因果归因上。但这一主张，似乎与互动中的行动者大部分是自我关注的说法相矛盾。

两个研究为现在的假设提供了启发性的证据。罗杰斯等 (Rogers et al., 1977) 表明，就特质形容词而言，当要求实验对象判断其自我关联性（如决定每一特质是否可描述自身）而不是判断其他维度（如同义词判断）时，他们更易回忆。这些数据意味着自我关联增强便利性。但是，他们并没有比较自我关联的形容词的回忆和他人关联的形容词的回忆。而这种比较，与现在的讨论更有关系。有学者 (Greenwald & Albert, 1968) 发现，在态度问题上，个体回忆其自身的论点，比其他被试的书面论点，还要准确。因为自我与他人的论点总是相左，因此，他们的发现能够令人信服地证明，与自身态度相一致的论点会更为熟悉，更易记住，而不是增强自发陈述的记忆（尽管有关态度一偏差的学习的证据是可疑的，如 Greenwald & Sakumura, 1967; Malpass, 1969）。

我们进行了一项先导研究 (pilot study)，来确定是否能支持便利性中的偏差假设。我们要求一个本科研讨班的学生估计每个同学在上节课中的发言时间。另外，我们还在自然情景中如餐厅及休息室中找了 26 个被试，他们两人一组。我们要求这些被试估计在互动中每人讲话时间在总时间中的比例。

我们假定，这些被试根据他们能够容易记起的谈话比例来进行时间估计。因此，如果个体在更易回忆自身陈述的方向上存在偏差，他们对自己发言时间的估计，应该超过组中其他被试对其发言时间的平均估计。

结果与此推论一致。在研讨班上，8 个学生中有 7 人对自己讨论时间的估计，超过其他被试对他时间的平均估计 ($p < 0.05$ ，符号检验)。类似地，13 对中有 10 对，自我的时间估计，超过对方对自己的估计 ($p < 0.05$ ，符号检验)。13 对中的偏差幅度是高度显著的， $F(1, 12) = 14.85$ ， $p < 0.005$ ；平均而言，参与者对自身发言时间的估计为总时间的 59%。这些数据初步表明，尽管并不直接，在日常情境中，存在以上所假设的便利性偏差。

一、实验 1

在这个实验中，我们希望考察在自然发生的持续关系中的自我中心偏差。已婚夫妇似乎是这一实验的理想目标群体。配偶们参与了许多重要程度各异的合作。这种场景似乎充满自我中心偏差的可能性。

因此，我们进行的第一个实验是为了：（1）确定婚姻关系中是否存在责任分配时的自我中心偏差；（2）用不同的因变量测量，来重复验证预研究中所获得的便利性中的自我中心偏差；（3）发现便利性偏差同责任偏差之间的相关。如果责任偏差由便利性偏差所引起，两组数据应有相关。

1. 方法

被试：实验被试由学生居住区的 37 对已婚夫妇构成，其中 20 对有子女。两名女实验助手入户招募这些被试，并概要描述实验程序。如果他们同意参与，就会签订一份协议。研究在实验对象的公寓内进行，每一对夫妇有 5 美元的报酬。

程序：基于对 6 对已婚夫妇广泛的初步访谈，我们编写了一份问卷。而在正式实验中，问卷由丈夫和妻子分别完成，并确保他们的匿名权。问卷第一部分要求实验对象对与已婚夫妻相关的 20 项活动的责任程度进行估计，估计方式是在一条 150 毫米长的直线上定位，直线两段分别标注“主要是妻子的”和“主要是丈夫的”。^[1]这 20 项活动包括：做早餐、洗碗、清扫房间、购物、照料孩子、闲暇时的活动安排、如何开销、选择住处、择友、影响二人的重要决策、争端、冲突解决、搞乱房间、洗衣服、同亲戚联系、关爱对方、倒垃圾、激怒对方、等待对方、决定是否要孩子等。

接下来要求实验对象简单记录他或者配偶对所有这些活动的贡献。随后考察他们的文字记录是否反映个体自身的记忆更为“便利”，或者说，他们记录的事例是否更加强调自己而非配偶的贡献？在不知道实验假设的情况下，一名统计者记录实验对象所提供的自己或其配偶的独立的事例，第二名统计者对数据的三分之一进行编码；信度（Pearson product-moment

correlation, 皮尔逊相关) 为 81%。

2. 结果

将配偶双方对每一个责任问题的反应汇总, 包括妻子视作其自身贡献的数目, 以及丈夫视作其自身贡献的数目。既然反应刻度有 150 毫米长, 可以分配 150 个“责任单位”。多于 150 的总和, 将表明在被感知的贡献中存在自我中心偏差, 因为配偶中至少有一方高估了自己在活动中的责任。为了评估配偶在每一项活动中所显示的高估或低估的程度, 在每一对配偶每一项活动的总数中要减掉 150。将 20 项活动进行平均, 从而得出配偶们的综合成绩 (如果配偶没有子女, 则为 19 项活动)。

将配偶双方作为分析单位进行方差分析, 可以看出综合成绩显著地大于 0, $M = 4.67$, $F(1, 35) = 12.89$, $p < 0.001$, 这反映出在被感知的贡献中存在自我中心偏差。37 对夫妇中有 27 对, 有高估倾向 ($p < 0.025$, 符号检验); 而且平均而言, 高估倾向表现在 16 个项目上, 甚至包括在消极项目上如两人争端, $F(1, 32) = 20.38$, $p < 0.001$ 。尽管平均而言, 高估的范围相对较小, 这显示出实验对象在试图控制量度的范围。大部分反应, 集中于量表中部稍高或稍低的区域。没有一个项目有显著的低估效应。

问卷第二部分的项目, 要求实验对象提供事例, 以记录自己或配偶对每一项活动的贡献。在 20 项活动中, 用配偶双方贡献的事例数目减掉自己贡献的数目, 可得出丈夫和妻子各自的平均值。大均值检验高度显著, $F(1, 35) = 36.0$, $p < 0.001$, 正如我们所预期的, 与配偶所提供的事例相比 ($M = 8.1$), 实验对象给自己提供了更多的事例 ($M = 100.9$)。自我—他人成绩差异和感知责任的初始测量之间的相关, 得到确定。正如假设, 对自我相关的行为回忆越多, 感知责任中的高估越大, $r(35) = 0.50$, $p < 0.01$ 。

实验对象所报告的每一行为事例中的字数, 同样被作为其回忆精致化或丰富性的评估尺度。每一事例是由自己 ($M = 10.0$) 还是由配偶 ($M = 10.1$) 所报告的, 两者平均字数并无差异, $F < 1$ 。而且, 这一测量与感知的责任测量之间, 没有关联, $r(35) = -0.15$, 没有显著性 (ns)。

总之, 对责任的测量以及反映相关行为便利性的测量, 显示了所假设的自我中心偏差。而且, 在便利性的偏差幅度和责任中的偏差幅度之间, 存在

显著相关。责任分配中的自我中心偏差由便利性中的偏差所控制，而这一发现同我们的假设也是一致的，即责任归因中的自我中心偏差受便利性偏差所调节。最后，所回忆的行为数量，而不是回忆的丰富性，更是重要因素。

二、实验 2

在实验 2 中，我们让 12 个大学院际篮球队的队员各自完成一个问卷，在问卷中，他们要回忆上场比赛的一个重要转折点，并且要评价他们球队胜利或失败的原因。

这是一次跨越，即从前一研究中的自我—他人比较跨越到我队—他队比较。然而，有很多理由可以预见，对归因者而言，我队的行动，较之他队的行动，会更为便利：我知道自己队友的名字，因此，我有便利的手段存储和提取与他们相关的信息；而我队在今后几场比赛中的胜利，则更多取决于自身进攻和防守的能力，而非对手的能力。其结果，我可能更密切地关注我队友的行动，而这会增强编码及储存。而且还存在信息的不一致：我队的策略比之对手更为显著 (Tversky & Kahneman, 1973, 11)。如果我队的行为动因是分化便利的，队员会依据我队的行动去回忆转折点，并且把比赛结果的责任归于自己球队。

1. 方法

被试：74 名大学院际女篮队员和 84 名大学院际男篮队员参与研究。通过电话与球队教练联系，他们与其队员商量后，全部同意其球队参加研究。

程序：在六场比赛后，我们对比赛的双方 12 支球队进行问卷调查。对入选的三场男子比赛而言，在所研究的六支男队中，三支对垒另外三支。类似地，三场女子比赛，则包括了所有参与研究的六支女队。在赛后的第一训练期间（赛后的一到两天），我们进行问卷调查。只有一次例外，由于比赛日程的缘故，需要在赛后及时收集数据（两支女队）。问卷由个人单独完成，保证参与者的匿名权。以现在的视角，有如下相关问题：

(1) 请概要描述上场比赛的一个重要转折点, 以及它发生的时段。

(2) 我队赢/输的原因是……

对第一个问题的反应, 是要确定转折点是否由我队、双方或他队所引发。而对第二个问题的反应, 则是来估计与自己或对手球队的行为相关的输赢原因的数目。统计数据, 由一位不知道实验假设的人员进行编码, 还有一位对 50% 被试的反应, 独立进行编码。对这两个问题, 两者编码均很吻合。

2. 结果

两次因变量测量, 不存在显著的性别差异。因此性别差异的结论不成立。因为队员的反应不能视作是独立的, 因此将他们的反应平均, 并将球队作为分析单元。

对于“转折点”数据的初步分析表明, 即使在同一球队内部, 队员也会回忆非常不同的事件。尽管如此, 有 119 名队员所回忆的转折点都由自己球队的行动所引发; 13 名队员所回忆的转折点由双方球队所引发; 而 16 名队员所认为的转折点则与对方球队有关 (剩下的 10 名队员没有回答这个问题)。他们所描述的事件, 如比赛最后两分钟的顽强防守, 一次防守反击, 进攻战略的转换等。

我们因此得出每支队伍中认为转折点由其队友引起的队员的百分比, 并将这 12 个数字同 50% 的机会预期进行对比分析。得到的分布与机会分布明显不同, $F(1, 11) = 30.25, p < 0.001$, 平均数为 80.25%。正如所假设的一样, 大部分报告强调了队员自己所在球队的作用。

将认为转折点由自己球队引起的队员的百分比同其球队的表现进行比较检验, 6 场比赛里有 5 场, 输队的平均百分比要高于赢队 ($p < 0.11$, 符号检验)。但输队 ($M = 88.5$) 与赢队 ($M = 72.0$) 百分比的平均差异, 并不显著 ($F < 1$)。

我们还考察了队员对球队输赢的解释。158 名参与者中, 只有 14 名队员提供了和对方球队作用相关的解释。平均而言, 实验对象所报告的与自身球队有关的胜负理由有 1.79 个, 而与对方球队有关的胜负理由则是 0.09

个, $F(1, 11) = 272.91$, $p < 0.001$ 。最后, 把原因归于自身球队的趋向, 与输 ($M = 1.73$) 赢 ($M = 1.65$) 的关系, 并不显著, $F < 1$ 。

3. 讨论

对转折点问题的回答表明, 被试队友的表现比对方队员的表现, 能更便利地提取。而且, 被试将比赛结果的责任, 更多归于其队友的表现 (有为或懈怠), 而非对方球员的表现。因此, 在群体水平上, 也存在便利性偏差和责任判断偏差。有人 (Rather & Heskowitz, 1977) 提供了关于群体自我中心的另一例证: “在 1968 年阿波罗登月之后, 哥伦比亚广播公司 (CBS) 的新闻成为绝对的头号。如果你是 CBS 人, 你会说我们的登月直播无与伦比。而如果你是美国广播公司 (NBC) 人, 你会将联队 (Huntley-Brinkley team) 的瓦解作为关键因素。” (p. 307)

三、实验 3

在实验 3 中, 我们试图改变个体注意的焦点来影响便利性。我们的实验操纵, 是为了促进与责任归因直接相关的信息的选择性提取。

在最初的分析中, 我们认为记忆中信息的选择性提取, 会产生自我中心的责任归因, 并且这种提取会受个人自我追问的各种问题的指引。实验 3 即用来验证这种假设。通过不同的问题提问的方式, 被试参与不同的提取任务。我们激发研究生来权衡他们的学士论文中自己的贡献或导师的贡献大小, 然后评估被试分配给自己或是导师的与论文有关的责任量。我们假设在导师中心而非自我中心的情境下, 实验对象对于研究成果会承担较少的责任。

1. 方法

被试: 被试由 17 名女性和 12 名男性心理学研究生组成。他们大多数在研究生院上了 1~2 年。所有学生, 在本科的最后一年, 做过学士论文所需的实验。

程序: 我们在其办公室里个别接触这些被试, 要求他们完成有关导师—学生关系的简单问卷。没有人拒绝。两种形式的问卷, 随机地

发给他们，并确保其反应的匿名性及保密性。

一种形式的问卷，要求被试说明自己在与自己学士论文有关的各项活动中的贡献。问题如下：“（a）研究中最后被采用的方法论的____%由我提出；（b）我提供了结果阐释的____%；（c）同导师进行的与论文相关的讨论____%由我发起；（d）在与论文相关的讨论中，我对讨论内容和过程的控制时间约占____%；（e）总的来说，我对整个研究成果的____%负有责任；（f）与系里其他人相比，你如何评估自己的论文？”

问卷的第二种形式与上述相同，只不过将问题1~5中的“我”（自我中心情境）换成了“导师”（导师中心情境）。要求被试对前5道题进行填空，而对第6题的回答，则是在一条150毫米长的直线上定位，直线两端标注“低劣的”与“优秀的”。

2. 结果与讨论

为了分析，我们假定在每个项目上，导师与学生的贡献总计为100%。尽管我们声称实验的目的是研究导师—学生关系，学生仍可能会在他们的评估中考虑到其他人的工作（如同学）。即便如此，现有的实验程序，还是能对实验假设进行保守的检验。例如，如果被试对“我”版本的问卷中的一个项目的反应是20%，这意味着他的导师的贡献是80%。但导师可能贡献了60%，其余的则由他人完成。然而，通过高估导师的贡献，我们的数据存在偏差以反对实验假设：版本1的问卷，会降低分派给导师的责任比例。

在“我”版本的问卷中，被试对前5道题的回答，从100中减掉，从而较大的数字能够反映导师在两种情境中较大的贡献。而第5道题，则涉及研究成果的全部责任。和预期一致，同自我中心情境（ $M = 16.5\%$ ）相比，在导师中心情境（ $M = 33.3\%$ ）中，研究对象将更多的责任归于导师， $F(1, 27) = 9.05, p < 0.01$ 。前4道题涉及论文的不同方面，其平均发言有类似的结果：在导师中心情景中， $M = 33.34$ ，而在自我中心情景中， $M = 21.82$ ； $F(1, 27) = 5.34, p < 0.05$ 。最后，在自我中心情境中，比之导师中心情境，被试对其论文的评价，更为积极：112.6对94.6，

$F(1, 27) = 3.59, p < 0.10$ 。

问题的文字差别,对被试的责任分配具有所预期的影响。同“我”版本的问卷相比,导师版本的问卷,能够使学生更多地回忆导师的贡献。这一分化的便利性,在随后的责任分配中也体现出来。不过,值得注意的是,问题的文字差别,并不能完全成功地控制被试的提取。在导师中心的情境中,导师对论文只被分配了 1/3 的责任。

根据现有的数据,实验 2 中篮球队员对比赛结果的责任归因,需要重新检验。回忆一下要队员完成的句子,“我们球队赢/输掉上一次比赛是因为……”这个问题产生了高度显著的自我中心偏差。事后来,很明显,问题的形式如“我们球队……我们上一次比赛”,会促使被试关注自己球队的行动,尽管文字表达并未排除提及对手。而实验 2 中的“转折点”问题的文字表达,则更为中立,不会导致这种模棱两可的阐释。

在这些研究中,主要问题由外部(如实验者)提起。然而,我们许多提取性提问是自我发动的,我们的回忆,由于我们有关提取问题的自我提问的方式,而存在很大偏差。如篮球队员的思考,更可能根据“为什么我们赢/输了”的方式,而不是依据中立的语句“哪支队伍该为比赛结果负责?”……

现在的研究表明,在便利性和责任判断中,普遍存在自我中心偏差。日常生活中,当合作工作并不要求明晰的责任分配时,这些自我中心倾向可能会被忽视。但是,如果清楚阐明责任分配,则会有潜在的纠纷;而个体也不会意识其判断的差异,是由对信息的诚实评价所引发,而这种信息是分化便利的。

本章原载《人格和社会心理学杂志》(*The Journal of Personality and Social Psychology*, 1979, 37, 322-336)。允许重印。

注释

[1] 在最初的访谈中,我们采用百分数估计。我们发现被试能记住自己所报告的百分数,而问卷后的百分数比较会在配偶双方引发强烈的冲突。而采用 150 毫米的量度,可避免这些困难;被试不会把自己在 150 毫米量度上的定位,转换为会在以后引发争执的精确百分数。

第 13 章

社会知觉和互动中的便利性偏差

泰勒 (Shelley E. Taylor)

190 每天，社会认知者都进行无数复杂的社会判断，如预测他人的行为、判别责任归属、将个人分类、评价他人、评估一个人的权势或影响力，以及进行因果归因。社会心理学的中心任务，就是确定社会认知者是如何做出这些判断的。直至最近，有关这一主题的研究，始终以理性主义的偏差为特征，即假设社会判断是深思熟虑和最优策略的结果 [例如请参见费希荷夫 (Fischhoff) 在 1976 关于这一点的讨论]。判断中的误差，被归结为两个源头：一是由于社会认知者可能未意识到的信息而带来的偶然误差；二是由认知者的非理性动机和需要而导致的误差。

在社会心理学领域，有关归因的研究体现了这一视角。早期的归因模型 (例如, Jones & Davis, 1965; Kelley, 1967)，假定认知者作为一个朴素的科学家，通过从周边环境多渠道地收集资料来进行因果关系的归因。一旦观察到与这一规范模型的偏离，人们就相

信它们源于一些偏差，诸如享乐主义的关怀或其他的自我中心的需求（参见 Miller & Ross, 1975）。

然而，历经多年不断增多的证据表明，人们的判断和决策，不但不及设想的那么完整和理性，而且也并非所有的误差都可以追溯为动机性因素。甚至在不考虑动机的前提下，判断也经常基于少量的数据；而这些数据似乎是无序的组合，并受到先见（preconceptions）的影响（参见 Dawes, 1976）。这些发现致使人们重新审视和修正认知系统。人被重新定位为能力有限的个体，即一次只能处理一小部分信息。认知者，不再被视为进行最优选择的朴素科学家，而是在进行“满意”选择，并借助捷径高效而准确地做出决策和判断。

在对判断过程的修正研究中，最具影响力的贡献之一，当属卡尼曼和特沃斯基关于认知启发式的研究（Kahneman & Tversky, 1973, 4; Tversky & Kahneman, 1974, 1）。根据特沃斯基和卡尼曼的研究（1974, 1），在重要信息不确定、不可获得或者不明确条件下，认知者将采用启发式。在非社会判断中，不确定性主要是因为有关特定判断的信息几乎总是不完备的。合适的事实材料可能无法接近，也可能无法被及时地收集以辅助决策，或者也可能因为数量过于庞大，无法在一次判断中被适当地组织和利用。

启发式是社会判断和非社会判断的可能策略。其主要原因如下：第一，在社会判断和非社会判断之间的区分是武断的，因为任何重要的判断实际上都有社会后果。第二，社会判断也牵涉到同类的不确定性，而它们则是非社会判断的特征。第三，社会判断包含不确定性的新来源。相对于有关物体和非社会事件的信息而言，有关人的信息，更为模糊、更不可靠，也更不稳定；因为人们并不像客观物体外显其颜色、形状或大小那样，并未外在地表现出其个人品质。从而，个人品质，必须通过推断而非直接地观察。人们的意向，并不都是可直接说明的。假定大多数重要的社会行动的发生，有种种理由，并将产生一系列后果，那么社会行动的意义，基本上是模糊的。尽管客观物体的属性，具有跨情境性和历时性，但是人的动机，则随情境的变化而变化，并且，其目标也会毕生随时变化；因此，即使特定情境下的判断是准确的，其预测效用也是很差的。想要获得完整、可靠

而又具预测价值的人和社会互动的信息，是不可能的；这种不可能性，促使人们在信息不充分、数据不可靠的情况下，采用启发式进行推理和预测。

1. 便利性启发式

启发式的一种，可以被称为便利性启发式 (availability)。“当一个人通过他可回忆的实例和联想的容易性 (ease) 来推测事件发生的频率和概率时，他便是在采用便利性启发式。” (Tversky & Kahneman, 1973, p. 208, 11)

1.92 这里包含了以下的假设，即容易忆起的例子是大量存在的；而容易生发的联想，既然是基于现实经验的，那么也是准确的。而且，人们在频率和概率推测时，依据的正是提取、建构和联想的容易程度，而不是对可能回忆到的实例和联想所做的总数统计。这样看来，使用便利性启发式和使用更精确的推理过程的一个重要差别在于，前者几乎不需要完成实际的信息提取和建构，而对于信息提取和建构过程的操作容易性的推测，就已足够作为推论的基础。

特沃斯基和卡尼曼 (1973, 11) 提出，便利性偏差会明显出现在两大类认知任务中：一类是实例和联想的建构过程，另一类是联想和实例的提取过程。这两类任务，也就是社会认知者所要完成的任务。在许多情况下，我们会被要求建构社会行为，比如试着猜测如果某个朋友的妻子将要离他而去，他会做何反应。在这些情形中，人就在建构社会实在，以此和实际的社会事件相比较。在其他的情形中，一个人则可能利用既有经验进行推论，如依据这个朋友以往在应对危机时的表现，来推测他现在将做何反应。

在建构或提取过程的难易程度与实例或联想的多寡之间所做的假设，在某种程度上是准确的；并且在假设准确有效的范围内，利用启发式可以做出正确的推测，或者说与详尽而又费神的认知程序相比，其准确性并不逊色。然而，在其他情况下，这些推论，很可能是不准确的，因为有关问题的可资利用的信息中，存在着偏差。在可资利用的信息中所存在的偏差，至少以三种途径，使后继的社会认知过程产生偏误。其一，高度显著性的信息更为便利，从而对判断过程产生或多或少的影响。其二，提取过程中的偏差本身，可能产生并不具有代表性的数据库。其三，认知者持久的认知结构，如信念和价值，会滋养先见；这会抬高某些证据的便利性，从而

使判断过程产生偏误。其中任何一种证据上或过程中的偏差，都可能导致偏差性推论。

2. 便利性和显著性偏差：一个实例

显著性偏差，所指的是一个事实，即色彩鲜艳的、活动的或者其他显著的刺激，或多或少地吸引了认知者的注意力，进而或多或少地对判断造成影响。这种偏差的一个例证，被称为基本归因误差（参看 Ross, 1977）。作为一种普遍的偏差，它把人看成是环境中的因果行动者（causal agents）。也就是说，当社会情境中的候选因素——个人变量或情景变量，都可能是某结果的原因所在时，通常的偏差是把人视为导致结果的原因，尤其是其持久的品性特征。

一些人比另一些人更为突出和显著。而在社会环境中，这一分化的显著性，也将使判断过程产生偏差。应用格式塔原理中图形比背景显著的许多研究（如 McArthur & Post, 1977; Taylor & Fiske, 1975; Fiske et al., 1979），揭示了一个酩酊大醉的、移动不定的（比如在摇椅上）、对比鲜明的（比如通过衬衣颜色等看来不起眼的琐碎安排）或者与众不同的，或多或少将会吸引相当的注意力。

一些研究，雄辩地证明了显著性偏差的社会后果。这些研究，考察了孤立地位（solo status）和象征性整合（token integration）对人们所持有的个体印象的影响。当一个团体欲废除种族隔离，并接纳少数群体的成员如黑人、女人或是伤残人时，在达到完全整合之前，将是一种试探性的协调阶段。在这一阶段，原被拒群体中的一两个成员，可能被纳入一个完全不同的白人男性工作群体，从而同时创造了孤立地位的实例。孤立地位的形成有多种原因。可能是由于特定的少数人群体中缺乏合格的申请者；或者可能为了避免大量少数群体成员对组织造成威胁；或者一个孤立者被用来逃避平权运动力量（affirmative action forces）的指责。为了象征整合而不论是什么原因，象征的或是孤立的个体，通常被视为他或者她所属群体的代表。据此，对他或者她的表现所作的评价，常常被套用于其所属群体的其他成员，以预测他们若是加入组织后的可能表现。孤立地位的重要性，就在于其新奇性。在一个其他均为男性的群体中，一位女性突显出来，就像

一位黑人在一个均为白人的环境中。这一独特性，造就了显著性偏差。

在对这种情景的实验模拟中，被试观看了一个小组（6个人）非正式讨论会的视听资料。部分被试观察的是一个黑人、一个男性或者一个女性分别在白人、女性和男性组的情况（Taylor et al., 1976; Fiske et al., 1978）。另一部分观察的则是比较组，即讨论的内容与孤立组完全一样，但是小组的性别和种族构成则是均分的（比如，三男三女）。被试所做的，是观察小组讨论，并记录下他们对小组中个体的印象。

与先前的论点一致，孤立地位或象征性整合，产生了显著性偏差。同一个黑人，被试对其印象，在平均混合组和孤立组的条件下，略微不同：194 他在后一组内的行为能被更好地回忆起来，并被认为参与了更多的讨论。^[1]由此，这种显著性偏差，依次导致了评价的极端性。当个体的同一种行为在组内处于孤立地位时，较之发生在混合组内部，所受到的或积极或负面的评价，将被极端化。当处境孤立时，一个讨厌的人将被认为更加令人生厌，一个友好的人也将被认为更加易于相处。

可以推断，利用便利性启发式作为解释框架，当被试者被要求评价群体中的一个成员时，他们会尝试使用与刺激个体相关的实例和联想；而回忆实例和联想的容易性，导致他们判断的极端化。就孤立个体的情形而言，因为存在较大的数据库，因此，相关行为有许多例证可资利用，由此导致评价的极端性。这些结果强有力地表明，一个与众不同的个体，比如孤立的、伤残的或者异于与之互动的其他个体，都将在其周围人中引发极端判断。这些发现，对于我们理解现实社会的启示，意义深远。举例来说，如果孤立者的表现作为废除种族隔离的决策基础，并且对孤立者的知觉是天然有偏差的，那么，误差的决策就会出现。对一个孤立者夸大的负面评价，会毫无根据地导致废除种族隔离计划的流产；另一方面，夸大的积极评价，则会导致对少数人群体中的其他成员的误差预期，而这种预期通常难以达到。

很显然，动机不能解释在感知显著他人时所产生的这些偏差。虽然一些动机性过程可能卷入了对伤残人或孤立者的反应，但在评价一个衣着怪异色彩或举动特异的个人显著性时，这一过程却很难被发现。由此，人们

必须仔细考察各认知因素，而便利性启发式，则为这些效应提供了一种可能的解释。

3. 便利性和提取偏差：两个实例

便利性偏差产生的第二条途径，源于信息的储存或提取方式。也就是说，以特定方式组织的记忆，它会促进特定种类的实例或者推论的提取或建构，并且会阻碍其他实例或者推论的提取或建构。阻碍的途径之一，在于记忆中可存储的信息量是有限的，而这会导致所存储的联想路径和实例产生混乱。

罗斯巴特和他的同事 (Rothbart et al., 1978) 在有关社会群体特质的研究中，论证了这个问题。首先，告知被试假定的群体成员的特质信息（如 Phil 是懒惰的），而这在两种不同的条件下进行：一种是由几个不同成员（如 Ed, Phil, Fred, Joe）构成的名单，它和特定的特质（懒惰的）配对给出；另一种是以等价次数的“姓名—特征”对的形式给出。随后，要求被试描述群体的整体特征。如果被试能够精确地记住姓名和特质的对应，同时好几个成员而不是仅仅只有一个都具有某一特定特征，那么对该群体的整体推断，将更加肯定。当被试接触到的“姓名—特征”对总数较少时，他将谨慎地推测群体特征。然而当“姓名—特征”对总数较多时，被试就很难记起哪些成员确实有哪些特征，于是转而认为同样的“姓名—特征”对的许多例证，和几个不同的成员姓名都具有那个特征的情形，两者提供的是等价充分的信息。因此，即使只有少数成员实际上是懒惰的，该群体也会被标定为“懒惰的”。

现实社会是瞬息万变的，信息通常扑面而来。这种情形，更多反映在罗斯巴特及其同事所界定的高信息负载的条件下，而非低信息负载的条件下。这些条件，利用部分个体的多频次行动而促进了群体刻板印象的形成。正如罗斯巴特他们所注解的例子，媒体更容易被负面事件而不是积极事件所吸引，由此而产生的所有这些偏差，会促成消极群体刻板印象的形成；尤其当成员身份被新闻媒体醒目提及时更是如此。

提取偏差对社会判断产生影响的第二个实例，就是自我中心的归因 (Ross & Sicoly, 1979, 12)。在很多情况下，个体必须判断谁该对既成事实

负责。论文的冠名权需要决定，咨询费用需要在同事间摊派，群体成功或失败的责任需要分明。在每个人都无法清楚地列出自己和他人分别所作的贡献时（曾经勉为其难地采取该策略的人，都深感沮丧），所有的利益方，都必须进行近似的估计，判断每个成员在合作中的努力大小。便利性启发式，为之提供了一种潜在的策略。一个人可以仔细地回想自己和他人分别作出贡献的具体实例，并在此基础上，根据所能想到的两者实例的多寡判定谁付出的更多。然而此时，偏差就可能出现，干扰评价的精确性。一种假设的偏差，就是自我中心的回忆，即在某种程度上更容易回忆起自己而不是他人的贡献的能力。这种回忆偏差，依次产生了责任认知偏差。举例来说，如果我能记得自己曾经倒了6次垃圾，而我的丈夫才3次，那么我就可能得出结论说，倒垃圾的事由我承担了。最近的研究（Ross & Sicoly, 1979, 12）表明，婚姻家庭的杂务琐事，是检验便利性偏差有用而中立的语境。³⁷ 对夫妇接受了相关访谈，访谈涉及已婚者必然要经历的各种行为活动，比如做早餐或者洗碗。访谈要求每对夫妻双方分别指出，某项活动主要是由双方的哪一方承担的。之后，每个人都回忆了合作过程中自己和配偶分别出力的具体实例。这个设计，是为了测量回忆自己和回忆配偶贡献度的相对便利性。

其结果清楚地表明，配偶双方都认为各自较对方在合作活动中贡献了更多。如果把双方的责任评价得分加和，总数就超出了可能承担的责任总数。同时，配偶双方所作贡献的具体实例，也分别加以计数，结果表明归功于自己的远远超过归功于自己配偶的。能回忆起来的贡献实例和责任归因之间是高度相关的（+0.51），它意味着回忆偏差很可能导致了责任认知偏差。

对这种效应，一种可能的解释建立在动机原则的基础上。该解释主张，人们有失公平地过分归功于自己是为了维持或加强积极的自我意象。如果这种说法成立，那么我们应该发现，若一项合作失败，人们将会否认个人责任并把责任归咎于他人。比方说，如果一对夫妇购买了一件次品，任一方都会把责任推脱给另一方。罗斯等人（1979, 12）检验了这种可能性，但发现自我中心归因的解释力很弱。在合作项目的失误和成功上，人们的

归咎和归功的比例，几乎没有差别。

而对此效应的认知解释，则注重信息的提取或存储如何产生可觉知的责任偏差。一种暴露偏差 (exposure bias)，也许存在。因为比较而言，一个人观察自己的贡献更近；从而在对责任进行认知重构时，自己的功劳更多地被提取。例如，如果夫妇俩同时工作，一方可能根本不在意对方的贡献，而只是专注于观察距离更近的自己的贡献。也有可能，一方没有意识到配偶的贡献，因为当配偶在分担自己的工作时，自己并不在场；因而，自己很可能低估了配偶实际投入的时间和精力。其二，如果一个人更为频繁地考虑自己的贡献，偏差同样会出现。特别是在合作项目牵涉到一系列思考、写作或其他复杂的工作时，比如筹划一次家庭装修，或者组织一次家庭派对。第三种可能性在于，自己的贡献，更为吻合其个人构念和认知图式，而这正是信息铭刻或编码的方式。与既存的图式相吻合的信息，似乎更容易被回忆，因此，所提高的回忆本人贡献实例的便利性，也就反映了提取偏差 (参看 Taylor & Crocker, 1979b)。总而言之，信息的存储和提取的偏差，会导致有偏差的社会判断。一个可能的中介，就是便利性启发式的使用。

197

4. 源于认知结构的便利性和偏差

便利性还可通过第三条途径使社会判断产生偏差，即社会认知者利用习以为常的规则、图式或其他认知结构。人们在处理新输入的信息时频繁地利用持久的认知结构，长此以往便将其作为认知习惯加以利用。比如我们知道，人们更可能利用一些特质来组织与人有关的信息，而忽略另外的特质。例如，学者经常用智力这种特质，来对人进行区分；而对运动员来说，是运动技能而不是智力，成为区别的因素。这种利用特定图示或认知构念的倾向即偏差，同样会使判断产生偏差。例如，如果一个人被问及他对某个特定个体的印象，他可能会以自己偏爱的特质作为回忆的线索，并且他的描述较少是依据该个体的实际行为，更多的则是依据那些他自己偏好的品质 (参见 D'Andrade, 1965)。类似的偏差，可能出现在社会行动的结构中，因为人们在面临新情景或者被要求对未来做出预测时，同样会借助于其个人构念和图式。比如说，一位学者将会依据他的朋友有多机敏来预

测其摆脱困境的能力，而体育运动员依据的，则会是他的朋友能跑多快。据此，一个人基于规则、图式和构念所做的推论，将与他人不同，除非他们之间分享同一的认知结构。

刻板印象，也许是这方面最为引人兴趣的例子，这一现象展示了持久的图式结构对信息认知的影响。预期一个人的特定行为，会导致他已经这样行动的推测。刻板印象，是特定的期望，其功能在于指导和形塑现实；而其功能的发挥，至少是部分借助便利性偏差。哈密尔顿和罗斯（Hamilton and Rose, 1978）在其刻板印象研究中，探索了这种可能性。在一次研究中，给被试一系列描述性的句子，而描述以这样的格式给出：“卡洛是一名图书管理员，迷人而严肃。”在每个句子中，对应于一种职业的一个198人，都是有两种品质。一些品质带有与该职业种类相关的刻板的联想，比如上例中的“严肃”是“图书管理员”这个职业的刻板印象；而另一些品质，对于该职业来说是非刻板性的（比如“迷人”对于“图书管理员”）。在所有的句子中，每种品质，以均等次数分别与每种职业配对。其中和刻板印象配对的比率，是1/3，和非刻板印象配对的比率是2/3。然而，当被试被要求估计每种品质描述每种职业中的成员次数时，他们有选择地遗忘了某些“品质—职业”对，以偏好刻板性的联结。比方说，他们更有可能记住是图书管理员而不是女招待，是严肃的。

虽然说动机因素对这种情况也可提供一种解释，但认知者却未必会这样行事。一些刻板印象，可能在个体的需要和目标中异常突显，但有关职业的刻板印象，诸如女招待或者图书管理员之类，已是相当老套，因而很可能并不明显。据此，这些结果，最好作为一种认知现象来理解。便利性启发式提供了一种可能的解释，其假设是，当要求被试回答品质和职业的配对频率时，他们是依据职业和品质间的关联强度进行估计的。在许多情形下，在许多次的配对后，关联强度会增大。然而，就这个案例而言，因为刻板印象存在于一些“品质—职业”配对中，这一关联强度，产生了偏差。其结果是，因为这些关联较强，认知者因此会推测刻板的“品质—职业”配对，比之非刻板的“品质—职业”配对，更为频繁。并且，这种偏差的社会内涵，是显著的。刻板印象一旦形成，就会使信息的收集和储存

以及随后的印象，都产生偏差。其后果就是，对社会群体或个人的所作的推论即使毫无根据，也可能在经验缺失的条件下永远存在下去。

5. 结论

在过去的几十年中，我们已经见证了研究视角的转变。判断一度被视为理性的、逻辑决策的产物。这种观点，由于受到偶然出现的非理性需求和动机的攻击，而趋向另一种观点。后者将人视为启发式使用者。有关非社会判断的经验研究表明，认知者使用捷径或启发式来尽可能迅速地解放认知能力和传递信息。而近期的社会心理学的研究也表明，这些认知过程同样适用于社会判断的形成和使用。回忆实例或联想的容易程度，提供了一种可能性估计（如频率或概率），而这又以此为其他的社会判断提供基础，如评价他人、因果和责任归因、描述他人特征、把他人归类或者自我描述等。

敏锐的批评家会指出，在上述研究中，有关应用启发式的证据是推论性的，而不是直接的。他们完全有理由要求更清晰的证据，以表明正是回忆实例或联想的容易程度，实际上在调节判断。三个原因，造成了这种暧昧不清。第一，这些社会研究，和有关便利性的认知研究不同，其中没有一项是专为检验便利性而设计的。更确切地说，它们是为了阐明一些本质上令人好奇的社会现象，而便利性启发式，只是作为一种可能的解释被提出来。第二，到目前为止，在便利性的测量问题上，还没有达成一致意见。有些研究把信息提取的速度作为测量便利性的方法（参见 Pryor & Kriss, 1977），而另一些研究，则关注能回忆的信息量作为衡量信息便利性的指标（参见 Ross & Sicoly, 1979, 12）。因为这两种测量并不总是有很好的相关（参见 Pryor & Kriss, 1977），因此测量的含混不清阻碍了经验研究的进展。

然而，有关测量的问题，仍是悬而未决。其部分原因，是由于第三个问题，即便利性这一术语在使用上的概念模糊性。有一种很浅显的理解认为所有的社会推论都是由便利性所调节的；而人的判断，总是基于其所能想到的。便利性的这种用法，必须和便利性启发式严格地区分开来。而这种区分，必须通过清晰的标准，以决定是否和如何利用启发式，以及实例和联想的便利性是否和如何调节后续的判断。社会心理学家的重心，

已经关注哪种信息是便利的，其原因何在？便利性启发式是如何与认知过程中的具体内容相联系的？对社会现象解释水平的提高，要求我们转而在这个问题上作进一步的努力。

便利性这一概念在社会心理学中的重要影响，毋庸置疑。它主要有以下几个原因。第一，它强调即使不诉诸动机性构念，也可以对信息处理中的偏差做出解释。这并不意味着动机对判断的影响不再重要，而仅仅是指出了主要偏差同样存在于认知中。第二，虽然理性主义理论已经提供了卓有益处的规范模型，用以比较实际的判断，但是与这些模型的偏离，是如此常见和明显，因而发展描述性的模型很有必要。在这里列举的和许多其他的例子中，便利性启发式已经提供了一种可能的描述。在某些情况下，便利性启发式的使用，带来了完全恰当的结论。然而，在便利的信息存在偏差的其他情形下，误差的推论便随之而来。具体说来，显著性偏差、提取偏差，和源于认知结构的偏差如图式、信念和价值，都可能在社会判断任务中使误差信息或误导信息的便利性大大提高。

既然便利性的偏差如此普遍，为什么还要继续使用便利性启发式呢？
200 一个显而易见的理由是，由此产生的正确答案多于误差答案，并且从成本收益角度考虑，我们成功地节省了时间和精力。第二种理由在于，由便利性启发式所产生的许多结论中的误差，是无关紧要的。举例来说，如果有偏差的印象对其将来发挥个人作用没有影响，一种情况是如果你对个人产生了误差的印象，而你们只有一次见面机会，那么这个误差就是无关紧要的。再有一种情况是发生这种便利性偏差的情境是固定不变的。例如，如果一个员工认为他的老板态度粗暴，同时如果老板仅仅在身处“老板”这一角色时才脾气暴躁，而职场又是该员工与他互动的唯一场合，那么该员工的看法并无利害关系。第三种理由，是许多误差能够被纠正。只要便利性偏差不是历时相关的，过程本身就将随着重复的遭遇而不断修正。举例说来，如果某人的好几个朋友都在最近离婚了，那他对离婚率的估计可能会暂时偏高。但假定他的朋友们并没有继续不断地离婚，那么他的估计，最终会和客观数据一致。最后的一种理由在于，误差通过交往可以被察觉。举例说来，如果一个人声称他完成了应承担的部分家务，却遭到其配偶的

愤怒反对，那么这种立场就有可能被修正。简言之，正常的社会交往，为人们真实检验自己的推论提供了基础，明显有误差的结论及其深远广泛的影响，很可能得到纠正。

但是在某些情况下，误差的认知及其所造成的严重后果，将会持续。举例来说，正如有关刻板印象的研究所证明的，如果在判断的形成中存在偏差，这种偏差，通过既已形成的认知结构（即所谓的刻板印象），依次导致维持这种判断的偏差。这些密切关联的偏差，作为结果，可能会高度抗拒相反的证据。在其存在的范围内，它们很可能对个体认知和社会政策都产生破坏性的影响。简而言之，不能忽视便利性偏差的后果；也正因如此，察觉和纠正偏差性推论的策略，是必须的。

本章的研究，受国家科学基金会（NSF）的研究基金 No. BNS77-09922 的资助。

注释

[1] 在这些研究中，没有收集孤立男性和孤立女性的回忆数据。



第 14 章

模拟启发式

卡尼曼和特沃斯基 (Daniel Kahneman and Amos Tversky)

201 我们之前对于便利性启发式的研究 (Tversky & Kahneman, 1973, 11), 主要讨论的是两类“能把事例带到脑海中”的心智操作: 对事例的提取以及对例子或情节的建构。回忆和建构, 是两种完全不同的把事例带到脑海中的方式; 它们用于解答不同的问题, 并且遵从不同的规则。过去的研究主要关注的是从记忆中提取事例, 却相对忽视了心智建构这一过程。

为了推进我们对于建构便利性 (availability of construction) 的研究, 我们现在简单勾勒出一个被我们称为模拟启发式 (simulation heuristic) 的心智操作。我们的出发点是基于普通的内省, 显然在生活中存在许多这样的情境, 即我们通过某种操作来解答关于某事件的一些问题, 而这种操作类似一个模拟模型的运行。这种模拟, 以几个方式受到抑制和控制: 该模型运行的起始条件, 是所模拟事件在现实中的系统设定值, 或者稍作调整以假定某特定突发事件; 模拟

的结果可以是任意的，或者也可设定一目标状态，而模拟的任务就是要找到一条连接起始条件和目标状态的路径。一次模拟，并不必然产生一段唯一的故事，它开始于起始状态，结束于某特定结果。相反，我们将模拟的结果作为一种对容易程度（ease）估计的方式，基于这种对容易程度的估计，这个模型在给定的初始条件和运行参数下能得出不同结果。因此，我们假定，这种心智模拟得出了对某人的能产生出不同结果的情境模型的品性（propensity）的测量，就好像蒙特·卡洛（Monte Carlo）技术能估计出某统计模型的品性一样。对某系统的模拟得出一特定结果的容易程度，最终能被用于判断现实中该系统达到那种结果的品性。 202

我们将论证，在一些判断任务中，会用到这种依据心智模拟估计品性和概率的方式，而且它们在一些情感事件中也扮演了重要的角色。我们首先会列出一些判断活动，其中显然与心智模拟有关。然后我们会介绍一个关于一些认知规则的研究，这些规则主宰着人们在脑海中取消过去发生事件的过程，或心智取消（mental undoing），我们还会简要地讨论到这些规则对于人们在比较现实情境与某个他期望的情境时所产生的情绪的意义，这个期望情境虽然在现实中未能实现，但人们很容易在脑海中想象它的达成。最后，我们通过对情节以及模拟启发式的运用所带来的偏见的评论，来结束对模拟启发式的简要的勾勒。

（1）预测。想象一下你非常熟悉，但他们彼此从未谋面的两个人的第一次会面。你如何得出诸如“他们会相处得很好”或“他们会激怒对方”的预测？

（2）估计某特定事件的发生概率。你如何估计为保护沙特阿拉伯的油田，美国在下个十年中会进行军事干预的可能性？请注意这个问题与前一个预测问题的差别。对当前这个问题的模拟，有一个很明确的目标状态，模拟的任务，是要得到某种对该目标状态在国际体系的现实模型局限中得以实现的容易程度的测量。

（3）估计条件概率。如果沙特阿拉伯爆发内战，可能带来什么样的结局？注意这个模拟活动不同于单纯的预测，因为它包含了一个特定的起始状态，而且这个起始状态多多少少与目前的现实不相符合。具体来讲，对

这种可能性比较小的突发事件的估计，包含了一种有意思的模糊性：在模拟进行之前，我们应该如何改动当前的模型？我们应该遵照基本的一致性要求，仅对模型做小小的修改，将这个特定的突发事件（如沙特阿拉伯爆发内战）包括进去就可以了吗？或者我们应该将由这个突发事件所引起的所有可能变化都纳入到模型中去？比如，就沙特阿拉伯的内战这个例子而言，我们关于政治体系的模型应该首先进行调整以使其内战的爆发尽可能地显得比较合理，而模拟时就应该采用这个修改后的模型的参数。

（4）反事实的估计（counterfactual assessments）。希特勒聘用的科学家们，在第二次世界大战时是否差点就制成原子弹了？如果他们在1945年2月就研制成功了，那么这场战争的结果会不同吗？在许多生活情境中也会用到这种违背事实的估计，如当你判断“如果她的孩子没有得病的话，她是能够应付这种工作处境的”时，就是采用的一种反事实的判断。

203 （5）估计因果关系。为了检验A是否导致了B，我们会在脑海中将A取消，并观察B是否仍然在模拟中出现。模拟也能被用于检验A是否极大地增强了B的趋势，甚至使B变得不可避免。我们假定在一些像“即使她没有提到她母亲，你也能知道她们吵过嘴”这样的例子中，纳入了模拟以检验因果关系。

一、有关取消的心智研究

我们最初关于模拟启发式的研究的关注点，集中于反事实判断。具体来讲，我们研究了人们判断某事件“差点发生”或者“几乎发生”的过程。这种空间隐喻是非常引人注目的，被许多哲学研究采用：谈论现实与一个曾经可能但未成为现实的世界之间的距离，似乎是合理的。下面这个例子，将用来说明对已发生事件与应发生事件之间的距离进行估计的心理意义。

克兰先生和蒂斯先生预定在同一时间和同一机场搭乘不同的航班。他们乘同一辆轿车出发，却遇上了交通阻塞。结果他们到达时比其航班预定起飞时间晚了半个小时。

克兰先生被告知他的航班按时起飞了。

蒂斯先生则被告知他的航班延迟了，五分钟前刚刚起飞。

谁是最沮丧的？

克兰先生_____ 蒂斯先生_____

在一群由学生构成的被试中，有 96% 的人都认为蒂斯先生是最沮丧的，这似乎一点儿都不让人惊讶。那么是什么使得这种一致如此显著呢？我们会注意到这两个人的客观遭遇是完全相同的，即他们都错过了航班。而且，由于两个人都已经预计到了他们会错过航班，他们之间的差异并不能作为失望的原因。从任何意义上讲，蒂斯和克兰的差别是无关紧要的。造成蒂斯先生更沮丧的唯一原因，在于他更有“可能”搭上航班。我们认为要出现这种情境所导致的标准情绪反应，需要两位旅行者进行一次模拟练习，通过这次模拟他们检测自己离准时搭上航班有多近。这种反事实建构所起的作用，跟预测差不多。虽然这个故事很清楚地告诉了蒂斯先生和克兰先生的预期并没有什么差别，但现在蒂斯先生却更失望，因为相比于克兰先生，要去想象自己如何避免 30 分钟的迟到来讲，他想象自己怎样提前 5 分钟到达会更容易。

这种例子，有一种艾丽丝奇遇记的味道，都包含了一种想象与现实的奇妙结合。如果克兰先生能够想象到独角兽的话——我们相信他能——为什么他会如我们所认为的一样觉得想象自己如何避免一个 30 分钟的迟到会相对困难呢？显然，想象的自由受到了限制，对心智模拟的心理分析，主要包括的就是对这些限制的研究。 204

我们对于心智模拟的规则的理解，仍处于初期，我们只能介绍有关这个似乎相当广泛而且颇有发展前景的领域的一些早期研究结果，以及一些尝试性的推论。我们已经对一些规则进行了初步观察，这些规则主宰着一类特定的模拟活动——取消过去发生的事情。我们对于取消的研究，集中于一种这类模拟活动相当普遍的情境——事故中的死者的亲属对于该致命性事故的反应。与前面蒂斯先生和克兰先生的例子一样，在这里，我们选择被我们称为情境的情绪版本（the emotional script for a situation）来进行研

究。下面的故事就是一个例子：

琼斯先生今年47岁，他是三个孩子的父亲，也是一个成功的银行经理。他的妻子生病在家已经好几个月了。

在事故发生的当天，琼斯先生按往常的时间下班。他有时候会应他妻子的要求早一点下班回家处理家务事，但在那天并不需要。下班后，琼斯先生并没有依照他常规的路线开车回家。那天的天气相当晴朗，琼斯先生在办公室时告诉他的朋友他会沿海滨路回家，以欣赏风景。

这场意外发生在一个大的十字路口，在黄灯亮的时候琼斯先生到达那里。目击者说他急刹车停在了路口处，虽然他完全可以在红灯亮前通过路口。他的家人知道这是他开车时经常发生的事。他在绿灯亮后开始通过路口的时候，一辆轻型卡车全速闯进十字路口，从左面猛撞到琼斯先生的车。琼斯先生当场死亡。

后来得知那辆卡车的司机是一个十几岁的小男孩，他当时服了药，药性发作。

跟类似情境中经常出现的情况一样，琼斯的家人和他的朋友在事故后的很长时间里经常想到和念叨的都是，“要是……”他们这些想法的下文是什么？请补充一个或多个可能的下文。

62名不列颠哥伦比亚大学的学生看到的，是这个版本的故事（称为“路线”版本）。另一组由61名学生组成的被试读到的，是“时间”版本的故事，这个版本是在原版本的基础上对第二段作了如下的改动。

在事故发生的当天，琼斯先生按比往常提前了一点儿时间下班，以应他妻子的要求回家处理一些家务事。下班后，琼斯先生依照常规的路线开车回家。在某个天气相当晴朗的日子，琼斯先生会沿海滨路开车回家，以欣赏风景，但那天一切照旧。

我们分析了被试给出的第一个对“要是”命题的补充,结果如表 14—1 所示。被试给出了四种回答:(i) 取消死者所走的路线;(ii) 取消死者离开办公室的时间;(iii) 琼斯先生在黄灯亮的时间穿过十字路口;(iv) 将服药的小男孩从整个场景中取消。

表 14—1 所显示的结果中让人印象比较深刻的一个方面,是有一类情况没有发生:没有一个被试提到要是琼斯先生早到十字路口两三秒钟,他就会安全地通过。这个发现很具有代表性:人们不会通过武断地更改连续变量的值,来在头脑中取消某事件。显然,被试在完成这个取消任务的时候,并没有取消掉先验概率很低的关键事件得以发生的必要条件,这个过程必然会使他们把注意力集中于两辆车在十字路口相撞的特殊巧合性上。因此那个人无论做什么,都不能很好地跟先验概率联系起来。

表 14—1 被试对“要是”命题所作的第一个补充的分析

| 回答类型 | 时间版本 | 路线版本 |
|------------|------|------|
| (i) 路线 | 8 | 33 |
| (ii) 时间 | 16 | 2 |
| (iii) 十字路口 | 19 | 14 |
| (iv) 男孩 | 18 | 13 |
| (v) 其他 | 1 | 3 |

人们对故事进行的改编,可以区分为三类:下降的,上升的,或者水平的改变。下降的变化,指的是去掉故事中某个令人惊讶的或者未预料到的方面,或者说增加它们之间的一致性。上升的改变指的是,引入某些不可能的事件。而水平的改变指的是,用另一个任意值取代原先的变量值,而前者可能并不比后者好。这个实验处理使得在一个版本中路线的改变是下降的,而在另一个版本中则是上升的,对琼斯先生的行程的时间所作的改动,也对应着上述特征。这个处理显然是成功的:被试更有可能通过保留某个变量的正常值而不是引入一个异常值来取消这个事故。总而言之,在被试的回答中,上升的改变是相对比较少的,而水平的改变则几乎没有。

下降或者上升的概念,借自越野滑雪者们的经验,它力图说明的是一个系统的某些可能状态之间的距离关系的特殊性质。这种关系的最重要的性质,在于它是不对称的。对越野滑雪者来讲,从 A 地点到 B 地点这样一

个短距离的下坡，通常伴随着从 B 到 A 的一个距离长并且颇吃力的上坡。在这个隐喻中，山峰代表异常的状态或者事件，而山谷则属于正常的状态或者事件。因此，我们假定，从一个异常情况到它所违背的正常情况之间的心理距离，要小于从正常情况到同一个异常情况的心理距离。对下降式改变的偏好，也许就是心智模拟所遵从的最主要的规则；它体现了对现实与反事实性幻想之间的转换的基本限制。

206 表 14—1 显示的结果中一个显著的特点是，有一类回答，在被试的所有回答中所占比例，相对较小，此类回答就是通过排除某个很自然地被视为事故原因的事件——服药的小男孩在十字路口的疯狂行为——来取消该事故。这个发现说明了心智模拟的另一个特征，我们称它为聚焦规则（focus rule）。这个规则是指，通常来讲，如果改变故事中我们所关心或留意的某主要目标的一些特征，整个故事就会发生很大变化。在前面这个例子中，我们注意的焦点当然是琼斯先生，因为故事在引导被试去强调他的家人。为了证明聚集规则，我们设计了关于此事故的一个新的故事版本，其中我们将最后一段作了如下改换：

后来知道小型卡车的司机是一个叫做汤姆的十几岁的小男孩。汤姆的父亲在此之前在家发现他服了药，而且药力已经发作。这是常发生的事情，因为汤姆严重依赖药物。他父亲因此跟他吵起来，在争吵过程中汤姆抓起放在起居室桌子上的车钥匙，疯狂地驾车出去了。他本人也在这场事故中受重伤。

我们要求看到这个版本的故事的被试，或者代表琼斯的亲属，或者代表汤姆的亲属，来继续补充“要是”命题。这里，我们仍然分析的是被试所给出的第一个回答。扮演汤姆的家属的大多数被试（68%），通常通过假定没有把那把致死的钥匙放在桌子上将汤姆从整个事故场景中去掉。相反，在代表琼斯的亲属的被试中，只有少数人（28%）在他们的回答中提到汤姆。

我们已经比较详细地描述了这个取消研究，尽管它还处于初期，但也

能说明主宰心智模拟的一些规则的令人惊讶的整齐性，并能证明有关适用于某些场景的反事实性幻想的一些广泛认同的规则的存在。我们相信，那些主宰着在脑海中取消某些事的容易程度的认知规则，对于很多被称为反事实情绪（counterfactual emotions）的研究，是非常有帮助的，这些情绪包括挫败感、后悔、某种愤慨、悲痛以及嫉妒，它们的产生，依赖于将现实与那些可能或应该可以发生的事情进行比较。这些令人反感的情绪状态的普遍特征，在于一个人享乐性的适应水平（one's hedonic adaptation level），要高于当前的现实水平，这就好像将未实现的可能事件按照它们在心智模拟中实现的容易度对应到不同的适应水平中去。

二、对情节的评论

对于不确定状况下的预测和计划来讲，心智模拟的谨慎操纵，显得非常重要，它可以被称作是一种独特的模拟启发式。最明显的例子，就是人们会明确地建构一段情节（scenarios）作为概率估计中的一个步骤。

怎样才是一个好的情节呢？情节这个词本身就说明了，一个好的情节 207 会采用一系列的中间事件连接起初始状态和目标事件，而且中间的变化通常采用下降的趋势，几乎没有明显的上升运动。非正规的观察结果显示，一段情节的合理性，在很大程度上，更多地依赖于其中最不牢固的连接合理性，而不是其所有连接的数量。如果一段情节中连接起始状态和结束状态的路径，并不是立刻变得很明显，这段情节就是相当令人满意的，因此中间阶段的引入，实际上增加了主观上判定的目标事件的概率。

任何情节都必然是概略性的、不完整的。因此有必要找出那些控制着如何选择情节中某些特定事件的规则。我们假设情节中的“连接点”，是一些冗余度很低，但因果显著性很高的事件。一个非冗余的事件，代表了结果预测中的局部极小化，某个重要的替代事件很可能在这一点上出现。某连接点事件的出现，若能改变整个链条中那些最终导致该情节的目标状态发生的其他事件的正常值，它就被称作是因果性显著事件（a causally significant event）。

在一个合理的情节里，对现实的初始状态和某特定的结果之间起连接

作用的那些细节，经常被用于支持该结果发生的概率非常高的判断。另一方面，如果某种结果，只能通过援引一些关于稀有事件或奇怪巧合的上升性假设才能实现的话，我们倾向于认为这种结果是不可能的。因此，对某个情节的好坏的评估，能作为一种判断事件概率的启发式。尤其是对于人们的计划来讲，情节，通常被用于评估该计划成功的概率，以及估计各种导致计划失败的原因的风险。

我们已经提到过，情节的建构，能作为一个启发式去估计事件的概率，这中间主要是通过一个中介性的估计，它评估的是一些因果体系能导致那些事件的品性。跟其他任何的启发式一样，模拟启发式也会导致一些特有的错误和偏差。这个领域的研究非常少，但下面的假设似乎很合理：（i）在情节建构中，对冗余度低和因果显著性高的“连接点”的搜寻，会导致人们对某种情节（和结果状态）的偏爱，在这种情节中，是用一些引人注目的事件来标示出因果转换。这就会相应地造成另一种倾向，即人们会低估那些由缓慢的和累积性的变化所导致的事件的发生可能性。（ii）在概率估计中，使用情节建构，还伴随着另一种偏差，即人们会偏爱那些能建构起一个合理情节的目标事件，而相应地排斥那些能以多种途径实现的事件。如果人们的计划活动存在这样的偏差，将会造成相当有害的后果，因为它会导致人们过度乐观地估计计划成功的可能性。一个计划，从本质上讲，是一条包含着许多合理连接点的链条。在这个链条的任一点上，都有理由期望事情会跟计划的一样圆满实现。然而，即使造成计划失败的每一个原因发生的概率，都可以忽略不计，但至少有一个致命性失误的累积概率是相当高的。计划失败，往往是因为一些令人惊讶的事件，在这些事件中，出现了未预料到的上升性改变。因此，使用偏向于下降性改变的模拟启发式，就必然伴随着犯巨大而系统性错误的危险。

本章源自1979年4月在密歇根大学（Ann Arbor, Michigan）的Katz-Newcomb社会心理学讲座的演讲，论题为“论可能世界的心理学”（On the Psychology of Possible Worlds）。本研究受海军研究办公室给斯坦福大学的资助，合同号为N00014-79-C-0077。

第**五**部分
PART FIVE

共变和控制



第 15 章

非正式的共变评估：基于数据的判断与基于理论的判断的比较

杰宁斯等 (Dennis L. Jennings, Teresa M. Amabile, and Lee Ross)

211

社会经验之流不断地要求我们承认经验共变 (empirical covariation) 的存在。有些时候, 这些共变仅仅是对于我们的观察能力的另一种检验, 对于我们没有什么直接的实用意义。而在其他时候——如当这些共变涉及问题的初期症状与后期表现, 或者运用的行为策略与得到的结果, 或者人、情境的相对明显的特征与相对隐秘的特征时——这种检测能力, 可能有助于决定我们在适应日常社会生活的要求方面的成功。更普遍地说, 共变检测, 在我们作为“直觉的科学家” (见 Nisbett & Ross, 1980; Ross, 1977, 1978) 评估和修正我们持有的关于我们自己、我们的同辈和我们的社会的假设的持续斗争中, 扮演了重要角色。因此, 一个明显的问题, 由此引发: 作为普通人 (laypeople), 我们在评估由经验证据所显示的经验共变时, 我们的水平如何呢?

在开始讨论过去的或者现有的研究之前, 我们应

该注意到，日常观察提供了大量的相关证据，也暗示了关于水平问题的答案可能远远不是那么简单。一方面，社会行为的普遍适应性与社会互动的普遍协调性都使我们相信，我们的文化参与者对于行为的原因和结果有着深邃的洞察力。事实上，当代的归因理论家（如 Jones et al., 1971; Kelly, 1967, 1971, 1973; Weiner, 1974）对这种洞察力的原理表现出独特的兴趣。他们似乎普遍地把这种检测共变的能力视为一种“给定的”能力，并且开始在推理中运用，而这种推理，由可觉知的共变构成。另一方面，日常经验所提供的雄辩的证据表明，还存在毫无根据的偏见、破坏性的行为策略，以及错误的因果或者相关理论的其他表现。因此，应该很明显，任何关于非正式的共变评估的充分描述，都必须解决普通人的表现的不均衡性（unevenness）问题。不管怎么样，它必须调和人们对一些函数关系的精确洞察与对其他关系的忽视或者错误印象。

一、关于共变评估的先期研究

或许，值得注意的是，即使是那些最坚定地为普通人作为直觉的科学家的能力辩护的学者（如 Peterson & Beach, 1967），也几乎没有人不承认他们过高地估计了普通人处理双变量（bivariate）观察资料的能力。我们将评论两种支持这样的客观描述的研究路线，对于每一种情况，我们集中关注其程序和结果，它们会证明与我们自己关于普通人共变评估的研究报告相关。

1. 阅读列联表：“生态效度”问题

先期研究的一条重要路线，是研究人们识别简单的 2×2 列联表（contingency table）中所呈现的函数关系的能力。通常，这种表格汇总了变量 X （如某种疾病）出现和不出现的个案的数目，以及据称与之相关的变量 Y （如某种症状）出现和不存在的状况。两个变量的数据，被如此巧妙地收集和“包装”起来，因此表面上看来被试的任务是再简单、直接不过了。然而，证据（如 Jenkins & Ward, 1965; Smedslund, 1963; Ward & Jenkins, 1965）表明，普通人在这种任务上通常表现得非常糟糕。

忽视四个单元中的一个或者更多单元的判断策略，被证明是大多数被

试所面临的困境的根源。如一种常见的缺点，与被试几乎完全依赖于“出现—出现”单元的样本规模在全部总体的比例直接相关。因此，许多被试，可能仅仅因为列联表报告了较大数目的人确实同时拥有某疾病及其症状，就认为症状 X 与疾病 A 相关。类似地，那些仅仅考虑四个单元中的两个单元的被试，可能也被误导了。如果注意到没有症状 A 的病人比有症状 A 的病人要多，他们可能推论这种相关关系是一种负相关；或者注意到有这种症状的人中患有该种疾病的比不患病的更多，他们可能推论说这种关系是一种正相关。

但是，由于缺乏正规的统计学训练，几乎没有人意识到，在这种情况下，只有在考虑了所有四个单元之后，才能够做出有效的推论。如一种令人满意的方法，可能需要比较比例（即比较表现出某种症状的病人与表现出该种症状而未患病的人的比例）。当然，当前的统计推论教材所指定的正规策略，同样需要考虑所有四个单元。特别地，这种策略，要求人们注意观测到的单元频次与“期望”频次之间的差异，而后者则基于相关的边缘频次或者概率（因而提供了计算卡方统计量所需要的信息）。

为普通人辩护的怀疑论者，可能试图拒绝这种论证作为人们拙于“阅读”列联表的证据，他们会认为，被试在这种任务中出错，主要是因为任务的新奇性和人工性。实验室表现与日常现实世界表现之间的关系——也就是“生态效度”问题——是一个重要问题，也是本章所描述的研究中不能够规避的一个问题。相应地，在此让我们开始处理这个问题，希望读者会预先得得到警告并且做好准备。

首先，尽管不可否认，表格阅读任务不是一种常见任务，并且从生态学意义上来说并不具有代表性，但是我们要注意，被试在这些实验室任务中所表现出的逻辑（和各种逻辑弱点），与各种日常推理中所表现出的逻辑相当一致。如考虑一个复杂的问题，“上帝会回答祈祷者吗？”“是的”。那些仅仅考虑出现—出现单元的普通人可能会说，“因为有很多次我向上帝要求一些东西，结果他给了我那些东西。”他的那些更加老练多疑的朋友可能会问，“但是又有多少次你向上帝要求一些东西而没有得到呢？”但是，对于目前这个推理任务，两个单元的比较，仍然绝对不够。即使加上第三

个单元——没有祈祷而出现满意的结果——也不会解决问题，因为在做出任何可靠的推论之前，都必须考虑所有四个单元。并且，即使是最老练的直觉心理学家，也可能会固执地拒绝这种建议，即“不存在—不存在”单元的数据（如没有祈求要得到满意的结果，并且确实也没有出现满意的结果），对于评估祈祷对尘世结果的影响，必不可少。

但是，我们甚至可以更强有力地回击“不具代表性”的这种指控。如果说四项表格（fourfold table）的任务，是一些独特的问题，那么，这同一种任务，也是被试在日常生活中被要求的，而它们正是日常共变评估的困难所在。这些要求，包括初始抽样、编码、存储和提取相关数据，然后以符合共变评估的形式整理数据。如考虑有一个人，他试图检验一种陈旧的观点，即红头发的人脾气暴躁（或者更准确地说，红头发与脾气暴躁之间存在正相关关系）。首先，他必须决定应该考虑哪些数据。他应该试图抽样或者回忆红头发且脾气暴躁的人的实例吗？或者他应该考虑一些有红头发的人，然后注意他们中有多少人是脾气暴躁的，或者考虑一些脾气暴躁的人，然后注意他们中有多少人是红头发的吗？他很可能会采取一些这样的策略或者策略组合，而从来没有考虑过性情平和且头发深色的人的相对频次可能与这个推理任务有关的可能性。这种情况，与四项表格任务形成对比，后者使得被试能够平等地利用所有相关数据，即使它并不要求所有数据都应该被给予适当顺序的考虑。

让我们假设不知什么原因，正被讨论的这个人确实承认，为了检验他的成见，他需要考虑所有四个单元的数据。他应该怎样选取出一个合适的样本从而得到相关的单元频次呢？他应该仅仅考虑碰巧首先浮现在脑海中的那组人吗？他应该考虑属于某个有限类别（如他的家庭成员、他的朋友，或者他所在公寓的房客）的所有人吗？或者他应该尝试采用某种“随机”方法从全部总体中得到一些例证吗？大多数非正规抽样方法的潜在偏差来源都很明显。根据首先浮现在脑海中的样本抽样，容易出现严重的“便利性”偏差（Tversky & Kahneman, 1973, 11）。朋友、家庭成员或者邻居样本，可能出现同样的偏差，也违反了独立性的要求（并不是一个微不足道的问题，因为一个人的家庭、他的朋友甚至他的邻居往往可能具有共

214

同的行为倾向、身体特征，甚至是他们之间的一致)。当然，“新奇的”四项表格程序，使得共变评估者免于面对这些复杂的和潜在的临界问题。

215 编码、存储以及提取的工作，给日常共变评估带来许多额外的因素和问题。最重要的是，红头发的人脾气暴躁这一隐含的假设，可能左右了关于谁脾气暴躁或者不暴躁以及谁的头发红或者不红的判断。本章随后会进一步考虑这个问题。目前，我们仅仅需要注意到，人们先前的预期或者假设，可能会持续潜在地影响每一阶段的信息处理，以致产生偏差。而当人们面对的是已经完成了全部处理的数据即包括编码、存储和提取时，这种影响就被排除了。更一般地，让我们仅仅重申，容易证明，与源自于列联表方法论的人工性或者新奇性的困难相比，现实世界中的这种困难，对于精确地评估共变来说，是更加难以克服的障碍。

2. 伪相关：先见对于感知的影响

当相关的观察资料是一个一个单独地而不是以列联表的形式给出时，人们处理二元分布的成功情况又是怎样的呢？现存的证据大部分来自于查普曼（Chapman）等人关于伪相关的开创性研究。对这一研究的详细说明见查普曼等人的研究（Chapman & Chapman, 1967, 1969），也参见本书第17章。在此，我们的简短评论，集中关注那些为我们自己的最新努力做出铺垫的特定问题和研究发现。

查普曼等人研究被试检测临床迹象和症状之间的共变的能力，或许对于他的研究发现和结论的最简单概括就是，被报告的似乎反映的真实共变，共变所远远少于基于理论或者基于语义的先入之见，而这些先入之见，是关于“应该”存在的关系的性质。尽管这些研究发现激动人心，充满争议，并且很大程度上与实践者直接相关，但通常认为先入之见对于观测到的联系的影响，从根本上来说很普通。

正如L. 查普曼自己在他的第一篇论文中（1967）所指出的，伪相关并不局限于临床判断领域。大多数的迷信，本质上是关于某种行为或者事件与随后的积极或者消极后果之间的关系的信念，而这些信念缺乏经验的证据。类似地，种族的、道德的、地区的或者职业的成见，也是关于共变的信念，是被强烈坚持的、尤其抗拒非支持性数据的影响力的信念（比较

Adorno et al., 1950; Allport, 1954; P. Goldberg, 1968; J. M. Jones, 1972; Taynor & Deaux, 1973)。这种情况，也可以引用关于“内隐人格理论”(implicit personality theory)的研究。大约半个世纪之前，桑代克(Thorndike)描述了所谓的成见效应(halo effect)，由此感知者期望并且报告了所有积极的人格特征之间的联系。随后的研究者(如 Koltuv, 1962; Norman & Goldberg, 1966; Passini & Norman, 1966)极大地拓展了这项研究，探索关于人格的概念图式和信念，它们使得感知者倾向于报告不同特质之间或者同一特质的不同指标之间的一致关系，而它们之间的经验联系非常弱，甚至根本不存在。

因而，在共变评估领域，有大量的经验和逸事证据证明，预期与证据之间的角逐有可能是一场不平等的竞争(比较 Bruner, 1957a, 1957b; Bruner et al., 1951)。直觉的科学家关于经验关系的先入之见，往往决定了他发现什么，他不能发现什么以及他看到的而实际上并不会被看到的东216西。然而，关于理论优于数据的这种肤浅的归纳，可能使我们忽略了查普曼研究中一个重要的含糊之处——或者，至少是一个未解决的重要问题。特别地，我们能够确认与直觉共变评估有关的两个相互独立的问题。某种意义上，在那些研究以及其后的后续研究中(如 Golding & Rorer, 1972; Starr & Katkin, 1969)，这两个问题被混淆了。第一个问题，涉及人们在检测即时可得的数据显示出的共变时所面临的困难——也就是说，当他们不受关于正被讨论的经验关系的任何特定理论或者预期“妨碍”时，他们识别并且评估共变的能力。第二个问题，涉及被试的直觉或者理论引导他们假定存在强的相关关系的倾向，而这种假定，被最佳的“客观”证据所隐藏。

这两个问题之间的分离，意味着需要采取两种非常不同的任务，来研究被试的共变评估表现。第一种任务，本质上是一个“心理物理学”任务，它产生关于主观判断与客观测量之间的相关关系的证据。在这个任务中，只向被试出示了几组二元观察，而他们不能事先形成任何关于这些观察的先入理论，然后要求被试评估每一组中的关系的强度。然后研究者就能够尝试衡量被试的任务的难度(主要是通过记录与某些判断相关的变异

程度)，判断那些联系主观估计与客观测量的“心理测量”函数的性质。

第二种任务，处理的是另一极的情况，与这种纯粹“基于数据”的评估对立；它关注的是完全，或者几乎完全“基于理论”的评估。因而，被试的直觉或者理论能够联系的成对变量或者测量，将会被指定，但是研究者不会提供任何相关的二元数据。因此，如果被试在这种评估中运用了任何数据，那么这些数据一定是由被试的日常经验和记忆提供的；这些数据，也同样容易受到我们前面的讨论中提到过的一大堆“加工”错误和偏差的影响。为了进行评估和比较，实验者也必须掌握一些方法，来评估与这种直觉估计相关的精确性，或者错误的性质。也就是说，他必须掌握一些相关关系的“客观”尺度，以测量指定给被试的每一对变量。

二、基于数据与基于理论的共变评估的实验比较

最近，我们的一项研究讨论了这个问题，并且运用了刚刚概述过的基本策略。在基于数据的任务中，运用了三种不同类型的二元分布，在下一部分我们会更具体地描述。至于基于理论的评估，实验者指定了成对的相关变量，但是没有给出任何实际的数据。为了方便比较基于数据与基于理论的评估，我们使用了一种被试内设计（within-subject design），在两种任务中，被试运用同一种主观等级量表（根据出示的顺序保持均衡）。

在每一种情况下，相关的共变评估任务使用的，都是连续变量而非离散变量（与 Chapman & Chapman, 1967, 1969 和 Jenkins & Ward, 1965 的研究形成对比）。之所以这样设计，是为了更加准确地把握大多数日常评估任务的性质。^[1]

被试与评估任务

一共有 64 名斯坦福大学的本科生，参加这项比较基于数据与基于理论的共变评估而设计的实验，他们中还没有人已经完成大学阶段的统计学课程。在两种类型的评估中，被试都使用一种简单的等级量表，来描述他们关于每一对变量之间的关系的方向和强度的主观印象。在使用这种量表时，被试首先估计问题中的关系是正还是负。然后，他们估计这一关系的强度，

并在一个百分制的等级量表上画“X”表示这种强度，量表两端的初始值分别是“完全相关”和“无关”。

(1) 基于数据的共变估计

实验的一个部分处理的共变估计，是基于实验者提供的二元观察资料集。运用了三组刺激物来产生相关的二元分布。对每一种类型的刺激物，设定一组二元分布，以代表两个变量之间不同的客观共变。被试被要求研究每一种分布，并且使用百分制主观等级量表估计分布所描述的关系。

我们预计，在这个任务中，被试估计的精确性，部分地依赖于他们在编码、处理和回忆数据时所面临的困难。为了检验这个假设，在被试进行共变估计之前，三种类型的刺激物对被试提出了明显不同的信息处理要求，特别是与估计刺激强度和存储以及从记忆中提取信息相关的要求。因而，219第一种类型给出的，是每组 10 个简单的数字对。第二种类型，给出的是图表，每张图表描绘了一组 10 个不同高度的人手持不同长度的手杖。最后一种类型的分布，是用录音磁带给出的，内容是一组 10 个人，每个人首先说出一些字母，这些字母在字母表上占据一定的顺序（以他们的姓氏的首字母为序），然后立即唱出不同长度的一段乐谱。

(2) 基于理论的共变估计

实验的另一部分（两个部分出示的顺序保持一致）处理的共变估计，是基于被试的先验预期或者理论，而非任何即时可得的二元数据。

指定成对的测量或者变量，然后要求被试首先估计每一对关系的方向，然后估计关系的强度，并且始终运用与基于数据的评估中所运用的相同的简单百分制主观等级量表。在这部分的实验中，一些变量对，处理的是个人特质的不同行为测量尺度，如哈桑等人（Hartshorne & May's, 1928）关于道德行为的跨情境一致性的经典研究中所使用的测量诚实的两种尺度，其他一些处理的是个人态度、习惯或者偏好。但是，在每一种情况下，我们获得的主观共变估计，都应该能够与从以往的经验研究中挑选出的“客观”相关关系比较（表 15—1 给出了这些变量对以及相应的相关关系的一个完整列表）。

218 表 15—1 指定给评估者的变量对和它们的客观相关系数

| 变量描述 | | 客观相关系数 |
|------|--|--------------------|
| A | 学生自我评估自由主义
学生评估过去十年中商界领袖的表现 | -0.28 ^a |
| B | 学生自我评估理智主义
学生评估过去十年中美国总统的表现 | -0.19 ^a |
| C | 学生自我评估羞怯程度
学生旅游美国州的数目 | -0.12 ^a |
| D | 学生自我评估抱负
学生的身高 | 0.01 ^a |
| E | 学生评估过去十年中大学校长的表现
学生每星期花在娱乐运动上的时间 | 0.08 ^a |
| F | 学生自我评估理智主义
学生的家庭收入 | 0.17 ^a |
| G | 根据谎报运动成绩测量的儿童不诚实程度
根据解谜中作弊的数量测量的儿童不诚实程度 | 0.18 ^b |
| H | 学生的家庭收入
学生评估过去十年中商界领袖的表现 | 0.28 ^a |
| I | 六年级学生“延迟满足”的能力
六年级学生抵御作弊诱惑的能力 | 0.31 ^c |
| J | 学生自我评估责任心
由室友评估这些学生的责任心 | 0.35 ^d |
| K | 学生自我评估理智主义
学生自我评估抱负 | 0.37 ^a |
| L | 学生评估过去十年中国会的表现
学生评估过去十年中劳工领袖的表现 | 0.40 ^a |
| M | 学生每星期花在娱乐运动上的时间
学生自我评估身体的健康 | 0.52 ^a |
| N | 学生评估过去十年中美国总统的表现
学生评估过去十年中商界领袖的表现 | 0.55 ^a |
| O | 自我评估政治保守主义
自我评估种族中心主义量表项目的组合 | 0.57 ^e |
| P | 学生的身高
学生的体重 | 0.79 ^a |

- a. 相关系数得自对斯坦福大学学生的一项调查 (N = 295)。
 b. 相关系数得自 Hartshorne 和 May 的研究 (1928)。
 c. 相关系数得自 Mischel 和 Gilligan 的研究 (1964)。
 d. 相关系数得自 Bem 和 Allen 的研究 (1974)。
 e. 相关系数得自 Adorno 等的研究 (1950)。

三、结果

1. 基于数据的共变估计

我们首先关注的事情，涉及被试针对即时可得的二元数据做出的共变估计。我们可以提出两个明确的问题：首先，在没有任何关于相关数据分布的“理论”的情况下，被试怎么能够容易地检测出各种强度的共变？其次，联系评估者的主观估计的均值与客观的共变测量的“心理物理学函数”的性质是什么？

对数据的考查表明，这种估计任务非常困难。每一组数据的标准偏差以及与全组估计相关的四分位差间距（interquartile range）都非常高（见图 15—1）。特别值得注意的是，许多被试，在仅仅确认正相关关系的存在时，就面临困难，即使是强度相当大的相关关系。事实上，只有当客观的相关关系达到 +0.6 到 +0.7 的水平时，四分位差区间，才彻底排除负的相关关系估计。三种类型的二元数据分布，在共变估计的不稳定性方面，似乎并不存在系统的差异，尽管与这三种类型相关的信息处理（随机或者系统误差的概率）在数量上存在明显差异。被试明显对于信息处理要求的增长不敏感，这一点，为了解被试的评估“策略”的本质，可能提供了一些线索；而被试的评估“策略”，就是一种概率，稍后我们在文中再来讨论。 220

图 15—1 也给出了与基于数据的评估的第二个问题相关的数据，即共变的客观与主观指标之间的“函数关系”问题。读者将会注意到，三种刺激域的均值，在结果描述中是共同给出的。检验以及显著性检验，都没有显示出任何有效的或者连贯一致的趋势，能够区分出三种刺激域的结果。与三种独立领域中的任何一种得到的结果相比，这种共同分担的结果，是一个更加平滑、更容易理解的函数。为了方便读者，我们拟合图形中的各点，得到一条曲线。应该强调，这条曲线是在这之后作出的，它不是来源于任何“心理物理学理论”。 221

通过检验，很明显，尽管对于个体估计存在组内变异性（within-group variability），但是联系主观估计的均值与客观的相关关系（Pearson r ）的函数却相当平滑。更加特别地，似乎存在一个联系两个变量的剧增函数（sharply accelerating function）。因此，心理学家通常关注的关系，即所测量

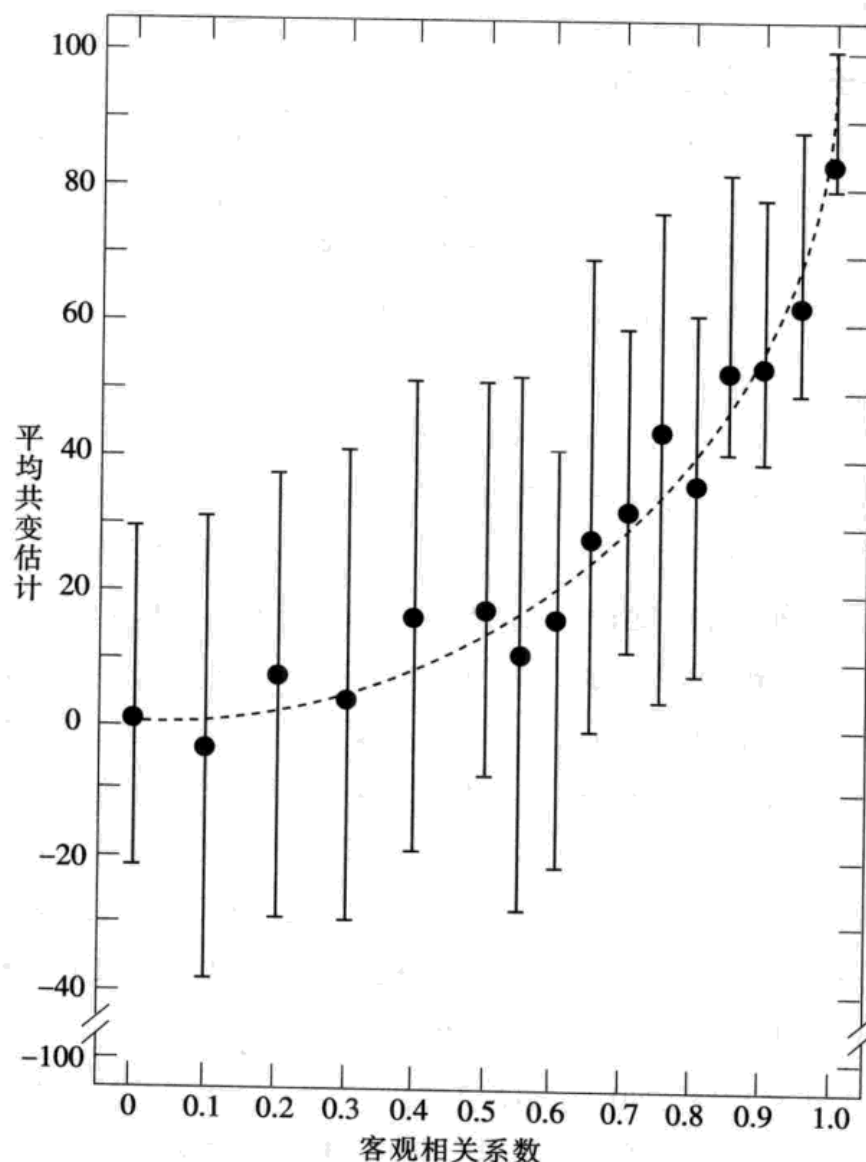


图 15—1 “基于数据的”共变估计的均值和四分位差间距，它是三种分布类型共同分担的结果，“心理物理学函数” $100(1 - \sqrt{1 - r^2})$ 以虚线表示。

的人格特质或者其他社会行为的跨情境一致性如 ($r=0.2$ 到 0.4)，几乎检测不到。得到的平均估计根据百分制量表在 4 到 8 之间。即使是这些心理学家认为非常强的关系（如 $r=0.6$ 到 0.8 ），得到的结果也是不太高的主观共变估计。如客观相关关系 0.7 ，得到的主观估计均值为 34——就百分制主观量表而言，是一个位于“相当弱”与“温和”之间的中间等级。只有当客观相关关系接近 0.85 的水平时，组平均值才达到主观量

表的中间值，并且只有超过这一点，被试才一致地把这种关系评定为正的强关系。

对估计更加仔细的考查表明，相关性或者共变的主观等级并不是 r 甚至 r^2 的一个线性函数。相反，表达式 $1 - \sqrt{1 - r^2}$ 似乎准确地描述了共同给出的估计均值的模式。[读者可能会把这个表达式看作是“离间系数”，一种测量标准误差的减少量的尺度，标准误差的产生与根据对变量 X 以及 X 与 Y 之间的关系的了解预测变量 Y 有关 (Huntsberger, 1967)。] 事实上，实际的估计均值与这个表达式描述的均值之间的相似性，相当令人惊讶——准确地说，基于离间系数的函数解释了 95% 的均值间变差。但是，我们要赶紧提醒读者不应该误解这种巧妙拟合的意义。图 15—1 所描绘的拟合优度 (the goodness of fit)，仅仅适用于组估计；而个体估计，则表明被试面临区分不同程度的共变的困难。

显然，关于联系主观回答与客观共变测量的“心理物理学”函数的精确性质，现在做出任何结论都还为时过早。毫无疑问，任务的细节、情境甚至可能是评估者的总人数，都会导致函数的不同，而只有通过引入合适的参数，才能捕捉到这种差异。但是，现在的结果，要证明人格评估小组 222 成员频繁报告的那种类型的适度相关关系（比较 Mischel, 1968, 1969），可能不够，但足以给那些由于缺乏基于理论的先入之见而遭遇这种共变的普通人留下最微弱的印象。

值得进一步再次强调的是，尽管它们对被试提出了迥然不同的信息处理要求，三种不同的基于数据的共变估计任务得到的结果非常类似。因而，正如图 15—1 所证明的，被试参与的三种任务中的每一种，心理物理学函数都精确地反映了主观估计与客观测量之间的关系。被试的实际估计与函数预测的估计的均值之间的相关关系从乐谱/字母对的 $r = 0.91$ 到数字对的 $r = 0.98$ 不等。

2. 基于理论的估计

至此，我们已经给出证据证明，为了得到即使是稍微不等于零的主观估计，就需要相对强的客观相关关系（根据皮尔逊相关系数 r ）。孤立地考虑，基于数据的评估的心理物理学函数只能证明，普通人的主观度量标准，

相对于大多数正规统计学家所运用的来说，是一种谨慎的或者保守的标准。（当然，几乎没有人会怀疑， r^2 与 r 相比是更加合适的相关性指标，因为它更容易与预测效用相联系，或者是某个变量的值的不确定性的减少量，根据对其他相关的变量的了解。）简单地说，尽管与这种估计相连的变异性使我们几乎不再怀疑估计任务的难度和不确定性，但是我们获得的基于数据的估计，并不能证明直觉心理学家有任何系统的弱点。

记住这些研究发现以及可能的解释，现在我们可以转向关于表 15—1 中列举出的关系的主观共变估计的结果。我们应该记得，这些估计，是在不存在即时可得的数据的情况下做出的。根据推测，它们是仅仅根据评估者的非正式理论或者直觉做出的，没有任何数据，除了在日常经验中记录下并回忆起来的高度“处理过的”数据。图 15—2 所提供的，是这些结果的一幅直观图。很明显，任何一个函数——线性的或者其他的——都不能描绘出主观估计与客观测量之间的关系。然而，不可否认，有一种趋势使

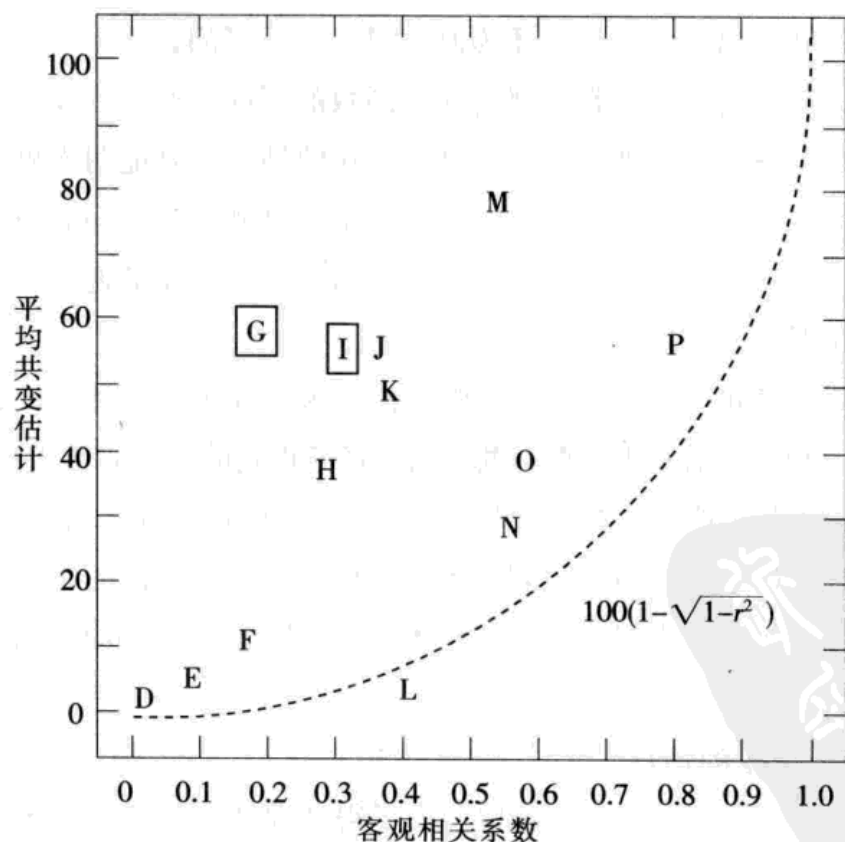


图 15—2 “基于理论的”共变估计的均值（具体变量凡例见表 15—1）

得被试基于理论的估计与客观证据之间表现出大致的对应。也就是说，正的经验关系被估计为正；负的关系被估计为负；与相对弱的关系相比，相对强的经验关系，通常会被估计的强一些，等等。 223

但是，或许这些基于理论的估计最惊人的特征在于，一旦解除了即时可得的数据强加的限制，被试就变得不再谨慎或者保守。他们心甘情愿地冒险进入主观百分制量表的中部甚至上部，即使是在处理那些客观上仅仅适度相关的变量对时。图 15—2 提供的图形证明，使我们理解了这一点。同一相关水平的几个变量对，在基于数据的评估的情况下得到的主观估计均值，几乎等于零（如 $r = 0.3$ ），而现在得到的估计，根据主观百分制量表，是 30、40 或者更高。

变量对的 G 和 I ，值得特别强调，因为它们处理的基于理论的估计是关于行为的跨情境一致性，因此证明了一个目前受到特别关注的理论问题（比较 Alker, 1972; Allport, 1966; Bem & Allen, 1974; Bem & Funder, 1978; Bem & Lord, 1979; Mischel, 1968, 1969）。关于这两种关系的主观估计所传达的信息，很明显，也十分惊人。当面对即时可得的客观数据时，这些数据给出的相关关系从 $r = 0.2$ 到 $r = 0.3$ 不等，被试给出的平均估计根据百分制等级量表平均在 0 到 10 之间。现在，在不存在可以直接利用的数据的情况下，处理相关关系同样是从 $r = 0.2$ 到 $r = 0.3$ 的双变量域时，被试固有的人格理论，使得他们给出的共变估计平均在 50 到 60 之间。在没有“理论”或者先入之见的情况下，为了得到类似的估计，将会需要即时可得的二元数据给出什么程度的共变？如同我们在图 15—2 中所能够看到的，答案是只有在 $r = 0.90$ 范围内的经验相关关系，才可能得到这样的估计。 224

因而，我们可以用一对经验概括来总结我们自己的研究结果。

(1) 在没有任何关于它们的相关性的理论或者先入之见的情况下来检查即时可得的二元数据，直觉心理学家面临检测出各种强度的共变有重大困难，这些共变可能构成日常社会经验提供的范围广阔的函数关系的特征。特别地，对那些探求行为的跨情境一致性的心理测量学家来说，他们所研究的强度范围内的共变，可能一直没有被发现，或者被认为几乎不存在。

(2) 当不存在任何客观的、即时可得的二元数据能够检查，但是可以运用先前的理论或者先入之见时，直觉的心理学家，倾向于期望和预测强度相当大的共变，它经常是强度远远大于以往的经验可能提供过的，或者未来的经验将会提供的。

综合来看，这些概括有助于突出目前基于数据的与基于理论的评估之间的对比。显然，我们持有的理论，使得我们期望并且预测比实际存在的关系更强的经验关系；除非我们已经期望去发现它们，否则，许多确实存在的这种经验关系，甚至是因果强度的关系，就可能一直埋没不被发现。

四、论理论的起源与生存

1. 共变评估策略

在概括各种评估任务的结果以外，开始思考可能构成这些结果的基础的判断过程和策略，这一点也很重要。再一次说明在基于数据的估计与基于理论的估计之间的区分具有决定性意义；而查普曼等人的开创性研究，也再次提供了一个有利的切入点。

(1) 数据与理论的比较

225 根据查普曼等人的研究，很明显，先入之见，胜过即时可得的数据提供的信息。但是，这种情况发生的方式并不那么明显。被试事实上“看到了”他们所报告的关系吗？或者他们仅仅报告了他们所期望的数据中会存在的关系，而没有任何相应的主观印象？或者被试折中他们的主观感知和预期，赋予各自一定的权重？根据报告的结果给出一个明确的答案，是不可能的。尽管有明显的证据证明，被试至少确实重视他们对数据的直接感知，并且他们的感知也不是完全为他们的预期所决定。特别地，我们注意到，查普曼等人的报告反复揭露不存在的相关关系，最终确实减少了伪相关的报告。但是，也很明显，数据从来没有能够彻底击败被试的先入之见，因为即使扩展到负相关关系，也不能完全排除伪正相关关系的报告。

因而，关于共变评估的直觉策略的问题，就尖锐地突显出来。被试怎样判断手头拥有的资料中存在的关系的程度呢？他们的直觉理论或者先入之见提出具有代表性的一组二元观察，他们又是怎样判断其中“应该”存

在什么程度的关系的呢？先入之见什么时候以及怎样转化成人们关于共变的主观经验呢？在开始思考这些问题的答案时，一开始就集中关注纯粹基于数据的与纯粹（或者至少是大部分）基于理论的评估，并且只有在那之后才转回到即时可得的数据与基于理论的预期相互冲突的情况，这种做法再次帮助了我们。

(2) 基于数据的评估的策略

关于基于数据的评估，可能最令我们吃惊的研究发现是涉及被试针对那些给出日常经验中可能会碰到的各种强度的共变的二元分布所做的回答，这些共变与非正式的社会理论有关——特别地，被试发现检测这种相关关系非常困难，并且平均而言，他们认为相关的共变非常接近于零。考虑到我们的被试可能运用的非正式共变评估策略，这些结果意味着什么呢？

在思考这个问题的时候，重要的是记住，个体评估者对于客观上弱的关系的估计通常不会接近于零。相反，他们给出一个相当大的估计范围，只有相关的这些估计的均值才接近于零点。再一次，很难准确地辨别出被试事实上“看到了”什么以及他们估计的是什麼，不管他们看到了什么。联系平均估计值与实际的相关关系的“心理物理学”函数令人吃惊的平滑和连贯，使得我们相信，被试确实利用了与客观的相关关系密切相关的数据特征。然而，他们的估计的变异总量表明，被利用的数据特征与客观的共变测量，仅仅是非常粗略的相关。 226

我们猜测，被试关于“相关性”的印象，并没有导致任何试图考虑有关的二元观察资料的全部样本的努力，一些参与任务的素朴评估者的内省，支持了我们的猜测。相反，被试可能依赖很少的特殊个案，可能是变量 X 或者变量 Y 的极端个案。因而他们可能仅仅观察与 X 的极值相关的 Y 的值的方向和 Y 的极值，反之亦然。影响注意力和记忆的因素，可能起一定的作用；因而描述的形象性和具体性，以及首因效应（primacy）和近因效应（recency），可能影响到被试会依赖哪些数据点。

照这样依赖于有限数目的特殊个案或者“检验”个案，特别是依赖极端数值，将会产生一定的结果，这些结果似乎与我们自己的研究发现高度一致：首先，设计刺激域（如数字对图画、对录音磁带）的操作，是为了

调节共变评估之前所必需的估计和记忆的强度，刺激域的操作本应该几乎不发挥作用，因为如果评估者仅仅依赖非常有限数目的特殊个案，那么这些额外的要求就无足轻重。其次，由于依赖于他们碰巧注意到的那些特殊个案，我们可以预期，处理弱相关关系的被试给出的估计值会是稍大于或者稍小于零，而不会等于零。与此相对，处理强相关关系的被试，应该给出相当一致并且相当极端的估计，因为高的正相关关系，使得与其他变量的极端值相关的变量的值的变动几乎不可能。我们的心理测量函数的形状，以及给出的最强经验相关关系的估计减少的变异性，都与假设依赖极端个案得到的这些结果一致，尽管这些证据远非结论性的。很明显，只有根据仔细操作极端个案或者其他比例不等的便利个案的数量的研究设计，才能得到更加明确的答案。

(3) 基于理论的评估的策略

当我们把注意力转向基于理论的评估时，将要提出的一个明显问题，涉及任何数据可能被利用的程度，除了纯粹的理论之外。当然，在一些个案中，可能根本没有参考过任何数据。如评估者可能仅仅依赖基于语义或者基于理论的直觉和预期；也就是说，在考虑每一个变量对中的两个变量之间的相关关系时，运用“代表性”标准（比较 Kahneman & Tversky, 1973, 4）。可能没有读者会怀疑这种论点，即人们持有——或者至少能够容易地产生——这样起源的社会理论。实际上，任何普通人都能够如此迅速地产生新的理论或者关于函数关系的预测，使得这些理论不能依赖于实际个案的生产和分析。

然而，我们并不怀疑一些基于理论的共变估计，可能会需要参考实际证据的可能性。但是，我们再次猜测，普通人更加可能依赖特别的检验个案——他们容易受到编码、存储和回忆过程中多种来源的显著偏差影响——而不是产生并且非正规地分析一些二元数据的总体样本。

因而一些关于特殊相关关系的理论或者信念，可能至少部分地基于数据。而其他的理论或者信念，可能基于对更为广泛的信念的演绎，而这些更为广泛的信念涉及世界、传统的民间智慧、语义联想，或者家庭、朋友以及大众媒介的劝说性交流。但是，很明显，无论它们起源于什么，这些

信念有许多是错误的，并且在面对那些会使任何不受偏见影响的观察者——当然是任何拥有一个计算器、一本统计学导论课本和一些关于如何使用它们的传统知识的不受偏见影响的观察者——在产生重大疑问的证据的情况下，能够存活下来，甚至盛行一时。因此，我们愿意简单地论述一对相互关联的问题，以此结束本章。这对问题，是关于直觉理论与经验共变的日常经验之间的不一致，而经验共变“检验”了这些理论：首先，面对有明显说服力的逻辑或者证据方面的质疑，使得那些关于函数关系的错误的或者过分夸大了的信念能够存活下来的机制是什么？其次，怎样调和关于普通人的弱点的新兴观点与生物体表现出的明显的学习能力（特别是操作性条件作用和反应性条件作用），以及那么多日常信念和社会策略明显的精确性和恰当性？

在此我们只能简单地涉及这些问题，我们建议感兴趣的读者参考本书第 9 章以及其他书中更为详尽的讨论（如 Nisbett & Ross, 1980）。

2. 理论持续的机制

关于经验关系如对特定个体的印象的理论，经常表现出一种面对经验质疑时的惊人存活能力。一组解释这种持续性的机制，源于这样一个简单的事实，即人之特征是根据他们的信念行动；这些行为，可能包括从简单的公开支持到人们在时间、精力、财富和声望方面的投资。这种行为承诺，使得行动者在面对新的证据时仅仅修正他的信念，也是极“不协调的”（Abelson et al., 1968; Festinger, 1957, 1964）。实际上，一个行动者的信念或者预期，可以构成一种自我实现的预言，由此最初毫无根据的信念，会导致行动者的行动，并使得产生的结果或者数据最终合法化的信念（Merton, 1948; Rosenhan, 1973; Rosenthal & Jacobsen, 1968; Snyder & Swann, 1976; Snyder et al., 1977）。

罗斯和他的同事最近进行的研究，阐明了可能构成信念持续性现象基础的其他机制或者程序。这些机制包括，外行人倾向于同化被考虑的新信息，使之与先前信念保持一致，并且超越所同化的数据以发展因果解释，而这种解释能够说明为什么所假设的事态应该维持（见本书第 9 章和 Anderson et al., 1980; Jennings et al., 1980; Lord et al., 1979; Nisbett &

Ross, 1980; Ross, 1977; Ross et al., 1975; Ross et al., 1977)。

3. 共变检测、条件作用和其他日常生活中的成就

乍一看，我们关于普通人检测和评估共变的能力的黯淡观点，似乎与众多实验证据，甚至是基础更为广泛的日常经验相冲突，而这些证据和经验，已证明了经典的和操作性的条件作用。每一只曾经学习过逃离梭形盒的老鼠，每一个曾经觉察出在他的早餐麦片中添加糖的好处的孩子，每一位发现诚实并不总是最好的策略的恋人，每一位曾经对孩子的哭声表现出泌乳反射的妈妈，都证明生物体确实在环境的刺激中识别出共变。因此，查普曼等人 and 目前的其他研究者的研究，所处理的是一些非常有限的、异常类型的、普通人唯一不擅长的共变评估问题吗？尼斯比特和罗斯（1980）讨论了这个问题，他们认为，答案是否定的。他们声称，反而是条件作用现象构成了“例外”，正是本章始终描述的检测和评估共变的有限能力，体现了“规则”。

我们不能具体地评论相关的论点，但是尼斯比特和罗斯的论文的要旨是，经典的和操作性的条件性作用，通常是在严格限制的，包含刺激突出性和特殊性、最优的刺激间距与试验间距，以及缺乏非相关或者分散性刺激事件的环境中，获得的。但是，他们列举的两个额外因素值得特别强调。首先值得注意的是，当条件作用能够在条件刺激（conditioned stimulus, CS）与非条件刺激（unconditioned stimulus, UCS），或者反应与强化之间存在相对低的相关关系下维持时，那么它最初几乎总是在完全共变条件下获得的——也就是说，条件刺激或者反应后总是伴随着非条件刺激或者强化，并且后者绝不会在前者不存在的条件下出现。因而，获得一种条件反应的必要条件与这种反应的维持或者“持续”之间存在的不对称，在一定程度上，也说明了本章探讨的基于数据与基于理论的共变评估之间的同一种不对称。也就是说，在维持期间，生物体可能继续期望，甚至可能感知到条件刺激与非条件刺激或者反应与强化之间有一种比被客观证明的更加强的共变。它可能信以为真地接受强化的试验，而把非强化的试验，归咎于环境的减弱、第三变量的影响，甚至是运气。

最后，或许也是最重要的是，不断增加的证据证明，操作性的和经典

第 15 章 非正式的共变评估：基于数据的判断与基于理论的判断的比较

的条件作用，都受益于在一些情况下甚至是必须的，要习得的偶然性与机体带到实验室的先前理论或者预期之间所存在的相当高的拟合。很明显，绝不是所有的条件刺激—非条件刺激或者反应—强化偶然性，都是同样可习得的。关于这一点，令人印象最为深刻的证明，可能是由加西亚和他的合作者给出的（如 Garcia et al., 1972）。这些研究者报告说，在一个单独试验中，老鼠能够学会避开伴随有肠胃不适的味道异常的食物，即使吃食与不适之间的间隔长达 12 小时；与此相对，在吃掉一种味道熟悉但是形状异常的食物几个小时之后开始不适的动物，并没有表现出这种回避学习现象。另一方面，当即时的疼痛取代延迟的不适作为非条件刺激时，得到了一种相反的结果模式；也就是说，当这种刺激伴随有即时的电击时，老鼠立即学会了避开形状异常但是味道没有异常的食物。如同尼斯比特和罗斯（1980）所概括的，老鼠的特性，可能拥有两种“理论”，这两种“理论”都非常适合它们的生态环境中实际的偶然性：（a）当伴随有延迟的（甚至是非常迟的）胃部疼痛时，应该考虑怀疑特殊的味觉刺激；（b）当伴随有立即的肉体疼痛时，应该考虑怀疑特殊的触觉或者空间刺激。

因而用更一般的话来说（比较 Testa, 1974），结论就是机体——人和老鼠——可能发现这些共变，并且只有这些共变，才是他们自己的历史或者他们的种群的历史使他们发现的。当没有“理论”的指引时，共变检测变得非常困难，而且只有当有关的相关关系达到一致，并且/或者当根据这一因素，学习的条件最优时，共变检测才可能出现。实验室中长期闻名的赫尔式（Hullians）的和斯金纳式（Skinnerians）研究方法，解释的就是这个因素。

当然，有些时候，日常环境，是最适宜于学习的。因而，在我们的日常经验中，我们学习数不尽的开关、操作杆、按钮和其他器械的用途，以及令人迷惑的种种迹象、符号和信号的意义，因为相关的共变，是如此地接近完美。或许同样重要的事实是，普通人和正规的科学家一样，能够经常“检验”他开始怀有的新的假设。他们能够仔细制造出样本数据，与意外经验和偶然回忆提供的样本相比，它们更加适合于所讨论的推理任务。

实际上，我们对于我们的环境的控制，逐渐依赖于我们用相对正式推理的工具替代非正式工具的能力。我们的成功，反映了世世代代的普通男女的遗产，他们小心谨慎地观察并且记录下他们的发现，而最近的，则是数不尽的受到正规的实验和统计分析训练的科学家们的研究发现。

我们感谢 Daniel Kahneman、Amos Tversky 和 Paul Slovic 富有建设性的评论和建议。我们也要感谢 Rita French 和 Julia Steinmetz 的坚定支持，他们是本章描述的一些研究的合作者。我们的研究，部分受到国家自然科学基金研究生奖学金对第一作者的支持，以及国家精神健康研究院第 MH-26736 号资金和国家自然科学基金第 BNS-78-01211 号资金对 Lee Ross 和 Mark Lepper 的支持。

本章的结构和内容显然归功于 L. Ross 和 R. E. Nisbett 早年对同一主题的论述，见《人类推理：社会判断的策略与弱点》（*Human Inference: Strategies and Shortcomings of Social Judgment*, Prentice-Hall, Inc., 1980。）

注释

[1] 我们的文献搜索显示，以前的研究者几乎没有运用过连续变量为特征的共变评估任务。那次搜索中所发现的一个例外，是一项很少被引用的研究（报告见 Erlick, 1966; Erlick & Mills, 1967），该项研究主要处理方差对于感知到的相关性的作用，几乎没有得到什么与目前问题相关的数据。



第 16 章

控制的错觉

兰格 (Ellen J. Langer)

尽管大多数人都同意技能和运气之间有很多重叠，但是对于两者之间的紧密关系，仍未得到完全的理解。大体而言，这两者间的区别是很明显的。在技能的情境下，行为和结果之间有着因果联系，所以在技能任务 (skill tasks) 中的成功，是可以控制的。另一方面，运气却是一种机会事件。要靠运气或是侥幸来获得成功的事件，很显然是无法控制的。我们目前所关心的问题，就是这种区别通常是否被人们意识到了。我们的立场是认为人们并没有意识到这一点。关于机会事件，人们虽然嘴上说得好听，但事实上他们的表现，却仿佛机会事件是可以控制的。如果这是正确的，那么确定在这种混淆中起作用的一些变量，将是非常有趣的。

关于人们把机会事件看成是可以控制的这一断言的观察资料，来自社会学家戈夫曼 (Goffman, 1967) 和亨斯霖 (Henslin, 1967)。在拉斯韦加斯研究赌博

的时候，戈夫曼发现那些运气不好的赌博者（dealers）经常会有失去工作的危险。亨斯霖则研究了掷骰子的游戏。研究表明，这些玩家很明显地表现出他们好像可以控制投掷行为的结果。如果他们想得低的分数，他们就会轻轻地投掷，如果他们想得高分，他们就会用力地投掷。他们相信努力和集中精力会有回报。打赌的时候，也会引发控制，比如说，人总是和那些看起来非常有控制力的人打赌。如果人们认为这些游戏是技能性的，那么这些行为就是非常理性的。

222 如果一个人想对机会事件实施控制，他就会在结果出来以前来施加他的影响。有学者（Strickl and et al., 1966）验证了这种说法。被试参加一个投骰子的游戏，在这里他们可以选择两种下赌注的方法，一种是在骰子投出之前，另一种是在骰子投出以后但最终的结果还没有出来之前。他们发现，被试在投掷前下的赌注比投掷后下的要高，即他们会选择高风险。

前边的研究显示，人们通常不能很好的区分可控制的和不可控制的事件。然而，支配这种“虚假控制”行为的因素，还没有被系统地研究。一种可以确定这些因素的方法，就是研究技能型情境的特点。在技能型情境中，人们热衷于各种公开的或私密的行为，为了使成功的可能性最大化，他们会采取几种措施：选择适合此情境的物质；选择该做出什么样的反应，使自己熟悉这些材料和反应；花一些时间来思考该采用什么样的策略来应对即将到来的任务；在积极地投身到任务的同时，继续努力来提高成功的可能性。此外，技能型情境中也有一些因素，它们并非必然由个人所引发，但为了扩大成功的可能性。竞争就是这样的一种因素。

这些与技能有关的因素，也许是导致控制错觉（the illusion of control）的原因。控制错觉，是指个人对成功可能性的期望不恰当地高于客观可能性所保证的。下面的这些研究的设计，用来评估这些与技能有关的因素在引发控制错觉方面的效果。具体来说，我们将描述的这些研究，是用来验证下边这个假设的：在机会事件中，通过鼓励和允许参与者投身于他将会在技能型情境中采取的行为，一个人就增强了此种行为的技能倾向，也就是说这人引发出了控制错觉。这样，这个人就会把前边提到的任何关于技能型情境中因素——选择、刺激，或反应的相似性、消极或积极地参与、

竞争——引入到偶然情境中去，在这种情境中，参与者对结果已经失去了控制力，并且引发一些更适应技能型场景的行为。

对这个假设的一个有力的证明，就是将这些因素引入到一些情境中去，如抽彩票，在这种行为中，结果完全是由运气决定的。如果这些因素成功地在这些机械的场景中引发了控制错觉，当这些因素被引入到一些已经有了控制因素存在的场景中去的时候，它们就会更加显著……

一、实验一：竞争对控制错觉的影响

人们在评估他们的能力的时候总是会热衷于竞争，所以通常假定，在机会性事件中，引入与技能相关的因素，会带来可控性的错觉（an illusion of controllability）。在技能的较量中，一个人事实上掌握的可以导致胜利的结果的控制能力，是作为他的对手的能力的一个函数来变化的。如果人们对机会事件的反应和他们对在由技能决定的场景中的竞争一样，那么控制错觉，也会作为对手的能力的一个函数来变化。 229

在下边的研究中，被试要和两种同谋进行竞争，一种是有吸引力并且自信的，另一种是笨拙的、紧张的。如果被试认为这个任务的结果是不可控制的，那么，除获胜的可能性之外的其他因素，就会在影响被试的打赌方面起到决定性的作用。在这种情况下，当被试和自信的同谋进行赌博的时候，他会倾向于下大的赌注，原因可能是因为被试认为同谋会下更大的赌注，而被试希望表现得和同谋一样；或者是因为这种冒险，在我们的社会中被认为是有价值的（Wallach & Wing, 1968）。当被试和一个笨拙的同谋赌博的时候，他可能也会下大的赌注，是为了表现得和他不同，或者也是因为这种冒险是有价值的。然而，他们有可能在跟笨拙的同谋赌博的时候下的赌注较少，因为他们认为这个同谋会下的赌注比较少，这样被试就可以少冒风险，但仍然有风险。在任何一种情形中，被试跟自信的同谋打赌时所下的赌注，总是比和笨拙的同谋赌博的时候下的赌注多。另一方面，如我们所预测的那样，竞争引发了一种技能定向，那么被试就会在成功的可能性的基础上来进行赌博。因为一个人的对手越没有竞争力，他获胜的可能性就越大，所以，与和那些有自信的同谋赌博相比，被试应该在和那

些笨拙的同谋赌博的时候下更多的赌注。

1. 方法

被试

被试是 36 名男性本科生，他们在耶鲁大学选修了心理学导论这门课程。他们是看到广告后来参加的，这个广告中许诺，如果来参加一个关于认知和心理反应之间关系的研究，可以提供学分并且有机会赢钱。他们被随机地分到两种不同的实验条件下，每一组 18 个人。

程序

当被试进入做实验的房间的时候，他发现已经有一个冒充被试的实验同谋等在那里。这个同谋也是一个本科男生，并且对实验的假设一无所知，他可以扮演自信或者是不自信的人。我们依次称之为自信条件组（dapper condition）和非自信条件组（schnook condition）。

自信条件组

在这种条件下，冒充被试的同谋，表现得自信并且非常开朗，穿着一件非常合体的运动服。他向被试介绍自己，并且指出屋里的一个指示，上边说实验者马上就会回来，要求被试在这段时间里填写一张问卷。为了使这个实验显得像一个心理学的实验，问卷上询问了关于饮食、家庭疾病等类似的问题。被试和同谋在这十分钟左右的时间内填写了问卷，并进行了交流。谈话是非结构性的，但是主要是关于体育运动的。在这个短暂的互动以后，同谋不经意地敲了一下墙（这堵墙把实验者和他们隔开了），以暗示实验者可以回到这个房间里来。

非自信条件组

234 在这种情景下，同谋显得非常的害羞，非常笨拙，并且总是紧张地抖动。他穿了一件非常小的运动服。其他的方面都和自信条件组中保持一致。

在两种条件下，同谋都在实验者进来之前脱掉了衣服。在实验者为迟到表示歉意以后，她指导被试坐好，并且在她准备实验材料的时候不要交谈。通过这些方法，使得实验者好像对实验前期的一些安排并不知情。被试和同谋面对面坐在一张桌子的两侧。在实验者把遥测伏特计、酒精、纱布海绵、电机凝胶、录音机放在桌子上以后，她朗读了下边的实验指

导语：

我们对一些运动反应和认知反应对生理反应的影响很有兴趣。具体来说，我们对皮表阻力作为有压力 and 没有压力的任务的函数的变化感兴趣。我们希望这个实验的设计可以使你们感兴趣，并且我也可以获得我想要的信息。你有机会赢得或者输掉一些钱，所以这会是有兴趣的——但是并不能保证当你离开这个房子的时候会获得多余的钱财。好，首先我要求你们做的就是把这些电极戴在手上，我希望你把它戴在那个你不用来写字的手上，你是左撇子还是右撇子？不要担心，不会受伤的。（实验者给两人戴上电极，插上了遥测伏特计，并拿出了一副纸牌。）第一个任务是纸牌游戏。规则就是你们每个人从这副牌里取出一张，谁取的牌比较大那么这一轮他就赢了。将会有四轮比赛，在每一轮开始之前，你要写下你愿意赌多少，每一轮你都可以从 0~25 美分之间任意选择。然后你把你写下的数字给我看，但是你们不要互相看。也不要看你选中的那张牌。这样你的赌注和结果就不会影响你下一轮的心理反应。我会把牌翻过来然后计算出你们赢的和输的数目，并在结束以后分别告诉你们每个人。这个比赛是在你们每个人和我之间进行，所以如果你赢了我就会付钱，如果你输了你就要付钱或者是时间（subject time）。你们愿意参加吗？（同谋马上回答说，“当然。”）好，那我们开始。当我说准备的时候，你们就写下你们的赌注，这样我就可以做基本观察。

实验者然后就指导被试记录下他们的赌注并且给她。这些赌注被记录下来，然后被试开始分别抽卡片，然后同时把牌给实验者看，实验者记录下这个结果，然后把卡片面朝下放在了旁边的一张桌子上。在每一步开始时，实验者都假装在记录皮表阻力的变化。这个过程被重复四次。

因变量和操作化监测

因变量就是每一轮中被试所下的赌注。

在纸牌的游戏结束后，被试被告知下一步的实验将要单独进行，所以

285 一个被试要到另一个房间中去，那里将有另一个实验者指导他的下一步实验。而且他们还被告知，一旦这个实验结束，这个实验者就会告诉另一个实验者纸牌游戏的结果，这样债务就会被确定下来。实验者告诉同谋离开这个屋子，并且要他们互道再见，因为他们的合作结束了。然后每一个被试被给予了一个篡改了的结果，这样他们每个人获得了两美元，无论他们前边的赌博如何。被试被要求数一罐软糖并确定它的数量，同时实验者测试他的皮表阻力的波动。然后被试要回答一份本质上来说是心理学的问卷。在他被问到另一个被试的存在是否影响了他的心理反应的时候，他被要求在六点量表（6-point scale）上给另一个被试打分，从1（并非非常地有能力）到6（非常地有能力）。其他的问题都是关于心理方面的。当这些测量结束以后，实验者感谢被试并且告诉他们如果他们想了解实验的结果和目的，他们可以在一个月以后打电话来向设计者询问。

2. 结果

在检验是否投入的钱数是同谋的竞争力的一个函数之前，一个很重要的问题，就是要确定在两种条件下同谋是否真的被认为是不一样的。当同谋被假定是衣冠楚楚的时候，他的能力的平均评定分数是4.8，当他伪装成一个非常笨拙的人的时候，他的能力的平均得分是3.17。两种情形之下几乎没有重合。均值之间的差异，是非常显著的（ $t = 5.46, p < 0.005$ ）。所以，我们可以很肯定地说，与非自信条件组相比，在自信组条件下，被试认为他们是在同一个更有竞争力的对手竞争。

我们可以回想起来，被试在四轮比赛中可以投入从0~25美分中的任何数目。这四次的赌博，会被平均起来给每一个被试一个分数。在自信组条件下，被试的下赌的平均数是11.04美分，而在非自信组的条件下其平均数是16.25美分（ $t = 2.39, p < 0.025$ ）。如果我们检验第一次的赌注，我们发现两组之间的差异更加明显，因为第一轮赌注，是紧跟实验操纵的。第一次赌注的平均值，在自信组条件下是9.28美分，而在非自信组条件下是16.72美分（ $t = 3.16, p < 0.005$ ）。

对实验操纵的概念性检验。为了证实我们的假设，即耶鲁的被试期待有吸引力的同谋比那些没有吸引力的同谋时会下更多的赌注，两份问卷被

分给从耶鲁大学的本科生中随机抽出的样本。第一份问卷对任务和参与者进行了描述，让被试来回答他们认为谁更容易下多的赌注。16 个人中，有 12 个人认为有吸引力的被试会下更多的赌注 ($\chi^2 = 4, p < 0.05$)。第二份问卷描述了任务，然后问被试他们会在每一轮当中下多少赌注。回答问卷的所有 15 个被试，都认为他们会下最大的赌注 (25 美分)。

二、实验二：选择对控制错觉的影响

再一次，我们假定，当一个机会情境很类似于一个技能情境的时候，²⁹⁶ 即使成功与否完全取决于运气，人们的行为表现，好像他们可以控制那些不能控制的事件。彩票，提供了一个研究这种控制错觉的非常好的工具，因为，除了决定要不要买彩票之外，所有的结果，都受运气控制。如果一个人可能控制六合彩的结果，他就增加了他对彩票的选择性。对于这个人而言，这张彩票，比没有这种控制力的人的彩票更有价值。如果这张彩票确实具有较高的价值，那么这个人就可以向潜在的买主索要更高的价格。

在接下来的这个研究中，一个抽奖的方法，被用来评估选择（一个在技能场景中非常重要的因素）对于控制错觉的影响，我们预测如果给被试自己选择他们的奖券，他们会对奖券索要一个更高的价钱。

1. 方法

(1) 被试

两个位于长岛的工厂里的成年男女办公室人员，可以购买彩票，这两个公司分别是保险公司和制造业的工厂。^[1] 因为各种各样的抽签和运动的选择，对于这些公司来说并不陌生，对于这次彩票活动的合理性解释，就是多余的了。除了 4 位女性以外，所有其他的人都允许票证公司来替他们买彩票。我们随机地将被试分配到两种场景当中，一种是选择场景，有 24 名男性和 3 名女性，而在没有选择的情境之中，有 23 名男性和 3 名女性。

(2) 材料

彩票是标准的 4 × 2 英寸的 (10.16cm × 5.08cm) 的足球卡。在每一张卡上，有一个足球明星、他的名字和他的球队。卡片首先是按照球队名字

的字母顺序排列的，然后是按照个人名字的字母顺序排列的。这里有两组互相匹配的卡片，每一组包括 227 张足球卡，每一个被试从一组中选出一张，而对应的另一组里的这一张被放入一个硬纸盒里，从中会产生获奖的卡片。

(3) 程序

这次抽奖，是在 1973 年的超级联赛之前的一个星期进行的，由保险公司的一个男性和制造业工厂中的一个女性来完成。这两个实验者，对于实验的假设都一无所知。他们分别处理他们各自代表的工厂中的员工的抽奖活动，他们负责问他们的同事是否愿意花一元钱购买彩票。被试被告知他们的彩票可以在他们的公司和另一个公司中买到（另一个公司的名字也提供了），并且赢者所获得的总金额是 50 美元，并且他们还被告知了抽签的日期。在同意的参加抽奖之后，第一个到达的被试，被给了一个卡片盒，并且让他从中任意选择。被试要说出卡片的名字，这样实验者就可以从另一堆中找出相应的卡片，放在一个封闭的硬纸盒里。同时，实验者会记下被试的名字和他选的卡片的名字。第二个来到的被试也被同样的对待，只有一点不同：他在同意参加抽奖之后，他所拿的一张卡片是前面被试所选择的卡片的匹配卡片。这样，被试就被分到了两种条件之下，第一种就是选择性的，第二种就是非选择性的。在第一个公司买了彩票的第二天，同样的程序在另一个公司里实施。

(4) 因变量的测量

在抽签的那天早上，实验者向每一个从他那里买彩票的人进行了提议。他分别告诉他们每一个人说：“另一个公司里有人也想买彩票，但是我已经不卖票了，所以他想让我问问你们愿意多少钱卖掉你们的彩票，对我来说你们卖多少钱是没有区别的，但是我该怎么跟他说呢？”所给出的量值，就构成了因变量的测量。如果被试说他不愿意卖，实验者就会不断地催促他直到最后他给出一个价钱，他会在记录中写下“不卖”，并且在旁边注明他最后愿意出的价钱。

2. 结果

诚如我们所预言，选择的介入，对于抽奖的价值有很大的影响。在选

择情境中，被试卖出他的票的时候要价的平均值是 8.67 美元，而在非选择的情境中被试要价的平均值只有 1.96 美元 ($t = 4.33, p < 0.005$)。虽然他们被询问的是你们愿意以多少钱卖出你们的票，而不是问你们是否愿意卖你们的票，15 个被试一开始就说他们不愿意卖。这其中，有 10 个被试在选择的情境中，5 个是在非选择的情境中 ($p < 0.10$)。然而前边所说的那种差异，并不仅仅是那些被试在被敦促以后所提供的数目的一个函数，因为他们给出的价位从 3 美元到全部的奖金额 53 美元，只有 3 个人给出了最后这个数字 53 美元。

在这个实验中，所获得的一个结果，是熟悉性对于控制错觉的影响，但是直到下一个实验才被具体证实。女性对于足球运动的熟悉程度不如男性。因此，她们不太容易一开始就参加到这个足球的抽奖活动中来，如果她们确实参加了，她们应该要的价钱更少一些。我们应该还记得有 4 个人拒绝参加这个抽奖，她们全是女性。在那 6 个参加了的女性当中，4 个人要了 1 美元，而两个人要了 2 美元。这样女性要价的平均数就是 1.33 美元，而男性为 5.89 美元 ($t = 2.14, p < 0.05$)。

三、意义和应用

在我们所呈现的证据基础上，被试不能区分机会事件和技能事件，似乎是因为他们对这两种事件的不同界定。客观的偶然性，并不是控制被试行为的关键变量。相反，一个事件是否被作为可以控制的事件来对待，主要取决于如下一些因素，如竞争、选择、熟悉程度和介入程度，这些可能和实际的偶然性没有任何关系。这种情况，在那些明显地被运气所控制的情形中（如抽彩票）也存在着。

为什么会这样？人们总是试图去控制他们的环境。在这种语境下，控制的重要性，已经被临床医学家和社会科学的研究者广泛地讨论过。无论它是被看成一种竞争的需要 (White, 1959)，还是一种想去控制的本能 (Hendrick, 1943)，或者是力争获得优越的地位 (Adler, 1930)，或者是为个人的原因去争取 (deCharms, 1968)，大多数社会科学家都认为人们有一种想掌握个人环境的动机。而一种完全的掌握，则包括“战胜一切不确定

性”的能力，也就是说要控制一切机会事件。一个问题越难，一个人在解决它以后就会获得更大的成就感。最大的满足和成就感，就来自于对似乎不可控制的事件的控制。

在控制的动机以外，对于可控制和不可控制的事件缺乏区分，还有一个原因。技能和机会因素在人们的经验中是紧密相关的，这是事实。也就是说，这里不仅存在一种不去区分的动机，而且在做出区分的过程中，确实存在着很大的困难。因为在每一个技能事件中，机会都会起到作用，而几乎在每一个机会事件中，技能也会起作用。前者是明显的，不需要进一步的解释。后者的例子，见诸在掷骰子中的良好的表现（比如知道偶然性），或者知道哪一个被操纵的老虎机会给出最高的回报……

本章原载《人格和社会心理学杂志》(*The Journal of Personality and Social psychology*, 1975, 32, 311-328)，有删节。允许重印。

注释

[1] 两个公司要匿名。



第 17 章

测验结果的随意性

查普曼和查普曼 (Loren J. Chapman and Jean Chapman)

每一天，精神病学家和临床心理学家都必须要做
出一些重要的决定：

289

他的问题在哪里？他是否该到精神病医院救治？他有自杀倾向还是有杀人倾向？这个病人是可以出院还是要继续治疗？

为了有助于做出决定，临床医师总要用到心理测验。
根据桑伯格 (Norman Sundberg) 的调查，最常用的两种测验，就是罗夏墨迹测验 (Rorschach inkblot test) 和画人测验 (Draw-a-Person test, 简称 DAP)。这两种测验都是投射测验 (projective tests)，其前提假设是，人们对模棱两可的、非结构性的情境 (situation) 进行反应时，会投射出他的一部分人格。举例来说，一个墨迹并没有一个客观形状，个人从中看出的任何东西，都可能反映出他的内在驱力、冲突和人

格。类似地，当人在白纸上画人像的时候，其作品也或多或少地映射出他自己。

一、自我

我们最近的研究表明，罗夏墨迹测验和画人测验，不仅仅在一个方面表现出它们是投射测验。在解释这些测验结果的时候，普通的临床医师，可能会把他自己的先见和假定（preconceptions and assumptions）投射到他对他的病人的描述中。

我们第一个研究是关于画人测验的。在这个测验中，临床医师给被试一支铅笔和一张白纸，让他在上边画个人像。马克娃（Karen Machover）在1949年发表了这个测验。她描述了尤其是那些有种种情绪问题的人画的画，并且解释如何把绘画中的一些特点当作理解他们人格特征的关键。比如，她说，“妄想狂会画较大的眼睛”，而“无论男女，将头画得比较大的人，在其性别内部，往往有与之相称的更高的理智和社会权威”。

马克娃的测验手册，就何人会画什么，有许多大胆的概括，但是她很少给出支持性的数据。

二、部分（parts）

一些临床医师不愿意采纳马克娃的说法，他们用实验来验证她的论断。例如，有学者（Jules Holzberg & Murray Wexler），就试图确定妄想狂是不是一定会画一个突出的眼睛。他们比较了18位妄想型精神分裂症患者和76位实习护士的作品，但是没有发现这两组人画的眼睛有什么区别。

很多类似的研究，也检验了马克娃关于其他一些图画特征的预测，如头部、耳朵、唇、衣着、嘴，等等。但是DAP测验的说法从未被证实。少数实验者发现，善于自我调适的被试似乎画得更好，但是研究所得出的压倒性的结论却是，特定的绘画内容，并不是人格特征的有效指标。

三、迹象（sign）

应该指出，不能完全歧视这类研究。如果50%的同性恋者都用某种方

式绘画，而只有 25% 的其他人用这种方式绘画，那么这种绘画方式仍可作为有效的诊断迹象，因为长远来看，它可能会给我们诊断同性恋提供信息。

多数的临床医师都了解那些否定画人测验的研究，但是成千上万的医师仍经常采用它，因为他们声称这些迹象（sign）在他们的临床实践中发挥了作用。一个临床心理学家说：“在相信那些杂志上的文章之前，我首先相信我自己的感觉。”另一个人说道：“我知道，在实验室里，妄想狂并不一定会画大眼睛，但是在我的办公室里他们确实会这样做。”

四、错觉

一些批评者认为临床医师是如此拘泥于他们的理论和传统，以致对事实视而不见。然而我们认为这里有另一种解释。那些面对反面证据却仍然相信画人测验的临床医师，也许正体验伪相关（illusory correlation），这种现象是我们几年前在词语联想的研究中发现的。

我们发现那些高度互联的词语，似乎被认为比实际更多地同时出现。在这些实验中，被试坐在舒服的椅子上，我们把一些词语对（word-pairs）（如 bacon-tiger）投射到他面前的一个大屏幕上。这些词语对每两秒钟变换一次。这对词语左边的单词总是从四个单词中选择：bacon、lion、blossoms、boat，每个词语出现的频率都和其他的词语一样（25%），只是他们总是出现在屏幕的左边。右侧的词语是 eggs、tiger、notebook，也以相同的频率出现。

我们系统地安排了词语对，这样左边的词语同右边的每一个词语同时出现的频率都是一样的。举例来说，当 bacon 出现在左边的时候，有三分之一的时间在右边出现的是 eggs，而 tiger 和 notebook 在右边出现的频率也分别为三分之一。但是，当我们后来讯问被试这些词语对的时候，他们说当 bacon 出现在左边的时候，右边有 47% 的时间出现的是 eggs；而当左边出现的是 lion 的时候，右边出现最多的是 tiger。虽然事实上这些词语对出现的频率是相同的，但被试却声称那些存在较强词语联想的词语对比其他词语对出现的频率更高。

认为两件事同时发生的频率比实际上多的这种倾向，我们称之为伪

相关。

学生声称有的词语对出现的频率比实际情况要高，而临床心理学家认为画人测验是有效的，但事实却证明是无效的。这两种现象之间，似乎有着本质的相似性。

五、描述

DAP 的迹象和解释，当然和马克娃 20 多年前所引介的已大不相同，所以我们关心的是现在的专业人士如何运用这个测验。我们给 110 个经常用测验来诊断的临床医师发了问卷。我们概要描述了六种类型的病人，并且要求每位临床医师告诉我们每种病人的绘画特征。这六种描述如下：（1）“他担心自己是否有男子汉气概”；（2）“他怀疑他人”；（3）“他担心自己的才智”；（4）“他担心由别人供养并被人照顾”；（5）“他有过性无能”；（6）“他非常担心人们说他的坏话”。我们让临床医师假定这六种人都是男人，并且他们在画一个男性。

242 我们收到了 44 份完成的问卷。很明显，就这六种人的绘画特征而言，这些医师的观点普遍一致。例如，大多数医师（91%）认为疑心重的人会画出大的或者是奇怪的眼睛。82% 的医师认为那些担心自己智力的人会倾向于画一个大的或者是突出的头部，参看表 17—1。

243 表 17—1 报告伴随六种病症描述的各种绘画特征的临床医师和普通学生的比例

| 绘画的特征 | 担心自己的男子气概 | | 怀疑他人 | | 担心智力 | | 担心被别人供养和照顾 | | 性无能 | | 担心人们说自己的坏话 | |
|-------------|-----------|----|------|----|------|----|------------|----|-----|----|------------|----|
| | C | S | C | S | C | S | C | S | C | S | C | S |
| 1. 宽肩膀，肌肉发达 | 80 | 76 | 0 | 6 | 0 | 8 | 0 | 12 | 25 | 31 | 0 | 6 |
| 2. 眼睛反常 | 0 | 0 | 91 | 58 | 0 | 6 | 0 | 3 | 2 | 2 | 43 | 26 |
| 3. 头大或者强调头部 | 0 | 5 | 0 | 13 | 82 | 55 | 2 | 7 | 0 | 3 | 9 | 10 |
| 4. 强调嘴部 | 0 | 0 | 7 | 5 | 0 | 1 | 68 | 8 | 2 | 1 | 5 | 5 |
| 5. 细致描述性征部位 | 14 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 55 | 8 | 0 | 0 |
| 6. 耳朵反常 | 0 | 0 | 55 | 6 | 0 | 3 | 0 | 0 | 2 | 0 | 64 | 7 |
| 7. 面部表情反常 | 0 | 17 | 18 | 44 | 2 | 21 | 2 | 21 | 2 | 14 | 18 | 52 |
| 8. 女性化、孩子气 | 23 | 22 | 7 | 12 | 2 | 11 | 32 | 39 | 23 | 25 | 11 | 13 |
| 9. 头发与众不同 | 23 | 13 | 2 | 2 | 2 | 8 | 0 | 1 | 11 | 6 | 0 | 3 |
| 10. 详细的图形 | 20 | 8 | 2 | 6 | 34 | 13 | 0 | 3 | 7 | 3 | 2 | 6 |

续前表

| 绘画的特征 | 担心自己的
男子气概 | | 怀疑他
人 | | 担心
智力 | | 担心被别人
供养和照顾 | | 性无能 | | 担心人们说
自己的坏话 | |
|---------------|---------------|---|----------|---|----------|---|----------------|----|-----|----|----------------|---|
| | C | S | C | S | C | S | C | S | C | S | C | S |
| 11. 被动的姿势 | 5 | 4 | 2 | 8 | 0 | 2 | 36 | 21 | 2 | 2 | 0 | 8 |
| 12. 衣服上的纽扣 | 0 | 0 | 0 | 0 | 0 | 0 | 23 | 1 | 0 | 0 | 0 | 0 |
| 13. 忽略性征部位 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 | 27 | 0 | 0 |
| 14. 阴茎状的鼻子、肢体 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23 | 2 | 0 | 0 |
| 15. 肥胖 | 0 | 2 | 0 | 1 | 0 | 0 | 7 | 16 | 0 | 4 | 0 | 1 |

注：C = 临床医师，S = 学生，表中列举的特征是，至少就一种病症，有不少于 15% 的临床医师或学生提及，几乎每一种绘画特征，临床医师最频繁联想的病症与学生是相同的。

这种一致性并非完美，但发人深省。整体而言，这些临床医师，在每种类型的病人的作品中，能发现两三个共同点。

六、配对

大多数的临床医师都有哲学博士学位，从事心理治疗的平均年限为 8.4 年。但我们想知道那些几乎没有工作经验的人从画人测验的迹象中可以观察到什么。

为此，我们收集了 45 张男性的画像，其中 35 张是由附近一个州立医院中的精神病患者画的，其余的 10 张是由临床心理学的研究生画的。我们测量了每一张图片上头部、眼睛等部位的大小，而且对这些绘画的更为主观性的特征如男性特质和女性特质等，也有独立的评定。

我们给每个图片附了两个诊断陈述，这些陈述是从前边的六个陈述中挑选出的，比如，“画这幅图片的人（1）怀疑他人，（2）有过性无能。”这六个陈述，因此有 15 种不同的配对方式，这样三张不同的图片，就会有相同的陈述配对。

我们将这些陈述配对系统地分配到所有类型的图片上。如“他担心自己的才智”会同时出现在那些有大脑袋和小脑袋的图片上。

然后，我们筛选了一组大学生，选出 108 名学生。他们声称自己从来没有听说过“画人”测验，并且丝毫不了解这个测验是如何被阐释的。

我们对这些学生进行了分组测试。在每一次测试之前，我们简要地介绍了画人测验的基本原理。我们告诉被试，他们将看到一组图片，图片上

配有对作者简单的描述。我们还说，画画的人中，多数有同样的问题，你们要仔细地检查这些图片，并且要找出那些有同样问题的人在绘画中所表现出来的共同点。然后，学生就根据事先安排好的随机顺序来看图片，每幅图片看30秒。

七、证据

244 虽然我们仔细地平衡图片和陈述，以避免它们之间产生客观的联系，但几乎每一个被试都报告说他们看出了其中的关联。学生们发现的这些关联，跟临床医师所报告的在每天的治疗实践中发现的关联，非常相似。当然这里也有一些区别，但是学生们却试图用与临床医师同样的方式来解释每种类型的病人的典型绘画。而在临床心理学的研究生的绘画中，我们知道这些迹象是虚假的，因为数据并没有显示出来。

我们前边关于词语对的测验，也暗示了一种解释：回想一下我们的发现，那些有很强的关联性的词汇常常被认为是同时出现的。也许在画人测验背后，有同样的机制。我们做了一份单词联想的问卷，以确定一些症状（怀疑、才智、无能等）怎样和身体的某个部位（眼睛、头部、性器官、肌肉等）相关联。问卷采用如下的形式：“怀疑使你联想到头部的倾向性（1）很强，（2）强，（3）适度的，（4）微弱，（5）很微弱，（6）根本没有联系。”

我们把问卷给了45名没有参加过其他部分实验的学生。他们所报告的词语联想的结果，和那些普通学生在症状和绘画特征之间所发现的伪相关几乎一致。而且词语联想和从业医师所报告的相关性更为匹配。

八、付费

在接下来的实验中，我们连续三天对56个被试进行测试。我们想知道，如果多一些时间来看测验材料，他们是否可以意识到在症状和绘画之间并没有真相关。但结果是，这种相关在第三天和在第一天是一样的显著。我们开始意识到这种伪相关是如此有力。我们想知道什么条件可以让人克服这种伪相关。

我们逐个地测试了 41 个新被试，并且让他们看那些图片，时间不限。为了鼓励他们仔细研究，我们答应付给每个做出准确判断的学生 20 美元。

但这也不起作用。学生看出的伪相关和以前一样强。

最后，我们去除了所有的阻碍因素，并且给被试提供我们所能够想到的每一个可以检验他的看法的机会。我们把所有的图画给每一个被试，让他们自己来看。我们告诉他们可以按照任意的顺序来看，并且没有时间限制。他们可以把图片分类，并且进行直接比较。他们可以把所有的“容易怀疑别人的人”的绘画放到一起，然后研究它们的相似性。我们给每人草稿纸、铅笔和橡皮；我们再一次答应奖给判断最为准确的被试 20 美元，而且最后我们也给每个被试一份问卷，这样他们就可以知道要回答的问题了。

九、男子气概

在这种慷慨的条件下，关于大多数症状的伪相关，真的有了显著的下降，但并没有消失。例如，在通常条件下，有 76% 的学生会发现在担心自己的男子气概和画一个肌肉发达的人之间有关联；而在新条件下，仍有 45% 的人声称这些并不存在的关联。伪相关异常有力，并且显著地对抗一切想改变它的企图。

学生们甚至宣称，当那些卡片被以相反的顺序安排的时候，他们发现了典型相关。如在一个研究中，我们把陈述“他担心他的才智”只放在那些头部比较小的绘画中；而把关于怀疑的陈述仅放在那些小眼睛的绘画上等。这在一定程度上有效地减少了伪相关，但并没有消除它。仍有 16% 的人认为那些担心自己的才智的人会画一个大脑袋，50% 的人仍然相信在担心自己的男子气概和画一个肌肉发达的人之间存在关系——即使真实的情况恰好相反。

从我们的研究中，可以清晰地发现，对画人测验的临床解释，似乎带有很强的伪相关的成分。而临床医师对其病人的任何决策，也许是医师自身先见的投射。

十、墨迹

我们想知道在最为常用的罗夏墨迹测验中是否也存在伪相关，如果有，它们是否被看成是真相关，因为在罗夏墨迹中，很少有迹象被发现是某种人格特征的有效指标。

在罗夏测验 50 年的历史中，很多临床医师报告说，一些同性恋者确实比其他人更容易做出一些反应。在 1949 年，惠勒（William Wheeler）总结出了 20 种关于同性恋的罗夏迹象。其他研究者也曾经验证过惠勒所指出的这些迹象，但只有几个研究者发现在这 20 个迹象中只有 2 个是有效的。其中一个（第 7 号）是对第四个墨迹“一个人或者动物——扭曲的、恐怖的或者是危险的”的反应；另一个有效的迹象就是惠勒迹象第 8 号，是对第五张卡片上的不明人兽图的反应。

十一、迹象

246 为了研究临床医师实际上是如何运用罗夏测验来诊断同性恋的，我们对 76 位临床医师进行了问卷调查，让他们描述同性恋者从罗夏测验的 10 个墨迹中看出的两种典型对象。在完整地回答了问卷的临床医师中，有 32 个人表示他们曾看过罗夏关于同性恋的测验方案。他们描述了罗夏提出的几种迹象，但是其中他们最经常提到的是（1）屁股或肛门，（2）生殖器，（3）女性化的衣着，（4）性别特征不明显的人，没有明确的男性或者女性特征，（5）既有男性特征又有女性特征的人。所有的这些都是惠勒提出却没有被证明的迹象。另一方面，只有两个临床医师提到了有效的迹象第 7 号——一个扭曲了的、恐怖的人，而没有一个人提到了另一个有效的迹象第 8 号——一个半人半兽图。

一些临床医师，似乎在罗夏测验中看到了并不存在的迹象，而没有看出实际存在的迹象。我们词汇联想的研究，可再次提供解释。这两个有效的迹象不是直觉性的：同性恋并不是轻易地就会使人想起咆哮的动物或者是半人半兽的杂种。但是同性恋确实和临床医师最经常的那五种迹象有高度的词汇关联性。某种程度上，人们会在直觉上合理地预期同性恋倾向于从墨迹中看到屁股、女性化的衣着或者是性别暧昧的形象。

十二、想法

为了客观地验证这些观念，我们研究了 34 个学生被试，让他们来评定“同性恋”这个词汇引发不同联想的强度。他们的评定，有共识性的在于：那些大众化的但无效的迹象，确实比那些生僻的却有效的迹象与同性恋一词有更高的词汇关联性。这意味着，临床医师所宣称的在罗夏测验中发现到的同性恋的迹象，也许只是反映了他们自己的假设和期待。

采用类似于验证画人测验中的设计方案，我们检验了这个论点。我们选了几张罗夏卡片，并在上边附上了一个反应内容——假想人从卡片上能看出的知觉图案。反应所指的区域，用圆圈圈住。

在一些墨迹上，给出了一个有效的同性恋的迹象（如“一个有着小胳膊的巨人”）；在另外一些墨迹上，给出的则是无效的迹象（如“一个女人的蕾丝文胸”）；而在另一些墨迹上，则是一些中性的迹象（如“一张西班牙地图”）。在这些回答的下边，又给出两句对回答者的个人描述。我们从可能的四种描述中穷尽了所有可能的排列：（1）“他对其他男人有性感觉”，（2）“他认为其他人在算计他”，（3）“他总是感到悲痛和沮丧”，（4）“他很自卑”。当然我们最感兴趣的还是第一个描述。

十三、混合

在画人测验研究中，我们系统地安排了症状描述和图片，这样在任何症状描述和图片中的迹象之间，就不会存在稳定的关系。

在学生看完了一系列的图片之后，我们讯问他们四种症状的病人会报告什么样的罗夏迹象。学生们报告说同性恋者会更多地看到屁股、生殖器 247 等。总之，和临床医师报告说的那五种无效迹象是一致的。没有一个学生看出同性恋者和那两个有效迹象之间的关系。

在后来的设计有些变异的研究中，我们有意地在测验材料中引入负相关，这样，“他对其他男人有性感觉”这个陈述，就从不出现在那些容易让人想到女性服饰和屁股等的图片上。但这并没有减弱伪相关——学生们认为的相关强度仍然和以前一样。

十四、联结

这些研究表明，相信两个独立事件之间存在联系是多么的容易，尤其是当这两个事件之间有主观的词语联想的时候。在这些简洁的、结构性的任务中，我们的被试在症状和投射测验的迹象之间发现了大量的伪相关。当然，临床医师的工作要复杂得多。一个真正的病人的问题是多种多样和模糊不清的——很少有病人只有两种清晰确定的症状。而且真正的病人，在投射测验中会给出不同的反应，而不止是一个。也有可能，在临床实践中，一个临床医师所观察到的伪相关，会被他的同行的报告所强化，而他的这些同行本身也受到了伪相关的影响。这种共识会强化每一个人的错觉（illusions）。另一方面，我们的学生，在测验中却不被允许互相说话，所以每个人只能靠自己来发现伪相关。基于这些原因，很有可能我们的临床医师所面对的伪相关，会比我们的被试所报告的更为强烈。

十五、难度

很多人也许会因此轻易地认为临床心理学家是不合格的，或对事实反应迟钝；但是我们的本意并不在此。我们的数据，并不表明临床医师的无能，而是要表明他们工作的极端艰巨性。和所有其他人一样，临床医师也会受到同样的错觉的影响。类似地，在穆勒-莱尔（Müller-Lyer）的错觉测验中，几乎每一个人都会认为下边的这两条平行线的长度不同，但是没有人会因此认为一个木匠作对距离的判断是不称职的，仅仅因为他也有错觉。



248 如果临床医师想对伪相关进行弥补，他们就必须意识到它的存在。理想意义上，临床医师应该首先体验这种伪相关。一种好的训练策略是，让每一个临床心理学的研究生来观察我们所描述的这些测验。那么，他就会仔细地检查他所体验的伪相关的范围和来源，这样，他才会有希望在未来的临床实践中学会避免这种错误。

这种经历也会提醒他，他的感觉也是会犯错误的，他的临床判断必须

要用客观的方法不断地进行检验，并且他的职业工作是所有心理学领域中最为复杂和艰难的工作之一。

本章原载《今日心理学》(*Psychology Today*, November 1971, pp. 18-22, 106-110)。允许重印。



第 18 章

临床医学中的概率推理：问题及机遇

艾迪 (David M. Eddy)

249 在很大程度上，健康保健的质量和花费，是由医生的诊断来决定的，而医生最终的目标，是设计并执行一套治疗方案以改善病人的健康状况。大多数决定，都涉及很多因素、很大的不确定性，以及一些艰难的价值问题。

这一章将考察这些诊断决策中一个方面的问题，即运用概率推理来分析特定问题：是否该对一个有乳房肿块的妇女进行组织活检，而这肿块可能是恶性的。我们尤其应该了解医生是如何处理乳房 X 射线照片的，一种用来诊断乳腺癌的 X 射线。已有证据表明，医生并不是非常擅长处理不确定性，许多医生在概率推理中会犯一些严重的错误，并且这些错误影响到医疗质量。

一、问题

乳房组织活检 (breast biopsy)，并不是一个无足

轻重的过程。其中最常用的（80%）一种，就是切除组织活检（excisional biopsy），可疑的肿块会被切除，病理学家会对它们进行显微镜检及组织学诊断。通常，病人被收入院，然后进行一系列术前检查。组织活检，通常都是在全麻下进行的（麻醉死亡率大约为每 10 000 人中有 2 人）。需要切一个小切口（1~2 英寸），然后像一个山核桃到李子那么大的组织被切除。多数情况下（1/2），切除后的组织缺损是看不出来的；其他的情况下，会留下一个凹痕。在一些极个别的情况下（1/200），会出现感染，或需要引流疏通。花费大概有 700 美元。这个过程在门诊，并且局麻下即可完成。作为切除活检的一种替换方式，一些外科医生愿意用细针穿刺来获得他们需要的组织。这可以在门诊中完成，不会留疤痕也不会有后遗症，并且便宜得多。然而，对于一些医生来说，这种方式是不可靠的，因为一些恶性的组织，可能被遗漏。 250

影响是否需要组织活检的一个重要因素，就是概率，即这个乳房肿块是癌症的概率。为了推测这种概率，医生需要列出可能的疾病，同时评估每一种疾病会发生的可能的症状和体征的频率，并且按照这些信息来对比病人的表现，并评估她患每一种疾病的机会。如果癌症或其他的可以治疗的疾病的可能性很高，就要进行组织活检。为了帮助医生，一些教科书描述了一些非恶性的疾病和癌症之间的区别。比如，接下来的这一段就是描述一种良性的疾病——慢性囊性病（chronic cystic disease）。

“慢性囊性病，经常和乳腺癌相混淆。它经常发生在生育过的且乳房较小的妇女身上。它经常发生在外上象限，但是有时候也会在其他部位，并且最终累及整个乳房。通常会疼痛，尤其在经期前，并且通常伴随着月经的紊乱。大约 15% 的病例会出现较为严重的乳头渗液，但乳头本身无改变。病变弥散，边界不清，不与皮肤固定。多个腺体坚硬、圆形、欠光滑，若含有清澈液体可有透光性。慢性囊性病变中单个巨大腺体类似肿瘤，但多表面光滑且边界清楚。腋窝淋巴结通常无肿大，罕见的慢性囊性病会出现大的蓝色的囊肿。通常囊肿是多发的且较小。”^[1]（del Regato, 1970, pp. 860-861）

相似的描述也同样适用于纤维腺瘤、脂肪坏死、外伤和其他很多乳房

的疾病，同样也适用于癌症。

这种概率性的信息，可以用来帮助医生分析病人乳房肿块的可能的病因。通过对可能的结果的评估（比如，正确诊断癌症，对良性的病变做多余的组织活检，或者是决定对非恶性的损坏不做活检），根据病人的症状和体征，医生可以确定他的病人是癌症的可能性，并因此选择如何行动。

二、乳房射线检查

251 其他一些诊断方法，也可以帮助医生来确定一个女人的乳房病变是恶性的可能性。通常最常用的也是最重要的，就是乳房的射线检测（mammography）。这种检查的价值，基于恶性细胞对 X 射线的吸收不同于非恶性的细胞对它的吸收。通过研究乳房射线检查，放射医师可能会看出一些征象（signs）在不同的病变中的不同的频率，通过这些信息可以对有问题的病变组织进行判断。通常，乳房射线检查，对癌症来说，可以分成阳性和阴性。偶然情况下，扩展的分类被使用，比如有一种包含三种分类：恶性的、可疑的、良性的。

这种检测并不完美，因为一些恶性的病变，被错误地划分为良性的，而一些良性的被化分成恶性的。这样，一个因素对于医生来说就非常重要，那就是诊断的准确性。

1. 概率推理

让我们来精确地研究一下这个概念。诊断检查的目的，是给医生提供关于病人状况的信息。医生用这些信息来修改关于病人状况的评估，并且根据新的评估来选择治疗方案。这种行动可能会成为下一步临床检查的指导，或者，如果医生对病人的状况非常自信的话，就会根据方案对病人进行治疗了。关键的问题就在于，医生对病人的状况要有足够的把握，医生要不断收集新的证据来确定他的自信，即病人是有癌症还是没有。并且当这种确定变得非常肯定的时候（如病情恶化，或治疗后预后改变），就会采取行动。

我们可以将医生认为病人有癌症的主观概率和确定性程度联系在一起。因此，对病人进行诊断检查如乳房射线检查的重要性，在于这种检查有力