

★★★ 教育部高等学校心理学教学指导委员会推荐 ★★★

约翰·肖内西

【美】 尤金·泽克迈斯特 著

珍妮·泽克迈斯特

张明

等译

第 7 版

心理学研究方法

Research Methods
in Psychology

John J. Shaughnessy

Eugene B. Zechmeister

Jeanne S. Zechmeister

 人民邮电出版社
POSTS & TELECOM PRESS

心理学研究方法

Research Methods
in Psychology

这本由约翰·肖内西等人编撰的《心理学研究方法》具有众多的特色。第一,这本书已经出到第七版,从内容安排到文字表达都比较成熟,事例丰富,通俗易懂。第二,与一些《实验心理学》教材相比,这本书更注重介绍实际研究中的思路、方法和过程,而不是具体的研究内容(如记忆,语言等)。我认为,这种对方法和思维方式的强调可能正是中国心理学在教学和研究中所缺乏的。只有加强对学生提出问题、思考问题、解决问题的训练,中国的心理学研究才能真正具有创新性。第三,在对心理学研究方法的具体介绍上,本书超出了一些传统教材所涉及的范围,例如本书较为详细地讲述了个案(single-case)研究方法。这种对传统范围的突破,有助于学生把所学的知识应用到心理学研究的各个分支(如神经心理学)中去,有助于他们认识到心理学研究的精髓和共通之处。第四,也是我要特别强调的,这本书对心理学研究中的道德问题和论文的写作有比较恰当的关注。

总之,我认为,这本《心理学研究方法》是一本优秀的教材。如果我们的学生能够真正领会书中介绍的内容和规范,我国心理学队伍的水平必将有一个较大的提高。

我热忱推荐这本书。

周晓林 教授
北京大学心理学系 系主任

北京新曲线出版咨询有限公司

网址: www.ncc-pub.com

ISBN 978-7-115-22326-5



9 787115 223265 >

ISBN 978-7-115-22326-5 / F

定价: 68.00元

分类建议: 心理学

人民邮电出版社网址: www.ptpress.com.cn



第 7 版

心理学研究方法

约翰·肖内西
【美】尤金·泽克迈斯特 著
珍妮·泽克迈斯特

张明 等译

人民邮电出版社
北 京

图书在版编目(CIP)数据

心理学研究方法:第7版/(美)肖内西(Shaughnessy, J. J.), 泽克迈斯特(Zechmeister, E. B.), 泽克迈斯特(Zechmeister, J. S.)著; 张明等译.

-北京:人民邮电出版社, 2010. 4

ISBN 978-7-115-22326-5

I. 心… II. ①肖… ②泽… ③泽… ④张… III. ①心理学研究方法 IV. ①B841

中国版本图书馆CIP数据核字(2010)第030002号

John J. Shaughnessy, Eugene B. Zechmeister, Jeanne S. Zechmeister

Research Methods in Psychology, Seventh Edition

ISBN 0-07-298622-0

All Rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including without limitation photocopying, recording, taping, or any database, information or retrieval system, without the prior written permission of the publisher.

This authorized Chinese translation edition is jointly published by McGraw-Hill Education (Asia) and Posts & Telecom Press. This edition is authorized for sale in the People's Republic of China only, excluding Hong Kong, Macao SAR and Taiwan.

Copyright © 2010 by McGraw-Hill Education (Asia), a division of the Singapore Branch of The McGraw-Hill Companies, Inc. and Posts & Telecom Press.

版权所有。未经出版人事先书面许可,对本出版物的任何部分不得以任何方式或途径复制或传播,包括但不限于复印、录制、录音,或通过任何数据库、信息或可检索的系统。

本授权中文简体字翻译版由麦格劳-希尔(亚洲)教育出版公司和人民邮电出版社合作出版。此版本经授权仅限在中华人民共和国境内(不包括香港特别行政区、澳门特别行政区和台湾地区)销售。

本书封底贴有 McGraw-Hill 公司和人民邮电出版社防伪标签,无标签者不得销售。

北京市版权局著作权合同登记号: 01-2006-4945

心理学研究方法(第7版)

- ◆ 著 (美)约翰·肖内西 尤金·泽克迈斯特 珍妮·泽克迈斯特
译 张明等

策划 刘力 陆瑜

责任编辑 刘力 常玉轩

装帧设计 陶建胜

- ◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街14号A座
邮编 100061 电子函件 315@ptpress.com.cn
网址 <http://www.ptpress.com.cn>
电话 (编辑部)010-84937150 (市场部)010-84937152
(教师服务中心)010-84931276

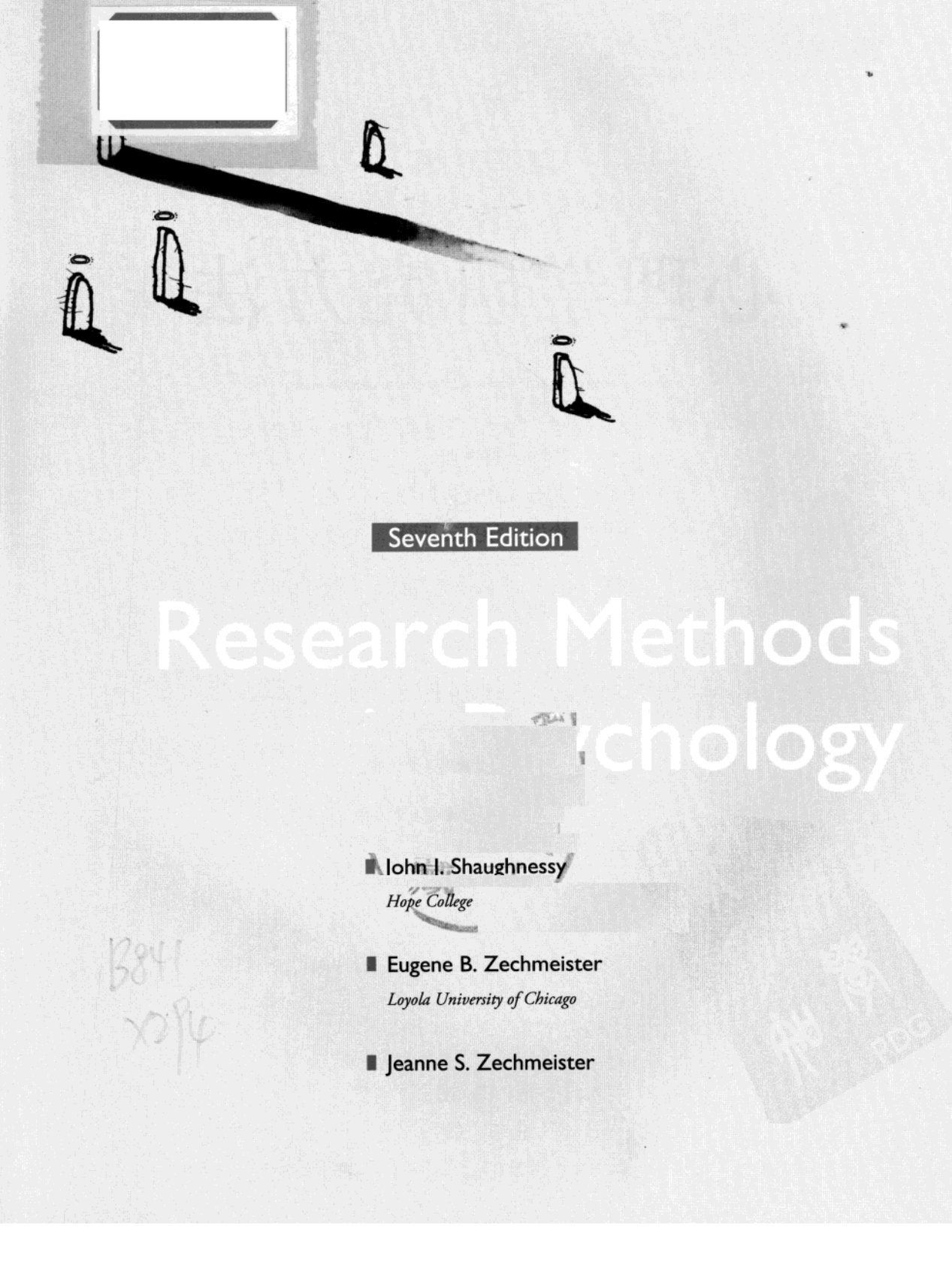
北京圣瑞伦印刷厂印刷

新华书店经销

- ◆ 开本: 850×1092 1/16
印张: 28
字数: 680千字 2010年4月第1版 2010年4月第1次印刷
著作权合同登记号 图字: 01-2006-4945
ISBN 978-7-115-22326-5/F

定价: 68.00元

本书如有印装质量问题,请与本社联系 电话:(010)84937153



Seventh Edition

Research Methods in Psychology

■ John I. Shaughnessy

Hope College

■ Eugene B. Zechmeister

Loyola University of Chicago

■ Jeanne S. Zechmeister

B841
x284

PDG

译者

(以章次为序)

东北师范大学	张 明 教授	第1、2章、前言、附录、专业术语表
河北师范大学	鲁忠义 教授	第3、4章
陕西师范大学	游旭群 教授	第5、6章
南京师范大学	刘 昌 教授	第7、8章
中山大学	高定国 教授	第9、10章
北京大学	吴艳红 教授	第11章
华东师范大学	郭秀艳 教授	第12章
首都师范大学	丁锦红 教授	第13章
辽宁师范大学	张 奇 教授	第14章

内容提要

这是一本供心理学专业学生使用的研究方法教材。

作者选择大量最新的研究实例，详细说明了心理学研究中经常要用到的各种描述性方法（包括观察法、调查研究、行为的隐蔽测量）、实验性方法（包括独立组设计、重复测量设计和复合设计）以及应用性研究（个案研究法和单被试设计、准实验设计和项目评估），并简单叙述了数据处理和统计的基本概念、心理学研究报告撰写和心理学学术交流中应该注意的问题。

本书适合心理学专业大学本科和研究生课堂学习使用，也可供心理学研究方法和实验心理学课程的教师以及从事心理学教育和科研的工作者参考，同时还可以供管理学、社会学、政治学等学科的教学和研究人员借鉴。

心理学研究方法:第7版/(美)肖内西(Shaughnessy, J. J.), 泽克迈斯特(Zechmeister, E. B.), 泽克迈斯特(Zechmeister, J. S.)

—北京:人民邮电出版社

ISBN 978-7-115-22326-5

要 容 内

I. 心… II. ①肖… ②译… III. ①心理学研究方法 IV. ①B841

中国版本图书馆 CIP 数据

John J. Shaughnessy,

Research Methods in

ISBN 0-07-298622-2

All Rights reserved. No part

reproduced, including with

www.wiley.com

This authorized Chinese

author is authorized for sale

Copyright © 2008 by

and Posts & Telecomm

版权所有。未经出版

制、录音、或经过任何

本授权中文简体字

属共和国境内(不包

本书封面贴有 McGraw-Hill

北京市版权局著作权

约翰·肖内西等著的《心理学研究方法》是一本非常优秀的心理学教科书,2004年人民邮电出版社曾以影印本的形式将这本书介绍给国内的读者。如今我们将其翻译成中文,是希望能有更多的读者可以从中获得帮助。关于本书的主要特色,在周晓林教授为影印本所作的推荐中已经进行过很好的归纳,故此处不再赘述。

刚刚开始从事心理学研究的学生常会为如何进行研究设计而感到困惑,一些学生有很好的理论素养并且有了很好的想法,但却不知道如何将其付诸实践;另一些学生尽管做出实验设计,却未能将其好的想法以恰当的方式转化为良好的研究成果。我以为造成这种窘境的原因其实是研究新手对以往的研究方法和规范没有一个透彻的了解,学生往往只凭自己的主观臆想或是其他人的只言片语来开展研究。这就需要有志于学术研究的学生花一些时间学习有关研究方法的知识。利用前人摸索出来的经验,他们可以更快地步入正轨,且能有效减少走弯路所浪费的时间。毕竟,站在巨人的肩膀上,才能看得更高更远。

同时需要强调的是,心理现象的复杂性决定了对它的研究将不会有一个固定的模式或方法,任何方法都是在发展中的。研究者掌握了一定的方法后,切不可盲目迷信,把某种工具或方法当作一种绝对的权威,这样做就会制约研究进展和新思想的产生。当然这些都是后话,这本书的作用在于给你一个推力,让你能够爬到巨人的肩膀上。

这本书由我主持翻译,参与翻译的同仁有:首都师范大学丁锦红教授、南京师范大学刘昌教授、北京大学吴艳红教授、辽宁师范大学张奇教授、华东师范大学郭秀艳教授、中山大学高定国教授、陕西师范大学游旭群教授、河北师范大学鲁忠义教授;此外,我的学生刘幸娟、温婷、刘亨荣、曹海丽、高丽在书稿编辑和审校过程中也做了大量的工作。借此机会对参与书稿翻译的教师和同学致以由衷的感谢,正是各位辛勤的劳动才使这本书的中文版得以与读者见面。

尽管我和各位译者花费了很多的时间和精力在此书稿之中,但译本中的错误在所难免,还希望读者诸君不吝指正。

张 明

东北师范大学教育科学学院心理学系 系主任、教授

2010年1月于长春

推荐序

近年来,“心理学”成为一个热门词汇,越来越多的人开始认识到心理学的重要性,国内各高校也不断开设心理学专业,心理学专业的招生规模一直在扩大。作为一个心理学工作者,一方面,我为此骄傲,因为它们反映了中国心理学事业的飞速发展,反映了中国社会对心理学的需求;另一方面,我也在担忧,担心我们培养出的学生学术水平每况愈下,招收越多,水分越大。事实上,这种担心不是没有根据的。如果每个研究生在几年的学习时间内只发表一篇文章,我们每年就应该有一千多篇论文,而实际情况是:大部分研究生的毕业论文远远达不到正式发表的水平。

依我看来,阻碍这些研究生水平提高的一个主要因素就是,这些学生(包括我们的一些教师)在心理学研究方法上所受到的训练薄弱,没有真正体会、掌握心理学实验设计与数据处理和推论的本质。一个学生,如果在本科阶段没有真正学好心理学研究方法、统计学和认知心理学这三门课程,就不是一个合格的心理学毕业生。

许多心理学界同行都持有与我相似的看法。现在市面上已有几本比较优秀的《实验心理学》、《心理学研究方法》之类的编写或翻译著作。与这些已经出版的书籍相比,这本由约翰·肖内西等人编撰的《心理学研究方法》具有众多的特色。第一,这本书已经出到第七版,从内容安排到文字表达都比较成熟,事例丰富,通俗易懂。第二,与一些《实验心理学》教材相比,这本书更注重介绍实际研究中的思路、方法和过程,而不是具体的研究内容(如记忆、语言等)。我认为,这种对方法和思维方式的强调可能正是中国心理学在教学和研究中所缺乏的。只有加强对学生提出问题、思考问题、解决问题的训练,中国的心理学研究才能真正具有创新性。第三,在对心理学研究方法的具体介绍上,本书超出了一些传统教材所涉及的范围,例如本书较为详细地讲述了个案(single-case)研究方法。这种对传统范围的突破,有助于学生把所学的知识应用到心理学研究的各个分支(如神经心理学)中去,有助于他们认识到心理学研究的精髓和共通之处。第四,也是我要特别强调的,这本书对心理学研究中的道德问题和论文的写作有比较恰当的关心。中国科学界普遍忽视学生研究和写作中的职业道德和规范问题。许多学生也不清楚什么时候、以什么方式承认前人的贡献,导致一些人贬低或故意忽略他人的研究来突出自己的“创新”之处,(我本人就多次受到这样的伤害)。更多的学生和教师不懂得心理学论文的写作规范,不知道如何从读者的角度思考,在论文

中把当前研究的问题、已有的前人研究和自己的实验设计、实验假设连贯起来。课堂教学中的实验报告和许多学术刊物只要求作者报告做了什么，而不要求作者说清楚为什么要研究这个问题、为什么要这样设计、你的研究与已有研究究竟有什么不同。我个人认为，长期以来，这种教学和刊物要求让我们的学生和研究者形成了肤浅和懒惰的思维模式，极大地阻碍了中国心理学研究水平的提高。《心理学研究方法》这本书把科研道德和写作规范作为心理学最主要教学内容的一部分，有助于我们改变现状，改善学生的综合研究素质。

总之，我认为，这本《心理学研究方法》是一本优秀的教材。如果我们的学生能够真正领会书中介绍的内容和规范，我国心理学队伍的水平必将有一个较大的提高。

我热忱推荐这本书。

周晓林

北京大学心理学系 系主任、教授

张明

2010年1月于长春

前言

我们编写这部研究方法教科书已经二十多年了，希望本书能够帮助学生发现心理学研究的乐趣，同时掌握心理学研究的方法。看到学生从刚上课时略带惊恐的神情向完成学习后的充实与自信转变，这正是我们编写这本书和讲授研究方法这门课所获得的最大满足。

第7版的修订采纳了学生和读过之前版本的心理学教师的意见和建议，我们也从中受益匪浅。尽管让所有人满意是不可能的事情，但是，我们还是希望在新版中的这些修订能够更好地介绍心理学的研究方法。对于初次使用本书的读者，请先阅读下面的“结构和方法”，而已经使用过本书先前版本的读者则可以直接跳到“本版的修订”来了解我们更新了哪些内容。

结构和方法

书中所讲的研究方法来源于我们多年的教学经验。作为教授心理学研究方法课程的教师，我们发现课堂上的大多数学生都是研究的消费者而不是生产者。无论哪种角色的学生都将从批判性思维技巧中受益。因此，我们相信可以在研究方法的学习过程中，通过问题解决的方式来更好地帮助学生锻炼批判性思维。研究者通常从有价值的问题切入，然后针对问题选择最好的研究方法。有些时候，痛苦的证据收集工作仅仅是研究进程的开始，但对于心理学研究进程中的结论部分，分析和解释证据同样重要。研究者（和学生）为了批判性地评估他们所获证据的性质，还必须分析他们所选用的研究方法的优点和不足。

另一个特点是，从上一版开始，我们为本书设计了一个网站。这里有许多交互式的练习和测验供学生检验学习效果，还有很多重要的心理学网站的链接。教师则会发现有用的教师指导、课程或讨论指南。学生和教师都可以通过这一网址联系到本书的作者。请访问我们的主页 www.mhhe.com/shaughnessy7（此项专为使用英文原版者提供——译者注）。

最后，我们认为使用已经发表的心理学研究来教授这门课程是最好的教学方法。因此，在探讨研究者实际使用的方法时，我们会继续使用大量心理学文献作为例子。更新研究例子的过程总是充满乐趣，特别是增添了一些能够帮助学生有效学习研究方法的重要的“经典”发现和研究。我们相信，激发学生追求知识的方法之一是向他们展示学习心理学研究方法能够得到哪些收获和回报。

本版的修订

我们将继续在每章中使用要点提示及每章结尾的复习题来帮助学生掌握每章

的重点,还会继续通过每章后的“挑战性问题”来协助学生学习运用学到的理论来解决问题。基于挑战性问题模式,我们还在大多数章节中插入了“拓展练习”,该专栏的目的也是供学生学习如何运用其所学知识解决问题。对统计学的大范围回顾仍然设在本书的结尾部分(第12和13章),但在本书的其他章节也会就统计问题进行适当的简要介绍。我们认为这种灵活的结构有助于教师自行确定何时以及如何研究课程中教授统计学知识。在这一版中,我们试图将统计学和研究方法更清晰地联系起来。我们还引入了一种叫做“统计小窍门”的教学方法,并非常欢迎教师能就这一新栏目的效果作出反馈。当然,我们也很重视这种研究方法与统计的紧密联系,并且有意地通过正文和“统计小窍门”栏目向学生强调这一点。希望我们的方法对学生学习和教师教学都能产生积极的作用。

以下是本版中的主要修订部分:

- 我们彻底修订了第1章,以使它对学生更有吸引力并且更具挑战性。我们以新的一节名为“科学心理学”的内容作为本书的开篇,其目的有两个:(1)强调科学研究方法的重要性;(2)通过突出“媒体暴力”研究的主要发现来激发学习的兴趣。此外,我们还添加了2002年诺贝尔奖获得者丹尼尔·卡尼曼的研究。第一章还提供了一份表格,该表格描述了研究过程的步骤,并为学生提供了一个非正式的学习框架。因为我们每版教材中采用的方法都是为了让学生会利用多种方法来解决,都是为了能最好地促进科学心理学的发展,所以本书的一个目的就是使学生的“工具箱”中装满各种心理学研究策略。最后,我们用要点小结、重要概念、回顾和挑战性问题来巩固第一章的内容。
- 第2章在描述科学方法的重要性上几乎没有改变。通篇我们都致力于消除那些容易让学生混淆的细微差别(例如,在介绍自变量时提到与个体差异有关的各种变量)。审阅者还建议对上一版中研究四个目标中的后两个目标进行重新命名,这四个目标现在是:描述、预测、解释和应用。
- 为了更好地帮助学生理解重要概念,我们在每章中首次涉及某一概念的地方(有少数例外)用黑体字标示。这表示一个被作为重要概念的概念只在相应的章节中出现(通常是它较早出现的章节),这和之前的版本不同。并且所有的重要概念都会在每章的最后单独列出。以下的概念已被加入本版的关键概念列表中:应用研究和基础研究,科恩 d 值,两均值比较,构念,民族中心主义,新异效应,虚无假设显著性检验(NHST)。
- 研究方法中最令人兴奋的进展是基于网络而进行的研究的增加。研究者在网站上实施实验,吸引来理论上数以千计的被试。然而,这种时髦的方法却引起了很重要的伦理争议。我们在很多的章节中都将讨论互联网在研究中的应用和伦理上的争论,特别是在第3章和第5章中。
- 第10章的标题改为“个案研究法和单被试设计”,以便更好地反映该领域中的研究,但其内容基本未作改动。
- 前面提到过,为了帮助学生更清楚地了解方法学和统计学的联系,我们在大多数章节中新增了“统计小窍门”栏目,用以把学生的注意吸引到统计分析

的问题上来。在一些案例中我们为学生解答这些问题；在其他情况下，我们让学生参考第 12、13 章的材料。我们希望这些内容能够对学生学习收集数据和分析数据之间的联系有所帮助。

- 我们在统计部分的一个改变是，我们强调自变量多于两个水平时应该做“后续分析”。我们强调“两个平均数”的比较，强调根据事先确定的系数来进行计划中的比较。在进行“后续分析”时，我们强调了置信区间对均数比较解释中的重要作用。许多观察者已经指出（Fidler, Thomason, Cumming, Finch, & Leeman, 2004），尽管有人建议减少对虚无假设显著性检验的依赖，但它还是研究者们首选的传统统计方法。我们希望学生们能够明白在他们的统计“工具箱”中有许多工具可用，而不仅仅只有一种。

致 谢

因为无法向为本书第 7 版的修订工作作出贡献的所有人员逐一感谢，在此我们谨向以下审阅者表示谢意，并为我们未能完全参照他们的修改意见而深表遗憾。

Stephen Burgess
*Southerwestern Oklahoma State
University*

Laura Carlson
University of Notre Dame

Lindsey L. Cohen
West Virginia University

Philip T. Dunwoody
Mercer University

Samuel Hill
The Sage Colleges

Rolf Holtz
University of Florida

Christine Elizabeth Hughes
*University of North Carolina at
Wilmington*

H. Durell Johnson
*Pennsylvania State University—
Worthington Scranton*

Elizabeth C. Lanthier
*Northern Virginia Community
College*

Corrinne Lim
Purdue University

Richard E. Mattson
*State University of New York at
Binghamton*

Mark O'Dekirk
Meredith College

Steven L. Schandler
Chapman University

Jerry I. Shaw
*California State University—
Northridge*

John Shelley-Tremblay
University of South Alabama

Michael Young
Southern Illinois University

我们也希望向芝加哥 Loyola 大学的同行 Emil Posavac 和 Patricia Rupert 致谢，感谢他们对本书的一贯支持和提供的宝贵建议。

我们为本书及本书之前版本的出版者麦格劳—希尔公司的编辑和出版人员致敬，特别向给予我们支持的 Melissa Caughlin 和 John Wannemacher 致谢。并向麦格劳—希尔公司的其他工作人员表示最深的谢意，他们是：责任编辑 Kirsten Stoller；项目经理 Cathy Iammartino；设计 Gino Cieslik；图片研究协调员 Nora Agbayani；生产总监 Jason Huls；美术编辑 Katherine McNab。

约翰·肖内西
尤金·泽克迈斯特
珍妮·泽克迈斯特

20 统计交互作用

详细目录

21 统计交互作用
22 统计交互作用

23 统计交互作用

第一编 一般问题

第1章 绪论

科学心理学

科学的背景

历史背景

社会和文化背景

道德背景

像研究者一样思考

评价媒体报道中的研究发现

开始做研究

小结

第2章 科学方法

科学方法和日常方法

一般方法

观察

报告

概念

仪器

测量

假设

科学方法的目标

描述

预测

解释

应用

科学理论的建构和检验

小结

第3章 心理学研究中的伦理问题

引言

研究前需要思考的伦理问题

APA 伦理标准

注解

风险收益比

风险判定

1

2

3

5

5

8

10

11

14

15

18

21

22

22

23

25

27

29

30

30

32

32

34

36

38

38

41

45

46

47

47

47

49

50

第五编 分析和报告研究结果

介绍第5编

92

第12章 数据分析与解释

描述数据、置信区间、相关系数

101

最小风险

风险应对

知情同意

APA 伦理标准

注解

心理学研究中的欺骗

APA 伦理标准

注解

事后解释

APA 伦理标准

注解

用动物做研究

APA 伦理标准

注解

心理学研究的报告

APA 伦理标准

注解

伦理决策的步骤

小结

第二编 描述性方法

第4章 观察法

概述

行为样本

时间抽样

情境抽样

观察法的分类

非干预观察

干预观察

参与性观察

结构化观察

现场实验法

行为记录

行为的定性记录

行为的定量测量

观察数据的分析

51

51

53

54

55

59

60

60

61

63

63

65

65

66

68

68

69

71

71

77

78

79

80

80

81

82

83

84

85

87

90

90

91

93

97

数据简化	97	问题与局限	165
观察者信度	99	伦理问题与隐蔽测量	169
批判性思考观察研究	101	小 结	170
观察者的影响	101		
观察者偏差	104		
小 结	106		
第 5 章 调查研究	110	第三编 实验方法	173
概 要	111	第 7 章 独立组设计	174
调查研究的使用	111	概 述	175
调查研究的特点	112	心理学家为什么要做实验	175
调查研究中的抽样	113	实验研究的逻辑	176
抽样的基本术语	113	随机组设计	180
抽样方法	116	随机分配	180
调查方法	120	对内部效度的威胁	182
邮寄调查法	120	实验结果的分析和解释	186
个人面谈法	122	数据分析在实验中的作用	186
电话访谈法	123	描述结果	188
网络调查法	124	确认结果所揭示的内容	190
调查研究的设计	125	数据分析所不能告诉我们的	193
横断设计	126	确立实验结果的外部效度	193
连续独立样本设计	127	匹配组设计	196
纵向设计	129	自然组设计	198
问 卷	131	小 结	199
作为调查工具的问卷	132	第 8 章 重复测量设计	204
自陈测量的信度和效度	134	概 述	205
编制问卷	136	研究者为什么使用重复测量设计	205
关于调查研究的批判性思考	141	练习效应在重复测量设计中的作用	207
报告行为与实际行为的一致性问题	141	练习效应的定义	207
相关与因果关系	143	在完全设计中平衡练习效应	209
小 结	146	在不完全设计中平衡练习效应	213
第 6 章 行为的隐蔽测量	150	重复测量设计的数据分析	217
概 述	151	描述结果	217
物理痕迹	152	确定结果所揭示的内容	218
基本原理	152	差异转换问题	220
物理痕迹的类型	153	小 结	221
问题与局限	156	第 9 章 复合设计	224
档案数据	158	概 述	225
基本原理	158	确定实验设计的指导原则	227
档案数据的类型	162	描述复合设计中的实验效应	227
内容分析	164	2x2 设计举例	229
		主效应和交互作用	231

描述交互作用	233	第五编 分析和报告研究结果	305
复合设计的数据分析	234		
存在交互作用的数据分析	237	第 12 章 数据分析与解释 (I):	306
不存在交互作用的数据分析	239	描述数据, 置信区间, 相关	306
解释交互作用	241	概 述	307
交互作用和理论检验	241	呈现研究结论	308
交互作用与外部效度	241	计算机辅助的数据分析	309
交互作用与天花板效应和地板效应	243	例证: 一个实验研究的数据分析: 均数比较	309
交互作用和自然组设计	244	第一阶段: 了解数据	310
小 结	245	第二阶段: 概括数据	314
		第三阶段: 用置信区间来判断数据所反映的意义	318
第四编 应用研究	249	例证: 一个相关研究的数据分析	328
第 10 章 个案研究法和单被试设计	250	第一阶段: 熟悉数据	329
概 述	251	第二阶段: 概括数据	329
个案研究法	252	第三阶段: 为相关构建一个置信区间	334
个案研究法的特点	252	小 结	334
个案研究法的优点	254		
个案研究法的缺陷	258	第 13 章 数据分析与解释 (II):	339
对个案研究证据的批判性思考	260	统计显著性检验与分析	339
单被试 (小样本) 实验设计	261	概 述	340
单被试实验的主要特点	262	虚无假设显著性检验 (NHST)	340
具体的实验设计	264	实验敏感性和统计效力	343
所有单被试设计共同的问题和局限性	271	NHST: 两个平均数的比较	344
小 结	272	独立组设计	345
		重复测量设计	346
第 11 章 准实验设计和项目评估	276	统计显著性和理论及实际意义	346
概 述	277	均数比较的建议	347
真实实验	277	均数比较的结果报告	348
真实实验的特点	277	多重条件的数据分析	349
在自然环境中进行真实实验的困难	278	单因素独立组方差分析	350
控制真实实验的内部效度	281	三个以上独立组设计的效应大小计算	354
真实实验也可能控制不到的问题	283	独立组设计的统计效力估计	355
准实验	285	多组实验设计的均数比较	355
非等控制组设计	287	单因素独立组实验设计的结果报告	357
非等控制组设计: 兰格和罗丁的研究	288	重复测量的方差分析	358
影响非等控制组设计内部效度的因素	289	独立组设计的双因素方差分析	361
外部效度问题	293	含有交互作用的复合设计分析	362
间断时间系列设计	293	无交互作用的方差分析	363
非等控制组的时间系列	296	双因素独立组设计的效应大小计算	364
项目评估	297	复合设计分析中置信区间的作用	364
小 结	301	双因素混合设计的方差分析	366
		复合设计的结果报告	368

小 结	368	参考文献	390
附录	391	附录	391
作者注释	391	作者注释	391
脚 注	391	脚 注	391
文稿各页的排序	392	文稿各页的排序	392
口头报告	392	口头报告	392
研究计划	393	研究计划	393
研究报告写作范文	395	研究报告写作范文	395
附录 统计表	415	附录 统计表	415
术语表	423	术语表	423
参考文献	433	参考文献	433

科学心理学

第一章

第一编

一般问题

第1章
绪论

第2章
科学方法

第3章
心理学研究中的伦理问题

第1章

绪 论

本章大纲

科学心理学

科学的背景

历史背景

社会和文化背景

道德背景

像研究者一样思考

评价媒体报道中的研究发现

开始做研究

小 结

科学心理学

- 心理学家通过建立理论和开展心理学研究来解答关于行为和心理过程的问题，对这些问题的解答能够对个体和社会产生影响。
- 科学方法是获取知识的一种方式，包括提问的方式、解答问题的逻辑和方法。
- 科学方法的两个重要特征是实证方法和怀疑态度。

你可能已经通过媒体和心理学课程了解了很多心理学方面的研究成果。如果你像本书作者一样对心理和行为充满了好奇，你可能会思考人类（或动物）的行为；你可能想知道为什么人们会有这样的行为，人们是如何形成自己独特人格的，独特的人格又将如何发展和演变；你可能还想知道自己的行为和心理的原理与规律。正是这些想法和思考使得你区别于一般的人——要知道并非所有人都对心理感兴趣，也不是所有人都会思考行为产生的原因。如果你有这样一颗好奇心，如果你的确想知道人和动物为什么会表现出他们特有的行为，那么你就已经踏上了神奇的、令人兴奋并时常带有挑战的心理学研究方法之旅。

许多学生进入心理学领域是因为他们有志于改善人类的生活质量。但到底采取什么样的方法和干预手段才能惠及他人呢？以心理治疗为职业目标的学生就必须学会鉴别不良行为模式，学会区别无用的心理干预和有用的心理干预。心理学家通过建立理论和开展心理学研究来回答有关行为的问题，并借此获得改善人们生活的知识和洞见。

先思考一个许多心理学家都研究过的问题：媒体中的暴力造成了什么影响？半个多世纪以来，心理学家从多个方面对这一问题进行了数百项研究。《公众兴趣心理科学》（*Psychological Science in the Public Interest*）（Anderson et al., 2003）刊登了对此问题所做研究的一个回顾，这个期刊主要发表一些公众感兴趣的行为研究。该刊物最近还刊登了创伤后障碍的早期心理干预（McNally, Bryant, & Ehlers, 2000），高自尊对行为绩效、成就、幸福感和健康的作用（Baumeister, Campbell, Krueger, & Vohs, 2003），以及压抑的预防和治疗（Hollon, Thase, & Markowitz, 2002）等方面的研究。尽管课题各异，但是这些发表在《公众兴趣心理科学》和其他高水平心理学刊物上的研究都有一个共同特点：这些研究非常强调调用好的研究设计和研究方法来回有关行为的问题。

经过数十年的研究，关于媒体暴力对行为、情绪和社会的影响，心理学家有着怎样的看法呢？安德森等人（Anderson et al., 2003）在回顾了一些对电视、电影、游戏、网络 and 音乐中暴力的影响的研究后，总结出几条重要结论：

- 媒体暴力会使攻击性和暴力性思维、情感及行为在短时或长期发生的可能性增加。
- 关于媒体暴力的影响作用，在采用不同的研究方法、不同种类的媒体间以及不同的样本间表现出一致性。
- 最近所做的一些长期跟踪研究，将童年期频繁接触媒体暴力与他们成年后的攻击行为联系起来，这些攻击行为包括身体攻击和家庭暴力。

- 心理学家的理论认为, 媒体暴力会“激活”(启动)人们的攻击认知和躯体唤醒; 媒体暴力会促进人们对于攻击性行为的学习, 使人们对暴力行为更不敏感。一些研究证据支持了这个理论。
- 媒体暴力引起攻击行为还可能受其他因素的影响, 这些因素包括观看者的特征(例如, 年龄及其自身的攻击性程度)、社会环境(例如, 父母对媒体暴力的监控)和媒体内容(例如, 暴力描述的真实性和暴力出现的后果)。
- 没有人能对媒体暴力的影响免疫。

大量的研究表明, 青少年用于观看媒体节目的时间仅次于其睡眠时间 (Lyte & Hoffman, 1972)。因此, 这些研究结果隐含这样一个含义: 要想减少社会中攻击行为和暴力行为造成的破坏性影响, 措施之一就是减少媒体中的暴力成分。事实上, 心理学对暴力节目过滤芯片 (V-chip) 的研究与推广起到了重要作用 (Anderson et al., 2003), 该芯片可以有效屏蔽暴力内容。

关于媒体暴力的研究还有很多, 其中一个重要的问题是被动观看(电视节目)与主动参与(游戏、网络)这二者间的区别(见图 1.1)。主动参与视频游戏获得暴力信息是否对个体的影响更大? 有可能积极参与时所获信息会导致更加强烈的攻击行为。还有一些研究关注采取哪些步骤减少社会暴力影响以及限制媒体暴力播放以提高社会和谐。也许这些问题有一天将成为你的研究课题, 也许你对毒瘾或偏见的根源感兴趣。毫不夸张地说, 还有数以千计的重要研究课题需要有人做。只要你继续从事心理学研究, 总有一天你会为改善人类的现状做出贡献。

心理学家试图通过科学方法对有关行为、思想和感受的问题做出回答。科学方法 (scientific method) 是一个抽象的概念, 指的是提出问题的方式以及获得答案的逻辑和方法。科学方法的两个重要特征是实证方法和怀疑态度, 科学家据此来解释行为和心理过程。在本章中我们将论述这两个重要特征; 在第 2 章中我们将介绍科学方法的其他特点。



(a)

(b)

图 1.1 媒体暴力影响在 (a) 被动观看电视和 (b) 主动参与视频游戏中会有不同吗

科学的背景

- 科学至少存在于三种背景之中：历史背景、社会文化背景和道德背景。

尽管科学方法这个概念看似抽象，但心理科学却做了许多实实在在的事，在许多层面上对人类产生了重大影响。心理学家能在个人层面（例如，对攻击行为做出治疗干预）、家庭层面（例如，帮助父母控制子女使用媒体）和社会层面（例如，努力减少电视网络中的暴力内容）对人们产生影响。然而，心理学家必须以前精心设计和实施的研究为基础，才能使这些影响得以实现。

人类行为在很大程度上受其所处背景影响，科学活动也不例外。我们认为至少有三个背景影响到科学活动：即历史背景、社会文化背景和道德背景。我们将依次简要介绍这些内容。

历史背景

- 实证方法强调依靠直接观察和用实验回答问题，因其持续推动着科学心理学的发展而备受推崇。
- 计算机革命是使心理学研究的主旋律从行为主义转向认知科学的关键因素。

实际上，我们并不知道心理学成为独立学科的确切时间。心理学是逐渐形成的，它植根于亚里士多德的思想（Keller, 1937）、存在于笛卡儿和洛克等哲学家的论述中，后来，又出现在19世纪的生理学家和物理学家的研究工作中。公认的科学心理学的诞生时间是1879年，威廉·冯特（Wilhelm Wundt, 1832~1920）在德国莱比锡大学建立了第一个正式的心理实验室。

19世纪末，心理学家面临的一个问题是：心理学应该更紧密地联系物理学还是应该继续作为哲学的一个子学科（Sokal, 1992）。随着心理物理学（特别以古斯塔夫·西奥多·费希纳为代表）和象征神经系统传导的反应时法（以赫尔曼·冯·赫尔姆霍兹为代表）的发展，心理学家们相信他们最终能够测量思维本身（Coon, 1992）。基于这些有效的观察方法，心理学开始成为一门定量的实验科学。心理学家希望他们对心理的研究能够取得和物理学、化学、天文学等比较完善的学科一样的知名度。

20世纪之初，科学心理学的发展遇到了一个障碍，社会大众对唯灵论和灵媒现象产生了强烈兴趣（Coon, 1992）。一般的公众认为这些“精神类”的话题才属于心理学领域，并且试图为神秘主义、超人洞察力、心灵感应以及与死者交流等现象寻求科学的解释。然而，许多心理学家希望能将这门年轻的科学与伪科学话题分离开来。为了建立科学的心理学，心理学家采用实证主义方法作为理解人类行为的有效手段。实证方法*（empirical approach）强调用直接观察和实验的方法

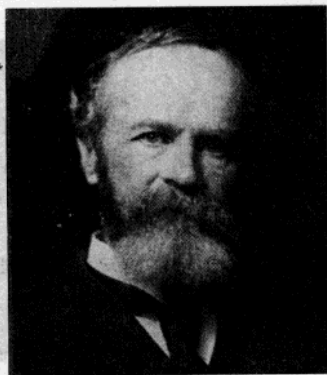
也可译作“经验方法”。在西方哲学传统中，经验主义（empiricism）是反对宗教迷信和形而上学的，但它同时又与理性主义（rationalism）相对立。近现代的逻辑实证主义（logical positivism）则是吸收了理性主义合理成分的经验主义新形式，故也称逻辑经验主义或科学经验主义。因此，在论述中常见“实证主义”和“经验主义”混用，英语语境和汉语语境均有。又因汉语现代语境中“经验主义”有时含有贬义成分，故大家更倾向于用“实证方法”一词。——译者注

来研究问题。这也许是科学方法最重要的特点。从这个角度出发,心理学家把能够直接观察的行为和经验作为主要的研究对象。

尽管心理学一直强调实证方法,但是心理学从一开始就不断进行着重要的转变。早期心理学家研究的主要兴趣是感觉和知觉,例如视错觉和想像。20世纪初,美国的心理学深受约翰·华生(John B. Watson, 1878~1958)的行为主义方法的影响,这一时期的心理学理论主要集中在学习方面,心理学家大多使用动物实验来验证他们的理论。对行为主义来讲,“心理”是一个“黑箱”,代表着刺激和反应之间的活动。行为主义在20世纪的前50年里是心理学的主流。然而,随着1967年乌尔里克·奈瑟(Ulric Neisser)《认知心理学》一书的出版,心理学的研究兴趣开始转向对思维过程的研究。认知心理学家也重新使用反应时这种早期心理实验方法来研究认知过程的规律。认知的观点现在仍然是心理学的主流,而且认知科学业已成为研究思维生理基础的神经科学的主要课题。科学心理学在21世纪必将有更大的发展。

认知心理学兴起的一个重要原因是计算机革命(Robins, Gosling, & Craik, 1999)。伴随着计算机的出现,行为主义的“黑箱”被表征为计算机模拟,心理学家谈论的是输入(刺激)与输出(反应)之间的信息加工、储存和提取。计算机为理解认知过程提供了有效的帮助。随着功能更强大的计算机的出现和应用,认知加工过程的研究视野得到进一步扩展,认知过程的测量也更加精确。目前,美国乃至世界各地的心理学实验室,对人类思维、情感和行为测量用的纸笔测验都已经被计算机技术所取代。同样,脑成像技术(功能性核磁共振成像)的不断发展将推动神经科学成为心理学和生物学领域的重要学科。

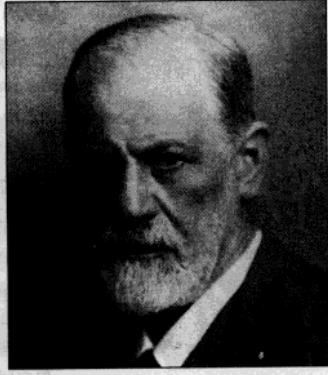
从行为主义到认知神经科学,这些心理学历史发展中的大趋势代表了20世纪心理学领域发生的主要变化。更进一步看,这种趋势还展现出科学心理学研究所涵盖的广泛领域和丰富内容。今天的心理学研究领域包括临床心理学、社会心理学、组织心理学、咨询心理学、生理心理学、认知心理学、教育心理学、发展心理学和健康心理学,所有这些领域的研究都将帮助我们理解人类行为和心理过程



(a)



(b)



(c)

图 1.2 许多有影响的人物推动了心理学的发展,包括 (a) 威廉·詹姆斯, (b) 斯金纳和 (c) 西格蒙德·弗洛伊德

专栏 1.1

心理学和诺贝尔奖

瑞典皇家科学院每年都要给在各个领域做出突出贡献的科学家颁发诺贝尔奖。2002年10月,丹尼尔·卡尼曼(Daniel Kahneman)博士成为第一个获得此殊荣的心理学家。他因直觉判断、理性思维以及不确定条件下的人类决策研究而闻名。他和他的长期合作伙伴阿莫斯·特韦尔斯基(Amos Tversky, 1937~1996)的研究因为对经济学理论产生影响而获奖(Kahneman, 2003)。卡尼曼和经济学家弗农·史密斯(Vernon Smith)共同分享了诺贝尔经济学奖,后者以发展经济学的实验室实验(本书的主要话题)而闻名。

还有一些获得诺贝尔奖的科学家虽然来自其他领域,但他们却是因为从事直接与人类行为有关的科学研究而获奖(Chemoff, 2002; Pickren, 2003),例如:

- 1904年,生理学医学奖:伊万·巴甫洛夫(Ivan Pavlov)因对消化问题的研究而获奖,此研究影响了他本人后来所做的经典条件反射研究。
- 1961年,生理学医学奖:物理学家乔治·冯·贝克西(Georg Von Békésy)因对心理声学——对声音的感知的研究而获奖。
- 1973年,生理学医学奖:三位生态学家,卡尔·冯·福雷契(Karl von Frisch),康拉德·洛伦兹(Konrad Lorenz)和尼古拉斯·汀伯根(Nikolaas Tinbergen)共同获得了第一个颁发给纯行为科学研究的诺贝尔奖(Pickren, 2003)。生态学是生物学的一个分支,研究者主要研究生物体与其所处环境之间的关系(参见本书第4章)。
- 1978年,经济学奖:赫伯特·西蒙(Herbert A. Simon)因其在组织决策方面的突破性研究而获得诺贝尔奖(Macloun, 2002; Pickren, 2003)。卡尼曼



在2002年获得诺贝尔奖时提到他引用西蒙的研究成果为自己的研究工具。

- 1981年,生理学医学诺贝尔奖颁发给了罗杰·斯佩里(Roger W. Sperry),一名动物学家。他通过“割裂脑”程序证实大脑两半球的不同作用。

这些科学家取得的成就证实了行为科学研究的广泛性及重要性。尽管没有“诺贝尔心理学奖”(同样也没有数学奖),这些科学家在多个领域的研究工作为我们对行为的理解做出了重要贡献。

的复杂性。

科学,尤其是心理学,会因天才个体的杰出思想而发展变化。伽利略、达尔文和爱因斯坦的思想不仅改变了科学家对其研究领域的看法,而且也改变了人们对自我和世界的理解。同样,很多杰出的个人也影响了心理学的发展进程(Haggbloom et al., 2002),其中包括一些诺贝尔奖获得者(见专栏1.1)。早期的美国心理学家威廉·詹姆斯(William James, 1842~1910)撰写了第一本心理学教科书《心理学原

理》，他采用内省的技术洞察心理过程。行为主义兴起之时，斯金纳（B.F. Skinner, 1904~1995）通过对行为的实验分析扩展了我们对反应和强化等概念的理解。与斯金纳一样，西格蒙德·弗洛伊德（Sigmund Freud, 1856~1939）也是心理学界最具有代表性的人物之一，但是这两人的观点和方法却截然不同！弗洛伊德关于人格、精神障碍和无意识的理论通过他倡导的自由联想法，戏剧般地把注意力从行为转移到心理过程上。还有许多人对心理学的特定领域产生了重要影响，包括发展、临床、社会和认知领域。我们真诚地希望你能多了解自己感兴趣领域中有影响的心理学家，无论是过去的还是现在的。

科学发展也是缓慢的，需要由许多个体的共同努力累积而成。描述这种渐变过程的一个方式是回顾心理学的专业化过程。美国心理学会（APA）成立于1892年，当时只有几十名会员；1992年APA庆祝成立100周年之际，它的会员已近7万人。成立于1988年的美国心理协会（APS）与APA一样关注心理学研究，APS更强调心理学的基础和科学问题。APA和APS每年召开会议，心理学家们可以从其中了解相关领域的最新进展。两个组织都面向会员和社会发行科学刊物，用以交流最新的研究发现。

你也可以成为创造心理学历史的一分子。APA和APS都鼓励学生入会，并为心理学本科生和研究生提供教育和研究的机会。要了解加入APA和APS成为正式会员或学生会会员的信息，可以访问它们的网站：

(APA) <http://www.apa.org>

(APS) <http://www.psychologicalscience.org>

APA和APS的网站均会登载近期的心理学研究发现、心理学出版物（包括为学生提供最低折扣的心理学刊物）等信息，另外网站还会提供许多心理学组织的链接。值得一看！

社会和文化背景

- 社会和文化背景影响到研究课题的选择、研究结果的接受及实施研究的地点。
- 当人们用适合本民族文化的标准和眼光去衡量其他文化时，会产生民族中心主义。

科学不仅受它所处历史背景的影响，还会受到社会和文化背景的影响。这种占主流的背景有时被称为时代精神（zeitgeist）。心理学研究及其应用和社会相互作用：研究对社会产生影响，社会反过来也影响研究。社会和文化背景会从三个方面对研究者产生影响：对研究课题的选择、支持研究的资源以及社会对研究结果的认同。比如，因为对妇女问题的关注（也因为从事研究工作的妇女数量的增加），研究者们提出一些新的研究课题：阻碍妇女在组织中晋升的“玻璃天花板效应”、双职工夫妇工作和家庭的相互影响、儿童看护质量对家长劳动生产率和儿童发展的影响，等等。社会和文化背景不仅影响研究者的研究内容，还影响其实施研究的方式。比如，社会对双语的态度会影响研究者的研究，社会的否定态度会

使研究者侧重于研究双语教育环境儿童成长中的问题,而反之则会使研究者侧重于研究接受双语教育的好处。

社会和文化价值观影响人们对心理学研究结果的态度。例如,关于性取向、童年期性虐待经历回忆以及媒体暴力等有争议的研究报告,通常会受到更多的媒体关注。有时,公众的兴趣越高,就越会引发对研究结果的解释及该结果对社会政策导向的争论,公众有时会做出过激的反应。发表在《心理学公报》(*Psychological Bulletin*)上的一篇关于儿童性侵犯的文章所引起的反应已经证明了这点(Rind, Tromovitch, & Bauserman, 1998)。作者在文章中回顾和分析了59个有关儿童性侵犯(CSA)问题的研究,林德等人(Rind et al.)总结这些研究结果时认为:“在忽略性别的大学生群体中的广泛调查表明,CSA经历并不会造成长期的伤害”(p.46)。这个研究结论被鼓吹恋童癖的网站大肆宣扬之后,“劳拉博士”(脱口秀节目主持人)将这篇文章理解为赞同成人与儿童发生性行为(其实这并不是研究者的意图),并批评美国心理学会在它的权威刊物上发表这篇文章(Dallam et al., 2003)。1999年,美国众议院一致通过对这篇研究报告的不信任决议来回应媒体的负面关注。当然,针对有争议研究的科学论战还在继续,各类批评和反驳被刊登在《心理学公报》和其他刊物及书籍上(Dallam et al., 2001; Ondersma et al., 2001; Rind, Tromovitch, & Bauserman, 2001)。虽然研究是基于可靠的实证科学,但公众对研究的批判仍然呈上升的趋势。因为固执地坚持个人信仰或是为了维护个人利益,反对者利用法律、行政和政治手段对研究结果进行攻击(Lof-tus, 2003),这些攻击不可避免地会妨碍正常与合理的科学质询和争论。

心理学家对诸如儿童性侵犯等社会问题的顾虑是致使心理学没有完全发展为实验室科学的一个原因。虽然实验室研究仍然是心理学研究的核心方法,但是,心理学家和其他的行为科学家们,还是在学校、诊所、企业、医院以及其他的非实验室环境中,包括在网络上,开展他们的研究。事实上,互联网已经成为了心理学家有效而普遍的研究工具(如Birnbaum, 2000)。美国的统计调查信息显示,截止2000年,5400万美国家庭(51%)至少拥有一台计算机;4000万家庭(42%)中至少有一人在家里使用互联网(Newburger, 2001)。这些数据明显低估了美国互联网用户的数量,因为这些数字是使用家庭用户而非个人用户来进行统计的,还没有考虑那些通过商务或教育设施接入网络的用户。而且,这些数字也没有反映出美国之外其他国家的互联网使用情况。据估计,到2005年,全世界的互联网用户已经从3亿人增加到了6亿人。这足以说明,用不了多长时间,行为科学家就会意识到存在一个巨大得惊人而又多样的“被试库”或“参与者库”(participand pool)*可以用于他们研究(Birnbaum, 2000; Gosling, Vazire, Srivastava, & John, 2004; Skitka, & Sargis, 2005)。由于20世纪90年代互联网和超文本语言的发展,心理学家很快就开始实施在线研究(如Musch & Reips, 2000)。网络可以进行任何以计算机为工具、以人为被试的心理学研究(Krantz & Dalal, 2000)。研究者招募被试的一种方法是在各研究网站发布有关信息。例如,

*美国心理学会(APA)规定,专指人类被试时可用“参与者”(participant)一词代替“被试”(subject)。subject可专指动物“被试”,也是人类参与者与动物被试的通称。——译者注

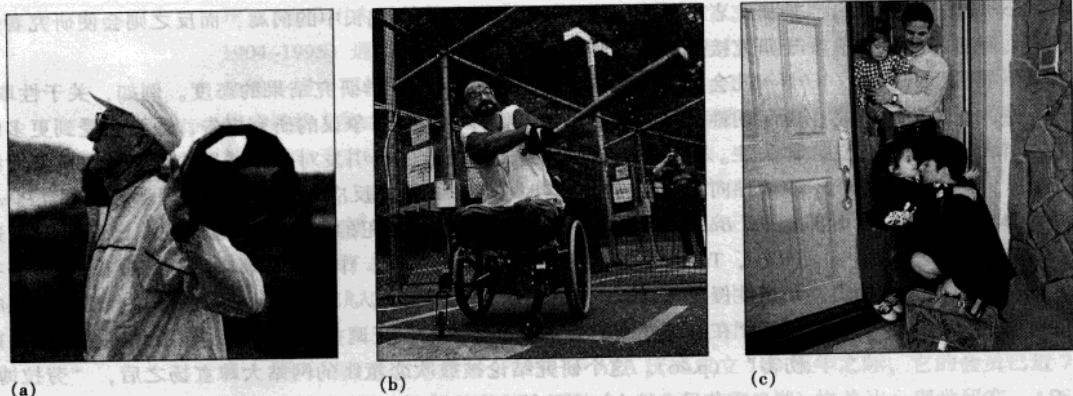


图 1.3 摘掉我们的文化眼镜，我们通过调查 (a) 强壮的老人，(b) 战胜残疾者和 (c) 居家父亲和职业妈妈获得了关于研究课题的新想法

APS 拥有一个以互联网用户为被试进行心理学研究的网站。要想成为研究被试，可以登陆这个网址：<http://psych.hanover.edu/APS/exponnet.html>。在我们后面章节中介绍某些心理学研究方法时，会对在线研究进行更多的介绍，其中关于这种研究形式引起的伦理学争论尤为重要（参见第 3 章）。

如果承认科学会受到社会和文化价值观的影响，那么就存在一个问题，即谁的文化正在产生影响——谁的文化应该产生影响。当我们试图通过本民族的文化思想或观念来理解不同民族文化中的个体行为时，产生了一个潜在的问题（见图 1.3）。这种问题的根源叫做民族中心主义（ethnocentrism）。关于民族中心主义的实例，你可以思考一下关于道德发展理论的争论。在科尔伯格（Kohlberg, 1981, 1984）提出的道德发展六阶段理论中，他认为在道德发展的最高阶段（后习俗发展），个体能根据自己定义的道德标准和对个人权利的理解做出道德判断。研究证据表明，科尔伯格的理论为美国和欧洲的男性的道德发展提供了良好的描述——他们的文化强调个人主义。在那些强调集体主义文化，如中国和巴布亚新几内亚，这一理论就不适用了，这些文化强调集体的价值。如果我们断言集体主义文化中的个体道德发展比较差，透过一个个人主义的标准和眼光来解释他们的行为，那我们就会犯民族中心主义的错误。跨文化研究能帮助我们避免只研究一种占统治地位的文化，并且提醒我们在研究中必须谨慎地使用不同文化背景下的标准。

道德背景

- 研究的道德背景要求研究者在伦理上保持最高标准。
- 《APA 伦理准则》可以指导研究，并帮助研究者评估伦理困境：如欺骗性策略的风险和收益，以及用动物做实验的问题。

科学是对真理的探求，科学家个人和团体必须确保科学活动符合伦理和道德的最高标准。欺诈、谎言和歪曲不应该出现在科学研究中。但科学也是一项人类

付出努力的工作,并且结果通常无法预料。科学家和科研机构以工作、金钱和名气为最终奖励来雇佣他人进行一场竞赛。大学教师发表的论文数量通常成为衡量其专业水平的依据,并对其职位的升迁产生影响。在这种环境中,学术上的不端行为令人遗憾地难以避免。

大量的不端行为是对科学诚实性的冒犯,这些不端行为包括捏造数据、学术剽窃、有选择地报告研究结果、否认真正对研究做出卓越贡献的人、滥用科研经费,以及不道德的对待人与动物被试(见 Adler, 1991)。有些违规行为比较容易被发现,如果数据完全是捏造,那么研究结果不能被再现(重复),或发表的报告中出现逻辑上的不一致。但是,还有更多狡猾的违规行为,如只报告符合实验目的的数据或误导报告结果,这些是很难被发现的。蓄意的不端行为和单纯的科学腐败间的界限有时不是很清晰。

为了教育研究者正当当地从事研究,帮助他们克服遇到的道德困境,大多数科学组织都制定了正式的伦理规范。在第3章中我们将介绍APA关于研究中涉及人和动物时应遵守的伦理准则。你将会看到,道德困境常常出现。我们以希思和戴维森(Heath & Davidson, 1988)的研究为例,他们让几组女大学生帮着准备一份预防强奸的小册子,并告诉她们这是“编写手册”的一部分工作。然而,他们让女大学生阅读手册的真正用意是以此分出控制组和非控制组,进而调查这些被试是否觉察到自己易受到强奸。研究结果显示,阅读了手册的非控制组女大学生报告了更多的焦虑,并且更注意采取防范强奸的措施。但是,研究者还发现,这些防范意识并非总能导致谨慎行为,因为有些女性可能认为对强奸的防范是无效的。该研究引发了一个重要问题:研究者应该在什么情况下蒙蔽参与者?从研究结果——妇女对于她们是否容易遭到强奸的知觉——中得到的信息比这一任务中的欺骗行为更重要吗?

欺骗仅仅是研究者要面对的道德争议之一。到目前为止,另一伦理争议是用来帮助理解人类精神病理学的动物被试。实验有时会将动物置于压力甚至痛苦条件下,而且有时要杀死这些动物进行尸体解剖。在什么情况下允许心理学研究使用动物被试?心理学研究会遇到很多伦理上的问题。因此,在职业生涯的早期阶段,你应该熟悉APA伦理准则以及它们的应用,只参加符合科学诚实性最高准则的研究(作为参与者、助手或主要研究人员)。我们希望通过研究方法的学习,你能学会做一些有益的研究,并能分辨哪些研究值得去做。

像研究者一样思考

- “像研究者一样思考”就是要以怀疑态度看待那些关于行为和心理过程的研究结论,即使这些结论是根据已“发表”的科学发现做出的。
- 对一个结论最有力的证据是多项研究得出一致性的结果,尽管科学家认为研究结论正确与否是有概率性的。
- 研究假说是对现象的试探性解释,它经常采用一种预言的形式以对可能的结果进行解释。

● 使用多种方法和测量来研究问题，是心理学家理解行为和心理的最好方法。

心理学专业的学生必须学会像研究者一样思考问题。更重要的是，科学家首先必须是怀疑论者，对行为和心理过程的因果关系持怀疑态度是科学心理学方法的一个重要特点。科学家对研究结论的态度是：不仅希望“眼见为实”，还希望一遍又一遍地证实，希望在他们选择的条件下能够证实。研究者最好从实验证据而非主观臆断中得出结论。对同一问题使用不同研究方法并得出相似结果是该结论最有力的证据。行为科学家之所以喜欢持一种怀疑态度，是因为他们知道行为是很复杂的，一种心理现象经常由多种因素交互作用产生，而找出所有的影响因素通常是一个艰巨的任务。有时做出的结论是不全面的，因为可能没有考虑到影响某一现象的所有因素，甚至有些影响因素根本没有被注意到。行为科学家还认识到科学是一种人类活动，而人会犯错误，人类的推理并不总是可信的。因此，科学家们对“新的发现”和不寻常的论断持怀疑态度，甚至对那些“已经发表”的研究也持怀疑态度。

怀疑精神使科学家比那些没有经过科学训练的人更谨慎。显然，普通人更容易接受那些从不充分和不可信的证据中得出的解释。对神秘学的广泛信仰已经证实了这一点。对于超自然的神秘现象，许多人倾向于毫无批判地接受，而非谨慎地思考。对公众看法的调查显示，美国绝大多数的人相信存在超感知觉（ESP），还有些人相信外星人到访过地球。有 2/5 的美国人相信占星术，1 200 万成年人在阅读了有关占星术的文章后改变了他们的行为（Miller, 1986）。尽管占星术正确性很低并且经常发现反面证据，但是这些信仰依然存在。

当然，科学家并没有想当然地认为神秘现象的非传统解释一定是错的。他们坚持只接受可以被检验的观点，拒绝那些一经提出就无法检验的主张。科学的怀疑精神能帮助公众避免遭到江湖医生、兜售“灵丹妙药”和偏方的人的欺骗，科学的怀疑精神还可以帮助公众鉴别“不劳而获”的思想以及对自然现象做出超自然解释的迷信。但在科学生涯中，信任与怀疑同样重要。科学家需要信任设备、被试、同行的研究报告和他们在实施研究时的专业判断。

要像研究者一样思考，你需要对证据和论断持怀疑的态度。如果你读过描述犯罪和侦探的书籍，或是看过一些法律类的流行电影和电视剧，那么你就应该对证据和推理有一定了解。侦探、律师和其他法律行业从业人员从大量的原始材料中搜集证据，并试图汇集证据以对人们的行为做出判断。少量证据也许足够怀疑某些人有罪，但是要想宣判其有罪，就需要从大量资料中搜集确凿证据。

心理学家的工作也差不多——他们搜集证据来解释行为和心理过程。分析以下的这些论断并思考每条论断中的隐含证据。

1. 在对美国成年人的一项调查中，96%的已婚人士报告说他们在过去的一年中忠实于自己的配偶。
2. 研究显示，青少年开始吸烟主要和同伴是否吸烟有关，而与父母是否吸烟关系不大。
3. 实验研究证明，从大学开始记录情感体验（与只记录鸡毛蒜皮事情相比）能使大学生取得更好的健康和学业成绩。

4. 研究者发现, 多次重复同一错误记忆实验, 大约在 3/4 的实验时间里, 被试在错误地报告实验中从未出现过的单词。

你可能已经注意到, 这些论断包含了一定范围的心理学研究课题, 从对婚姻的忠实度到错误记忆的形成。研究者用来收集证据的测量方法(例如调查、吸烟、健康和学业成绩、错误记忆)随他们所研究的领域不同而变化。通过对以上四个论断的思考, 你可能已经注意到, 它们在证据的范围和质量上有所不同。和法律专家需要不同种类的证据才能怀疑和定罪一样, 研究者也需要各种证据来对行为及其产生的原因做出解释。

如果你继续学习研究方法, 你将发现有一些重要的科学原则可以用来报告调查统计或行为观察的结果, 用来确定因素(或变量)之间的关系, 或者推断变量间的因果联系。本书着重讨论导致各种证据和结论的不同研究方法。最有力的科学证据类似于在审判中多方面搜集到可以定罪的证据。举例来说, 几个不同的实验, 采用相同的程序, 不同的被试, 但得出了相似的结果, 那么人们对“错误记忆”这一结论的信心就增加了。尽管研究者通过重复实验获得了极其有利的证据支持他们的结论, 但这种情况就如同陪审团根据相似的情景给某人定罪, 研究者和陪审团都要寻找真相, 但如果证据单一, 他们的结论最终只能是或然性的, 他们常常很难做出确定无疑的结论。

拓展练习

思考下面 4 个研究结论显示的证据类型, 并和后面列举的类型配对。你发现哪几项相匹配, 为什么? 原文出处已列出, 你也可以获得更多的相关信息。

研究结论

1. 在一项研究中模仿被试的行为和姿态, 那些被模仿的被试与没有被模仿的被试相比, 更多地表现出助人行为。
2. 一研究综述表明, 药物、人际关系心理治疗和认知行为疗法都对抑郁治疗有效。
3. 根据对 1996-2000 年间的交通死亡事故的分析, 估计 2001 年 9 月 11 日后的三个月中, 道路交通死亡事故增加 353 起, 这可能是由于人们害怕乘坐飞机出行而造成的。
4. 七八年级有体重压力的青春期的女孩, 在一年后问她们有关自己身材的问题时, 她们更多地对自己的身材表示不满意。

证据类型

- a. 跨研究的发现
- b. 对事件的观察
- c. 因果关系研究
- d. 两个变量间的关系

参考文献

1. van Baaren, R. B., Holland, R. W., Kawakami, K., & van Knippenberg, A. (2004). Mimicry and prosocial behavior. *Psychological Science*, 15, 71-74.
2. Hollon, S. D., Thase, M.E., & Markowitz, J. C. (2002). Treatment and prevention of depression. *Psychological Science in the Public Interest*, 3, 39-77.
3. Gigerenzer, G. (2004). Dread risk, September 11, and fatal traffic accidents. *Psychological Science*, 15, 286-287.
4. Stice, E., & Whitenton, K. (2002). Risk factors for body dissatisfaction in adolescent girls: A longitudinal investigation. *Developmental Psychology*, 38, 669-678.

当你思考上述四个研究论断的证据时，你可能已经对它们的结果做出了反应。也许你因较高的婚姻忠诚度而欢欣鼓舞，也许你不相信婚姻忠诚度有如此之高。如果是这样，则说明你已经真正开始像研究者一样思考了。和侦探一样，研究者靠直觉引导他们收集证据，陪审员和研究者也可能在评估证据时，发现前期构想和最初印象是错误的。但是，法律裁决和研究结论最终是根据证据而非我们的主观判断做出的。研究者需要使用科学方法约束自己的主观判断，并根据证据得出恰当的结论。

通过像研究者一样思考，你能够掌握两项重要的技能。第一项使你能够成为科学发现的有效用户（或消费者），能够做出更有见解的个人和专业决策；第二项技能将使你学会如何去做一项研究，如何为心理学做出更多的贡献。本书通篇重点阐述科学方法的这两方面内容，本章只对此做简要概述。我们首先要说明为什么在评价媒体报道的研究时像研究者一样思考非常重要，其后将描述当研究者想通过科学方法搜集证据时该如何开始。

评价媒体报道中的研究发现

心理学研究者在专业期刊的印刷版或电子版上发表他们的研究成果。大多数人都是通过媒体——网络、报纸、杂志、电台和电视上了解到心理学研究发现的，大部分的此类研究都是有价值的。心理学研究能在许多领域使人们受益，例如，帮助人们与患有早发性痴呆（阿尔茨海默病）的亲属交流、避免争吵，或者学会如何去原谅别人。但是，通过媒体来报道研究可能出现两个问题。第一个问题是媒体报道的并不总是好研究，批判性的读者需要区分研究的好坏——即哪些研究是可靠的，而哪些研究还没有被证实。读者还需要判断哪些研究结果可以应用到实际生活中，哪些研究结果还需要采取观望的态度。客观地说，各种媒体报道的心理学研究大多数都不尽如人意，所以我们有理由质疑从媒体中获得的相关报道。

第二个可能出现的问题是，媒体在报道科学研究时会有“翻译过程中的信息遗漏”。媒体只报道原始研究的概要，而关键的研究方法、结果或对研究的解释不会出现在媒体报道中。你对科学方法的学习越深入，就能越好地辨别媒体报道研究的质量和缺失的重要信息。下面，以具体媒体报告为例，让你感受一下解读媒体报道时应该想到的问题。

几年前，一个名为“莫扎特效应”的话题得到广泛宣传，“古典音乐有益于婴儿大脑发育”的大标题十分普遍。这些标题特别引起了初为父母者的注意，媒体报道说一些父母给他们的婴儿播放古典音乐，希望提高孩子的智力，100万新生儿母亲得到了一张免费的“聪明交响乐”CD和配套方案，显然，分发者和许多新科父母都相信“莫扎特效应”的真实性。

听音乐能提高新生儿的智力分数是一个很让人好奇的观念，如果你在媒体报道中看到类似的问题时，首先最好去直接查看这一研究的原始资料。在这个案例中，原始文章刊登在声望非常高的《自然》杂志上。劳舍尔、肖和凯（Rauscher, Shaw & Ky, 1993）描述了一个实验，实验中，大学生在进行空间测验前听10分钟莫扎特的音乐，静坐10分钟或听10分钟放松指导。听音乐组的大学生测验成

绩比其他两种条件下的都好,但是这种效应只能持续10到15分钟。

原始材料报告的效果可能是真实的,但对结果的推广却很不可靠。100万妇女受到鼓励给她们的婴儿播放“聪明交响乐”,而其理论仅仅建立在大學生身上检验到的且只能维持15分钟效果的方法上!虽然后来也做了一些针对儿童的研究,但是所有研究中模棱两可的结果都表明,从原始的研究报告到广为人知的“莫扎特效应”,一些信息在广为传播的过程中丢失了(由媒体造成的)。那些具有怀疑精神的人会在听到或读到媒体中的研究报道时产生疑问,而且那些有足够知识阅读原始研究材料的人也较少受到误导。你要做的就是要有怀疑精神,我们要做的是在本书中为你提供足够的知识,使你能够批判性地阅读那些研究的原始材料。

开始做研究

当你开始学习心理学研究者如何搜集证据时,我们会讲述专家们提出的关于研究的最基础问题——如何开始研究工作。开始研究之前他们通常会问三个问题,我们就以这三个问题来组织本节内容:

- 我要研究什么?
- 我如何建立一个研究中要检验的假设?
- 我研究的问题有价值吗?

在进行心理学研究之前,你需要做许多决定,第一个要做的决定当然是要研究什么课题。许多学生因为对精神病理学和心理健康感兴趣而加入这个领域;还有些人受人类认知之谜,如记忆、问题解决和决策等内容的吸引;还有人对社会心理问题和心理问题感兴趣。心理学为研究者提供了一套自助餐,这已成为成千上万发表心理学研究的科学刊物所证实。你可以快速浏览专门介绍心理学书籍的目录来找到许多心理学研究领域的信息。当然,还可以从你们的大学或学院的心理学系开设课程中找到更详细的信息,如变态心理学、认知心理学、社会心理学。

学生开展的第一个研究课题通常受到导师的影响,许多教授主持研究并且希望学生加入他的科研小组,你要做的仅仅是向他们询问一下。心理学系也会为学生们开展实验计划提供资源,其中的一个机会是通过“研讨会”的形式实现的。研讨会上,研究者向同事和学生介绍其理论和研究发现,有时研究者也来自其他大学。留意一下你们心理学院的通知吧。

第二个问题有一点困难,研究者开始研究时,必须首先提出他们的研究假设。假设是对现象的试探性解释,假设(hypothesis)通常是对一些结果的预测和对这些预测的解释。我们在前面提到过一个假设,玩电子游戏比被动观看电视受到媒体暴力的影响要大,因为游戏者主动去参与攻击性的活动,主动参与增加了他们的攻击性(另一种假设则是电子游戏对攻击性行为的作用比较小,因为玩家有机会在游戏中宣泄攻击性冲动,而被动观看电视者没有这样的机会)。

麦圭尔(McGuire, 1997)提出了49个简单规则(启发)用来产生科学检验假设。在此不能逐一呈现这49个规则,但我们可以列出一些例子来了解他的思想。他建议可以通过下列方式来提出假设:

— 思考一般趋势或原则的偏差（异常或例外）；

— 想像一下在一项任务中或面对特定问题时我们会如何反应；

— 考虑已经解决的与之类似的问题；

— 持续地、有准备地观察人或现象（例如，进行“个案研究”的时候）；

— 针对显而易见的结论提出反例；

— 借鉴其他学科的思想或理论。

不论你在何地、以何种方式开展研究，你都需要仔细查阅已经发表的心理学研究报告。在开始自己的研究前检索心理学文献有以下几个原因：一个很明显的原因是你要研究的问题可能已经得到解答，有人可能已经思考过这一问题，并且给出了答案或者至少给出了部分答案。你也很可能发现和你实验有关的研究结果。虽然你可能对你要研究的问题已经被解答而感到失望，但是考虑到有人对此想法已经进行了研究，更证明了你这个想法的重要性。不对已知文献进行仔细查阅就做研究可能很有趣或很好玩（当然也很简单），也许你可以称之为“消遣”，但是我们不能把这个叫做科学。科学是一项不断累积的事业——当前的研究是建立在以往研究基础之上的。

一旦你发现有文章和你的研究想法有关，那就应该仔细阅读，这个阅读过程会使你发现其中不一致或相矛盾的地方。你也可能发现研究结果受被试选择或实验条件的限制，或者发现有一个心理学理论需要验证。在做了以上的工作之后，你就已经找到了确凿的研究线索，一条可以继续的道路。

当阅读心理学文献并思考可能的研究问题时，你也需要思考如何运用这些心理学研究的结果去解决社会问题。在你学习如何做心理学研究时，可能已经想到了如何用这些知识使人类生活更美好。

查阅心理学文献已经不像过去那样困难，计算机辅助技术以及互联网的使用，已使文献查阅变成一个相当简单甚至充满乐趣的事情。在本书第14章中，我们会简要介绍如何查阅心理学文献，包括如何在你的研究中使用计算机数据库的方法。

最后，像斯滕伯格（Sternberg, 1997）指出的那样，选择一个研究问题绝对不能轻率。有些问题根本不值得去研究，因为它们的答案不会推动科学心理学的发展。可以说这样的问题是无意义的或者充其量是琐碎的。斯滕伯格（1997）建议新加入到心理学研究领域的学生在确定好的研究课题前应该思考以下问题：

— 为什么这个问题可能是有科学价值的？

— 这个问题的范围是什么？

— 如果我开展了这项研究，结果可能是什么？

— 解答了这一问题可以在多大程度上使心理学得到发展？

— 为什么会有人对这个问题的研究结果感兴趣？

在你的研究开始时，可能需要研究顾问和成功从事过研究的其他人帮助你找到这些问题的答案。我们也希望你通过对心理学理论和研究的深入学习，通过在本书中读到的有趣而富有深意的例子提高自己回答这些问题的能力。

当然，确定研究课题并不能告诉你如何进行研究。你真正想知道的是什么呢？

表 1.1 研究进程的步骤

步 骤	方 法	章 节
提出研究问题	• 对民族中心主义有所觉悟和防范。	1
	• 获得个人的研究经验。	1
	• 查阅心理学文献。	1, 14
产生研究假设	• 阅读与你的课题相关的心理学理论。	1, 2
	• 回顾个人经验；思考先前研究中的例外和不一致。	1
形成操作定义	• 看一看在先前的研究中其他人是如何定义相同或相似的概念的。	
	• 确定研究变量。	2
选择研究设计	• 确定被试样本。	4, 5
	• 决定你的研究是寻求描述、预测，还是确定因果关系。	2
	> 为描述和预测选择明显和相关的设计。	4, 5, 6
	> 为因果问题选择实验设计。	7, 8, 9
	> 处理小样本或个体推断时使用单因素设计。	10
	> 对实验控制不严格的因果研究问题采用准实验设计。	11
评估研究的道德性	• 明确研究的潜在风险和收益，保护被试的权利。	3
	• 服从道德观察委员会的建议。	3
	• 获得官方的许可。	3, 11
收集和分析数据；得出结论	• 了解数据的意义。	12
	• 概括数据。	12
	• 确认数据揭示的含义。	13
	• 在心理学会议上报告研究结果。	14
报告研究结果	• 向心理学期刊提交研究报告。	14

要回答这一问题意味着你必须做出我们将在本书中讲述的其他决策。作为一个研究者，你要问自己这样几个问题：我是要做定性的研究还是定量的研究？我想观察的变量的性质是什么？我如何对行为进行准确、有效的测量？什么研究方法最适合我的研究课题？需要使用哪种统计分析方法？我选择的方法符合道德和伦理标准吗？这些问题以及其他的方法和步骤见表 1.1。不必因为对这些问题和表中的项目不熟悉而担心，当你开始通过本书学习心理学研究方法时，你将会学到这些研究步骤。如果你开始着手自己的研究，表 1.1 将是一个有用的指导。

本书向你介绍心理学家使用科学方法的方式。要知道，心理学是一个有许多研究领域和难题的学科，单一的研究方法不能够解释心理学家关于行为和心理过程的所有问题。因此，解答我们的问题最好采取多种方法（multimethod approach）——即运用多种研究方法和不同的行为测量方法来解决一个问题。本书的目标是帮助你装满实施研究用的“工具箱”。在本书中你将学到，任何一种方法或行为的测量都可能存在缺陷或不能完整地回答研究问题。当采用多种方法进行研究时，任何一种特殊研究方法的缺陷都可用其他方法以“填补裂缝”的方式克服。因此，多种方法的一个优势是研究者能够借此更完整地理解行为和心理过程。我们的希望是，通过使用这些工具——本书中描述的研究方法——你将踏上通往解决研究问题的心理学征程。

小 结

心理学家通过建立理论和进行心理研究来寻求对行为和心理过程的理解。心理学研究对个人和社会具有重要的影响作用，其中的一个例子就是：研究表明媒体暴力会产生负面影响。研究者运用科学方法，即强调实证的方法来理解行为，这一方法依赖直接观察和实验来解答问题。科学实践存在于历史、社会文化和道德背景之中。从历史角度来说，计算机革命成为促进心理学从行为主义转向认知心理学的工具。不论是过去还是现在，许多心理学家都为不同领域的心理学发展做出了贡献。

社会文化背景对心理学研究的影响表现为两方面：研究者的选题和社会对研究结果的认同。当存在民族中心主义时，文化也会对研究产生影响。在这种偏见下，人们试图以自身所处的文化准则或观点来理解其他文化背景中的人们。道德背景要求研究者遵守伦理行为的最高准则。对科学诚实性的违背包括伪造数据、学术剽窃、否认真正对研究做出卓越贡献的个人、滥用科研经费以及不道德地对待人类和动物被试。美国心理学会的伦理规范指导研究，并帮助研究者评估研究中欺骗行为所带来的风险和收益，以及使用动物被试的伦理困境问题。

研究者必须持怀疑的态度对待有关行为和心理过程的结论。虽然心理学家认为所有的研究结果都只是可能性的而不是确定性的，但研究结果最有力的证据是多个研究得出相似的证据。媒体在报道研究时会带来两个问题：或者研究不符合规范，或者只报告典型的结果。评估媒体报道的一个重要步骤是追溯到原始材料，并且更深入地了解研究的方法和程序。

开始研究的第一步是提出研究问题，学生们从教材、课程以及与教师的讨论中获得研究想法。下一步是做出研究假设，研究假设是对将要检验的现象的推测性的解释，经常是以预测和对预测结果进行解释的形式出现。虽然研究假设可以有多种产生方式，但是这一步骤中最重要的是查阅相关的心理学文献。最后，评价所研究问题是否有利于心理学家对行为和心理过程的理解，这也是很重要的。

多种方法的意思是使用多种研究和测量方法来解答研究问题，以此获得对行为和心理过程更加完整的理解。科学家认识到任何一种研究方法或行为测量工具都有缺陷或不完善之处；多种方法允许研究者去“填补”特殊的方法留下的“裂缝”。本书的目的就是介绍心理学家使用的各种研究方法。

重要概念

科学方法
假 设
实证方法

多种方法
民族中心主义

复习思考题

1. 描述科学方法的两个重要特征。
2. 为什么早期心理学家使用实证法作为心理学研究的首选方法?
3. 指出计算机在 20 世纪影响心理学发展的两个重要方式。
4. 举出一个 (1) 社会文化因素影响心理学研究课题的选择以及 (2) 社会文化因素影响着对研究课题接纳的例子。
5. 描述民族中心主义如何成为研究的障碍, 并且提出一种使研究者不犯此错误的方法。
6. 研究要在“道德背景”下进行是什么意思?
7. 描述研究者在进行研究时可能面对的两个道德困境。
8. 解释研究者为何要对研究结果持怀疑态度, 并解释他们与一般大众有何区别?
9. 确定研究者在做出有关行为和心理过程论断时必须借助哪些类型的证据。
10. 列举两个原因说明为何要批判性评价新闻媒体中(如电视、杂志)的报道。
11. 研究者在开始一项研究时最初的三个步骤是什么?
12. 列举两个原因说明开始研究时检索心理学文献为何如此重要。
13. 描述多种方法做研究的概念并指出其主要优点。

挑战性问题

1. 思考下面这个假设: 玩电子游戏比被动观看电视更易使人具有攻击性。
 - A. 你如何检验这个假设? 也就是说, 你通过什么方法来比较两种情境下的不同体验?
 - B. 你如何判断暴露于暴力之后人们出现了攻击行为?
 - C. 你要增加什么因素来保证暴力是引发攻击行为主要的原因?
2. 你已经知道获得知识有很多种方法。例如, 阅读文艺作品时, 人们通过作者和他/她塑造的人物来看待事物。这种方法与心理学家做研究的方法有何不同? 每种方法的优势和缺点又是什么?
3. 纵观心理学研究历史, 我们见证了心理学研究从强调感知觉到行为主义再到认知心理学的转变。在心理学的不同分支领域(如, 临床、发展、神经科学和社会心理学), 研究课题的数量以惊人的速度在增长。
 - A. 你最感兴趣的心理学领域是什么, 为什么?
 - B. 在你们的图书馆里阅读 3~4 种属于你感兴趣领域的学术期刊(如《发展心理学》、《咨询与临床心理学报》、《人格与社会心理学报》)。(向你的老师或图书馆的管理人员咨询其他刊物的名称。)研究者调查些什么课题? 你能察觉到的这些课题的研究趋势或正在进行的研究的类型吗? 描述你的发现。
4. 指出民族中心主义如何影响研究类型, 并在下述研究项目和被试选择时起到哪些作用。
 - A. 男人对女人
 - B. 多数民族对少数民族
 - C. 保守政治对开明政治
 - D. 18~25 岁、35~45 岁、55~65 岁、75~85 岁的被试

拓展练习的答案

1c, 2a, 3b, 4d

最令人信服的研究证据出现在研究结论 2 中, 因为它总结了包括药物、人际关系心理治

疗和认知行为疗法的一些研究。跨研究的证据一致性——在这个案例中，就是对抑郁的治疗——使我们对结果非常自信。第二个令人信服的证据是在因果关系研究中（我们将在第2章以及后续的章节中加讨论）。结论1是关于如果（仅在如果的前提下）被试的行为被模仿，他们就会表现得乐于助人，即模仿使人们变得乐于助人。令人信服程度再次之的证据是两个变量有关系，如结论4所述，（1）青春期女孩想变苗条的压力和（2）她们对自己身材不满意的程度这两个变量中。要知道描述这一关系可以说因受到变苗条的压力引起女孩们的不满，也可以说因为对身材不满才使她们产生了变瘦的压力。最后，最低可信度的证据即对事件的观察。在结论3中，对2001年10月到12月间的交通死亡事故的分析并不能说明为什么死亡事故增加了，但是可以提出猜测：这个数据也许是由于人们对911恐怖袭击事件的反应造成的。

挑战性问题的答案

- A. 检验这一假设的一种方法是选取两组被试。一组玩暴力视频游戏，另一组观看暴力电视节目。另一种方法是让同一组被试在不同时间段分别尝试两种实验条件。
- B. 要判断人们是否会在玩视频游戏或观看电视节目之后表现出更多的攻击性，我们需要测量攻击行为。潜在的测量方法是无限的，有限的只是研究者的创造力。比较合理的是使用其他调查者使用过的量表，你还可以将你的结果与以往的结果进行比较。对攻击性的测量包括让人们报告他们如何对待一些愤怒情境，或者观察他们在暴力条件后如何对待实验者（或其他人）。在第二种情况下，研究者需要一份表格或其他方法记录被试的暴力（或非暴力）行为。需要注意的是攻击性可以定义为好多种，包括身体行为，口头表达甚至想法（但是要测量后者的困难性）。
- C. 确保两组被试——电视和视频游戏——除了电视和视频游戏这两个实验条件外，其他条件都相似是很重要的。例如，确保你的研究有两组被试：一组看电视，另一组玩视频游戏。当然，你还要保证你的研究结果证实玩视频游戏的被试比看电视的被试更有攻击性。

可能存在的一个问题是玩视频游戏的被试本身比看电视的被试更具攻击性，如果是这样，那就不能判断在实验中观察到的攻击性差异是因暴力引起的，还是因被试特点变量引起的。因此，你还需要在实验之前确定两组被试的攻击性特点相似。本书的后面部分会教你如何保证两组被试间的相似性。

你可能还希望保证被试的其他方面也相似，例如，保证每组被试在暴力情境下的时间相等。此外，还要保证电视中的暴力和游戏中的暴力水平相当。还有被试在其他的因素上应该无差别，如是否有其他人在场，实验的时间等。为了证明玩视频游戏比看电视更能增加（或减少）攻击性，最重要的一点即两组被试惟一的差别就是分别处于不同的暴力条件下。

第2章

科学方法

本章大纲

科学方法和日常方法

一般方法

观察

报告

概念

仪器

测量

假设

科学方法的目标

描述

预测

解释

应用

科学理论的建构和检验

小 结

科学方法和日常方法

- 科学方法以实证为基础，要求系统的、控制的观察。
- 科学家在实施一项实验时对实验对象的控制最为严格，在实验中，研究者操纵自变量以确定它们对行为的影响。
- 因变量是对行为的测量，用来评估自变量的效果。
- 科学报告是客观、无偏的；使用操作性定义可以促进概念的准确交流。
- 科学仪器准确且精细，物理测量和心理测量应该可信且有效。
- 假设是指对现象试探性的解释，可检验假设有非常明确的概念（操作性定义），这些概念均为非循环定义的概念，并且均可被观察。

100 多年以来，科学方法一直都是心理学研究的基础，科学方法不是一定需要特殊的设备，也不是非要求特殊的程序和技术不可。正如第 1 章所描述的，科学方法包括科学的提问方式、获得答案的逻辑性和方法。已经有许多行之有效的方法可以获得有关我们自身和所处环境的知识，比如哲学、宗教研究、文学、艺术和其他学科。科学方法与很多其他方法是有区别的，但它们都有一个共同的目标——寻求真理。科学方法是寻求真理的一种途径，而我们理解科学方法的最佳办法就是将它与我们理解问题的“日常”方式相区别。就像放大镜和显微镜能辅助视力一样，科学方法能提高我们日常生活中的认识能力。

科学方法和日常方法的差别见表 2.1。表 2.1 使我们对科学家的思考方式和人们日常的思考方式之间的差别一目了然。非正式的、随便的方式通常就是我们日常思考的方式，接下来我们将对科学方法和日常方法的差别进行简要概述。

一般方法

第 1 章中提到，要想像研究者一样思考，就必须持有怀疑的态度。心理学家谨慎小心地接受有关行为和心理过程的论断，并且在接受论断之前对证据进行批判性的评估。然而按照日常的思考方式，人们在接受证据或论断时是很少对证据

表 2.1 科学方法和非科学（日常）方法的特征和区别^{*}

	非科学的（日常的）	科学的
一般方法：	直觉的	实证的
态度：	非批判性的，认可的	批判性的，持怀疑态度的
观察：	随意的，非控制的	系统的，控制的
报告：	有偏差的，主观的	无偏差的，客观的
概念：	模糊的，带有冗余含意	定义清晰，操作具体
仪器：	不精确的，不准确的	精确的，准确的
测量：	不可信的，无效的	可信的，有效的
假设：	不可检验的	可检验的

^{*} 部分参考马克思（Marx，1963）的区分方法。

进行评估甚至不评估的。一般情况下,我们更多地凭直觉作出日常判断,这意味着我们的行为以“感觉正确”或“似乎合理”为基础。直觉不是以正式决策过程为基础。我们通过直觉获得的很多日常推断和结论是个人领悟的结果,是在个人经验基础上迅速将其知觉为正确的结果。尽管在某些时候,如几乎没有任何其他信息时,直觉是有价值的,但直觉并不总是正确。例如,直觉认为应该对视频游戏、电影和电视节目中的暴力和性相关内容进行评估,直觉认为这类评估是防止接触暴力内容的有效工具,而实际上却出现了截然相反的结果。研究表明,这些评估诱使青少年观众去观看有暴力和性内容的节目——布什曼和坎托(Bushman & Cantor, 2003)称之为“禁果效应”。评估并没有起到限制接触暴力和性内容节目的作用,反而会增加这类接触,因为“评估成了找到这些内容的捷径”(p.138)。

我们直觉的事物并不总是与实际正确的事物相一致,因为我们无法意识到我们的知觉会因认知偏差而歪曲,或者我们忽视了对获得的证据进行适当的衡量(Kahneman & Tversky, 1973; Tversky & Kahneman, 1974)。2002年,丹尼尔·卡尼曼因研究认知偏差如何影响人们的经济选择而获诺贝尔奖。有一种认知偏差叫做错觉相关,即当什么关系都不存在的时候,我们倾向于知觉事物之间的关系。萨斯坎德(Susskind, 2003)认为,儿童在判断男人和女人行为时更容易受到这种偏差的影响。向2年级和4年级的儿童呈现许多男人和女人的常规行为图片(例如一位正在编织的女人)、反常规图片(例如一位正在编织的男人)和中性行为图片(例如一位正在读书的男人或者女人),之后让他们估计每幅图片出现的频率。结果表明儿童高估他们看到的常规行为图片出现的次数,该结果说明儿童容易受到错觉相关的影响。也就是说,儿童对男人和女人常规行为的预期,导致他们相信这些类型的图片比实际呈现的图片数量要多。对错觉相关偏差的一种可能解释是,我们更可能注意与我们信念相一致的事件,而不是与我们信念相冲突的事件。

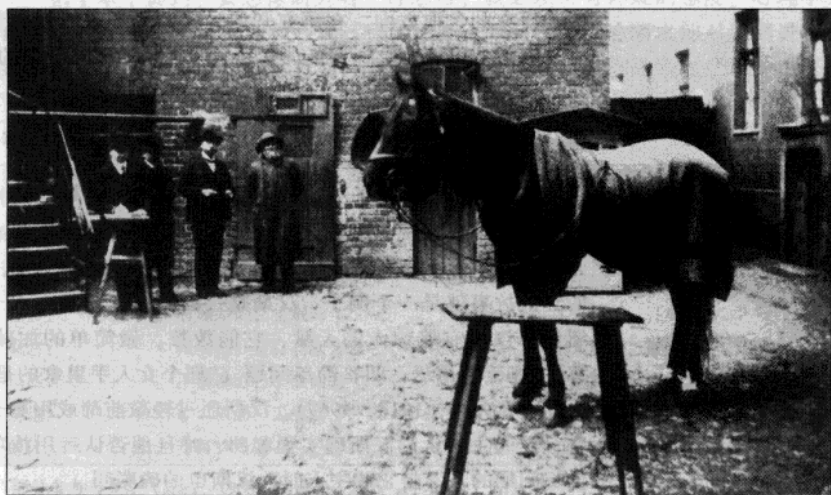
科学方法是以实证而非直觉为依据的。实证研究强调直接观察和实验,但这并不意味着直觉在科学中不起作用。任何一位科学家都可能会讲述因直觉暗示而获得实验结果的故事。然而,同样一位科学家也可能得到很多与直觉相反的结果。关于研究的选取方向,在研究开始的时候科学家可能会受直觉的引导。然而最终科学家则会积极地受到直接观察和实验提供的实验证据的引导。

观 察

我们可以通过对他人活动的简单观察来了解他们的行为。然而,日常观察却做不到仔细或系统,多数人在观察的时候没有尝试控制或者排除可能影响事件的因素。结果,我们经常基于随意观察而得出不准确的结论。例如,“聪明汉斯”的经典故事就是一个例子。汉斯是一匹马,据它的主人(一位德国数学教师)介绍说,汉斯有着惊人的天赋,它能数数,做简单的加减运算(甚至包括分数运算),能读懂德语,回答简单问题(“那个女人手里拿的是什么?”),知道日期以及时间(Watson, 1914, 1967)。汉斯通过轻敲前蹄或用鼻子指向选择项目来回答问题。它的主人认为汉斯确实很聪明,并且他否认运用技巧引导汉斯。实际上,即使提问题的人不是它的主人时,汉斯也的确聪明。

报纸报道了汉斯的表现，数以百计的人前来观看这匹惊人的马（见图 2.1）。1904 年还专门成立了科学委员会，旨在弄清汉斯能力的本质。令其主人失望的是，科学家观察到汉斯在两种情况下不再聪明。第一，如果提问者不知道问题的答案，那么汉斯也不能回答问题。第二，如果汉斯看不到提问者，那么它也不再那么聪明。科学家们观察到了什么？他们发现，汉斯的回答其实是对提问者的轻微移动做出的反应。提问者一个轻微的向前弯曲的动作会使汉斯轻敲前蹄，提问者任何向上或者向后的移动都会使汉斯停止轻敲。委员会证实，提问者无意的动作是汉斯做出回答的线索。这似乎说明，与那些观察汉斯的人相比，汉斯是一个更好的观察者。

图 2.1 上图：聪明的汉斯在
很多观看者面前表演。下图：
汉斯在不能看到提问者的控
制条件下接受测验



这个著名的“聪明汉斯”的报道证明，科学观察（不像随意观察）是系统的和有控制的。确实，这一案例已经表明控制（control）是区分科学与非科学必不可少的成分（Boring, 1954; Max, 1963）。以“聪明汉斯”为例，研究者每操纵一个条件实现一次控制。例如，提问者是否知道问题的答案，汉斯是否能看见提问者（见图2.1）。通过有控制的观察，科学家对产生现象的原因有了更清晰的认识。对“聪明汉斯”仔细且系统的观察是科学家利用控制获得对行为理解的一个例子。

科学家在做实验时需要极其严格的控制。在实验（experiment）中，科学家操纵一个或者多个因素，并观察这种操纵对行为的影响。研究者为了确定什么因素影响行为而控制或者操纵的因素被称作自变量（independent variable）。¹在最简单的研究中，自变量有两个水平：处理和不处理。处理的条件一般称为实验条件，不处理的条件一般称为控制条件。例如，如果想要研究饮用含酒精饮料对加工复杂信息的速度和准确性的影响，那么自变量就是饮料中是否含有酒精。实验条件下的被试喝含有酒精的饮料，而控制条件下的被试喝同样的饮料，但是不含酒精。操纵这个自变量之后，研究者要求被试玩一个复杂的视频游戏，用以观察他们能否加工复杂的信息。

用来评估自变量对行为的影响作用而测量的指标被称作因变量（dependent variable）。在我们以酒精对加工复杂信息的影响研究为例，研究者会测量控制组被试和实验组被试在玩较难的视频游戏时的出错次数，而错误次数就是因变量。

科学家试图确定他们观察到的任何因变量的差异是否都是由自变量的不同条件引起，这里的不同条件即例子中的有酒精还是无酒精。然而，为了得出清晰的结论，科学家还必须运用适当的控制技术。“聪明汉斯”的故事就表明了科学家是如何通过仔细的和有控制的观察来排除对一个现象的可能解释（例如，“聪明汉斯”是真的聪明）。本书的每章都会强调研究者是如何运用控制技术来研究行为和心理过程的。

报 告

假如你错过了一节课，你想通过他人了解这节课，你很想准确地知道课上发生了什么。或者，你的两个朋友在你错过的聚会上争执不休，你很想通过其他人了解究竟发生了什么事情。可以想像得到，个人偏见和主观印象经常会掺杂在我们收到的日常报告中。当你让其他人描述一件事情时，你所获悉的细节（不总是正确的）往往带有他们的个人印象。你也会发现，对同一件事情，你报告的细节和别人向你报告的细节不同。我们经常按照自己的兴趣和态度报告事情，很明显，这些兴趣和态度并不总是与其他人一致的。在下次的班级考试后，调查几个同班同学对这次考试的印象。他们的报告可能完全不同，这主要依赖于下面的因素：

¹ 有时，自变量水平是由研究者选择的而不是操纵的。个体差异变量是人与人之间的性格与特质的不同。例如，被试的行为是否会因性别（男，女）不同而出现差异。当研究者研究被试的性别差异时，他们选择男性和女性作为个体差异变量来检验这个因素。我们会在第7章中看到操纵和选择自变量之间存在的重要差别。

拓展训练

在这个练习中，看完研究报告的简短描述后，你需要回答几个问题。

研究报告的题目是：“如果电视节目血腥暴力，对广告的记忆就会减退”（Bushman & Phillips, 2001）。研究者认为，一半以上的电视节目是暴力的，并且数以百计的研究已经表明，暴力电视节目会增加暴力行为。最后，他们指出广告商至今还在继续赞助暴力节目。

研究者将他们的研究集中于对 12 个实验的分析，这 12 个实验检验电视节目类型对人们记忆电视节目中插播的商业广告的影响。在典型的实验研究中，向观看暴力节目和观看非暴力节目的被试呈现相同的广告。用

回忆和再认测验测量被试对广告商标名字和商业广告细节的记忆。主要结果是，观看暴力电视节目的被试对商业广告的记忆要差于观看非暴力电视节目的被试。对于这些结果，研究者的解释是，暴力节目使人们愤怒，而愤怒干扰了记忆。研究者还提出，他们的结果表明“赞助暴力节目并不是广告商明智的投资策略”（p43）。

1. 找出这个研究的自变量（及其水平）和因变量。
2. 研究结论认为，暴力电视节目会减少对广告内容的记忆。你能从另外一个角度，即非暴力电视节目会提高对商业广告的记忆，来解释得到的结果吗？

图 2.2 你怎样描述这幅画面



他们考前准备的如何，他们学习时注意力集中于哪些方面，以及他们对讲师将要强调的考试内容的预期。

科学家报告其研究结果的时候，试图将观察到的结果和基于这些观察得出的推断区分开。例如，仔细观看图 2.2 的照片，你是怎样描述你看到的每一个人的？

描述这个场景的一种方式，两个人跟着他们前面的另外一个人沿着小路跑步。你也可以这样描述这个场景，三个人正在赛跑。如果你用第二种描述，你报告的是从你看到的场景中得出的推断，而不仅仅是你所观察到的结果。科学报告往往会趋向于三个人一起跑步的描述。

报告中的描述和推断之间的区别可能会很极端。例如，将图 2.2 描述为跑步也可以被认为是推断，实际观察到的是三个人大步地以某种速度移动他们的腿。这样缺乏想像力的描述也是不适合的。关键是，在科学报告中，观察者必须提高警觉，避免迅速做出推断的倾向。另外，应该充分详细地描述事件，但不包括琐碎的不必要的细节。观察和报告观察的恰当方法将在第 4 章讨论。

科学报告要尽量无偏和客观，检查一个报告是否无偏的一个可接受的方法是，报告是否被一个以上的独立观察者证实。例如，测量观察者一致性在观察研究中很普遍。不幸的是，很多偏差是细微的，未必总会被检测到，甚至在科学报告中也未必能检测得到。有一种鱼，它们的卵在公鱼嘴中孵化，直至小鱼被孵化出来。如果首先观察到鱼卵消失在其父亲嘴中的科学家立即就假定它们的父亲正在吃它们，这可以原谅，因为我们通常认为嘴是用来吃食的。但是，仔细的观察者会耐心等待，密切注意预期外的结果，并且不会想当然。

概 念

我们所说的概念 (concepts) 涉及事物 (有生命的和无生命的)、事件 (活动中的事物) 以及事物或事件之间的联系，还有它们的特征 (Marx, 1963)。“狗”是一个概念，“吠”和“驯服”也分别是一个概念。概念是我们进行交流的符号，清晰的、不模糊的思想交流需要概念有明确的定义。

日常交流中，我们不必为如何定义一个概念而担忧。比如说，很多词是生活中经常使用的，并且很容易理解。但是人们在交流过程中并不知道这些词的确切涵义，也就是说，人们频繁地交流，但却并不完全明白他们正在讨论着什么！这听起来似乎很荒谬，但是为了证明我们的观点，请阅读下面的内容。

如果你问人们，智力主要是通过遗传获得的还是通过后天学习获得的，你可能会饶有兴趣地和那些与你持有相反观点的人辩论。讨论智力的起源之后，再问问他们“智力”是什么意思。你可能会发现，即使刚讨论完智力的起源，多数人还是很难定义这个概念。然而人们通常愿意争论有关智力的重要观点，甚至明确地支持某一个论点，但是却不能说究竟什么是“智力”。当一个人提出一个定义时，他可能与另外一个人给出的定义不完全一样。也就是说，“智力”对于一个人意味着一件事，而对于另外一个人意味着另外一件事。为了回答智力主要是遗传获得的还是后天习得的问题，我们必须有一个准确的且所有参与辩论的人都能接受的定义。

“概念”的学习在心理科学中是如此重要，以至于研究者用一个特殊的名字来提及概念：构念。一个构念 (construct) 就是一个概念或是一种思想，心理构念的例子包括智力、抑郁、攻击性和记忆等等。科学家给构念赋予意义的一种方式是在操作定义。操作定义 (operational definition, 也译作“操作性定义”) 即仅根据可

图 2.3 如果在鼻子上平衡球是智力的操作定义的话，那么海豹岂不是比人类更聪明



观察的程序来解释概念，这个观察程序是用来产生和测量概念的。例如，智力可以用一种纸笔测验来进行操作定义，这种纸笔测验强调对逻辑关系的理解、短时记忆以及对词义的熟悉性。有些概念可能与智力的操作定义不一样，但是，一旦某个特殊测验被定义，至少不会再引起这样的争论：根据定义，智力到底是什么。操作定义促进了交流，至少在知道如何且为什么应用操作定义的人们之间增进了交流。

尽管通过操作定义传达了准确的意思，但是这种交流方法仍然受到了批判。其中一个问题已经很明显了，那就是如果我们不喜欢智力的某种操作定义，那我们必然会给出智力的另外一种操作定义。这是不是就意味着有多少种操作定义就有多少种智力？每当给智力纸笔测验增加一系列新问题时，我们是否需要重新定义智力？很不幸的是，我们真的不知道答案。为了确定一个不同的程序是否会产生一个新的智力定义，我们必须寻找大量的证据。例如，在一个测验上得高分的人在第二个测验中的得分是否也高？如果是，那么新测验可能与旧测验测量了相同的构念。

对使用操作定义的另一个批评是，定义并不总是有意义的。例如，如果根据在鼻尖上平衡球的时间来定义智力的话，那么多数人会发现这个操作定义并无意义。我们怎样确定构念的定义是否有意义呢？解决的办法仍旧是求助于其他证据。用鼻尖平衡球的任务与其他被普遍接受的智力测量任务相比有何不同？我们更愿意在此借助常识，人们通常会认为平衡球的能力是智力的证据吗？科学家也意识到了操作定义的局限，然而，应用操作定义的一个主要的优点是它们有助于科学家们交流他们的构念，这个优点比缺点更重要。

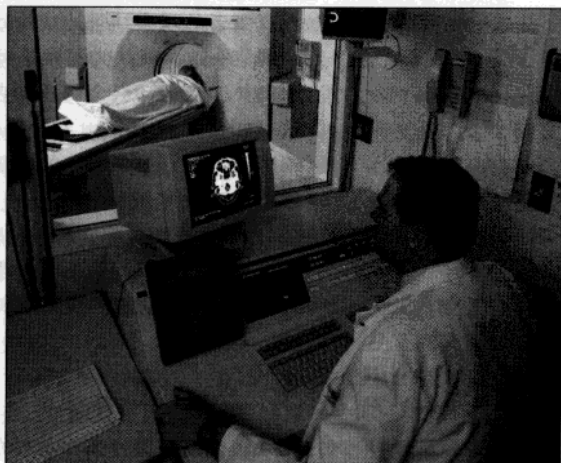
仪 器

你经常通过仪器来做测量，其广泛性远远超过了你所意识到的程度。汽车的速度计、卧室内的钟和用来测量体温的温度计都是我们离不开的仪器。如果这些仪器不准确，就会出现问題。准确性涉及仪器的显示与实际准确值之间的差异。不准确的钟会使我们迟到，不准确的速度计会使我们收到交通罚款通知单，不准确的体温计会导致我们在生病的时候相信自己没病。通过校准刻度或者与已知准确的另外一个仪器核对的方式可以确定仪器的准确性。速度计的准确性可以通过观察路旁距离标识和用准确的秒表测量花费的时间来进行核对。

测量存在不同的精确性水平。以0.1秒为单位的时间测量没有以0.01秒为单位的時間測量精確。轿车上的汽油表是一种不能精确测量的仪器。尽管汽油表的测量相当准确，但是它不能给出非常精确的读数，虽然很多人曾经指望汽油表能帮助他们确定是否有足够的油到达下一个服务站。

我们也需要仪器来测量行为。1879年，冯特建立了第一个正式的心理实验室，这标志着科学心理学的诞生。冯特用反应时仪器来测量认知过程的时间。过去的100年来，这种仪器的精确性和准确性得到了极大改进。如今，以毫秒（千分之一秒）为单位的电子计时器提供了精确的反应时测量。当代的心理学研究还采用了很多其他仪器。生理心理学实验（例如评估一个人的唤醒水平）需要能精确测量内部状态的仪器，如测量心率和血压的仪器。对焦虑测验有时要使用测量皮肤生物电反应（GSR）的仪器。其他的行为测量方法是各种各样的纸笔测验，问卷和测验是心理学家最常用的行为测量工具，量表也是观察者使用的测量工具。例如，7点等级量表对儿童攻击性等级的测量相对准确（尽管可能不太精确），范围从没有攻击性（1）到非常有攻击性（7）。行为科学家的责任就是运用尽可能准确和精确的仪器。

图2.4 心理学采用的科学仪器大大增加了测量的精确性和准确性



测量

为了研究事件和现象，科学家用仪器来进行测量，测量提供了以科学方法为特征的仔细和可控制的观察记录。物理测量是科学测量的一种类型，它包括一致的量度标准和实施测量的仪器，例如，长度是一个能用物理测量来评级的量度，并且有一致的长度单位（例如英寸、米）。同样，重量单位和时间单位都属于物理测量。

尽管心理学研究者也使用物理测量，但是多数心理研究中的测量不涉及物理测量。不存在测量诸如漂亮、攻击性或者智力之类心理构念的尺子。对于这些维度，我们必须用第二种测量类型——心理测量。在某种意义上，人类观察者就是心理测量的工具。更具体地说，几个观察者之间的一致性心理测量的基础。如果几位独立观察者对某一特定行为的攻击性用7点等级量表评级的得分都是3，那么我们可以说它就是攻击性行为的心理测量。

仪器应该准确和尽可能的精确，同样重要的是，测量也应有效且可信。一般来说，效度（validity）是指测量的“真实性”。构念的有效测量是测量了人们想要测量的东西。在提及智力的可能操作定义时，我们就讨论过测量的这个问题。前面已经提到，智力可根据在鼻尖平衡球这个任务的表现来定义。根据“操作性”原则，这是完全允许的。然而，多数人会质疑这种平衡行为是否真的能测量智力。换句话说，我们质疑这是否是对智力的有效测量。智力真的可以通过将球保持在鼻尖上的时间长短来测量吗？正如前面所说的，与定义有效性有关的证据必须来自于多方面。测量的有效性可以通过人们在使用不同的独立测量方法针对某一构念测量时是否能获得预想的相同结果来确定，如果做得同样好就认为测量有效。如果平衡球花费的时间是对智力的有效测量，那么，一个人在平衡任务上做得好，他也应该在词汇量测验、推理能力测验和其他已获得认可的智力测验上表现不错。

测量的信度（reliability）是指测量的一致性。下面区分几种不同类型的信度。说到仪器信度时，我们要讨论仪器的工作是否具有有一致性。车有的时候能启动，有的时候不能启动，这时我们认为引擎不十分可信。如果两个或者两个以上的独立观察者的观察结果达成一致的话——也就是说如果一个观察者与另一个观察者的观察一致，就认为观察是可信的。当几位科学家要求学生1992年夏天巴塞罗那奥林匹克运动会金牌获得者的“幸福感”进行评级时，他们发现评级者高度一致（Medvec, Madey, & Gilovich, 1995）。他们还发现，铜牌获得者（第三名）比银牌获得者（第二名）的幸福感强，这有点违反直觉。这个结果可以用“反事实思维”理论来解释，很显然，得到奖牌的人（例如获得奖牌）比失去奖牌的人（例如没有得到金牌）的幸福感强。

测量的效度和信度是心理研究的重要问题。在接下来我们向你介绍不同的研究方法时，你会遇到很多用来确定信度和效度的方法。

假设

假设是对某件事试探性的解释。假设通常试图回答“怎么？”和“为什么？”

的问题。在简单水平上,假设仅表示怎样把特定变量联系起来。例如,在流行文化中,我们经常将白的或者明亮的联想为“好的”,或将黑的或者黑暗的联想为“坏的”(Meier, Robinson, & Clore, 2004)。例如,在影片《星球大战》中,天行者卢克和莱娅公主穿白色服装,入侵者达斯·瓦德穿纯黑色服装。对很多宗教(例如佛教、基督教、印度教、索罗亚斯德教)来说,明亮与上帝或善良之间存在联系,黑暗与恶魔或邪恶之间存在联系。一件事的好与不好涉及到情感判断,换句话说,我们体验的明亮(黑暗)是一种感知。在迈耶和他的同事们的工作中,他们假设情感判断和对明亮感知之间的联系是自动的——即人们自动地判断较明亮的客体为好的,而较暗的客体为不好的。

为了检验他们的假设,迈耶和他的同事们(Meier et al., 2004)做了一系列实验,他们要求参与者判断呈现在计算机上的100个词是消极的还是积极的。50个是先前被评为积极情感的词(例如甜、爱、漂亮、熟睡),50个为消极情感词(例如苦、癌症、魔鬼、粗鲁)。研究者操纵的变量是,词以白色字体呈现还是以黑色字体呈现的。结果表明,当词的情感与字体颜色相冲突时(例如,爱用黑色呈现),相比词的情感与字体颜色相匹配时(例如,爱用白色呈现),参与者做出判断花费的时间要长,并且错误更多。

在复杂的理论水平上,假设可以提供特定变量联系起来的原因(“为什么”)。例如,迈耶和他的同事(2004)认为,人脑进化的结果使概念思考与物理知觉自动连接起来,例如情感判断(Barsalu, 1999; Laoff & Johnson, 1999)。基于这些理论,研究者认为,如果没有首先自动地考虑词的物理特征,例如明亮,人们不能判断一个词(或任何一个客体)的情感。在他们的实验中,当明亮与正确情感判断相冲突时,人们需要额外的加工(例如时间、思考)才能消除他们的自动联想,对词的褒贬做出正确的判断。

几乎每个人都曾经提出过解释人类行为的假设。为什么人们要做明显无意义的暴力行为呢?为什么人们要吸烟?为什么一些学生的学业成绩要比其他学生优异?区分随意的日常假设与科学假设的一个指标是可检验性。如果一种假设不可检验,那么它对于科学来说就是无用的(Marx, 1963)。有三种类型的假设不能通过“可检验性检验”。一种是假设的构念未被恰当地定义;第二种是循环假设;第三种是假设受个人观念的影响,而没有得到科学的承认。

如果假设提及的概念没有被恰当地定义,那么此假设就不可检验。一种假设说,一个人想要刺杀美国总统或其他著名人物,是因为他精神失常。这个假设不可检验,除非对“精神失常”的定义达成一致。遗憾的是,心理学家和精神病学家对专业名词的理解并非总能达成一致,比如“精神失常”,因为这些概念没有可接受的操作定义。操作定义除了有助于促进交流,还提供了一种方法来评估假设是否包含科学的、可接受的概念。

循环假设不可检验。当一个事件被用来解释事件本身时,就产生了循环假设(Kimble, 1989, p.495)。例如,“一个8岁男孩上课分心,并且阅读困难,因为他有注意缺陷障碍。”因为注意缺陷障碍本身就是用无法集中注意来定义的,所以这个陈述只不过是说男孩无法注意是因为他无法注意——即一个循环假设。

如果是不被科学承认的观念和信仰,假设也不可检验。科学是可观察的、可

证明的和实证的。假定人们做出极令人讨厌的暴力行为是来自于魔鬼的指令，这是不能被检验的，因为它援引的原理不在科学范围内。这样的假设可能对哲学家或者宗教研究家来说有价值，但对科学家来说没有意义。

科学方法的目标

- 科学方法试图实现四个目标：描述、预测、解释和应用。

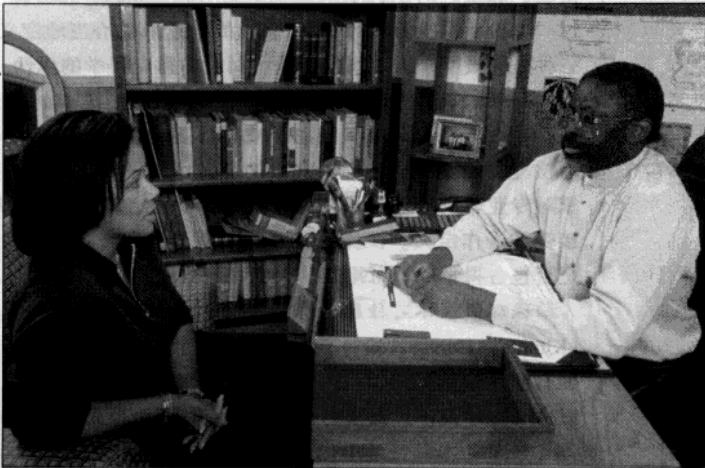
本章第一部分对日常思考方式与科学方法的不同点进行了分析。一般来说，科学方法以实证研究、系统和控制的观察、无偏和客观的报告、操作定义的构念、精确的仪器、有效且可靠的测量以及可检验的假设为特征。本章的下一部分将对科学方法的目标进行分析，心理学家们用科学方法来实现四个研究目标：描述、预测、解释和应用。

描 述

- 心理学家试图描述事件和变量之间的关系；很多时候研究者用一般规律研究法和定量分析法。

描述指的是研究者用来定义、区分、分类或者把事件以及它们之间的关系归类。例如，临床研究为职业医生提供了区分精神障碍的标准。这些基本上都可以在美国精神病学协会的《精神障碍诊断和统计手册》中找到（4th, Text Revision, 2000），也称作 DSM-IV-TR（见图 2.5）。用来定义分裂性神游症（原心因性神游症）的标准就是一个例子。

图 2.5 我们经常采用美国精神病学联合会制定的《精神障碍诊断和统计手册》作为精神障碍的诊断标准



分裂性神游症诊断标准

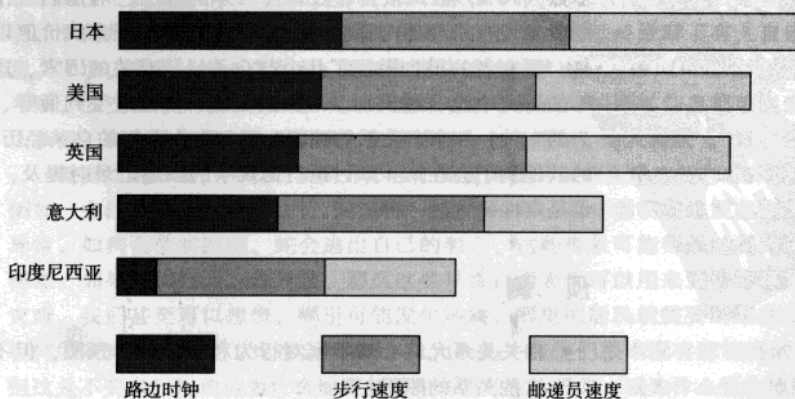
- A. 突然的、措手不及的离家出走或者离开平时的工作岗位，并且不能回忆过去。
- B. 对自己的身份感到迷惑或者毫无根据地设想一个新身份（部分或者全部）。
- C. 此症状不一定绝对发生在“人格分裂”情形下，并且也不是由于物质（例如药物滥用、药剂）或者一般生理损伤（例如，颞叶癫痫）所产生的心理影响。
- D. 症状引起严重的压力或者社会功能、工作能力以及其他重要领域的功能受损（DSM—IV—TR, 2000, p.526）。

用来定义分裂性神游症的诊断标准为这种障碍提供了操作定义。如同很多其他不同寻常的精神障碍一样，分裂性神游症相对罕见，因此，我们总是基于对人们表现出的症状来描述这类障碍，这些描述被称作“个案研究”。研究者们试图为临床医生提供精神障碍的常见症状描述，以及各种症状的出现与性别和年龄等其他变量之间的关系。例如分裂性神游症，根据 DSM-IV-TR (2000) 的描述，主要发生在成年阶段，尽管现在相对罕见，但是在“极端压力事件期间，如战争和自然灾害”的情况下发生较频繁 (p.524)。

科学，特别是对现象进行特定描述的心理学会用到常规规则研究法。用常规规则研究法 (nomothetic approach)，心理学家试图建立适用于多样性总体的广泛概括和一般规律。为了实现这个目标，心理学研究经常需要大量的被试。研究者们试图去描述一组人的平均水平或者有代表性的表现。这个平均水平可能描述了组内个体的表现，也可能没有描述组内个体的表现。

例如，通过记录户外路边时钟的准确性和测定步行者步行 100 英尺的速度，莱文 (Levine, 1990) 描述了不同文化背景和不同国家人民的“生活节奏”。这个研究的结果呈现在图 2.6 中。总的说来，日本人的生活步行速度最快，美国人位居第二，印度尼西亚人的生活节奏最慢。然而，并不是所有的日本人和美国人生活节奏都快。实际上，莱文 (Levine, 1990) 和他的同事发现，美国的不同城市间的生活节奏存在很大差异，这取决于国家的不同区域。美国东北地区城市（例如，波士顿、纽约）的生活节奏比美国西海岸城市快（例如，萨克拉门托、洛杉矶）。

图 2.6 测量每个国家路边时钟的准确性、步行者的速度以及邮递员完成一项任务的速度来描述一个国家的生活节奏。在图中，较长的条块代表时钟更准确或者步行速度更快或完成任务的速度更快



当然，城市里的市民也存在个体差异，并不是所有的洛杉矶市民的步行速度都慢，也并不是所有的纽约市民的步行速度都快。然而，日本人的平均步行速度比印度尼西亚人快，西海岸美国人的平均步行速度比东北部慢。

运用常规法则研究法的研究者承认个体之间存在较大差异，然而，他们试图强调相似性而非差异性。例如，一个人的个性有别于其他人，但他的心脏与其他人一样，同样位于胸腔左上方。同理，描述一个人的行为如何受强化方式（例如奖励、惩罚）影响时，我们不能否认那个人的个性作用。研究者描述生物一般性的特征都会以一组不同生物体的平均表现为基础。

有些心理学家，特别是戈登·奥尔波特（Gordon Allport, 1961）认为常规法则研究法是不恰当的——他认为独一无二的个体不能用平均值来描述。使用个人特质研究法（idiographic approach）的研究者的研究对象是个体而不是群体。这些研究者认为，虽然个体的行为方式遵循一般规律或规则，但是也必须描述个体的独特性。个人特质研究法研究的主要形式是个案研究法，我们将在第10章介绍。

研究者根据他们所要研究的问题来决定描述一组个体还是一个个体的行为。尽管很多研究者在研究中主要用常规法则研究法或者个人特质研究法，但是有的研究者会同时使用两种方法。例如，一位临床心理学家可能会主要用个人特质研究法研究治疗中的几位患者，而用常规法则研究法来研究大学生群体。研究者必须要做的另外一个决定是使用定量研究还是使用定性研究。定量研究涉及的研究结果主要来自统计汇总和分析。定性研究是用语言总结研究结果，而不是统计汇总或者分析。正如心理学研究用常规法则研究法要多于个人特质研究法一样，定量研究也比定性研究用得更多。

社会学家和人类学家广泛运用定性研究法（例如，见 Seale, 1999）。定性研究的数据最主要来自于访谈和观察，可以用来描述个体、群体和社会运动（Strauss & Corbin, 1990）。定性研究通常是“自然发生的、自然情境中的平常事件”（Miles & Huberman, 1994, p.10）。对于定性研究来说，重要的是，研究者要求被试用他们自己的方式来描述他们的经历，而不是要求被试使用理论或先前研究已经建立的范畴和维度（Kidd, 2002）。基德和克拉尔（Kidd & Kral, 2002）用定性研究法研究了29个多伦多街头青年（年龄17~24岁）的经历。访谈中绝大多数（76%）被试报告有过试图自杀的经历，对他们所作的叙述进行分析后，研究者发现自杀经历与孤独感、拒绝和背叛、低自我价值以及卖淫有关系。重要的是，研究者的报告揭示了几个与自杀经历有关的因素，这些因素在以前对街头青年的研究中没有被识别。“失去控制、卖淫时受到侮辱、药物滥用（慢性自杀行为的一种）和亲密关系的破裂”都与街头青年的自杀经历有关（p.411）。定性研究的其他例子会在第4章讨论行为观察的叙述记录时提及；第10章描述的个案研究也是定性研究的一种形式。

预 测

- 相关关系允许心理学家对行为和事件做出预测，但不允许心理学家推断引起这些关系的原因。

预测是科学方法的第二个目标，对事件及事件间关系的描述是预测的基础。心理学中有许多重要的问题都要求做出预测，例如，单亲家庭的孩子是否更容易抑郁？攻击性过强的儿童是否与成年人一样存在情绪问题？生活压力事件是否会导致躯体疾病的增加？研究结果将对诸如此类的问题做出肯定的回答，这些回答不仅为心理学增添了有价值的学科知识，而且有助于预防和治疗情绪障碍。

依据先前各类标准测验成绩对之后的成绩（如工作中的、学校里的、或特殊领域中的）进行预测，这是许多心理学家的一项重要工作。例如，美国研究生入学考试 GRE 成绩和平均学分积点 GPA，都可以用来预测学生在研究生阶段的表现。斯滕伯格和威廉姆斯（Sternberg & Williams, 1997）发现，GRE 能相当好地预测研究生第一学年的成绩水平，但却不能预测其他重要指标，如导师对学生创造能力、教学能力和科研能力等的评价。GRE 被广泛地认为是学生职业发展的预报器（例如，《美国心理学家》“评论”部分，1998，53，p.566-577），但上述研究未出预料地引起了关于 GRE 的预测效度（预测的准确性）的争论。

当一个变量的分数能用来预测另一个变量的分数时，可以认为两个变量间存在相关。当相同的人、事或物的两个不同的测量指标一起改变时，我们就说存在相关（correlation）。也就是说，当一个变量的一系列分数趋向于和另一个变量的一系列分数相关联时，此时的两列分数称为“共变”。例如，已知压力和疾病相关，人们经历的生活压力事件越多，就越可能产生躯体疾病。

回想一下曾在班级里进行的教师和课程评估，你可能经历过很多次这样的评估。在大学某门课程即将结束时，通常会要求大学生对授课教师及授课内容进行评估。课程结束时，你可能对授课教师形成一定印象（例如，这个教师是否乐于助人、是否热情、是否讨人喜欢），毕竟，你在这个教师的课堂上学习了 12 或 14 周（或者说 30 多个小时）。安巴比和罗森塔尔（Ambaby & Rosenthal, 1993）对没有参加课程学习所做的教师评估和在学习课程后进行的期末教师评估的相关程度进行了研究。他们给一组在校女大学生播放一段录像（无声），有趣的是，播放时间仅为 30 秒、10 秒或 6 秒（分别进行了几个研究）。结果发现，依据“简短的非语言行为片段”做出的教师评估和学生在学期末所做的评估存在较高相关。也就是说，在课堂上受到更多积极评价的教师，其在录像中的行为评分也更高；反之，在课堂上受到更多消极评价的教师，其在录像中的行为评分也更低。这些结果表明，只简短表现出人（这里指教师）的非语言行为，就已经暴露了有关自己的信息。同理，我们（作为观察者）能够非常快速地对情感行为做出相对准确的判断。当然，安巴比和罗森塔尔的发现并不意味着教学评估中的所有信息都可以这种方法获得，因为他们只是将判断聚焦在情感行为上（例如，讨人喜欢）。

需要指出的是，有效的预测并不在于知道为什么两列变量之间存在某种关系。例如，中国人曾依靠对动物行为的观察来预测地震。某些动物在地震前行为明显异常，如狗会乱吠跑圈，蛇会逃出自己的洞穴，这些现象可能都是地震的有效预测器。如果动物行为表现异常，那么这些异常行为表现可以用来预告即将发生的灾难。我们甚至可以想像，哪里可能发生地震，哪里的居民就需要对某些动物进行持续观察（就像矿工观察金丝雀），从而预测一些他们根本没有意识到的情况。但这并不要求我们理解为什么地震前某些动物的行为异常，或为什么发生地震。

有趣的是,莱文(Levine, 1990)发现,市民的步行速度可以用来预测死于心脏病的概率。但是,我们仅仅只能推测为什么这些测量之间相关。莱文对上述相关的解释是,生活节奏快的人有很多不利于健康的行为,如吸烟和不良饮食习惯,从而增加了患上心脏病的风险(Levine, 1990)。安巴比和罗森塔尔(1993)对由课堂外的学生和课堂内的学生对教师进行评估的相关性做出的解释是,人们能够“闻声即知地”快速获取一个人的某些特征信息,因为这些信息在现实生活的决策中具有重要的适应价值。但是,如果没有额外的信息,对这两种现象的这些解释则只能是推测性的。

解 释

- 心理学家理解一个现象的原因需要满足三个条件:共变、时序关系,以及排除其他貌似真实的可能原因。
- 实验方法就是研究者操纵自变量以确定其对因变量的影响,研究者可借此建立时间序列关系并清晰地确定共变趋势。
- 如果研究中没有混淆因素,那么就可以排除其他貌似真实的可能原因。
- 研究者尝试对研究结果进行推广,以使其研究发现可以描述不同的总体、情境和条件。

尽管描述和预测是科学研究的重要目标,但它们只是理解和阐释现象的第一步。解释是科学方法的第三个目标。只有找出现象产生的原因,才能理解和解释这种现象。研究者通常用实验的方法来找出现象的原因。实验研究不同于描述和预测(相关)研究之处在于,研究者实验研究中可以做到对环境的严格控制。研究者控制某种情境时,他们一次操纵一个自变量,以确定自变量对因变量即他们所感兴趣现象的影响。通过操纵控制严格的实验,心理学家推断出产生现象的原因,得出结论。实验对心理学家做出因果推论来说尤为重要,所以本书的第7、8和9章将详细讲解实验方法。

科学家设置了三个重要条件来做出因果推论(causal inference):事件的共变、时序关系,以及排除其他貌似真实的可能原因。举个简单的例子来说明这三个条件:假如你的头撞在了门上并感觉头痛,你可能会推断是因为头被撞了所以才引起头痛。因果推论的第一个条件是事件的共变,即如果一个事件是产生另一个事件的原因,那么这两个事件必定一起发生变化。也就是说,一个事件发生了变化,另一个事件也必定发生变化。在我们所举的例子中,这种共变表现为,头撞在门上与产生头痛的感觉都发生了。

因果推论的第二个条件是时序关系(也称附带事件),即推测原因(撞头)必须发生在假定结果(头痛)之前。如果头痛发生在撞头之前,就不能推断出是撞头引起了头痛。换句话说,对于之前的撞头事件,头痛只是偶然性事件。最后,只有当其他可能导致结果的原因被排除了——当貌似真实的可能原因被排除时,对原因的解释才能被认可。在所举的例子中,这意味着在做出撞头导致头痛的因果推论时,还必须考虑和排除其他可能引发头痛的原因(例如,阅读深奥难懂的书

籍等)。

不幸的是,人们存在一种不良的倾向,这种倾向表现为只要符合因果推论的第一个条件,人们就会得出结论,而不去考虑因果推论的其他两个条件。例如,一贯严厉并采用体罚的父母比不太严厉和不体罚的父母更可能教养出具有攻击性的儿童,父母的惩罚和儿童的攻击性存在明显的共变。父母对儿童行为的影响可能使我们想到:时间序列条件已经出现了——父母采用体罚的行为和儿童出现攻击性的结果。然而,同样是这个例子,婴儿活跃性和攻击性之间也存在差异,而且婴儿的行为对父母的教育方式有着极大的影响。换句话说,一些儿童的攻击性可能是天生的,这种天生的攻击性导致了父母的严厉惩罚,而不是父母的严厉惩罚导致了儿童的攻击性。因此,因果关系的方向可能和我们首先想到的正好相反。

只有证明了共变,事件的原因才能被确定,认识到这点很重要。科学方法的第一个目标:描述,适合在单一情境下描述事件。然而,理解研究对象比描述研究对象的要求更高。例如,如果一个教师想证明“主动学习方法”(如,讨论、小组表演)有助于学生学习,那么他应该先教会学生使用主动学习方法,然后描述接受这种特殊指导的学生的学习成绩。但是,为了证明主动学习方法能提高学生的成绩,她还应该了解些什么呢?也许另一组使用其他学习方法的学生也可以达到和主动学习相同的学习效果。所以,在这个教师提出主动学习会引起学生成绩提高之前,她应该对这种学习方法和其他一些学习方法进行比较。也就是说,她应该找到采用主动学习方法的学生和未采用主动学习方法的学生在学习成绩上的差异,这样的比较结果才能说明教学方法和成绩之间的共变。实验控制实现后,伴随着自变量和因变量共变关系的产生,时间序列关系也产生了。因为研究者先操纵自变量(例如,教学方法)并随后测量因变量(例如,学生的一次学习测验),所以因果推论的时间序列条件也就出现了。

研究者做出因果推论时必须排除其他可能的解释和原因,这是因果推论过程中最具挑战性的问题。回想之前提到的评估两种教学方法(例如,主动和被动学习)效果的研究,假如研究者将所有男生分配到一种教学方法中,把所有女生分配到另一种教学方法中,那么此时得出的两组成绩的差异,既不能归因于教法的差异也不能归因于性别差异。因为研究者不能分辨两组学生成绩的差异是自变量(主动或被动学习)产生的结果还是学生的性别导致的结果。说得正式些,即教法自变量和性别自变量相混淆了。如果两个潜在自变量同时发生变化,就会产生混淆(confounding)。如果研究被混淆了,就不可能判断成绩差异是由哪个变量引起的。

研究者通过实验来探究现象产生的原因。然而,即使实验控制严格,研究者可以根据实验结果形成因果推论,也还存在其他问题。其中一个重要的问题就是因果结论的推广,即实验结果只适用于参加实验的人群,但研究者经常要将实验结果推广至没有参加实验的人群。

心理学研究中的大多数被试是心理学专业的大学生,难道心理学家的研究结论只适用于大一大二的学生?同理,实验室研究对条件的控制比自然环境中多得多。因此,判断实验室研究结果能否推广到“现实世界”中去是科学家的一个重要任务。一些人机械地认为实验室研究没有用,或者认为实验室研究和现实世界

的生活没有联系。然而,通过学习本书中的研究方法不难发现,那些认为实验室研究和现实世界之间没有联系的观点是没有意义的,或者说是令人不满意的。相反,心理学家应该认识到:实验室研究得出的结论有助于解释现象,而且这一知识适用于研究和干预现实世界中的问题。

应用

- 在应用研究中,心理学家运用他们的知识和研究方法改善人们的生活;心理学家进行基础研究是为了获得有关行为和心理过程的知识,以及对理论进行检验。

心理学研究的第四个目标是应用。虽然心理学家乐于对行为和心理过程进行描述、预测和解释,但这种知识不能只停留在理论层面上。相反,这些知识应该用于现实生活之中,帮助人们摆脱心理障碍的困扰、暴力和攻击性行为的迫害以及陈规和偏见对人们生活的影响。虽然现实世界中的各种问题有时看来是无穷无尽的,但我们不该因此丧失信心。心理学家研究问题和结论的广泛性为研究者帮助人们适应生活的方方面面、改变个人生活提供了多种途径。作为 21 世纪的心理学家,我们从“积极的心理学”的角度获得最大化的成长和潜能实现,可能发现更多积极和创造性的方法 (Seligman & Csikszentmihalyi, 2000)。

关于创造性变化的研究通常叫做“应用研究”。在应用研究 (applied research) 中,心理学家实施研究是为了让人们的生活更加美好。例如,对那些饱受心理障碍之苦的人来说,心理治疗技术方面的研究可以改善他们的生活。因此,应用心理学涉及不同类型的干预,包括以学校学生、各类工作人员以及社区个体的生活改善为目标的干预。另一方面,进行基础研究 (basic research) 是研究者理解行为和心理过程的重要途径。人们通常将基础研究描述为“出于个人兴趣的知识探索”,基础研究是在典型的实验室环境中完成的,其目的在于检验理论。

科学理论的建构和检验

- 理论是用来解释现象的原因的,它在解释的范围和水平上存在差异。
- 科学理论是按逻辑组织起来的一组命题,用来定义事件、描述事件间关系和事件的发生过程。
- 成功的科学理论能够对实证知识进行概括,通过提出可检验的假设,对研究进行引导并且能通过严格的检验。
- 研究者通过判断理论的内部一致性、检验理论过程中是否出现假设的结果,以及基于严密的解释能否做出明确的预测等方面对理论进行评价。

理论是关于物质世界如何工作的“观念”,心理学家提出有关行为和心理过程的本质理论,以及人和动物行为与思考的模式。心理学理论可以从不同角度展

开,例如,理论可以从生理学角度展开也可以从符号学角度展开(见 Anderson, 1990; Simon, 1992)。精神分裂症的生理学理论认为,精神分裂症是生理原因如特殊的遗传载体导致的;而基于符号学角度展开的精神分裂症理论更可能提出心理因素方面的解释,如情感冲突或压力。当然,精神分裂症的某个理论也可能既包含生理原因也包含心理原因。理论中的命题可以用语言说明,也可以用数学方程或计算机程序表示。

理论通常在其试图解释的现象范围上存在差异。一些理论试图解释特殊的现象,例如,布朗和库利克(Brown & Kulik, 1977)的理论试图解释“闪光灯记忆”现象。闪光灯记忆即我们对特殊的个人境遇的记忆,尤其是对生活中怪异事件和情感事件的记忆,例如2001年可怕的9·11事件。而另一些理论的范围可能更广,因为它试图描述和解释更复杂的现象,如爱(Sternberg, 1986)或人的认知(Anderson, 1990; 1993; Anderson & Milson, 1989)。一般说来,理论的范围越广,理论本身就越复杂。当代心理学的多数理论在范围上趋中,只试图说明限定范围内的现象。

理论是科学家的直觉、个人观察和发现的知识(已知事实和概念)的有机结合。著名哲学家卡尔·波普尔(Karl Popper, 1976, pp.268-269)认为,创造性理论的产生需要两个条件,一个是“对问题的强烈兴趣(因此而准备一次又一次地尝试)”,一个是“批判性的想像力”。批判性的想像力是一种批判性的思考能力,它存在于很多方面。批判性的想像力意味着思维超越了公认的思考问题的界限——通常指“问题之外”的思考。假设我们有解决某个问题的强烈愿望,那么建构理论最有效的办法是,批判地检查对于这个问题我们都了解了什么,并针对问题找出现有知识体系中存在的不足和从未见过的错误。此方法和我们在“开始做研究”部分讨论的“提出假设”相似(McGuire, 1997)。

尽管理论在其解释水平和范围上存在差异,但在这些差异之中存在定义所有理论的共性。我们可以为科学理论(theory)下一个正式的定义:科学理论是一组命题(断言、陈述、主张)的逻辑性组织,这些命题用来定义事件(概念),描述事件间关系和解释事件的发生过程。例如,闪光灯记忆的理论首先需要准确地陈述什么是闪光灯记忆,其中包括对闪光灯记忆和传统记忆的区别的描述;其次,这个理论需要包括情感投入程度和记忆数量间关系的描述;最后,这个理论还需要解释为什么在某些情况下人们的闪光灯记忆明明是错误的,但人们对这种记忆却相当自信(见 Neisser & Harsch, 1992)。塔拉瑞可和鲁宾(Talarico & Rubin, 2003)对学生2001年9·11事件记忆的调查结果就是这种情况:尽管随着时间的推移记忆的准确性会下降,但被试对他们的鲜活记忆仍然很自信。

理论的主要功能是组织实证知识和指导研究(Marx, 1963)。即使像闪光灯记忆研究这样相对特殊的领域,也已经进行了大量的研究。随着理论的范围逐渐增至诸如爱或人类认知等领域,相应的知识主体也会扩大。科学理论的重要贡献是呈现个别研究结果的逻辑组织,找出研究结果之间的重要关系。除了提供一系列研究结果的逻辑组织外,科学理论还通过提出可检验的假设来引导研究。

理论通常要求提出干预过程来解释观察到的行为(Underwood & Shaughness, 1975)。这些干预过程把研究者操纵的自变量和随后测量的因变量联系起来。因为

这些过程“处于”自变量和因变量之间，所以称之为中间变量。如果你回想一下计算机的使用方法，你可能会理解中间变量。当你在键盘上按键或点击鼠标时，你将看到（或听到）显示器、打印机和音箱呈现不同的结果，但这些结果并非你按键或点击鼠标的直接结果，中间变量就相当于连接按键和显示器输出结果间的无形的软件。

中间变量像是计算机软件，计算机软件连接按键和显示器上输出的结果，中间变量连接自变量和因变量。举个心理学的例子，如“渴”的构想。研究者控制被试的饮水剥夺时间，并在这个特殊时间段后测量被试的饮水量。在禁止被试饮水后和允许被试喝水前这段时间内，被试是“渴的”——需要补充身体水分的心理体验。渴，是一种允许理论家连接“饮水剥夺时间”（自变量）和饮水量（因变量）的构想。和计算机软件一样，渴也是一种无形的连接，就是说，观察人们被剥夺饮水一段时间之后会做什么，渴就是依据此观察推断出的构想。诸如渴这样的中间变量不仅连接了自变量和因变量，还解释了变量为何被联系起来。因此，当研究者用理论解释结果时，中间变量起着重要的作用。

中间变量和理论都是有用的，因为它们允许研究者找出看似不同的变量间的关系。其他自变量也可能对渴产生影响，例如，盐的食用量。表面上看，这两个自变量——剥夺饮水时间和盐的食用量——非常不同。然而，两者都影响饮水剥夺后的饮水量，两者都能通过中间变量得以解释。其他和饮水量相关的自变量包括运动量和温度，运动越多或温度越高，人们就越渴，饮水量越大。尽管这些例子强调自变量对理论发展的作用，但因变量对理论发展所起的作用也很重要。因此，研究者宁愿将饮水量作为因变量，也不愿将其他和渴相关的心理体验作为因变量。例如，当个体被剥夺饮水时，个体可能会努力去获得饮用水甚至喝一些带苦味的水。因此，为获得水所付出的努力和水中苦味的含量也可以作为因变量。

在心理学理论发展中中间变量至关重要。在上述例子中，饮水剥夺研究中看似不同的变量：盐的食用量、运动、温度、饮水量，为获得饮用水付出的努力和水的味感，能依赖“渴”的中间变量构成理论。中间变量的例子——和理论——在心理学中有很多。例如，“抑郁”的中间变量将导致抑郁的理论因素（如神经病学因素、某些部位的损伤）和抑郁的不同症状（如悲伤、无助、睡眠紊乱和食欲下降）相连。“记忆”作为中间变量用来解释学习时间和学习成绩的关系。正如你将在心理学中学到的一样，中间变量就是打开变量间复杂关系的钥匙。

如何对科学理论进行评价和检验是心理学和哲学中最难的问题之一（例如，Meehl, 1978, 1990a, 1990b; Popper, 1959）。金布尔（Kimble, 1989）提出了一个简单易行的方法。他说，“最好的理论是经过逻辑和实证检验之火考验的幸存者”（p.498）。也就是说，科学家评价一个理论，首先要看它是否具有逻辑性，即判断理论是否有意义，判断理论的命题间是否存在矛盾。理论的逻辑一致性是通过科学界的洞察力来检验的。例如，心理学文献中对闪光灯记忆的确切定义就有过激烈的争论（例如，Cohen, McCloskey, & Wible, 1990; Pillemer, 1990; Winograd & Neisser, 1992）。

金布尔（1989）提到评价理论的第二把“火”，是用实证检验理论中提出的假设。成功地检验一个假设有助于增强理论的可接受性，不成功地检验一个假设则

会减弱理论的可接受性。根据这种观点,最好的理论应该是顺利通过了所有假设检验的理论。但是,对假设进行检验的过程困难重重,因此,要想证明理论成立或不成立也是困难重重。例如,一个理论,尤其是复杂的理论,可能存在许多特定的可检验假设。理论不可能因为某个单一检验而被推翻(例如, Lakatos, 1978)。此外,理论还可能包含没有充分定义的陈述和概念,或者还可能有以复杂甚至匪夷所思的方式提出的与行为相关的干预过程。这种理论可能经久不衰,但它的科学价值却值得怀疑(Meehl, 1978)。根本上说,理论是否具有权威性取决于科学界的评价。

一般来说,能做出精确预测的理论更有实用意义(Meehl, 1990a)。例如,一个理论预测儿童在12岁时就已经掌握了抽象推理,而另一个理论预测12~20岁是抽象推理的发展时期,那么,第一个理论在预测上就比第二个理论更精确和更具有可检验性。建构和评价理论时,科学家也会考虑成本(Marx, 1963)。在多个可供选择的解释中接受一个最简单的解释就是遵循吝啬原则,科学家更喜欢那些对现象做出简单解释的理论。在1932年的创造性思维研究中,艾瑞克·弗罗姆曾给100位著名的科学家和哲学家写信,以了解他们“产生最具有创造性思想的认知过程”(Fromm, 1998, P.1195)。在阿尔伯特·爱因斯坦的回复中,他描述了自己提出相对论时的思维过程,简要说明了吝啬原则:

一直在探索简单的实证关系,始终坚信存在一个简单的逻辑涵义,并受此信念推动不断进行探索。(Fromm, 1998)

总之,好的科学理论能通过最严格的检验。与直觉稍有不同的是,理论假设的证伪检验比证实检验能提供更多的信息(Shadish, Cook, & Campbell, 2002)。对一个特殊理论的证实检验的确为理论提供了支持,但这种逻辑的证实并不能排除对同一现象的其他理论解释。证伪检验是消除理论分歧的最好方法。理论的建构和评价是科学工作的核心,也是科学心理学健康发展的根本。

小 结

作为研究方法,科学方法不是根据直觉进行描述的,而是根据实证程序和控制现象的影响因素来进行描述的。科学家在实验时对实验的操纵和控制最为严格。在实验中,由实验者系统地操纵以探查对行为产生影响的因素称之为自变量,用来确定自变量产生影响的行为测量称之为因变量。

科学家客观地报告实验结果,对概念进行操作定义时无偏性和客观性显得尤为重要。心理学研究人员把概念看作是“构念”,他们采用尽可能准确和精确的仪器,用物理测量法和心理测量法对现象进行量化,追求测量的信度和效度。假设是对事件试探性的解释,但对科学家来说,假设还必须具有可检验性。缺乏恰当定义的假设、循环假设、凭空猜想的或超出科学范围的假设都不具有可检验性,假设一般由理论衍生而来。

科学方法的目标是描述、预测、解释和应用。定量研究和定性研究都可用来描述行为，观察是科学描述的重要基础。当两项测量存在相关时，我们可以通过了解一项测量的数值预测另一项测量的数值。只有找出产生现象的原因才能理解和解释现象，理解和解释现象时要求有事件共变、存在时序关系，以及对其他可能原因排除的证据。当两个潜在的效果变量存在共变时，例如，无法分辨两个自变量对行为产生的影响时，我们就说研究被混淆了。即使实验控制严密，研究者可以根据实验研究做出因果推论，但还存在能否将实验结果推广到其他人群或其他环境的问题。在应用研究中，心理学家努力运用知识和研究方法改善人们的生活。基础研究的目的在于获得有关行为和心理过程的知识，在于对理论进行检验。

对心理学来说，科学理论的建构和检验是科学方法的核心。理论即一组命题的逻辑组织，用来定义事件、描述事件间关系和解释事件的发生过程。理论的重要作用表现在两方面：一是组织实证知识，一是通过提供可检验假设引导研究。中间变量在心理学理论的发展中至关重要，因为这些概念允许研究者对自变量和因变量间的关系做出解释。

重要概念

控制	信度
实验	相关
自变量	因果推论
因变量	混淆
概念	应用研究
操作定义	基础研究
效度	理论

复习思考题

1. 对下列概念，分别从科学方法和日常方法两方面加以区别：一般方法、观察、报告、概念、仪器、测量和假设。
2. 比较自变量和因变量，并分别举例。
3. 在心理学中采用操作定义的主要好处是什么？在哪两种研究方法中不能采用操作定义？
4. 区别测量仪器的准确性和精确性。
5. 测验的效度和信度有何差别？
6. 可检验假设的三个特性是什么？
7. 科学方法的四个目标是什么？简要描述科学方法每个目标想要实现的内容。
8. 区别常规法则研究方法和个人特质研究方法在研究对象和归纳类型上的差异。
9. 两列变量存在相关有何意义？也就是说，两列变量存在相关允许研究者做什么？
10. 以书中提到的研究为例，举例说明因果推论的三个条件。（可以用一个例子说

明一个条件,也可以用例子说明多个条件。)

11. 基础研究和应用研究有何差别?
12. 什么是中间变量?提出一种心理概念,该概念可以充当“攻击行为”(出现/不出现)和“攻击性反应”间的中间变量。通过提出包含中间变量的假设来解释这些变量间的联系。
13. 描述逻辑一致性和实证检验在评价科学理论时的作用。
14. 解释证伪比证实能提供更多信息的原因。

挑战性问题

1. 请找出下列研究中的自变量和因变量,你在每个研究中至少能找到一个因变量。
 - A. 某心理学家研究食物剥夺对运动行为的影响。她将四组小白鼠分配到四种实验条件中:食物剥夺0小时、8小时、16小时、24小时。食物剥夺时间结束后测量各组动物在活动轮上消耗的时间。
 - B. 某体育教师研究室外运动器械(如滑梯、秋千、攀登墙)对运动协调能力的影响。实验时间为8周。实验中,学前儿童使用室外运动器械的时间为每周4小时、每周6小时和每周8小时,8周后从儿童蹦、跳、单脚站立三个方面来测量他们的运动协调能力。
 - C. 某发展心理学家研究他人在场对幼儿语言行为的影响。被试为3岁幼儿,观察幼儿语言行为的时间为30分钟。将被试分成两组,其中一组幼儿在观察期间和成人在一起,另一组幼儿在观察期间和非被试幼儿在一起。心理学家测量每个被观察幼儿的语言表达数量、持续时间和复杂性。
2. 某生理心理学家研制出了一种药物,他认为这种药物将引起骑马竞赛的变革。他将这种药物命名为Speedo,并认为这种药物能让马跑得更快(出于对这个假设问题的好奇,我们忽略给赛马喂药的不合法性)。她挑选了两组赛马,并给其中一组赛马注射Speedo,每周一次,共注射四周,由于Speedo会对马的消化系统产生副作用,因此给注射Speedo的赛马配给高营养饮食。而那些没有注射Speedo的赛马则维持常规饮食。4周之后,测量所有赛马完成2英里竞赛的时间,结果发现注射Speedo的赛马组的平均速度显著快于没有注射Speedo的赛马组。心理学家认为他的药物有效。
 - A. 找出上述研究的自变量(及水平),可能与首要自变量发生混淆的潜在自变量,详细地解释这种混淆是如何发生的。
 - B. 给上述研究做出准确合理的结论。
 - C. 最后,设计研究Speedo效果的合理方法。
3. 《纽约时报》报道了卡内基梅隆大学一项历时两年、耗资150万美元完成的研究。研究经费由国家自然科学基金和几个大的技术公司提供。研究对象为匹兹堡地区的169名被试。研究者调查了网络使用和心理健健康之间的关系。该研究的一位负责人声明,该研究不涉及对网络的过度使用,被试是正常成人和他们的家庭成员。报告认为,使用网络越多的人心理越不健康。例如,每周上网一小时会轻微增加抑郁和孤独,会减少和家庭成员之间的互动。研究者得出结论,使用网络将导致心理健康水平下降。他们认为,网络导致使用者和他人情感联系下降。
 - A. 研究者得出结论:使用网络会导致心理健康水平下降,仔细阅读这篇报道,什么条件导致做出这种推论,还缺少哪个必要条件?
 - B. 在形成结论之前还应该寻求什么信息以证实之?(你可能要从《纽约时报》1998年8月30日刊登的《孤独的网络》和《华盛顿邮报》1998年9月7日刊登的《网络抑郁研究评论》这两篇报道着手。)
 - C. 部分经费由几个大的技术公司资助,这在你评价该研究时会产生什么影响?
4. 在发展心理学课程中的学习中,做笔记是否会影响期末测验成绩?学生在整个学期中在一本125页的《学习指导》中记录笔记,《学习指导》包括书中和课堂上关于课程内容的问题。学生的笔记从三个维度进行测量:完整性、范围、准确性。研究结果表明,笔记

(续)

越准确的学生在问答题和多项选择题中成绩越好。基于这些发现,研究者认为,教师在课堂中应采用稍作暂停和提问等引导技巧。研究者认为这些技巧能够提高学生课堂笔记的准确性,而准确的课堂笔记将显著增加学生的学习成绩。

- A. 结合学生笔记准确性和测验成绩间的关系,说明本研究中具备了因果推论的什么条件?缺乏什么条件?(务必清楚地指出因果推论的三个条件。)
- B. 结合本研究结果指出科学方法的目标。

拓展练习的答案

1. 该研究的自变量是插播广告的电视节目类型,自变量水平为暴力节目和非暴力节目。因变量是通过回忆和再认记忆测验测量出的被试对广告内容的记忆。
2. 如果自变量有两个水平,那么因变量随自变量水平不同而发生变化时,就存在两种可能的解释,要么是一种条件下成绩更好,要么是另一种条件下成绩更差(如果两个人身高不同,相当于说一个更高,另一个更矮)。在这个研究中,研究者对研究结果的解释是,暴力节目降低了被试对广告的记忆。但这种自变量只有两个水平的研究不能推翻此研究结果的另一种解释——非暴力节目增强了被试对广告的记忆。分辨上述两种解释的办法是,将自变量水平设为三个。在此研究中,可以在三个水平上变化节目性质——即自变量的类型:积极节目,中性节目和暴力节目。当积极节目和中性节目对记忆的影响没有差异,而暴力节目的记忆更差时,就可以接受暴力节目降低了被试对广告的记忆的解释;反之,如果暴力节目和中性节目对记忆的影响没有差异,而积极节目的记忆更好时,就可以接受非暴力节目增强了被试对广告的记忆的解释。当然,此研究的自变量分为三个水平时也可能产生其他的结果。这个练习告诉我们,操纵自变量对如何解释结果影响非常大。

挑战性问题1的答案

- A. 自变量:食物剥夺时间,分为四个水平;因变量:动物耗在活动轮上的时间(以分钟为单位)。
- B. 自变量:在室外器械上运动的时间,分为三个水平:每周4小时、每周6小时和每周8小时;因变量:运动协调性的测试分数。
- C. 自变量:他人是否在场,分为两个水平:在场和不在场;因变量:儿童语言表达的数量、持续时间和复杂性。

第3章

心理学研究中的伦理问题

本章大纲

引言

研究前需要思考的伦理问题

APA 伦理标准

注解

风险收益比

风险判定

最小风险

风险应对

知情同意

APA 伦理标准

注解

心理学研究中的欺骗

APA 伦理标准

注解

事后解释

APA 伦理标准

注解

用动物做研究

APA 伦理标准

注解

心理学研究的报告

APA 伦理标准

注解

伦理决策的步骤

小结

引言

好的科学必须要有优秀的科学家，科学家的专业能力和诚实是保证高质量科学的根本。科学研究过程需要诚实，这既是科学家个人的责任，也是科学团体（如美国心理协会和美国心理学会等专业组织）的责任。每个科学家都担负着探求知识和改善生活质量的道德责任。迪纳和克兰德尔（Diener & Grandall, 1978）概括出以下几种具体责任，他们认为科学家应该：

- 以积极的态度从事研究；
- 准确报告研究结果；
- 诚实对待研究资源；
- 学术交流中，公正地认可那些曾经为研究做出贡献和付出努力的人；
- 思考研究给社会带来的后果；
- 以科学和专业的方式公开谈论社会关注的问题。

科学家要担当这些职责就会面对挑战，面对一些道德上的两难问题。为了指导心理学者做出正确的道德决策，美国心理学会制定了《APA 伦理准则》，为心理学者“提供一套专业科学工作所通用的原则和标准”（美国心理学会 2002, p.1062）。这些标准可以概括为五个通用原则：善意和避免伤害；忠实和责任；正直；公正；尊重人权和尊严。以这五个原则为基础，《APA 伦理准则》阐明了心理学研究、治疗、教学及管理四个方面的道德行为标准（见美国心理学会，2002）。《APA 伦理准则》涵盖性骚扰、心理服务资费、借助媒体向公众提供建议、测验编制及课堂教学等诸多方面。

正如在《APA 伦理准则》序言中所阐述的，心理学者应该“承诺并且终生负起研究的道德责任；敦促学生、被指导者、助手、同事光道德行为；与他人协商道德问题”（美国心理学会 2002，第 1062 页）。对所有的心理学专业的学生来说，不遗余力地遵守这些理想目标和行为准则同样重要。你可以登陆 APA 网址（<http://www.apa.org/ethics>）查阅《APA 伦理准则》。

《APA 伦理准则》中有很多内容是直接针对心理学研究的（主要见标准 8.01~8.15），包括心理研究中如何对待人和动物。大部分伦理准则适用于一般规律，所以在特殊情境中需要专门界定。可应用于特定研究情境的道德标准不止一种，而有时标准之间会彼此矛盾。例如，伦理准则要求保护被试免受身体伤害，然而某些有关药物或其他侵入性处理的研究可能会使被试处于身体伤害的风险中。动物被试应得到保护，然而有些研究会使其疼痛或遭受其他痛苦。解决这些进退两难的道德难题并非易事，需要深思熟虑和尽职尽责地来制定道德决策方案。

互联网改变了许多科学家的研究方式，心理学家也不例外。例如，全世界各地的研究者经常共同研究一个科学问题，并通过网络便捷地交换彼此的想法和发现。大量档案信息可以通过官方网站获取（如美国人口普查局）。20 世纪末的几十年中，研究者通过万维网搜集人类被试数据，一项研究竟可有数百万被试！网上的心理研究类型包括简单观察（如记录在聊天室内的“行为”）、调查（问卷，包

括人格测验)以及操纵变量的实验。

网络为行为科学家提供了诸多机会,但同时也带来了许多道德问题。主要问题在于,在线研究情境中研究者不在场,获得充分的知情同意和提供事后解说比较困难,在被试隐私权保护上也有难度(另参阅 Kraut et al., 2004; Nosek, Banaji & Greenwald, 2002, 以寻求此类问题的建议及解决方案)。我们将在本章讨论一些道德问题,并在描述具体研究方法的章节中继续讨论有关内容。

做道德决策时最好向同事,尤其是那些在专门领域更具经验和学识之士咨询。事实上,在某些情况下,没有参与实验的人对研究进行评价是规则所要求的。本章其余部分将明确心理学研究的道德准则,同时就某些方面的问题做一些简明注解,并提供一些假设的研究例子来说明研究中道德方面的问题。通过研究解决道德问题的例子,你将逐渐学会使用道德标准来应对挑战,并克服道德决策的困难。请你与同事、教授及其他已具有心理学研究经验的人讨论这些计划。

研究前需要思考的伦理问题

- 在实施研究之前,必须审查该研究是否符合伦理标准。

研究者在开始一项研究之前,必须首先考虑与道德相关的问题。只有在研究前谨慎计划并向适宜的个人及群体咨询才可能避免出现道德问题。不符合道德标准的研究会危及整个科学进程,阻碍知识的进步,损害公众对科学和学术团体的尊重,还可能导致个体或机构接受严重的法律惩罚和经济损失。《APA 伦理准则》中的条目(美国心理协会, 2002, pp.1069~1071)阐述了研究者在研究前必须遵循的重要原则。

APA 伦理标准

8.01 机构批准 在实施研究之前,心理学者提供其研究计划的准确信息并获得批准,心理学者按照批准的研究议定书开展研究。

注 解

- 机构审查委员会(Institutional Review Boards, IRBs)审查心理学研究,以保护人类被试的权利和福利。
- 动物实验管理委员会(Institutional Animal Care and Use Committees, IACUCs)审查以动物为研究对象的研究,确保人道地对待动物。

1974年,美国《国家研究法案》生效。在其要求下,建立了在生物和行为研究中保护人类被试的国家委员会。该法案要求向联邦专门机构申请研究经费的社会机构必须组建委员会,由委员会对该机构承办的研究进行审查。大专院校和综

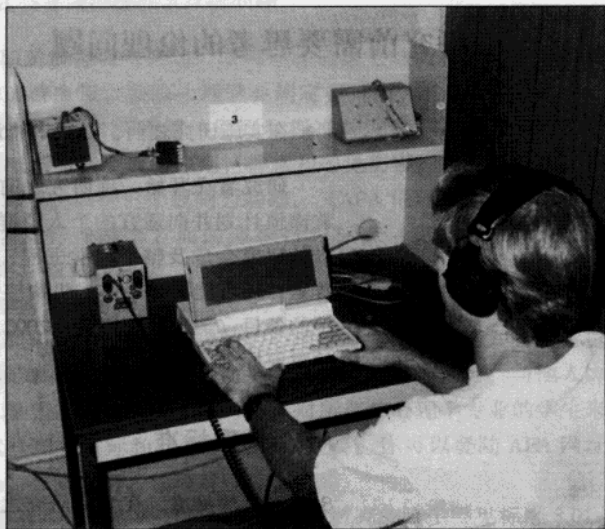
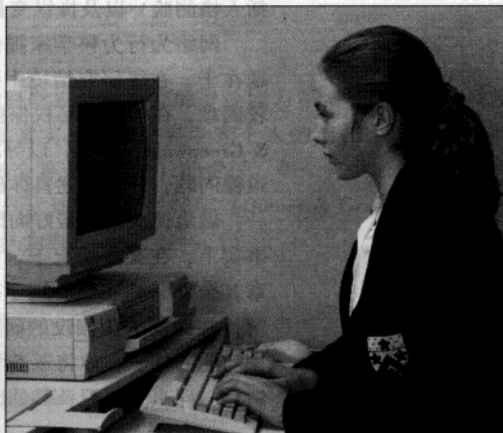
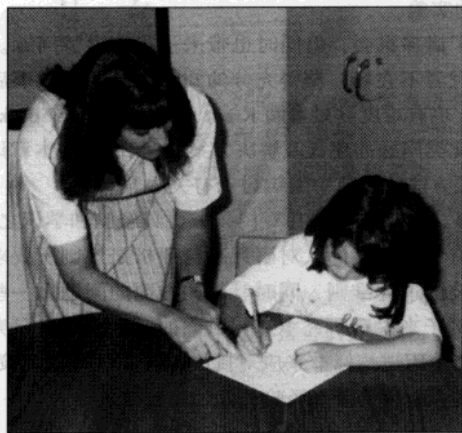


图 3.1 用人进行研究会凸显很多道德问题

合性大学已组建的委员会被视为前面所提到的机构审查委员会 (IRBs)。可登陆网址 (<http://www.hhs.gov/ohrp>) 查阅有关机构审查委员会的一些联邦法规。经过机构审查委员会审查, 确保研究者保护被试免受伤害并维护其权利。联邦法规对机构审查委员会的成员资格及其职责做了具体要求 (见 *Federal Register*, 1991 年 6 月 18 日)。例如, 机构审查委员会必须由 5 个具有不同技术背景和分属不同技术领域的人组成, 其中既有科学家也有非科学家, 并且至少有一位机构审查委员会成员不隶属于该机构。某些社区人员, 如牧师、律师、护士等, 经常被邀约加入这些委员会。

一位以人为研究被试的心理系学生很可能会向负责审查心理学研究的系委员

图3.2 二战后，纽伦堡军事法庭指控德国医生犯有违背人道主义的罪行，包括在当事人没有同意的情况下对其进行医学实验。法庭对此类案件的判决导致了纽伦堡法则(Nuremberg Code)的产生，该法案制定了对人进行实验的规则



会递交研究计划，根据系委员会和机构审查委员会之间的协调，研究在各种层面上同时接受系委员会和机构审查委员会审查。机构审查委员会有权批准或否决该研究计划，或要求修正计划后再实施该研究。机构审查委员会对审查的公平性负有道德责任，而这种公平性是从机构、研究者、被试的角度出发来考虑的(Chastain & Landrum, 1999)。

1985年，美国农业部及公共卫生组织制定了新的保护实验室动物的方针(Holden, 1987)。动物研究机构必须具备一个动物实验管理委员会(IACUC)，这些委员会一定要有一位科学家、一位兽医和一个不隶属于该机构的成员。动物实验管理委员会不仅审查研究程序，还关注动物的生存环境，监督与动物直接接触的人必须接受适当培训(Holden, 1987)。

几乎每所学院和大学所承担的所有研究在某个阶段均应接受独立委员会审查。涉及人或动物的研究如果违背了联邦法规，就会导致研究终止、联邦经费削减以及巨额罚款(Holden, 1987)。基于联邦法规和各机构审查政策的复杂性，任何研究者都应在研究前向权威机构咨询审查程序。提交研究计划的学生会得到有关IRB(McCallum, 2001)和IACUC(LeBlanc, 2001)方面的一些有益建议。

风险收益比

- 对研究项目进行风险及收益的评价以确定是否应该实施该研究。

除了核查研究是否遵循了适当的道德原则，机构审查委员会还需权衡研究的风险收益比。研究能够增加新知识或改善人们的生活，社会和个体可以从中受益；如果不进行研究，则会付出相应的代价：我们会错过获取新知识的契机，失去改善人类环境的机会。对于个体被试来讲，研究也是有代价的。例如，当被试处于有潜在伤害的环境中时，就承担了被伤害的风险。当然，审查者必须首先考虑这些潜在代价和收益。机构审查委员会是由有识之士组成，并且他们对研究没有个

人兴趣偏好，所以更适于判定风险收益比，最终决定是否批准该研究计划。

风险收益比 (risk/benefit ratio) 提出这样一个问题：这个研究值得做吗？对此问题不存在数学上的答案。相反，在对研究计划进行审核时，机构审查委员会力求分析该研究对被试及社会的风险和收益，并在评价结果上达成一致。当风险超过收益时，机构审查委员会否决该研究；当收益超过风险时，机构审查委员会会批准该研究。很多因素会影响研究风险和收益的权衡，其中最关键的是风险的性质、收益的大小以及研究的科学价值和社会价值 (Fisher & Fryberg, 1994)。如果我们事先就确定研究对个体有明确和即时的利益，或者研究具有明显的社会价值和科学价值，我们就可以容忍较大的风险。例如，一项行为疗法的研究项目或许需要被试承担风险，但如果该治疗极具公益性，那么对个体和社会的可能收益就会超过研究中存在的风险。

在判定风险收益比时，研究者还须考虑到是否会产生有效的、可解释的结果。罗森塔尔 (Rosenthal, 1994b) 充分证明了研究质量应该作为道德审查的一个因素。“其他一切条件都相同的前提下，研究的科学质量越高，研究就越具道德保护性” (p.127)。更具体地说，“如果由于一项低质量科学没有产生收益，那么我们如何证明被试的时间、注意力、精力及金钱、空间、供应及其他资源的使用价值呢？” (Rosenthal, 1994b, p.128)。也就是说，研究者必须力求研究在科学角度上符合最高的质量标准。

当存在潜在风险时，研究者必须确保此程序风险最低，并且没有可替代的方法，还要保证此研究问题在前人的研究中尚未被成功解决。如果不认真查阅心理学文献，研究者可能会重复已有的研究，而将被试置于没必要的风险中。

风险判定

- 心理学研究中的潜在风险包括身体伤害、社会伤害及心理或情绪压力等。
- 评价风险必须以被试的日常行为、身体和心理健康状况及其能力为依据。

判定被试是否“处于风险中”在道德决策时是一个很大的难题。生活本身就是冒险。工作或上学时往返乘车、横穿马路、乘坐电梯等都有风险，并非仅仅心理学实验存在风险。如果心理学实验的人类被试永远不面对任何危险，所有的研究就会停滞不前。要判定风险的构成就必须考虑日常生活中的风险。

判定风险还应考虑被试特征。某类行为对一些人来说可能具有严重风险，而对其他人则意味着没有风险。爬一段台阶或许会有引发老年人心脏病的风险，但对青年人则没有风险。同样，严重抑郁或焦虑的被试对某些心理任务的反应或许比其他人更强烈。因此，当考虑风险问题时，研究者必须考虑参加研究的被试群体或个体的特征。

说到风险，我们常常想到的是身体伤害的可能性，实际上，社会科学研究中的被试经常要冒的风险是社会伤害或心理伤害。在心理研究中，被试信息被泄露给他人就是这类潜在的社会风险之一。如果公开被试的异常性行为问卷，对于该被试来说就会导致尴尬等社会风险。心理学研究中搜集的个人信息可能包括智力、

人格特质、政治信仰、社会信念或者宗教信仰等,被试或许不想把这些个人信息透露给老师、老板或同事。如果研究者不能很好地保护被试的隐私,就会增加这些被试的社会风险。有人认为在以网络为基础的研究中,被试面临的**最大风险**是把个人信息暴露于研究情境之外(Kraut et al., 2004)。还有的研究者认为,尽管网络提供了“匿名感”(Nosek et al., 2002, p.165),但在某些情况下,这种感觉是假的,研究者必须考虑在数据传输、存储及研究后与被试联系时保护其隐私。

在一些心理学研究中,被试会体验到强烈的心理或情绪压力,这样的研究可能就存在心理风险。例如,当被试正等待实验时,屋子里忽然充满了烟雾,可以想像一下她可能体验到的压力。研究者希望用房间的烟来模仿一种紧急情境,被试或许会一直体验相当程度的焦虑,直到研究者说明烟的真实性质。另外,仅仅参加心理学实验就能引起某些被试的焦虑。曾经有一个学生确信,学完一系列无意义音节词表后研究者对他已了如指掌,该生猜想心理学者可能会通过他所使用的单词联想来研究他的人格。事实上,他参加的只不过是一个简单的有关遗忘的记忆实验。研究者有义务尽最大可能保护被试免受情绪压力或心理压力,包括被试对心理任务的误解而产生的压力。

最小风险

- 当研究过程和研究内容与被试日常生活经历相似时,该研究的**风险最小**。

有时需要区分“处于风险中”的被试与“处于最小风险中”的被试。**最小风险(minimal risk)**意味着被试在研究中所体验到的伤害或不适没有超过他们在日常生活中或在常规的生理或心理测验中所体验的伤害或不适。以心理学实验室研究中的纸笔测验为例,纸笔测验旨在评价各种心理能力,要求被试快速完成测验并接受成绩反馈。尽管此种情境也可能产生压力,然而心理伤害的风险仅相当于一名学生所面临的风险。因此,此类研究对大学生来讲仅具有最小的风险。当伤害超过最小限度时,则认为被试处于风险中。如果研究使被试处于风险中,那么研究者保护被试安全**的责任随之增加**。

风险应对

- 无论是“处于风险中”还是“处于最小风险中”,被试必须得到保护。风险越大,需要的保护越多。
- 为保护被试**不冒社会风险**,被试应该匿名;如果做不到匿名处理,研究者应为其信息保密。

即使潜在风险不大,研究者也应努力将风险降到最低以保护被试。例如,在记忆实验开始前,简单说明该记忆任务不用来测量智力或人格,那么被试体验的压力水平就会降低。如果研究中伤害的可能性大于日常生活中伤害的可能性,那么研究者保护被试的责任也相应增大。例如,在心理学实验中,如果被试处于强烈的情感压力下,那么机构审查委员会可以要求安排一个临床心理学家来对被试

拓展练习

判断下列每种研究情境呈现的是“最小风险”（也就是说，不超过日常生活中的风险）还是“处于风险中”。如果你判断被试“处于风险中”，你会给研究者提供什么建议来降低风险以保护被试？在思考过程中，你可能会预期到本章接下来将讨论的一些道德问题。

1. 要求大学生在一份词单列表中选择能够描述他们目前情绪的形容词。研究者借此识别抑郁的学生，以便为与抑郁相关的认知缺陷研究选择被试。
2. 老年公寓的老人在他们的休息室内完成一系列成就测

验。心理学者试图判断心理机能是否随年龄增长而逐渐降低。

3. 心理学研究方法课堂上的学生们看见有一个学生在上课程期间进入教室，并愤怒、大声地与导师讲话，然后离去。作为目击证词研究的一部分，要求学生描述这个闯入者。
4. 一位研究者在心理学入门课上招募学生参与酒精对认知功能影响的研究。实验要求一部分学生在玩电脑游戏前喝下2盎司酒（掺有橘汁）。

进行咨询。可以想像，在线研究在这一点上处于道德两难困境之中，被试在网络研究情境中就像在实验室研究中一样会体验情感压力，因此，在某些研究情境中在线被试可能会面临风险，机构审查委员会必须进行风险和收益评估。然而，由于研究者不在实验情境中，“尤其是研究者或许无法监控在线被试并补救研究导致的伤害”（Kraut et al., 2004, p.112）。一种可能的解决方法就是先获得被试基本数据以确定哪些人可能会处于风险中，并在研究开始前将其剔除。但是，有时高风险的研究在网上根本不按道德准则操作（Kraut et al., 2004）。

如果被试遇到的风险超过最小风险，那么这样的研究应该被禁止，除非已探究出低风险的数据搜集方法。某些情况下，应使用观察或问卷等描述方法而不使用实验处理。研究者也可以利用没有实验压力的自然发生的“处理方法”。例如，安德森（Anderson, 1976）采访了被飓风洪涝毁了生意的经营者。他发现适宜的压力水平导致有效的问题解决及应对行为，而高于或低于此压力水平，则问题解决成绩反而降低。实验诱导产生压力的实验室研究也证明了这样的结果。

为保护被试免受社会伤害，搜集数据时应该告诉被试切勿使用姓名或其他有关身份的信息以保证其反应的匿名性。如果做不到，研究者应除去被试反应记录中的身份信息以保护其反应的隐私性。研究者必须多次测验被试或追踪某些个体时，可以在研究开始前将数字与被试随机匹配，被试的反应记录上只出现这些数字。如果被试所提供的信息敏感性很高，编码设计或许会有利于社会风险的最小化。可以用某些程序将编码数字与被试匹配，在名单上一个名字对应一个数字，这份名单的使用是受一定权限限制的。在线研究者要谨防电子窃听或黑客攻击数据，并且应该采取适当的预防措施（见 Kraut et al., 2004）。

匿名性或隐私保护能使被试更诚实、更开放地做出反应，对研究者也是有益的（Blanck, Bellack, Rosnow, Rotheram-Borus, & Schooler, 1992）。如若被试不担心他们的反应结果被泄漏，那么他们就不大可能撒谎或隐瞒信息。

图 3.3 美国公共卫生组织在 1932 年至 1972 年期间调查了未经治疗的梅毒发展历程,梅毒患者是来自阿拉巴马州美肯郡的贫穷非洲裔美国人。该组织没有对梅毒患者给予知情同意,患者没有意识到自己患了梅毒,致使疾病未得到治疗



知情同意

- 研究者和被试经常通过知情同意程序来达成一种社会契约。
- 研究者的道德责任有:清楚地描述研究程序、明确澄清可能影响到被试意愿的任何潜在风险,并解答被试对该研究的任何疑问。
- 被试的道德责任有:在研究中不撒谎、不欺骗、不进行其他欺诈行为,以恰当的行为方式做出反应。

心理学研究的成功与否依赖于学生、病人、客户及其他社区成员参加科学调查的意愿。有些研究中被试会因其耗费的时间和付出的努力而获得金钱或其他补偿。通常人们自愿参加研究,不需要任何补偿。无论哪种情况,研究者和被试都达成了一种社会契约。作为该契约的一部分,研究者有道德责任向被试澄清研究的细节,及被试可能面临的风险;研究者必须在研究过程中尊重被试的权利。研究者必须不遗余力地确保获得被试参加这项研究的知情同意。

恰当的行为方式是同意参加研究的被试应负的道德责任。例如,被试应该注意指导语并按研究者的要求完成任务。泰勒和谢泼德 (Taylor & Shepperd, 1996) 描述了一项研究,举例说明不负责任的被试行为可能导致的后果。该研究中,实验者告知被试独处时彼此之间不能讨论该实验。然而在实验者离开后,被试谈论起实验并彼此交换信息,结果导致实验无效。而且,当实验者后来询问被试对实验程序和研究目的知晓程度时,没有人承认他们交流过该实验并获得了重要信息。此事例说明了一个重要原则:被试撒谎、欺骗及其他欺诈行为违背了科学研究的严谨性。

APA 伦理标准

3.10 知情同意

(a) 心理学者亲自或者通过电子技术或其他方式进行实验或提供评估、治疗、咨询或顾问服务时，他们通过合理的、可理解的语言来获得被试的知情同意。除非法律、政府规定或者道德准则允许，否则不可以进行未经知情同意的研究。不同情况下知情同意内容有所不同，但一般要求被试：(1) 有同意的能力；(2) 已获知可能影响参加研究意愿的信息，包括保密限制和金钱及其他成本或偿付；(3) 明白自愿参加的性质并知道可以不受影响地自由地表达；(4) 有机会提出问题并得到答案。(另参阅标准 8.02，研究中的知情同意；9.03，评估中的知情同意；10.01，治疗中的知情同意)

(b) 对于法律上不能给予知情同意的人，心理学者则：(1) 给予适当解释；(2) 寻求个体认可；(3) 考虑到个人偏好和兴趣爱好；(4) 如果法律要求，那么征求合法监护人的适当同意。如果法律不要求该监护人的同意，那么心理学者应采取合理步骤保护被试的权利和安全。

(c) 如果心理服务是法庭要求的或其他机构允许的，心理学者在实施前要告知被试预期服务的性质，包括该服务是否法庭要求或许可的，以及任何保密限制。

(d) 心理学者应适当地用文件证明书面或口头同意、允许和赞成。(另参阅标准 8.02，研究中的知情同意；9.03，评估中的知情同意；10.01，治疗中的知情同意)

8.02 研究中的知情同意

(a) 当获得如标准 3.10 所要求的知情同意时，心理学者告知被试：(1) 研究目的，预计持续时间及程序；(2) 被试拒绝参加及参加后退出研究的权利；(3) 拒绝或退出后可预见的后果；(4) 合理的、可预见的影响参与意愿的因素，如潜在风险、不适或不利影响；(5) 可预期的研究收益；(6) 保密程度；(7) 参加研究的诱因；(8) 有关研究及被试权利问题的联系人，他们可以提出疑问和获得解答的机会。(另参阅标准 8.03，研究中使用声音和影像记录的知情同意；8.05，研究中的无需知情同意；8.07，研究中的欺骗。)

(b) 心理学者进行包括实验处理在内的干预研究，在研究开始时应向被试说明：(1) 实验处理的性质；(2) 控制组可能接受或不接受实验处理；(3) 实验处理的分配方式及控制组的选定；(4) 如果有被试不愿参加该研究或者研究开始后希望退出，有效的替代方法；(5) 参与研究的补偿和金钱成本，是被试偿付还是由第三方偿付。(另参阅标准 8.02a，研究中的知情同意)

8.03 研究中使用录音和录像的知情同意

心理学者以录音或录像的方式进行数据搜集前应征得被试的知情同意，除非：(1) 研究只包括公共场合的自然观察，使用记录信息预计不会导致身份确认或他人伤害；(2) 研究设计包括欺骗，知情同意在事后解释阶段告知被试。(另参阅标准 8.07，研究中的欺骗)

8.04 客户或病人、学生和下属被试

(a) 当以客户或病人、学生和下属作为研究被试时, 心理学者应采取措施, 避免遇到拒绝或退出研究的负面结果。

(b) 参加研究属于课程要求或算作选修学时, 被试可以在其他均等的研究活动中进行选择。

8.05 研究中无需知情同意

心理学者只有在下列情况下才可以无需知情同意: (1) 假定研究不会产生痛苦或伤害, 包括: (a) 教育情境中实施的常规教育实践、课程或课堂管理方法的研究; (b) 匿名问卷、自然观察或者档案研究, 透露被试的反应不会使被试处于刑事或民事责任或危害他们的经济地位、任职资格、声望等风险, 且隐私权得到保护; 或者 (c) 在组织情境进行的有关工作或组织效益相关因素研究, 对被试的任职资格没有风险, 隐私权被保护或者 (d) 法律、法规或机构规则允许。

8.06 向研究参与者提供诱因

(a) 心理学者应努力避免给参加研究被试提供过度的或不恰当的经济或其他诱因, 否则这些诱因会强制被试参加研究。

(b) 提供专业服务作为研究参与的诱因时, 心理学者要说明服务性质及风险、责任和局限性。(另参阅标准 6.05, 与客户/病人的交易。)

4.01 保密 心理学者有责任采取措施以保护获得的私密信息, 必须认识到法律或机构规定或职业、学术关系规定的不同的保密范围和保密限制。(另参阅标准 2.05, 工作授权)

注 解

- 必须使潜在研究被试知晓可能影响参与意愿的有关研究的各方面信息。
- 必须允许被试随时退出而不进行任何惩罚。
- 禁止强迫个体参加研究。
- 对于无法提供知情同意的个体, 必须获取其合法监护人的知情同意 (例如儿童、受精神损伤的个体); 得到其认可后被试才能参加研究。
- 研究者判定是否需要知情同意时 (如在公共情境中进行研究), 应该向其他有识之士咨询, 包括机构审查委员会。在这些情境下从事研究应特别注意保护个人隐私权。
- 隐私权指个体决定怎样向他人公开自己个人信息权利。

研究中与伦理准则相关的大部分问题是知情同意问题。这很合情合理, 因为知情同意是研究者和参与者之间的社会契约的最基本成分。知情同意 (informed consent) 是被试在充分理解研究性质、不参加的后果、影响参加意愿的所有因素后, 明确表达参加研究的意愿。合乎道德标准的研究要求充分告知被试可能影响参加意愿的所有因素。被试应该明白他们同意做的是什么, 被试还应明白他们可以随时退出而不受任何惩罚或歧视。研究者还必须做出努力来合理解答被试关于

图 3.4 儿童参与研究时知情同意问题尤为重要



研究的任何问题。这样，被试才可以做出参与研究的知情决定。被试的同意必须是自愿做出的，没有被过度引诱或强迫。研究者征求被试的知情同意应该贯穿研究始终。被试面对的风险大于最小风险时，书面知情同意是绝对必要的。

某些个体不能提供真正意义上的知情同意，如精神受损或情绪障碍者、儿童以及能力有限无法理解研究性质和潜在风险的人。只要可能，“同意”应该始终由被试自己做出。然而，在上述情况下，正式的知情同意必须由被试父母或其法律监护人做出。这类涉及道德方针的知情同意就出现了方法学上的问题（见 Adair, Dushenko, & Lindsay, 1985）。例如，一项研究发现，父母没有表示同意的孩子在学业成功和受欢迎程度上都不如父母表示同意的孩子。（Frame & Strauss, 1987）。此研究提出的问题是，父母同意的孩子所获得的结果是否适用于父母没有同意的孩子。然而，毫无疑问的是，按照程序考虑研究在道德方面的问题比这些方法论问题都重要。

在线研究再次提出了此领域的特殊道德问题。请思考一下，在线被试大都在电脑屏幕上按键表明已阅读或已理解同意条款内容，但这能否构成具有法律约束力的被试“签名”？研究者从何知道被试果真具有合法年龄并充分理解了知情同意的条文呢？对于判定被试是否已理解知情同意的条文，有人主张对其内容进行简短测验；在提供只有成人才能获得的信息之前，先用测验区分出成人和儿童（Kraut et al., 2004）。

一位研究生在对那些正在从家庭计划诊疗中接受服务的青少年进行访谈时遇到了一个两难困境（Landers, 1988）。十几岁的青少年去诊所无需得到家长许可，然而对未成年人实施实验前必须征得其父母的同意，这是标准的符合道德的做法，也是联邦法规所要求的。如果研究者请求父母允许他们的孩子参加研究，就会透露其子女接受了诊疗服务。该研究生向 APA 人类被试保护委员会寻求忠告，以决定在此困境下如何满足道德程序的要求。学生利用专家的建议，系统地编制程序并将其作为论文研究的一部分。无论何时出现道德难题，向知识丰富的专业人员

寻求忠告都不失为明智之举,但是,研究道德的最终责任还需要研究者来承担。

判定什么是不适当的引诱或强制参与并非易事。通常认为给参加心理实验的大学生每小时9美元的报酬不属于不当强制,而以同样的报酬去大街上招募极端贫困或底层的人们或许更具强制性,是不可接受的(Kelman, 1972)。囚犯或许会认为拒绝参加心理学实验会成为不合作的证据,使他们很难获得假释权。如果将大学生参加心理学实验(假设是具有教育意义的体验)作为一项课堂作业时,必须给不愿参加研究的学生提供其他获得学分的途径,参加其他替代性活动所需的时间和精力都应该和参与研究所需的均等。常用的替代任务包括阅读或总结期刊上的研究论文,进行非正式领域的行为观察,参加研究生或老师研究结果的报告,自愿进行社区服务等(Kimmel, 1996)。

机构审查委员会要求研究者证明以人类为被试的研究都遵循了正当的知情同意程序。我们在本书上一个版本里提供了一个用于不含欺骗的最小风险研究的同意样表。然而,认可联邦人类研究保护办公室确定的方针是很重要的:“知情同意是一个过程,不仅仅是一种形式。”一位机构审查委员会主席曾说过,她让研究者想像与被试坐在一起并解释该研究。因此,我们在新版里提供一些获得恰当知

专栏 3.1

获得知情同意的提示

恰当的知情同意应明确表明研究目的或问题,研究者本人及助手身份,遵循的程序,参与研究的风险收益比,报偿(如果有的话),被试的代价(如果有的话),研究参与的选择余地,保密程序,自愿参与并可随时退出而不予惩罚,问题或疑问的解答途径,研究者及被试签名,研究日期。有时机构审查委员会会根据情况附加其他要求。联邦人类研究保护办公室(OHRP)出版了可供研究者参考的指南。以下为OHRP指南节选,全文及相关文献可登陆网页获得:<http://www.hhs.gov/ohrp/humansubjects/guidance/ictips.htm>

- 知情同意书应使用易于理解的语言,避免使用科学术语或专业术语。
- 避免使用第一人称(如我明白……或我同意……),因其会被理解为带有暗示性或被错误的用来取代充足的客观信息。如:“如果同意参加,您需要做下列事情”会更妥当些。“将文件主要当作是一种教学工具而非法律手段。”
- 描述可能经历的所有体验(例如,这个实验是如何进

行的)以及可预见的伤害、不适、不便和风险。

- 向被试说明参与研究的收益。如果研究通常仅具有社会及科学公益性,也应予以说明。
- 描述可供被试做出的其他任何选择。如果从大学生中选择被试,必须向他们解释参与心理学研究以外的其他可替代性途径。
- 必须告知被试个人信息信息保密的方式。搜集高度敏感信息时,机构审查委员会或许会要求保密证书等额外保护措施。
- 研究中相关伤害超过最小风险时,必须对自愿参加的报偿及待遇予以解释。
- 不能忽视被试的合法权利。
- 必须指定熟悉研究的“联系人”,以便解答被试研究后所提出的问题。在以下三个领域随时都会出现问题,所以在知情阶段必须予以明确阐述:研究体验、被试权利、相关伤害。有时需要不止一个联系人,也可以指导被试向机构审查委员会或机构成员请教。
- 必须包含自愿参与的声明,以强调被试无论何时退出研究均不会受到惩罚。

情同意的提示(见专栏3.1)并省略了样表,因为样表有可能被误解为“适用于所有情况。”恰当的同意程序和书面文件随被试人群和研究情境的不同而不同。关于怎样以符合道德方针及保护被试权利的方式获取或者证明知情同意,机构审查委员会的建议是最佳的资源。

一些情境不要求研究者获取知情同意。最明显的例子就是研究者在公共场合观察个体行为而不加任何干预。例如,一位研究者通过观察相同和不同种族群体穿越校园的频率来搜集大学校园内种族关系的证据。研究者不必在观察前征求学生的许可,但如果要记录特殊个体时则需要知情同意。

判定什么时候行为是公开的或隐私的并非易事。迪恩纳和克兰德尔(Diener & Crandall, 1978)提出三个主要维度,研究者可借以参考来帮助判定隐私信息:信息敏感性、情境敏感性和信息传播方式。很明显,一些信息比另一些信息更具敏感性。就性习惯、宗教信仰或犯罪行为进行的访谈与那些就谁会赢得世界杯进行的访谈相比,接受访谈的个体更关注访谈信息将如何使用。

情境对判定公开行为和隐私行为也起了一定的作用。比如参加音乐会等行为是公开的。在公共场所中,人们会暴露一定程度的隐私,因此很难区分发生在公共场所的一些行为是公开的还是隐私的。当你驾驶自己的汽车、使用公共浴室或在公园享受家庭野炊时,这些行为是公开的还是隐私的?在“聊天室”内交流是公开的还是隐私的?此种情境中,对研究行为的道德判定取决于所搜集的数据的敏感性及使用信息的方式。

当信息以平均方式或按比例方式传播时,就不会反映出具体的个体信息。其他情境中,可以利用编码系统保护被试隐私。未经个体或小组许可,传播其敏感信息是严重违背道德的。在个体不知晓的情况下(例如,由隐匿的研究者进行观察)搜集可能敏感的信息,研究者可以事后与个体取得联系,并询问是否可以使用该信息。若个体不允许,则研究者就不能使用该信息。关于隐私权最难判定的情况是在一个维度上存在明显道德问题但另两个维度上却不存在,或者三个维度

图 3.5 判定公开行为和隐私行为并非易事



拓展练习

《APA 伦理准则》说明，如果研究属于自然观察，心理学者可以忽略知情同意（见标准 8.05）。但正如我们所见，判定公共场所的观察是否“自然”并非易事。思考下列研究，说明并判定研究者在研究前是否应取得被试的知情同意。如果你想从研究者那里获得更多信息，那么，在决定是否需要知情同意前还需要哪些附加信息呢？你会发现，知情同意对研究情境产生极大影响，例如，会加大研究者在“自然”情境中行为记录的难度。这就是道德决策的两难问题。

1. 在大学生饮酒行为研究中，一位本科生实验助手参加同学聚会并记录聚会上其他学生的饮酒量。

2. 作为同性恋群体研究的一部分，一位同性恋研究者参加了同性恋棒球队，目的是记录赛季球队竞争情境中被试的行为。所有比赛均在城市娱乐联盟间进行，有很多普通观众。

3. 由隐藏在公共浴室中的男女观察员分别观察公共浴室的男人和女人的行为，如便后冲水、洗手、乱丢垃圾、乱涂乱画等。

4. 一位研究生要调查大学生的作弊行为。在礼堂内以大班级为单位组织考试时，他藏在投影仪后面，借助有利位置及望远镜，可以看到多数学生的行动。该研究者记录大学生的头部活动、交换试卷、传递纸条及其他的可疑行为。

上都存在一点道德问题。例如，电影院阴暗情境中个体行为易产生敏感性信息，但是这种情境被归类为公开的。

隐私权（privacy）指个体拥有的决定有关本人信息是否可以被传播的权利。《APA 伦理准则》明确规定：“心理学者有责任采取合理预防措施来保护被试的隐私信息。”（APA, 2002, p.1006）。无论何时都应向参与心理研究的被试说明保密的方式，这样被试可以自行判断隐私保密措施是否恰当。履行知情同意要求研究者在研究人类行为的需求和人类被试的权利之间寻求一种平衡。

心理学研究中的欺骗

- 心理学研究的欺骗指的是研究者故意隐瞒信息或有意误导被试对实验的理解和认识。就其性质而言，欺骗违背了知情同意的道德原则，但在心理学某些领域，这却被认为是必要的研究策略。

研究中最具争议的道德问题就是欺骗。有些人极力主张无论何时都不应该欺骗被试，因为伦理实践要求，研究者和被试之间的关系应是公开和诚实的。有些人认为欺骗是道德所不容的，它无异于撒谎。**欺骗（deception）**可以通过省略和伪装的手段，即隐瞒信息和就研究某方面有意误导被试。任何一种欺骗都与知情同意的原则相抵触。尽管近几十年来，研究者对欺骗的关注逐渐升温，但心理学研究中欺骗手段的使用并没有减少，它依然是一种很受欢迎的研究策略（Sharpe, Adair, & Roses, 1992）。斯基卡和瑟吉斯（Skitka & Sargis, 2005）调查了一些以网络为数据搜集工具的社会心理学者，发现 27% 的研究对网络被试存在欺骗。

图 3.6 20 世纪 60 年代, 在斯坦利·米尔格拉姆 (Stanley Milgram) 实验中, 没有告知被试实验目的是观察人们对权威的服从, 结果很多人遵循研究者的指示对他人实施了强烈的电击



APA 伦理标准

8.07 研究中的欺骗

(a) 心理学者不可实施带有欺骗的研究, 除非他们能证明研究具有重大的科学意义、具有教育或实用价值, 而且不采取欺骗程序就没法进行此项研究。

(b) 当研究会带来生理疼痛或极度情感焦虑时, 心理学者不得欺骗被试。

(c) 心理学者应及早向被试解释欺骗是实验设计及实施的要求, 解释时间最好在他們刚参加完实验时, 但不可迟于整理数据时。一旦被试不愿意, 他们也可以在那时收回数据 (还可参阅标准 8.08, 事后解释)。

注 解

- 欺骗是一些心理研究中必要的研究策略。
- 为达到使被试参加研究的目的而欺骗他们是不道德的。
- 研究者考虑使用欺骗手段时, 必须审慎地衡量欺骗的代价与研究可能获得的收益。
- 研究者有责任尽早向被试解释欺骗手段的使用。

如果不向被试隐瞒某些信息, 有些研究就无法进行。还有一种情况下, 为了使被试采取某种态度和行为, 必须对他们进行误导。例如, 卡森和卡基尔 (Kassin & Kiechel, 1996) 想知道哪些因素会使人们违心承认实际上并没有做的事情, 他们研究的目的是了解什么因素导致犯罪嫌疑人假意承认犯罪。在实验中, 被试的任务是在键盘上敲击他们听到的字母, 他们告诉被试在敲击时不能碰到 “Alt” 键, 因为那样会损坏电脑。一段时间过后, 电脑出现故障, 研究者指责被试敲击了 “Alt” 键。尽管没有被试敲打过 “Alt” 键, 但将近 70% 的被试在书面材料上签字, 承认他们敲击了。如果事先告知被试该程序的目的是诱发其违心的承认, 那么他们或许就不会承认。知情同意公开化的话, 就不可能进行违心承认的研究。尽管有时使用欺骗手段可以调查一些重要问题, 但是为了让被试参加超过最小风

险的研究而欺骗是不合道德的。正如道德准则中所阐述的：“当研究会带来生理疼痛或极度情感焦虑时，心理学者不得欺骗被试”（标准 8.07b）。

8.07 研究的目的是观察人们的正常行为。之所以使用欺骗手段，有这样一个基本假设：有时隐瞒实验的真实性质能够使被试的行为像正常情况下一样，否则其行为会与实验者的指示一致。然而经常和随意使用欺骗手段就会引发问题（Kalman, 1967）。如果人们认为研究者经常误导被试，他们在参加心理学实验时就会预期自己被欺骗，被试对研究的怀疑会妨碍其表现出正常行为（见专栏 3.2），这恰恰与研究者的初衷背道而驰。有意思的是，埃普利和赫夫（Epley & Huff, 1998）在实验后的事后解释中，直接对比了告知欺骗的和未告知欺骗的两组被试的反应。结果，告知欺骗的被试比没有意识到被欺骗的被试对心理学研究持有更深的怀疑态度。随着在线研究的增多，研究者应特别注意研究中欺骗的使用。因为欺骗可能不仅加剧社会成员对研究者的不信任，而且还会成为“毒害”人们获得社会支持和彼此联系的一种手段（如网络）（Skitka & Sargis, 2005）。

凯尔曼（Kalman, 1972）主张研究者在使用欺骗手段前必须认真考虑：（1）研究对科学的重要性，（2）可替代的、无欺骗手段的可行性，（3）欺骗的危害性。最后一项指欺骗的程度及对被试可能的伤害。凯尔曼（p.997）认为：“只有非常重要的研究，且没有可行的替代方法时，超过最小等级的欺骗才可能是合理的。”使用欺骗手段时，研究者必须在实验后的事后解释中告知被试使用欺骗手段的原因，讨论被试可能具有的错误观念，消除欺骗所产生的任何有害影响。事后解释的目的之一就是使被试明白欺骗的必要性。卡森和凯基尔（1996）在研究违心承认的实验中，被试在事后解释中获悉了使用欺骗手段的原由。被试报告说：他们认为研究是有意义的，而且他们的贡献是有价值的。

事后解释

- 实验结束后，研究者仍担负着使被试受益的道德责任。实现此目标的最佳途径之一就是给被试提供完整的事后解释。

数年来，很多研究者陷入了一种怪圈，认为被试只是达到目的的一种手段，被试似乎是研究者获得数据并达到实验目的的“客体”。有时，研究者认为搜集完最后数据，他们对被试的责任也随之结束了，握一下手或者说一句“谢谢”标志着研究阶段的结束。这样做被试可能会带着对研究情境的种种疑虑和自己在研究中所起的作用迷惑地离开。计划或实施研究时，考虑研究结束后对被试造成的影响是很重要的，应该时刻寻找被试参加研究所能获得益处的方式。这些问题遵循道德准则中的两项道德原则，即善行和尊重人类权利与尊严。

专栏 3.2

欺骗还是不欺骗：这是个难题

在心理学研究中研究者一直在使用欺骗手段 (Sieber, Lannuzzo, & Rodriguez, 1995), 科学界对使用欺骗手段的争论也一直没有停止 (Broder, 1998; Fisher & Fryberg, 1994; Ortmann & Hertwig, 1997)。参与辩论的人有时对欺骗的界定也感到困惑, 所以这是个很复杂的问题 (见 Ortmann & Hertwig, 1998)。费舍和费赖伯格 (Fisher & Fryberg, 1994) 将争论总结如下: “关于是否应该使用欺骗手段的道德争论, 一直围绕潜在的社会收益、是否违背了善意的原则, 对个体的尊重及心理学者对被试负有的受托责任而展开” (p.417)。此话举足轻重, 让我们试着将其分解开来。

道德原则的“善意”指研究应该为个体及社会带来益处。如果欺骗手段有害于个体或社会, 则该研究的“善意”遭到质疑。“对个体的尊重”原则 (应将被试当作“人”来对待而非研究的“客体”) 表明, 人们有权对其参与的研究的程序和目的做出自己的判断 (Fisher & Fryberg, 1994)。心理学者的“受托责任”指被信任者应对他人负责, 哪怕是暂时性的。心理学研究中, 研究者要对被试安全负责并对其参与实验的后果负责。

鲍姆林德 (Baumrind, 1985) 的论点可以很好地解释这些思想和原则。她声称: “研究情境中欺骗手段的使用从科学上来讲是不道德的、不谨慎的和无保证的” (p.165)。她认为欺骗对于被试、心理学专业和社会来讲代价太大, 因而无法保证继续应用。她的观点既长又复杂, 我们在这里仅进行简要概括。首先, 根据鲍姆林德的观点, 欺骗被试付出了代价, 因为它削弱了被试对自己的判断及对“受托者” (获得他人信任者) 的信任。鲍姆林德坚信, 被试发现自己被欺骗后, 会对自身现有的了解及曾经确信能为其提供有效信息和忠告的个体 (如社会科学家) 产生怀疑。心理学专业所付出的代价是指被试 (甚至社会) 很快就意识到心理学者是欺骗者, 并且不再相信心理学者提供的指导语。正如凯尔曼 (Kelman, 1972) 提出的观点, 如果被试怀疑研究者撒谎, 那么欺骗能否起到研究者所预期的作用就会受到质

疑。鲍姆林德还认为使用欺骗表明研究者有意撒谎, 这似乎与寻求真理的目标相矛盾。最后, 对社会的危害是指欺骗削弱了人们对专家的信任, 使他们怀疑所有的发明创造。

当然, 并非所有心理学家都持这种观点 (Christensen, 1988; Kimmel, 1998)。例如, 米尔格拉姆 (Milgram, 1977) 认为, 心理学家的欺骗做法其实只是一种“技术手段”, 为了科学的需要应予以许可。毕竟, 有时现实生活情境中, 为了让人们相信某事也会制造错觉。人们收听广播节目时, 一般不会被雷声或者马蹄声所干扰, 这仅仅是声音效应专家所创造出来的技术上的错觉。米尔格拉姆主张在科学需求情况下应允许技术错觉。我们欺骗孩子去相信圣诞老人, 科学家为何就不能制造错觉来帮助他们了解人类行为呢?

米尔格拉姆指出, 正如在实际生活中经常制造错觉一样, 在一些情境中, 可以暂时放弃一般的道德原则。如果获悉一项罪行, 从道德上讲我们定会向当局汇报。而另一方面, 律师肯定会偏向于委托人提供的信息, 哪怕该信息证明委托人是负有罪的。外科医生会对我们的身体进行检查, 尽管其行为在医务室内是道德允许的范围, 但同样的行为在医务室外就不可以。米尔格拉姆认为, 为了科学的利益, 有时应该允许心理学者暂时放弃真实性和诚实性的道德原则。

为欺骗手段辩护的人们发现, 被试没有因为受到欺骗而产生消极反应 (如, Christensen, 1988; Epley & Huff, 1998; Kimmel, 1996)。尽管人们对心理学研究的“怀疑”会增加, 但总体影响并不大 (见 Kimmel, 1998)。金梅尔 (1998) 对欺骗的底线进行了很好的总结: “若绝对禁止在所有心理学研究中使用欺骗手段, 会妨碍研究者实施大量重要研究, 从而导致严重后果。” (p.805)。科学界内没有人认为可以轻率采用欺骗做法, 然而, 对于很多科学家来讲, 不具备该类研究的经验比使用欺骗手段更有害。

你认为心理学研究中应该使用欺骗吗?

APA 伦理标准

8.08 事后解释

(a) 心理学者应提供机会让被试了解研究的性质、结果及结论等信息，并采取合理措施来纠正他们所意识到的被试可能会有错误的观念。

(b) 如果科学或人道价值方面的考虑要求推迟或取消事后解释，心理学者也应采取合理措施来降低危害的风险性。

(c) 心理学者意识到研究程序会伤害被试时，他们应采取合理步骤使伤害最小化。

注 解

- 事后解释既有益于被试又有益于研究者。
- 事后解释使被试了解研究性质、自身在研究中的作用及研究过程。事后解释首要的目的是使被试对自己的参与过程感觉良好。
- 事后解释使研究者获悉被试对研究过程的观点，从而对研究结果更具洞察力并对未来研究产生新思路。

前面我们讨论了保护被试隐私权既有利于被试（保护被试免受社会伤害），又有利于研究者（如提高被试如实反应的可能性）。同样，研究结束后的事后解释（debriefing）对被试和研究者也都有益（Blanck et al., 1992）。若研究中使用了欺骗手段，就需要在事后解释中向被试解释欺骗的必要性，纠正被试可能具有的对于参与实验的错误观念，消除欺骗所产生的有害影响。事后解释还有助于被试理解研究过程（理论基础、方法、结果），并能够让被试对参与研究有积极体验。研究者应该给被试提供机会，使他们更多地了解自己对所做的特殊贡献，并更大程度体会到自己的参与感。例如，研究者可以在研究结束后给被试发电子邮件报告研究结果。在线研究结束后，研究者可以在网站上发出事后解释，有新结果时及时更新资料内容（见Kraut et al., 2004）。

事后解释使被试从总体上更好地了解研究。例如，被试可以明白研究中的个体成绩并非其能力的直接测量。再如，他们在记忆测验中成绩的好坏不仅受其记忆能力的影响，还受研究者提供的记忆材料及施测方式影响。心理学入门课的大量学生因研究的教育意义而自愿参加，所以研究者有责任保证参与研究对大学生来讲是具有教育意义的体验。有时老师在课堂上就给学生打下了有关事后解释的基础，他们让学生书面简要概述研究目的、所使用的技术及该研究对了解行为的重要性。对该做法的评价结果表明：写过报告的学生比没写过报告的学生具有更满意的研究体验，并获得了更大的教育收益（Richardson, Pegalis, & Britton, 1992）。

事后解释可帮助研究者了解被试对研究过程的想法。研究者希望弄清被试是否以研究者所预期的方式来看待研究程序（Blanck et al., 1992）。假如有一项关于人们面对失败如何反应的研究，其中必须设计根本不可能完成的任务，然而，如

图 3.7 事后解释对于确保被试对参与研究产生愉悦体验很关键



果被试不把自己的表现看做失败，研究者的假设就无从验证。事后解释可以使研究者弄清被试到底认为他们失败了还是认为他们缺乏运气。

在试图了解被试对研究的看法时，研究者不应该急功近利。总地来讲，被试希望有助于科学进步，他们或许知道在心理学研究中研究者隐瞒了某些信息，他们甚至担心如果暴露自己的确不了解研究的重要细节，就会“破坏”该研究（例如，实际上不可能的任务）。为避免这种可能存在的问题，事后解释应该是非正式的、非直接的。最佳解决途径就是设计开放性问题（例如，你认为这是一项关于什么的研究？你觉得你在研究中的体验怎么样？）。研究者可以再加上关于研究过程的更具体的问题，这些具体问题尽可能避免任何暗示，以免被试按期望进行反应（Orne, 1962）。

事后解释还因“指引未来研究并确认当前研究中存在的问题”而利于研究者（Blanck et al., 1992, p.962）。换句话说，事后解释可以对被试的行为原因提供线索，帮助研究者更好地对结果进行解释。研究者在事后解释中了解的内容可以给他们提供未来研究的思路。最后，被试有时会觉察出研究材料中的错误——例如，有遗漏信息或者指导语含糊不清——他们可以在事后解释中报告给研究者。正如我们所说过的，事后解释既有益于被试又有益于研究者。在线研究情境中研究者不在现场，使得合适的事后解释过程有些困难，从而更增加了网络研究的道德难题（Kraut et al., 2004）。在线被试可以很容易地随时从研究中退出，这也是极其令人头疼的问题。一种建议就是，将实验程序设计为被试一旦草率地关闭窗口就自动呈现事后解释主页（Nosek et al., 2002）。

研究者的教育责任并不局限于给被试提供有效的事后解释，科学的目标是改变并且改善人类生活条件，科学家有责任公开研究问题，科学家要担当该责任就会面临道德挑战。心理学家对社会关注的重要问题进行研究，如歧视、暴力行为、儿童性侵犯等。社会焦点问题常常是持不同政治观点的人们激烈争论的焦点。将研究结果向媒体公开的研究者可能要承受反对派的批评甚至敌对反应。美国心理

协会 (APS) 正竭力让社会了解心理学研究对社会关注问题所作出的建设性贡献。APS 针对每个社会关注的重要问题, 委派该研究领域的专家组成研究小组, 这些研究小组得出的结论和提出的建议发表在《公众兴趣心理科学》(*Psychological Science in the Public Interest*) 上。这些问题包括班级人数对学生成绩的影响、暴力预测以及癌症诊断等心理科学。

用动物做研究

每年实验室研究要测试数百万只动物以探究大量重要问题。新药物在投入使用前要先在动物身上进行测试; 排放到环境中的物质要先通过动物测试来确定其影响; 各种治疗手段也需要在生病的动物身上实验; 新的外科手术——尤其是关于脑部的——经常先在动物身上进行尝试。行为研究中也用到很多动物, 例如, 生态学家和实验心理学家的研究。压力和糖尿病之间关系的动物模型帮助研究者理解了糖尿病的心理影响因素 (Surwit & Williams, 1996)。这些研究的结果对人类的健康有很大贡献 (Miller, 1985), 但是在研究进程中, 很多动物要承受疼痛和不适、压力, 甚至面对疾病和死亡。啮齿类动物, 尤其鼠类是实验室动物研究中的最大群体。研究者在其实验中也会使用猴子、鱼、狗和猫等各种动物。有些特定物种经常被选用的原因是因为他们能够提供良好模型来说明人类的反应, 例如对听觉感兴趣的心理学者有时用南美栗鼠作被试, 因为栗鼠的听觉过程与人类非常类似。

APA 伦理标准

8.09 研究中的动物使用和人道关怀

(a) 心理学者获取、照顾、使用及处置动物都要符合现有的联邦、州或地方的法律、法规及专业标准的要求。

(b) 受过研究方法培训并有动物实验经验的心理学者, 应该对动物研究的所有过程进行督导, 并保证使动物得到舒适、健康和人道的对待。

(c) 心理学者确保给予其督导下的所有动物研究者有关研究方法、动物照顾、动物保养等方面的指导, 以促使研究者尽其责任。(还可参阅标准 2.05, 对他人的工作授权)

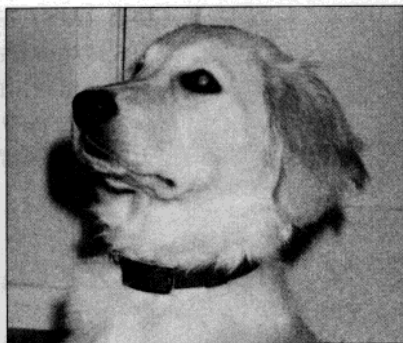
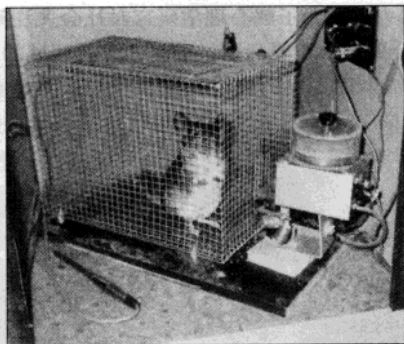
(d) 心理学者应努力使受试动物的不适、感染、疾病和疼痛最小化。

(e) 心理学者只有在没有可替代程序, 而且证明研究具有可预期的科学的、教育的或实用的价值时才可应用给动物带来疼痛、有压力、或剥夺的实验程序。

(f) 心理学者实施外科程序前要对动物进行适当麻醉, 遵循技术要求, 避免传染并在术中及术后使疼痛最小化。

(g) 需要终止动物生命时, 心理学者动作要迅速, 努力使疼痛最小化, 并符合公认程序。

图 3.8 关于动物研究的伦理指导说明了在研究前、研究中、研究后应怎样对待动物



注 解

- 研究中使用动物是为了获得有益于人类的知识。例如，动物研究可以帮助人类治愈某些疾病。
- 在道德上，研究者必须遵循联邦、州和地方的现行法律、法规以及专业标准要求来获得、照顾、使用和处理动物。
- 在研究中使用动物会涉及复杂的问题，使用动物也是一个受争议的话题。

人们通常认为，把动物作为实验被试是理所当然的。实际上，为证明使用动物进行实验的合理性，人们有时会引用圣经中提到的人类对所有生物的“统治”（Johnson, 1990）。然而更多时候，我们需要获得知识而不使人类处于危险中，这种需要证明了在研究中使用动物被试是合理的。大部分药剂、药品、疫苗或疗法都是通过动物试验研究出来的（Rosenfeld, 1981）。马埃斯特里比得里和卡罗尔（Maestriperieri & Carroll, 1998）也指出，通过研究自然条件下猴子虐待幼猴的现象，科学家们可理解人类儿童的受虐和被忽视的问题。

然而，实验室研究中，使用动物被试引发了很多问题（Novak, 1991; Shapiro, 1998; Ulrich, 1991）。其中最根本的问题是在心理研究中是否应该使用动物，此外还包括一些照顾和保护动物被试的问题。根据 APA 伦理规范章程，使用动物被试的研究者在获得、照看、使用和处理动物上负有道德义务，即必须遵循联邦、州和地方的现行法律、法规以及专业标准要求。使用动物进行研究是一个被高度监控的事情，这个事情的最高目标是保护研究中动物的安全。只有当个体具备研究、管理和照顾某种动物的资格时，才能使用这种动物进行研究。如果研究者使动物疼痛或不适，那么就必须用研究可能具有教育或实用的意义来证明研究中使用动物的合理性。如我们前面提到的，动物审查组织（如动物实验管理委员会）作为研究机构向公共卫生组织申请资金，这些委员会决定降低疼痛、执行安乐死、为动物提供住所以及训练人员的适当程序。动物实验管理委员会还判定实验设计对获得重要的新信息是否有效、使用动物的模式是否恰当，或能否使用替代性的无动物的研究模式（Holden, 1987）。

20 世纪 80 年代，研究者们必须满足许多联邦、州以及地方的要求，包括自美

国农业部的兽医对动物研究室的检查（见 National Research Council, 1996），这样做的部分原因是为了响应动物权利团体成员的相关声明，这样的规定常受到科学团体成员和许多隶属于保护实验室动物团体的动物研究者们的欢迎。APA 已制定了一套心理学研究中使用动物被试时须遵循的特殊方针，这些方针在 www.apa.org/science/anguide.html 中可以看到。

然而，由于这个问题仍很敏感，因此在使用动物进行研究时必须做出一些妥协。例如，在发现替代动物研究的方法前，人类为战胜疾病和苦难需要使用动物进行研究，但必须保护实验室研究中动物的安全（Goodall, 1987）。就像 APA 前任首席执行官雷蒙·福勒（Raymond Fowler）指出的那样，即使研究价值没有立即显现，也不能禁止使用动物被试，这一点很重要（Fowler, 1992）。“关于动物研究没有价值的指控是因为没与潜在的实用价值联系起来看，这是针对所有基础研究的指控。”这种对心理学的指控“威胁着理性和科学基础”，同时威胁着“科学家和技术工作者”。

为防止动物遭受不必要的痛苦，限制是必需的。虽然极少有科学家否认这点，然而很多科学家都希望避免流于形式的限制。因为有时候限制导致的高额费用会毁了研究。芬尼（Feeney, 1987）认为，严格的限制和高额的费用，以及动物行为组织中的激进分子指向个体和研究机构的负面宣传（以及偶发的带感情色彩的例证），可能会阻止青年科学家们进入动物研究的领域。如果这种情况真的发生的话，（当前）不可治愈的疾病或永久性的瘫痪很可能攻克无望，因为这些疾病的治愈依赖于科学研究。关于动物研究结果运用于人类的适用性的争论很多且很复

专栏 3.3

人类和非人类动物的道德地位如何

道德决策常使相反的哲学观点针锋相对，研究中是否可以使用动物的争论就在证明这一点。争论的焦点是人类和非人类动物的“道德地位”。澳大利亚哲学家彼得·辛格（Peter Singer, 1990, p.9）指出，两个普遍接受的道德准则是：

1. 所有人在道德地位上都是平等的。
2. 所有人在道德地位上都高于非人类动物。

据此，辛格继续指出，“根据这些原则，我们普遍认可人类的利益应该高于非人类动物遭受的痛苦；这种假设反映在很多领域中人类对待动物的方式上，包括农业、狩猎、实验和娱乐等方面”（p.9）。

然而，辛格不同意这些普遍接受的观点。他认为

“没有理由让非人类动物承受人类的苦楚”（p.9）。按照辛格的观点，“作为人类”没有特殊的道德地位，除非我们诉诸宗教观点（辛格拒绝把这一点作为多元社会中作决定的基础）。这种立场源于实用主义的哲学传统，这始于大卫·休谟（1711~1776）和杰里米·边沁（1748~1832）以及约翰·斯图尔特·穆勒（1806~1873）的著作（Rachels, 1986）。总地来讲，这种观点认为，只要我们有选择，我们应选择能使所有个体获得最好结果（产生最大“幸福”）的操作方式。这种观点的实质是正被讨论的个体是否体验到快乐/不快乐、舒适/疼痛，而与个体是人类还是非人类无关（Rachels, 1986）。

你怎样考虑人类和动物的道德地位？它与心理学研究有什么关系？

杂(见专栏 3.3)。乌尔里希(Ulrich, 1992)说得好,有关这些问题的讨论必须通过“智慧和权衡”来处理(p.386)。

心理学研究的报告

- 研究者通过同行评论的科学期刊交流他们的研究结果,《APA 伦理准则》为该程序提供了指导。

当研究负责人将研究初稿交给心理学领域的科学期刊时,一个已完成的实验研究就开始了它成为科学文献的历程。在心理学期刊上发表研究的主要目的是同科学团体和社会大众交流自己的研究。在期刊上发表研究是提高研究者声望的一种途径,甚至是提高此项研究的赞助机构声誉的一种途径。但是科学研究结果的发表并非易事,尤其是研究者想发表在一个享有更高声望的科学期刊上时。APA 赞助的期刊,比如《心理咨询期刊》、《教育心理学期刊》、《实验心理学期刊:学习、记忆和认知》和《心理学评论》,它们的平均拒稿率为 71%,最高拒稿率高达 90%(APA, 2001)。

提交的发表初稿必须按《APA 出版手册》的格式要求撰写,要遵循严格的方法和标准,并且必须合乎所投稿期刊对文稿的要求。如果上述三方面的某一方面出现严重问题,就会被退稿。但是下列情况更容易被退稿:研究方法出现问题、统计处理出现问题,或该研究对该期刊相关领域没有明显贡献。初稿的可接受性由期刊编辑决定,这种决定通常以同行评论为基础。同行评论是指该领域专家的评价,这些专家应编辑邀请评价文稿的科学价值。研究者为了获得专业发展而发表研究结果,面对这些压力,所有科学家都要严格遵守《APA 伦理准则》,这是强制性的。

APA 伦理标准

8.10 研究结果报告

(a) 心理学者不得伪造数据。(另见标准 5.01a, 避免虚假或欺骗性陈述。)

(b) 如果心理学者在他们发表的数据中发现明显的错误,他们要通过合理的步骤,采用修正、收回、勘误或其他适当的方式来更正错误。

8.11 学术剽窃 即使他人的作品和数据可以间或被引用,心理学者也不得将他人的研究或数据中的任何部分作为自己的成果展示或提交发展

8.12 发表署名

(a) 心理学者仅对他们真正从事的或做出过充分贡献的研究承担责任和拥有署名权,包括作者署名权。(另见标准 8.12b, 发表署名。)

(b) 主要作者和其他参与人员的署名准确地反映了每个个体对相关科学或专业的贡献,这与他们的相对地位无关。仅仅拥有一个机构内的地位——如部门主

席——不能证明是原创作者。可适当地对为研究和论文撰写做过较少贡献的人表示答谢,如在脚注或介绍性的综述中答谢。

(c) 除特殊情况外,任何主要基于学生博士论文的多作者论文,都要将该学生列为主要作者。指导老师应尽可能早地和学生讨论发表署名问题,此讨论还应贯穿在整个研究和发表阶段。(另见标准 8.12,发表署名。)

8.13 复制发表数据 心理学者不得将已发表过的数据再次作为原始数据发表。也不得再次发表已得到正式认可的数据。

8.14 分享研究数据以作为验证

(a) 研究结果发表后,当其他有能力的专业人员需通过再分析来验证其主要观点,且仅为该目的需使用这些数据时,假如被试的隐私能得到保护,而且法律没有不允许其发表,心理学者不得保留得出结论的数据。这不排除心理学者要求个人或团体为使用数据而付费。

(b) 心理学者从其他心理学家那里索求数据来证明自己的观点,必须经过数据的重新分析,且只能用于澄清观点的目的。索求数据的心理学者须事先获得书面同意后才可以使用。

8.15 专业审稿人 审阅其他人提交来的陈述、出版物、认证材料或研究建议的心理学者,应尊重提交者的隐私和其材料的所有权。

注 解

- 发表署名根据每个人在学术上所做贡献的重要性决定。
- 出于道德要求,研究报告必须标明引用和参考文献以识别他人的贡献,否则可能犯剽窃的错误。
- 直接引用原作者材料,或未查阅到原始材料而引用的第二手材料,均须使用引用标记标明。

当科学研究完成后,研究者通常会把初稿提交给某个专业期刊。报告研究结果的标准似乎比我们先前讨论的道德标准更明确。然而,即便如此,关于如何安排署名和学术剽窃等问题的伦理决策并不总是十分明确。一项研究工作的顺利完成通常会涉及很多人,同事们为研究设计提供建议,研究生或大学生帮助研究者测试被试和整理数据,工程师们建造专门的仪器,专家给予统计分析建议。当为发表准备初稿时,是否这些人都应是这项研究的“作者”?发表署名过程中需要确认为该研究作出过显著贡献的哪些人可作为作者。因为发表科学研究的作者身份经常被用来衡量个体在某科学领域的能力和声望,所以,公正地承认那些为研究作出贡献的人很重要。

一个人对研究的贡献可保证他成为研究论文的“作者”?还是以相对不重要的方式(如在脚注里)承认他的贡献?这种决定总是不那么容易。另外,一旦被认可为作者,那么就应决定作者的署名顺序。一篇多作者的文章中,“第一作者”通常意味着其贡献比“第二作者”大(“第二作者”比“第三作者”贡献大,依此

类推)。作者排序主要根据其学术贡献重要性(例如,在研究概念方面有帮助),而不是根据在研究中投入的时间和精力来决定(如, Fine & Kurdek, 1993)。

与作者署名相关的道德问题有很多形式。例如,不仅系里的教员把学生的研究据为己有是不道德的,给学生不应得的作者署名也是不道德的。例如,导师给其学生一个不应得的研究论文作者署名,会产生一种误导,使该学生在争取一个竞争性的研究生项目中占据优势。凡恩和科德克(Fine & Kurdek)认为,给予学生不应得的作者署名,可能错误地评价学生的专业水平,让该学生得到相对同辈来说不公平的优势,也可能导致其他人对其产生过高的期望。他们还认为,所有成员和学生应共同决定作者署名,在研究早期就应该讨论何种程度的参与会被列入作者名单。由于参与学生的能力和地位不同,教研组全体成员应该主动发起关于作者署名的讨论(见 Behnke, 2003)。

还有一个需要补充的与署名有关的道德问题。如前所述,研究者从被试那里获得信息,必须保持其机密性以保护被试的隐私权。当作者形成一个研究构想时,他同样也有隐私权。审稿人对出自发表或获得研究资助目的而提交的初稿进行审查,审稿人必须尊重作者的隐私权,要保持初稿信息的机密性。

对专业人员和学生来说,与研究报告有关的相当麻烦的一个问题就是学术剽窃(plagiarism)。道德标准似乎已经很明了:不得将他人作品的重要内容或要素当作自己的成果炫耀或发表。但是“重要内容或要素”由什么构成,一个人怎样才能避免剽窃他人的成果?做这样的决定就像走钢丝,一方面是因学术贡献而被认可的个人目标;另一方面是认可前人贡献的道德责任。专业人员和学生都会犯学术剽窃错误的事实说明,许多人经常为追求自己的荣誉而不在著作中给他人应得的署名,因而失去了在钢丝上行走的平衡。

有时,学术剽窃行为是因为马虎造成的(例如,没有查清原始资料来源)。这种错误仍然属于学术剽窃;无知不是合法的借口。所有人都会犯错。例如,研究者(和学生)偶尔会问,引用“多少”文章内容才可以不用引用标记或标明来源。文章的重要元素可能是一个单词或短语,只要该元素能被证明是作者思想成果的关键观点或概念。由于没有明确的规章规定著作的重要元素由多少材料构成,故而学生在引用他人著作时要特别小心。很多时候——特别对学生来说——剽窃是由于直接从原始材料中引用时未使用引用标记造成的。无论何时引用原始材料,都必须用引用标记作为标识,而且要注明出处。在论文中解释材料时要列出文章中材料的出处,这也很重要。伦理准则规定,当使用确切的词语以及解释的时候,必须列出观点的来源。

引用第二手材料时不标明出处也会产生学术剽窃。第二手材料是指讨论他人(原创的)的研究文献。第二手材料包括教科书以及发表在类似《心理学通报》等科学杂志上的研究评论。如果你的观点或研究结果仅来源于第二手材料,那么从要求查阅原始著作上来说,发表这些信息是不符合职业道德的。查找和阅读原始材料远好于引用第二手材料。如果没办法这样做,当涉及原始著作的时候,你必须用类似“如在……中引用的”这样的短语,告知读者你没有阅读原始资料(见第14章关于第二手材料的引用的例子。)通过引用第二手材料,你告诉读者你在呈现另一个人对原始材料的解释。忽略引用的适当形式同样不是一个可以接受的

借口,有时很不幸,研究者——包括教授和学生——眼看着自己的事业因被指责为学术剽窃而毁于一旦。

伦理决策的步骤

为了搜集冷漠的旁观者信息,就应该置被试于严重风险中吗?心理学者能否使用欺骗手段来了解违心承认的原因?为了了解人类药物成瘾,就应该让动物受到伤害吗?这些问题需要答案,然而,到现在为止,你可能已经知道回答这些问题并不容易。什么是“正确的”答案,甚至是否有正确答案也是不清楚的。

面对这种不确定性,研究者和机构审查委员会及动物实验管理委员会可以通过全面的计划而做出伦理决策。按照一系列步骤,研究者可以批判性地思考伦理问题。批判性思考伦理问题将有助于保护研究中人类以及动物的权利和安全。以下步骤是基于阅读伦理文献,并和参与伦理决策的哲学家在讨论基础上制定的。*

1. 弄清研究情境的所有因素。也就是说,确定研究中存在哪些重要因素,如被试的特性和计划中的程序。
2. 确定相关的伦理问题。这个问题的一个重要部分是参考已有的伦理准则,例如我们在本章已摘选的《APA 伦理准则》,也可参考不同专业机构的政策条文。另外,了解适用于该研究领域的本地、州和联邦的相关法律和法规。
3. 确定对参与实验的所有成员(被试、研究者、资助者)构成危险的因素。这意味着须采取不同的观点,例如从科学观点、社会观点、被试角度以及全面的伦理观点来确定什么是危险的。
4. 确定可供选择的其他方法或程序,讨论每一种选择的后果,包括它们的相关伦理问题。作为讨论的一部分,要考虑不进行研究的后果。检验每一种选择的现实局限性。
5. 决定要采取的行动。不是根据它是否会让你快乐来判断决定的“正确性”,而是根据一定的程序来决定的。在给定的环境下,它是最佳的选择吗?

小结

心理学研究会引起很多伦理问题。因此,在开始一项研究前,既要考虑《APA 伦理准则》中的特殊的伦理问题,又要考虑与研究相关的法律和法规。开始做研究前,大多数情况下要获得正式的制度上的批准,例如机构审查委员会或动物实验管理委员会的批准。机构审查委员会的功能之一就是对提议的研究的风险

*我们要感谢莱奥那芝加哥大学伦理道德中心的多名成员对本章所做的贡献,特别是 David Ozar、Mark Waymack, 以及 Patricia Werhane 等成员。

收益比达成一致意见。风险可能包括身体的、心理的或社会的伤害。大多数心理学研究必须得到人类被试的知情同意。当出现超过最小风险的情况时,研究者必须采取特殊的安全措施来保护人类被试,还要在测试完成之后提供适当的事后解释。研究者对被试隐瞒信息或有意误导被试会引起严重的伦理问题。当使用欺骗手段时,要在事后解释中告知被试使用欺骗手段的原因。事后解释有助于使被试对研究有更深入参与感,还有助于研究者了解被试是怎样理解实验处理和任务的。在线研究给研究者提出了新的道德难题,所以,在计划一项研究前很有必要向机构审查委员会成员以及对网上数据收集有经验的咨询者咨询。

测试动物被试的心理学者必须遵守联邦和州的多种法律和法规,而且通常必须保护动物的安全。只有没有其他可选择的操作,并且研究在科学、学术或应用方面有重要价值时,才可以将动物置于疼痛或不适中。发表心理学研究报告时应给对研究有贡献的每个人恰当的署名。如果先前发表的著作或论文对研究者的研究有帮助,那么研究者必须承认该贡献,并标注那些研究者的姓名,不这么做就会造成学术剽窃这种严重的伦理问题。

重要概念

风险收益比

欺骗

最小风险

事后解释

知情同意

学术剽窃

隐私权

复习思考题

1. 解释为什么研究者在开始一项研究前要向机构审查委员会或动物实验管理委员会递交研究提议,简单描述这些委员会在研究过程中的作用。
2. 解释在伦理判定中是如何进行风险收益比判断的。简要描述用来评价潜在收益时主要考虑研究项目的哪些特征。
3. 解释为什么研究不能完全规避风险,描述研究者用来检验被试是否“处于风险中”的标准。简要描述被试的哪些特征影响风险评价。
4. 心理学研究中可能出现三种不同类型的风险:身体的风险、心理的风险、社会的风险,它们有什么不同?研究者采取什么安全措施避免可能的社会风险?
5. 在线研究引发的三个重要的伦理问题是什么?
6. 为保证被试的知情同意,研究者有伦理责任向被试说明哪些信息?《APA 伦理准则》指出在什么情形下可以无需知情同意?
7. 当研究者决定信息是公开的还是私密的时候,迪恩纳和克兰德尔(1978)建议考虑哪三个维度?
8. 解释为什么在心理学研究中欺骗有时是必要的?简要描述研究者在使用欺骗手段前要问的一些问题,并描述通常在哪些条件下欺骗被试是不道德的?
9. 事后解释以什么方式让参与者受益?事后解释以什么方式让研究者受益?

10. 《APA 伦理准则》要求在研究中使用动物的研究者应承担哪些具体的道德责任?

11. 《APA 伦理准则》要求满足哪两种条件后才可以将动物置于压力和疼痛中?

12. 解释研究者如何决定一个个体可以作为作者在一份发表的研究报告中署名?

13. 描述当作者从第一手材料或第二手材料中引用信息时, 为避免学术剽窃必须遵守的程序。

挑战性问题

注意: 与其他章节不同的是, 本章给出的挑战性问题或拓展练习没有答案。为了解决道德难题, 你必须会运用恰当的道德标准, 还必须能在和不同背景及不同知识结构的人讨论提议后就研究的性质达成共识。因此, 你将不得不考虑与你的观点不同的观点。我们希望你能够将这些问题作为小组讨论的一部分。

本章的前两个挑战性问题包括一个假设的研究提议, 这个研究提议包含的理由和方法与真实发表的研究类似。为回答这些问题, 你需要熟悉本章提出的《APA 伦理准则》和其他关于伦理决策的材料, 包括在本章结束时列出的伦理决策可采取的步骤。正如你将见到的, 你的任务是判定研究程序是否违背了特定的伦理标准, 同时还要对提议的研究给些建议, 包括最基本的建议——是否允许研究者进行该项研究。

1. 机构审查委员会的提议

介绍 假设你是机构审查委员会的一名成员。除你外, 委员会成员还包括临床心理学家、社会学家、社会工作者、哲学家、新教牧师、历史学教授以及受人尊敬的当地企业经理各一名。下面是一个提交给机构审查委员会审查的研究摘要。要求你考虑向研究者提出的问题, 以及根据规定, 你是否会同意该研究以它提出的形式来进行, 或是否同意修改后进行, 或者是否否该提议 (一个提交给机构审查委员会的真的研究提议包含的细节比我们这里提的更多)。

理由 人们在缺乏明显的原因或表面证据矛盾的情况下接受他人的观点或判断时, 会发生心理从众现象。先前研究已经考察了容易产生从众现象的条件, 并且以得出一些结论, 例如, 在人们预期到不快乐的事件 (如震惊), 以及来自个体认同的人的压力时,

从众行为会增加。一项研究提议考察十几岁的学生之间讨论酒精消费情况时产生的心理从众。研究目的是识别导致学生主动参加为学生提供酒精和允许酒后驾车的社会活动的影响因素。此研究试图在自然情境下研究从众行为, 如不迎合同伴压力就可以避免不愉快事件 (例如, 法律惩罚、学校开除、受伤甚至死亡) 的情形。

方法 此研究将有 36 名 16~18 岁的高中生, 他们自愿参加“当今高中生的信仰和态度调查”的研究项目。将被试分为 4 人一组进行讨论, 小组的每个成员需要回答 20 个相同的问题, 要求回答问题前同组内成员讨论每个问题。其中 4 个问题涉及未成年人酒精消费和可能减少未成年人饮酒和驾车的举措。

研究负责人指定小组中的一个作为讨论的引导者, 在被试不知情的情况下, 将他们随机安排到三个不同的小组中, 三个小组分别没有、有 1 名或 2 名实验助手来协助研究者。研究者事先指导每个“实验助手”在小组讨论关于未成年人饮酒的关键问题时应说什么 (将在第 4 章讨论心理研究中实验助手的利用)。需特别指出的是, 要求实验助手按一个脚本去做, 该脚本提出一个论点: 大多数达到法律允许的饮酒年龄 (16 岁) 的人, 以及所有能在全美选举中投票和服兵役 (18 岁) 的人, 都可以对饮酒做出自己的决定; 此外, 因为这个决定取决于个体自身, 所以, 如果低于法定饮酒年龄的人选择饮酒时, 其他人无权干涉。每个实验助手“承认”至少在先前的两个场合下饮过酒。这样, 实验操纵了小组中分别没有、有 1 名或 2 名成员, 他们的观点是学生无需为避开为未成年人提供酒精的场合而负责, 或在其他人选择饮酒和驾车时

(续)

他们没有干预的责任。

在此实验中,将评价这种论点对真实被试作答的影响。另外,这些时段会在被试不知道的情况下作录音,并对录音内容进行分析。实验结束后,向被试解释欺骗的性质和录音的原因。

2. 动物实验管理委员会的提议

介绍 假设你是动物实验管理委员会的一名成员。除你外,委员会成员还包括兽医、生物学家、哲学家以及受人尊敬的当地企业经理各一名。下面是一个提交给动物实验管理委员会审查的研究摘要。要求你考虑向研究者提出哪些问题,以及根据规定,你是否会同意该研究以它提出的形式来进行,或是否同意修改后进行,或者是否否决该提议(一个提交给动物实验管理委员会的真的研究提议包含的细节比我们这里提到的多得多)。

理由 该项研究试图研究边缘系统的皮层下组织在适度情绪和攻击时所起的作用。此研究提议基于该实验室先前对猴子脑皮层组织的研究,这些研究表明了猴子的脑皮层下不同区域的损伤与进食、攻击和其他社会行为(如求爱)的变化存在明显相关。研究的脑区是试图控制过度攻击和攻击偏好而被神经外科切除的脑区。另外,本研究关注的特定皮层下区域有可能与某种性行为的控制有关,而控制性行为有时是心理治疗(如过度性行为)的主题。先前的研究不能确定控制该种行为的确切区域;此提议有望能在该方面有所进展。

方法 被试为两组恒河猴。一组(N=4)为控制组。对这些动物做假手术,包括麻醉和在头骨上钻孔。然后以相同的方式对实验组和控制组动物进行测试和评价;损伤实验组被试的皮层下组织的一小部分——杏仁核,两只猴子的损伤部位为该组织的一侧,而另外两只猴子的损伤部位为该组织的另一侧。痊愈后,所有动物将完成一系列的任务以测量它们的进食情况、与同性和异性猴子(正常的猴子)在一起的社会行为以及情绪反应(例如对新异刺激的反应:实验者戴着小丑面具出现)。猴子被关在现代化的实验室内,一位职业兽医实施手术并监控伤愈情况。进行测试后,实验组猴子将会被宰杀,它们的大脑会留作组织学检查。(用组织学来确认损伤的部位和程度是必需的。)控制组猴子不会被宰杀,它们被遣回群体以供日后研

究用。

3. 斯坦利·米尔格拉姆的顺从研究引起了心理学研究中使用欺骗的道德问题的争论(见专栏3.2)。顺从指一个人遵循权威人物指示的程度。在这个问题的A部分,你会读到一个描述米尔格拉姆实验基本程序的摘要。假设你是机构审查委员会的成员之一,你把这篇摘要当作一篇提交给机构审查委员会的研究提议那样来看待。这个问题的第二部分,你要考虑在B部分中米尔格拉姆用这种范式研究顺从的附加信息。然后解释为什么你改变或者不改变基于A部分评论所作的决定。

A. 两个人来到心理学实验室,他们认为是要来参加一项学习实验。他们被告知,本实验研究的是学习中的惩罚效应。抽签决定谁是“教师”、谁是“学习者”。其中一个人实际上是研究者的实验助手,并且抽签程序是人为操纵的,以便让真正的被试每次都抽到教师角色。被试看着学习者被带到一个邻近的房间并捆在椅子上,手腕接上电极。然后,研究者对被试说,学习者在学一系列词对时如果出错,将会受到电击。然后将老师带入实验室,在那里配有一个看起来令人印象深刻的带有30级电闸的电击产生器。每级电闸标有伏特数(范围从15伏特到450伏特),电闸的旁边是描述电击量的词语标签,例如“轻微电击”、“强电击”、“危险的强电击。”最后两个电闸仅标有xxx。向老师做一个示范性的电击,告知老师只要学习者犯错就实施电击。学习者的反应可以通过四个开关点亮电击产生器顶端的灯来传达。要求老师在学习者每次错误反应后给予更高一级的电击。随着实验的进展,学习者对电击提出抗议。透过墙壁可以听到学习者的抱怨和由于强烈电击产生疼痛时的尖叫声,以及后期学习者要求实验者停止实验程序的声音。当老师把电击调到180伏特时,学习者叫喊“我受不了了”,调到270伏特的时候,学习者极度痛苦地尖叫。在300伏特时,“我不回答了,”学习者仍然尖叫。当按下相应的330伏特的电闸后,听不到学习者的任何声音了。实际上学习者并没有被电击,这里主要的因变量是被试执行实验者的“命令”而可能给出的最大电击量。实验后,向所有的被试给出事后解释,而且有时研究者会和被试谈一段时间。所有

(续)

被试还都接受了追踪问卷调查。在实施实验前，米尔格拉姆向37位心理学家描述了他的计划程序，没有一位心理学家预测被试会给予最大电击。

- B. 米尔格拉姆用这种方法进行了十几次实验（见Milgram, 1974）。在其中一个实验中，被试能听到学习者的尖叫声而看不见学习者，接近60%的被试给予了学习者最大电击。在这样一个出乎意料的发现后，继续此研究的一个主要原因是，实验没有明显地伤害到被试，并且占很大比例的多数被试（84%）说他们很愿意参与这样的实验。许多被试（74%）在追踪问卷调查中回答说，他们从这个经历中获得了某些自我价值。通过这一系列的实验，米尔格拉姆发现，被试顺从的可能性是受情境因素影响的。例如，当学习者和教师呆在同一个房间里时，被试可能较少顺从；当教师能自主选择电击的伏特时，被试更不太可能实施最高电击。对这种独创性发现的一种解释是，人乐意顺从——他们像谚语中的绵羊一样行为。通过整个系列的研究，证明了一个迥异的观点。米尔格拉姆阐述了人类对要求他们顺从的情境的许多方面很敏感。这里仍然存在一个问题，基于米尔格拉姆的发现，我们学到了人

们有顺从的倾向性，这种获益是否能让他的实验范式承担的风险有了正当理由？更概括地说，当不能用研究的成果来评价其潜在利益时，机构审查委员会怎样才能最佳地评估提议的研究的潜在利益？

4. 考虑下面情境，这是由凡恩和科德克（Fine & Kurdek, 1993）提出的，这是他们讨论发表作品的作者署名问题时碰到的一个案例。

一位研究生请求一位心理系教员指导其做论文。学生提出主题，系教员改进研究方法，学生收集和输入数据，系教员对数据进行统计分析，学生使用部分统计分析的资料来写论文。学生在系教员的悉心指导下完成了论文。论文完成后，系教员认为研究得出的数据很有趣，可以作为一个整体来发表。因为学生缺乏给杂志写整个论文的必要能力，所以系教员做了这个工作。学生的论文大约占了杂志文章中的近三分之一的比重。

- A. 为了确定学生是作为发表作品的作者还是应该在文章注脚中被提及，你需要考虑哪些因素？
B. 如果你决定学生为作者，解释你认为学生应是该文章的第一作者还是第二作者。

第二编

描述性方法

第4章
观察法

第5章
调查研究

第6章
行为的隐蔽测量

第4章

观察法

本章大纲

概 述

行为样本

时间抽样

情境抽样

观察法的分类

非干预观察

干预观察

参与性观察

结构化观察

现场实验法

行为记录

行为的定性记录

行为的定量测量

观察数据的分析

数据简化

观察者信度

批判性思考观察研究

观察者的影响

观察者偏差

小 结

概述

我们每个人每天都在观察行为。我们走路或开车时一旦离开了观察就会发生危险,甚至可能威胁到生命。出国旅行时,我们可以通过观察当地人的行为来避免尴尬。我们可以通过观察人们在商店内排队付账、乘公交车以及在教室里的行为获得很多信息。同样地,科学家也可以通过观察来了解许多有关行为的信息。然而,日常观察同科学观察在很多方面都不同。在随意观察中,我们很可能不会注意使观察发生偏差的因素,而且也几乎不会对观察做正式的记录。相反,我们依靠的是对事件的记忆,尽管我们的经验(以及心理学家的观察)表明我们的记忆并不完美。从另一方面来说,科学观察却需要在精确界定的条件下,以系统、客观的方式进行,同时认真记录所观察的信息。

系统观察不仅是心理学家的重要工具,也是人类学家、社会学家以及生态学家的重要工具。本章将主要讨论自然情境下的观察。值得一提的是,科学观察需要遵循的原则不仅适用于自然情境,在实验室中也同样适用。

观察法的主要目的是描述行为。科学家致力于尽可能全面而且准确地描述行为。要想达到这一目的,研究者面临着很大的挑战。显然,研究者不可能观察到一个人的所有行为。因此科学家依靠观察人们行为的样本来描述行为,但是,他们必须确定这些样本能够代表人们的普遍行为。本章我们将描述心理学家是如何选取行为样本的。

为了全面地描述行为,心理学家还面临另一个挑战:行为经常随其发生的背景改变而改变。设想一下我们自己的行为:你在教室里的表现和你在家里、学校里或者宴会上的表现一样吗?你能通过对他人的观察得出背景很重要的结论吗?你难道没有观察到孩子与父亲在一起和与母亲在一起的时候行为会发生变化吗?举例来说,社会心理学家很久前就注意到,仅是他人在场就足以影响一个人的行为。这种影响作用可能是社会助长(他人在场时成绩提高)或社会抑制(他人在场时成绩降低)(Bond & Titus, 1983)。因此,对行为进行全面的描述就需要在不同情境、不同时间进行观察。

观察法为行为假设提供了丰富的资料来源,因此观察也是研究者探索人类行为方式及其成因的第一步。例如,卡根、雷茨尼克和斯奈德曼(Kagan, Reznick, & Snidman, 1988)观察了2~3岁儿童群体对不熟悉的人或物的反应。根据观察结果,他们将这群儿童分为害羞、安静、胆小和好交际、健谈、情感丰富两种类型。当这些儿童7岁时再对他们进行观察,结果发现每一类型中的多数孩子都表现出了类似的行为。根据这些观察,研究者提出一个假说:儿童时期的羞怯与成年期的极度社会焦虑相关,他们认为两者均是先天气质差异造成的。在一个相似的研究中,卡斯皮(Caspi, 1977)和他的同事发现,在儿童3岁时所做的观察,可以成功预测其长大后威胁健康的行为(如酗酒)。这些研究结果支持了儿童早期的气质差异与成年后的行为模式相关的假说。

通过本章的学习,你将明白科学观察者并非总是被动地记录行为的发生。本章将探讨为什么科学家要为观察创造特殊的情境,此外还会介绍记录和分析观察

数据的方法。最后，本章还将阐述用观察法做研究时解释结果的困难和挑战。

行为样本

- 当不能获得所有的行为记录时，研究者会寻求具有代表性的行为样本。
- 观察的推论程度（外部效度）取决于行为样本的抽样。

在进行观察研究前，研究者必须决定何时何地进行观察。在多数观察研究中，研究者不能观察到所有的行为，只能观察到特定时间、特定情境以及特定条件下发生的一些行为。换句话说，必须对行为抽样，该样本可用来代表一个更大总体的所有可能行为。研究者只有通过选定时间、情境和条件，对代表总体的行为样本进行观察，才能把他们的结果推论到这个总体中。时间、情境和条件的抽样，很大程度上影响着抽样的最重要方面，即被试。观察结果只能推论到与实施观察的研究相类似的被试、时间、情境和条件中。代表性样本的关键特征是，他们与所在总体中的被试相似，在时间、情境和条件等方面也具有相似性。例如，学年初对课堂行为的观察结果并不能用来等同于学年末课堂的典型行为。

外部效度（external validity）指研究结果能推论到不同的总体、时间、情境和条件中的程度。效度与“真实性”有关。考察一项研究的外部效度，需要确定本研究的结果在多大程度上可以解释除本研究之外的其他人、情境和条件。我们将在本节描述时间抽样和情境抽样是怎样提高观察结果的外部效度的。

时间抽样

- 时间抽样指研究者系统地或随机选择进行观察的时间间隔。
- 当研究者对那些并不经常发生的事件感兴趣时，他们就会对行为样本采用事件抽样。

研究者通常结合时间抽样和情境抽样来确定代表性样本。在**时间抽样**（time sampling，也可译作“时间取样”）中，研究者通过选择不同的时间间隔来寻找代表性样本。时间间隔的选择可以是系统的（例如，每周的第一天观察）、随机的，或者二者兼有。

不妨思考一下，如何用时间抽样来观察儿童的课堂行为。在这样的研究中，研究者每天观察儿童两个小时。如果研究者把观察限制在一天的某个特定时间内（比如说只在早上），那么他们的结果将不能被推论到其余的在校时间。获取代表性样本的方法是在整个教学日内系统地安排观察时段。比如可以每隔两个小时开始一次历时30分钟的观察，共进行4次观察。第一个观察时段在上午9点开始，第二个时段在11点开始，依此类推。另一种可能的安排是在教学日内每隔半个小时安排历时10分钟的观察。在此情境中也可以运用随机时间抽样，在整个教学日内随机分配4次历时30分钟的观察（或者12次历时10分钟的观察）。这样每天

都有一个不同的观察时间表。虽然每天的观察时间段不同,但是从长远来看,在整个教学日内发生的所有行为都有相同的抽样机会。电子装置为使用随机时间抽样提供了极为有利的条件。可设定电子寻呼机按照随机时间表给被试发信号(正常睡眠时间除外)。例如,在对中学生的研究中,拉森等人(Larson, Richards, Moneta, Holmbeck, & Duckett, 1996)在“16 477 个随机时间点”采集了青少年被试对其生活的自我报告。

研究中常常同时采用系统时间抽样方法和随机时间抽样方法,比如可以系统地设计观察时间间隔,而在每个时间间隔内随机选择时间点进行观察。例如,每天的同一时间段内设定四次历时 30 分钟的观察(例如,上午 9 点,上午 11 点等),然后观察者在每次 30 分钟的观察内,随机进行 20 秒钟时间间隔的观察。无论使用哪一种时间抽样方法,观察者都必须仔细考虑每一种方法的优势和局限。

当感兴趣的事件发生频率较低时,时间抽样就不是一个有效方法。对发生频率较低的事件使用时间抽样的研究者可能会彻底错过事件。如果事件持续的时间很长,那么时间抽样可能导致研究者错过事件的重要部分,如事件的开始或结尾。对发生频率较低的事件来说,事件抽样是较有效的抽样方法。在事件抽样中,观察者会记录符合预先定义的每个事件。例如,研究者想要观察儿童对学校特殊事件的反应,如假日活动,这时就会使用事件抽样。特殊事件决定了什么时候进行观察。研究者不会对特殊事件采用时间抽样,因为特殊事件未必能在随机选定的观察时间段内发生。

事件抽样除了适用于观察有严格时间规定的事件(如学校活动),对情境的研究也十分有用。研究者经常对不能预期的事件感兴趣,比如自然或人为的灾难。只要可能,研究者都力图在感兴趣的事件发生或可能发生的时候在场。例如,在对儿童“扭打”活动的研究中,研究者站在操场的角落来观察某幼儿园一个班级中的孩子(Smith & Lewis, 1985)。由于这种行为发生的频率相对较低,该研究选择了事件抽样的方法。只要发生扭打,研究者就进行观察,一直观察到扭打场面结束。尽管事件抽样是观察非预期事件的有效方法,但是使用事件抽样容易把倾向性带入行为记录中。例如,事件抽样可导致观察者在最“方便的”时候或仅在事件一定会发生的时候抽样。这样观察到的行为样本可能不能代表其他时候发生的此种行为。很多情况下,观察者倾向于仅使用某一种形式的时间抽样来获得行为样本。

情境抽样

- 情境抽样是指在不同地点、不同环境和条件下研究行为。
- 情境抽样会提高研究结果的外部效度。
- 在情境中,可使用被试抽样来对情境中的某些人进行观察。

研究者可通过使用情境抽样来显著地提高观察结果的外部效度。情境抽样(situation sampling,也可译作“情境取样”)要尽可能多地在不同地点、不同环境和条件下观察行为。通过不同情境的抽样,研究者可避免其研究结果只适用于某

些特定的环境和条件下所特有的情况。动物在动物园与在野外的行为表现不同。儿童和父亲在一起的行为表现与和母亲在一起的行为表现并不总是一样的。通过对不同情境的抽样,研究者也可增加被试样本的多样性,与只观察一些特殊类型的个体相比,能够获得更大的普遍性。

为了对大学生喝啤酒情况进行自然观察,研究者有意对提供啤酒的不同场合进行行为抽样,选取了五个城镇酒吧和一个大学生中心(Geller, Russ, & Altomari, 1986)。在另一项研究中,拉弗朗斯和梅奥(Lafrance & Mayo, 1976)调查不同种族人群在目光接触上的差异时也采用了情境抽样。他们在大学自助餐厅、商业区快餐店出口、医院和机场休息室以及饭店观察成对个体间的目光接触。通过采用情境抽样,研究者可以调查到不同年龄、不同社会经济地位、不同性别和种族的人群。这种对目光接触文化差异的研究结果具有很高的外部效度,比那些只对某一特定情境中某些特定类型的被试进行观察所得到的结果要好得多。

多数情况下,研究者会观察在抽样时间和地点出现的所有个体的行为。例如,研究者想观察学生早上和晚上在图书馆学习时的专心程度,他们会在指定时间观察图书馆内的所有学生。然而,有些时候同时发生的事件较多,就会影响到观察的有效性。例如,如果研究者要观察食堂高峰期学生对食物的选择,那么他们就不可能观察所有的学生。在这种及其他类似情况下,研究者会使用被试抽样来确定观察哪些学生。与时间抽样程序类似,研究者可系统地选择学生(每次观察第十名学生)或随机地选择学生。同样,被试抽样的目标也是为了获得在食堂吃饭的学生的代表性样本。

观察法的分类

- 观察法可分为“干预观察”和“非干预观察”。
- 记录行为的方法可按记录行为的多少分类。

观察法可在两个维度上进行分类(Willems, 1969)。按第一个重要维度是干预观察和非干预观察。第二个维度涉及记录行为的方法。观察研究可按照是记录了所有的(或几乎所有的)行为还是仅记录了特定的行为来区分。在某些情况下,研究者会寻求对行为进行全面的描述。他们通过影片、录音带或较长的口头描述来达到这一点。更多的情况下,研究者会记录与特定研究目标相关的特定行为单元。例如,钱伯斯和阿西奥纳(Chambers & Ascione, 1987)对儿童玩攻击性或亲社会性视频游戏时的行为进行了观察。他们在观察中发现,相对于玩亲社会性游戏的儿童而言,玩攻击性游戏的儿童放入捐款箱的捐款要少,而且帮助别人削铅笔的行为也较少。在这个研究中,研究者只记录了与儿童亲社会行为相关的特定反应。

我们将先根据观察者干预的程度,然后再根据记录行为的方法来讨论观察法。

非干预观察

- 自然观察的目的是描述正常发生的行为并检验变量间的关系。
- 自然观察有助于建立实验室研究结果的外部效度。
- 当伦理和道德上的考量导致无法进行实验控制时，自然观察就成为重要的研究策略。

在自然情形下观察行为，研究者没有任何的干预，通常称之为**自然观察法**（naturalistic observation）。在这种方法中，研究者扮演了被动记录行为的角色。事件自然发生，观察者没有操纵或控制，因此很难精确地定义自然情境（Bickman, 1976），我们通常认为自然情境就是事件正常发生，而且没有为达到某一个观察目的而做特定的安排。然而，在心理学实验室内进行的观察就不能称之为自然观察，因为实验室情境是为了研究行为而专门创建的，从这个意义上说，实验室是**人造**的情形而不是自然情形。实际上，自然情形观察的功能之一是为实验室研究结果建立外部效度——把实验室发现带入到“真实世界”中。观察网上讨论群和聊天室的行为是研究者使用的另一种描述自然发生行为的方法。然而，这种新近的“自然”观察方式会引发严重的道德问题，我们在第3章中已提到这一点，在本章的后面还会继续讨论（Kraut et al., 2004）。

自然观察的主要目的是描述正常发生的行为，并且调查行为所涉及变量间的关系。例如，哈塔普（Hartup, 1974）用自然观察法研究了明尼苏达州圣保罗儿童看护中心的学龄前儿童出现攻击行为的频率和类型。他区分了敌意攻击（指向人的）和器物攻击（针对物品、领地或权力）。尽管总体上男孩比女孩更具攻击性，但他在观察中没有发现证据证明攻击类型存在性别差异。因此哈塔普得出结论，男孩和女孩的敌意攻击没有“显著的”差异。

哈塔普关于儿童攻击性的研究说明了为什么研究者会选择使用自然观察，而

图 4.1 动物研究者，如简·古多尔（Jane Goodall），经常根据自然观察来获得被试行为的信息



不是操纵与行为有关的实验条件。出于道德和伦理考虑,我们对人类行为的某些方面不便进行控制。例如,一些研究者对儿童早期隔离与后期情感和心理发展的关系感兴趣。然而,我们会强烈反对人为地将儿童从父母身边带走。如果想研究这一问题,就必须思考另外的收集数据的方法。例如,采用动物实验来研究早期隔离对后期发展的影响(Harlow & Harlow, 1966);对由动物抚养长大的称之为野人的儿童进行描述(Candland, 1993);对遭受父母不寻常的隔离待遇的儿童进行个案研究(Curtiss, 1977);以及对送进收容机构的儿童进行系统观察(Spitz, 1965)。对儿童攻击行为的研究也需要道德和伦理的认可,我们不愿看到仅为了记录儿童的反应就有意置他们于激怒和挑衅状态。然而,观察过儿童的人都知道,儿童有很多自然发生的攻击行为。哈塔普的研究表明了道德和伦理的约束范围内,自然观察如何成为获取儿童攻击信息的有用方法。

并非只有心理学家在自然情境下观察行为,观察法也是生态学的基础研究方法。尽管生态学与心理学有关,但还是被普遍认为是生物学的一个分支(Eibl-Eibesfeldt, 1975)。动物行为学家研究与环境有关的生命物体的行为。他们采用比较法来解释行为,试图以在进化等级较低的物种中观察到的本能行为模式为基础来解释一个动物物种的行为。关于本能机制决定人类行为的推测是生态学家常用的。生态学研究关注的通常是某种生物的习性谱,这是有关某种生物所有行为模式的完整分类,包括行为频率、持续时间以及每个行为发生的背景信息。另外,研究已证明,生态学观点对理解人类正常和异常行为都是很重要的。一个有趣的例子是,对精神病学问题的生态学分析是以对精神分裂症患者面部表情的全面描述为基础的(Pitman et al., 1987)。

干 预 观 察

- 大多数心理研究采用干预观察。
- 干预观察的三种方法是参与性观察、结构化观察和现场实验。

让我们来直面这种情况:科学家们喜欢“摆弄”自然。为了观察效应,或许也为了去验证理论,他们喜欢进行干预。干预比非干预更能表现大多数心理学研究的特色。尽管心理学家使用的干预类型太多而无法分类,但干预理由一般下面的一种或几种:

1. 促成或引发一个自然条件下不常发生的事件,或通常在不易观察的条件下发生的事件。
2. 通过系统地改变刺激事件来研究有机体的反应限度。
3. 获得一种途径以进入那些通常不向科学观察开放的的情境或事件。
4. 设置条件,对先前发生的重要事件进行控制,观察随后出现的行为。
5. 通过操纵一个或几个自变量来检验它们对行为的影响,并进行比较。

当研究者在自然情境中进行干预的时候,通常采用三种主要的观察方法:参

与性观察、结构化观察和现场实验。三种方法的干预性质和程度各不相同。我们将逐一介绍。

参与性观察

- 无伪装的参与性观察常被用来了解群体文化和行为。
- 当研究者知道个体会因为被记录而改变行为时，通常使用伪装的参与性观察。
- 参与性观察能让研究者观察到通常不能被直接进行科学观察的行为和情境。
- 参与性观察有时可能会有失客观，或不适当地影响到所要记录的行为个体。

在参与性观察 (participant observation) 中，观察者扮演着双重角色。他们观察人的行为同时又积极参与到观察情境中去。在无需伪装的参与性观察中，被观察的个体知道观察者在收集他们的行为信息。人类学家经常使用这种方法，他们通过和群体成员一起生活和工作来了解群体的文化和行为。

在伪装的参与观察中，被观察者不知道他们在被观察。当人们知道自己的行为被记录时，他们的行为表现就与平常不同。例如，政治家对媒体经常发表不同的声明，这取决于媒体是否记录他们的发言。当我们意识到被关注时，我们的行为可能会受到影响。由于存在这种可能性，研究者会决定伪装观察者的身份，因为他们知道，如果被观察者知道正在记录自己的行为，他们的表现就会和平时不同。伪装的参与观察引起的道德问题（如隐私和知情同意）必须在实施研究前先行考虑。我们在第3章中已探讨过这些道德问题，我们还将在本章的后面进行深入讨论。

罗森汉 (Rosenhan, 1973) 使用伪装参与观察法来调查精神病的诊断和住院治疗。8个人（包括几名心理学家、一名儿科医师和一名家庭主妇）隐瞒了他们的姓名、职业，他们向12所不同的心理医院寻求住院治疗。每个人都声称相同的症状：他或她一直不断听见各种声音。他们中的大多数被诊断患有精神分裂症。

住院治疗后不久，研究者就不再抱怨任何症状，也不再出现异常行为（除了暂时的，继续表现出害怕被“抓住”的焦虑）。这个研究除了观察全体病人与医护人员的相互作用外，还对一个“健全”的人要用多久才能出院感兴趣。研究者住院治疗了7到52天不等，当他们出院时，他们的精神分裂症据说已“好转”了。显然，一旦伪装的病人被贴上精神分裂症的标签，不管他们后续行为的性质如何，人们都会坚持这种标签。然而，我们有理由去继续验证这个特殊的结论以及罗森汉 (1973) 的其他研究结果（见专栏4.1）。

参与性观察法使观察者得以在通常不对科学观察开放的情境中进行观察。例如，一个研究者以“好奇的新教徒”的身份进入各种“白种人的网上聊天室”，分析针对非洲裔美国人的仇恨犯罪 (Glaser, Dixit & Green, 2002)。当然，通常无法对那些鼓励暴力的犯罪地点进行科学观察。参与其中的观察者可能会与被研究者有一样的体验，这种体验有助于洞察和理解被研究的个体或群体。例如，在罗森汉的研究中，伪装的病人体验了被贴上精神分裂症患者标签的感受，以及不知道要多久才能重返社会的心态。罗森汉 (1973) 研究的一个重大贡献是，它论证了

专栏 4.1

对“不正常环境中的正常者”的批判性思考

在罗森汉(1973)的“不正常环境中的正常者”文章中,他质疑了精神病诊断和住院治疗的问题。正常人是如何被诊断为患有最严重的心理疾病——精神分裂症的呢?为什么医院的医护人员分辨不出病人是在伪装?经过数天或数周的住院治疗,为什么医护人员没认识到伪装的病人是“心智健全的”呢?

这些都是非常重要的问题。罗森汉的文章在《科学》(Science)杂志上发表后,许多心理学家和精神病学家就罗森汉所提问题相约集会并撰写文章(如,Spitzer, 1976; Weiner, 1975)。下面列举一些对罗森汉研究的批评。

- 我们不能批评医护人员诊断错误:一个基于伪装症状的诊断当然会出现错误。
- 伪装的病人不止有一种症状:他们焦虑(害怕被“揭穿”)、他们报告说感到压抑、他们寻求住院治疗。主动要求进入精神病院,这“正常”吗?
- 伪装的病人在医院里的表现真“正常”吗?或许正常的行为就是说类似这样的一些话:“嗨,我只是假装有病来看看我是否能住院,但是说实在的,我撒谎了,现在我想回家了。”
- 精神分裂症患者的行为不总是错乱的;“真正的”精神分裂症患者经常表现“正常”。因此,医护人员花了很多天来确定伪装的病人不再表现那些症状,

这并不奇怪。

- “好转”的诊断是相当少见的,它反映了医护人员认可一个伪装的病人不再出现症状。然而,关于精神分裂症的研究说明:一旦一个人表现出精神分裂症的症状,他或她比其他入更有可能再次体验这些症状。因此,“好转”的诊断可以指导心理健康专业人员了解一个人的后续行为。
- “心智健全”和“心智不健全”是法律术语,不是精神病学术语。法律上确定一个人心智是否健全需要判断一个人是否能区别对和错,这与本研究无关。

就像你看到的,罗森汉的研究颇具争议。现在,许多专业人士都相信,该研究不能帮助我们了解精神病诊断。但是罗森汉的研究从长远上看来具有重大的意义:

- 心理健康专业人员会延迟诊断,直到收集到更多关于病人症状的信息,这叫做“延期诊断。”
- 心理健康专业人员更容易意识到他们的理论和个人偏向可能影响他们对病人行为的解释,因而努力避免出现偏向判断。
- 罗森汉的研究描绘了许多病人在心理健康中心中人格丧失和无能为力的经历。他的研究促使心理健康领域中的工作者检验他们的工作并改善病人就医的条件。

发生在公共机构环境中的去人类化现象。类似地,在1959年秋天,约翰·霍华德·格里芬(John Howard Griffin)开始调查“在歧视黑人地区黑人的般境遇。”白人格里芬为了冒充黑种人,用药物和色素染黑了皮肤。他用一个多月的时间穿越密西西比州、阿拉巴马州和乔治亚州(步行、搭便车和乘公车)。在《我这样的黑人》(Black Like Me, 1960)一书中,他描述了作为一名参与观察者的经历,这成为20世纪60年代民权运动的一个重要导火索。

在实施一项成功的研究时,参与性观察者在一个情境中的角色可能会引起严重的问题。例如,如果观察者把自己和被研究的个体视为一体,那就可能会失去有效研究所要求的客观性。参与性观察者有时变化很大,又不易预见。一名犯罪学家曾经使用无伪装参与观察法研究工作中的警察,他的经历可以证明这一点。柯卡姆(Kirkham, 1975)像新警察那样通过了警察学校的培训,成为了一名巡

警，被安排到一个大约 50 万人口的高犯罪率的城市。他每天忙于警察的日常事务，这导致了他的态度和人格发生了显著的改变。柯卡姆自述道：

我的新职业——一名贫民窟警察。随着时间的流逝，我慢慢地而又无情地开始变得不能区分自己和共事警察的态度和行为……按照我家人、同事和朋友的说法，我日益表现出与我先人格不同的态度和行为——惩罚，玩世不恭和不信任他人，易怒和飘忽不定的敌意，种族主义，以及因犯罪率上升和有强迫倾向罪犯带来的威胁造成的弥散性人格焦虑。先前我是死刑制度的反对者，而现在却极力主张对谋杀、绑架和杀害警察等重罪犯执行死刑——尽管作为一名犯罪学者，我仍一直认为死刑对犯罪不会有威慑力（p.19）。

参与观察者必须意识到卷入会威胁研究的客观性，这种威胁必然会随着卷入程度的增加而增加。

观察者卷入的另一个潜在问题是，观察者对被研究者行为的影响。观察者卷入必定会和人们相互影响：做决定、开展活动、承担责任以及像情境中其他人一样的行为。无论何时，观察者卷入自然情境时必须考虑清楚他对被观察者和事件的影响程度。有观察者卷入与无观察者卷入条件下所观察到的情况是否一样？如果卷入产生的行为只是研究者创建的条件和事件中的特定行为，那么就很难将结论推广到其他情境中去。

观察者卷入对被观察行为的影响程度不易评价，因为必须考虑许多因素，比如参与者伪装与否、所参加群体的大小以及观察者在群体中的角色等。在罗森汉的研究中，伪装的参与性观察法看起来是成功的：罗森汉和他的同事假装成病人似乎没有显著地影响医院的环境。然而，尽管没有一个医院职员发现伪装病人心智健全，但是有一些病人却发现了，他们认为这些观察者在对医院进行检查。

如果观察的群体较小，或参与观察者的表现突出，那么观察者更有可能显著地影响参与者的行为。几个社会心理学家混入了一个声称和外太空生物有联系的团体（Festinger, Riecken & Schachter, 1956）。该团体的一位领袖称，他已接收到外星人的信息，预测在某个特定的日期会发生一场灾难性的洪水，洪水会从北极圈延伸到墨西哥湾。鉴于该团体成员对“不信者”的态度，观察者不得不虚构一些奇异的故事来进入该团体。这个方法很奏效，甚至有一个研究者被认为是向大家发布信息的外星人。研究者已经在无意中强化了团体的信念，还以一种不确定的方式影响了后续事件的发展进程。顺便说一下，那场洪水最终也没发生，但是至少一部分成员开始用这种不确定来巩固最初的信念。他们声称他们的信念阻止了洪水的发生，并以此来招募新成员。

因而，尽管参与性观察法可能允许观察者进入通常不能直接进行科学观察的情境，但是观察者在运用这种技术时，必须尽量防止不客观并避免对研究的行为产生潜在影响。

结构化观察

- 结构化观察法用于记录那些自然观察法难以观察的行为。

• 临床和发展心理学家经常使用结构化观察法。

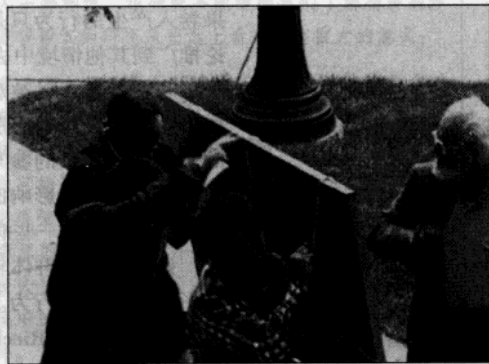
• 当不同的观察或观察者没有严格遵循同样的观察程序，或没有控制重要的变量时，解释结构化观察就会出现问题的。

使用干预技术的观察法有很多种，很难对它们进行分类。这些程序不同于自然观察法，因为研究者通过干预的方式对他们观察的事件施加控制。然而，这种对事件进行干预和控制的程度小于现场实验（下节我们会简要地描述，第7章会详细描述更多细节）。我们把这些程序称为**结构化观察（structured observation）**。观察者经常会利用干预来引起一个事件或“创设”一个情境，这样比不进行干预时更容易记录事件。在有些情况下，观察者可能要设计相当精细的程序来更全面地研究某个特定行为。

西蒙斯和莱文（Simons & Levin, 1998）使用结构化观察法研究了“变化盲”的现象。当人们注意不到周围环境的变化时就出现了变化盲。我们周围有很多状况，我们不可能注意其中每个细小的变化。然而，西蒙斯和莱文证明了人们甚至也经常察觉不到他们正在关注的事情所发生的变化。在他们的研究中，研究者使



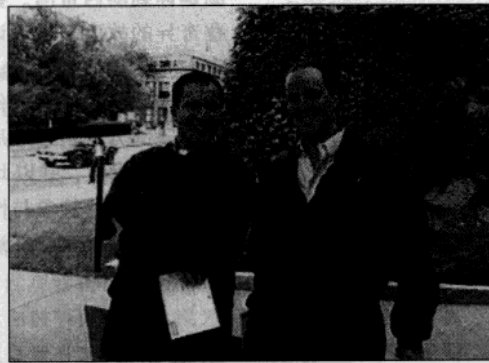
a



b



c



d

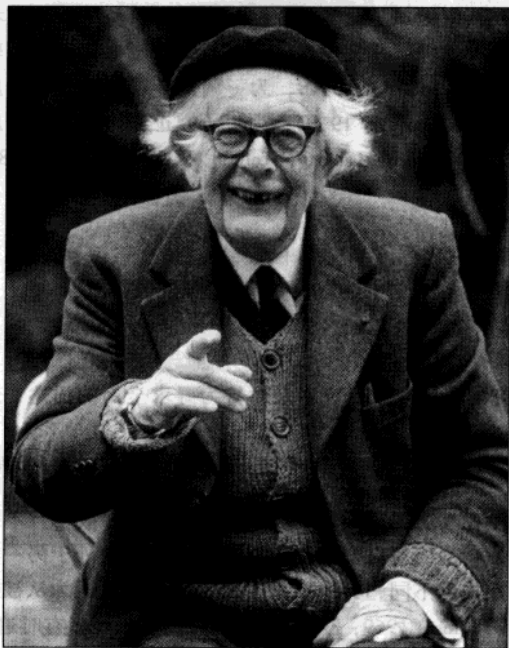
图 4.2 实验录像的几幅画面：画面 a—c 显示转换的顺序，画面 d 显示肩并肩的两个同谋者

用了“同谋者”，就是说，为了创造出观察行为的情境，引导研究情境中的个体按某种特定的方式行事。一个同谋者扮演一名走在校园里问路的路人。问路对话大约15秒时，另外两个同谋者抬着一扇门粗鲁地从他们中间穿过。这时候，问路同谋者同抬门的另一个同谋者交换了位置。这个结构化观察创建了一个环境的变化，现在，新的同谋者同那个路人进行交谈（见图4.2）。新的同谋者——和原先的同谋者的身高、声音和衣着都不同——与那个路人进行目光接触。只有大约一半的路人会察觉到这种改变，一半的路人对这种变化视若无睹。

可以像西蒙斯和莱文（1998）研究一样在自然情境中使用结构化观察法，也可以在实验室情境中使用结构化观察法。在进行亲子互动行为评价时，临床心理学家常常采用结构化观察法（例如，Hughes & Haynes, 1978）。动物研究者也经常构建一些观察的情境来引发在自然环境中很难观察到的事件。对于这一点，一个有趣的例子是关于猩猩和人的认知能力差别的一系列研究。（达尔文的观点对吗？它只是一个程度的问题吗？）一种结构化观察法是评价黑猩猩凝视后紧接着的动作（见Povinelli & Bering, 2002）。一个管理人员和猩猩进行眼神交流，然后看猩猩的上方。观察表明，猩猩事实上也随着管理人员的目光的变化（例如，向上看）而变化。但是，关于这个行为的真正意义还存在争议。猩猩是否“懂得”管理人员的心理状态（例如，管理人员看见某些东西时的心理状态）？

发展心理学家也经常使用结构化观察法。让·皮亚杰（Jean Piaget, 1896~1980）或许是其中最著名的一位。在皮亚杰的许多研究中，首先让孩子解决一个问题，然后让他解决此问题的许多其他变形，从而测量其理解能力的极限。观察

图4.3 让·皮亚杰（1896~1980）运用结构化观察法研究儿童的认知发展



者让孩子们透彻理解问题的特性,然后问一些问题从而探明儿童推理的过程。这些结构化观察提供了大量有关儿童认知的信息,是皮亚杰“智力发展阶段理论”建立的基础。皮亚杰型任务也常常被动物研究者用于动物认知研究。有一个类似的研究,在经典的客体永久性测试中,猩猩表征客体永久性(看不见的物体仍然存在)的能力优于松鼠猴(Blois et al., 1998)。

结构化观察法代表一种折衷,一种介于自然情形下的被动型非干预和具有实验室特色的系统操纵及精确控制之间的折衷。这种折衷的优点是,它允许在更自然的情况下进行观察,而不是在实验室人为的条件下进行。不过,这需要付出一些代价。每次进行观察时不遵循相同的程序可能会让他人在研究同样的问题时很难获得相同的结论。未控制的、也许是未知的变量,可能对观察到的行为产生重要影响。例如,不同的观察者运用皮亚杰的结构化观察法进行观察时并不总是遵循相同的程序,这样一来,这种在不同的观察研究中观察程序的不一致性就成为使用这些方法的潜在问题(见 Brainerd, 1978)。

现场实验法

- 在现场实验中,研究者操纵自然情境中的一个或多个自变量以确定它们对行为的影响。

观察者在自然情境中操纵一个或多个自变量以确定它们对行为的影响,这一方法被称为现场实验(field experiment)。之所以在这一章中提到现场实验,是因为它代表了观察法中最极端的干预形式。现场实验与其他观察法的本质区别在于研究者在其中施加了更多的控制。在现场实验中,研究者操纵一个自变量以创造两种或多种条件,并测量自变量对行为的影响。现场实验常用于社会心理学的研究(Bickman, 1976)。例如,为了调查旁观者对犯罪的反应,实验助手装作抢劫者(Latané & Darley, 1970);为了研究正在排队的人的反应,实验助手插入等候的队伍中(Milgram, Liberty, Toledo, & Wackenhut, 1986)。我们将在第7章继续讨论实验方法。

行为记录

- 研究的目的决定研究者是对行为寻求全面的描述,还是只对选定的行为进行描述。
- 如何概括、分析和报告研究结果取决于最初对行为观察的记录。

各种观察法因观察者干预的程度而不同。被动且不做干预的自然观察法处于维度的一端,而对自变量积极操纵的现场实验处于另一端。不同的观察法中行为的记录方式也不同,其中最重要的一个特征是行为从被观察的情境中抽取出来的程度(Willams, 1969, 全面讨论观察法)。研究者有时追求对行为及其发生的情境

拓展训练

请先阅读一个观察研究的描述，然后回答后面的问题。

研究方法课上的同学做了一项观察研究，调查学生学习时集中注意的能力是否受学习地点的影响。具体而言，他们对大学校园里两个场所（图书馆和学生会的一个休息室）的学生进行了观察。他们装作在图书馆或休息室学习以便观察那里的学生。他们只观察每一场所中单独坐着的同学，他们面前的学习材料（如课本或手册）是打开的。观察者以5分钟为单位，记录这期间学生总的学习时间（这种学习表现为查资料或写东西）。观察者确保在观察期间对每位学生只做一次记录。观察者的预期是，在图书馆学习的学生比在学生会休息室中的学生更集中注意力。

在同一天（周一）晚上的9点到11点，5名学生观察者对60个在图书馆的学生和50个在休息室的学生进行了观察。他们发现在5分钟内，图书馆内学生的平均

学习时间是4.4分钟，学生会休息室内的学生平均学习时间是4.5分钟。他们对结果感到惊奇：第一，在5分钟的学习时间内学生们几乎90%的时间都在学习；第二，与预测相反，两个场所的学习时间居然没有差异。

1. 指出学生在研究中使用哪种类型的观察方法，并解释你是根据哪些特征来做出这一判断的。
2. 你认为学生对“注意力”这一变量的操作定义怎么样，它在多大程度上“捕获”了学生的注意力？他们对变量的操作定义怎样促成了这两种研究结果：即学生在两个场所的学习时间都很高而且两个场所间缺少差异？你会提出什么建议以改进“注意力”的操作性定义。
3. 为什么时间抽样计划在这类研究中特别重要？为了提高外部效度，应当如何改进研究中使用的抽样计划。
4. 假设学生在图书馆比在学生会休息室能更好地集中注意力，思考一下：学习材料的不同是否会导致两个场所学生的“注意力”之间没有差异。

进行全面描述，而更多的时候则只集中于某些特定行为或事件。是观察环境中的所有行为，还是只观察所选定的行为？这取决于研究的目的和研究者的意图。如何记录行为，也取决于调查者是做定性研究还是定量研究。定性研究的结果主要以言语描述和逻辑论证的形式呈现，定量研究的报告主要强调数据的统计描述和数据分析，用以支持研究结论。最重要的一点是，如何记录行为决定了最终如何测量、概括、分析和报告研究结果。

行为的定性记录

- 对行为做叙述性的书面记录、录音带和录像带都是比较好的定性记录形式。
- 研究者对叙述性记录的数据进行分类和组织，以检验他们对行为的假设。
- 叙述性记录应该在观察行为期间或者观察后不久进行，观察者必须在接受严格训练后根据已有的标准记录行为。

当研究者试图对行为进行全面记录时，他们经常使用叙述性记录。叙述性记录（narrative records）对行为发生的原始形式提供了不同程度的忠实再现。为了建立一个叙述性记录，观察者可以用文字描述，也可以用录音带、录像带和胶卷来记录行为。生态学家对行为事件的每一个细节都感兴趣，因此他们经常使用录像记录行为（Eibl-Eibesfeldt, 1975）。例如，生态学家通过对自然社会情境中的个体进行67小时的拍摄，研究了社会交互作用中“扬眉动作”的功能（Grammer, Schiefenhoevel, Schleidt, Lorenz, & Eibl-Eibesfeldt, 1988）。他们发现，在三种不

同文化背景中,伴随着微笑的扬眉动作是一种普遍的社会交往信号(如,表示“同意”的信号)。

一旦建立了叙述性记录,研究者就可以对记录进行分类、组织和研究。通过检验数据,可以验证与所观察行为有关的特定假设或预期。叙述性记录与其他测量和记录行为的方式不同,因为它对行为的归类是在观察后完成的。因此,研究者必须确保叙述性记录捕获了用于检验研究假设的信息。

作为儿童攻击性研究的一部分,哈塔普(Hartup, 1974)获得了叙述性记录。他从多个角度研究了儿童攻击性这一问题,包括攻击行为之前的突发事件和攻击性质之间的关系。下面是哈塔普研究中的部分叙述性记录(p.339):

玛丽安(7岁)……正在对所有人报怨大卫(大卫也在场)向她身上洒水,水弄湿了她的裤子。她说“我也要这样对他,看他是否喜欢。”随后她把一个容器装满水。大卫跑到老师那儿叙述玛丽安对他的威胁。老师从玛丽安那里拿走了容器。玛丽安动手打大卫并用力拽他的头发。大卫大声哭喊并且向玛丽安挥舞拳头,老师制止了他,接着把他带上楼……过了一会,玛丽安和伊莱恩走上楼,进入大卫和老师所处的房间。大卫拿起一本书扔向玛丽安,老师要求玛丽安离开,玛丽安用脚踢了大卫,然后离开。大卫大哭并尖叫:“我要离开这儿,他们欺负我。”

哈塔普(1974)要求观察者使用精确的语言描述行为,不对被试的意图、动机和情绪做任何推理。注意上面一段描述没有告诉我们大卫为什么要向玛丽安扔书或玛丽安受到攻击后的感受。哈塔普认为某种先前行为与特定的攻击类型有关。通过严格排除观察者的任何推断或感想,对叙述的编码就不会受到观察者推断的影响。因此,研究人员就能够以更客观的方式对叙述性记录的内容进行分类和编码。

不是所有的叙述性记录都像哈塔普所获得的那样集中,叙述性记录也不总能防止观察者的推理和感想,有时也不总是对行为的全面描述。例如,现场记录(field notes)只包括观察者对被试、事件、环境和行为的连续描述。记者、社会工作者、人类学家、生态学家和其他人都使用现场记录。这些记录并不总包括对每件事的精确记录。观察者记录下特别感兴趣的事件和行为,并且有可能依据观察者特定的知识或专业技能做出解释。例如,一位生态学家可能记录一个物种的行为是怎样与另一物种的行为相类似。现场记录往往被高度个性化(Brandt, 1972),但它们可能比其他类型的叙述性记录更常用。科学记录的有效性取决于内容的准确性和精确性。准确性和精确性严格地依赖对观察者所做的培训,以及观察记录被独立观察者和其他研究方法证实的程度。

实际操作性以及方法学的问题决定了做叙述性记录的方式。叙述性记录应当在行为观察过程中或在观察之后不久立即记录,这是个一般性原则。时间的推移使细节变得模糊不清,并使再现行为的原始顺序更为困难。阿德章(Adjang, 1986)使用便携式盒式录音机对逗惹幼年黑猩猩的研究做了口述叙述性记录;然后,他“尽可能快地”把这些口头报告抄写在纸上(p.139)。当在自然环境中做观察记录时,这一规则有时难以遵循。生态学家试图记录自然状态下生活的动物行为,有时会受到恶劣天气、动物迁徙、白昼缩短等情况的阻碍。记录必须得马上

做——甚至当观察者很忙时也一样。然而有些时候，观察者不得不回到帐篷中去做书面的行为记录。

在观察行为之前，必须决定叙述性记录应当包括哪些内容。例如，我们已经看到，言语叙述可能因观察者推理程度的不同或记录完备性的不同而不同。因此，一定要在开始研究之前决定叙述性记录的内容（如，Brandt, 1972）。叙述性记录的内容一旦决定，一定要训练观察者依据已经建立的标准来记录行为。在收集“真实”数据之前，一定要进行观察练习，并由多个研究者对记录进行评析。

行为的定量测量

- 研究者为了描述具体的行为或事件，经常需要使用定量的测量，例如事件的频率或持续时间。
- 行为的定量测量可以使用四种不同水平的测量量表：称名、顺序、等距和比率。
- 常用来测量心理维度的等级量表常被当作等距量表来使用，尽管通常所做的是顺序测量。
- 在自然环境中可以用电子设备记录行为，有时可以用寻呼机向被试发出信号，提示他们报告自己的行为（如在问卷调查中）。

研究者经常只对个体或环境的某些行为或特定方面感兴趣。他们可能对预期的行为有具体的假设，并对正研究的行为有明确的定义。例如，假设你想做一个自然观察研究，研究正常人对那些有明显身体残疾者的反应。为了做研究，就有必要界定什么是对身体残疾者的“反应”。例如，你对帮助行为、趋近/回避行为、目光接触、交谈时间或一些其他的行为反应感兴趣吗？在考虑使用什么行为来定义人们的“反应”时，也必须决定如何测量。例如，假定你计划通过正常人与残疾者之间的目光接触来测量反应，那么你就需要严格地确定如何测量目光接触。你是简单地测量是否有目光接触？还是测量每次目光接触的持续时间？当然，你可以使用不止一种方法进行测量，只要可行，任何方法都可以作为一种研究策略，选择依据在于研究的特定假设或目标。在做决定时，应当查阅以前发表的、使用了相同或类似的行为测量的研究，并利用从中获得的信息。例如，汤普森（Thompson, 1982）就测量了正常人对身体残疾者的反应，发现这些反应经常可被归类为不受欢迎之列。我们现在要阐述行为测量的四种方式。

测量量表 不同的测量量表决定了不同的定量测量方法，因此，熟悉行为研究中使用的测量量表很重要。测量量表（measurement scales）代表了行为能被量化的不同水平，不同的测量量表影响后续的数据分析及使用。物理学和心理学使用的量表有四种：称名量表、顺序量表、等距量表和比率量表。表 4.1 描述了每种测量量表的特征，更多细节描述参阅专栏 4.2。在为分析研究结果而选择统计方法时，需要谨记这四种测量量表的特点。在这一节中，我们要描述观察研究中如何使用测量量表。

一览表（checklist）常用来记录称名量表测量。观察者可以在一览表上做记

专栏 4.2

“不同水平上的”测量

测量的最低水平为称名量表 (nominal scale), 即把一个事件分到许多不同类别中。例如, 我们可以通过把人们的眼睛分为“棕色眼睛”或“蓝色眼睛”来测量眼睛的颜色。当研究人们对明显身体残疾者的反应时, 研究者可以使用称名量表测量被试对有明显身体残疾的人是否有目光接触。

概括和分析称名量表测量的数据是受限制的。在称名数据中可以执行的惟一数学运算是“相等”和“不相等”的关系。概括称名数据的一种常见的方式是以每一类别中事件的比例或百分比记录频率。

测量的第二个水平称为顺序量表 (ordinal scale)。顺序量表包括对测量事件进行排序或排列等级。顺序量表在测量过程中增加了数学关系“大于”或“小于”。竞赛结果是一个常见的顺序量表。我们知道一个奥运会长跑运动员获得银牌, 但我们不知道的是, 她到达终点时由于与第一名非常接近以致于要通过照片才能判断, 还是落后于金牌选手 200 米。

测量的第三个水平称为等距量表 (interval scale)。等距量表是指两个事件在一个既定维度上相距多远。在顺序量表中, 排列在第一和第三的两个事件之间的差异不一定等于第三和第五两个事件之间的差异。例如, 第

一名和第三名赛跑运动员相差的时间不一定与第三名和第五名运动员相差的时间相同。然而, 在等距量表中, 相同差异的测量值其意义是相同的。例如, 态度测验中 50 和 70 个正确回答之间的差异与 70 和 90 个正确回答之间的差异相等。等距量表中缺少的是有意义的绝对零点。例如, 在言语能力测验中某人的得分是零, 他或她不一定完全没有言语能力 (毕竟, 我们已假定此人有足够的言语能力来做测验)。重要的是, 我们能够在等距量表中测量的数据上进行标准的数学运算 (加、减、乘、除)。因此, 一旦有可能, 心理学家就会使用等距量表测量心理维度。

测量的第四个水平称为比率量表 (ratio scale)。比率量表具有等距量表的所有特点, 并且比率量表有绝对零点。从数学运算的角度来看, 零点使量表数值的比率具有了意义。例如, 摄氏温度计表示的温度代表了测量的等距量表, 摄氏 0 度并不真的意味着绝对没有温度。因此, 说 100 摄氏度是 50 摄氏度的两倍, 或 20 摄氏度是 60 摄氏度的 1/3 是没有意义的。而另一方面, 开氏温标确实有绝对零度, 并且量表数值可以进行有意义的计算。测量时间、重量、距离的物理量表通常被视为比率量表。

录, 例如, 被观察者是否与身体残疾者有目光接触、教室中的儿童有没有在说话、人们是否使用安全带, 等等。一览表经常被用来记录被试的特征——如年龄、种族、性别; 也可以用来记录环境特征, 如时间、场所以及是否有他人在场等。研究者经常对作为被试和环境变量函数的行为感兴趣。男性比女性更多地使用安全带吗? 驾驶便宜汽车的人比驾驶昂贵汽车的人更可能合伙使用汽车吗? 这些问题可以通过观察不同类别的被试和环境 (男性和女性、昂贵汽车和便宜汽车) 中的某种行为 (使用安全带或合伙使用汽车) 是否出现而得到回答。

塔斯纳瑞和汉森 (Tassinari & Hansen, 1998) 使用等距量表测量男女大学本科对女性形体线条画的反应。形体在物理维度上发生变化, 如身高、体重和臀部大小。大学生依据吸引力和生育能力给线条画排列等级。根据进化心理学理论, 女性的吸引力可以定义为身体吸引力和生育能力的综合效应。根据该理论所做的一个预测是, 对吸引力和生育能力两者来说, 腰臀比例都应当是一个不变的知觉线索。与预期相反, 研究的结果显示, 臀部的相对大小和体重与生育能力的等级

表 4.1 测量量表的特征

量表类型	运 算	目 的
称 名	相等/不相等	把刺激分成不同的类别
顺 序	大于/小于	在单一维度对刺激的等级进行排序
等 距	加/乘/减/除	在给定的维度上确定刺激间的距离
比 率	加/乘/减/除/数值的比例形式	在给定的维度上确定刺激间的距离并表示测量数值的比例

呈正相关，而与身体吸引力呈负相关。

为了在研究中对行为进行量化，研究者有时要对行为和事件评定等级。观察者通常根据对一些特质或条件上的主观判断来评定等级（参阅 Brandt, 1972）。例如，迪克（Dickie, 1987）让观察者评定某研究中父母—幼儿之间的相互影响（设计该研究是为了评价一个父母培训项目的效果）。她的观察者拜访了被试的家，并在父母与他们的孩子相互影响时评价母亲和父亲两者的等级。在观察期间，观察者与幼儿坐在房间里，并让父母“尽可能正常地做事——就好像我们（观察者）不在这儿”。她还使用了对每位父母分配角色的结构观察法。评定父母—幼儿在 13 个不同维度上相互影响的等级，包括言语、身体和情绪相互影响的程度。对每个维度来说，该连续体被定义为代表了这一变量在 7 点量表上被评定的不同等级。等级 1 代表特质缺乏或非常少，较高的等级代表特质逐步增多。

表 4.2 列出了观察者在这一研究中使用的一个维度：对儿童的温暖和喜爱。注意，为了帮助观察者对特质的不同程度进行定义，给出了四个奇数等级的精确的言语描述。观察者用偶数（2、4、6）来界定介于两个更明确界定的值之间的事件。研究者发现，参加此“与幼儿相处能力”培训项目的父母在许多变量的评价等级上要高于未被培训的父母。

乍一看，迪克所使用的等级量表似乎代表测量的等距量表。没有绝对零点，且间距似乎相等。实际上，许多研究者把这样的等级系统看做是等距量表。然而，

表 4.2 用于测量父母对幼儿的温暖和喜爱的等级量表的例子*

量表值	描 述
1	缺乏温暖、喜爱和愉快。先前占主导的是过度的敌对、冷漠、疏远和孤立的孩子。关系处于攻击水平。
2	
3	相互交流中偶尔有温暖和愉快。父母对孩子几乎没有表现出自豪感，即使有，也是与孩子的出格和异乎寻常的行为有关。父母的相关举止是设计出来的、理智的，而不是自然流露的。
4	
5	相互交流中有中度的温暖和愉快。父母在某些方面显示出愉快，而在其他方面没有。
6	
7	温暖和愉快是与孩子相互交往的特征。孩子有愉快和自豪的明显表现。对孩子行为的愉快反应是恰当的。

*引自 Jane Dickie 提供的材料

更进一步的检验显示,研究者用来在心理维度上评价人或者事物的多数等级量表,实际上提供的仅是顺序信息。对一个标准的等距评定系统来说,测量等级2与等级3的距离,应当等于等级4与等级5或等级6与等级7的距离。人类观察者对温暖、愉快、攻击或焦虑等特质可以做出主观判断,但以这种方式在等级间产生精确的等距距离几乎是不可能的。然而,大多数研究者在使用等级量表时假定测量为等距水平。决定哪种量表适用于哪种行为的测量并非易事。如果你拿不准,就应当从经验丰富的专家那里寻求建议,这样就能对数据的统计描述和分析做出适宜的决定。

拉夫朗斯和梅奥(LaFrance & Mayo, 1976)提供了一个比率测量的例子。他研究了不同种族的个体在交谈时目光接触的差异。在自然环境下观察黑人组和白人组被试,记录各组中每个成员看其他成员面部的时间。目光接触的持续时间代表了测量的比率水平,因为时间单位(如秒)有相等的间距,而且“零”是一个有意义的值(即没有目光接触)。研究者发现,在听别人说话时黑人盯着另一个人看的时间要少于白人。拉夫朗斯和梅奥认为,目光接触的微妙差异可能是导致社会误解的一个根源。白人谈话者会觉得黑人听众较少的目光接触是不信任或不感兴趣的表现,而这可能仅仅反映了种族间的文化差异。

行为的另一种重要测量是记录事件发生的频数(frequency)。可以使用一览表对相同个体或团体进行重复观察,以测量他们在一段时间内特定行为发生的频数。在每次观察中记录特定的行为是否出现。在这些情况下,可以假定反应的频率代表了测量的比率水平。即,如果计算某种行为“单位”(如儿童离开教室的座位的次数),那么零代表着特定行为没有出现。例如,只要一个有20个单位的个体是有10个单位的个体的2倍,那么量表数值的比率就是有意义的。例如,麦金尼等人(McKinney, Mason, Perkeron, & Clifford, 1975)使用教室行为规范一览表(SCAN),把教室行为分成27个类别。在一定的时间间隔内,观察者重复观察并记录教室中的每个孩子都在做什么。对同一孩子的多个观察记录可以发现这个孩子最频繁发生的行为。研究者发现,把SCAN一览表观察儿童行为获得的分数与从IQ测验中获得的信息联系在一起,比单独使用观察行为信息或IQ分数对儿童的学校成绩的预测更准确。

电子记录和追踪 有时使用电子记录和追踪装置测量行为。例如,在关于大学生应对策略与血压关系的一项研究中,被试佩戴了可随身携带的血压监视器(Dolan, Sherwood, & Light, 1992)。大学生在两个“具有代表性的”在校日子戴着电子记录装置,其中的一天进行一场考试。被试还完成了一个关于应对策略的问卷,并详细记录他们每天的日常活动。研究者比较了一天中不同时间的血压读数,并把它作为应对风格的函数。在考试过程中或考试后,归类为“高自我应对策略”的学生(即那些在压力环境中显示出“自我负责和/或责备自己”倾向的学生, p.233)比那些低自我应对策略的学生有更高的血压反应。电子记录在研究中的使用可能会随着个人数字化助手(PDAs)技术的发展而增加。

研究者还使用了一种被称为经历抽样法(experience sampling method, ESM)的创新程序。他们给被试提供电子寻呼机,时间通常为一周,要求被试在寻呼机给出信号时汇报他们的活动(如, Csikszentimihalyi & Larson, 1987; Larson,

1989)。在一项关于青少年与家庭成员相互关系的研究中,220个来自中产和工人阶级的年轻人提供了他们生活中5年级到8年级和9年级到12年级这两个时期的报告(Larson et al., 1996)。当被随机地“呼叫”时,这些年轻人根据核查表的提示回答他们和谁在一起,并在一个7点量表上回答他们的情绪状态和对与他们在一起的人的感觉,还要回答有关他们活动的开放式问题。正如预期,数据显示,从5年级到12年级这些年轻人与父母在一起的时间呈急剧减少趋势。然而,令人吃惊的是,家庭冲突的数量与在家时间两者之间没有关系,这一点没有在其他灵长类动物研究中看到。在其他灵长类动物中冲突经常导致从家庭中分离出去。研究者提出,对于人类青少年来说,“家庭负面经历可能是一个刺激,它会促使成员更注重家庭中的交往——这一点在一个充满矛盾和麻烦的家庭中可以看到——就好像这个刺激能把他们再次拉回家庭一样”(p.752)。ESM程序也被用于研究慢性精神病人的情绪状态(如Delespaul, 1995)。

帕克、阿莫莱和泰恩(Park, Armeli, & Tennen, 2004)使用“互联网日记法”对康涅狄格大学本科生的压力和应对进行了测量。总共190个学生在28天内每天登陆安全的互联网网站参加这项研究。(每天发E-mail提示)被试记录每天最大的压力事件,然后评定自己对事件的控制程度,并回答一些有关他们情绪和应对策略的问题。积极的情绪更多地与以问题为中心的应对策略相关,而不是与回避策略相关,特别是当被试认为压力可以控制时。

ESM和日记法都依赖于被试对情绪和活动的自我陈述,而不是对他们行为的直接观察。因此,设计一种技术来识别数据集中的偏差(如,被试对个人活动可能存在遗漏和歪曲)就很重要了(参阅Larson, 1989,使用ESM方法可能出现偏差的讨论)。这些问题需要权衡直接观察与非直接观察两种方法在时间和人力上的花费与收益(如,Barker, Wright, Schoggen, & Barker, 1978)。

观察数据的分析

数据简化

- 观察法所得到的数据可以通过简化的过程加以概括。
- 研究者依据特定的标准(例如,通过对行为进行分类)对行为进行编码,把叙述性记录中的数据量化。
- 可以使用描述性测量(如频数、平均数和标准差)对数据加以概括。

描述性记录的分析 描述性记录可以提供自然环境中丰富的行为信息。信息收集之后,研究者如何进行概括?在分析描述性记录的内容时,简化数据常常是重要的一步。**数据简化(data reduction)**是抽象和概括行为数据的过程。研究者总是力图使用定性分析对他们的观察记录进行言语概括,并形成一种理论,以解释描述性记录中的行为。定性分析中,在研究者对信息进行言语概括、确定主题、分类、分组,并记录他们自己对描述性记录的观察时就产生了数据简化。

数据简化经常涉及编码(coding)过程,即依据特定标准对行为或特定事件进行确认。例如,哈塔普(1974)对儿童攻击行为进行为期10周的观察后发现了758条攻击信息。作为学前儿童行为学研究的一部分,麦克格鲁(McGrew, 1972)确认了115种不同的行为模式。他依据身体部位建立了分类行为模式的编码图式,范围从面部表情(如呲牙咧嘴、露齿而笑的脸、脸皱了起来)到运动行为(如飞奔、爬、跑、跳和踏步)。编码者在观看幼儿园孩子们的录像时使用编码图式对这些行为进行分类。

编码通常以与研究目的有关的行为或事件为基础。例如,哈塔普使用9种分类对攻击事件的性质进行编码。用编码来简化数据,可以使观察者确定特定行为类型与这些行为的先前事件之间的关系。例如,麦克格鲁(1972)发现,儿童在争夺玩具失败后会露出“撅嘴表情”。这位行为主义观察者注意到,幼年黑猩猩在试图与母亲重聚时会露出类似的表情。儿童经常在刚受挫之后(并且经常在刚要哭泣之前)“脸会皱起来”。有趣的是,在非人类的灵长目动物里似乎没有这种表情的记录。

描述性测量 当运用定量数据分析时,会使用描述性测量对观察数据进行概括。当事件被分到不相容的类别中时(称名量表),最常用的描述测量是相对频数。各种行为发生的次数与观察到的事件总次数的比率就是相对频数测量。相对频数测量以比例或百分数表示。例如,詹妮等人(Jenni & Jenni, 1976)观察了6所大学校园里的学生,她们报告82%的女大学生用一只或两只胳膊抱着书(或者是将书的一小部分靠在臀部,或者是放在身体前面),只有3%的男大学生使用这种特殊的拿书方法!

用等距量表或更高级的比率量表记录行为时,就会使用到包含信息更丰富的描述性统计。用等距量表记录观察结果,或用比率量表测量时间(持续时间、反应时)时,会使用一种或多种集中趋势测量。最常用的集中趋势测量是算术平均数,或简称平均数。平均数描述了一组分数中的“代表性”分数,并提供一种有效测量来概括一群人的成绩。为了更完整地描述一群人的成绩,研究者要报告关于离散性的测量,或分数围绕平均数的离差。标准差近似为一个分数与平均数的平均距离。

现在是复习集中趋势和离散性测量的好时机,也是系统复习数据分析一般准则的好时机。第12章的前面几页将专门探讨这些问题。

拉夫朗斯和梅奥(1976)研究了交谈中同种族的黑人和白人组之间目光接触的平均数和标准差。他们记录各组中的每位听众注视说话者面部的时间。表4.3中的平均数显示,白人听众看白人说话者面部的时间多于黑人听众看黑人说话者面部的时间。相同性别组和混合性别组都得到这一结果。标准差表明男性组与女性组或混合组相比显示出较低的离散性。集中趋势和离散性测量有效概括了这一研究中大量的观察记录。

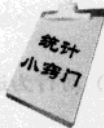


表 4.3 听者注视说话者面部的平均时间(秒)及标准差*

组 别	平均数	标准差
黑人谈话者		
男子组	19.3	6.9
女子组	28.4	10.2
男子—女子组	24.9	11.6
白人谈话者		
男子组	35.8	8.6
女子组	39.9	10.7
男子—女子组	29.9	11.2

* 引自拉夫朗斯和梅奥(1976)。

观察者信度

- 观察者间信度指不同的独立观察者针对同一观察所做记录的一致性程度。
- 提高观察者间信度可以采用明确界定所观察行为和事件、培训观察者和提供差异性反馈等手段。
- 观察者间信度高有助于提高研究者对行为观察准确性(有效性)的信心。
- 观察者间信度是通过计算一致性或相关的程度来评估的,它取决于如何测量和记录行为。

分析观察数据的另一重要方面是评价观察的信度。除非观察是可信的,否则就不能获得任何有意义的信息。研究者评价观察者信度的一种方法是问:“观察相同事件的另一位观察者也得到了相同的结果了吗?”

观察者间信度 两个独立的观察者相一致的程度被称为观察者间信度(interobserver reliability)。当观察者记录不一致时,我们会不确定正在测量什么以及实际上发生了什么行为和事件。当没有明确界定要记录的事件时,可能导致低观察者间信度。想像一下,如果哈塔普(1974)让观察者记录儿童的攻击事件而没有确切定义攻击性会如何?究竟什么是攻击性?一些观察者可能把攻击性定义为一个儿童对另一个儿童的身体打击;其他观察者可能把攻击性定义为言语侮辱。什么是顽皮地推,什么是愤怒地操?没有明确定义要记录的行为或事件,使得观察者不总一致,从而降低了观察者间信度。除了提供精确的言语定义,还要给出一种现象的具体例子,这样通常有助于提高观察者之间的信度。向观察者呈现攻击和非攻击事件的照片或录像带,这样可以提高区分攻击行为的能力。通过培训观察者以及让他们做观察练习也能提高观察者信度。在培训以及练习期间,向观察者提供与其他观察者有差异的反馈也能提高观察者信度(Judd, Smith, & Kidder, 1991)。

高度可靠的观察者的观察不必一定是精确的观察。试想两名观察者看到的事情确实一致,但两者“错误的”程度相同,两个观察者都不能提供精确的行为记录。例如,两者可能都以相似的方式,受其观察研究所期望的结果影响。媒体有

时会报道几个观察者宣称他们看到了同样的事物（如不明飞行物 UFO），而结果证明，他们与其他人看到的事物或物体是一样的，只是说法不同而已（如一个气象气球）。而且，当两个独立的观察者一致时，我们一般会更倾向于相信他们的观察要比基于单个观察者观察的数据更为准确和有效。为了确保观察者的独立性，每位观察者不能知道其他人记录了什么。两个观察者受到结果预期、疲劳或厌烦影响的程度一般不会完全相同，所以我们可以确信所报告的事情确实发生了。当然，越多的独立观察者一致，我们也越确信。

信度测量 评价观察者间信度的方法取决于如何测量行为。当事件根据不相容的种类（称名量表）分类时，通常使用一致性百分比评价观察者信度。计算观察者之间一致性百分比的公式是：

$$\frac{\text{两个观察者一致的次数}}{\text{总的观察次数}} \times 100$$

哈塔普（1974）使用一致性百分比指标比较了法官们的观察信度（法官们依据攻击类型和先前事件的性质对描述性记录进行编码），该一致性百分比在 83% 和 94% 之间。尽管没有严格的规定说明低观察者信度的数值，但在已发表的文献中，研究者一般报告的信度估计值超过 85%，这表明比之更低的一致性是不被接受的。

在许多观察研究中，数据是由几个观察者在不同的观察时间内收集的。在这些情况下，研究者只使用一个观察样本来测量信度。例如，要求两名观察者根据时间抽样程序记录行为，在这个时间子集中两个观察者都在场。两个观察者都在场时的一致性大小可以表示整个研究的可信度。

当观察数据至少是等距测量时，如时间作为被测量的变量时，观察者信度可以用皮尔逊积差相关系数 r 评价。例如，拉夫朗斯和梅奥（1976）获得了观察者记录交谈中听众盯着说话者面部的时间的信度测量，他们发现记录目光接触时间的两组观察者之间的平均相关是 0.92，则该研究中的观察者信度是令人满意的。

当对相同的人、事件或事物所做的两个不同的测量一起变化，即一个变量的分数与另一个变量的分数共变时，我们说存在相关。**相关系数**（correlation coefficient）是这一共变程度的量化指标。如同上面提到的，收集等距或比率数据时，可以使用皮尔逊相关系数 r 获得观察者间信度的测量。这一测量告诉我们两位观察者一致的程度。

相关系数表示相关的方向和强度。方向可以是正的或负的。正相关表示当一个测量的值增加时，另一个测量的值也随之增加。（例如，吸烟和肺癌的测量是正相关。）负相关表示当一个测量的值随之增加时，另一个测量的值减少。（例如，看电视的时间和数学测验的分数是负相关。）很明显，评价观察者间信度时我们是在寻找正相关。相关的强度是指存在共变的程度（或者有时这样说，预测关系的强度），因为相关是对行为和事件进行预测的基础。（第 5 章和第 12 章对相关进行了更充分的讨论。）

相关系数大小在 -1.00（完全负相关）到 1.00（完全正相关）之间变化。数值 0 表示两个变量之间没有相关（因此没有做预测的基础）。相关系数越接近 1.0 或 -1.0，两个变量之间的相关越强。请注意，相关的符号只表示它的方向；相关系数 -0.46 比 0.20 表示更强的相关。我们主张，超过 0.85 的观察者间信度的测量表示令人满意的观察者一致性

(当然再大更好!)。如同你已看到的,拉夫朗斯和梅奥(1976)报告了平均0.92的观察者内部相关,显示出非常好的观察者一致性。

在第12章我们会更充分地讨论相关,包括如何用散点图描述两个变量的关系,如何计算皮尔逊积差相关系数,以及如何很好地解释相关的意义。如果你想深入了解相关,请参考第12章的内容。

批判性思考观察研究

观察者的影响

- 如果个体知道他们正在被观察而改变了自己的行为(“反应性”),这样的行为将不再代表一般行为。
- 被试可能会根据研究情境中的需求特征,引导他们的行为做出反应。
- 控制反应性的方法包括隐蔽(非反应性)测量、适应(习惯化、脱敏)和间接观察。
- 研究者在试图控制被试反应性时必须考虑道德问题。

要想做好观察研究就必须考虑如何对行为和事件抽样、如何选择适宜的观察方法,以及如何记录和分析观察数据。在了解观察方法的基本要点之后,我们还需要了解可能产生的问题。当人们知道自己正在被观察时经常会改变行为,这是第一个问题。观察者的偏差影响他们对行为的选择记录,这是第二个问题。

反应性 观察者在场会导致人们改变行为,因为他们知道自己正在被观察。当观察者影响观察的行为时,反应性(reactivity)问题就出现了。当观察者在场时的个体做出“反应”,他们的行为可能不再代表观察者在场时的行为。安德伍德和肖内西(Underwood & Shaughnessy, 1975)讲述了一个学生观察人们在有停车标志的交叉路口是否把车停下来的实验(这是一个课堂作业)。观察者手拿笔记本站在街角,不久他注意到所有的汽车都停在停车标志处。后来他意识到自己的存在可能正在影响司机的行为。当他在交叉路口附近隐藏起来时,他发现司机的行为发生了改变,他可以收集数据来检验他的假设了。

研究中当被试注意到他们的行为正在被观察时,会以非常微妙的方式做出反应。例如,被试在参加心理学研究有时会感到不安和焦虑。测量唤醒水平,如心率和皮肤电反应(GSR),可以简单显示出观察者在场时的效应。让被试戴上一个电子寻呼机(它用来提醒被试何时记录行为和情绪)也可能影响被试行为(如Larson, 1989)。

观察者在场时,被试经常试图以他们想像中研究者所期望的方式做出反应。他们知道自己是科学研究的一部分,所以通常想要合作并且成为“好”被试。他们经常试图猜测什么行为是被期待的,并且会使用所获得线索和其他信息指导自

己的行为 (Orne, 1962)。研究情境中的这些线索被称为**需求特征** (demand characteristics)。奥恩 (Orne, 1962) 提出, 个体通常问自己: “我在这儿被期望做些什么?” 为了回答这一问题, 被试会注意环境中出现的线索、研究程序本身及研究者给出的隐含线索。当被试试图猜测什么是被期待的时候, 他们会相应改变自己的行为。被试对研究情境中需求特征的反应会威胁到心理研究的外部效度。当被试的行为方式不能代表其在心理研究情景之外的行为时, 研究结果的推广 (外部效度) 和对研究结果的解释也受到了潜在的威胁。因为被试可能会无意地使一个研究变量看起来比实际更有效, 以至抵消另一个变量的作用。可以通过限制个体对他们在研究中角色的了解以及他们对本研究要检验的假设的了解来减少需求特征的影响。研究者不让被试了解研究细节的目的是为了获得更具代表性的行为。用这种方法解决需求特征的问题确实存在道德问题, 比如知情同意问题。

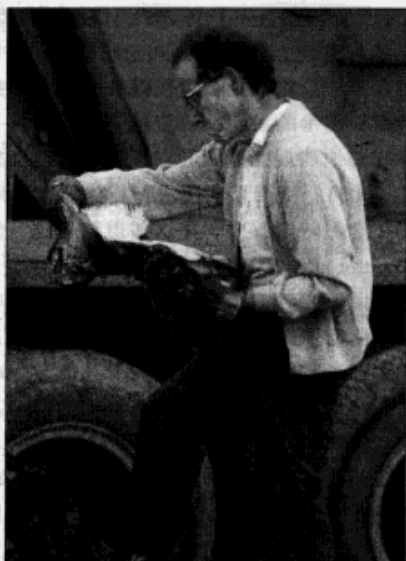
控制反应性 研究者可以用若干种方法来控制反应性问题。他们可以不让被试发觉观察者在场来消除反应性。当被试不知道他们正在被观察时对其行为的测量称为**隐蔽** (非反应性) 测量。隐蔽测量可能需要隐蔽观察者或机械记录设备, 如磁带录音机和录像机。拉夫朗斯和梅奥 (1976) 背着交谈的人对他们进行了观察。观察是在各种自然环境中进行的, 比如饭店和候车室。我们能够想像, 观察者为了获得行为的隐蔽测量, 必须把秒表和记录表隐藏在菜单和盆栽植物后边。然而对研究者来说, 另一种方法是采用伪装的参与性观察, 作为情境中的一个角色, 而不是作为观察者。你可能记得社会心理学家用这一程序研究了个体的行为, 那些个体宣称他们与太空的外星人有联系 (Festinger et al., 1956)。研究者使用隐蔽测量时, 由于被试不知道观察者在场, 可以假定他们会像平常一样反应。在第 6 章, 我们会阐述使用隐蔽测量的研究方法。

研究者处理反应性的另一种方法是让被试适应观察者在场。这样的假定是合理的: 当被试适应了观察者在场时, 他们就会逐渐显示出正常行为。适应可以通过习惯化或脱敏完成。在习惯化程序中, 观察者在多种不同场合简单地把自已引入一个情境, 直到被试不再对他们的在场做出反应。在 20 世纪 70 年代早期, 公共电视频道放映了一部题名《一个美国家庭》的记录片。为了拍摄这部片子, 研究者 (带着他们的摄像机) 确确实实地进入一个加利福尼亚家庭, 并记录了这个家庭 7 个月的活动。尽管不可能区分他们的行为多大程度上受到了观察者在场的影响, 但是显露的事件以及家庭成员所做出的评定都表明发生了习惯化过程。在拍摄过程中, 这个家庭解体, 母亲要求父亲离开家。当把离婚一事通告给成千上万的电视观众后对家庭进行访谈时, 父亲承认他们本可以让拍摄组离开, 但是那时, “我们已经适应了” (《新闻周刊》, 1973, p.49)。

作为一种控制反应性的方法, 脱敏与在恐惧症行为治疗中使用的脱敏方法类似。在治疗情境中, 有特定恐惧 (如对蜘蛛的不合理恐惧) 的个体首先面对一个强度非常低的恐惧刺激。可能会要求病人想像与蜘蛛有关的事物, 比如满是灰尘的房间和蜘蛛网, 同时治疗师帮助病人放松。刺激的强度逐渐增强, 直到病人能够忍受真实的刺激本身。

生态学家常使用脱敏方法使动物被试适应观察者在场的情形。在去世之前,

图4.4 对人们行为的隐蔽（非反应性）测量可以通过寻找他们的废弃物进行物理痕迹的研究，但必须考虑与隐私相关的道德问题



佛西 (Fossey, 1981, 1983) 在她深爱的学科领域对非洲的山地大猩猩进行了卓越的观察研究。在一段时间内她逐渐地向大猩猩靠近，所以它们习惯了她在场。她发现通过模仿它们的动作——比如用力咀嚼叶子和抓挠自己——可以让大猩猩更加自在。最后她能够坐在大猩猩中间，并在它们触摸她和考察她使用的研究设备时对它们进行观察。

最后，行为的隐蔽测量可以通过间接观察行为而获得 (Webb, Campbell, Schwartz, Sechrest, & Grove, 1981)，这包括检查物理痕迹或检查档案信息。档案信息是由社会保存的有关个人和事件的记录。一位研究者通过计算垃圾桶中的空酒瓶数调查生活在官方“禁酒”的城市中的人们的饮酒行为 (参见图4.4)。另一位研究者使用图书馆保存的记录评价了电视的引入对一个社区的影响，他发现读者借阅小说的记录减少了，但对非小说的需求没有影响 (参阅 Webb et al., 1981)。物理示踪物和档案数据是重要的隐蔽测量，它们能成为非常有价值的信息资源，这些方法将在第6章更详细地介绍。

道德问题 无论何时，研究者试图通过背着人们对其进行观察的方法来控制反应性，都会出现严重的道德问题。例如，未经同意就对人们观察可能会严重侵犯其隐私权。判定什么构成了隐私侵犯并非易事 (如我们在第3章讨论的)，而且必须考虑到信息的敏感性、观察的环境以及所获得信息的传播方法 (如 Deiner & Crandall, 1978)。

如同我们已经看到的，最近使用互联网进行的行为研究遇到了新的道德难题。当研究者假扮成参与者进入互联网聊天室，旨在弄清什么使种族主义者提倡种族暴力时 (Glaser et al., 2002)，他们获得的信息可能被看做是背着回复者收集犯罪证据，这与“刺探”没有什么不同。当然，问题在于如果获得了知情同意，那么

回复者就可能非常不合作。在另一个例子中,机构审查委员会(IRB)批准了该研究,同意研究者在聊天室建立一个“公共论坛”。在这个论坛里这些话题是常见的,并且研究者建立了充分的安全装置以保护回复者的身份(如,仔细地把名字与回答分离,这个聊天室中的人们一般使用假名)。另一方面,在这几个例子中,当人们知道研究者背着他们观察他们的在线讨论时,会感到隐私受到了侵犯(参阅 Skitka & Sargis, 2005)。使用互联网进行行为研究刚刚开始,研究者和机构审查委员会的成员都在学习阶段,要解决这些道德问题,他们都需要学会创造性地解决问题(参阅 Kraut et al., 2004)。

当个体参与一个由研究者精心安排的情境中时(这可能发生在结构化观察或现场实验中),可能会出现把被试置于风险之中的有关道德问题。例如,有一个研究,调查大学生对种族骚扰的态度如何受其他学生或支持或谴责种族主义的影响(Blanchard, Crandall, Brigham, & Vaughn, 1994)。来自不同大学的200多名白人女大学生是“真被试”。当这些女大学生穿过校园时,一名白人访谈者接近她们,并请她们回答一系列简短问题,如“她们的学院对种族主义行为如何反应”(p.994)。一名女实验助手(装作一名学生)接近访谈者同时假装成被试。访谈者问这两名“学生”相同的五个问题,然而,访谈者总是先提问那名实验助手。这时,实验助手的回答是或支持或谴责种族主义者的行为。真正有意义的是这些回答对真被试回答相同问题的影响。结果很明确:相对于没有影响的控制组来说,听到另一名学生谴责种族主义行为的被试产生了更多的责备反应,听到另一名学生支持种族主义行为的被试产生了更多的谅解反应。因而,如同作者提出的,该结果“意味着一些直言不讳的人会在不同方向上影响种族互动的正常气氛”(p.997)。

真被试“处于风险中”吗?如果你认为被试处于风险中,风险程度有多大?研究的目的是可能获得的知识会比研究中涉及的风险更重要吗?尽管在这一研究中向被试“立即做了事后解释”,但是这足以使天真的学生们面对种族主义言论时能处理好自己的行为吗?能重建因欺骗而失去的对科学研究的信任吗?这些问题显示出道德判定的难度。(为了回答这些道德问题,第3章结尾列出的道德判定过程中推荐的方法可能会有帮助。)

观察者偏差

- 当观察者的偏见决定了选择哪些行为进行观察,当观察者的期望在确定和记录行为时导致了系统误差,这时,观察者偏差便产生了。
- 当观察者知道研究结果的假设或以前的研究结果时,会产生期望效应。
- 控制观察者偏差的第一步是承认它可能存在。
- 可以通过不让观察者知道研究的目的和假设(“盲”)而减少观察者偏差。

在本章的开始我们描述了罗森汉(1973)和他的同事们的一个研究,在研究中他们观察了精神病院里的工作人员和病人之间的相互作用,并且发现了在工作人员方面出现的严重偏差。病人一旦被贴上精神分裂症的标签,人们就只会根据

这一标签对他们的行为进行解释。心智健全者的正常行为表现,被工作人员解释为是病人精神错乱的证据。研究者后来得知,参与性观察者公开做的记录,被工作人员当作那些伪病人的病态例子加以引用。因此,工作人员倾向于根据病人的标签解释他们的行为。这个例子清楚地解释了观察者偏差(observer bias)——由观察者期望导致的观察过程中的系统误差——的可能风险。

期望效应 在许多科学研究中,观察者对在一个特定情境中或一个特定的心理治疗之后的行为会有一些期望。这些期望可能基于过去的研究结果而产生,也可能通过观察者自己对这一情境中的行为的假设而产生。如果期望导致了观察中的系统误差,它们可能就是观察者偏差——期望效应——的一个来源(Rosenthal, 1966, 1976)。

科达罗和爱森(Cordaro & Ison, 1963)设计了一项研究证明了期望效应。这一研究要求大学生观察者记录两组扁形虫转头和身体收缩的次数。研究者引导观察者为两组扁虫的转头和收缩率产生不同的期望。实际上,两组虫子基本相同,只是观察者对它们的期望不同而已。结果显示,高期望率时观察者记录的转头数是低期望率时的两倍,所记录的身体收缩次数是低期望率时的三倍。很明显,学生们根据自己的期望对虫子的反应进行了不同的解释。

其他偏差 观察者对于研究结果的期望不是观察者偏差惟一的来源。你或许认为使用自动装置(如电影摄像机)会消除观察者偏差。尽管自动化装置降低了观察者偏差的机会,却不一定能消除。考虑一下实际情况,为了用胶卷记录行为,观察者必须决定拍摄的角度、位置和时间。观察者的个人偏差对研究的这些方面都会产生影响,在某种程度上,这样的决定可能将系统误差引入到研究结果中。阿尔特曼(Altmann, 1974)描述了对动物行为的一项观察研究。在这项研究中,每当动物不活动时观察者就会午间休息,这使结果产生了偏差,明显地缺少对动物们不活动期间的观察记录。此外,使用自动化装置一般只会拖延分类和解释的过程,并且在叙述性记录进行编码和分析时很有可能导致观察者偏差效应。

控制观察者偏差 观察者偏差不能被消除,但可以用几种方法降低其发生的概率。如同我们提到的,使用自动化装置可能会有用,尽管仍然可能存在偏差。处理观察者偏差的最重要因素可能是意识到它的存在。即了解这一偏差的观察者更可能采取措施降低偏差的作用。

也可以通过限制观察者所获得的信息来减少观察者偏差。当哈塔普(1974)分析他对儿童攻击行为的观察研究结果时,不允许做分析的个体看到先前的叙述性记录。当对攻击行为的性质分类时,先前事件被抹去;当对先前事件编码时,攻击行为的性质被抹去。因此,在做分类时,编码者不会受到与他们正编码的事件有关的信息的影响。在某种意义上说,编码者对研究的某些方面是“盲的”。当观察者不知道为什么做观察或不知道研究的目的时,他们是“盲的”。此外,拉夫朗斯和梅奥(1976)调查自然环境中交谈的个体之间的目光接触时,观察者不知道他们探测的是种族的差异。使用“盲”观察者极大地降低了观察者期望导致的系统误差。

小 结

研究者很少能观察发生的所有行为。因此,研究者必须采取某种行为抽样模式,比如,时间抽样和情境抽样。抽样的一个重要目的是获得行为的代表性样本。观察方法可以根据两个维度分类:观察者干预的程度以及记录行为的方式。自然环境中没有观察者干预的观察被称为自然观察。有干预的观察可以采取参与性观察、结构化观察(发展心理学家常用)和现场实验(社会心理学家常用)。在观察研究中,可以对行为进行全面描述,也可以只记录某种预先界定的行为的某些方面。当需要对行为进行全面描述时采用叙述性记录,当研究者对某一特定行为是否发生(以及在什么条件下发生)感兴趣时采用一览表。行为的频率、持续时间和等级是观察研究中常用的因变量。

怎样描述和分析定量数据取决于使用什么样的测量量表。心理学家使用的四种测量量表是称名量表、顺序量表、等距量表和比率量表。当做叙述性记录时,通常使用某种编码系统作为数据简化过程的一步。比较典型的是使用描述统计(如平均数和标准差)对测量的频率、持续时间和等级加以概括。在报告观察研究的结果时,有必要提供观察者信度的测量。根据所使用的测量的水平,可以使用一致性百分比或相关系数来评价信度。

在任何观察研究中,都必须控制反应性或观察者偏差可能导致的问题。最后,研究者在开始一项调查研究之前必须重视道德问题。当观察研究涉及到某种形式的欺骗时(如伪装的参与性观察或使用不公开身份的实验助手),道德问题尤其突出。互联网研究提出了新的道德难题,这需要研究者和机构审查委员会成员共同努力解决。

重要概念

外部效度	测量量表
时间抽样	数据简化
情境抽样	编 码
自然观察法	观察者间信度
参与性观察	相关系数
结构化观察	反应性
现场实验	需求特征
叙述性记录	观察者偏差

复习思考题

1. 指出科学观察不同于日常观察的三个特征。
2. 解释研究者在观察研究中为什么使用抽样法,并描述正确抽样是为了达到什么目的?
3. 如何根据干预的程度和记录行为的方法对观察方法进行分类?
4. 描述一个出于道德考虑不允许研究者对人类行为的某些方面施加控制的研究情

境,并说明此时自然观察法是有用的。

5. 指出参与性观察中研究者需要考虑哪三种因素以确定观察者对被观察行为的影响程度。

6. 结构化观察法代表了自然观察法和实验室实验的一个折衷,这一折衷的主要优点和可能代价是什么?

7. 使用四种测量量表的一种,举例说明研究者如何测量交谈中人们彼此间的目光接触。

8. (a) 当使用称名量表测量时和 (b) 当行为是用等距及等比量表测量时,最常用的描述性测量是什么?

9. 描述以下每一因素对观察者间信度的影响:被观察事件的定义、训练、反馈练习。

10. 通过了解相关系数的符号和数值,你获得了哪两种信息?

11. 指出需要以相关系数评价观察者间信度的测量量表,并解释在这种情况下负相关意味着什么。

12. 高观察者间信度是否能保证观察的准确、有效?

13. 被试对需求特征的反应为什么会威胁到心理研究的外部效度以及威胁到对研究结果的解释?

14. 当研究者使用隐蔽测量时(如隐瞒观察者在场),隐私和风险等道德问题是如何出现的?

15. 描述心理学研究中观察者偏差(期望效应)产生的两种方式。

16. 减少观察者偏差的最好方法是什么?

挑战性问题

1. 发展心理学实验课上的同学做了一个父母与婴儿在家中相互影响的观察研究。在4天时间里,他们每天都要观察一个指定的家庭,与父母和婴儿(以及家里的其他孩子)打招呼。他们告诉被观察家庭成员按照平常的样子行事,并对当天的活动提出一系列问题,以确定当天是“正常”的一天或者是有一些不同寻常的事发生了。学生们努力让家庭成员感到舒适自在,但他们也要努力减小与家庭成员间以及学生彼此间的相互影响。在任意给定的2小时观察期间,每个被观察者家庭一直有两名观察者在场,他们相互独立地做记录。将6对学生随机分配到12个家庭中(每一对观察两个家庭),12个家庭是自愿参加研究的。同一对观察者一直观察一个指定家庭,以获得那个家庭整整8小时的观察记录。观察者使用等级量表在不同维度上记录行为,如父母—婴儿相互影响中的温情和喜爱

的表示。

A. 指出学生们用于保证观察结果信度而采用的两种具体办法。

B. 指出对研究结果外部效度的可能威胁;并从提供的描述中列举一个具体的例子。

C. 指出学生们的研究过程中,能够表明其对反应性问题敏感的一个具体方面。他们用来处理这个反应性问题的方法是什么?

2. 大学生在由大学资助的一个酒馆中做了一项观察研究,评价环境对饮酒的影响。他们观察了21岁以上的82名学生。观察者使用一览表记录被试是男性还是女性,以及被试是与另一个人在一起还是与由两人或多人组成的团体在一起。他们周一到周六每天做观察,每段观察期都是从下午3点到凌晨1点,观察为期3个月。在任一观察期间,两名观察者都在场。从

(续)

每位被试(他或她)第一次点啤酒开始,对其进行1小时的观察。记录每小时喝的啤酒数。结果显示,男性比女性喝得多,喝得快。男性与其他男性在一起时喝得更快,女性在与其他男性在一起时喝得也较快。男性和女性在团体中时都比只与一个人在一起时喝得多。这些结果确实表明,饮酒发生的环境对饮酒的性质和程度起着重要作用。

- 指出该研究使用的观察方法,并解释为什么选择该观察方法。
 - 指出该研究的自变量和因变量,并描述自变量每一水平的操作定义。
 - 该研究的研究者是怎样控制反应性问题的?他们的方法中可能出现什么道德问题?
 - 指出该研究中可能提高观察信度的方法。
 - 指出该研究可能降低研究结果的外部效度的方法。
3. 你的一位朋友特别相信交谈时友好所带来的积极影响。他根据每日的观察得出了这一结论。你说服他要证实他的假设需要一个系统的观察。你的朋友(一直微笑着)对友好交谈进行了操作定义,记录他参加的50次交谈时对方的态度(用等级做评价)。结果显示,75%的交谈被评价为“非常友好”,20%被评价为“友好”,5%被评价为“中等”。你的朋友回到你的身边——现在他更加确信友好气氛对交谈有积极的作用。尽管你的朋友会对你不满,但你还是得向他解释,作为证实他观点的研究并不可靠。在你给朋友所做研究评论中一定要提到与研究有关的以下问题:
- 解释在他宣称友好对交谈有积极影响之前他需要什么样的对照观察组。解释这一对照观察组甚至会导致这样的结论:即你的朋友对友好性交谈有消极影响。

致这样的结论:即你的朋友对友好性交谈有消极影响。

- 解释为什么你的朋友不能是观察交谈内容的人,也不能是记录交谈友好程度的人。
 - 对你朋友研究所做的评论能使你推断出你的朋友对他参加的交谈没有积极作用吗?为什么?
4. 四名学生在他们大学的社会科学研究院做实习生。研究院有一个合同,要为大学附近的一个小城商业区发展机构做一系列交通安全的研究。分配实习的学生做其中的一个。具体而言,他们要做一个研究,以确定汽车在商业区人行横道带有停车标志的交叉路口是否确实把车停下来。你需要回答学生们在计划他们的研究时所考虑的以下问题。
- 学生们不想仅用“是”或“否”的分类来区分汽车是否停车。他们应当怎样形成汽车停车的操作性定义,这个定义应当包括汽车完全停下来、慢速停车以及根本不停车。
 - 在开始为实际的研究收集证据之前,学生们应当采取什么步骤以提高观察者间信度。
 - 学生们对确定商业区步行街行人稀少时和拥挤时汽车停车的可能性感兴趣。为了研究这一点,学生们应使用什么样的时间抽样计划?
 - 学生们感兴趣的是每辆车是否可以不受其他车辆停与否则自己能够停下来。为了研究汽车的独自停车,学生们应该怎样对所观察的汽车抽样?学生们应该记录什么信息就可以让他们顾及到抽样的所有汽车,又能确定汽车独自停车的可能性?

拓展练习的答案

- 这一研究中学生们使用了自然观察法。他们没有干预观察的情境,研究也是在自然环境中(图书馆和学生会)进行的。
- 学生们选择使用5分钟的观察时间,这可能限制了他们测量学生注意力的有效性。5分钟的观察时间(特别是如果观察者对5分钟的时距开始计时时,学生正在查找材料或写字)可能太短而不能“显示”注意的转移。如果这样不能显示出注意的转移,那么两个场所间就不可能观察到注意的差异。改进该操作定义的一种可能方法就是,把观察时距延长到20或30分钟。
- 学生们集中注意的能力可能因不同日期和一天中不同时间段而不同。学生们选择在星期日晚上9到11点观察是因为他们认为应当在学生的“黄金”学习时间观察。时间抽样计划可以通过在一天中不同的时间、一周中不同日期并跨越学期的各周做观察而得到改进,

以提高外部效度。

4. 如果我们假定图书馆的学生比学生会休息室的学生更能集中注意,那么我们需要找到某种方法解释为什么在这一研究中两种场所的学习时间相同。一种可能是两个场所的学生们选择了不同的学习材料。如果学生会的学生在学习“较容易的”材料,那么他们就会像图书馆里学习较难材料的学生一样集中注意力。做自然观察的一个难题是研究者不能控制那些可能影响观察结果的因素(是因为自然观察法的非干预特征)。

挑战性问题1的答案

- A. 学生们提高信度的方法如下:对每个家庭观察8小时,使用两个独立的观察者,并使用一览表提供操作定义。
- B. 对结果的外部效度的一个可能威胁是12个家庭自愿参加研究,这样的家庭可能不一定是具有代表性的家庭。
- C. 学生们努力减小与家庭和彼此之间的相互影响,表明他们对反应性问题是敏感的。他们所使用的其他两种方法是习惯化和脱敏。

第5章

调查研究

本章大纲

概 要

调查研究的使用

调查研究的特点

调查研究中的抽样

抽样的基本术语

抽样方法

调查方法

邮寄调查法

个人面谈法

电话访谈法

网络调查法

调查研究的设计

横断设计

连续独立样本设计

纵向设计

问 卷

作为调查工具的问卷

自陈测量的信度和效度

编制问卷

关于调查研究的批判性思考

报告行为与实际行为的一致性

相关与因果关系

小 结

概 要

你最喜欢的食物是什么？你的睡眠时间是多长？你开什么款式的车？你使用校园的图书馆吗？你对目前的生活满意吗？也许你曾被别人询问过诸如此类的问题，这其实就是调查的一部分。心理学家、社会学家、政治学家、市场调查人员、杂志社编辑以及许多其他的人，都会运用调查方法来获取有关人们想法和感受方面的信息。调查的结果常被用于描述人们的观念、态度和偏好。例如，选举前民意测验的结果能够显示选民对候选人的偏好以及对选举的看法。除此之外，调查结果还可用于对人们的行为做出预测。例如，市场调查人员也许会利用调查结果预测哪些家庭更可能会使用某种新产品；心理学家也许会设计一份问卷，用以确定人们的自杀风险。相关研究（correlational research）能为这样的预测提供基础。为了确定预测关系，需要对自然发生的变量间的关系进行评估。在本书第4章，我们曾经谈到相关系数是确定变量间关系的方向和大小的量化指标。在本章中，我们将在调查研究的背景下阐述相关研究。

调查研究的特点就是对所选取的样本人群开展研究。因此，在本章中，我们将首先介绍抽样的基本概念和技术——即从人群中抽取子集以代表总体人群的操作过程。我们还将阐述各种调查研究方法和调查设计的优劣。在调查研究中最主要的工具就是问卷，因此我们将阐述设计一份好问卷的要领。我们还将在本章中讨论调查研究中一个非常重要的问题，即“人们真的会像他们在调查中所回答的那样做吗？”本章最后，我们将讨论一个更广泛的问题，“当两变量间存在相关时，该如何推导因果关系？”

调查研究的使用

- 调查研究可用于评估人们的想法、观点和感受。
- 调查研究的目标在范围上可以是具体的、有限制的，也可以是更综合的、更全面的。
- 确定调查结果是否存在偏差的最佳方式是检查调查的程序和分析方法。

在第4章我们曾经讨论过心理学家如何运用观察法推断个体做出某种行为的想法或感受。与观察法不同的是，调查研究则对人们的想法、观点或感受的实质进行更直接的考察。从表面上看，调查研究似乎太过简单。如果你想知道人们在想什么，提问就是了！同样地，如果你想知道人们在做什么，观察就是了！然而，当我们想要从人们的行为中推导出普遍原理时，所采用的观察法必须比日常的随意观察复杂得多。同理，调查研究也要比简单的提问复杂得多。

社会科学家，如政治学家、心理学家、社会学家会出于各种原因在他们的研究中运用调查研究，无论这是出于理论需要还是应用需要。除此之外，媒体、政治候选人、公共卫生官员、专业团体、广告和市场总监等也会运用调查研究来满

足其较实用的需要。换句话说,调查研究的范围和目的可以是有限制的、具体的,也可以是更综合、全面性的。例如,美国心理学会(APA)下属的一个专门关注临床心理学的机构曾经开展过一项非常具有针对性的研究(Peterson, 1995)。该机构对全美国1087位成年人展开调查,旨在了解公众对心理学家的认识,包括:你认为心理学家是做什么的、心理学家的受教育水平如何、心理学家在什么样的环境中工作。该调查引发了许多人的关注——其中46.7%的受访者报告自己或家中其他成员曾经咨询或求助过心理健康方面的专家;80%的受访者认为心理学家通常供职于私人机构。事实上,虽然超过半数的心理学家确实从事心理健康领域的工作,但其中大多数的人都不在私人机构工作。上述调查结果对该机构制定公共教育方面的策略非常有用。

迈尔斯和迪纳(Myers & Diener, 1995)开展了一项全球范围的调查研究,考察人们对幸福的个人感受。除南极洲外,他们从各大洲具有代表性的24个国家中抽取样本。研究中关心这样一个问题:较富裕国家人民的幸福感是否高于相对贫穷国家人民的幸福感?即“国力昌盛”是否与人民的“幸福感”相关?调查结果显示,国家的富裕程度(以人均GNP为衡量标准)与国民的幸福感呈正相关(.67)。然而两者的关系并非如此简单,因为国家富裕程度也与其他变量存在相关,这些变量也与国民幸福感存在高度相关,如国家奉行民主精神的持续时间与幸福感存在显著相关(.85)。由此可见,调查研究也可用于考察较复杂的全球性课题。

在使用调查研究方法时有一个问题涉及伦理,必须在此指出。当调查结果与从事该研究的主办方存在利益关系时可能产生伦理困境。克罗森(Crossen, 1994)曾经这样说道:“……我们在日常生活中进行购买、选择、治疗等活动时所依赖的越来越多的信息并不能增加我们的见识,这些信息之所以被制造出来只是为了增加产品销量,或是为了提高说服力”(p.14)。克罗森列举了此方面的一些事例,如由锌工业协会组织主办的一次全国调查表明,62%的调查者想要保留便士硬币(其中含锌);某项由手机制造商发起的调查研究则表明,70%的受访者(他们全都使用手机)认为使用手机的人在商业上更容易获得成功。

那么,我们是否据此就可以认为所有与主办方利益相关的调查结果都是不可信的呢?当问题涉及伦理层面答案很简单,但对此问题的回答却比较复杂。当主办方能从调查结果中获益的时候,我们也同样可以开展高质量、并且符合伦理的研究。因此,在评估调查结果时必须对该调查的主办方有所了解。当然这还不足以判断此研究是否存在偏差,因为更重要的是必须考察样本的代表性、问卷中的措辞是否合适、数据的报告与分析是否全面。以上所有方面都可能造成研究结果的偏差。如果研究者不负责任,也可以通过歪曲上述各方面造成“结果正确”的假象。因此,揭穿某些不负责任的研究者和低劣研究伎俩的最好方法,就是仔细检查该调查的程序和所使用的分析方法。

调查研究的特点

- 调查研究需要抽取一个或一组样本,使用一套预先确定的问题。

虽然有各种不同的调查研究方法,但这些方法具有一些共同的特点,这使得调查研究在描述人们的态度和观念时更出色。首先,调查研究涉及抽样,抽样几乎是所有的行为研究都会涉及的问题。在第4章讲授观察法时,我们曾经在阐述时间与情境抽样时介绍过此概念。在本章下一节我们将讨论调查研究中的抽样问题。其次,调查研究还会为所有调查对象准备一套问卷,问卷中的问题已经事先确定。调查对象对问卷做出的口头或书面回答就是调查研究所能获得的最主要的数据。研究人员会让调查对象以相同的顺序对同样的问卷做出回答,这样进行数据总结时就更有效。

当研究者用一套相同的问题询问具有代表性的样本时,我们就能根据他们的回答描述该样本所代表人群的态度。并且,使用相同的问卷对于我们比较不同人群的态度或考察态度随时间的变化大有裨益。例如,普罗斯(Plous, 1998)曾经先后两次运用同样的问卷考察动物权利保护者的态度随时间的变化情况。他们在1990年开展了首次调查,当时华盛顿特区正在举行全国性的游行。1996年又开展了第二次调查,当时全国也在举行相似的游行。两次调查表明,六年后动物权利保护者的态度发生了改变。尤其值得一提的是,1990年大多数的动物权利保护者认为动物研究是他们急需解决的问题,1996年他们则把矛头指向了畜牧业。上述研究表明,研究者可以在不同时间向选择的样本呈现相同的问题,以此来获取人们态度改变方面的讯息。

调查研究中的抽样

- 仔细选择调查样本才能确保从样本获得的结果可以推广至其代表的总体。

假设你计划开展一项调查研究,你对问卷中的所有问题成竹在胸,也明确了要选择什么人群开展调查,那么下一步就是——调查对象是谁?这涉及了该如何从你所感兴趣的人群中选取有代表性的样本。调查研究者安格斯·坎贝尔(Angus Campbell, 1981, p.17)曾对抽样的有效性进行过比较恰当的总结:

如果想描述庞大的异源群体的多样性,那么毫无疑问应该选择抽样的方法。自从统计学家和社会科学家获知如何从庞大的人口中进行抽样的方法后,从数千人的样本中所得的数据对全国人口总体情况进行准确描述已成为可能。因为准确运用抽样能够确保人群中每位成员都有平等的被抽取的机会。

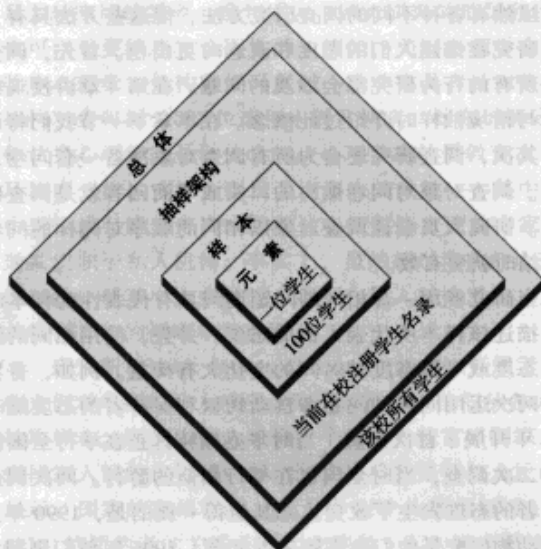
无论我们要对全国人口进行描述,或是对更小群体(如某高校的大学生)进行描述,获取代表性样本的程序是相同的。我们将在以下内容中进一步探讨坎贝尔所提出的如何进行抽样的问题。

抽样的基本术语

- 在所有抽样技术中,最核心的是确认和选择组成样本的元素;样本选自抽样

图 5.1 抽样中四个基本术语

间的关系解释



架构，或是感兴趣的抽样总体的所有成员。

- 研究者不只对调查对象的反应感兴趣，实际上，他们对如何用样本所得结果对更大的总体进行描述更感兴趣。
- 样本在多大程度上能推论至总体主要取决于样本的代表性。
- 有偏样本是指该样本与总体存在系统性差异。
- 当抽样程序中发生了样本选择偏差，就会导致样本对总体的某些部分代表过度或代表不足。

在我们开始讨论抽样技术前，必须先澄清四个术语：总体、抽样架构、样本和元素。图 5.1 对这四个重要术语间的关系进行了概括。总体（population）是指你所感兴趣的人或事物的全体。例如，如果你感兴趣于自己所在高校的学生对计算机服务的态度，那么所谓的总体即指你们学校所有的学生。显然，我们不可能与每位学生接触。因此，研究者通常会从总体中选择一部分人作为代表。

为了能够科学地从总体中抽取对象，我们就需要有总体成员的详细名单。这份详细的名单就是抽样架构，在某种意义上，这是对你所感兴趣总体的操作性定义。如果想调查学生对计算机服务的态度，那么抽样架构就可能是从学校注册办公室获得的所有在校生名录。抽样架构在多大程度上能真实反映你所感兴趣的总体，决定了我们能否适当的抽取样本。如果你是从注册办公室获得的名录，那么它应该能为我们提供良好的抽样架构。但该名单也可能会遗漏一些学生，如某些延误注册的学生。

从抽样架构中选取的人群构成了所谓的样本（sample）。如在调查学生对计算机服务的态度的例子中，我们可以从学生注册名录中选取 100 位学生作为我们的样本。这 100 位学生在多大程度上能够代表全校学生完全取决于抽样的操作。样

本中的每位成员则被称为元素。在所有的抽样技术中,最核心的就是确认和选择元素。

需要强调的是,研究者对样本本身并无兴趣,或者兴趣较低。学校不会因为100名学生的需要增添新的计算机设施。同样,社会心理学家不会只关注所调查的50位对象的种族态度,市场总监也不会只关注200位被调查的消费者的消费偏好。总体,而不是样本,才是研究者的兴趣所在。从样本中所得的调查结果能在多大程度上描述总体是基于如下假设:样本中获得的调查结果可应用于其代表的总体。

样本调查结果能够推广至总体的能力取决于样本的代表性(representativeness)。显然,总体中的个体间在许多方面存在差异,不同总体间也存在很大差异。例如,某总体的性别构成为40%的男性,60%的女性;另一总体的性别构成则是75%为女性,25%为男性。判断样本具有代表性的标准就是:样本必须与总体的特征具有相同的分布状况。如果一个200人的代表样本中有男性80人,女性120人,那么你认为该样本能够代表我们上述所举例子中的哪个总体?在实际操作中确定代表样本还有一些其他注意事项,详见专栏5.1。

专栏 5.1

如何确定代表样本

左侧为四种总体的情况。请在右侧中选择与各总体对应的样本。

总 体

样 本

1. 60%女生, 40%男生

90%年龄在18-22岁, 10%年龄超过22岁

70%大一/大二本科生, 30%大三/大四本科生

2. 80%女生, 20%男生

60%年龄在18-22岁, 40%年龄超过22岁

70%大一/大二本科生, 30%大三/大四本科生

3. 75%女生, 25%男生

65%年龄在18-22岁, 35%年龄超过22岁

80%大一/大二本科生, 20%大三/大四本科生

4. 80%女生, 20%男生

90%年龄在18-22岁, 10%年龄超过22岁

70%大一/大二本科生, 30%大三/大四本科生

A. 132名女生, 44名男生

114名年龄在18-22岁, 62名超过22岁

141名大一/大二本科生, 35名大三/大四本科生

B. 244名女生, 61名男生

183名年龄在18-22岁, 122名超过22岁

213名大一/大二本科生, 92名大三/大四本科生

C. 48名女生, 12名男生

54名年龄在18-22岁, 6名超过22岁

42名大一/大二本科生, 18名大三/大四本科生

D. 150名女生, 100名男生

225名年龄在18-22岁, 25名超过22岁

175名大一/大二本科生, 75名大三/大四本科生

答案

1. D 2. B 3. A 4. C

资料来源:摘自 Zechmeister, Zechmeister & Shaughnessy, *Essential of Research Method in Psychology*, McGraw-Hill, 2001, p.124.

对于样本代表性而言,可能存在的最大威胁就是样本存在偏差。有偏样本是指该样本的特征分布与目标总体存在系统性差异。例如,某100人的样本中包括了80位女性和20位男性,但如果其目标总体中的60%为女性、40%为男性,那么该样本就是有偏样本。由于在该样本中女性所占比率高于总体中的比率,因此会导致该样本中的女性代表过多,男性代表不足。导致样本偏差有两种原因:样本选择偏差和反应偏向。在抽样过程中,当对总体的某些方面代表过度,或相反地,对总体中的某些重要方面代表不足甚至忽略了这些重要方面时,选择偏差(selection bias)就会出现。在下一节的“调查方法”中我们将探讨反应偏差问题。

查尔斯·马迪根是《芝加哥论坛报》的一位资深作家,他曾经指出美国1994年大选调查中存在样本选择偏差。他通过将投票后民调样本中的人口统计学特征与美国人口普查中的资料进行比较后得出该结论。他指出,在全美人口普查中不仅包括了投票人,还包括了未投票人。但在此次投票后民调样本中,却只包括了投票人。马迪根指出,投票后民调样本与美国总体人口相比,无论在年龄、种族、教育水平、收入、职业女性所占比例等方面都存在系统性差异。上述差异表明,由于存在样本选择偏差,因此该样本无法代表全体美国人。虽然该样本也许能够准确反映投票人的兴趣与态度,但却无法正确描述全体美国人口的特征(因为剔除了未投票人)。显然,政治家不应该根据这样一个有偏本来制定政策。

从该例子中我们还能获得另一启示,即代表样本应该从研究者所感兴趣的总体中选取。如某高校想要了解学生中的有车族对校园泊车的观点,那么总体应该是开车上学的学生,而不是该校所有学生。因此,非偏差样本应该是能代表在学校有车一族的学生总体。

抽样方法

- 选取调查样本的方法有两种:非随机抽样和随机抽样。
- 非随机抽样(如便利抽样法)无法保证总体中的每个元素都有同等被抽取的机会。
- 随机抽样(如简单随机抽样和分层随机抽样)能使研究者评估样本所得结果与总体间的区别。
- 简单随机抽样能够保证总体中的每个元素都有同等被抽取的机会。分层随机抽样则把总体划分为几个子群体(即“层”),并从不同的层中获得随机样本。
- 随机抽样是获得代表性样本的一种选择方法。

抽样的方法有两种——非随机抽样和随机抽样。非随机抽样(nonprobability sampling)无法保证每个元素都有被抽取的机会,也无法评估每个元素被选入样本的概率。在我们前述的学校计算机服务的调查中,如果研究者对进入图书馆的前30位学生进行调查,那么所用的就是非随机抽样法。因为在此操作中,显然不是所有学生都会在该时刻进入图书馆,有些学生也许根本没有机会被抽取。

与此相反,如果研究者从学校注册办公室的在校生名录中随机抽取30名学生,那么所用的就是随机抽样法。在随机抽样(probability sampling)中,所有注册学生(元素)都有同等机会被选入样本。之所以能将研究者的此种抽样方法称

为随机抽样,是因为抽样过程(如从某事先确定的名单中进行随机选择)能使所有学生都有被选入该调查的平等机会。上述两种抽样方法,我们更推荐大家选用随机抽样,因为该方法能使研究者评估样本所得结果与总体间的区别。在被选择样本对总体的代表性方面,随机抽样的优势远高于非随机抽样。因此,从注册名单中随机抽取30位学生与抽取进入图书馆的前30位学生相比,显然前者的样本更具有代表性。

非随机抽样 非随机抽样最常见的形式是便利抽样。便利抽样(convenience sampling)主要根据获得调查对象的容易程度和调查对象参加调查的意愿所进行。例如,各种报刊经常采访“街头行人”,然后在此基础上发表评论。我们也许只能以调侃的方式面对此类评论,而不能认为这些路人的观点代表了更广大群体的心声。由于便利抽样是非随机抽样,因此无法保证每个人都有机会被选入样本,因此通过该方法所选取的样本缺乏代表性。除此之外,读者对某些杂志上的民意调查的回应也属于便利抽样。因为这首先要要求读者拥有此杂志,其次还要求读者愿意对此调查进行回应并将自己的答案寄给杂志社。许多大专院校的心理学家们经常会将选修心理学基础课程的学生作为调查对象,这也是一种便利抽样。

克罗森(Crossen, 1994)曾描述过另一种便利抽样的缺点,即“来电调查”。电视及广播媒体经常采用此法来获得观众或听众的意见。在此种调查中,样本是由那些正在收看或收听此节目并且愿意打进电话(有时还需要支付电话费)的人们组成。因此,我们必须记住的是,这些愿意向电台或电视台打电话参与调查的人与总体人群存在差异。不仅因为他们只代表那些正在收看或收听该节目的人,还因为他们只代表那些愿意打电话参与调查的人群。同样,那些在登录某网站主页后参与该网站弹出页面调查的网民也无法代表那些在看见该调查后不做回应的人,当然更无法代表那些没有电脑上网的人们。

美国有一档黄金时段播出的新闻节目曾经采用来电调查方法征询公众对于联合国所在地的意见(Crossen, 1994)。与此同时,有研究者采用随机抽样方法对500人就同一问题展开调查。该问题是:您认为联合国总部是否应该继续设在美国?电视台共收到186 000名观众的来电反馈,其中67%的人希望联合国总部搬出美国。采用随机抽样法的研究结果则显示,500人中72%希望联合国总部继续设在美国。为什么两项研究的调查结果差异如此巨大——甚至是相反?我们是否应该因为电视台的调查参与人数众多,而更信任他们的结果呢?绝对不行!因为无论便利抽样获得的样本数量有多么庞大,其暴露的问题与所有其他采取便利抽样的调查毫无二致。因此,我们应该牢记以下准则:便利抽样将导致有偏样本,除非你有足够的证据证明该样本具有代表性。

随机抽样 随机抽样最重要的特点就是:研究者可以确定总体中每一元素被选入样本的可能性。两种最常见的随机抽样形式是简单随机抽样和分层随机抽样。简单随机抽样是随机抽样的一种基本抽样技术。简单随机抽样(simple random sampling)最常见的定义是:在该抽样方法中,每个元素都有平等的机会被选入样本。专栏5.2描述了简单随机抽样的操作步骤。

在进行简单随机抽样的过程中,首先需要确定样本大小。简单地说,就是随

专栏 5.2

随机抽样范例

下列名字代表缩减版的样本架构，资料来源于某高校注册办公室。我们将在此名录的基础上阐述如何进行简单随机抽样和分层随机抽样。

Adamski	女	大三
Alderink	女	大四
Baxter	男	大四
Bowen	女	大四
Broder	男	大二
Brown	男	大三
Bufford	男	大二
Campbell	女	大一
Carnahan	女	大二
Cowan	女	大一
Cushman	男	大四
Dawes	男	大三
Dennis	男	大四
Douglas	女	大一
Dunne	男	大二
Fahey	男	大一
Fedder	男	大一
Foley	女	大二
Grossman	女	大三
Harris	女	大三
Hedlund	女	大二
Johnson	女	大一
Klarren	女	大三
Ludwig	女	大一
Nadeau	女	大四
Nowaczyk	男	大三
O'Keane	女	大四
Osgood	男	大二
Owens	女	大二
Penzien	女	大三
Powers	男	大四
Ryan	男	大一
Sawyer	男	大三
Shaw	男	大四
Sonders	女	大四
Suffolk	女	大二
Taylor	女	大一
Thompson	男	大一
Wattersen	女	大三
Zimmerman	男	大二

简单随机抽样：

步骤 1. 为样本架构中的每一元素编号：Adamski 为 1，Harris 为 20，Zimmerman 为 40。

步骤 2. 确定你想使用的样本大小。由于这是一个范例，故我们选择样本大小为 5。

步骤 3. 从附录 A.1 中的随机数字表中选择一个起点（闭上双眼在表上随意用手指确定一个起点亦可——此处我们将起点定于第 8 列、第 22 行的 26384）。由于我们的样本架构编号为 1 至 40，因此我们决定在使用随机数字之前，先选用每组 5 个数字中左边的两个数字，并从左至右开始选择。当然你也可以从上至下或者从右至左。我们也可以选用每组 5 个数字中的中间两位数字或者最后两位数字。但你必须在使用随机数字之前做好决定。

步骤 4. 通过移动表格，确定样本大小。我们在此得到 26, 06, 21, 15 和 32。请注意在此范例中样本大小不能超过 40。

步骤 5. 根据所选数字列出对应名字。在此范例中，样本将包括 Nowaczyk、Brown、Hedlund、Dunne 和 Ryan。

如果采用系统抽样也能得到随机样本，而且操作更简单。首先将样本架构的人数除以你所需要的样本人数就得到了 k 值。然后就从随机样本架构中选择第一个人，再选择该人后面的第 k 个人，依此类推。在此范例中，我们需要从 40 人的样本架构中抽取 5 人，因此 k 为 8。因此我们就可以随意选择第一位，然后选择其后的第 8 个人，依此类推。如果我们选择 Alderink 作为第一个被选择出来的人，那么其余的人依次是：Cowan、Foley、Nowaczyk、Shaw。注意，如果样本架构是按照一定规律排列的，那么就不能用此种方法。例如，你有一份名单，该名单根据住宿生的宿舍号码进行排序，并且每隔 10 个宿舍就是靠角落的房間。你马上就会发现，如果你的抽样间隔为 10，那么你会把所有靠角落的房间里的人选中，或者根本无法选到住在角落房间里的人。

(续)

大一	大二	分层随机取样:
1. Campbell	1. Broder	步骤 1. 将样本架构按层进行排列。在此例中, 我们按照
2. Cowan	2. Bufford	年级进行排列, 因此样本架构将变成左侧的样子。在此
3. Douglas	3. Carnahan	例中, 每一层的人数相等。但在实际操作中不必一定如
4. Fahey	4. Dunne	此。
5. Fedder	5. Foley	步骤 2. 将每一层中的元素进行编号, 编号后的结果见左
6. Johnson	6. Hedlund	侧。
7. Ludwig	7. Osgood	步骤 3. 确定样本大小。在此例中, 我们需要的样本大小
8. Ryan	8. Owens	为 8。
9. Taylor	9. Suffolk	步骤 4. 从每一层抽取相同数量的样本, 最后将获得整个
10. Thompson	10. Zimmerman	样本。在此例中, 需要从每一层抽取 2 个。
大三	大四	步骤 5. 按照随机样本抽取的步骤进行, 每一层抽取时步
1. Adamski	1. Alderink	骤相同。在此我们采用前面所确定的起点 (附录 A.1 随
2. Brown	2. Baxter	机数字表中的第 8 列、第 22 行), 但在此例中, 我们将
3. Dawes	3. Bowen	选用每组 5 个数字中的最后一个数字。因此, 每一层抽
4. Grossman	4. Cushman	取的编号分别为: 大一 (4 和 1)、大二 (6 和 4)、大三
5. Harris	5. Dennis	(7 和 9)、大四 (2 和 9)。
6. Klarren	6. Nadeau	步骤 6. 根据所抽取的编号列出名单。该例中最后选取的
7. Nowaczyk	7. O'Keane	分别为: Fahey, Campbell, Hedlund, Dunne, Nowaczyk,
8. Penzien	8. Powers	Sawyer, Baxter 和 Shaw。
9. Sawyer	9. Shaw	
10. Watterson	10. Sonders	

机样本的大小必须能够代表总体, 这取决于总体的同质性和异质性。例如, 在学习能力方面, 常春藤盟校的大学生总体就比全美大学生总体更具同质性。一种极端情况就是, 同质性最高的总体中的所有人员应该完全相同, 即无论总体人数多少, 总体中的一个元素就能代表该总体。另一种极端状况就是, 在异质性最高的总体中, 每个人之间在各方面都完全不同, 即无论该总体人数多少, 没有一个元素能够代表该总体, 也就是在抽取样本过程中, 不得不包括总体中的每位个体。幸运的是, 实际情况中的总体总是处于这两种极端状况之间。因此, 就如同品尝一块巧克力蛋糕时越吃越美味一样, 样本的代表性在一定程度内将随着样本数增加而增加, 直到达到最佳的某一点。

除此之外, 我们还可以采用分层随机抽样法来提高样本代表性。分层随机抽样 (stratified random sampling) 是指先将总体分为不同的子群体, 即“层”, 再从每一层中抽取随机样本。在此有两种方法可用于确定从每层中抽取的样本数量。第一种方法 (见专栏 5.2 中的最后一个例子) 是从每一层抽取同等大小的样本。第二种方法是在比例的基础上选取样本元素。例如, 某大学本科生总体中, 大一新生占总体人数的 30%, 大二学生占总体人数的 30%, 大三学生占 20%, 大四学生占 20%。那么, 在该例子中, 学年就是“层”。如果采用分层随机抽样方法得到 200 人的样本, 那么该样本中必须有大一新生 60 名, 大二学生 60 人, 大三学生 40

拓展练习

两位学生被要求开展一项调查,了解大学生对校园里的男生联谊会和女生联谊会的态度。该校共有3200名学生。其中25%的学生属于希腊社团,75%的学生不属于任何组织。两位学生在采用何种抽样方式上意见不一。其中一位学生认为应该采用分层随机抽样方法选取200人的样本;其中100名是属于希腊社团,剩下的100名来自不属于任何组织的学生。第二位学生认为应

该采用简单随机抽样,抽取100名学生作为样本。

1. 请从样本代表性以及他们能有效测量希腊社团学生观点两个方面,就此两种抽样方法发表你的看法。
2. 如果你认为两种抽样方法均不理想,那么请给出你的抽样方法。

人,大四学生40人。与此不同的是,如果是从每一层中抽取同样数量的样本,那就意味着从每一年级的学生中各抽取50名。值得注意的是,在分层随机抽样中,只有在比例基础上抽取的样本才具有代表性。

分层随机抽样方法除了能够提高样本代表性之外,还有助于研究者描述总体的特定比例分布。例如,想调查某高校2000名大学生的态度,那么采用简单随机抽样从中选取100名学生就可以了。但如果假设该校化学系学生有40名,但你的样本只包括了其中2名,而你却想调查不同专业学生的态度,那么问题就出现了。也许你的样本能够准确反映该校化学专业人数在全校人数中的比例,但若只用两名化学系学生来代表该系40位学生就存在较大风险,因为人数实在太少了。面对此种情况(或更宽泛地说,当利用分层抽样会导致某层的样本数过小),你就应该从化学专业中抽取更多的人数,才能得到更合理的调查结果。在此我们无法给出确切的样本数目,因为过去的经验告诉我们,样本大小能否代表总体取决于总体的异质性程度。

调查方法

- 获得调查数据的四种主要方法为:邮寄调查法、个人面谈法、电话访谈法和网络调查法。

抽取样本只是开展调查研究中的一个步骤而已。除此之外,还有其他的任务要完成。在确定样本后,你还必须决定如何从调查对象获得你需要的信息。在此我们介绍四种主要方法:邮寄调查法、个人面谈法、电话访谈法和网络调查法。在实际调查的过程中,你会发现无法找到适合所有情况的最佳方法,每种方法都有其优劣。你需要做的是从中选出一种最适合你的调查方法。

邮寄调查法

- 虽然邮寄调查法快捷方便,但如果接受调查的人没有完成调查问卷或没有寄

回调查结果时就会发生反应偏差。

- 反应偏差可能导致邮寄调查的最终样本无法代表总体。

邮寄调查通过将自填式问卷邮寄给个体的方式展开,该方法要求调查对象自行填写问卷。该方法的优点之一就是比较快捷。因为采用的是自填式问卷,因此避免了调查中可能出现的访谈者偏差(参见下一节内容)。在四种调查方法中,邮寄问卷最适合调查一些高度私人化和令人尴尬的主题,在不要求调查对象提供真实身份的情况下,其优势尤为明显。

不幸的是,邮寄调查存在诸多缺陷。缺陷之一就是没有其他方法严谨。因为调查对象在填写问卷的过程中如果遇到疑问时无法提出问题,更不可能得到指导,因此所使用的问卷必须便于调查者理解。缺陷二是研究者无法对调查对象的作答顺序进行控制。问卷的顺序往往会影响到调查对象在某些问题上的反应。邮寄调查法还存在一个较为严重的问题——反应偏差。

反应偏差(response bias)是指由于一些调查对象没有完成调查,从而降低了样本代表性的现象。有很多原因会导致反应偏差。例如,调查对象如果存在阅读困难、学历较低或视力问题,就可能无法完成调查。因此,在确定最终样本时,应该将有上述问题的个体排除。还可能的原因是调查对象太繁忙因此无法寄回问卷,或调查对象对该调查兴趣不大所以没有完成问卷。在邮寄调查中,低反应率(如没有完成问卷或没有将问卷寄回)是导致反应偏差的主要因素。

由于低反应率的问题,往往导致最后研究者得到的样本少于期望值。通常,样本的大小还不是最严重的问题,最严重的是这可能意味着反应偏差。由于反应偏差,原本精心抽取的随机样本可能变成非随机样本——即可能会变成便利抽样,因为个体的意愿决定了他们会否完成调查。例如,在一项有关心理学基础课的调查中,泽克迈斯特等人(Zechmeister & Zechmeister, 2002)依据1998年版的《皮特森四年制高校指南》(*Peterson's Guide to Four-year Colleges*),抽取了全美250所高校,随后将调查寄给了各校的心理系。结果得到的反应率只有56%——仅有140所高校的心理系寄回了问卷。

无论之前你如何严谨地抽取样本,只要回收率或反应率没有达到100%,就存在反应偏差的可能。因此,研究者必须在研究报告中对反应率加以说明,即必须说明最终提供反馈的样本能否代表总体,是否存在对总体的某些部分代表过度或代表不足的问题。例如,虽然他们的调查只有56%的反应率,但泽克迈斯特(2000)利用《皮特森四年制高校指南》中的信息证明了其调查结果依然能够代表总体。

一般而言,邮寄调查的回收率只在30%左右。但是也可以通过一些方法来提高回收率。

当你注意到以下事项,回收率一般将有所提高:

- 让问卷显得更具针对性(例如,可以在问卷中写上调查对象的名字,而不是简单标以“居民”或“学生”);
- 尽量减少调查对象的工作量;
- 让调查主题更能引起调查对象的兴趣;
- 在一定程度上让调查对象与主办调查的机构或个人的关系更密切。

个人面谈法

- 虽然个人面谈法比较耗费人力物力，但研究者能更好地控制调查过程。
- 如果结果记录不准确或访谈者对受访者进行了不必要的引导，这些都可能导致访谈者偏差。

利用个人面谈法收集调查资料通常是在受访者家中或购物中心等公共场合进行，调查过程中由训练有素的访谈者进行提问。与邮寄调查法相比，个人面谈法在提问的过程中允许更大的灵活度。在提问过程中，当受访者认为问题不明确时可以马上得到访谈者的解答。在此过程中，访谈者控制着提问的顺序，并能确保所有的受访者都回答了所有问题。通常，个人面谈法的反应率要高于邮寄调查法。

个人面谈法的优势显而易见，但它也有一些缺点。如今，城市居民对犯罪行为的警惕性提高了，再加上越来越多的家庭白天都无人在家，这使得在家庭中开展面谈不再那么有效。除此之外，个人面谈法还有一个较严重的问题就是费用昂贵。个人面谈法必须使用一些访谈者，对这些访谈者的培训既耗时间又耗金钱。也许，个人面谈法最严重的问题是可能存在访谈者偏差。在面谈的提问与受访者作答的过程中，要求访谈者始终保持中立态度。但是当访谈者只记录受访者的某些回答，或为了迎合受访者对提问做出更改时，就会产生访谈者偏差（interviewer bias）。例如，某访谈者向被调查者询问他们对电视节目的看法。受访者回答：“电视节目最大的问题是暴力内容太多。”如果访谈者这时候记下“暴力是电视节目问题”而没有把受访者的原话完整记录，此时就会出现访谈者偏差。在接下来的追问中，如果访谈者问“你说的暴力是指凶杀、抢劫还是强奸？”，此时也出现了访谈者偏差。访谈者应该以更中立的方式进行提问，例如他可以这样提问：“您能更具体的描述一下您所说的暴力吗？”

避免访谈者偏差的最佳方法是选用一些对调查非常感兴趣的人们作为访谈者，并支付给他们较高的报酬。除此之外，更应该训练他们准确提问、正确记录、明智追问的技能，并向他们提供一份如何处理面谈中各种特殊状况的详细说明。最后，还应该由该调查项目的负责人密切监督访谈者的执行情况。个人面谈法的调查任务比较繁重，因此单枪匹马很难完成。戈洛波克等人（Golombok et al., 2003）曾经用个人面谈法对女同性恋家庭中的孩子进行调查，让我们来了解一下他们的研究过程。他们先从英国埃文镇的出生登记局获得了女同性恋和异性恋家庭的家长及其孩子的名单，并从中抽取样本。随后访谈者们对样本家庭展开上门调查，这些访谈者事先都已经过培训，他们对标准的访问程序非常熟悉。面谈时间控制在一个半小时至两小时之间，面谈过程进行了录音。他们记录了受访者在回答问题过程中的行为表现，并对其进行评分（如，表现得很热情等）。访谈者们还会定期讨论评估过程以减少不同访谈者在评估中存在的差异。访谈者还对这些家庭中的孩子（7岁左右）开展调查，并让孩子们完成一份简短的人格问卷，这需要访谈者将问卷大声朗读给孩子听。该研究还通过随机抽取访谈者计算了评估者间信度（0.73~1.00之间），此部分内容我们在第4章已有过详细论述。研究发现，“与异性恋家庭中的孩子相比，女同性恋家庭中的孩子在性别化行为方面没有差异。”该

虽然个人面谈法调查比较方便，但如果接受调查的人没有完成调查问卷或没有可

结果“与那些认为异性恋父母对孩子性别化行为的获得是不可或缺的观点相矛盾”(p.31)。由于此类研究刚处于起步阶段,因此究竟哪种说法更可信也许需要更进一步的研究。未来通过考察更大、更具多样性的样本也许能够为我们揭示真相。

现今,计算机技术飞速发展。通过电脑将自填式问卷调查与个人面谈法相结合已成为可能。受访者可以先倾听电脑中访谈者的提问录音,然后对电脑上的问题做出回应。在该技术中,每位受访者听到的都是同一位访谈者的声音,提问方式也完全一致,这就降低了访谈者偏差。除此之外,该项技术还能让受访者回答某些涉及一定隐私的个人问题(Rasinski, Wills, Baldwin, Yeh, & Lee, 1999)。

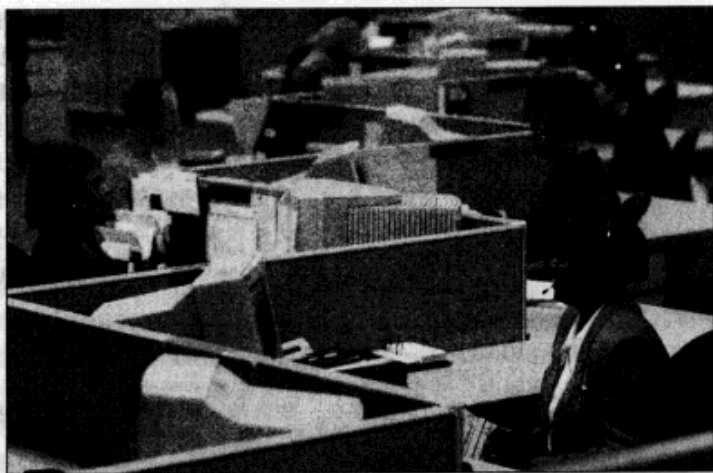
电话访谈法

- 虽然电话访谈法存在一些弊端,但此法仍经常在一些较简短的调查中得到运用。

个人面谈法不仅耗费财力,而且很难对访谈者的调查过程展开有效监督。因此一些调查研究者更倾向于采用电话访谈法或网络调查法。电话访谈法在选择潜在受访者的样本架构上存在较严重的问题,因此在诞生之初就遭受了很多严厉的批评。许多人并没有把电话号码登记在黄页上,一些穷人和生活在偏远地区的人们也许家里根本没有电话。直至2000年,美国97%的家庭才拥有了电话(U.S. Census Bureau, 2000),而且调查者通过随机数字拨号技术有可能拨叫到那些没有登记在黄页上的家庭。随机数字拨号技术可使研究者更有效地从全美国电话使用者总体中获取更广泛的代表样本。除此之外,电话访谈还能使研究者接触到个人面谈中无法调查的群体,如犯罪率高发区的人们、大门紧闭的家庭,以及晚上才回家的人们(你与否在晚餐期间接到过想对你开展简短调查的来电呢?)。电话访谈能在较短时间内完成,并且由于所有的访谈者都在同一地点进行电话调查,因此也更容易对他们的工作进行监督(见图5.2)。

当然,电话访谈法也和其他调查方法一样存在缺陷。当受访者只局限于拥有电话的人群,那么就可能产生选择偏差,同时也还可能存在访谈者偏差。除此之外,受访者的耐心有限,而且由于访谈者没有出现在受访者面前,这也可能会导致与面谈法不同的反应。再加上目前手机用户的增加,可能会对调查产生其他不可知的影响。例如,打电话给手机用户时,他们通常正忙着某事,或者正在商务场合不便接听。时代的变化可能导致电话访谈法的反应率降低。除上述问题外,现在一些中产阶级以上的人群可能拥有多个电话号码,那么利用随机数字拨号技术还可能致这群人代表过度。希普勒和施瓦茨(Hippler & Schwarz, 1987)认为,在电话访谈中人们进行判断的时间很短,这可能导致他们无法有效记住访谈者提出的选项。此外,人们还经常收到兜售商品、请求捐献类的电话,因此很多人都不愿参加调查。来电过滤和语音信箱等技术也使人们可以轻松地拒接某些不受欢迎或陌生的电话。除此之外,还有许多人由于工作繁忙很少回家,因此很难通过拨打他们的住宅电话取得联系。虽然电话访谈法存在上述诸多弊端,但在比较简短的调查中研究者们仍经常采用该方法。

图 5.2 随机数字拨号技术可使研究者有效获取更广泛的代表样本,并对他们展开简短调查



网络调查法

- 网络调查法有诸多优势,例如能够高效率和低成本地获得大量多样化的样本,还能对研究中关注较少的群体展开调查。
- 网络调查法的主要缺点包括:可能的反应偏差和选择偏差,无法控制调查情境等。

可以说调查研究是最早被用来研究基于网络的人类行为的方法。在网络调查法中,参与者只需完成在线问卷后点击“提交”按钮。该方法依托复杂的软件程序,能够自动记录和汇总数百万受访者的回复结果。此外,程序还能按照实验要求操控变量、随机分配被试。(参见: Fraley “新手指南——基于网络的心理学研究”, Kraut et al. “有用的网络资源”, 2004)

下面我们将细数网络调查的诸多优势。其最突出的优势就是高效和低成本(参见 Buchanan, 2000; Skitka & Sargis, 2005)。因为使用该方法只需简单的操作就能联络到数以千计的受访者,而且他们在年龄、种族甚至国籍上都各不相同,因此与邮寄调查法、电话访谈法、个人面谈法相比,该方法能够节约大量的时间和人力。由于网络调查法中采用的是在线问卷,因此能够保护自然资源,节约成本。受访者可以在有空时填写问卷,而且对完成问卷的场合亦无限制,无论是家中、办公室、宿舍还是其他地方,只要是能上网的场合均可。

网络调查法除了能够取得大量多样性的样本外,斯基特卡和萨吉斯(Skitka & Sargis, 2005)指出,该方法还能对心理学中得到较少研究的群体展开调查。网络上形形色色的聊天室、新闻组、论坛等都为研究者提供了接近那些在传统调查研究中无法轻易接触到的人群(如宠物主人、仇视团体、癌症康复者、各种罪行的受害人等)的机会。除此之外,全球来自不同文化和国家的人们都越发热衷于上网,因此,网络调查法还将为跨文化研究提供新的契机(如 Gosling et al., 2004)。

当然,基于网络的调查也有其劣势。其中最大的劣势是样本偏差(Birnbaum,

2001; Kraut et al., 2004; Schmidt, 1997)。网络调查中可能出现反应偏差和选择偏差。与其他研究方法一样,当调查对象没有完成调查时就会出现反应偏差。事实上,网络调查法的反应率低于邮寄调查法或电话访谈法(参见 Kraut et al., 2004; Skitka & Sargis, 2005)。我们知道,参与并完成网络调查问卷的人们与那些没有参加调查的人群间存在较大差异。由于参与网络调查的人们在一定程度上也是便利样本,而且只有能上网的人们才能参与,因此该方法还存在选择偏差问题。一般而言,美国收入较高的家庭更有可能安装上网设施,这些家庭中的孩子上网的机会也更多。在美国,具备上网条件的白人与亚裔家庭的数量几乎是黑人或西班牙裔家庭的两倍(Newburger, 2001)。与众多心理学研究选用的大一和大二学生样本相比,美国的网民的代表性(代表全体美国人)更差。

在招募参与者的过程中,如果所采用的方法不当还可能增大样本选择偏差。研究者可以通过以下方式招募参与者:在网站上张贴研究告示(如第1章中描述的与APS网站进行链接),或创建相应的网页(如人格调查)等待网民登录(Krantz & Dalal, 2000)。上述方法比较被动,研究者还可采取更积极的策略,如向那些可能会对该调查感兴趣的团体或个人发送该研究项目的告示。然而,正如斯基特卡和萨吉斯(2005)所指出的那样,不仅网民无法代表总体,而且从特定网民群体中所选取的对象也无法很好的代表其所在的特定群体(如遭受性虐待的受害者们)。迄今为止还没有一种方法能够对网民进行随机抽样(Kraut et al., 2004)。

无法全面控制调查情境也是网络调查法的主要缺陷之一(Birnbaum, 2000; Kraut et al., 2004)。我们在第3章曾经谈到这种缺乏控制可能会导致较严重的伦理问题。参加网络调查的网民通常无法全面了解研究内容,此外,如果该研究可能对参与的网民造成伤害,通常网民也无法获得相应的保护(如,网民在回答问题后可能出现的悲伤情绪)。在网络调查中,由于研究者并未出现在网民面前,因此很难确定调查对象是否真正理解了指导语,他们对该调查的态度是否认真,他们会否轻视还是恶意捣乱,有否有一人多次提交问卷的问题,等等(Kraut et al., 2004)。此外,研究者也无法得知问卷是由一个人还是几个人完成,调查对象所处环境是否嘈杂等(Skitka & Sargis, 2005)。有一位从事网络调查的研究者还担忧调查对象有否遵从指导语,例如,在有关概率与风险的调查中,虽然研究规定不能用计算器,但是调查对象很有可能漠视该规定(Birnbaum, 2000)。

尽管网络调查法存在上述问题,但我们依然认为该方法利大于弊。相信随着技术的日新月异,随着IRB委员会找到保护网民参与者的切实途径,网络调查法也将日臻完善。

调查研究的设计

● 调查研究的设计有三种类型:横断设计、连续独立样本设计和纵向设计。

研究者最重要的决定之一是选择研究设计的类型。调查研究的设计是对整体研究计划或结构的设计,主要有三种类型:横断设计、连续独立样本设计和纵向设计。需要指出的是,没有一种设计能满足所有的研究目标。因此,研究者应该在明确研究目标的基础上选择相应的设计类型。

横断设计

- 横断设计要求研究者必须在同一时间从总体中抽取一个或多个样本。
- 横断设计可使研究者描述总体的特征,或者描述两个或多个总体间的差异。利用横断设计所得的相关分析结果,研究者可以进行预测。

横断设计是最常使用的调查研究设计方法之一。横断设计(cross-sectional design)要求研究者必须同时从总体中抽取一个或多个样本。横断设计的重点是描述,即描述总体的特征,或者描述两个或多个总体间的差异。例如,古德曼(Goodman, 1987)及其同事曾在了一项全国性研究中采用了横断设计,调查全美9~17岁青少年的精神障碍问题。该研究旨在确定该年龄段存在精神障碍的青少年中,接受了心理健康治疗的青少年与未接受治疗的青少年之间是否存在系统差异。先前研究所使用的青少年样本基本来自于临床,古德曼等人质疑这些样本能否代表那些没有获得治疗,却患有精神障碍的青少年群体。因为他们认为在所有患有精神障碍的青少年群体中,只有一小部分得到了治疗。

为了解决这个问题,古德曼等人对美国心理健康研究院(National Institute of Mental Health, NIMH)资助的一项有关儿童青少年精神障碍研究中的数据进行了检验。他们以社区为单位,对来自康涅狄格州、佐治亚州、纽约州以及波多黎各的1285对家长及孩子展开调查,并对他们进行入户访问(由此可见,横断设计不一定简单,过程也不一定快速)。在此基础上,古德曼等人描述了接受治疗和未接受治疗的两个群体间存在的诸多差异。

在该研究所选取的样本中,747名(58%)青少年在过去半年中没有精神障碍,但其中的25人(3%)接受过心理治疗。535名(42%)被诊断患有精神障碍,但其中只有69人(13%)得到了心理健康服务。那么,这69位青少年和其余466位青少年之间究竟存在哪些区别呢?古德曼等人对一些变量进行了考察,结果发现,与其他族裔相比,非西班牙裔白人青少年得到心理治疗的人数较多;处于青春前期的少女接受治疗的人数较少。研究结果发现,接受治疗的青少年还被认为较缺乏能力、被诊断患有两种或两种以上精神障碍。除此之外,两个群体的家长也存在差异。研究发现,得到心理治疗的青少年的家长教育水平较高,对家庭生活满意度较低,不监督孩子,家长自己接受心理健康治疗的也比较多。基于上述数据,古德曼等人认为这些得到治疗的青少年不能代表所有患有精神障碍的青少年。

横断设计非常适合那些旨在进行描述与预测的研究。然而,调查研究有时还需要评估随着时间的变化人们在态度或行为上的改变,有时还需要确定某些自然发生的事件的影响(如:提高饮酒年龄的法案通过后对社会的影响)。如果旨在对

类似问题进行研究,那么就不应该采用横断设计,而是应该采用另两种设计方法。我们将在以下两节分别探讨这两种方法。

连续独立样本设计

- 在连续独立样本设计中,要求来自总体的不同样本在一个时间段内完成调查。
- 连续独立样本设计可使研究者考察总体随时间的变化。
- 连续独立样本设计无法推断个体随时间的变化情况。
- 如果从总体抽取的样本间不具有可比性,即样本对总体的代表性不同,那么连续独立样本设计就会出现问題。

连续独立样本设计 (successive independent samples design) 是指在连续的一段时间内开展一系列的横断调查。在该设计中样本是独立的,因为不同样本必须在不同的时间完成调查。在使用此方法的过程中,必须遵循以下两点:(1)必须让每一个样本完成同一套调查问卷;(2)必须从同一总体中抽取不同样本。只有做到上述两点,研究者才能对来自不同时间的调查结果进行合理比较。该设计方法最适宜于考察某总体的态度或行为随时间的变化情况。例如,公众舆论研究人员经常询问美国公众中的独立样本对美国总统的满意程度(即总统支持率)。研究人员可以通过考察支持率随时间的变化来描述美国公众对总统行为的评价。

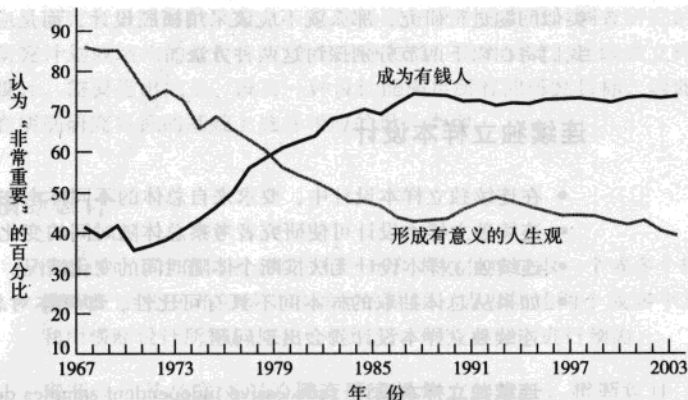
让我们再来举个例子。有一项研究自从1966年以来每年都会开展一次。每年研究人员都会对来自大约700所高校的350 000名全日制大一学生进行考察(Sax et al., 2003)。该研究项目是全美高等教育研究中规模最大、持续时间最长的实证研究。在38年的时间中,来自1 500多所高校的一千多万名大学生参与了该项研究。研究人员向学生提出涵盖不同方面大约40道问题。在几十年间,研究人员虽然会对某些问题做出改动,但几乎每年都向学生提出一些相同的问题。因此,该研究堪称连续独立样本设计的典范。

那么,该研究如何描述几十年间学生的价值观和生活目标的变化呢?每年,该研究项目都会要求学生针对不同价值观进行评定。萨克斯等人(Sax et al., 2003)报告了部分调查结果。研究人员对其中的两个价值观最感兴趣:“形成一个有意义的人生观的重要性”,“成为有钱人的重要性”(pp.6-7)。图5.3显示了认为上述价值观“非常重要”的学生的百分比。由图可见,20世纪60年代后期,超过80%的学生认为形成一个有意义的人生观非常重要——事实上,这也是该时期学生认为最重要的价值观。与此形成鲜明对比的是,在该时期不足45%的学生认为成为有钱人很重要,在60年代后期的大学生中,该价值观在所有价值观中的排名是第五或第六位。

但是,在2003年,这两个价值观的地位却完全颠倒了。73.8%的学生认为“成为有钱人”非常重要,认为“形成一个有意义的人生观”非常重要的比率则降至研究以来的最低点,只有39.3%。观察图5.3我们可以发现,20世纪70年代初这两种价值观在大学生心目中的地位开始发生改变,1977年左右两种价值观的地位持平,进入80年代后期,两者的地位完全颠倒了,而且直至今日仍未发生较

图 5.3 1966 年至 2003 年
美国大一学生在价值观上的
变化趋势

资料来源: Figure 7 (p. 7),
Sax et al. (2003) .



大改变。萨克斯等人(2003)认为这“反映了这一代大学生中外在价值和内在价值之间的持续博弈”(p.7)。除此之外,研究人员还发现2003年可被称为“志愿精神年”,因为83.1%的大一学生声称曾在高三参加过志愿活动,他们认为大学是从事志愿工作的好时机,并表示要在大学期间多参加志愿服务或社区服务工作。

连续独立样本设计也存在不足之处。让我们假设某项采用连续独立样本设计的调查结果如下:1974年,35%的大学生声称不信任美国政府、25%声称很难说、40%的大学生声称信任美国政府。2004年,对大学生提问相同的问题,调查结果则是:50%的大学生声称不信任政府、20%声称很难说、30%的大学生声称信任政府。那么,请问我们该如何解释上述数据?2004年大学生对政府信任度的三种态度的比例为:50:20:30,我们能否据此认为1974年持“很难说”态度中的5%大学生以及持“信任”态度中的10%大学生到了2004年改变了看法,从而变成了不信任政府呢?当然不是!那么,为什么呢?

在解释那些采用连续独立样本设计的调查结果时,我们一定要牢记:在我们刚才假设的调查中,1974年被调查的大学生和2004年被调查的大学生并不相同!如果想要解释特定个体随时间的变化情况,那么就一定要对该个体展开调查。因此,在连续独立样本设计中,我们无法了解是哪些人的观念或态度发生了改变、也无法了解改变的程度。也许你在观察萨克斯等人2003年的调查结果(图5.3),并试图对此进行解释的过程中遇到了类似的问题。图中显示了1966年至2003年大学生在两种价值观上的变化,你可能产生这样的疑问:导致大学生态度改变的原因是什么?显然,该图无法向我们提供这方面的答案。连续独立样本设计的目标是描述总体特征随时间的变化,它无法描述单个调查对象随时间的变化情况。因此,如果希望从图5.3中得到导致变化的原因显然不可能,连续独立样本设计并不胜任此类工作。(下一节我们会介绍另一种调查设计方法——纵向设计,该方法比较适合处理此类问题。)

如果抽取的连续样本无法代表相同的总体,那么就出现了连续独立样本设计中常见的第二个问题。让我们回到刚才假设的美国学生对政府信任度的例子。如果1974年抽取的是来自农村的小型学院的学生,2004年抽取的学生则来自美国大

城市的大学,那么试图比较两者随着时间迁移在对待政府态度上的差异就没有任何意义。我们不能据此认为随着时间变化,大学生们对政府的信任度降低了,因为就读于农村小型学院的学生与就读于城市大学的学生也许在对政府的信任度上本身就存在着差异。因此,该因素也可能会导致1974年与2004年调查结果上的差异。该例子说明,郊区样本和城市样本是不具可比性的连续样本。只有当连续独立样本能代表相同的总体,我们才能准确描述总体随时间的变化情况。因此,确保连续独立样本具有可比性对该设计至关重要。尽管我们也许可以通过复杂的统计程序解决样本间可比性不足的问题,但是,如果我们可以仔细选择连续样本,保证他们能够代表同一总体,那无疑是最佳解决之道。

纵向设计

- 纵向设计在一段时间内调查相同的对象,其目的是研究个体调查对象随时间的变化情况。
- 调查数据揭示的是变量间的相关关系,因此很难确定导致个体随时间发生变化的原因。
- 如果在纵向设计中调查对象中途退出(即样本流失),那可能意味着最终样本与最初样本不再具有可比性,或可能意味着最终样本无法代表总体。

纵向设计(longitudinal design)的突出特点就是对同一样本调查多次。例如,在某项针对抑郁母亲的纵向设计中,研究人员在不同时期对同样的母亲样本进行了数次调查,分别是:怀孕期间、产后三个月、孩子1岁、孩子4岁,以及孩子11岁时(Hay, Pawlby, Angold, Harold, & Sharp, 2003)。研究人员发现,母亲产后抑郁与孩子11岁时的暴力行为存在一定关联。

纵向设计具有两大优势。优势一:研究人员可以确定个体调查对象变化的方向与程度。由于能够对个体的变化进行评估,因此就更易于考察导致态度或行为改变的原因。优势二:如果研究者意欲评估某些自然发生的情境造成的影响,那么纵向设计无疑是首选。例如,研究者可以在大选前后对人们的态度展开调查,由此就能描述该事件的影响。

希瑟顿等人(Heatherton, Mahamedi, Striepe, Field, & Keel, 1997)曾经运用纵向设计考察了大学至成年早期与进食有关的态度与行为的变化情况。尽管人们对青少年和大学生的进食障碍已颇为了解,但是希瑟顿及其同事想要考察进食障碍是如何随着个体的成长而变化,如在个体成家、立业、抚养后代以及获得了更好的自我认同后,其进食障碍的演变情况。该研究的假设是:在个体二十多岁时,随着角色与生活目标的改变,对外貌的关注度将会下降,进食障碍问题也将随之减少。为了验证假设,他们在1982年从美国东北部的某私立大学中随机抽取了800名女性和400名男性,并对他们进行有关进食与节食的调查。女性的反应率为78%($N=625$)、男性的反应率为69%($N=276$)。1992年,研究者在该校的校友办公室的协助下与这些被试再次接触,用同样的问卷,要求他们重新回答有关饮食态度与行为的问题。由此可见,纵向设计的显著特点就是在研究的每一阶段对同一样本开展调查。

不难想像,纵向设计要求研究者付出巨大的努力。当然艰辛的付出最终能让研究者对个体随时间发生的变化有所了解。在希瑟顿等人(1997)的研究中,调查发现两性随着年龄增长,在对待进食与体重增加的态度上存在差异。研究显示,自跨入大学校门后的十年间,女性进食障碍的症状、长期节食习惯,以及对自己体型的不满意程度都呈下降趋势。尽管如此,女性依然对自己的体型非常不满意、依然有很强烈的减肥意愿。男性则不同,研究发现,男性在大学期间很少出现进食和体重上的困扰。然而十年后,他们的体重开始增加(平均增加12磅左右,女性则平均增加4磅)。因此,在1992的调查发现,男性在节食与进食障碍方面的症状有所增加,但与女性相比男性在此方面的症状较轻。

希瑟顿等人(1997)的研究也许能够加深我们对纵向设计的理解。他们认为女性进食问题的减少反映出女性日渐成熟,角色上发生了改变,而且离开了大学校园(大学校园中更强调女性要拥有苗条的身材)。当然,还可能存在其他因素造成她们发生改变。他们还运用连续独立样本设计,在1982年和1992对独立样本进行调查。希瑟顿、尼克斯、马哈迈迪和基尔(1995)指出,在进食障碍症状和对体型的不满意程度方面,1992年的大学生在此方面的问题少于1982年的大学生。上述结果表明,对待进食障碍的态度与行为上的改变也许反映了十年间社会在此方面的改变(例如,媒体有关进食障碍的报道增多)。然而,纵向设计可能存

图 5.4 类似希瑟顿等人(1997)的调查研究旨在考察随着女性的年龄增长,进食障碍如何对她们造成影响



在的缺憾是较难确定造成个体发生改变的确切原因。¹

纵向设计的另一缺憾是较难获得愿意参加这种研究的调查对象。即使调查对象同意参加,也可能存在其他的问题。读者也许认为,纵向设计调查的始终是相同的样本,这就解决了样本间不具可比性的问题,每次调查的样本也都能代表相同的总体。不幸的是,在实际研究过程中,很难保证每位调查对象都能从始至终参加每一次调查。例如,在希瑟顿等人(1997)的研究中,共有901人参加了1982年的调查,但在1992年的研究中,仅有724人(80%)再次参与了调查。

在纵向设计中,除非所有的调查对象都能参与每一阶段的调查,否则就会存在样本流失的问题。样本流失是纵向设计最大的缺陷,因为如果每一阶段都流失一部分调查对象,那也就意味着最终样本与最初抽取的样本间的差异会越来越大。幸运的是,通常我们能够在纵向设计中确定最终样本与最初样本之间的可比性。在后续研究阶段虽然会存在样本流失,但由于他们至少参与了首次调查,因此我们可以了解到缺失样本的特点。因而,研究者可以通过考察他们在首次调查中的反应,确定这些缺失的样本与其他坚持参加调查的样本是否存在差异。

纵向设计的优势是通过多次调查能够确定个体调查对象的变化情况。但在纵向设计中,此优势有时反而会变成棘手问题。因为有些调查对象可能会在多次调查中刻意保持自己的一致性。如果研究目的就是考察调查对象的态度改变,那出现此类问题则会令研究者非常苦恼。因为即使他们的态度确实发生了改变,但为了表明自己始终如一,他们可能依然报告在第一次调查时的态度。纵向设计可能存在的另一问题是,首次调查中涉及的主题也许会使调查对象以后对该问题过于敏感。例如,某纵向设计想了解学生对校园犯罪的看法。在参加首次调查后,调查对象可能会比常人更关注校园犯罪。由于调查对象知道自己在参与此方面的研究,因此其对待此问题的态度或行为也可能发生变化。

与那些想要刻意保持一贯态度的调查对象不同的是,在希瑟顿等人(1997)的研究中,调查对象可能很不情愿承认自己在大学期间的进食问题始终没有得到改善。因此,研究者们在十年后所观察到的进食问题减少也许是由于“这些即将步入三十岁的女性认为进食障碍是青春期的典型问题,要承认自己还有这样的问题令她们感到难堪”(p.124)。在要求调查对象报告他们的态度和行为时,研究者必须留意调查对象的报告与其行为是否一致,如果不一致,研究者必须思考其原因何在。本章后半部分我们还将谈及此问题。

问 卷

即使调查所抽取的样本能完全代表总体、反应率为100%、研究的设计完美无缺,但如果所使用的问卷非常糟糕,那么调查结果仍将毫无价值。在此我们将讨

¹ 希瑟顿等人指出(1997),在纵向设计中调查对象进食障碍的减少程度高于连续独立样本设计所得结果,因此他们认为除了社会在此方面的改变外,个体的成熟过程也可能会减少进食障碍问题。

论调查研究中最常用的工具——问卷。如果研究者希望调查具有价值，那么，问卷在测量样本的人口统计学变量以及测量自陈量表部分的个体差异上，必须具有足够的信度和效度。虽然在如何准备一份好问卷方面我们没有现成的经验可供大家借鉴，但是你必须熟悉构建问卷过程中的几条基本原则。我们将在此部分阐述准备问卷的六个基本步骤，此外，还将向大家介绍有关记录与管理单个问题的一些具体方法。

作为调查工具的问卷

- 大多数的调查研究都采用问卷对变量进行测量。
- 人口统计学变量描述了调查对象的特征。
- 只有在问卷构建过程中做到既专业又细致，问卷才能既准确又有效。
- 自陈量表可用于评估人们的倾向或态度。

调查研究（包括所有研究）的价值最终都取决于研究者所做测量的质量。测量的质量又取决于进行测量时所使用的工具的质量。在调查研究中最主要的研究工具就是问卷（questionnaire）。表面上看来，问卷也许不像许多现代科学研究所使用的高科技工具，但如果问卷构建良好、使用正确，那么它也是测量不同的变量的强有力的科学工具。

人口统计学变量 在调查研究中，人口统计学变量是一种经常被测量的重要变量。人口统计学变量常被用于描述调查对象的特点，如种族、民族、年龄、社会经济地位等都是非常典型的人口统计学变量。在调查研究中无论我们是出于研究或是出于其他考虑，都不可避免地需要对人口统计学变量进行测量。恩特威斯尔和阿斯通（Entwisle & Astone, 1994）认为在调查研究中考察民族和种族非常重要，因为美国人口的构成日趋多样化。恩特威斯尔和阿斯通（1994）指出：“本世纪中

图 5.5 虽然族裔背景是人口统计学中的一个重要变量，但要人们对人们进行分类却并非易事。



叶以来,美国人口的种族和族裔构成日趋多样化,因此在考察美国人的种族时,只能把大多数的美国人粗略地划归为“非白种人”(p.1522)。通过询问调查对象的种族,我们就能对样本的种族构成有更清晰的了解。如果种族与调查问题有关,那还将有助于我们比较不同种族群体间的差异。

测量诸如种族等人口统计学变量乍看上去非常简单。最直接的方法就是用开放式问题向调查对象进行询问:“请问您是什么种族?”此种方法非常直接,但得到的结果也许并不令人满意。例如,有些调查对象可能会混淆“种族”与“民族”。也许调查对象和研究者都不太了解不同民族之间存在的重大差异。如“西班牙人”就不是一种种族划分,它是指所有来自母语为西班牙语的国家的人。因此,如果某人出生在西班牙,那么我们可以认为他是西班牙人。拉丁美洲人则是指所有来自南北美洲国家的人(加拿大和美国除外)。类似区分非常容易造成人们的混淆。例如,某人具有欧洲西班牙血统,自认为是白人而不是拉丁美洲人,那么他的民族就是西班牙人。

总之,如果在调查研究中采用“快速草率”的方法进行测量,那么得到的数据结果必定杂乱无章,研究者也将很难对数据进行分析和解释。例如,许多人可能认为自己是“多种族的”,但如果研究者对该变量不够重视,那么可能导致调查对象信息有误,或者导致研究者根本对此问题视而不见。恩特威斯尔和阿斯通(1994)推荐了一种测量种族时可以借鉴的方法,该方法非常细致而有效。他们总共提出了9个有关种族测量的问题。其中一个问题如下:你认为自己属于什么种族?其他问题还包括你的祖先来自哪国?如果你是拉丁美洲人,具体是来自墨西哥、波多黎各、古巴还是其他地区?……这些题目有助于研究者更准确、更清晰地测量调查对象的种族。列举这个测量种族的例子是想说明一个更普遍的原则:只有在问卷构建过程中做到既专业又细致,才能保证调查研究中所用问卷的准确性和精确性。

偏好与态度 在调查研究中经常会评估个体的偏好和态度。例如,市场研究人员也许对顾客在咖啡品牌上的偏好非常感兴趣,政治团体也许对潜在投票人在具有争议性的公共议题上的态度比较感兴趣。心理学家们则长期对测量人们在众多问题上的观点与感受兴趣十足,他们经常编制自陈量表,并要求人们对量表上的项目做出口头或者书面的回答。

自陈量表(self-report scales,也译作“自我报告量表”)常用于测量人们对问卷项目的反应(如咖啡、离婚、政治候选人、生活事件等),从而确定人们在量表所测的某些维度上的差异(如人格特质、压力大小等)。例如,要求调查对象根据他们所察觉到的压力大小对不同的生活事件进行评定。在此基础上,研究人员就能制定出一份代表着不同压力大小的生活事件量表。此类量表强调的是项目间的差异,而非个体间的差异。如果要测量个体差异,研究人员可能要求调查对象根据压力事件量表,报告他们去年经历这些压力事件的频率。通过将调查对象在每个项目上的反应进行相加,研究人员就能得到每位调查对象的总分。然后,根据不同调查对象所得总分,就能对他们加以比较。²专栏5.3对李克特方法进行了介绍,

² 与等距测量或测量刺激有关的领域是心理物理学;与个体差异测量有关的领域是心理测量学。

专栏 5.3

编制个体差异量表

假设某位研究者想研究个体原谅他人的意愿。如果该研究者决定采用李克特方法,那么第一步就是写下可能反映个体原谅他人意愿的项目。第二步,通常是要求参与者在五点量表的基础上,对每个项目进行评定。参与者可以在其认为合适的数字上画圈。以下为测量个体原谅他人意愿的一个项目:

当人们伤害我时,我通常能原谅他们。

1 ——— 2 ——— 3 ——— 4 ——— 5
非常不同意 非常同意

在采用该方法时要求有一定数量的参与者(至少100人)对上述类似的项目进行回答。最后将每位调查

对象的反应相加就能得到总分。随后研究人员剔除那些区分度较低的项目,从中筛选出最能代表所测范围的项目(在此例中是最能代表原谅他人意愿的项目)。每个挑选出的项目应该与量表总分存在显著相关。项目选择完毕后,将每位调查对象在这些项目上的反应相加就能得到每人的量表总分。编制量表的一个重要目标就是信度。如果采用的是李克特方法,那么研究人员必须保证量表项目每次测量的都是同样的内容(下一节我们将更详细地讨论信度)。需要牢记的是,并非所有的五点评定量表都是李克特量表,只有遵循我们之前描述的方法所得到的量表才是合格的李克特量表。

该方法常用于编制个体差异量表。

自我报告测量或自陈测量是问卷的一种形式,它是心理学研究中最常见的一种测量工具。由于它们在心理学研究中非常重要,因此在进行量表编制时应该非常认真。评定自陈量表优劣的两个关键指标就是信度和效度,当然这也是评定所有测量的重要指标。

自陈测量的信度和效度

- 信度是指评价测量一致性的程度指标,在调查研究中经常用重测信度方法进行评估。
- 一次测量中包含许多相似的项目、对多个样本施测并遵循标准的施测程序能使信度有所提高。
- 效度是指对测量的真实性和准确性的估计,即是否测到了想测的东西。
- 结构效度是指量表能测量所构造概念的准确程度。聚合效度和区分效度可用于验证结构效度。

可靠的自陈测量如同可靠的观察者或其他可靠的测量一样,评价其是否可靠的重要指标就是一致性程度如何。可靠的自陈测量在每次测量后都能得出相似的(一致性)结果。假如要把自陈测量作为进行预测的基础,那么就要求其必须足够可靠。例如,如果某研究想要通过个体的生活压力程度预测与压力有关的健康问题,那么就要求对个体生活压力的测量是可靠的。测量信度有多种方法,最常见的方法之一就是计算重测信度(test-retest reliability)。重测信度是指用同一问卷在

不同时间对某大样本人群施测两次。要确定一个问卷是可靠的,不要求人们在两次测量中得到相同的分数,但是,一个人在两次测量中的分数在分数分布中的相对位置应该是相近的。通过计算样本中的每个人在问卷中的两次分数的相关系数,就能判断这种相对位置的一致性程度。重测信度系数一般要求在.80以上,但是,相关系数的大小还依赖于其他因素,如测验题目的类型和数量等。

刚才我们谈到自陈测量的信度受题目数量的影响,既然如此,也就意味着题目数越多信度越高。如果仅观察棒球手的一次击球就想判断他的击球能力,那么显然所得结论会很不可靠。同样,只向人们提一个问题也无法真正了解他们对死刑的态度。因此,只有当我们在对棒球手的击球进行大量观察、用许多问题询问人们对死刑的态度后,才可能大幅提升测量的信度 (Epstein, 1979)。

除此之外,如果施测的被试较为多样,也将有助于提升测量的信度。测量的目的往往是确定人与人的差异程度。因此,如果被试间差异大,那么对他们进行区分就更容易,如果被试间差异小,想要对他们进行区分就会比较困难。让我们举个例子,假设我们想评估足球运动员的传球水平。如果我们的样本幅度更宽,即多样性丰富,如有职业球员、高中生球员,还有少年球员,就比较容易对他们进行区分。如果只是对职业球员进行评估,显然较难做出可靠评估,因为他们都踢得太棒了!这些职业球员的技术水平差异幅度不大。因此,如果想要问卷具有较高的信度,那么在选取被试时就应该考虑增加被试的多样性,而不是在狭窄的范围内选择被试。

影响信度的第三个因素也是最后一个因素是施测情境。如果施测情境能保证被试安心作答,如果给出的指导语足够明确,那么测量信度也将增加。我们都参加过考试,如果考试期间出现噪音,或者考生觉得试题不知所云,这都可能影响他们的考试成绩。在调查研究中也是同样道理。

与信度相比,确定一个调查测量的效度较为困难。效度的定义看似简单,所谓效度就是正确测量研究者所要测量的变量的程度。很多人都听过学生们抱怨考试试题与平时讲课内容无关。这反映的就是效度问题。

测量效度有许多方法,在此我们主要讨论其中的构念效度。构念效度 (construct validity) 是指问卷能准确测量所构造概念的程度。有两个指标可用于验证构念效度:聚合效度和区分效度。我们将通过下面的例子帮助您理解这些概念。

表 5.1 中的数据来自一项有关“生活满意度”的测量,我们将用这些数据说明如何评估构念效度。卢卡斯、迪纳和苏 (Lucas, Diener, & Suh, 1996) 指出,现在心理学家们开始越来越关注考查诸如快乐、生活满意度、自尊、乐观主义以及主观幸福感等指标。然而,这些不同的指标是否都在测量一个相同的概念(如主观幸福感)?还是每一个指标在测量各自的概念呢?为了回答这些疑惑,卢卡斯及其同事开展了几项研究,在研究中他们要求被试完成这些有关幸福不同指标的问卷。由于我们在此只是想借用该例子说明有关构念效度的概念,因此我们只选取他们第三个研究中的一小部分数据。在卢卡斯等人的第三个研究中,他们要求被试完成三个量表,分别是:生活满意度量表 SWLS、五项目生活满意度量表 LS-5 以及一个正性情感量表 (PA)。在该范例中,我们想要确定的是,生活满意度的概念能否与更广义的幸福(积极情感)的概念相区分。

表 5.1 构念效度范例*

	SWLS	LS-5	PA
SWLS	(.88)		
LS-5	.77	(.90)	
PA	.42	.47	(.81)

*数据来源:卢卡斯等(1996),表3

注:SWLS=生活满意度量表;LS-5=五项目生活满意度量表;PA=积极情感量表

表 5.1 中的数据构成一个相关矩阵,此相关矩阵能清楚表示出多个变量间的相关。请先观察对角线上括号内的数值,这些数值代表了三个量表(生活满意度量表、五项目生活满意度、积极情感量表)的信度。三个量表的信度都在.80 以上,由此可见,三个量表都具有较好的信度。但在该范例中,我们的重点是评估“生活满意度”的构念效度,所以我们应该观察表格中的其他数据。

生活满意度量表(SWLS)和五项目生活满意度量表(LS-5)都旨在评估生活满意度这一概念,因此它们之间存在相关不足为奇。事实上,卢卡斯等人发现两个量表间的相关系数为.77,这说明两者存在相关,这为它们的聚合效度提供了证据,说明两个量表在测量生活满意度这个概念上存在“聚合”。

如果量表间存在区分效度,则能更有力地证明生活满意度的结构。在表 5.1 中,两个生活满意度量表与积极情感量表之间的相关较小:生活满意度量表(SWLS)和积极情感量表(PA)间的相关为.42;五项目生活满意度量表(LS-5)与积极情感量表(PA)的相关为.47。这表明生活满意度与另一理论概念(积极情感)间不存在显著相关,即两个生活满意度量表所测的概念与积极情感所测的概念不同。因此,在生活满意度与积极情感之间就存在了区分效度。这说明对生活满意与广义的幸福并不相同。在该范例中,因为既有聚合效度的证据又有区分效度的证据,说明了生活满意度量表的构念效度良好。

编制问卷

- 编制问卷包括确定要探究的信息、确定问卷类型、拟定问卷初稿、问卷前测以及确定正式施测程序。
- 问卷措辞应该简明扼要,尽量使用简单、常见和直截了当的词汇。
- 应该仔细考虑问卷中问题的编排顺序,因为提问的顺序可能会影响被试的回答。

问卷编制步骤 编制一份具有良好信度和效度的问卷并非易事。在此我们将提供一些问卷编制的指导原则。如果您第一次在研究项目中编制问卷,我们希望以下内容能对您有所帮助。在您准备问卷的过程中,可以遵循以下六个步骤:

1. 确定要探究的信息
2. 确定将使用的问卷类型
3. 拟定问卷初稿

4. 检查和修改问卷

5. 问卷的前测

6. 编辑问卷并说明具体施测程序

步骤一：在编制问卷时，第一步就是确定需要寻找什么信息，这也是整个调查研究的第一步。也就是说，你必须首先确定问卷中包括哪些类型和性质的问题。先预测提出这些问题可能出现什么结果，随后判断这些“结果”是否能够回答研究所提出的问题，这一步非常重要。通常，调查研究的时间非常紧迫，这往往使经验不足的研究者失去耐心。这时候我们需要记住的是：一份构建和编制粗劣的问卷与一份构建和编制完善的问卷在施测和分析时都需要投注同样的心力和时间。两者的区别在于：构建和编制完善的问卷能使研究结果具有价值，而一份构建和编制粗劣的问卷其价值仅在于让人得到一个教训，使他们最终明白计划阶段的深思熟虑是何等重要。

步骤二：确定需要探寻的信息后，下一步就是确定问卷的类型。例如，是采用被试自己作答的方式，还是让训练有素的访谈者提出问题？这主要取决于所选取的调查方法，如果你决定采用电话访谈，那么就必须选用训练有素的访谈者。在设计问卷的过程中，研究者还可以考虑选用其他研究者编制的问卷。例如，你想评估种族偏见问题。如果前人已经编制了一份相关的问卷，而且信度和效度良好，那么你就没有必要再去重新编制一份。此外，运用前人的问卷进行调查，你还能够将自己的研究结果直接与前人的结果进行比较。

步骤三：如果你发现没有现成的问卷能够满足你的研究目的，那么就不得不进入第三步——拟定问卷初稿。在拟定初稿过程中需要注意措辞及项目顺序。我们将在后面对此作进一步论述。

步骤四：此步骤涉及问卷的检查和修改，这是问卷编制中的重要一环。编制者认为客观和清晰的问题，也许在别人眼中却不够客观明确。因此在此步骤，最好能够请一些在调查研究方法方面以及对你所研究领域熟悉的专家共同检查问卷。例如，你想调查学生对校园餐饮服务的态度，那么你最好能够让学校负责餐饮工作的人员检查你的问卷，听取他的意见和建议。如果你想研究的是富有争议的主题，那么听取两方面代表的意见就显得尤为重要，因为这可以避免一些可能导致偏差的项目。

步骤五：在编制有效问卷的过程中最关键的一步就是问卷前测。前测是指尽可能在与正式测试相同的情境下，用小样本被试进行测试。前测所选取的被试也必须能够代表最终样本中的被试。如果你的最终样本是大学生，但你却在前测中对居住在疗养院中的老人进行调查，那么这种前测就没有任何意义。不过，前测与正式施测确实存在一点不同。那就是，在前测中，要对被试进行详尽的调查，了解他们对每一道题目以及整个问卷的反应。这有助于研究者剔除一些意思含混或者唐突的题目。

对于访谈者而言，前测就是对未来正式施测的“彩排”。在前测过程中，应该

对访谈者进行密切监督,保证他们理解了施测程序,并监督他们在此过程中有否严格遵循施测步骤。如果在前测后需要做出较大改动,那么就可能需要再进行第二次前测,以确保第一次前测出现的问题已经得到了解决。

步骤六:前测完成后,编制问卷就进入了最后一步,即编辑问卷,并确定正式施测中的具体程序。要想成功完成最后一步,需要注意问卷的措辞与项目的顺序。

问卷措辞的指导原则 律师们深知问题的表达方式会对回答造成巨大影响。我们从事调查研究的人们也能对此给予同样关注。洛夫特斯(Loftus, 1979)曾在有关头痛治疗的研究中遇到了此问题。她在研究中发现,不同的提问方式会得到调查对象不同的答案。当向调查对象提问“你是否经常头痛?如果有,请问频率如何?”,和问“你是否偶尔头痛?如果有,请问频率如何?”时,在第一种提问方式中,调查对象报告的次数更多。在调查中发现,一字之差可能会导致完全不同的结果。然而,表达方式的不同会对被试作答造成多大影响往往只有在测试结束后才得以知晓。

克拉克和肖伯(Clark & Schober, 1992)指出,被试在面对提问时往往会对问题的含义做出“想当然”的理解。这一点对选择问卷的措辞具有重要的启示。例如,问卷中某个词语的含义比较模糊,面对此种情况,被试可能会根据自己的嗜好、根据他“想当然”的理解对该词作出解释。因此,当面对诸如“很少”、“经常”、“能源危机”等词汇时,不同的被试可能会有不同的理解。此外,被试还可能假定调查中的问题与他们所在的文化或亚文化一致。例如,在大众文化中的“坏”与“好”在某些方面是否会有一致的意思呢?克拉克和肖伯(1992)曾提到这样一个例子:某调查人员想询问居住在尤卡坦半岛的墨西哥居民“你有多少孩子”。在把该问题翻译成西班牙文的过程中,他将“孩子”翻译成了niños,但是在当地居民的理解中,niños还包括曾经夭折的孩子。此外,被试还可能想当然地认为调查所提问题肯定是他们能够回答的问题。受此影响,当被试面对一些没有答案或者没有正确答案的问题时,他们都有可能给出答案。例如,要求被试对世界上不存在的国家发表意见时,他们仍可能回答得振振有词。

虽然我们知道问卷的措辞会对测量造成影响,但要找到有效对策却并非易事。因此,在描述调查结果时,至少应该说明问卷中的一些重要问题的表达方式。正因为问卷措辞可能会对测量造成影响,所以在调查研究中人们才会选用多种方法以减少此方面的影响。

调查研究人员在编写问卷的过程中经常会选用两种提问方式。第一种方式为自由反应式问题(即开放式问题),第二种为封闭式问题(即选择题)。自由反应式问题类似日常考试中的论述题,要求你在一定范围内自由作答。如,“请说出你对堕胎合法化的看法”就是一道自由反应式问题。封闭式问题则要求在备选项中进行选择,如某项有关社区警察工作质量的调查中有这样的问题:“你觉得社区警察的保护工作做得如何?”选项有:很好、好、无所谓好坏、不太好、很糟糕。

自由反应式问题的主要优势在于能够给予被试较大的灵活度。然而,在记录被试的回答和对被试反应进行记分的过程中,往往会发现该优势转变成了劣势。如,为了能够对被试多样性的回答进行总结,研究人员需要对被试的反应进行大

量的编码。封闭式问题的作答则较为迅速和容易,也不需要大量繁重的记工作。然而,封闭式问题的主要缺点在于降低了被试的自主性。此外,问卷有时所提供的备选项没有一项符合被试的情况,此时被试因为没有其他选择,只能退而求其次,勉强从备选项中选择一项。由此可见,采用封闭式问题的问卷可能无法准确反映被试的真正选择。

在问卷编制过程中除了考虑问题类型外,还需注意问卷中的词汇应该尽量简单、直接,尽量选用所有被试都熟悉的字眼。问题应该清晰具体,避免双重提问。何谓双重提问?让我们举个例子来加以说明。如这样的提问:“你最近有否感觉头痛、恶心?”。如果被试最近没有头痛也没有觉得恶心,那么他会回答“不”;如果最近出现了其中的一项症状,那么被试可能就会回答“是”。因此,当研究者对此类双重提问的反应进行分析时,往往无法确定被试真正的答案是什么。解决双重提问非常简单,将其拆成两个分开的问题便可。此外,应该在保证提问明确的基础上尽量缩短问题的长度。大多数调查问卷中的问题字数都控制在20字或20字以内。研究人员还应该反复斟酌问题的可读性。如果提问中有条件句,那么应该将条件或假设放在句子开头。请看以下两种提问方式,方式一:“如果你现在被迫离职,请问你会找什么工作?”;方式二:“你会找什么工作,如果你被迫离职?”显然,第一种方式在可读性上更胜一筹。

问卷中还应该避免引导性和暗示性提问。下面的提问方式就是引导性提问,“大多数人都赞成使用核能源,请问你怎么看?”。为了避免引导性提问可能带来的偏差,在提问中要么罗列出所有的观点,要么一个也不提及。例如,可以这样提问:“有些人赞成使用核能源,有些人反对使用核能源,还有些人对此不置可否,请问你怎么看?”或者直接问:“请问你怎么看待核能源的使用?”暗示性提问是指在提问中包含了某些带有感情色彩的词汇。如激进、种族主义者就带有一定的感情色彩,应该避免使用此类词汇。避免使用暗示性提问的最好方法是请一些持不同的社会和政治观点的人们检查你的问卷。

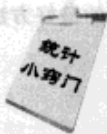
总之,一份好的问卷应该做到以下几点:

- 选用简单直接且所有被试都熟悉的词汇
- 明确具体
- 避免引导性、暗示性提问和双重提问
- 控制问题字数(20字以内)
- 条件句放在句子开头
- 确保问题的可读性

问题的顺序 编制问卷过程中的另一重要环节就是确定问题的顺序。问卷的最初几个问题尤为重要,因为它们不仅确立了其余问题的基调,而且也影响被试回答其余问题的意愿和态度。如果是自陈量表,那么研究者就应该将最有趣的问题放在开头以吸引被试注意。在自陈量表中,询问被试个人信息的问题应该放在最后。如果采取的是个人面谈法或者电话访谈法,那么要首先询问调查对象的个人信息,因为这些问题很容易回答,这样做可以提高调查对象的信心。此外,如果调查中要涉及某些敏感话题,先询问调查对象个人信息还有助于与调查对象建立融洽关系。

有时问卷中某些问题的位置会对调查造成巨大的影响。比如舒曼、普雷瑟、路德维希 (Schuman, Presser, & Ludwig, 1981) 等人的研究中就曾提到此类问题。该研究所用的一份问卷中有两道问题涉及堕胎, 一个较为一般, 一个较为特殊。较为一般的问题是: “如果某孕妇已婚, 而且不想再要孩子, 你认为她是否可以合法堕胎?”; 另一题较为特殊: “如果某孕妇腹中的胎儿很有可能存在某些方面的先天不足, 你认为她是否可以合法堕胎?” 研究者发现, 如果把较为一般的问题放前面时 60.7% 的人回答“是”; 如果放在后面, 则只有 48.1% 的人回答“是”。如果把较特殊的那个问题放在前面时, 84% 的人回答“是”; 如果放在后面, 83% 的人回答“是”。有一个公认的方法可用于解决此类问题, 即运用漏斗式提问法: 就某一设定的主题, 先提问最一般性的问题, 逐渐过渡到提问更特殊的问题。

在设计问卷的提问顺序上还可以考虑使用过滤法, 即先向被试询问比较一般性的问题, 在此基础上确定是否还需要向他们进一步询问更具体的问题。例如, “你是否想有一辆车?” 的问题就应该放在诸如“养车费用多少”等问题的前面。在该例子中, 只有当调查对象回答“是”时, 才有必要进一步询问更具体的问题。在自陈量表中, 针对被试的不同回答, 问卷中会有相关的指示告诉被试接下来该回答哪一题。如果过滤式问题中包括对客观信息的提问, 比如“你超过 65 岁了吗”, 此时进行过滤就变得较为直接。在将行为或态度问题作为过滤式问题时必须格外小心。史密斯 (Smith, 1981) 曾在调查中询问被试是否同意“在任何情况下”都可以打人? 理论上来说, 对大多数这样的一般性问题的否定回答可能意味着在与此有关的具体问题上的回答也是“不”。但是, 在上面的例子中, 对此一般性问题回答“不同意”的人中, 超过 80% 的人在后面的报告中认为在某些特定情况下, 如自我防御时可以打人。虽然采用过滤法应该谨慎, 但由于运用此法有助于缩短施测时间, 因此在问卷编制中这仍是一种重要的手段。



一个好的调查研究能够有效完成描述和预测等研究目标。即使无法找到几百位调查对象, 但通过向几十位调查者发放一份长度适中的问卷, 也能迅速获得几千个问题的数据。而且我们知道, 运用网络调查还能够使研究者在短暂的时间内获得百万份数据。但有个条件! 我们该如何处理如此众多的数据呢? 答案是: 通过缜密的计划!

在确定调查项目之前, 研究者就必须考虑到如何对问卷中获得的数据进行分析。问卷采用开放式问题吗? 研究目的主要是进行描述吗? 描述事件在目标总体中的百分比还是进行变量间的相关研究? 项目反应采取“是/否”的形式, 还是采取“是/也许/否”的形式呢? 采用自陈量表形式吗? 上述不同的形式都将导致不同的数据结果。采用开放式问题所得到的是定性数据, 对此类数据进行分析必须先进行编码, 还要确保编码员间信度。如果项目反应采取“是/否”形式, 那么获得的就是称名数据, 这是分类数据的一种, 但是量表最常提供的是等距数据 (参见第 4 章的量表类型)。在进行数据分析时, 数据类型不同, 所得结果也将不同。

在前一部分我们谈到, 必须先预测信息在问卷中可能出现的结果, 随后才能判断它们是否与本研究密切相关。在“预测”可能出现的结果时, 你必须确保数据分析能够解决你的研究问题。换句话说, 在施测前你就必须有一份数据分析计划。为了确保今后能够得到正确的数据分析, 我们建议你在计划阶段向调查研究领域的专家进行咨询。

至于数据分析的步骤,请参见本书第12章和第13章。如果你对于分析分类变量(称名变量)间的关系非常感兴趣,我们建议你翻阅其他书目。考察称名变量间相互关系的常用统计分析方法是卡方检验。几乎在所有的统计书中都会介绍该检验方法(例如,Zechmeister & Posavac, 2003)。如果你想分析等距变量间的关系,那么可以采用皮尔逊积差相关分析方法。本书第4章讨论观察者间信度时曾经介绍过此类分析方法。本章最后,我们还将进一步探讨相关分析的方法。计算皮尔逊相关系数的方法将在本书第12章进行介绍。

关于调查研究的批判性思考

报告行为与实际行为的一致性问题

- 在调查研究中,由于调查对象知道自己的反应会被记录,因此就可能伴随反应性效应。
- 社会赞许是指被试有时并不表现出自己的真实想法,而是迫于压力报告“应该如此”的想法。
- 研究者可以通过将所得结果与档案数据或者行为观察结果相比较来评估调查的准确性。

无论你在收集和分析调查数据时多么仔细,如果调查对象的回答不真实,那么这些数据都将毫无价值。那么,我们能否相信调查对象的回答反映了他们真实的思想、意见、感受与行为?有关调查研究中口头报告的真实性问题始终极富争议,迄今还未出现统一的说法。然而,在日常生活中,我们往往认为他人的口头报告是真实的。如果你的朋友告诉你喜欢某部小说,你可能会问他喜欢的原因,但是通常你不会怀疑朋友喜欢此小说的真实性。在我们的日常生活中,其实在某些情况下我们有充足的理由去质疑他人所说内容的真实性。例如,当我们想要买一辆二手车时,通常我们不太信任销售人员的花言巧语,然而我们通常比较相信他人的面部表情。对于来自调查对象的信息,我们也应该采取同样的标准。

从本质上说,调查研究会伴随测量上的反应性效应。调查对象不仅知道他们的反应会被记录下来,而且可能怀疑他们的反应会对社会、政治或商业行为产生影响。由于调查对象感受到了这方面的压力,因此他们可能报告那些他们认为“应该如此”的想法,而不报告他们的真实想法。描述此类压力的术语就是社会赞许(social desirability)。另一个术语——政治正确(political correct)也会导致类似的压力。例如,如果询问调查对象是否应该向穷人提供帮助,大多数人会说应该提供帮助,因为他们认为这是最被社会接纳的态度。在调查研究和观察研究中,对待此类问题的最佳方法就是充分意识到它的存在。

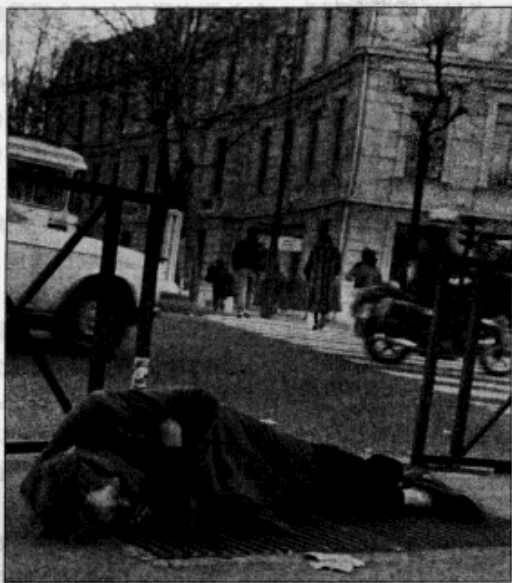
有时,研究者可以直接考察口头报告的真实性。例如,贾德等人(Judd et al., 1991)曾描述过帕里和克罗斯利(Parry & Crossley)在1950年开展的一项研究,

在该研究中,先是由经验丰富的访谈者收集调查对象的反应,随后将所得结果与从相关部门获得的这些调查对象的档案记录进行比较。结果发现,40%的调查对象在一道关于为联合基金(United Fund,一个慈善组织)捐款的问题上撒了谎,25%的调查对象声称他们登记并投票了(其实没有),17%的调查对象谎报了年龄。悲观主义者可能由此认为调查研究中的谎报几率太高了。乐观主义者则会发现,即使人们在被问及慈善捐款时会面临巨大的社会赞许压力,但大多数调查对象的口头报告仍是真实的。

评估口头报告真实性的另一种方式是直接观察被试的行为。拉塔内和达利(Latané & Darley, 1970)曾经在他们的研究中使用过该方法。他们发现旁观者在独自一人的情况下更可能帮助受害者,而当多人在场时,这种救人的意愿反而会降低。随后,他们询问第二组被试他人在场是否会影响他们救助受害者的可能性。被试们一律回答不会有影响。因此,个体的口头报告与行为并不一致。上述研究表明,当我们仅以口头报告为依据试图对调查结果进行推断时必须非常谨慎。当然,当我们仅以直接观察为依据试图对调查结果进行推断时也必须同样谨慎。观察到的行为与口头报告间的这种潜在差异又一次说明:只有采用多元方法才能使我们更全面、更准确地理解人们的行为和心理过程(图5.6)。

关于调查研究的价值,我们最后想说的是:做一个调查研究的过程与写一篇文章存在很多类似之处。在写作过程中,首先需要确定写作主题,在此过程中逐渐缩小范围,否则就会对文章失去控制。在缩小范围的过程中,你有时可能感觉最初的主题已经变得微不足道了。没有一篇文章能够涵盖原先主题的所有方面,同样,没有一份问卷能够涵盖所有的研究主题,无论该问卷是20个项目还是200个项目。调查研究与其他科学事业一样,是建立在这样的信念之上——只有在综

图5.6 面对此情此景,人们经常言行不一



合一系列有限研究的可靠结果基础上,才能增进我们对那些我们面临的更广泛问题的理解。

相关与因果关系

- 如果两个变量间存在相关关系,那么我们就能够对它们进行预测,但是我们无法推断它们之间的因果关系。
- 如果两个变量间的关系可以通过第三个变量进行解释,那么这种关系就被称为“假相关”。
- 采用多种方法考察变量间的相关将有助于研究者确定影响行为的可能原因。

相关研究非常适合那些试图对现象进行描述与预测的研究。在相关研究中经常采用调查法。例如,鲁宾及其同事(Lubin et al., 1988)曾经向人们证明了相关研究在心理学研究中的用途。他们对居住在美国的1543名成年人的生理健康与心理健康之间的关系展开调查。该调查要求被试评定自己的生理健康状况,并向研究人员提供与健康有关的信息,如:去医院的频率、有否接受药物治疗、接受药物治疗的频率,等等。除此之外,还要求每位被试完成一份有关积极情感(如:快乐)与消极情感(如:抑郁)的自陈问卷。调查发现,心理健康与生理健康间存在相关。例如,研究结果表明,积极情感与生理健康的自评之间存在正相关,快乐的人身体更健康等。类似上述的相关结果有助于人们预测生理健康与心理健康问题。

变量间的相关证据为预测提供了可靠的基础。然而,我们都熟悉这句忠告,“相关不代表因果关系”。当我们试图在只有相关证据的基础上推导变量间的因果关系时,必须意识到相关证据存在的巨大局限性。例如,有研究表明,爱好交际与生活满意度之间存在正相关(Myers & Diener, 1995)。然而,我们不能仅根据该相关证据就理所当然地认为喜好交际将使人们对自己的生活更满意。虽然喜好交际确实可能使人们更满意自己的生活,但颠倒两者的因果关系后也许推导依然成立(如:因为人们对自己的生活感到满意,所以更愿意参与社交活动)。相关证据无法说明变量间何为因、何为果。它无法说明究竟是由于喜好交际使人们对生活满意,还是因为对生活满意所以才喜好交际。由此可见,我们不可能仅根据变量间的相关证据就确定其因果关系。

无法确定两个变量间何为因、何为果只是相关研究的缺憾之一。在两个存在相关的变量间还可能存在着另一种因果解释。例如,我们可能找到第三个变量,如朋友数量越多可能导致人们更喜好社交,从而对生活更加满意。这种可以由第三个变量进行解释的相关关系称为假相关(spurious relationship)(Kenny, 1979)。在上述例子中,“朋友数量”就可能是说明喜好交际与生活满意度之间的第三个变量。拥有朋友较多的个体可能会比那些朋友较少的个体更愿意参与社交活动,对自己的生活更满意。这并不是说原来在爱好社交与生活满意度之间的正相关不存在(实际上它确实存在),它只是意味着也许其他变量(例如朋友数量)能够解释两者之间为什么存在这种关系。

专栏 5.4

连接相关与因果的路径

在处理相关数据时可以采用路径分析法 (path analysis), 该方法是一种较为复杂的统计技术 (Baron & Kenny, 1986; Holmbeck, 1997)。路径分析能够确定中介变量和调节变量。中介变量用于解释两个变量之间的相关。调节变量能够影响两个变量之间的方向或强度。在瓦伦丁纳、霍拉汉、穆斯 (Valentiner, Holahan, & Moos, 1994) 的研究中曾经对上述两种变量进行了说明。在该研究中, 研究者采用了相关研究的方法, 试图探究父母支持与适应大学生生活之间的关系。对父母支持的评估是通过向大一学生进行提问, 如询问他们“请问你的母亲/父亲理解你的感受吗?”, “你的母亲/父亲对你苛刻吗?”, 等等。在随后的两年中, 研究者对大学生的主观幸福感与忧虑程度进行测量, 观察父母支持与大学生心理调适之间的相关。大学生从父母那里感受到的积极支持与他们在随后两年中的主观幸福感与低忧虑程度存在正相关。此种关系称为“直接关系”, 因为一个变量 (父母支持) 与另一个变量 (心理调适上的变化) 之间存在直接的相关关系 (参见图 5.7 中的路径 a)。

瓦伦丁纳及其同事试图解释为何存在上述的直接关系。他们认为也许学生应对压力的方式是介于父母支持与适应之间的变量, 或者它有助于解释两者间的关系。为此, 他们特意测量了学生过去 12 个月中处理重大问题时, 采用“积极应对”多于“消极回避”的程度 (p.1096)。“积极应对”是指个体主动解决问题或改变对问题的看法; “消极回避”则指想要忘记和忽视问题的存在, 或者采取回避问题本质的行为 (如: 一味发牢骚等行为)。通常, 我们认为与“消极回避”相比, “积极应对”更具适应性、更能适应环境。研究结果与瓦伦丁纳等人的假设相一致。研究发现, 父母的支持与学生采用“积极解决方式”的百分比之间存在正相关 (参见图 5.7 中的路径 b)。由此可见, 父母的支持与大学生采用更具适应性的“积极应对”之间存在相关。

上述发现还只是研究结果的一部分而已。除此之外, 瓦伦丁纳等人的下一步研究目标则是试图探究“积极应对”与“积极的心理调适上的变化”之间的关系 (参见图 5.7 的路径 c)。只有探究此部分关系才能真正了

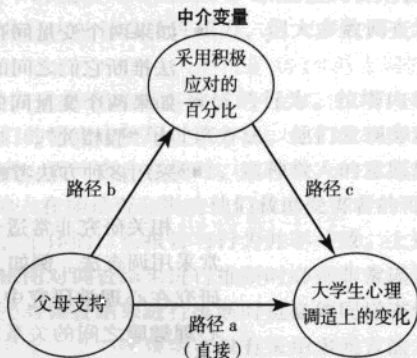


图 5.7 父母的支持与对大学生生活的心理调适上变化之间的直接关系 (路径 a) 和间接关系图解 (路径 b 和路径 c)

摘自瓦伦丁纳等人 (1994) 研究中的图 1。

解变量间的相关关系。研究表明, 那些较多采用“积极应对”的大学生与较多采用“消极回避”的大学生相比, 前者在两年中体验到更多的幸福、更少的忧虑。根据该相关模式, 瓦伦丁纳及其同事就能够做出这样的推论: 父母的支持确实与学生的适应之间存在直接关系 (路径 a)。此外, “积极应对”在父母支持与学生适应之间扮演着中介者的角色, 或者说它能够对父母支持与学生适应之间的关系做出解释 (路径 b 和路径 c)。因此, 父母的支持与学生良好适应之间为何存在关系的一个可能的理由是: 父母的支持也与学生采用更具适应性的“积极应对”策略有关。更正式的说法是: 父母的支持与学生的心理调适之间存在间接的关系, 两者的中介变量是学生在处理重大问题时采用的应对策略。

运用路径分析, 瓦伦丁纳等人 (1994) 还能对调节变量加以确定, 该调节变量影响了父母的支持、学生处理重大问题时时的应对策略以及学生心理调适上的变化之间可观察到的关系。在该研究中, 调节变量是学生对其过去一年中“最重大的问题”的可控性的认识。由此, 瓦伦丁纳等人将样本分为两组, 其中一组认为自己在过去一年中所面临的“最重大问题”的发生是可以控制的; 第二组的学生则认为他们无法控制“最重大问题”

(续)

的发生。随后,瓦伦丁纳及其同事重复以前的相关分析方法,分别对两组学生进行相关分析以确定两组学生是否能够得到同样的相关模式。研究者这样做主要是为了确定调节变量,因为只有确定了调节变量才能了解变量间的相关关系哪些情况下成立、在哪些情况下相关较弱或者相关不成立。

对那些报告“最重大问题”是可以控制的学生的相关分析得到了什么结果呢?研究者得到了有关这组学生的一个非常清晰的相关模式。在该模式中,父母的支持不再与心理调适上的变化直接相关(路径a)。然而,父母的支持与适应之间的间接关系依然存在,即“积极应对”的中介作用依然存在(路径b和c)。由此可见,在该组中,父母的支持有助于学生在面临重大问题时运用更有效的应对策略,这与学生心理调适上的积极变化存在相关。

对另一组报告“最重大问题”是不可控的学生的相

关分析又得到了什么结果呢?分析结果表明,在该组中,父母的支持对学生两年中的心理调适起到了直接的作用(路径a)。然而,研究发现学生处理重大问题时的应对策略与学生的心理调适之间不存在相关关系(路径b和路径c)。为何在该组中适应性的应对策略与学生的心理调适没有关系呢?瓦伦丁纳等人认为可能当学生认为事件无法控制时,采用何种应对策略解决问题都无关紧要,就像人们用头撞墙般的无济于事。当学生认为事件并不由他控制时,也许最佳的策略不是努力解决问题,因为在他们眼中这些问题根本无法解决。因此,在此种情况下,父母的支持对学生心理的积极调适就显得尤为重要了。

路径分析以及其他复杂的统计技术能够极大提高我们对相关研究结果的解释能力。在此需要牢记的是:基于对中介变量与调节变量的分析,我们还是不能明确变量间的因果关系。

我们必须理解为什么仅根据两个变量间的相关证据不能推论其因果关系的原因。同样,我们也必须认识到,相关证据对于我们确定那些导致行为产生的潜在原因十分有用。运用复杂的统计技术,有助于我们对相关研究做出因果性的解释(参见专栏5.4)。

安德森(Anderson, 2001)对有关“热假说(heat hypothesis)”研究的总结,为相关结果如何用于确定因果关系提供了一个好例子。热假说认为“高温天气会增加人们的攻击性动机和攻击行为”(p.33)。该假设暗示着这样的因果关系:高温会导致人们的攻击性动机与攻击行为有所增加。安德森认为,以前围绕该假设开展的研究得到的都是相关证据。例如,前人的研究发现夏季的攻击行为多于其余几个季节。但在此情况下,我们也可以通过第三变量进行解释。例如,因为夏季人们在户外活动的时数多于其他三个季节,因此增加了人们之间发生冲突的机会。这表明:此相关证据尚不足以证明高温是导致攻击性行为和攻击动机增加的原因。

近年来有关该方面的研究已经克服了相关研究无法说明因果关系的困境。研究者利用较复杂的统计技术(如路径分析)分析相关研究的结果(Baron & Kenny, 1986; Holmbeck, 1997)。早期的相关研究结果还促使研究者通过对人和动物进行实验来检验该假设。采用多种方法能够向人们提供更有说服力的证据,这些研究已经证明了高温不仅与攻击性有关,而且高温会导致攻击性的提高。由此可见,在探讨变量间因果关系的过程中,相关证据功不可没。相关证据不仅为因果推论提供了坚实的基础,而且它还是确定行为原因的第一步。

小 结

调查研究有助于我们准确有效地描述人们的特征(如:人口统计学变量)以及人们的思想、意见与感受。除此之外,分析自然情境下发生的变量间的相关关系还有助于我们对此做出进一步的预测。调查的目的与范围虽不尽相同,但是所有的调查都涉及抽样问题。通过仔细的抽样能够使我们对本样本所代表的总体进行描述。调查还包括采用一系列预先确定的问题,通常以问卷的形式出现。

在抽样技术中,最重要的是从抽样架构中确认和选择元素,研究者必须确保所选择的元素能够代表相应的总体。只有样本能够代表相应的总体,即样本与总体具有相同的特征,我们才能将研究结果推断至总体。保证样本代表性的方法是随机抽样方法而不是非随机抽样方法。简单随机抽样是最常见的随机抽样方法,该方法能够保证总体中的每个元素都有同等被抽取的机会。如果研究者对总体中的子群体感兴趣,那么可以采用分层随机抽样方法。

调查研究的方法主要有四种:邮寄调查法、个人面谈法、电话访谈法和网络调查法。邮寄调查法可以避免访谈者偏差,该方法特别适合于调查较私人的问题和较尴尬问题。但是,邮寄调查法最严重的问题是存在反应偏差。个人面谈法与电话访谈法通常具有较高的反应率,而且能够给予被调查者较大的自由。电话访谈法经常被用于进行简短的调查。网络调查法非常高效,而且节约成本,它为调查研究提供了新的机遇。然而,网络调查法容易存在样本偏差,并且由于研究者对调查情境缺乏控制,容易引发方法上以及伦理上的问题。

调查研究的开展必须遵循一份全面的计划,即研究设计。调查研究有三种设计方法:横断设计、连续独立样本设计和纵向设计。横断设计的重点是描述,即描述同一时间总体的特征,或者描述两个或多个总体间的差异。如果研究者想要探究人们的态度或观念随时间的变化,则就可以考虑采用连续独立样本设计或纵向设计。通常研究者更倾向于采用纵向设计,因为该方法能使研究者考察特定群体的变化,同时避免连续独立样本设计中样本间不具有可比性的问题。

调查研究中的主要工具就是问卷。问卷可用于调查样本的人口统计学变量,还可用于调查人们的偏好或态度。为了编制具有良好信度和效度的问卷,研究者必须确定所要寻找的信息、问卷的类型、最有效的提问方式等。除此之外,研究者还必须确保问卷中的问题简明扼要。

与其他口头报告类似,调查研究所得的结果可能并不完全可靠,例如,研究者必须考虑到该调查有否对被调查者施压,是否会使他们只表现出被社会赞许的反应。此外,人们经常言行不一,因此调查研究无法取代直接的观察。但是在考察人们的态度与观念上,调查研究无疑是一种极佳的研究途径。

对相关证据进行解释中存在的最大挑战是理解相关与因果间的关系。两个变量间存在相关并不必然意味着两者间存在因果关系。然而,通过采用多种研究方法、结合复杂的统计技术(如:路径分析中对中介变量与调节变量的分析),研究者能够把相关证据用于确定变量间的因果关系。

重要概念

相关研究

总体

样本

代表性

选择偏差

非随机抽样

随机抽样

简单随机抽样

分层随机抽样

反应偏差

访谈者偏差

横断设计

连续独立样本设计

纵向设计

问卷

社会赞许

假相关

复习思考题

1. 简述相关研究与实验研究的区别，并说明调查研究的主要目标。
2. 假设某调查结果与该研究的主办方之间存在利益关系，请简述如何检查该调查的结果有否存在偏差。
3. 调查研究的目标虽然各异，但有两点是共同的，请问是哪两点？
4. 为什么在调查研究中采用便利抽样可能严重影响结果的解释？
5. 请解释总体同质性与为了确保代表性而必须保证的样本大小之间的关系。
6. 如果你要开展调查研究，请说明你选择邮寄调查法、个人面谈法、电话访谈法或网络调查法的原因。
7. 某调查研究的反应率为 50%，我们不能简单认为该调查存在反应偏差，为什么？
8. 简述网络调查法的主要优缺点。
9. 在连续独立样本设计中，为了能够解释群体随时间的态度变化，对样本之间的关系有一定的要求，请问是什么要求？
10. 假设你对评估个体的观点随时间在方向与程度上的变化很感兴趣，那么你该选择何种调查设计方法？为什么？
11. 分别简述一种确定自陈问卷信度和效度的方法。
12. 简述在调查研究中影响自陈测量信度的三种因素。
13. 如果有人告诉你调查对象没有真实作答，请问你该怎么办？
14. 为什么“相关不代表因果关系”？并说明怎样才能将相关证据用于确定行为的原因。

挑战性问题

1. 想要出色地完成一项调查研究并非易事。想要调查人们对待性的态度以及人们的性行为则更不容易。某人想在其著作中谈论有关女性性欲的问题，为此他向全美43个州的不同团体和组织中的女性邮寄了总共100 000份问卷。这些团体包括了女权组织、教会团体、园艺俱乐部等等不一而足。问卷包括127道题目。最后，作者收到了4 500名女性的回复。在这4 500人中，70%结婚五年或五年以上的女性报告曾经有过婚外情，其中95%的女性觉得丈夫对自己已经感觉厌倦了。

A. 该调查得到的样本较大（4 500名女性）。请问这是否就能保证该样本具有代表性？如果不能保证，那么请问是哪些因素降低了样本的代表性？

B. 在对问题A回答的基础上，是否可以认为该调查研究所得的任何结论都是不正确的？如何判断结论是否正确？

2. 有两个不同组织各自对高等教育进行了调查研究，询问教授是否认为学生已经为上大学做好了准备。当两项研究的结果刊登在《高等教育年报》时引起了人们的关注，因为他们所得的结果差异极大。研究基金会A的调查结果发现，几乎75%的教授认为他们的学生“准备非常不充分”。研究基金会B的调查结果则发现，只有18.8%的教授认为他们的学生“没有充分准备好”。通常，不同的调查所得结果确实可能存在差异，但是上述两项研究结果的差异如此之大，不免使人怀疑调查结果的可靠性和可信度。在怀疑之前，让我们先来考察一下两项研究的一些具体细节。（注：该问题基于NCRIPTAL Update的一项报告，Spring 1990，Vol.3. No.1，pp.2-3）

A. 调查对象是谁？基金会A的调查对象是在各种机构中给本科生和研究生授课的10 000名大学教授。其中54.5%的教授对调查作了反馈。基金会B没有对研究型大学中的教授展开调查（占基金会A样本的25%），其所采样本人数2 311（反应率为62%）中大约90%是教基础水平学生的教师。请问基金会A和基金会B所采用的样本特点如何影响了两项研究的结果？

B. 调查了什么？基金会A的调查题目是“我接触到的

本科生在诸如写作与口语交流等方面的基本技能上的准备非常不充分。”被调查的教授可以在“非常同意、比较同意、不确定、比较不同意、非常不同意”中进行选择。基金会B的调查题目是“该课程的学生在所需基本技能的准备上_____”选项为：“没有完全准备好、有一定的准备、准备很充分、准备非常充分”。请问这些问题会对两项调查研究的结果造成什么影响？

C. 结果如何报告？基金会A（调查结果为：75%的教授报告学生的准备非常不充分）在报告调查结果时，将“非常同意”与“比较同意”选项上的数据结果进行了合并。基金会B（调查结果为：18.8%的教授认为学生没有充分准备好）在报告调查结果时，只统计了“没有完全准备好”选项上的结果。你认为如果基金会A仅对“非常同意”选项上的数据进行报告，结果会怎样？

3. 某小型文科学校在教导主任的指导下成立了特别工作小组，旨在考察学生对该校的看法。特别工作小组决定调查学生们对学校宿舍区的规章制度的了解程度以及学生对此的看法。问卷中包含了一些私人问题，如要求学生描述当他们违反了学校规章制度时的感受，或者让他们描述当得知其他同学违反规章制度时的感受。特别小组采取分层随机抽样法从所有住宿和走读的全日制学生注册名单中进行抽取。全校共有全日制学生2 000名，他们从中抽取了400名学生。最后160名学生完成了调查，反应率为40%。其中一项调查发现，超过三分之一的学生认为学校的制度不公平。特别工作小组不知是否该将此结果向教导主任汇报。

A. 请问最初的400名学生的样本能否代表该校所有的2 000名全日制学生？为什么？

B. 该调查研究中存在一个潜在问题，该问题可能导致最终样本无法代表所有2 000名全日制学生。请问是什么问题？

C. 该调查最后的反应率仅为40%，特别工作小组由此认为最终样本无法代表该校全体学生。另外他们认为“超过三分之一的学生认为学校的规章制度不公平”的结果不正确，是一种过高的评估。你同意特

（续）

别工作小组的结论吗？为什么？

D. 特别工作小组在讨论最终报告时，一位成员提出学生的报告未必可信，因此调查结果没有任何用处，根本不应该汇报。如果该工作小组的负责人让你对此进行回应，请问你会说什么？

4. 迈尔斯 (Myers, 2000) 总结了欧洲与北美的众多调查研究，这些研究一次次地表明，与未婚人士相比，已婚人士报告自己更幸福、对生活更满意。因此，婚姻与幸福之间存在稳定的正相关。

A. 请问我们能否根据婚姻与幸福之间的相关关系，确

定两者之间因果关系的方向？为什么？

B. 请思考以下假设：如果是幸福导致人们更愿意结婚，而不是婚姻使人们幸福。那么，由此可得，最幸福的人就会结婚越早、结婚的次数越多；而不太幸福的人则会在年纪比较大时才结婚。那么在此假设下，该调查研究中报告幸福的已婚人士与未婚人士会怎样？

C. 如果该调查中报告幸福的已婚人士与未婚人士不会随着年龄的增长发生改变，那么婚姻与幸福之间的因果关系又该如何理解？

拓展练习的答案

1. 第一位学生认为应该采用分层随机抽样，从“希腊社团”抽取 100 名，从不隶属于任何组织的独立群体中抽取 100 名。在该计划中，各层所抽取的样本大小相同，因此能够代表各自的总体。但是该计划存在一个较严重的问题，即所得的 200 人的样本不能代表总体中希腊社团与独立人群的比例（在总体中，来自希腊社团的人群在总体中的比例为 25%，独立人群在总体中的比例则为 75%）。如果采用该方法会导致偏差样本，因为来自希腊社团的人群过多。第二位学生认为应该采取简单随机抽样，从总体中抽取 100 名学生。这也可能产生偏差样本，因为会导致来自希腊社团的人数过少（我们大致需要 25 名来自该团体的学生），从而无法充分了解他们的观点。
2. 比较恰当的抽样计划是：利用分层随机抽样，但要求希腊社团与独立人群的样本大小与总体中的比例相同。如果抽取 200 名学生样本，那么就应该从独立人群中抽取 150 名学生，从希腊社团中抽取 50 名学生。

挑战性问题 1 的答案

- A. 一般而言，样本越大样本的代表性越高。但在该研究中，虽然最后的样本很大，但是反应率很低，只有最初样本（100 000 人）的 4.5%。由于反应率过低、加上本调查主题的特殊性，导致该调查中存在反应偏向问题。因此我们认为 4 500 名妇女不能代表总体。
- B. 存在反应偏差问题并不能得出该研究不正确的结论。我们不能在此基础上简单地判定结论正确与否。至少有一个好的方法可以判定这个调查结果是否正确。你可以从文献资料中查找有关妇女性态度和性体验的调查报告。当然这些调查必须确保其样本具有代表性。然后你可以比较这些调查结果与此研究的结果，只有他们一致时，才能说明此研究的结果是正确的。当然你也可以开展你自己的调查，前提是避免此研究中已经出现的问题，然后再判断你的研究结果是否与此研究的结果相同。

第6章

行为的隐蔽测量

本章大纲

概述

物理痕迹

基本原理

物理痕迹的类型

问题与局限

档案数据

基本原理

档案数据的类型

内容分析

问题与局限

伦理问题与隐蔽测量

小结

概述

在本书的第4章,我们介绍了观察法,即研究人员如何在自然情境下直接观察和记录人们的行为。在第5章中,我们则阐述了调查研究法,即研究者如何通过直接询问人们的观点与态度,从而了解他们的思想与感受。本章所要介绍的方法与上述两章均不相同,它不要求对人进行直接观察,也不要求向人直接提问,研究者只需考察行为的物理痕迹(如,涂鸦)和档案记录(如,高中年鉴)就能获得有关人们行为的重要信息。这种考察行为的非直接的方法就被称为**隐蔽测量(unobtrusive measures)**或**非反应测量(nonreactive measures)**。在运用该方法时,被考察的个体不会意识到观察者的存在,研究者从物理痕迹或者档案记录中获得信息更为隐蔽,因为研究者在搜集数据时,被考察的个体通常并不在现场。此方法不仅能够得到重要的结论,而且还可用于检验那些来自直接观察和调查研究的结果。因此,隐蔽测量无疑能在那些采用多种方法的研究中发挥重要的作用。

隐蔽测量的一种方式考察人们留下的物理痕迹。例如,假设拉丁美洲人与北美人在“守时”观念上存在差异,想要对此文化差异进行考察可以采用多种方法:可以运用调查研究法直接询问他们迟到的次数,或者询问他们对于音乐会、体育比赛等没有准时开始的感受等。但是,如果想要采用隐蔽测量考察两种文化下人们的时间观念则完全不同。比如,研究者可以观察两种文化下的公共时钟的准确性。事实上,有研究者就曾经发现,巴西一些城市的公共时钟(如,银行的时钟)的准确性就低于美国同等规模城市的公共时钟(Levine, West, & Reis, 1980)。虽然对此现象可以有多种解释,但是,其他物理痕迹的测量结果支持了“守时”观念上存在着文化差异的假设。莱文(Levine, 1990)(在本书第2章曾介绍过其有关生活节奏的研究)也曾对全世界不同城市与国家的公共时钟的准确性进行的考察。莱文的研究发现,城市居民的行走速度(直接观察)与该城市公共时钟的准确性(隐蔽测量)之间存在正相关。

档案记录是指那些能够描述个体、组织、政府与其他群体活动的资料,其中有些是公开的资料,有些则是不公开的。通过考察类似的档案记录就能得到档案数据。例如,研究者对死亡率进行分析后发现,在重要事件发生之前(如,生日、宗教节日)的死亡率会降低,而在这些重要事件结束后,死亡率则会升高(死亡起伏假说)。研究死亡率与重要事件之间关系的步骤大致为:首先研究者在考察档案数据后发现个体的死亡日期与某些重大事件有关(参见:Phillips, Van Voorhees, & Ruth, 1992; Schulz & Bazerman, 1980)。随后,研究者开始对三种重要事件(生日、大选、宗教节日)与死亡率的关系展开调查。例如,在一项有关死亡起伏假说的早期研究中,有人曾经利用《四百位美国名人》(*Four Hundred Notable Americans*, Schulz & Bazerman, 1980)一书中的名单,对名人们的出生日期与死亡日期进行调查。人们也许会认为,死亡具有偶然性,因而在一年的12个月份中,都有可能发生。然而,该研究则发现,在生日之前的几个月的个体死亡率低于其他月份。因此,该研究结果支持了死亡起伏假说。根据舒尔茨和巴泽曼的解释,人们会在重要事件(如:生日)之前延长生命或延迟死亡。

菲利普斯及其同事发现死亡起伏假说更适合于女性 (Phillips et al., 1992)。他们的研究发现, 在生日来临之前, 女性的死亡率会偏低, 而在生日结束后的几周内, 死亡率则升高。与女性不同的是, 男性的死亡率在生日来临之前会有短暂的升高。由此, 研究者得出结论: 生日也许有助于延长女性的寿命, 但是对于男性而言, 过生日可能反而提高了他们的死亡率。除此之外, 菲利普斯及其同事还对死亡率与宗教节日之间的关系进行了考察, 研究结果支持了死亡起伏假说。他们发现, 在犹太教的逾越节前后, 犹太教信徒的死亡率存在较大的波动 (Phillips & King, 1988); 在中国的中秋节前后, 中国人的死亡率也出现了类似的起伏 (Phillips & Smith, 1990)。

虽然上述研究支持了死亡起伏假说, 但导致该现象的原因我们仍然不甚明了。一种解释是: 当人们对此类重要事件有所期待时, 能够产生一种积极的状态, 进而可能会使神经系统中的化学物质产生有益的变化。另一种解释则认为老年人在重要事件临近时, 对自己的健康更为关注, 因此也会更加配合治疗, 更注重养生, 由此也可能降低死亡的机率。

在本章中, 我们将介绍运用物理痕迹和档案数据的基本原理。我们还会列举某些类型的物理痕迹, 并用案例的形式说明如何在心理学研究中加入运用。除此之外, 我们还将介绍几种常见的档案资料及档案数据的不同类型。另外, 本章还会对运用物理痕迹或档案数据的方法中的一些主要优缺点展开论述。

物理痕迹

基本原理

- 在隐蔽测量中 (如对物理痕迹进行考察), 由于人们不知道自己被观察, 因此不存在反应性的问题。
- 采用多种方法检验假设时, 物理痕迹测量是很有价值的一种研究方法。

阅读过侦探小说的人应该知道, 过去行为残留下的物理痕迹能够为侦破案件提供重要线索。例如, 案发现场留下的脚印能够告诉我们此人的身高与年龄, 同时, 脚印之间的距离能够说明此人当时是在行走还是在奔跑。

对于研究者而言, 物理痕迹具有很高的价值, 因为它能够向研究者提供非反应性的行为的测量 (Webb et al., 1981)。反应性行为是指当被试知道自己正在被观察时其行为会发生变化, 在采用直接观察法时经常会出现这种问题; 在采用调查法时, 该问题更为严重。当被试意识到自己的反应正在被记录时, 他们也许就会表现出与平常不一致的行为。因此, 我们应该警惕来自调查研究的数据, 因为被调查者的回答很有可能是为了迎合调查人员的喜好, 或者为了使自己看上去更好而说谎。例如, 社会科学家在对人们的政治活动展开调查, 如调查人们过去的投票行为时, 发现人们会刻意提高自己的形象 (如: Endersby & Towle, 1996)

与直接观察和调查研究相比, 物理痕迹测量是非反应性的。因为这些数据是

间接获得的,收集数据时被试往往不在现场。因此,当通过直接观察和调查研究出现反应性测量时,隐蔽测量就是一种有价值的替代选择。在某些研究中,可能只能对行为进行物理痕迹测量。在第1章我们曾经谈到,研究者可以通过采用多种方法获得大量的研究结果。在心理学研究中,我们推荐大家采用多种方法,因为这样能够减少单一研究中可能存在的问题(Webb et al., 1981)。在社会科学领域没有一种研究方法是完美的。因此,必须对研究结果进行细致的考察(如,反应性测量是否影响研究结果等),即便如此,我们依然可能不经意忽略了研究中存在的某些纰漏。如果某项研究是在结合了多种方法的基础上完成的,那么其所得的研究结果就会更具说服力。

在本章开篇我们曾经提到的,考察不同文化下人们时间观念的研究采用的就是多种方法。研究者通过调查研究方法考察了居住在拉丁美洲与美国的人们,得到了被调查对象迟到的频率以及他们迟到时的态度(Levine et al., 1980)。与居住在美国的居民相比,居住在巴西的被调查者报告的迟到次数更多,而且迟到后更少感到歉意。但是,采用该方法可能存在的问题是:人们也许只是报告了那些符合其所处文化的行为,而不是他们的真实行为。而且,我们之前曾经提到巴西城市中的公共时钟准确性低于美国城市中的时钟。此外,研究者还对美国人与巴西人所戴手表的走时精确性进行了观察。结果发现美国人的手表走时更准确。由此可见,观察两国城市中的公共时钟以及观察两国人所戴手表的走时准确性所得的结果,都证实了调查反应的效度。这些隐蔽测量与调查结果相结合,为该研究假设提供了更有力的证据。

物理痕迹的类型

- 物理痕迹的两种类型是“使用痕迹”与“产品”。
- 使用痕迹反映了使用(或未使用)某物品后的物理证据,可以在自然使用和控制使用两种条件下被测量。
- 通过考察人们拥有的或者某种文化下产生的产品,研究者能够了解人们的态度和偏好。

物理痕迹(physical traces)是指过去行为的残留、碎片和产品。两种主要的物理痕迹是“使用痕迹”与“产品”。顾名思义,使用痕迹是指人们使用(或未使用)某物品后留下的物理证据。人们对时钟的设定就是一种物理使用痕迹,烟灰缸中残留的烟头、书本上的记号等都是使用痕迹。产品是指人类行为的创造物,如建筑物以及其他的人造物品。人类学家经常对古代文化遗留下的产品感兴趣,他们通过考察先民遗留的器皿、绘画以及其他人工制品,往往就能精确地描述几千年以前人们的行为模式。心理学家也会通过考察“物理产品”来描述行为或者检验假设。例如,研究动物行为的心理学家可以通过考察动物的巢穴了解不同物种的行为方式。纽黑根和安塞尔(Newhagen & Ancell, 1995)曾经通过考察居住在华盛顿郊区的黑人与白人对车贴的使用,来研究情绪表达上是否存在种族差异。

使用痕迹 物理使用痕迹可以分为自然和控制(或计划)条件下两种。自然使用

痕迹是指没有调查者干预下的使用痕迹,通常在自然情境中发生。控制条件下的使用痕迹(或计划条件下使用痕迹)是指在研究者一定程度的干预或者操纵下的物理使用痕迹。弗里德曼和威尔森(Friedman & Wilson, 1975)的研究表明了两种使用痕迹间的差异。

研究者想要考察大学生对教科书的使用,他们同时对自然使用痕迹和控制条件下的使用痕迹进行了测量。研究者在学生们的购买教科书之前,在书本中某些邻近的书页间粘上了小的封条。学期结束后,研究者将这些书本回收,记录有多少教科书的封条被撕开了;如果被撕开了,位置是在哪里。由于研究者控制了书中封条的出现的位置,因此这就构成了被控制的使用痕迹。研究者还考察了学生在教科书中划线的频率与特点。由于在教科书中划线是一种自然现象,因此这就构成了对自然使用痕迹的测量。研究结果表明,学生对书中前面的章节阅读得更多,而对后面的章节阅读较少。

表 6.1 列举了如何对所感兴趣的变量进行物理使用痕迹测量。表中所举例子有些在真实研究中被采用过,有些则被认为能对行为进行非反应测量(参见 Webb et al., 1981)。在阅读该表时,请思考其他可用于测量表中所列变量的物理痕迹。

产 品 在心理学研究中,心理学家更多的是对物理使用痕迹进行考察,而对物理产品研究的较少。不过,这种类型的物理痕迹经常以巧妙有趣的方式,被用于

表 6.1 物理使用痕迹及其所测变量举例^{*}

痕 迹	变 量
自然使用痕迹	
公共场所墙壁上的涂鸦	对性的关注度
书上的指纹或污迹	书的使用
垃圾桶中的酒瓶	酒的消费
杂 物	不准乱扔杂物的告示的有效性
垃 圾	食品消耗与生活方式
车子有无上锁	对财产的关注程度
烟蒂的长度	癌症致死率上的文化差异
地板或楼梯的磨损程度	步行的数量
图书馆书籍的磨损程度	使用频率
所消费的食品	饮食行为
由于摩擦导致的雕像或宗教器物上的斑点	宗教信仰的水平
控制条件下的使用痕迹	
书上撕破的封条(在分发之前粘上的)	阅读的页数
博物馆展出时玻璃上的鼻印(每晚会擦拭)	受欢迎程度(频率)、参观者的年龄(及身高)
孩子鞋子的磨损(在两个间隔时间点进行测量)	孩子的活动水平
广告或告示上的“撕开”标签	注意点或兴趣点
在特定区域铺设垫子或地毯后观察其磨损	步行的数量
在雕像或宗教器物上涂抹对触摸敏感的材料	迷信行为

^{*}资料来源: Webb et al. (1981)。

检验行为假设。例如，最近研究者想要了解“法兰西之谜”，于是他们对美国人与法国人消费的食品进行了考察（Rozin, Kabnick, Pete, Fischler, & Shields, 2003）。“法兰西之谜”是指：人们发现法国人中由心脏病导致的死亡率远低于美国人。但事实上，法国人与美国人相比，食用更多高脂肪食物，但美国人比法国人更容易长胖。那么为何会出现“法兰西之谜”呢？有研究认为，这可能是由于法国人喝红酒比美国人多的缘故。还有研究认为，法国人与美国人在新陈代谢上存在先天性的差异，而且法国人的生活压力低于美国人（参见 Rozin et al., 2003）。还有人认为可能是由于法国人吃得较少。为此，罗津恩等人研究了法国人与美国人的食品消费，想藉此找到支持该假说的证据。他们对美国与法国规模相当的餐厅中食物的份量进行了现场观察。此外，他们还考察了两国超市中食品的份量。结果发现，美国餐厅中的食物份量平均比法国餐厅多 25%，美国超市货架上的食品份量也要普遍大于法国超市中的食品。上述研究结果为我们提供了另一个有关“法兰西之谜”的解释，即由于法国人吃得较少，因此他们体重较轻，心脏病的死亡率也较低。

布兰特（Brandt, 1972）指出，人们所拥有的产品能向研究者提供有关其行为方式与生活风格的重要线索。例如，人们在选购不同车型时会反映出人格上的差异。人们所拥有的产品不仅能够用于测量人格或反映社会地位，还可用于检验某些口头报告的真实性的。例如，人们可能在调查中声称自己很注重节约能源，那么他们所拥有和使用的产品有否体现了该态度呢？

让我们思考一下美国人对车贴的兴趣。你可能会看到有的车贴这样写道：“枪不杀人，人杀人！”或者“堕胎，停止一颗跳动的心。”纽黑根和安塞尔（Newhagen & Ancell, 1995）认为，车贴“允许车主向公众表达一些比较极端的个人观点，同时又不需要对此负责”（p.313），在这方面车贴的作用是独一无二的。同时，它又与涂鸦不同，涂鸦一般是被禁止的、不受社会欢迎，但人们可以通过车贴来表达自己的情绪，而这种方式又被社会所接受。纽黑根和安塞尔在考

图 6.1 考察车贴是一种对人们态度的隐蔽测量



察华盛顿特区郊区的白人与黑人使用车贴的频率时发现,低收入群体(无论白人或黑人)使用车贴的频率最高。高收入的白人群体使用的车贴所表达的大多为积极的、强烈的情绪。研究还发现,在高收入的黑人群体中车贴的使用频率最低,即使有,所表达的情绪也异常温和。因此,研究者认为上述结果表明了不同种族与收入的群体在表达情感上存在着不同形式。该结论看上去似乎与安德比和托尔(Endersby & Towle, 1996)有关车贴的研究结论相一致,后者认为车贴反映了“具有相同信念群体的自我认同与团结精神”(p.307)。

问题与局限

- 在断定某物理痕迹的测量能够正确反映人们的行为和态度之前,研究者必须保证所选取的材料不存在偏差。
- 研究者可以通过考察复合证据确定物理痕迹测量的效度。

物理痕迹测量方法为研究者考察人们的行为提供了一种有价值的和新颖的研究途径,而这种测量的有效性则取决于研究者的才智。如果研究者在研究中只采用了该方法,那么在考察和检验物理痕迹的效度时必须格外谨慎。物理痕迹留下的方式以及痕迹随时间的变化可能产生误差。如果某条向右拐的路走得很多,我们能否由此推断这是因为人们对通往该方向的事物感兴趣呢?还是仅说明人们天生喜欢向右拐呢?城市的公共时钟走时不精确就说明当地居民的时间观念差吗?还是说明了当地制作时钟的工艺不够先进?抑或说明了当地对时钟缺乏足够的维护?当地运动队近期的成绩又会对车贴的使用频率产生多大影响呢?

物理痕迹测量究竟存在哪些问题?让我们通过一个经典实验来说明。该实验采用“丢信法”对公众的诚信度进行调查(参见Merriitt & Fowler, 1948)。研究者在美国东部与中西部一些城市的不同地点,假装把一些明信片 and 信封掉在了沿街邮筒外。结果发现其中两种类型的信封“遗失”。一种信封里面是一张字条,另一种信封里面是 50 美分大小的铅片。所有信件都写了明确地址、支付了足够的邮

图 6.2 经典的“丢信法”说明了在分析物理痕迹时可能存在的一些问题



专栏 6.1

互联网上丢失的“邮件”

斯特恩和弗伯 (Stern & Faber, 1997) 曾利用互联网开展了电子邮件版的“丢信法”。在一项研究中, 他们调查了人们对 1996 年总统候选人露丝·佩罗特的态度。研究者从全国的电子邮件地址中随机选择, 假装写错邮件地址, 向所选出的邮箱发出了信件。信件上的收件人是“斯蒂夫”, 信中说明希望“斯蒂夫”为露丝·佩罗特帮忙募集资金。研究者非常好奇人们会作出何种反应。最后所得到的回信表明, 几乎没人支持露丝·佩罗特。如, 有些人在收到该邮件后, 回复说已经给其他总

统候选人投票了。而且结果中没有一个人声称想要给露丝·佩罗特投票。

传统的“丢信法”不存在伦理问题, 但是电子邮件法则引发了伦理上的争议。例如, 把信件不小心掉在路边是现实生活中经常出现的一幕, 而且参与者是匿名的。然而, 在电子邮件版的“丢信法”中, 在参与者的回复中会显示他们的电子邮件地址, 因此构成了非匿名的反应。你认为该如何解决这种方法存在的伦理问题?

资, 但都未写回信地址。85%的内装纸条的信封被人们放进邮筒, 但仅有 54%含铅片的信封放进邮筒, 而且其中超过 10%的在被放回之前曾被打开过。

在研究者得出信件回收率能够有效测量公众诚信度的结论之前, 必须首先考虑采用该物理痕迹进行测量时可能存在的偏差。例如, 与密封的信封相比, 明信片更少被送回。产生这种结果的原因也许是, 明信片与较大的信封相比, 更容易受其他自然条件的影响, 如它们更容易被风刮走等。另一方面, 含铅片的信件比较重, 不太容易被风刮走, 但它们被寄回的几率依然不高。某些区域 (如: 孩子比较集中的区域), 或在某些时刻假装丢失 (如: 夜幕降临之前) 的邮件被送回的几率低很可能是由于其他原因, 而不是公众的诚信度比较低或公众比较冷漠。

研究者在假装把信件掉在路边时采取了一些方法以避免该方法可能存在的某些偏差。在分析最后结果时, 他们注意到了其他可能存在的偏差 (如: 是否被风吹走等)。他们还为一组故意掉在路边的信件样本进行了观察, 结果发现, 信件被捡起时 90%都是被看见该信件的第一个人捡起。这说明这些掉在路边的信件并没被人忽略, 而是很快就到了行人的手中。

该方法可能存在的偏差揭示了一些更普遍的原则。例如, 我们应该尽可能获取一些能够证明某物理痕迹有效的其他证据 (参见 Webb et al., 1981), 必须考虑导致物理痕迹发生改变的其他可能性。在搜集资料时, 也要容许其他假设的存在。在比较不同研究之间的结果时, 必须仔细确认不同研究所采取的测量是否足够相似。例如, 当纽黑根和安塞尔 (1995) 研究车贴时, 他们只关注了汽车保险杠上的车贴。安德比和托尔 (1996) 对车贴的操作性定义则是: 所有能够从汽车后面看到的车贴, 包括汽车后窗玻璃上的车贴。概念的操作性定义不同, 得到的结论也很可能不同。

拓展练习

研究人员想要证明公立高中学生的思想比私立高中学生的思想成熟。为此他选择了两所在同一社区的高中。其中一所为教会高中，另一所为公立高中。两所学校间的距离仅为1英里左右。两所学校都包括9~12年级。他决定观察自然物理痕迹与控制条件下的物理痕迹。他对“思想成熟”的操作性定义为：学生在厕所里说脏话的次数（他事先拟定了一份脏话列表）。自然使用痕迹法评估的是每周末在厕所墙上新出现的脏话。他还在校方的同意下在每所学校的厕所隔间中放置了便笺

簿和笔，用于测量被控制的物理痕迹。每份便笺簿上写着“请给我留言”。每周他会更换便笺簿，并观察上面是否出现脏话。

1. 请评论在该研究中可能影响物理痕迹测量效度的偏差是哪些？
2. 如果让你就该研究的外部效度提问，你会提哪些问题？
3. 如果采用多种方法进行研究，请问研究者还可以考虑使用哪些方法？

档案数据

基本原理

- 档案数据包括个人、团体、协会和政府活动的记录或文件。
- 在多种方法中，档案数据可用于验证假设、为实验结果提供外部效度，并可用于评估自然事件的影响。
- 对于心理学研究而言，档案数据意味着一个充足的、丰富的观察来源。

我们每个人在出生后就会有一份出生证明。出生证明中可能包括了我们出生的城市、州、出生的日期和时间、父母和我们的姓名等。我们死亡后，又会有死亡证明。死亡证明可能包括死亡原因、死亡的日期与时间、死亡时的年龄等。在出生与死亡之间，还会有无数对我们行为所做的记录。例如，医生会记录我们的看病情况，医院会有我们入院和出院的记录；学校会记录我们的年级、课外活动等情况；公司会记录我们迟到或者无故缺勤的次数等。一些行政机构会记录我们结婚的时间、结婚对象，还会记录我们买房的时间、购房费用等。联邦政府还会记录我们的纳税情况等。

不仅个人的活动会被记录，国家、社会团体、城市、企业等的活动也会被记录。例如，国家的国民生产总值、进出口情况、国防预算、人口分布状况、民众拥有电视机的数量等都会被记录。至于企业，则需要向股东汇报利润。除此之外，各州的选举活动、联邦立法、每座城市用于社会福利的经费等均会被记录。此外，各种大小媒体上都充斥了当地事件和全球事件的分析与评论。媒体的各种报道、各种文献资料、所有已出版与未出版的书籍，都向我们提供了当前的流行时尚、各种偏见、信念的变化模式以及社会中坚力量（当然也包括非主流群体）的观念。此外，互联网的出现和普及使各种记录的数量呈几何式增长，使人们可以轻松获得大量记录。

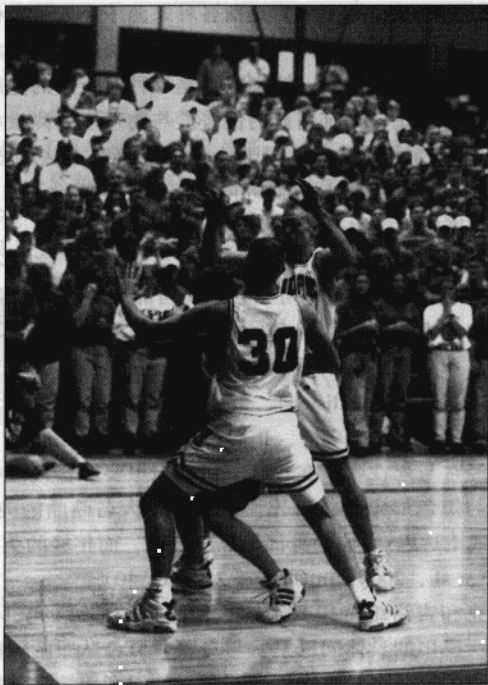
档案数据 (archival data) 是指各种来源于个人、机构、政府以及其他团体的活动记录或文献资料。作为一种行为测量方法的研究对象,档案数据与物理痕迹一样具有某些特殊的优势。档案数据是一种对行为的隐蔽测量,因此与调查研究以及直接观察相比,具有明显的优势。与物理痕迹测量相同,在那些采用多种方法的研究中,档案数据也可用于检查其他研究方法的效度。弗兰克和吉洛维奇 (Frank & Gilovich, 1998) 的研究向我们极好地展示了如何在采用多种方法的研究中将档案数据分析作为一种检验假设的方法。

弗兰克和吉洛维奇 (1998) 对黑色与“邪恶”之间在文化上所存在的关系进行了考察。他们指出,我们可以在老式西部片中轻松识别出好人与坏人:“好人”总是戴着白色帽子,而“坏人”总是戴着黑帽子。黑色与文化间的联系还体现在其他方面,如在很多表示不幸的词语中经常包括“black” (黑) 一词,如说某人被列入了黑名单 (blacklisted)、被勒索了 (blackmailed)、被诽谤了 (blackened)。如果某天发生了可怕的事情,我们会说这是黑色的一天 (black day)。在许多文化中,黑色经常与死亡联系在一起。当我们想要形容人性中的善恶面时,也经常将白色形象与黑色形象进行对比。例如,在大名鼎鼎的美国电影《星球大战》中,邪恶的黑武士达恩·韦德全身都是黑色装束,而年轻的英雄路克则是白色或浅色装束 (Meier et al., 2004, in Chapter 2, p.39-40)。

弗兰克和吉洛维奇想要了解这种黑色与邪恶的文化联系是否影响人们的行为。他们所提出的具体问题是:与身着非黑色队服的球队相比,人们是否认为身着黑色队服的球队更具攻击性。于是,他们利用档案数据法从美国国家橄榄球联盟 (NFL) 以及美国国家冰球联盟 (NHL) 的办公室获得了有关的档案数据。弗兰克和吉洛维奇分析了 1970 年至 1986 年间参加橄榄球赛与冰球赛的职业球队的受罚记录。他们对攻击性的测量如下:在橄榄球比赛中是被罚码数;在冰球比赛中则是犯规者在受罚席中的时间。他们没有采用犯规次数而是选用上述测量标准,因为与简单的犯规相比,上述衡量标准所代表的是更严重的攻击性行为。有关“黑色队服”的操作性定义则为:只有一半以上的颜色为黑色才是黑色队服。(美国橄榄球联盟中芝加哥熊队的队服虽然是深蓝色,但是由于许多人错误地将其回忆成黑色,因此也被放入了黑色队服组)。研究者预测,那些身着黑色队服的球队所得到的处罚更多。结果证实了他们的预测! 队服为黑色的球队,如奥克兰、芝加哥、辛辛那提的橄榄球队、温哥华冰球队遭受的处罚确实要多于身着非黑色队服的球队。因此,该结果表明了黑色确实与人们所认为的攻击性之间存在联系。

那么如何解释该现象的心理过程呢? 弗兰克和吉洛维奇认为可以用社会知觉与自我知觉等概念来进行解释。从社会知觉角度来看,其他人 (尤其是裁判员) 看到身着黑色队服的球员时会认为他们更具攻击性,因此对他们的处罚就会更多。当球员穿上黑色队服后,自我知觉则会使他们表现得更具攻击性。为了能够进一步证实他们的假设,弗兰克和吉洛维奇决定采用多种方法进行验证。他们在档案数据的基础上,开展了两项实验室实验。实验一要求大学生和裁判员观看一场橄榄球晋级赛 (两个球队分别身着黑色或白色队服), 然后让他们对球队的攻击性进行评分。在实验二中,研究者先让大学生假想自己要参加比赛了,要求他们选择

图 6.3 档案数据使心理学家能够研究体育比赛中的“主场优势”



自己所要穿的队服颜色（黑或白），然后再让他们从一份球队列表中选择一支球队（列表中包括了攻击性较强和攻击性较弱的球队）。实验一的结果表明，裁判员和大学生更倾向于认为身着黑色队服的球队更具攻击性。实验二则表明，假想自己身着黑色队服的学生更倾向于选择攻击性较强的球队。弗兰克和吉洛维奇在整合了档案分析与实验结果后，为这种有趣的文化现象提供了强有力的解释。

档案数据还可用于检验实验结果的外部效度。研究者已经用实验法对归因（即人们对行为结果的解释）进行了考察。有关归因的实验研究的主要发现是：人们对自己的成败有不同形式的归因。对于自己的成功，人们倾向于进行内部归因，即认为成功得益于自己本身的某些特点；对于失败人们则倾向于进行外部归因，认为失败是由不可控的外在因素造成的。劳和拉塞尔（Lau & Russell, 1980）想要检验这些来自实验室的结果能否解释真实世界。于是，他们对八份日报中的体育专栏进行了分析。他们分析了体育新闻记者或队员对于棒球和橄榄球比赛结果的归因，发现其归因规律与实验室研究的结果相一致。具体来说，劳和拉塞尔对 33 场重要体育比赛的 594 种解释进行了分析。结果发现，胜利球队在进行归因时，75% 为内部归因；而失败的球队在进行归因时，仅有 55% 为内部归因。

研究者还可利用档案数据实现另外两个研究目的——检验过去行为的假设，或者评估自然发生的事件对社会或个体的影响。例如，在研究体育比赛中的“主场优势”时，档案数据就能发挥重要的作用（参见，例如 Courneya & Carron, 1992）。在整个赛季中，球队有时需要在主场进行比赛，有时必须在客场比赛。那

么主场或者客场会如何影响球队的表现呢？主场优势是指当球员在家乡比赛时，往往比在异地更容易赢得比赛。鲍迈斯特（Baumeister, 1995）曾发现了一个有趣的现象，该现象也许与我们的直觉认识相悖。他发现即使球队在家乡更容易赢得比赛，但是他们在关键性比赛（如：争夺冠军赛）中并不一定能发挥出良好的水平。“主场失利”（home choke）现象也许并不能支持上述档案数据所得的结果（Schlenker, Phillips, Boniecki, & Schlenker, 1995a, 1995b）。

档案数据除了用于检验过去行为的假设之外，还可用于评估一些自然发生的事件对社会或者社会中的特定个人造成的重大影响。由于我们无法预料这些事件何时会发生，因此研究者必须运用多种行为测量的方法评估事件的影响，这其中就包括档案数据的分析。恐怖主义活动、股市的剧烈变化、新法案的通过等就是此类事件，它们可能会对人们的行为造成重大影响。此外，亲人亡故、父母离婚、或者婚姻危机等虽然不对社会构成重大影响，却会对身处其中的个人造成重大影响。如果研究者想要考察类似上述事件的影响，那么就可以考虑采用档案数据。

1921年，刘易斯·特曼（Lewis Terman）开始了一项有关智力的纵向研究，他从旧金山、洛杉矶的学校中招募了1500多名智商较高的孩子（参见Terman & Oden, 1947），其中包括856名男孩和672名女孩，当时他们都处于青春期以前。从1921年开始，特曼及其同事每隔5~10年会调查他们一次。弗里德曼等人（Friedman et al., 1995）利用了特曼在此项研究中的档案数据，调查了“心理与行为对长寿的预测力”。弗里德曼及其同事还访问了那些参加过特曼研究的依然健在的被试。

弗里德曼等人（1995）运用来自特曼研究中的档案数据调查了父母离婚对被试寿命的影响。特曼研究中的被试都是一些智商较高的孩子（IQ超过135），几乎都是白人，大多来自中产阶级家庭。心理学研究中大量的证据表明，与父母没有离婚的孩子相比，那些经历过父母离婚或父母冲突的孩子存在更多的行为问题、更多的学习困难，以及身体更不健康（参见Tucker et al., 1997）。弗里德曼等人（1995）则想要研究父母离婚是否也会影响被试的死亡率。研究结果表明，父母离婚与子女成年后过早死亡存在相关。他们发现，那些来自父母离婚家庭的男性死亡年龄预计在76岁，而来自完整家庭的男性死亡年龄预计在80岁；来自父母离婚家庭的女性，她们的死亡年龄预计在82岁，来自完整家庭的则是86岁。毋庸置疑的是，父母离婚与孩子的寿命间的关系是非常复杂的，我们现在还无法对此有全面的了解。例如，研究发现，来自离婚家庭的孩子日后也更容易离婚；在他们长大以后，男性更容易死于身体伤害，女性则抽烟较多（Tucker et al., 1997）。上述档案数据的分析表明，“父母离婚会引发一系列的消极事件，会对孩子的健康造成影响”（Tucker et al., 1997, p.389）。

在研究中考虑采用档案数据的另一个理由很简单，因为档案数据往往非常丰富。我们在此之前已经指出，社会记录了大量有关个人、团体、机构的信息。我们曾谈到，现在研究者可以利用互联网更轻松得到档案数据。通过仔细分析档案数据，研究者还可以获得大量的假说。除此之外，档案数据在实际操作上也有其优势。例如，研究者不需要花费大量时间收集数据，因为很多数据唾手可得。档案数据中甚至已经包括了初步的概要总结。在运用档案数据进行研究时，也不

太需要考虑伦理问题,因为档案数据通常是公共信息的一部分,这些信息在报告时往往不涉及特定的个人。我们撰写本章的目的之一,就是想提醒读者在开展心理学研究时,千万别忽视了档案数据这一丰富的研究资源。

档案数据的类型

- 档案数据的类型包括:连续型记录、新闻媒体以及其他类型的记录。档案数据包括公共档案与私人档案。

由于档案资源异常丰富和多样,因此想要对此进行分类不免有些武断。在档案数据中,有一些数据经常会被不断更新,通常我们将此类记录称为连续型记录。税收记录、各种政府部门的记录、有关体育赛事的大量记录都属于此种类型。由于此类数据能得到连续不断的更新,因此它们尤其适合纵向研究,也适合有关趋势分析类的研究。其他数据,例如个人档案,则是不连续的数据,它们通常是由一系列片断构成(见 Webb et al., 1981)。我们还可根据档案数据对公众的开放性程度进行划分。政府部分的大量记录是很容易获得的,然而,大多数社会团体或者企业的档案则不向公众开放。因此,对于研究者而言,想要获得此方面的信息就必须极富耐心,并须付出大量的努力。

新闻媒体是档案数据的另一种重要来源。报纸、电视和网络会报道各种信息,从股市行情到犯罪统计。媒体的各种报告也是档案记录的一种形式,也可用于研究分析。在本章中我们曾经谈到有人利用报纸中的体育专栏检验归因理论。同样,也有人利用报纸中失物招领专栏中的信息考察了公众的利他行为(参见 Goldstein, Minkin, Minkin, & Baer, 1978)。研究“法兰西之谜”的调查人员则考察了在美国费城和法国巴黎的餐馆指南,统计了其中提供自助餐的餐馆数量。调查发现,在巴黎没有餐馆提供自助餐,在费城则有 18 个餐厅提供自助餐(Rozin et al., 2003)。

菲利普斯(Phillips, 1977)曾利用加州高速公路巡逻队的车祸记录以及加州报纸中公布的自杀数据考察了车祸与自杀之间的关系。他的假设是:让自己在车祸中丧生是人们寻求自杀的一种方式,因此高速公路上的许多车祸由此引起。为了检验假设,菲利普斯考察了在报纸发布的自杀故事公诸于众后,车祸是否增加。那么如何判断该自杀故事是否引起民众关注呢?他采用了报纸的每日发行量以及自杀报道后的天数作为衡量标准。结果发现,车祸发生次数与每次自杀故事的报道之间存在显著相关。在自杀故事引起民众关注后的第三天,车祸发生率达到了高峰(参见图 6.4)。为了能够进一步证实结果,菲利普斯还进行了其他分析。他比较了实验期间(自杀故事报道的一周内)和控制期间(没有报道自杀故事)车祸发生频率的变化情况。他让实验期间和控制期间在时间上完全匹配,包括两者在星期几上必须一致、在这段期间有节假日也要一致。最后菲利普斯得出结论:“报纸上的自杀故事会激发起人们模仿自杀的浪潮,其中有些人会通过制造车祸假象进行自杀”(p.1464)。

表 6.2 罗列了档案数据的来源以及这些数据的性质。在表中,我们将档案数据

图 6.4 自杀故事报道前、报道时(第0天)、报道后的每日车祸死亡率在两周内的波动

(资料来源: Phillips, 1977)

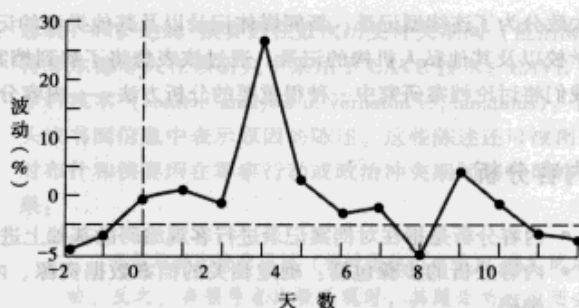


表 6.2 档案信息的可能来源及相应数据所能解释的内容*

来 源	解释的内容
连续型记录	
国会议事录	对特定问题的观点
电话号码簿	社区民族团体身份
教师或公务员的薪金	社区支持情况
政府部门的记录(劳工部、商务部、农业部)	生活趋势
法院记录	反社会行为判决的公正性
Moody 手册	公司的财务结构
美国名人录	成功人士所获成就的特点
税收记录	各地区生活模式的差异
城市预算报告	对各种活动的支持状况
媒 体	
大城市报纸上的社会专栏	中产阶级上层以及上层社会中的部分人士的活动
书店中的儿童书籍	榜样的品质(男英雄和女英雄)
报纸中的电影告示	禁忌与诱惑的变迁
招聘广告	雇主的动机
讣告专栏	慈善倾向
公开演说	政治、社会、经济态度
报纸头条新闻	媒体偏向
其他记录*	
缺勤和迟到的记录	工作习惯或积极性
延长服役期限	士气指标
产量大小	员工、部门等的表现
销售记录	销售效果、激励计划的有效性
球类比赛中花生的销售情况	观众兴奋性的指标(第7局后花生的销售额会上升)
航空保险数据	公众在空难前后的态度
不同颜色婴儿用品的销售情况(蓝色或粉红色)	不同社会阶层对不同性别婴儿的偏好情况
保险统计记录: 出生、受洗、死亡记录、结婚证	人口统计学数据的比较(职业、宗教、出生日期、死因等)
墓地文獻、土葬墓地的记录	家庭成员

*资料来源: Brandt (1972)。

*机构、企业、医院、学校等可能不会向公众开放。

大致分为了连续型记录、新闻媒体记录以及其他类型的记录,其中包括来自企业、学校以及其他私人机构的记录。通过该表您将了解到档案数据的丰富性。接下来我们将讨论档案研究中一种很重要的分析方法——内容分析法。

内容分析

- 内容分析是指在对档案记录进行客观编码的基础上进行推断的过程。
- 内容分析的步骤包括:确定相关的档案数据资源、内容抽样以及对内容进行编码。
- 内容分析中的编码包括将档案记录中的事件和行为进行分类,将它们转化成具有明确定义类目,并记录事件和行为出现的数量或时间等。
- 训练有素的编码人员会利用等级量表对档案记录的内容做出定量判断。

确定档案数据的资源固然重要,但是这些资源最终的价值则取决于对其内容的分析。在某些简单情况下,也许只需对档案资源作极少的分析就足够了。例如,记录各州立法机关成员的得票数也许简单到只需将立法机关文件中的投票数或将政府网站上的相关数据抄写下来即可。然而,在多数情况下,从档案资源中收集相关数据可能需要研究者遵循详细的步骤,并且需要研究人员对资源的内容进行复杂的分析。此外,分析档案资源就如同分析来自直接观察的数据,必须关注抽样、信度、效度等关键问题。

研究者在对信息进行客观分析后,就能确定信息的某些特征,并以此为基础做出推断,这种技术被称为内容分析(content analysis)(Holsti, 1969)。内容分析主要针对书面材料,但它也可用于分析各种形式的信息,包括电视广播节目、演讲、电影、访谈、电子邮件信息等。例如,韦格尔、卢米斯和索佳(Weigel, Loomis, & Soja, 1980)曾经运用内容分析考察了电视黄金时段的节目中中对种族关系的描述。他们分析了三家电视台的晚间电视节目中黑人与白人出场的频率。此外,他们还对节目中黑人与白人的互动,以及白人间的互动进行了评估。

许多内容分析研究都是对书面材料进行考察。对材料的定量分析通常包括考察单词、句子、段落、主题等(Holsti, 1969)。例如,有研究者对一些妻子死于艾滋病的男性进行了访谈,随后用内容分析法对访谈记录进行了分析(Pennebaker, Mayne, & Francis, 1997)。如果访谈中这些男性使用一些表示原因和领悟的字眼,那就预示着他一年之后就能走出丧妻之痛。如果想对报纸上的内容进行分析,那么经常采用的方法就是测量“空间”——如讨论某问题的栏目的版面数量。菲利普斯(1977)就曾运用该方法研究了报纸上报道的自杀故事与车祸频率之间的关系。如果想对电视或广播节目的内容进行分析,人们则经常用“时间”作为测量的单位。例如,韦格尔等人(Weigel et al., 1980)曾根据出场时间百分比考察了电视节目中黑人的出场情况,以及黑人和白人在电视节目中共同出场的情况。

此外,定性分析也是内容分析的一部分。例如,萨特菲尔德和塞利格曼(Satterfield & Seligman, 1994)曾经运用定性分析考察了乔治·布什(第41任美国

总统)和萨达姆·侯赛因在数次历史冲突期间(包括海湾危机)中的解释风格。萨特菲尔德等人在该研究中采用了CAVE技术。CAVE技术的全称为逐字解释内容分析技术(content analysis of verbatim explanations)。该技术能帮助研究者确定口头或书面信息中表示原因的陈述。这些陈述还可被用来评估其解释风格。这项针对布什和侯赛因在军事行动或政治冲突期间解释风格的研究得到了一个有趣的结果:

当领导者比较悲观时(相对于他们自己),他们随后的行为会表现得更加谨慎和被动。反之,当领导者比较乐观时,其随后的行为则会表现得更具攻击性和冒险性(Satterfield & Seligman, 1994, p.79)。

萨特菲尔德和塞利格曼意识到他们所做的是相关研究,这意味着“在断定解释风格导致谨慎还是冒险行为之前应该注意寻找第三个变量,因此无法简单推断两者间存在直接的因果关系”(p.79)。例如,还需进一步考虑解释风格中的政治要求或者修辞上的文化差异等因素。尽管如此,研究数据表明:“解释风格的改变看上去至少预示着相应的攻击与冒险行为的改变”(p.80)。档案研究有助于我们更好地理解政治决策中的心理过程,甚至可能有助于我们预测重大事件的进展。

我们知道,在对行为进行直接观察时,可以通过向观察者隐瞒研究目的,从而降低观察者误差。同样,在对档案数据进行内容分析时,也可以向那些负责对档案进行编码或评定的人员隐瞒研究的重要内容,如不向他们透露研究的主要假设、档案来源等。例如,在萨特菲尔德和塞利格曼关于解释风格的研究中,只有在“消除了可能存在的所有偏差”(p.78)以及资料来源被随机分配数字编码后,才对领导人的阐述进行评定。在内容分析中,还须报告评估者内部信度。(有关观察者内部信度的内容详见本书第4章)。

问题与局限

- 选择性保留和样本残存会造成档案数据的偏差。
- 选择性保留是指在对档案数据进行记录时产生的内容和记录方式上的偏差。样本残存是指记录数据丢失或者不完整的现象。
- 在保存过程中,档案记录经常出现误差和变化。
- 尽管档案数据是对非反应性行为进行的测量,但反应性仍可能对档案记录尤其是公共记录的产生过程造成影响。
- 研究者在对档案数据进行相关分析时,应该警惕存在假相关的可能性。

在讨论物理痕迹的效度时,我们曾谈到在物理痕迹的获得方式以及它们随时间的变化中可能存在偏差。由这些原因所导致的偏差分别称为选择性保留和样本残存,它们在档案记录中也存在(参见 Webb et al., 1981)。上述偏差可能会对研究结果的推广性造成严重影响。

在获得档案资源时假如存在偏差就会出现选择性保留(selective deposit)问题。让我们来思考一下自杀笔记中可能存在的选择性保留问题。奥斯特德和沃克

专栏 6.2

内容分析的三个基本步骤

内容分析有三个基本步骤：确定相关的数据资源、从资源中进行抽样、为定性和定量分析进行编码。

确定相关资源。确定相关档案资源的关键是界定何为“相关”资源。什么是相关的资源？相关资源当然应该与研究的目标、研究者的假设有关。有时只须明确研究目标就能确定哪些档案资源为相关资源。例如，某位研究者如果想研究碑文上的幽默感，或者研究电视节目中的种族关系，那么研究初衷就已经说明了相关的资源是什么。在其他情况下，对相关资源的界定则取决于研究者的独创性。例如，劳和拉塞尔（Lau & Russell, 1980）运用报纸中的体育专栏检验了前人有关归因的实验结果的外部效度。该相关资源的确定就体现了研究者的独具匠心。同样，卡罗尔和拉塞尔（Carroll & Russell, 1997）为了考察电影演员对情感的演绎，将几部好莱坞电影[如电影《凡夫俗子》（1980）和《死亡诗社》（1989）]作为他们内容分析的资源，因为“这些电影对现代生活进行了真实刻画，并且充满了真情实感”（p.167）。

从档案资源中进行抽样。内容分析的第二个步骤就是从相关档案资源中进行适当的抽样。在观察法或调查研究中也涉及抽样问题，抽样的目的是为了获得具有代表性的样本。档案研究所得结果的推广性取决于所选用样本的代表性。例如，在本章前半部分我们曾谈到特曼对智商较高孩子的研究。在该研究中，研究对象的特点

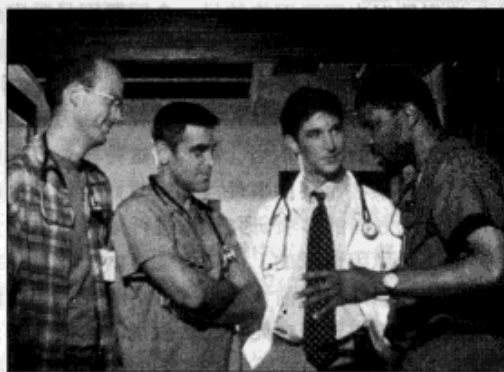
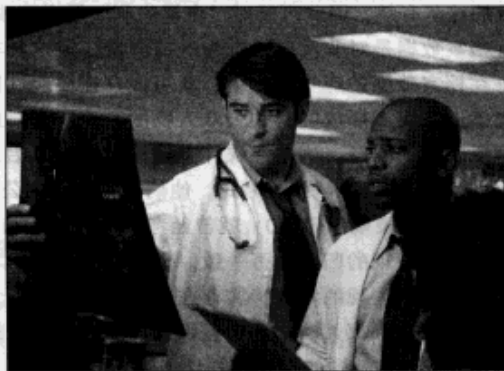
是高智商、主要是白人、来自中产阶级家庭。在随后有关寿命的档案研究中也选用了同样的样本。由于选取的样本的同质性较高，因此，在将研究结果进行推广时必须注意其局限性。此外，我们可能还会对有关寿命的档案研究存有另一疑问：该研究的样本来自20世纪初特曼研究中的调查对象。20世纪20年代，社会中的离婚率较低，而且离婚不被社会认可，因此与现在相比，那个年代父母离婚对孩子健康造成的影响可能与现在存在较大差异。

分析内容的编码。内容分析中的第三步，也是最后一步是编码。我们曾在第4章谈到描述性记录的记分，该步骤也同样需要确定相关的描述种类，以及相应的分析编码（参见 Holsti, 1969）。如何确定相应的描述种类与研究目的有关。韦格尔等人（1980）有关种族关系的研究说明了如何将编码中的类型与研究目的相联系。他们确定了四种类型：（1）人们出场的时间；（2）黑人出场的时间；（3）黑人和白人同时出场的时间；（4）黑人与白人互动的的时间。每一种类型都有明确的操作性定义。例如，“黑人与白人互动的的时间”是指“黑人与白人在屏幕上积极的互动（互动包括：对话、身体接触或者其他明确的非言语交流）”（p.886）。只有对描述种类进行精确的定义，编码者才能对事件做出可靠的判断。

（Osgood & Walker, 1959）曾经对真实的自杀笔记、普通信件和伪造的自杀笔记进行比较。如果你认为这些笔记中透露的思维和情感代表了所有自杀者，那么你就犯了选择性保留的偏差。因为在所有自杀的人群中，只有不到四分之一的人会在自杀前留下笔记，而大多数人并未留下只字片言。因此，很可能情况是，那些留下自杀笔记的人不能代表所有自杀的人（Webb et al., 1981）。

如果在完成最终的档案记录之前，相关人员有机会对档案进行编辑或者更改，同样可能产生选择性保留。虽然美国的国会议事录记录了国会在发表公开演说之前的言论（Webb et al., 1981），甚至是一些未公开的言论，但是在公开发表之前，国会仍可能对此进行修改。因此，研究者在分析档案数据时，必须充分意识到在该过程中可能产生的偏差。研究者必须思考哪些类型的偏差可能会导致诸如在犯罪率统计、收入支出账目以及销售数据上的选择性保留问题。

图 6.5 对不同种族电视演员之间互动的内容分析揭示了美国媒体描绘的种族关系



白人更倾向于攻击黑人，黑人之间亦存在种族关系（例如，Johnson & Johnson, 1991）。第三个变量，例如，不利经济条件造成的经济衰退，是讨论和解释暴力行为的一个因素。由于失业率的增加，许多研究者对福利制度与暴力行为之间的关系进行了研究（例如，Johnson, 1991）。

在得出最终结论时，研究者必须意识到不能简单地通过数据来确定和寻找变量间是否存在更多的证据，并对档案数据的质量和单一档案数据的局限性，档案数据更适合为研究提供背景信息，而不是提供因果关系的证据。研究者必须意识到，档案数据更适合为研究提供背景信息，而不是提供因果关系的证据。研究者必须意识到，档案数据更适合为研究提供背景信息，而不是提供因果关系的证据。

研究者必须意识到，档案数据更适合为研究提供背景信息，而不是提供因果关系的证据。研究者必须意识到，档案数据更适合为研究提供背景信息，而不是提供因果关系的证据。研究者必须意识到，档案数据更适合为研究提供背景信息，而不是提供因果关系的证据。

如果档案数据丢失或者不完整，就可能产生样本残存（selective survival）偏差，其中有些可能研究者意识到了，有些甚至根本就没有意识到。因此，在对某些记录进行选择时，研究者必须仔细考虑其中是否存在系统偏差。例如，是否丢失了一些对某个体或群体而言至关重要的内容？某些类别的数据是否受到损毁或者被放错地方？那些“致编辑的来信”能否代表所有来信？舍尼曼和鲁巴诺韦茨（Schoeneman & Rubanowitz, 1985）提醒人们在分析报刊杂志上的读者建议专栏时

应该谨慎,因为编辑只选登了读者来稿中的很小一部分。正如一些研究者所言,人们在考察档案数据时必须意识到“档案出于某些原因、为某些人而建,虽然承认这一点让人苦恼,但这就是事实”(Webb et al., 1981, p.140)。

除了上述可能存在的偏差外,研究者还面临其他问题。如档案保管过程中可能存在的失误、档案随时间可能出现的变化等。研究者必须证明在档案保管的过程中,档案是相对稳定不变的。如果研究的是连续记录,那么可能不得不中途更改档案类型的定义。有关此方面的一个有趣例子是研究者曾经围绕“主场失利”假说所进行的争论。有些研究者认为,有关球队在主场参加关键比赛的档案数据不支持该假说(Schlenker et al., 1995a, 1995b)。还有研究者则认为对档案数据的分析表明该假说成立。鲍迈斯特(Baumeister, 1995)指出,在档案收集阶段,世界职业棒球锦标赛中一项有关指定击球员的规则发生了改变。该规则改变后将主场作战的球队有益,因此可能对“主场失利”假说造成了影响。

尽管档案数据是对非反应性行为进行测量,但这种非反应性仅适用于分析档案记录。在档案数据的产生过程中,就可能存在反应性问题。例如,在各大报刊或其他媒体上刊登和播放的有关公众人物的言论,在对此类档案进行评估时必须考虑到反应性的问题。这些在媒体上频频曝光的政治人物或者其他人物都非常清楚如何将媒体“为我所用”。因此,他们在媒体上的公众形象很可能与其日常生活不相吻合。劳和拉塞尔(1980)就曾经考察过运动员与教练员在媒体上与私底下所说的有关球队表现的归因是否一致。他们认为运动员与教练员在公众场合与私人场合下有关球队表现的归因确实存在差异,但不能因为这些差异就认为所有来自实验室的研究结果都无法描述真实世界。控制档案数据中的反应效应的惟一方法就是充分意识到该问题的存在,并且,如果可能的话,寻找其他形式的证据。

在分析档案数据中需要面临的另一个问题则是确定是否存在假相关。如果研究者在分析证据时没有认识到两个或多个变量间的相关关系都是由第三个变量导致的,那么这就是假相关(参见本书第5章)。产生该问题的原因是由于采用了不恰当或不正确的统计分析。例如,有些研究者认为死亡率与重要事件之间存在相关。如果研究者采用的统计不当,那么,这种相关可能就是假相关。许多人可能认为个体为了经历某些重要的事件(如生日)会延长自己的生命。舒尔茨和巴泽曼(1980)在对此方面的多项研究数据进行分析后认为,两者间存在相关的原因可能是由其他统计原因造成的,例如可能是由于测量死亡前后的方式造成的。由此可见,舒尔茨和巴泽曼对前人数据的分析将死亡起伏假说的证据都否定了。但是,他们指出我们不应该否定死亡起伏假说。他们认为,“如果我们采用恰当的方法并谨慎的选择重要事件的话,依然有可能证明该假说”(p.261)。

当变量之间恰好存在联系时,最容易产生假相关的问题。在第5章我们曾经谈到,两个变量之间的相关很有可能是由第三个变量引起的,而人们往往对该变量一无所知。假设某研究发现冰激凌的销量与犯罪率存在正相关。在我们就此结果下定论之前,应该考虑两者间是否存在第三个变量,即是否是由于温度升高从而提高了冰激凌的销量,同时也使犯罪率上升了。至于冰激凌的销量与犯罪率上升间也许根本不存在任何关系,之所以两者间会存在正相关,可能是由于温度对它们的共同作用所导致。

菲利普斯(1983)的研究发现,美国的重量级职业拳击赛与杀人罪行之间存在相关。有些研究者于是试图确定两者之间是否为假相关(例如,Baron & Reiss, 1985)。针对菲利普斯的研究结果的质疑在于是否存在第三个变量。例如,菲利普斯在其研究中提到无法解释拳击赛后第三天杀人罪行出现峰值的现象,巴伦和赖斯认为这可能是由于一般赛事都安排在周末或假期,或是由于失业率的波动导致。虽然人们提出了上述可能存在的第三个变量,但是在研究者们对菲利普斯的原始数据进行了大量的重新分析后,却发现媒体上的暴力报道与人们模仿暴力之间的确存在联系(参见 Miller, Heath, Molcan, & Dugoni, 1991)。

由此可见,研究者在试图对单独来自档案研究的数据下最终结论时,应该采取谨慎的态度(例如可以参见 Baumeister, 1995)。研究者必须意识到不能简单地通过变量间的相关确定两者间的直接关系。由上述例子可见,确定和寻找变量间是否存在假相关并非易事。遇到此种情况,研究者可以考虑收集更多的证据,并对档案数据进行多种类型的分析。然而,即便研究者这样做了,依然很难从单一的档案数据中得出确切的结论。在采用多种方法对某现象进行考察时,档案数据更适合为研究提供补充证据。

伦理问题与隐蔽测量

- 隐蔽测量中的物理痕迹以及档案数据是非常重要的研究方法,因为心理学家可以运用它们实现他们的伦理责任,进而改善个人和社会的现状。

美国心理学会公布的《APA 伦理准则》中清楚地说明了“心理学家应致力于提高科学的精确性、真实性、诚实度,”“应关注其他心理学家在进行研究中有否遵循道德准则”,并“应致力于增加有关人类行为的科学知识,提高人们对自己及他人的理解,并将这些知识用于改善个体、组织和社会的状况。”(参见 Preamble, Principles B and C, of the APA Ethics Code, American Psychological Association, 2002, p.1062)。上述准则当然适用于所有心理学家,无论其采用何种研究方法。然而,我们在本章中涉及的一些研究采用的是隐蔽测量的方法。例如,我们曾经阐述了一些有助于我们理解诸如自杀、种族间的关系、父母离婚对孩子所造成的影响、国际间的争端、社会暴力问题、性别歧视等问题的研究。这些研究“增加了我们对行为的认识,并且有助于提高人们之间的相互理解”;此外,它们还可能“改善个人、组织和社会的状况。”这都归功于那些采用了物理痕迹或档案数据的心理学家们的研究。这些都说明了心理学家希望通过其研究改善个人与社会现状。

如果一个社会忽视研究的价值,其代价可能是惨重的(参见第3章)。如果我们通过研究能够对一些社会问题(诸如自杀、暴力、种族关系)有更好的了解,那么无疑将有助于改善很多人的生活。然而,心理学家在尝试对类似问题进行研究时,可能需要考虑风险收益比,因为某些研究会存在较为严重的伦理问题。要解决该问题可以考虑寻找其他风险较低的数据收集方法。例如,在试图考察某些

重要的心理学问题时,与其他更具干预性方法相比,隐蔽测量中的物理痕迹法和档案数据法可能会使研究面临最少的伦理问题。由此可见,如果研究者想要以最少的伦理风险考察重要的社会问题,那么隐蔽测量法无疑为其提供了福音。

小 结

隐蔽测量(如物理痕迹法与档案数据法)是对直接观察与调查研究的一种重要补充。物理痕迹是指过去行为的残留、碎片和产品。调查者可以没有任何干预地考察自然情境中人们的使用痕迹,也可以对其进行一定的操纵。物理痕迹法是对行为的非反应性的测量,因此可以将物理痕迹法单独使用,也可以将它与其他研究方法结合使用。在考察人们的行为时,多种方法备受推荐,因为这将最大限度降低研究过程中存在的某些偏差。在获得物理痕迹时,研究者必须充分意识到痕迹的收集过程中以及痕迹随时间推移时可能存在的偏差。

档案数据是指各种来源于个人、组织、政府以及其他群体的活动记录或文献资料。档案数据具有非常重要的价值,因为它们有助于我们考察那些来自于实验室的研究结果的外部效度,有助于评估“自然处理”下发生的事件(如:离婚)的影响,还有助于我们分析交流的内容。档案数据也是一种对行为的非反应性测量,可在多种方法中用于验证假说。分析档案数据特别要求某些形式的内容分析。与第4章中描述性记录的分析不同,档案数据中的内容分析包括从记录中抽样及对内容进行编码。在运用档案数据时,必须考察是否存在选择性保留和样本残存偏差,并且在研究中必须说明所观察到的变量间的相关不是假相关。

与更多的干预性方法相比,采用隐蔽测量可使研究面临最少的伦理问题。因此,在采用多种方法考察重要的社会问题时,研究者可以考虑将隐蔽测量作为其中的一种方法。

重要概念

隐蔽测量
物理痕迹
档案数据

内容分析
选择性保留
样本残存

复习思考题

1. 为什么在考察人们的行为或态度时,物理痕迹和档案数据尤其适合于作为直接观察和调查研究的补充方法?
2. 试解释为什么研究者应该采取多种方法来研究一个问题。
3. 说明心理学家经常采用的两种物理痕迹方法,以及两种方法间的差异。
4. 说明物理痕迹法中可能存在的偏差,以及如何确定测量的效度。

5. 研究者采用档案数据进行研究有哪三个理由？请分别举例说明。
6. 区分各种档案来源的两个维度是什么？
7. 简述档案研究中内容分析的三个步骤。
8. 说明在对使用从档案资源中获得的研究结果进行解释时，选择性保留、样本残存、反应性行为的作用。
9. 试述档案数据的分析中可能存在假相关的两个原因。

挑战性问题

1. 假设你需要对以下每一个变量进行物理痕迹和档案数据的测量，请给以下每个变量设定操作性定义。
 - A. 公众对文化事件的兴趣
 - B. 学生的政治态度
 - C. 社区对犯罪行为的关注
 - D. 居民对学校的态度
 - E. 流行服饰
2. 以下为档案来源，确定两种可用于心理学研究的变量，并给出所选变量的操作定义。
 - A. 新闻周刊
 - B. 肥皂剧
 - C. 报纸分类广告
 - D. 城市预算报告
 - E. 学生年鉴
 - F. 大学捐赠人名单
 - G. 电视广告
3. 假设有一位心理系的女研究生同时得到了在《新闻周刊》和《时代》周刊的工作机会。两家公司给出的薪金、工作条件和工作要求都相仿。为了决定该去哪家公司工作，她想了解哪家杂志更尊重女性。因此，她决定用内容分析法对这两家杂志展开研究。如果她向你咨询，那么在以下内容分析的各个步骤中你会向她提出哪些建议？
 - A. 抽样
 - B. 编码
 - C. 信度
 - D. 定性和定量测量
4. 某位教育专家确信在校园斗殴中男孩比女孩更暴力。为了收集该假设的证据，她考察了自己学校的校医院的记录。她发现，在100例由斗殴所导致的受伤中，75%都是男孩，仅有25%是女孩。由此，她认为自己的假设是正确的。如果真实情况恰好相反——校园斗殴中女孩比男孩更暴力，那么请仔细思考该专家的结论。（请注意该专家推导结论中所依据的数据是否充分。）

拓展练习的答案

1. 如果两所学校的学生在对待写脏话的行为方面存在差异，那么就可能影响物理痕迹测量的效度。由于调查者在该研究中是每周做一次记录，因此学校中的保安人员或者其他学生和教工（并不了解该研究目的）可能会擦掉厕所墙上的脏话，还可能拿走厕所隔间里的便笺簿。这种可能出现的偏差尤其会增加两校间进行比较的难度。
2. 有关该研究的外部效度可以提出的问题是，该研究只调查了一所私立高中和一所公立高中，而且两所学校的地理位置非常接近。此外，该研究人员只对特定学生进行了调查。因此，该调查的研究结果也许只能推广到与其情况类似的学校。
3. 该研究只采取了物理痕迹法，可以考虑采用其他方法和直接观察法来观察两所学校的学

生；也可以考虑对两所学校学生的态度和行为进行问卷调查。然而，如果想要就此话题在学校中开展研究，必须要处理研究中涉及的伦理问题。

挑战性问题 1 的答案

在此我们只提供每一变量的物理痕迹测量 (PT) 和档案数据 (AD) 的可能来源。它们并非最终答案，只是参考答案。我们只提供了操作性定义的大致内容，大家可以在课堂讨论中得出更全面的操作性定义。

- A. 公众对文化事件的兴趣：(PT) 艺术博物馆中相关展厅地板的磨损情况；(AD) 过去一年中去剧院观看演出、聆听音乐会的人数。
- B. 学生的政治态度：(PT) 在发放竞选传单之后，校园中对不同政治团体的反对数量；(AD) 学生参加地方性会议、州会议以及全国性会议的记录。
- C. 社区对犯罪行为关注：(PT) 社区中每户居民防盗锁的数量；(AD) 警方有关该社区的居民声称“有可疑”活动的记录。
- D. 居民对学校的态度：(PT) 社区小学的操场设施的情况；(AD) 家长教师联谊会的成员名单，如果家长积极参加活动，说明他们对学校的评价更积极。
- E. 流行服饰：(PT) 人们捐献给相关部门的衣服，这代表了曾经流行的服饰潮流；(AD) 杂志、报纸、电视上有关服饰的广告说明了目前的服饰流行趋势。

概

章7章

在第2章中,我们介绍了设计研究的四个阶段:设计、数据、分析和应用。心理学家通常自然而然地倾向于使用实验法对行为进行细致的描述。设计法使心理学家能够了解人们的态度和内心世界,并揭示出他们行为背后的原因。设计法在心理学研究中具有广泛的应用,但是设计法并不是一种单一的研究方法。设计法可以应用于许多不同的研究领域,包括心理学、社会学、教育学、医学、政治学、经济学、历史学、地理学、环境科学、计算机科学、工程学和生物学等。设计法在心理学研究中的应用,主要是为了揭示行为背后的原因,并预测行为。设计法在心理学研究中的应用,可以分为实验法、调查法、观察法和访谈法等。设计法在心理学研究中的应用,可以分为实验法、调查法、观察法和访谈法等。设计法在心理学研究中的应用,可以分为实验法、调查法、观察法和访谈法等。

第三编

实验方法

第7章
独立组设计

第8章
重复测量设计

第9章
复合设计

第7章

独立组设计

本章大纲

概 述

心理学家为什么要做实验

实验研究的逻辑

随机组设计

随机分配

对内部效度的威胁

实验结果的分析和解释

数据分析在实验中的作用

描述结果

确认结果所揭示的内容

数据分析所不能告诉我们的

确立实验结果的外部效度

匹配组设计

自然组设计

小 结

概 述

在第2章中,我们介绍了心理学研究的四个目的:描述、预测、解释和应用。心理学家通常在自然场景中使用观察法对行为进行细致的描述。调查法使心理学家能够对人们的态度和主张进行描述。当他们发现测量和观察结果存在共变(相关)时,心理学家就能做出关于行为和心理过程的预测。对科学研究来说描述和预测是至关重要的,但是这还不足以理解引起行为的原因。心理学家还寻求解释——行为“为什么”发生。只有知道了—一个现象发生的原因时,我们才可以说对此获得了科学的理解。第7、8、9章集中探讨鉴别因果关系最好的研究方法——实验法。我们将探讨如何用实验法来检验心理学理论,以及如何用实验来回答具有重要实际意义的问题。

如我们所强调的,做研究最好多种方法并用。在使用不同的方法获得比较的答案后,我们对结论会更加确信,那时我们的结论就可以说具有了聚合效度。每一种方法都各有不同的局限性,但是多种方法具有互补优势,因此就能克服这些局限缺点。实验法的一个特殊优点是它更有利于确立因果关系。

本章中,我们将讨论研究者进行实验的原因,并考察实验研究的思维逻辑基础,关注广泛使用的实验设计——随机组设计。我们将描述随机组的划分程序,并介绍应用随机组设计时所面临的解释上的风险。我们将描述研究者用来分析和解释实验结果的程序,我们还探讨研究者如何确立实验结果的外部效度。本章最后还会专门讨论匹配组设计和自然组设计,这两种设计与独立组设计有关系。

心理学家为什么要做实验

- 研究者为了检验有关行为原因的假设而做实验。
- 实验能帮助研究者确定一个处理或一个程序是否能有效地改变行为。

心理学家做实验的主要原因之一是为了对源于心理学理论的假设进行实证检验。例如,佩恩贝克(Pennebaker, 1989)提出了一个“压抑理论”,这个理论认为“有泪不轻弹”会损伤身体。佩恩贝克和他的同事们做了很多实验,在这些实验中,他们让一组被试写出关于创伤性情绪的事件,让另一组被试写一些非常一般的主题。与从压抑理论中得出的假设相一致,写出创伤情绪主题的被试比那些写一般主题的被试的健康结果更好。然而,不是所有的结果都与压抑理论一致。例如,被要求用跳舞来表现创伤的学生们就没有和另一群既被要求用跳舞又被要求写出创伤的被试体验到一样的健康好处。佩恩贝克和弗朗西斯(Pennebaker & Francis, 1996)对这个理论作了进一步的检验,并且证实写出情绪体验时所发生的认知变化是解释实验结果的关键。

从上述结果可以看出心理学家用实验来检验理论假设的一般步骤。如果实验的结果与假设所预测的一致,那么这个理论就会获得支持。另一方面,如果结果

与所期望的不同,那么基于理论所提出的解释可能就需要进行修订,然后就会提出一个新的假设并用另一个实验来检验。这种根据实验结果检验假设、修改理论的过程,有时候会是一个漫长和痛苦的经历,有点类似于用细碎的图片拼出一幅完整的图画。实验结果与解释之间的相互自我修正或磨合,也是心理学家用来理解人类思考、感受和和行为原因的一种基本工具。

除了可以对理论假设进行检验,实验还可以用来检验一种处理或程序的有效性。例如,实验经常被用来检验针对特殊病症的医学治疗效果。实验在医学中的作用由来已久(Thomas, 1992)。在大约19世纪初,伤寒症和震颤性谵妄通常是致命的病症。那时候针对这两种情况的标准医学治疗就是放血、清肠和采用其他类似的方法“治疗”。有一个实验对这些一直以来被认为有益的治疗方法进行了检验。一组随机分配的被试接受放血、清肠等治疗;另一组随机分配的被试,除了卧床休息、好的营养和密切观察外,不接受任何治疗。实验的结果非常出人意料:治疗组的情况比不给予任何治疗组的情况还要差。这意味着用19世纪初的这些方法来进行治疗还不如什么都不用!这些实验说明了许多疾病发展到一定阶段后就会自然结束。人们能够不通过治疗,自然而然地从许多心理疾病中恢复,这个过程就是所谓的自然康复。

好的实验能够在多个领域中提供关于处理和程序是否有效的重要信息。例如,塞利格曼(1995)和他的助手做了一系列令人印象深刻的实验,证明儿童在学习了认知性和社会性问题解决技巧后对抑郁有防治作用(Gillham, Reivich, Jaycox, & Seligman, 1995),这个过程被他们称为“心理免疫”。从这个例子来看,实验不仅能够在检验理论时发挥作用,还能够帮助我们寻找解决社会问题的方法。

实验研究的逻辑

- 研究者在实验中通过操纵一个自变量来观察它对行为的影响,其影响效果通过因变量来评估。
- 控制是实验最关键的核心;实验控制通过操纵、保持条件恒定和平衡获得。
- 实验控制使研究者能够做出因果推论,即,自变量导致了因变量的变化。
- 当满足因果推论所需的三个条件时,一个实验就具备了内部效度。这三个条件是共变、时序关系和其他可能原因的排除。
- 当发生混淆时,所观察到的共变就存在其他可能的解释,这将降低实验的内部效度。通过保持条件恒定和平衡(constant and balancing)可排除其他可能的解释。

一个真正的实验需要操纵一个或多个因素,并且测量(观察)这种操纵对行为的影响。如同你在第2章中看到的,研究者控制或者操纵的因素被称为自变量。用于观察自变量效果的因素被称为因变量。

实验能有效地检验假设,通过实验我们可以较高程度地控制某一情景。研

研究者通过实验控制才能确保所观测到的因变量变化是由该自变量引起的。做因果推论需要满足的三个条件是共变、时序关系和其他可能原因的排除（参见第2章）。

当我们观察到实验中的自变量和因变量存在关系时，就满足了共变的条件。当研究者操纵一个自变量，然后观察到随后的行为变化时（也就是说行为的变化取决于实验控制），时序关系即确立。最后，通过使用控制程序，主要是保持条件恒定和平衡，来排除其他可能的原因。当满足了因果推论的这三个条件时，就可以说实验具有内部效度（internal validity），我们就能说是自变量引起了测量到的行为（因变量）的变化。下面，我们首先介绍一个不满足这三个条件的实验。然后再介绍一个已经公开发表的、满足了这三个条件的实验，后面这个实验已具有内部效度。

在一类被称作单组前测—后测设计的研究中，因果推论的条件没有得到满足（Campbell & Stanley, 1966）。这种设计通常包含一组被挑选出来接受处理的被试。在处理前（前测）和处理后（后测），分别对行为进行观察。这就可以类比成用一种新的数学教学方法（处理）来教一个教室里的孩子，在教之前和教之后都对他们进行相关的数学测验。这个设计可以描述为：

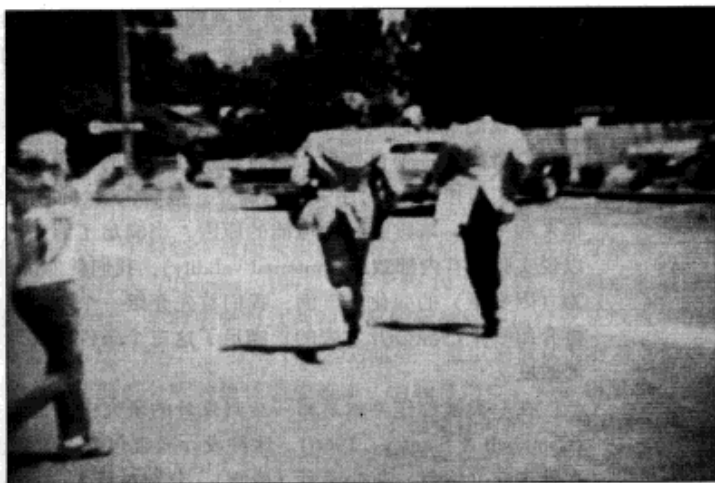
第一步	第二步	第三步
O_1	X	O_2

这里 O_1 指第一次对群体进行的观察，即前测；X指处理； O_2 指第二次观察，即后测。

虽然这种研究设计有时候会在心理学研究中被使用，但是这种设计的内部效度很低，很难解释这类研究设计中的结果。你能看出这类设计没有控制那些影响内部效度的因素吗？例如，如果前测和后测结果存在差异，我们怎样才能确保这个差异是由处理导致而不是由处理之外的其他事件导致的？或者这种差异不是由于被试群体的自然变化引起的呢？这种设计对于因果推论的价值不大，它有时被当作一种“前实验设计”被提到，或者被当作一种“糟糕的实验”用来说明对内部效度可能的威胁（Campbell & Stanley, 1966）。在本章的后半部分，特别是在第11章中我们讨论自然情境中的研究的时候，还将更多地谈到影响内部效度的实验情形。现在，让我们来考察另一种内部效度较高的研究设计。

实验设计的原理及内部效度控制技术可通过洛夫特斯和伯恩斯（Loftus & Burns, 1982）所做的一个目击行为实验来说明。此实验通过让被试观看一场含有暴力场面的电影，进而调查被试对于震撼性场面开始前几秒钟所呈现信息的记忆。实验在华盛顿大学进行，226名自愿参加实验的学生被分成若干组，每组被随机分配观看两部电影中的一部。大约有一半的学生（115名）被分配看了暴力版的电影。在电影的末尾，学生看到一个抢银行的劫匪跑向一辆汽车准备驾车逃走，同时转身朝两个追他的男孩开了一枪，男孩中一人面部中枪并倒在血泊中。另一半学生观看了非暴力版的电影，这个版本与那个暴力版在开枪以前的内容是一样的，所不同的是，非暴力版电影末尾播放的是，银行经理告诉顾客和员工发生了什么，

图 7.1 学生所观看的电影场景的黑白画面，显示男孩所穿足球运动衫的号码是 17
(From Loftus & Burns, 1982.)



并且要求他们保持镇静。电影的暴力和非暴力代表了实验中自变量的两个水平。

看完电影后，两组被试都回答了 25 个与所看电影相关问题。其中最关键的问题是：在银行外的停车场里玩耍的男孩，他穿的足球运动衫上的号码是多少？穿足球运动衫的男孩（图 7.1）在电影里的枪击（暴力场面）发生前出现了 2 秒，在银行场景（非暴力场面）发生前出现 2 秒。因变量是正确回忆运动衫上号码的学生比例。被试所看电影的版本确实影响了他们的回忆。在暴力条件下，仅有 4% 的被试正确回忆出号码。然而在非暴力条件下，差不多有 28% 的正确回忆率。洛夫特斯和伯恩斯得出结论，震撼性事件会削弱被试对发生在事件前的一些细节的记忆。

洛夫特斯和伯恩斯（1982）是如何应用控制技术保持条件恒定的呢？在本实验中，几个可能会影响被试记忆的因素在两种条件下保持了一致。除了关键事件不同外，学生们观看的电影完全一样；在实验开始时，给予他们一样的实验说明；实验中被试收到完全一样的问卷。研究者通过保持条件恒定来确保自变量是唯一系统变化的因素。

假如两组被试除了影片的差异外，还在其他因素上有差异，那实验结果将变得难以解释。例如，如果被试在暴力版本和非暴力版本的电影中，看到的那个男孩穿着不同的运动衫，那么我们将难以确知究竟是电影的版本不同还是运动衫的不同导致了回忆的差异。当拟研究的自变量和其他潜在的变量一起改变时，就会产生混淆。没有这种混淆的实验才具有内部效度。当实验具有内部效度时，我们才可以确定是自变量的变化导致了因变量的变化。

保持条件恒定是研究者用来避免混淆的一种控制技术。通过将两个版本的电影中玩耍的男孩和运动衫（以及其他事物）保持恒定，洛夫特斯和伯恩斯避免了这些因素的混淆。总体上，一个保持恒定的因素不会随着指定的自变量发生变化，且不会与因变量一起变化。如此，研究者可以排除该因素对观察结果的潜在影响。

需要注意的是，我们仅仅选择那些我们认为可能会对研究产生影响的因

素——也即我们所说的貌似真实的可能原因进行控制。例如，洛夫特斯和伯恩斯通过在同一间屋子内对被试进行检测来保持环境恒定，但是他们不可能专门对室内的温度进行控制以保持不同组所在的室内温度恒定，因为室温可能不会影响记忆（至少在小范围内变化时）。他们以“小群体”的方式测试了学生，但是没有打算保持群体大小恒定，因为看电影的群体大小的微小变化不太可能会影响记忆。这就是说，我们仅仅控制那些我们认为可能相关的因素，也就是那些可能会影响行为的额外因素。不管怎样，我们需要时刻警惕，在我们的实验中可能存在一些混淆因素，这些因素对实验的影响是我们尚未预料和考虑到的。

在独立组设计 [independent design, 也称为被试间设计 (between-subjects design)] 中，每组被试代表自变量所界定的一种条件。上述的目击行为实验是一个独立组设计的例子，其自变量（即电影的版本）在不同的被试组间发生变化。由洛夫特斯和伯恩斯所做的这个特殊的独立组设计被称为随机组设计 (random groups design)。要确保独立组设计具有内部效度，关键是要平衡不同实验组间的个体差异。我们拟控制的因素并不能都保持恒定，所以还需要采用平衡这一控制技术。也就是说我们必须确保两组被试在智商、动机、责任心、性别、专业等特征上没有显著的差异性。在洛夫特斯和伯恩斯关于目击行为的研究中，如果观看非暴力版本电影的被试更加专心或者具有更好的回忆能力，那么实验结果就会存在其他可能的解释。有可能是专注性（或者记忆能力）而不是电影的版本导致被试回忆得更好（用研究者的术语来说，就是发生了混淆）。但如果像专注性这样的个体差异在组间得到平衡，我们就可以排除这种混淆，排除组间的因变量差异是由被试特征差异所引起的。若一个独立组设计中的平衡是通过将被试随机分配到不同实验条件下获得的，则我们将此设计叫做随机组设计。

下面，我们将进一步考察随机组设计，并且介绍在这种设计中如何通过随机分配获得平衡。

拓展训练

以下是一个研究的简要叙述，请回答后面所附的问题。

在一个现场实验中，要求学习研究方法的学生作为观察者和主试同谋两个角色，考察个人空间受到侵犯时人们的反应。实验中采用主试同谋盯着被试看，以此作为干扰被试个人空间的手段。被试是在校园中心附近的路上独自走过的学生。主试两个人一组，一个人作为同谋坐在长凳上，另一个作为观察者观察并计算时间。实验分三种条件，每组主试只观察一种条件下的一个被试，条件是随机安排的。实验中的三个条件与主试同谋坐在长凳上的行为有关：主试同谋或者在被试面前拿一

张报纸读；或者当被试接近长凳的时候盯着被试看；第三种条件是在没有主试同谋坐在长凳上时，主试对被试进行观察。充当观察者的学生，分别计算了被试走向长凳和离开长凳的时间。研究方法课的学生预测，当有人盯着被试看的时候，被试会走得更快。然而，实验结果却与他们的预测相反，在被盯着看的情况下，被试走向长凳的时间比其他两种条件更慢，但三种条件下被试离开长凳的时间没有差异。

1. 通过处理，学生对实验的哪些方面进行了控制？
2. 通过保持条件恒定，学生控制了实验的哪些方面？
3. 通过使用平衡技术，学生控制了实验的哪些方面？

随机组设计

随机分配

- 随机组设计中有比较组，每一个组只接受自变量一种水平的处理，其他方面都相同。
- 随机组设计可使研究者做出自变量对因变量产生影响的因果推论。
- 将被试随机分配至各条件下，以形成比较组。其作用是为了在操作自变量的各条件下，平衡或平均被试特征和个体差异。
- 随机区组设计可以平衡被试的个体差异和实验操作过程中可能发生的混淆，并使每组人数相等。

最简单的实验要求一个自变量一个因变量，其中自变量有两个水平，两个不同水平的自变量使因变量出现差异。在独立组设计中，每一组被试参与其中一水平（即每组彼此独立）。例如，在洛夫特斯和伯恩斯（1982）的实验中，一组被试被分配看电影的暴力版，而另一组被试看同一个电影的非暴力版，然后比较两组被试对电影细节的记忆。实验设计的逻辑简单明晰：实验一开始两组被试的所有重要特征都是相似的，实验中两组被试除了自变量的水平不同外（暴力电影对非暴力电影）其他方面都受到了相同的处理。因此，两组被试成绩（对细节的记忆）的任何差异都必定是由自变量引起的。

实验法使得研究者能够解释自变量的作用。将被试分成几个比较组，每一组仅接受自变量一个水平的处理，除此之外在其他所有方面都相似。因此，从逻辑上讲，两组行为上的任何差异都一定是由自变量变化所导致的。很明显，此逻辑成立的关键在于实验开始时形成比较相似的组。在随机组设计中，操纵自变量变化之前，先将被试随机分配（random assignment）以形成比较组和实验组。随机分配的目的是为了平衡各条件下的个体差异。这也就是我们看到的洛夫特斯和伯恩斯在他们的目击行为研究中所遵循的程序。随机组设计的过程可以描绘如下：

第一步	第二步	第三步
R_1	X_1	O_1
R_2	X_2	O_1

这里 R_1 和 R_2 指将被试随机分配到实验条件下（即 1 和 2）； X_1 指自变量的第一个水平（例如，电影的暴力版本）， X_2 是自变量的第二个水平（例如，电影的非暴力版本），有时也可能是不做处理（控制）。然后对每一组被试进行观察（ O_1 ）。与单组的前测—后测设计不同，随机组设计是一种好的实验范例。

随机分配被试的主要优点是可以平衡或者平均两组被试特征上的差异。我们以前指出，如果看不同版本电影的两组被试在记忆能力上有差别，那么我们会认为观察到的对电影细节的回忆差异是由于被试记忆能力的差异所导致，而不是他

专栏 7.1

洛夫特斯和伯恩斯的目击行为实验概述

实验程序概述：大学生被分配看一个电影的两个不同版本。在看完电影后，被试被问及与影片内容相关的一些问题。

自变量：被试所看电影的版本（暴力版或非暴力版）

因变量：关于男孩运动衫号码的回忆

对控制程序的解释：

保持条件恒定：两组学生观看除了关键事件不同、其他方面完全一样的两个版本的电影。实验开始时呈现了同样的指导语，实验结束时收到完全一样的问卷。

平衡：通过将学生随机分配到不同的实验条件下平

衡学生间的个体差异。

提供因果证据的实验逻辑解释：

共变：被试对电影细节的记忆因实验条件改变而改变。

时序关系：对电影版本的操纵先于对记忆力的测量。

排除其他可能原因：通过随机分配保持条件恒定和平衡个体差异避免了混淆。

结论：震撼性事件（暴力版电影）会削弱人们对事件发生前的一些细节的记忆。

（洛夫特斯和伯恩斯，1982）

们所观看的电影版本不同造成的。因为洛夫特斯和伯恩斯在他们的实验中将随机分配到各条件下形成了比较组，而且其他重要的因素都保持恒定，两组之间变化的惟一因素是被操纵的自变量。因此，研究者就能够推论，由于看了不同版本的电影导致了学生记忆成绩的差异。专栏 7.1 总结了洛夫特斯和伯恩斯是如何将实验法应用到他们对目击行为的研究中去的。

随机分配的一种常用程序就是**随机区组（block randomization）**。首先，我们将详细介绍怎样实施随机区组法，然后再看看它能获得什么样的结果。假设一个实验有 5 种条件，一个区组就是 5 种条件的随机顺序中的一种。

一个区组的 5 种条件 → 5 种条件的随机顺序中的一种

A B C D E

C A E B D

使用随机区组时，一个被试被分配到区组的一种条件下，之后第二个被试再被分配到另一条件下。也就是，我们每次将不同被试分配到一个区组的不同条件下。例如，如果我们希望 5 种条件的每一种都有 10 个被试，那么在区组随机分配表中就会有 10 个区组。每一区组将按 5 种条件进行随机分配。具体而言，首先选择参加实验的 5 人，将每一名随机分配到 5 种条件中的一种中去，然后再选取 5 人并把每一个分配到 5 种条件中的一种中去，以此类推，直到这样做完 10 次。开始的 11 个被试的分配程序可示例如下：

随机区组设计

10个区组	被试	条件	
1) CAEBD	1) 卡拉	→ C	第一区组
2) ECDAB	2) 安迪	→ A	
3) DBEAC	3) 雅克	→ E	
4) BACED	4) 莫莉	→ B	
5) ACEDB	5) 艾米莉	→ D	
6) ADEBC	6) 艾里克	→ E	第二区组
7) BCADE	7) 安娜	→ C	
8) DCAEB	8) 劳拉	→ D	
9) EDBCA	9) 萨拉	→ A	
10) CEBDA	10) 莉莎	→ B	
	11) 汤姆	→ D	

等等—共 50 个被试

随机区组是一种非常有用的随机分配被试的方法，因为这种方法在控制与时间相关变量的同时，产生大小相同的组。每组的观测次数影响每组描述统计的可靠性。我们希望不同组的观测结果是可靠的，而随机区组能做到这一点。随机区组之所以有用，还因为实验通常需要大量的时间才能完成。实施实验过程中，被试通常会受到此间所出现的一些事件的影响，随机区组有助于控制这些事件的影响，因为每次将不同被试分配到一个区组的不同条件下。因为每个被试在一个区组的一种条件下接受施测，所以，实验的所有条件下会有相同数量的被试经历实验过程中所发生的事件。随机区组还能平衡其他时间相关变量，如实验者的变化甚至是被试群体的变化。例如，使用随机区组可以很满意地从秋季和春季学期中选取学生进行同一实验。随机区组的优点在于它能平衡（或平均）实验中所有条件下被试特征上的差异（包括时间相关因素的影响）。

如果你想练习随机区组分配的程序，你可以做本章末的挑战性问题 1。

对内部效度的威胁

- 由于不同被试群之间原本存在的差异，将不同的被试群随机分配到自变量的不同条件下会导致潜在的混淆。
- 随机区组设计通过平衡自变量各条件下的额外变量能够提高内部效度。
- 额外变量能否通过平衡或者保持条件恒定得到控制，关系到实验的外部效度和灵敏度。
- 选择性被试缺失会威胁实验的内部效度，而技术性被试缺失则不会。
- 安慰剂控制组用于控制“需求特征”问题，双盲实验既控制需求特征也控制实验者效应。

我们已经知道，内部效度指因变量的变化在多大程度上能清楚明确地被归结于自变量（而非其他未控制变量）的影响。这些未控制变量通常会成为内部效度的威胁（threats to internal validity）。这些威胁会影响到研究结果的解释。为了得到清楚的因果推论，必须控制威胁内部效度的因素。我们下面将介绍实验研究中

威胁内部效度的一些问题以及控制它们的方法。

整群检测 随机组设计中,随机分配的方法可以确保形成有效的比较组。然而,即使采用了随机分配法,有时还是会出现不具可比性的组。当将不同的整群(而不是个体)随机分配到实验的不同条件下时就可能发生这种问题。整群是那些在实验开始前就已形成的群体。例如,一门心理学导论课的不同班级就是不同的整群,学生没有被随机分配到这门课的不同班级(虽然有时候课的安排似乎是随机的)。学生们通常根据这门课的上课时间、任课教师、上这门课的朋友和其他因素来选择其中的某个班级。

然而,当不同的整群间个体存在系统差异时,就不能形成真正的比较组,这时就会出现混淆。例如,选择早晨8点上课的学生可能与喜欢上11点课的学生不同。将这些整群随机分配到不同实验条件下,并不足以平衡整群间的系统差异。两个整群间的系统差异几乎肯定会威胁到实验的内部效度。解决这个问题的方法很简单——在随机组设计中不使用整群。

平衡额外变量 实验研究过程中,操作上的考虑可能会导致大量因素发生变化。例如,为了更快地完成一个实验,研究者可能会决定使用几个不同的主试来施测小群体被试。群体的大小和主试本身就成为了可能混淆实验的潜在相关变量。如果实验组的被试都由一个主试施测,控制组的被试都由另一个主试施测,自变量的效应就与两个主试的效应混淆了。我们不能确定观察到的差异到底是由自变量导致的,还是由于不同主试对实验组和控制组分别进行施测所导致的。

研究者不感兴趣但仍然使实验发生混淆的潜在变量,称为额外变量。但不要被这个名称所欺骗!如果混淆变量本身具有相当大的影响力,那么一个实验被额外变量所混淆的程度会非常高。例如,埃文斯和唐纳斯坦(Evans & Donnerstein, 1974)研究了这样一个额外变量——在学期一开始就自愿参加实验和学期快结束时自愿参加实验的被试间的差异。他们的结果显示,那些早期自愿参加实验的被试在学习上更主动,而且更具有内控性(强调自己对自身行为的责任,而不是强调外部因素)。这意味着,以学期开始时对被试的施测作为实验条件,而以学期结束时对被试的施测作为控制条件,这一做法就是不明智的,因为这会潜在地将自变量与被试的控制点、学习兴趣等特点混淆起来。

随机区组为平衡组间额外变量提供了一种简洁而有效的控制方法。我们要做的就是额外变量的每一个水平上测试所有区组的被试。例如,如果有四个主试,那么每一个主试应施测所有区组的被试。因为每一区组都包含了实验的所有条件,这种策略保证了每个主试能够测试到每一种条件。通常,我们习惯给每个主试分配同样数量的区组,但是这并非必需的。必需的是,所有区组均应在额外变量的每一个水平测试,本例中的额外变量就是主试。当有几个额外变量时,使用平衡法会棘手一些,但是通过仔细的安排能够避免这些因素所产生的混淆。

被试缺失 我们已经强调,随机组设计要求实验组仅仅在自变量的水平上不同。我们已经知道,实验一开始形成比较组是随机组设计的另一个重要特征。同样重要的是,除了实验时的自变量不同外,实验结束时各组间也要具有可比较性。但当被试参与了一个实验却没有完成,实验的内部效度就会受到影响。有必要区分

被试不能完成实验的两种情形：技术性被试缺失和选择性被试缺失（Underwood & Shaughnessy, 1975）。

技术性被试缺失（mechanical subject loss）发生在当被试由于设备问题而不能完成实验时（在这里，也把主试作为设备的一部分）。如果发生了电脑故障，或者主试读错了指导语，又或者某个人无意中打断了实验，技术性被试缺失就会发生。技术性缺失是一个不太严重的问题，因为这种缺失不与被试的任何特征有关。这样的技术性缺失不会导致各实验条件下完成实验的被试特征的总体变化。技术性缺失可以理解为是偶发事件的结果，且在不同组间出现的概率相同。所以，如果是由于技术性缺失导致被试必须退出实验时，内部效度不会特别受到影响。当技术性被试缺失发生时，需要备有说明，并记录下退出的被试名字和退出原因，同时补充新的被试以取代退出的被试。

选择性被试缺失则是一个严重得多的问题。选择性被试缺失（selective subject loss）发生在：（1）当被试在各实验条件下不同程度地缺失；（2）当由于被试的某个特征导致了缺失；（3）当这个被试特征与因变量相关时。选择性被试缺失破坏了对于随机组设计而言特别重要的比较组，从而使实验难以被解释。这个问题会在多种实验情形中发生，如关于手术和药物治疗效果的生理学研究、长达数日或数周的记忆力保持实验以及调查研究中的纵向设计（第5章）。

我们可以通过一个假想的例子来说明选择性被试缺失。设想某社区健身中心的主管想要知道他们进行“健康一月”活动的效果。他们找到了80名自愿参加实验的人，通过随机分配形成两组，每组40人。随机分配所产生的比较组是通过平衡诸如体重、健康水平、动机等特征获得的。控制组的成员仅仅需要在月末时来到中心，给他们做一次健康测试。那些实验组中的被试在月末参加测试前，已经参加了为期一个月的严格的健身计划。假定控制组的40名被试都参加了健康测试，但实验组只有25名被试坚持完成了一个月的健身计划。假定实验组中这25个人的平均健康分数明显高于控制组中40人的平均值。基于这一结果，该健身中心的主管宣称：“一个基于科学的研究表明，我们的活动增进了健康。”

图 7.2 许多人开始进行严格的健身计划，但有些人没能坚持完成。从某种意义上，仅仅是那些最健康的人坚持了下来。如果对不同类型的健身计划进行比较，这种情形可能引起解释上的困难



你认为这个健身中心的说法合理吗？不合理。这个假想研究代表了选择性被试缺失的一种典型情况，该研究结果不能支持该健身中心的观点。被试的缺失在不同情况下有所不同，本研究中只有实验组的被试缺失了。值得一提的是，如果两组的人数本来不同，且实验结束时两组人数保持不变，这并不影响实验结果的解释。如果开始有 25 人被随机分配到实验组，40 人被随机分配到控制组，并且所有被试都完成了实验，那么结果是可解释的。选择性缺失之所以是一个问题，是因为那些完成了健身计划的 25 个实验组被试和 40 个控制组被试不具有比较性。实验组中另外 15 个不能完成健身计划的被试可能比那些完成了计划的 25 个被试更不健康（甚至是在计划开始前就如此）。实验组中被试的选择性缺失破坏了实验开始时通过随机分配形成的比较组。事实上，即使他们没有参加健身，那 25 个实验组被试的最终健康分数可能仍比控制组的要高，因为他们在活动开始时就更健康！因此，本实验中的被试缺失符合选择性被试缺失的其他两种情况，也就是，这个缺失是由于被试的特征——他们最初的健康水平导致的，而且这个特征是与本研究的结果相关的。

如果直到实验完成后才发现选择性被试缺失，除了记录下这个已经实施的、无法被解释的实验过程，其他的也无能为力。然而，当研究者提前认识到选择性缺失有可能发生时，是可以采取预防性措施的。一种做法是进行预测验，排除可能会缺失的被试。例如，可以先对健康状况进行一个初始测试，只有那些在得分上高于某个最低水平的被试才能参加实验。用这种方法筛选被试可能要付出一些代价，研究的结果可能将只能适用于那些高于最低健康水平的人。但这么做是值得的，因为研究结果是可解释的，虽然这降低了结论的普遍性，但仍然要好于一个无法做出解释的研究。

当可能发生选择性被试缺失时，研究者还可以使用第二种方法，即先对所有被试进行预测验，随后将所有被试随机分配到不同的条件下。如果实验组中缺失一个被试，则可以从控制组中删去一个预测分数与之可比的被试。在某种意义上，这种方法是试图恢复两组之间的可比较性。研究者必须能够预见可能导致选择性被试缺失的因素，而且要确保使用的预测验能够测到这些因素。在我们的例子中，做一个健康状况的预测验就要比做一个视敏度预测验更有效，因为后者可能与本研究中的被试缺失没有关系。

安慰剂控制和双盲实验 我们所要描述的最后一种对内部效度的威胁来自被试和主试的期望。“需求特征”是由被试期望导致的一种偏差 (Orme, 1962)。需求特征是指在心理学研究中，被试用于引导自身行为的线索和暗示信息（参见第 4 章）。例如，那些知道自己被给予了酒精的被试，可能预期会体验到一定的反应，如放松或者眩晕，然后他们可能会表现得符合其预期，而这或许不是对这些酒精本身起了反应。潜在偏差在实验中也会因主试的期望产生。用于描述这些偏见的常见术语是实验者效应 (experimenter effects) (Rosenthal, 1963)。如果主试以不同方式对待不同组被试，而没有按照操纵自变量的要求去做，那么实验者效应可能就会引起问题。例如，在一个有关酒精的实验中，如果主试对那些给予了酒精的被试念指导语的速度大大慢于没有给酒精的被试，那么实验者效应就可能出现。如果由于主试用不同方式对待不同组的被试而发生了实验者效应，那么就会出现

自变量的混淆。实验者效应出现的另一种方式是,当主试知道被试接受了何种处理后,会做出相应的倾向性观察。例如,在这个酒精的研究中,如果主试更加注意“喝酒者”不寻常的动作或者含糊不清的语音(他们“期望”喝酒的人是这样的结果),就会出现由实验者效应导致的倾向性观察。(参见第4章关于期望效应的讨论。)

研究者永远不能消除需求特征和实验者效应问题,但是有特殊的研究设计可以控制这些问题。研究者可使用安慰剂控制(placebo control)组作为控制需求特征的一种方法。安慰剂(来源于拉丁语,意为“我喜欢”)是一种看起来像药而实际上是没有药性的物质。一些研究显示安慰剂本身就具有治疗作用(如,Kivsch & Sapirstein, 1998)。皮斯肯和舒雷(Pishkin & Shurley, 1983)在一个测试两种药品(羟嗪和多虑平)效果的实验中,使用了一个安慰剂控制组来控制需求特征。测试此药品是想确定它们是否降低了精神病患者的压力唤醒水平。安慰剂组服用的安慰剂胶囊与这两种药物外观相同,但是安慰剂胶囊中只含有乳糖。安慰剂控制组的病人并不知道他们的药丸中没有药,所以组中所有人都认为他们服用了药,由此产生了相似的期望。也就是,各组的需求特征是相似的,所有被试都期望体验到药物的作用。实验组和安慰剂组之间的任何差异都可以合理地归于实验组被试吃药产生的真实效果,而不是被试对吃药所产生的期望。

使用安慰剂控制组并结合双盲程序能控制需求特征和实验者效应。在一个双盲(double-blind)程序中,被试和观察者都对正在进行的处理茫然不知。在皮斯肯和舒雷(1983)的研究中,需要两个研究者来完成双盲程序。第一个研究者需要准备药丸,并且用某种方式把每个药丸编上号码;第二个研究者需要把药分发给被试,并记录下每个被试所分药丸的编号。这个程序确保记录下了每个被试所服用药丸的编号,但是被试和主试(分发药丸并观察反应的那个实验者)都不知道被试接受的是什么样的处理。因此,主试对于处理的期望就被控制了,因为主试不清楚谁接受了处理、谁接受了安慰剂。同样,需求特征也得到了控制,因为被试不知道他们接受的是药品还是安慰剂。

当评估一个处理的需求特征十分重要时,采用安慰剂控制组的实验就是一种有价值的研究工具。然而使用安慰剂控制组会引发特殊的道德问题。当期望得到药品的被试实际上服用的是安慰剂时,研究者必须考虑到这可能面临的风险。通常实验开始前要在知情同意书中指出实验的道德规范,告知被试他们可能会获得药品或者安慰剂,只有同意服用任一物质的被试才能参加研究。一旦实验药品被证明有效果,研究者在伦理上就有必要为安慰剂条件下的被试提供真实药物治疗。

实验结果的分析 and 解释

数据分析在实验中的作用

- 研究者能否得出自变量影响行为的结论,数据分析和统计在其中起关键作用。



能力图

好的实验开始于一个好的研究问题，所有的研究都是这样。我们已经讨论了研究者如何使用控制技术来设计和操作实验，以搜集说明性证据来回答所研究的问题。这有点类似于侦探在犯罪调查中所做的工作。侦探仔细地搜集证据，以确定嫌疑犯是否是真凶，然而，最彻底的调查也并不足以判定嫌疑犯有罪。原告代理律师必须向陪审团出示证据，而且论据必须能够经受得住被告代理律师的反驳。同样，研究者不能仅仅通过做实验就下结论，他们必须通过一种可信的方式呈现他们的证据，来表明他们的研究结果支持其结论。数据统计和分析在实验结果的分析和解释中发挥关键作用。

罗伯特·埃布尔森 (Robert Abelson, 1995) 在其著作《按原则说理的统计学》一书中，认为数据分析的主要目标是确定我们的观察和测量结果能否支持我们的观点。也就是说，我们是否能够根据我们在实验中搜集的证据来确定我们做出的结论。我们将在第12和13章详细介绍如何进行数据分析和统计。在这里我们将介绍一些数据分析的基本概念。我们首先提及一种能够确保研究者研究结果有效的非常重要的方法。

确定实验结果是否可信的最好办法是重复实验，看两次实验能否获得同样的结果。重复 (replication) 指的是重复该实验的程序，看是否能再次获得同样的结果。但完全一致的重复几乎是不可能的，重复实验中的被试会与初次研究不同，实验场所和主试也可能不同。尽管如此，重复仍然是判定一个研究是否可靠的最好办法。可是，如果我们要求通过重复来确定每个实验的可靠性的话，研究过程会迟缓而低效。实验被试是一种稀缺资源，而且重复意味着我们难以开展关于行为问题的新实验。数据分析和统计为研究者提供了重复实验外的另一个选择，可以通过数据分析和统计来确定一次实验的结果是否能作为自变量影响行为的证据。



实验的数据分析包含三个阶段：(1) 了解数据；(2) 总结数据；(3) 确定数据揭示的内容。在第一步中，我们试图定义数据，寻找错误，确保数据有效；第二阶段，我们使用描述统计和图示来总结结果；第三阶段，我们寻找数据所揭示的有关行为的证据。我们能够得出什么结论？在这一阶段，可能会使用不同的统计技术帮助我们做决定。

下面我们对这些不同的数据分析阶段做个简要介绍。更详细的有关数据分析的介绍在第12章和第13章（重点参看第12章中的专栏12.1）。如果你需要阅读和理解科学杂志上发表的心理学实验结果，或者当你决定自己做心理学实验的时候，后面的这些章节将尤其重要。

我们将使用一个姓名学习实验的例子来说明数据分析的过程。学习并记忆我们遇到的人的姓名，是我们日常生活的一部分。不幸的是，不能记住人们的姓名也是我们日常经历中很常见的令人沮丧的事情。科恩 (Cohen, 1990) 曾经说，姓名比其他信息更难记住，因为姓名通常难以和记忆中的其他信息联系起来。兰多尔和比约克 (Landauer & Bjork, 1978) 介绍了一种称作扩展提取的记忆术，这种记忆术能够帮助人们记忆姓名。扩展提取技术要求，在你被介绍给某个人后，每隔一定时间间隔努力回忆这个你想要记住的姓名，并且这个时间间隔逐渐增加。

图 7.3 研究的呈现材料,包括照片、姓名、个人描述,以及在名字回忆实验中使用的三种可能的提取线索(仅有照片,仅有描述,既有照片也有描述)



在众多测试扩展提取效果的实验中,有一个实验要求学生在学习其他信息的同时学习一些姓名。实验模拟了一个社会情境,即人们在交谈时学习人名。在这个实验中,给予大学生一个基本的记忆任务,要求记忆 50 个不相关的事实。在事实呈现的中途,被试被“介绍”给几个人。在学习完事实和人的名字后,测试了被试对事实和姓名的记忆。

图 7.3 显示了研究过程中的情况以及用于回忆的三个提示线索。在研究中,给被试呈现一个人的姓名和照片以及一段简短的介绍。在姓名回忆测试中,被试被分成三组,每组 36 名,给予每组一个不同的线索来回忆姓名,其中一组仅呈现照片,一组呈现简短描述,一组同时呈现照片和描述。这个实验研究的关键问题是,当被试集中于学习时,扩展提取方法是否有效?结果是有效的!现在,我们讨论的问题是,对姓名的回忆是否因给予的线索不同而不同。

描述结果

- 用来概括实验结果的两种最常见的描述统计量是平均数和标准差。
- 效应大小显示了自变量和因变量之间关系的强度,而且它们不受样本大小的影响。
- d 是一个常用的效应大小指标,它考察与实验的平均变异 (variability) 相关的两组平均数差异。
- 元分析使用效应大小指标来总结关于相同自变量或因变量的诸多实验结果。

数据分析应该始于对数据的仔细观察,特别要注意可能的错误或者反常的数据。第二步是描述数据能发现什么结果。在这一步,研究者想要知道的是,“实验中发生了什么?”。为了回答这个问题,研究者使用描述统计。两个常使用的描述统计值是平均数(集中趋势量)和标准差(变异性量)。表 7.1 给出了姓名学习实验中回忆成绩的平均数和标准差,结果显示在照片线索提示条件下回忆成绩最高(3.98),描述条件下的成绩最低(2.40),照片和描述条件下的回忆成绩(3.80)与照片条件下的更接近。所有条件下的回忆上限是 8 个姓名,我们由此也可以说被试平均回忆量在总数量的 1/3 到 1/2 之间。

在一个设计与实施良好的实验中,每一组的标准差反映的应该仅仅是被随机

表 7.1 姓名学习实验中三种条件下回忆成绩的平均数、标准差和置信区间

线索条件	N	平均数	标准差	.95 置信区间
照片	36	3.98	2.02	3.31—4.65
描述	36	2.40	1.84	1.79—3.01
照片和描述	36	3.80	1.71	3.23—4.37

注：最多能回忆 8 个姓名。

分配到那个组中的被试间的个体差异。每一组中的被试应该被同等对待，而且实施所分配的自变量的水平时也应以同样的方式对待该组内的每个被试。表 7.1 中的标准差显示，每一组个体的回忆成绩围绕平均数存在一个变化范围，三组中的变化差不多相同。

在描述一个实验结果时，研究者想要知道的一个重要问题是，自变量对因变量有多大程度的影响？效应大小（effect size）能够用来回答这个问题，因为它能够反映自变量和因变量之间的关系强度。效应大小的一个优点是它不受实验中所测试样本大小的影响。效应大小考虑的不仅仅是实验中两种条件之间的平均数差异。两组之间的平均数差异通常与被试分数的平均变异性有关。一个常被用来测量效应大小的指标是科恩 d 值（Cohen's d ）。科恩（Cohen, 1992）提出了一个现在被广泛使用的标准，那就是 d 值在 0.20、0.50、0.80 上分别代表自变量的小、中、大三种效应量。我们可以通过比较表 7.1 中所描述的姓名学习实验的照片和描述两种回忆线索，来说明科恩 d 作为效应大小指标的用处。在照片条件（3.98）和描述条件（2.40）之间差异导致的 d 值为 .87。该 d 值表明，两种不同线索变量对于姓名的回忆产生很大效应。效应大小为研究者描述实验结果提供了有价值的信息。

在第 12 章中，我们描述了各集中趋势与变异趋势指标、效应大小指标等，同时简要叙述了这些指标的计算步骤，并且对它们做了解释和讨论。

研究者也在元分析程序中使用效应大小指标。元分析（meta-analysis）是一种统计技术，用来总结和分析不同独立研究的效应大小，这些研究均是针对同样的自变量和因变量。元分析可以回答这样的问题：男女之间在服从上存在性别差异吗？班级大小对学业成绩有影响吗？认知疗法对治疗抑郁有效吗？单独的一个实验无论做得多好，都不足以解答这些重要的一般性问题，我们需要考虑针对每个问题的大量文献和很多实验研究。（参见 Hunt 1997 年有关元分析的精彩介绍。）

元分析使我们能够得到更强有力的普遍性结论，这些结论只有考察多个实验结果之间的关系后才能发现。绳索是由很多股绳结成的，结成的绳索的力量却大于任何一股绳。同样，一个严密的实验增强了对一个特定研究结论的信心，任何单一实验的结果都是组成更强有力的普遍性结论的其中一股绳。例如，罗森塔尔（Rosenthal, 1994a）使用元分析方法总结了 345 个实验结果中的实验者效应问题。实验者效应是一种更普遍意义上的个体间期望效应，指的是人们的期望对他人行为的影响。表 7.2 呈现了 8 个研究期望效应例子，列出 8 个不同领域每个研究的平均效应大小。345 个研究的平均效应大小为 0.70，大于科恩（1992）提出的

表 7.2 元分析中平均效应大小示例

领域	<i>d</i>	研究类型举例
实验室访谈	.14	感觉剥夺对幻觉体验的影响
反应时	.17	特定刺激词对新词联想的影响
学习与能力	.54	智力测验分数、词语学习
个人知觉	.55	对他人成就的知觉
墨迹测验	.84	动物与人罗夏反应比
日常情景	.88	符号学习、运动成绩
心理物理判断	1.05	辨别音调的能力
动物学习	1.73	迷宫和斯金纳箱中的学习
总体平均数	.70	

资料来源: Adapted from Rosenthal, R. (1994a). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science*, 3, 176-179.

中等程度的效应大小(即 0.5)。8 个例子中的 4 个其平均效应大小比 0.80 大, 显示出非常强的效应。罗森塔尔的元分析研究让我们对这样一个普遍的心理学原理更有把握, 即个体间期望能够影响人们的行为。

元分析为总结大量的实验结果提供了有效途径。然而, 只有当用于分析的研究数据以合适的方式搜集时, 元分析中使用的复杂统计技术才是有效的。如果采用了内部效度很差的实验数据, 元分析结果就会产生误导作用。因此, 对于元分析而言, 一个重要的问题是: 元分析中应该分析哪些实验? 应该只分析那些发表在高水准学术期刊上的实验? 还是也分析那些没有经过编辑审查的研究报告? 总体上, 元分析中所采用的实验数据的质量好坏将决定元分析结果的最终价值(参见 Judd, Smith, & Kidder, 1991)。

确认结果所揭示的内容

- 研究者使用推论统计来确定自变量对因变量是否有显著影响。
- 根据样本数据做出推断的两种方法是虚无假设检验和置信区间。
- 研究者使用虚无假设检验来确定实验的组间平均数差异是否比误差变异导致的差异大。
- 一个统计上显著的结果是指虚无假设为真的可能性很小。
- 研究者通过检查实验中不同样本的置信区间是否有重叠来确定一个自变量是否对行为有影响。

研究者做实验时总是希望看到自变量对因变量有影响。换句话说就是, 研究者想要确认自变量使行为产生了变化, 但仅仅靠描述统计不足以做出这种基本结论。

研究者需要使用推论统计来确认一个自变量在实验中产生了效果。因为实验通过随机分配产生了比较组, 研究者需要使用推论统计。如我们以前指出的, 随



随机分配并不能消除实验组间被试的个体差异。随机分配仅仅将被试的个体差异分配到不同的组中，也就是不允许被试的个体差异在组间出现系统改变。每一组内被试的个体差异导致的非系统变化叫做误差变异。误差变异的存在引起了一个潜在的问题，即实验中不同组的平均数之间的差异可能仅仅是由于误差变异所导致，而不是由于自变量的影响而产生。因此，即使是控制得最好的实验所得出的平均数出现差异，也不能由此得出自变量对行为产生了影响。推论统计使研究者能够考察出组间平均数的差异是由于自变量的影响导致的，还是由于偶然因素（误差变异）所导致。研究者使用虚无假设检验和置信区间这两种类型的推断统计来决定自变量是否有作用。

即使实验控制得很好，也不能就断然认为自变量对行为产生了影响，这一点可能让人气馁。换句话说，你迄今所学的研究方法还不够！不幸的是，即使利用数据分析工具，我们仍然不能给你一种方法来得出究竟是什么导致行为发生了变化。但是我们能给予你一种（实际上好几种）方法来考察最有可能是什么导致了行为变化。这一结论建立在概率的基础上——也就是，通过概率帮助你确定行为变化是由于自变量的影响还是由于偶然因素所导致。但使用时很容易迷失在虚无假设检验和置信区间的复杂性中，下面两个关键点需要牢记：

首先，最重要的是，行为变化可能仅仅是由偶然因素（通常是误差变异）引起的。你想要知道的是，你观察到的变化有多大可能仅仅由偶然因素而不是自变量导致的？实际上，你真正想要知道的是，自变量产生了多大的作用？然而，我们得不到这样的信息。后面你将会看到，统计性推断是间接的（见第13章中的专栏13.1）。

第二，你收集的数据代表你从一个总体中获得的样本。但是在某种意义上说，真正重要的是总体而不是样本。（如果仅仅是样本平均数重要，那么你可以仅仅通过看样本平均数来看它们是否不同。）不同实验条件下的样本平均数差异可用来推断总体平均数差异。当你做统计性推断时，你是在使用样本平均数推断总体平均数间的差异。我们再次推荐你看第13章有关这一问题的完整讨论。

虚无假设显著性检验 研究者通常使用虚无假设显著性检验（null hypothesis significance test, NHST）来确定实验中自变量是否产生了影响。虚无假设显著性检验始于这样一个假设：自变量没有产生影响。如果设定虚无假设为真，我们可以使用概率理论来确定实验中观察到的差异偶然发生的概率。一个统计显著性（statistically significant，或译为“统计上显著”——译者注）的结果是指虚无假设为真的可能性很小。一个统计上显著的结果仅仅意味着，实验中我们观察到的差异大于误差变异所导致的结果差异。

实验结果通常以不同实验条件下的平均数差异来表示。我们怎么能够知道一个实验结果发生的概率？研究者通常使用推论统计检验，如 t 检验或 F 检验。当自变量有两个水平时使用 t 检验，当自变量有三个或更多水平时使用 F 检验。当设定虚无假设时，每一个 t 值或 F 值都有一个相关的概率值。一旦研究者计算出统计值，也就确定了相应的概率。

虚无假设出现的概率要小于多少才意味着统计上显著？科学家倾向于认为小

于 5% 的概率 ($p < .05$) 被判定为统计上显著。研究者用来判定一个结果统计显著性的概率值, 被称为显著性水平。显著性水平由希腊字母 α (α) 表示。

现在我们能够通过分析我们前面描述过的姓名学习实验 (参见表 7.1) 来说明虚无假设检验的程序。我们要问的第一个问题是, 作为自变量的提取线索是否有作用? 也就是, 三种提取线索下回忆姓名的平均数是否有差异? 这个检验的虚无假设是, 由不同实验条件下的平均数所代表的总体平均数之间没有差异 (注意, 虚无假设假定自变量不产生影响)。对于该实验中提取线索的作用, F 检验计算出的 p 值为 0.001。因为 F 检验的 p 值比 0.05 的显著性水平小, 提取线索的总体效果在统计上是显著的。为了解释这个结果, 我们需要参考表 7.1 中的这个实验的描述统计量。我们看到, 三种提取线索的回忆平均数是不同的。例如, 照片线索下的回忆最好, 描述线索下的回忆最差。 F 检验的统计上显著结果表明, 提取线索对姓名回忆产生了影响。

研究者通常都想获得更详细结论, 而不仅仅是得出自变量对行为有影响这一结论。平均数总体差异的 F 检验能够告诉我们实验中发生了什么, 但是不能告诉我们发生的程度。为了获得自变量作用的更详细信息, 研究者可以使用置信区间。

使用置信区间检验平均数差异 表 7.1 给出了姓名学习实验中三组的置信区间。置信区间是与概率 (通常为 .95) 相联系的, 这个区间包含了真实的总体平均数。区间的宽度告诉我们预测的精确度怎么样 (越窄越好)。置信区间 (confidence intervals) 还可以用来比较两组总体的平均数差异。我们可以使用表 7.1 呈现的 .95 置信区间探讨提取线索对姓名回忆产生影响的更具体问题。我们可通过考察不同组的置信区间是否重叠做到这一点。当置信区间不重叠时, 我们可确定两组的总体平均数是不同的。例如, 照片组平均回忆的最低估计值为 3.31, 描述组最高估计值为 3.01。因此, 照片线索组的置信区间与描述线索组的置信区间不重叠, 由此我们能下结论说照片线索组的回忆成绩要强于描述线索组的回忆成绩。

然而, 当我们把照片线索组的置信区间与同时给予照片和描述线索组的置信区间进行比较, 得到了不同的结论。这两个组的置信区间确实重叠。即使样本的平均数 3.98 和 3.80 不同, 我们还是不能得出总体平均数不同的结论, 因为置信区间重叠。我们可以提供下面的简单方法来解决此问题: 如果区间轻微重叠, 那么我们还不能确定是否有差异, 应推迟判断; 如果区间重叠, 且一组的平均数处在另一组的区间内, 我们可以推论说总体平均数没有差异。在这个例子中, 照片组的平均数就处在同时给予照片和描述线索组的区间内。我们可以下结论总体平均数没有差异。不过, 这一说法的真正含义是, 我们没有足够的证据表明它们存在差异。我们将在下一节中讨论这一点。

第 12 章讨论了置信区间以及 t 检验的逻辑和计算程序。第 13 章讨论了 F 检验。置信区间及其解释规则在第 12 章 (重点参阅专栏 12.5)。



数据分析所不能告诉我们的

我们已经间接提到了一件数据分析不能告诉我们的事情。即使我们的实验具有内部效度，并且实验结果也具有统计显著性，我们也不能肯定地说我们的变量有效果（或者没有效果）。我们必须学会接受概率陈述方式。我们的数据分析结果同样不能告诉我们，研究结果是否具有实际价值，甚至是否有意义。针对价值不大的研究问题去做实验是简单的（参见 Sternberg, 1997 及第 1 章）。做一个坏实验很容易（可能是太容易！）。缺少内部效度的坏实验很容易出现统计显著性结果和不重叠的置信区间，然而，其结果不具有可解释性。

当一个结果达到统计显著性时，我们就认为自变量对行为产生了影响。然而，如我们已经表明的，我们的分析不能确保我们的结论，即使我们得到的结论“找不到任何疑点”。当一个结果没有达到统计显著性时，我们也不能肯定地说自变量对行为没有影响。我们只是没有足够的证据来表明自变量对行为没有影响。在应用研究中，确定一个没有作用的自变量甚至更加重要。例如，一种药物的效果是否真如其商标上标明的那样？试图回答这些问题的实验所采用的标准要高于那些试图确认自变量确实发生了作用的实验。我们将在第 13 章说明这些标准。

因为研究者依靠概率来作出有关自变量效果的判定，这就免不了有出错的时候。当研究者使用推论统计时会出现两种类型的错误。当我们称一个结果具有统计学上的显著性且虚无假设为真时，我们会犯 I 型错误。I 型错误就像错误报警——当没有火时说有火。当我们没有足够的证据来拒绝虚无假设，而实际上却不是如此的时候，我们会犯 II 型错误（第 13 章将详细介绍 I 型和 II 型错误）。如果我们确切知道虚无假设是真还是假时，我们就不会犯这些错误。我们应注意到数据分析可能会导致错误判断，但我们也要记住数据分析同样能得到正确的判断。对于研究者来说最重要的事情是，推论统计永远不可能取代重复性实验来作为实验结果可靠性的最终检验。

确立实验结果的外部效度

- 当一个实验结果能超越特定的实验环境应用到其他个体、场景和条件中时，这个实验结果就具有外部效度。
- 在一些研究（如验证理论）中，研究者可能更强调内部效度；但对另一些研究者来说，他们可能会选择抽样或重复的办法来增加外部效度。
- 研究者能够通过做现场实验增强他们在现实世界中所做研究的外部效度。
- 部分重复是确定实验外部效度的有效方法。
- 研究者常常寻求变量间关系的普遍性结论，而非一些特定条件下、操纵下、选择场景和样本等情况下的具体结论。

如你在第 4 章中学到的，外部效度指实验结果能够超越特定实验范围，推广应用到其他个体、场景和条件中的程度。对高度控制性实验的一种常见批评是，

它们缺少外部效度。也就是说,在一个得到很好控制的实验室实验中观察到的结果,可能仅仅描述了特定场景、特定条件和特定被试群体下发生的事情。再次回想一下洛夫特斯和伯恩斯(1982)关于心理震撼性事件对记忆影响的实验。在这个实验中,大学生在一个实验控制条件下观看了一部电影。实验室采用了一些控制程序以确保实验的内部效度。可是,这个结果能够用来帮助我们理解自然场景吗?比如在一个现实的银行里发生的目击行为吗?当经历一个不同类型的创伤性事件又会怎样?当目击者是年长者时又会怎样呢?这些都是有关外部效度的问题,且由此产生了一个更普遍性的问题。如果实验室的实验结果是如此特别,那么它们能为社会提供什么好处呢?

首先,对于这个问题的回答不一定完全让人满意。穆克(Mook, 1983)认为,当实验的目的是为了检验源于心理学理论的一个特殊假设时,实验结果的外部效度问题是无关紧要的。做实验通常是为了确定能否以一定方式引导被试产生一个行为。被试在自然环境中是否还是那样行事与实验中的问题相比是次要的。实验的外部效度问题不是一个新问题,正如赖利(Riley, 1962, p.413)所说:“通常,实验室实验不是为了模拟自然情境中最典型的事实。相反,实验室实验是为了回答实验者感兴趣的某些特殊问题。”

当然,研究者通常确实想要获得能够超越实验边界之外的普遍性结果。为了寻求普遍性结果,研究者将相应的自然情景特征纳入到实验中。例如,切奇(Ceci, 1993)描述了一个研究项目,他与同事做了一个儿童目击证词的实验。他们做这个实验的部分动机是因为以前关于这个问题的研究没有包含真实目击情境的所有维度。切奇描述了他们的研究是怎样将诸多因素纳入到实验中的,诸如多重启发性会谈、非常长的回忆间隔以及压力情境下的紧张等。将这些因素纳入到实验情境中,使这个实验更能代表儿童作证时所处的情境。

但是,切奇(1993)也指出,实验和现实生活情境的最重要的差异在于:

“高强度的压力、对受害者身体的攻击以及缺乏控制,是法庭调查研究中典型的特征。虽然这些因素在我们其他的研究中出现过,但是我们不会在儿童受害者身上模拟他们遭受过的攻击性行为,虽然那些非常相近的研究如药物疗效与针对儿童的性侵害不同,已经得到了社会和儿童父母的批准。”(p.41-42)

图 7.4 如何使一个实验与诸如法庭上的儿童目击证词这样的现实生活情境相似



正如切奇所揭示的, 在一些情境中, 如涉及有关卑劣行为的目击证词时, 确定实验的外部效度可能会存在重要的道德约束。

对于研究者来说, 在心理学实验中取得更加具有代表性情境样本可能比获得更具有代表性的人群样本要相对容易。心理学研究结果的外部效度由于被试的特征常常被质疑。如你所知道的, 许多心理学研究以大学生进行实验, 这些实验是他们上的心理学导论课程的一部分。其中道斯 (Dawes, 1991) 就认为, 大学生是一个特别的群体, 他们不可能为建立关于人类行为和心理过程的一般结论提供一个好的基础。他认为大学环境倾向于保护学生免于遭遇外界问题, 从而提高他们对于生活的控制感。生活在大学环境之外的人可能没有和大学生一样的控制感。因此, 以大学生为被试所进行的有关“控制点”研究结论不能推广到大学环境之外的人群中。道斯和其他人的观点告诉我们, 对以大学生为被试得到的实验结论进行推广时要十分小心。

休 (Sue, 1999) 也提到了有关心理学研究结果的外部效度问题。他认为研究者关注内部效度甚于外部效度, 就会减少对心理学研究中人群的代表性的关注。他还进一步指出这样做会阻碍少数民族研究的发展。如果心理学家认为他们的研究结果可以超出其被试群体并能推广到各个群体, 那么就没有理由再通过检验少数民族群体来对结果进行证实。基于群体的研究结果的外部效度问题在应用研究中尤显重要。例如在医学研究中, 对于男性有效的疗法可能对于女性无效, 对于成人有效的疗法可能对于儿童无效。

我们在第4章中简单提到的现场实验是增加外部效度的一种方法。通过该方法能获得有用的结论。例如克鲁索和韦策尔 (Crusco & Wetzel, 1984) 调查了身体接触对饭店消费者产生的影响。女服务员给顾客找零钱的时候, 有意接触顾客的手或者肩膀。研究者预测, 手的接触将使顾客对服务员产生好感, 但接触肩膀将会被看作是一种支配性的信号而产生负面影响, 尤其是对于男顾客而言。研究者将114名顾客随机分配到自变量的三个水平: 短暂地接触手, 服务员找零钱时接触顾客手掌0.5秒; 接触肩膀, 服务员找零钱时把手放在顾客肩膀1~1.5秒; 不接触, 服务员与顾客没有身体接触。因变量是所给小费的数量。与不被接触相比, 男女顾客在被接触后都明显地付给服务员更多的小费。然而, 与研究预期相反的是, 不同类型的接触之间没有差异, 不论何种接触, 男女顾客都给了同样多的小费。因为这个实验是在自然场景中做的, 它更能代表“现实世界”的情况。因此, 相对于在实验室进行的模拟情境实验而言, 我们更确信这一结果能推广到其他现实场景中。

实验结果的外部效度可以通过部分重复实验来确定。部分重复实验通常作为研究过程的一个例行部分, 考察现象稳定发生的条件。部分重复实验使用一个略有不同的实验程序以期望出现相似的实验结果, 并以此来确立实验的外部效度。假定同时在一个大城市的公立大学和一个乡下的小型私立学院进行同一个实验, 实验的被试和环境将是非常不同的。如果在不同的被试和环境条件下得到了一样的结果, 我们可以说这个结果适用于这两种群体和环境。注意, 任一单独实验都谈不上具有外部效度, 只有这样两个实验出现了相同结果才意味着实验具有外部效度。

研究者还可以通过概念性重复来确定实验的外部效度。我们希望从一个研究概括出变量间的概念关系,而不仅仅是一些特殊的条件、操纵、场景或样本(见Banaji & Crowder, 1989; Mook, 1983)。安德森和布什曼(Anderson & Bushman, 1997)列举了一个关于概念性重复的例子。假定我们用5岁的儿童来做一个研究,以确定一种“辱骂”是否会引起愤怒和攻击。随后我们可以重复一次,看同样的辱骂对35岁的成人是否会导致同样的结果。如安德森和布什曼所陈述的,5岁儿童的结果可能不会被35岁的成人所重复,因为对于5岁儿童和35岁成人来说,“辱骂”具有不同的效力。尽管如此,如果我们想要确定“侮辱会增强攻击行为”这个观点的外部效度,我们就必须使用不同的词语,让不同被试群体都能感受到侮辱含义。

当安德森和布什曼(1997)在概念水平上研究与侵犯相关的变量时,发现从实验室中得到的结果与现实场景中的相关研究结果非常相似。他们得出结论,认为实验室实验提供了有意义的信息,因为实验室实验得到了与现场研究同样的结论。而且,实验室实验允许研究者分离出攻击的潜在原因,考察攻击发生的边界条件。

当实验室研究结果与现实场景的研究结果不同时该怎么办呢?安德森和布什曼(1997)认为,这些差异应该有助于我们提炼关于攻击的理论,而不是证明某种方法有缺陷。也就是说,这些差异应该使我们认识到,不同场景下的心理过程可能有所不同。当我们对这些差异更加理解后,将会对攻击的概念理解得更清楚。

通过进行部分重复或者概念性重复实际上不可能确定一个心理学研究结果的外部效度。如果我们认真地对待道斯(1991)和休(1999)等人的观点,那么我们将会面对一个不可能完成的任务。例如,我们如何能够证明从大学生群体得到的实验结果可以推广到年长群体、职业人员群体、低文化群体?安德伍德和肖内西(Underwood & Shaughnessy, 1975)提出了一个值得参考的方法。他们的观点是,除非我们有理由做出其他假定,否则我们应该认为行为在不同时间、不同个体以及不同情境下具有一定的连续性。从根本上说,研究结果的外部效度可能更多地是由相应的科学共同体的一致判断确定,而不能仅由有限的实证证据来确定。

匹配组设计

- 当被试太少,无法采用随机分配时,可以使用匹配组设计以组成比较组。
- 通过因变量任务匹配被试,是组成比较组的最好方法,但是匹配任务都必须和因变量任务相关。
- 被试通过匹配任务匹配后,需要被随机分配到不同的自变量条件下。

随机组设计要有效地进行,需要足够大的样本以确保被试间的个体差异通过随机分配得到平衡。也就是说,随机组设计背后的假设是,个体差异要在组间达到平衡。但是需要多少被试才能确保平衡是有效的呢?答案是视情况而定。与同质群体相比,当样本是从异质群体中抽取出来时,需要更多的被试以平衡组间个

体差异。

当从异质群体中抽取数量不大的样本时,随机分配不能有效地平衡组间被试的个体差异。这确实是许多心理学领域中的研究者所面临的真实情况。例如,发展心理学家中有人研究新生儿,有的研究老年人,这些心理学家常常只拥有有限的被试,且新生儿和老年人代表的是不同的群体。

在这种情形下研究者可以选择使用重复测量设计(将在第8章中进行讨论),将所有的实验条件应用于所有的被试。然而,一些自变量要求不同水平使用不同的被试组。例如,假定要对不同组的早产婴儿进行两种类型的产后照顾,在类似这样的情形中,你需要在实验中对不同的组做不同的检测。

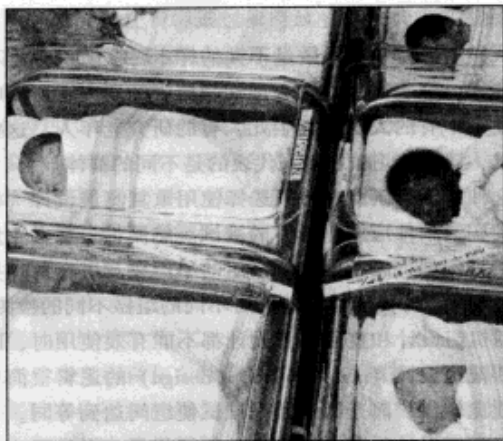
当随机组设计和重复测量设计都不能有效使用时,匹配组设计是一个很好的选择。匹配组设计(matched groups design)的逻辑很简单,研究者不采用随机分配来形成比较组,而是通过匹配被试使组间达到等同。一旦通过匹配形成了比较组后,匹配组设计与随机组设计的逻辑就是一样的了。采用匹配组设计时,多数情况下需使用前测任务来匹配被试。选择前测任务(也叫匹配任务)的关键在于,让组间在与实验结果相关的一个维度上达到相同。只有找到一个好的匹配任务,匹配组设计才是有效的。

最好的匹配任务就是实验任务本身。例如,如果实验的因变量是血压,只需要在实验开始前对被试的血压进行匹配。先测量所有被试的血压,然后将具有相同或相似血压的被试进行二等分、三等分或者四等分(具体根据实验条件的数量),如此完成匹配过程。因此,在实验开始前,不同组的被试的血压平均数是相等的。这样在实验后研究者就可以将组间的血压差异合理地归结为处理的作用(假定组间其他潜在变量已保持恒定或平衡)。

然而,在某些实验中,因变量是不能用来匹配被试的。例如,假定有一个实验研究不同的解迷方法。如果前测是测量被试解迷所需的时间的话,被试可能会在前测期间就已经学会解迷的方法。这样一来,实验可能就不能观察到不同被试组解迷速度的差异了。在这种情况下,最佳选择是使用一个与实验任务同类的任务作为匹配任务。在我们的问题解决实验中,可以根据被试解其他谜的成绩对被试进行匹配。退而求其次的选择是使用一个不同于实验任务种类的任务作为匹配任务。对于我们的问题解决实验,可以根据一些一般能力测试,如空间能力测验结果来匹配被试。但是,在使用这些替代任务时,研究者必须确保匹配任务的成绩与实验中所使用任务的成绩是相关的。通常,当匹配任务和因变量之间的相关降低时,匹配组设计相对于随机组设计的优点也就减少了。

有些实验即使有好的匹配任务,匹配也不足以形成比较组。例如,假定想比较两种不同的早产儿护理方法,这类实验一般采用匹配组设计比较有效。可以根据体重或者婴儿运动协调测验分数来把六对早产儿进行匹配。这个实验可以说明匹配组设计如何才能有效,即仅有少量被试,而且有合适的匹配任务。当我们拥有的被试太少,不能通过随机来形成比较组的时候,我们就需要使用匹配组。匹配组设计能使我们根据与匹配任务相关的维度(例如,早产婴儿的体重)来形成比较组。尽管如此,仍然存在其他潜在的被试特征没有得到匹配。例如,两组早产婴儿可能在一般健康程度、对父母的依恋程度上难以进行匹配。因此,在匹配

图 7.5 当从异质群体中选取少量的被试(例如新生儿)进行测试时,随机分配在平衡组间被试的个体差异上不太有效。在这种情形下,研究者可考虑采用匹配组设计



组设计中通过随机分配来平衡匹配任务之外的其他潜在因素是很重要的。具体就是,在体重和运动协调性方面对婴儿进行匹配后,再将配对的婴儿随机分配到两组中。总的来说,当拥有一个好的匹配任务并且仅仅只有少量的被试参与实验,而且实验的每一个条件需要不同的组参与时,匹配组设计是一个比随机组设计更合适的选择。

自然组设计

- 自然组设计通过选择(不是操纵)个体差异变量(被试变量)而形成。
- 自然组设计是一种相关研究,研究者寻求自然组变量和因变量之间的共变。
- 不能根据自然组变量的效应做因果推论,因为群体差异可能存在其他解释。

许多心理学领域中的研究者都对个体差异变量(individual differences variables)或者被试变量感兴趣。个体差异变量是在个体间变化的一种特性。宗教信仰就是一种个体差异变量的例子。研究者不可能通过将人们随机分配到天主教、犹太教、佛教、新教或其他群体中。研究者通过系统选择天然归属于这些群体的个体来控制这个变量。个体差异变量,如性别、内外向、种族、年龄或者攻击性等,都是人格心理学、变态心理学、发展心理学和社会心理学中重要的自变量。

实验中的自变量究竟是通过选择还是通过操纵得到,这一点非常重要。一个实验中的自变量若是通过选择得到的,如个体差异变量那样,那么这类实验就属于自然组设计(natural groups designs)。由于道德和现实的制约,我们不能直接操纵自变量,这时常常使用自然组设计。有些创伤性事件,如大的外科手术是否会导致抑郁产生?无论我们对此多么感兴趣,但从伦理方面我们决不能胡乱行事,绝对没有人将上心理学导论课程的学生随机分成两组,一组实施手术,然后将随

后的抑郁症状与没有实施外科手术的那一组进行比较。与此类似,如果我们对离婚和情绪障碍的关系感兴趣,我们也不能够按照实验的随机分配原则而让一些人离婚。但是,通过使用自然组设计,我们能够对那些做了手术的人和没有做手术的人进行比较,也可以对离婚的人与婚姻正常的人进行比较。

研究者使用自然组设计首先达到两个目的——描述和预测。例如,有研究显示,分居和离婚的人比那些已婚、丧偶或者依然单身的人更有可能接受精神治疗。以此研究结论为基础,我们可以从情绪障碍角度描述离婚和已婚的个体,而且我们可以预测哪组更易体验到情绪障碍。

然而,以自然组设计的结果做因果结论时会出现比较严重的问题。例如,离婚的人比已婚的人更容易接受精神治疗,这个结果体现了这两个因素的相关。这个结果可能意味着离婚产生情绪障碍,从而需要进行精神治疗。但是,在我们做出是离婚导致情绪障碍这个结论前,我们必须确保满足做出因果推论的时序条件。离婚到底在情绪障碍之前还是之后?自然组设计并不能告诉我们。

即使我们努力排除其他可能原因,满足进行因果推论的第三个条件时,自然组设计也仍旧会带来问题。自然组设计中的个体差异常常被混淆,自然组除了用来区分他们的变量外,其他许多方面都有可能不同。例如除了婚姻状况不同外,离婚个体和已婚个体可能在许多特征上都不相同,比如他们可能在宗教行为和经济条件上就会很不同,离婚和已婚个体间存在的任何差异都可能是由于这些因素导致的。由“自然”完成的这种“操纵”,很难使我们以之建立一个实验的内部效度,也不是实验研究所要求的那种类型的控制。

自然组设计也可以通过一定的改变得出因果推论,比如可以将个体差异与自变量的操纵结合起来研究。这些内容就需要使用混合组设计,这将在第9章中介绍。现在我们可以认识到,基于自然组设计进行因果推论是危险的。虽然我们有时也把这样的设计当作“实验”,但以个体差异变量为基础的实验和以操纵变量为基础的实验之间存在重要差别。

小结

研究者做实验是为了检验源于理论的各种假设,但是,实验还可以用来检验各种处理在实际情境中的效果。当合理使用操纵、恒定和平衡等控制技术时,实验是发现因果关系的理想方法。

在第7章中,我们着重考察了这些控制技术在随机组实验设计中的运用。在这类实验中,不同被试组接受同一自变量的不同水平的处理,通过随机分配程序在实验开始时形成比较组。随机分配是最常用的形成比较组的方法。通过将特征均衡地分配到不同的实验条件下,就可以保证被试间的差异在实验的不同组间达到平衡。随机分配最常用的技术是随机区组设计。

独立组实验会遇到几个影响内部效度的问题。将不同组被试随机分配到不同条件下时,应该避免使用“整群”被试,因为使用“整群”被试会导致混淆。应避免不同的实验场所和不同主试等额外变量与我们感兴趣的自变量相混淆。

在随机组设计中,被试中途退出实验会对内部效度产生严重的威胁。选择性被试缺失发生在不同组被试出现不同程度的缺失,以及一些与实验结果相关的被试特征出现缺失时。要防止选择性被试缺失,我们可以只挑选那些最有可能成功完成全部实验的被试,或者有选择地从没有缺失的对照组中剔除相应的可比被试。通过使用合理的实验程序,将需要特征和实验者效应降到最低,而最佳的控制方式则是使用安慰剂控制和双盲程序。

要确定一次实验的结果是否能作为自变量影响行为的证据,数据分析和统计提供了重复实验外的另一个选择。数据分析包括描述统计和推论统计。描述一个实验结果通常使用平均数、标准差、置信区间和效应大小等指标。元分析使用效应大小指标针对某个同样问题的大量实验结果做出量化总结。

推论统计在数据分析中是非常重要的,因为研究者需要一种方法来确定实验中获得的差异到底是由偶然因素导致的,还是由自变量的影响导致的。置信区间和虚无假设检验是研究者用来分析实验的两种有效的统计技术。然而,统计技术并不能保证实验结果是有意义的还是具有实践意义的。重复实验是检验实验结果可靠性的最基本方式。

研究者还努力建立研究结果的外部效度。当实验用来检验心理学理论的时候,研究者倾向于强调内部效度而忽略外部效度。一种提高研究外部效度的有效方法是在推广研究结论的所有维度上选择代表性样本。通过做现场实验,研究者能够增强其研究结果的外部效度。部分重复实验和概念性重复是研究者用来建立外部效度的两种常用方法。

当只有少量的被试,但拥有好的匹配任务,且实验的每一种处理需要不同组被试时,匹配组设计可取代随机组设计成为一个最佳选择。匹配组设计遇到的最大的问题是,被试组只需要在匹配任务所能测量到的特点上进行了匹配。在自然组设计中,研究者根据个体差异或者被试变量来选择自变量的水平,并寻找自变量和行为间的系统关系。需要注意的是,自然组设计所得出的被试特征和行为之间的关系是一种相关关系,这种相关关系在进行因果推论时会遇到问题。

重要概念

内部效度

独立组设计

随机组设计

随机分配

随机区组

内部效度威胁

技术性被试缺失

选择性被试缺失

实验者效应

安慰剂控制

双盲

重复实验

效应大小

科恩 d 值

元分析

虚无假设显著性检验 (NHST)

统计显著性 (或“统计显著”)

置信区间

匹配组设计

个体差异变量

自然组设计

复习思考题

1. 描述心理学家做实验的两个原因。
2. 操纵、保持条件恒定和平衡等控制技术是如何满足因果推论所必需的三个条件的?
3. 为什么单组前测—后测设计研究谈不上有内部效度?
4. 为什么比较组在随机组设计中如此重要?研究者如何获得比较组?
5. 如果你预测到选择性被试缺失可能会给你的实验带来问题,你将会采取哪些预防手段?
6. 如何使用安慰剂控制和双盲技术来控制需要特征和实验者效应?
7. 为什么元分析能使研究者得到有关心理学原理的更强有力的结论?
8. 一个统计上显著的推论统计检验结果,能够告诉你有关实验中自变量作用的哪些信息?
9. 当你检验一个实验中两种条件下的平均数差异时,如果置信区间不重叠,你能得到什么结论?
10. 简要描述研究者用来确定一个研究结果外部效度的四种方法。
11. 简要解释一下匹配组设计的逻辑。在哪三种条件下,匹配组设计要优于随机组设计?
12. 个体差异变量与被操纵的自变量有什么不同?为什么这种差异使得从自然组设计得出的结论难以进行因果推论?

挑战性问题

1. 某实验者计划做一个随机组设计实验来研究刺激的呈现速率对刺激再认的影响。自变量是呈现速率,有四种实验条件:非常快、快、慢和非常慢。请你对实验的以下方面提出帮助和建议:
 - A. 准备一个随机区组分配表,要求四种实验条件下每种安排四名被试。你可以使用下面的随机数字来完成这项工作。随机数字从附录(表A.1)中取得。
1-5-6-6-4-1-0-4-9-3-2-0-4-9-2-3-8-3-9-1
9-1-1-3-2-2-1-9-9-9-5-9-5-1-6-8-1-6-5-2
2-7-1-9-5-4-8-2-2-3-4-6-7-5-1-2-2-9-2-3
 - B. 实验者打算选择那些能够通过严格的反应时调查的人群,以保证他们能在非常快的呈现速率下顺利完成实验任务。实验者在做出这个决定时需要考虑哪些因素?如果仅仅使用这些严格挑选的被试进行测试,会存在哪些风险?
 - C. 如果必须要在两个不同的房间检测被试,实验者应该怎样安排这两个房间的测试条件,以避免额外变量可能产生的混淆?
2. 某研究者做了一系列实验来考查外部因素对人们坚持进行健身活动的影响。在其中一个实验中,当被试骑脚踏车的时候,研究者使用了三种类型的分心刺激物,分别为全神贯注于自己的内心思想(全神贯注组)、听音乐(音乐组)、观看人们参加户外娱乐的视频(视频组)。因变量是在被试结束运动时脚踏车的强度(此时人们通常会趋向于增加骑脚踏车的强度,以增加运动强度)。一门心理学导论课的120个学生自愿参加实验,研究者将40个被试随机分配到三种分心条件下。研究者预测视频组的平均强度分数最高,音乐组次之,全神贯注组最低。
在脚踏车上运动两分钟后,被试可以选择退出实验

(续)

验。这样短的时间,是为了确保让被试感到疲劳前就有权利选择退出实验。全神贯注组有15个学生退出实验,音乐组有10个退出,视频组没有人退出。所有中途退出实验的被试数据没有纳入到实验数据的分析中。实验结果不支持研究者的预测。完成实验的学生的平均强度分数以全神贯注组最高($n=25$),音乐组次之($n=30$),视频组最低($n=40$)。

- A. 找出这个实验中对内部效度的可能威胁因素,并说明这个问题怎样导致了预期之外的结果?
 - B. 假定对120个被试进行了前测,测量了每个被试的运动持续性程度。怎样使用这些前测分数来确认上一个问题是否已经发生?
3. 一个新闻标题为“研究表明:运动使各年龄的人受益”的文章,概述了一个医学杂志研究报告。文章中描述了一项长达10年涉及约10 000名男性的研究。这些男性在1970年和1989年间进行了骑脚踏车的测验。在第一次测验的5年后,又进行了第二次测验,并且继续对他们的健康监视了5年。在两次骑脚踏车测验中都被判定为不健康的男性,在第二个5年中的死亡率为122/10000。在两次脚踏车测验中都被判定为健康的男性,在第二个5年中的死亡率仅为40/10000。最有趣的是,第一次测验时被判定为不健康,但第二次测验时被判定为健康的男性的死亡率为68/10000。仅仅考察心脏病发作导致的死亡率,发现运动的好处甚至更大。由此,新闻标题说“运动使各年龄的人受益”。
- A. 为什么这篇文章的标题可能存在误导性?
 - B. 为什么研究者仅仅测试了男性?

C. 用两种不同的方法来确定这个研究结果是否可应用于女性,一种采用上述已经发表的研究,另一种是做一个新的研究。比较两种不同的方法是否取得了相同的结果。

4. 某实验拟检测一种药物是否对治疗慢性焦虑症有效。从社区健康门诊鉴别出50个慢性焦虑患者,他们都在被给予了知情权后参加到实验中来。25个被试被随机分配到实验组接受药物治疗;另25个被试被随机分配到控制组接受了安慰剂。在随后6个星期的治疗阶段,医生和临床心理学家对两组被试进行了持续观察。治疗期结束后,被试提交了一个可靠的20点自我评定量表,报告他们感受到的焦虑水平(高分显示高焦虑)。实验组的平均自我评定等级为10.2($SD=1.5$),控制组的平均等级为13.5($SD=2.0$)。实验组的自我评定等级平均数的.95置信区间为9.9到10.5,控制组的为13.1到13.9。
- A. 请解释为什么双盲程序在这个实验中有用?请描述一下这个实验是怎样执行双盲程序的?
- B. 在描述在这个实验中“发生了什么”的时候,你会以什么样的信息开始?
- C. 统计检验表明两组平均数差异的概率是 $p=.01$ 。以此概率为基础,你将得出关于治疗效果的什么结论?基于置信区间的比较,实验得出两组总体平均数的估计值,以此为基础,你会得出什么结论?
- D. 本实验治疗变量的效用大小是 $d=.37$ 。除了你从统计显著性检验中和置信区间比较中得知的信息,这个效用大小告诉了你关于药物效果的什么信息?

拓展练习的答案

1. 学生们在他们的研究中操纵了自变量。自变量有三个水平:没有人坐在长凳上;同伙在长凳上阅读;同伙坐在长凳上,并且盯着被试看。
2. 学生使几个因素保持恒定:路、长凳、仅仅观察独自行走的被试。
3. 通过随机顺序调查不同条件,学生们平衡了三个条件的被试特征。观察者和同伙组成的搭档也在各条件间进行了平衡,每对研究搭档只观察一种条件下的一个被试,每一组被安排的条件是随机的。

挑战问题1的答案

- A. 第一步是将1、2、3、4分配给各条件:1=非常快;2=快;3=慢;4=非常慢。然后,

使用随机数字, 选择数字 1 到 4 的四种序列。在这样做的时候, 你需要跳过任何比 4 大的数字, 和任何已经被选入到序列中的数字。例如, 如果你选择的第一个数字是 1, 你就要跳过所有重复的 1, 直到你选择了从 1 到 4 的序列的所有数字。按照这个程序, 利用随机数字表, 从左到右进行搜索, 我们获得了随机区组分配表的四个区组的四种序列。每一个区组中条件的顺序也都被呈现。区组随机分配表详细说明了对于一个实验最初的 16 个被试进行检测的条件顺序。

区组 1: 1-4-3-2 非常快, 非常慢, 慢, 快

区组 2: 4-2-3-1 非常慢, 快, 慢, 非常快

区组 3: 1-3-2-4 非常快, 慢, 快, 非常慢

区组 4: 2-3-4-1 快, 慢, 非常慢, 非常快

- B. 研究者采取了合理的措施来避免选择性被试缺失, 但是将被试限制在那些通过严格的反应时测试的人群中, 有可能会降低研究结果的外部效度。
- C. 可根据每一个房间里用于测试的随机区组分配表来分配所有区组, 以对房间进行平衡。通常, 分配给每个房间的区组数量是一样的, 但是这并不是必需的。然而, 为了达到有效的平衡, 在每个房间里都应该检测每个区组。

研究者为什么使用重复测量设计

1. 重复测量设计适用于以下情况:
 - (1) 当研究者希望比较同一组被试在不同时间点的表现时。
 - (2) 当研究者希望比较同一组被试在不同条件下的表现时。
 - (3) 当研究者希望比较同一组被试在不同任务上的表现时。
2. 重复测量设计的优点:
 - (1) 可以减少被试间差异的影响。
 - (2) 可以增加统计检验的效力。
 - (3) 可以减少被试流失的影响。
3. 重复测量设计的缺点:
 - (1) 可能会导致被试疲劳或练习效应。
 - (2) 可能会导致被试顺序效应。
 - (3) 可能会导致被试期望效应。

第 8 章

重复测量设计

本章大纲

概 述

研究者为什么使用重复测量设计

练习效应在重复测量设计中的作用

练习效应的定义

在完全设计中平衡练习效应

在不完全设计中平衡练习效应

重复测量设计的数据分析

描述结果

确定结果所揭示的内容

差异转换问题

小 结

1. 描述条件: 1. 2. 3. 4. 分配给的条件; 1 = 非常快; 2 = 快; 3 = 中等; 4 = 非常慢。然后,

概述

到目前为止,我们已经探讨了被试只参与实验中一种条件的情形。被试在随机组和匹配组设计中被随机分配到一个条件中,或者在自然组设计中被选择到一个组中。独立组设计是研究自变量作用的强有力工具,然而有时如果能让每一个被试参与到所有的实验条件中去的话,实验将会更加有效。这样的设计被称为**重复测量设计**(repeated measures design)[也称为被试内设计(within-subjects designs)]。在独立组设计中,控制组与实验处理组是不同的组。在重复测量设计中,被试自己就是控制组,因为他们既参与实验条件,也参与控制条件。

本章我们首先讨论研究者使用重复测量设计的原因,然后对重复测量设计中的关键特征进行描述。在重复测量设计中,被试接受重复测量时会发生变化。例如,被试可能通过练习而得到改进,因为他们学到了更多的关于任务的知识;或者他们在实验情境中更加放松;或者他们也可能由于疲惫或积极性下降等原因,练习后成绩反而变得更差。这些暂时性变化都称为**练习效应**。

我们在第7章中讲到,随机组设计并不能消除被试间的个体差异,但是研究者能够通过随机分配来进行平衡。与此类似,重复测量设计中重复测验导致的练习效应也不能被消除。但是,如同随机组设计中的个体差异一样,练习效应也能够重复测量实验的不同条件间得到平衡。当练习效应被平衡后,重复测量设计实验的自变量效果就具有了解释性。平衡不能消除重复测量设计的练习效应,平衡只能对不同条件下的练习效应进行平均,这样我们就能够避免将实验条件与练习效应混淆。

本章的主要目标是讨论平衡练习效应的各种技术,同时还会介绍重复测量设计的数据分析方法,最后我们会对重复测量设计中出现的问题进行总结。

研究者为什么使用重复测量设计

- 研究者使用重复测量设计是出于以下考虑:(1)实验时被试太少;(2)更加高效地开展实验;(3)增加实验的灵敏度;(4)研究被试在时间维度上的行为变化。

当研究者使用重复测量设计时,他们会得到几个好处。第一,重复测量设计需要的被试少,所以仅有很少数量的被试时,这样的设计是非常理想的。在对儿童、年长者或者具有脑损伤的个体进行实验的研究中,被试数量通常就很少。

甚至当被试的数量足以进行独立组设计时,研究者也会使用重复测量设计,因为重复测量设计常常更加方便和高效。例如,路德维希等人(Ludwig, Jeeves, Norman, & Dewitt, 1993)做了一系列针对大脑两半球之间联系的实验。研究者向被试迅速呈现两个字母,要求他们判断是否相同,同时测量判断所花的时间。字母包括 AaBb。当字母相同(Aa, aa, Bb)时,被试应该按“匹配”键,当字母不

同时 (AB, ab, Ab) 时, 被试应该按“不匹配”键。在四个实验中字母对呈现的方式有好几种, 但有两个主要条件, 或者两个字母被呈现给一个半球 (单侧), 或者被分别呈现给两个半球 (双侧)。在四个实验中均发现, 双侧呈现的反应速度比单侧呈现的反应速度快。在这些实验中, 大脑两半球的双侧比单侧表现更好。

路德维希等人 (1993) 做的每一个小实验都仅仅需要几秒钟就可完成。他们也可以在单侧和双侧条件下独立测试不同的被试组, 但这样做是低效的。与实验所需的时间相比, 实验前向被试介绍实验任务可能要花更多的时间。在重复测量设计中, 每一个被试分别在单侧和双侧条件下被测试, 这为理解大脑如何处理信息提供了更方便有效的途径。

重复测量设计的另一个重要的优点是, 它通常比独立组设计更加灵敏。实验的灵敏度 (sensitivity) 是指实验能够探测出自变量对因变量产生作用的能力, 即使这个作用很小。同一实验条件下被试都反应的个体差异越小, 那么这个实验就越灵敏。也可以说, 在任何一个实验条件下的被试不会按照完全相同的方式进行反应。这种误差变异可以归结于每次实验操作过程中发生的变化, 或者是被试的个体差异。通常, 重复测量设计中相同被试自身随着时间变化而发生的改变, 要比随机组设计中不同的被试间所发生的改变要小。换种说法就是, 群体间的差异通常要比群体内的差异大。因此, 重复测量设计中误差变异通常较小。误差变异越小, 就越容易检测出自变量的效果。重复测量设计的这种高灵敏度, 很适合研究自变量对行为仅有很小影响的情形。

研究者使用重复测量设计还有一个原因就是—一些心理学领域需要使用它。当研究问题所涉及被试的行为在时间维度上发生变化时, 比如说学习实验或者纵向设计实验 (第 5 章), 就需要使用重复测量设计。另外, 一旦实验程序需要被试比较两个或更多刺激的相互关联时, 必须使用重复测量设计。例如, 如果一个研究者想测量一束光变得更亮时光强度的最小增量, 或者让被试对一系列照片的吸引

专栏 8.1

重复测量和重复测量设计

在不同的情境中, 区分哪些情境中研究者对被试进行了重复测量是很重要的。研究者对被试的测量通常多于一次, 这样做是有重要原因的, 这些原因通常与测量的信度有关。研究者可能获得一个个体的两个 (或更多) 测量值, 以确立一个测量的信度 (一致性)。例如, 一群被试可能在两个不同的时间, 接受同样的一个态度测验, 以评估该态度测验的重测信度。可使用相关系数来确定态度测验的信度。重复测量被试的另一个原因是为了获得一个更加稳定而可信的行为量值。例如, 有一个关于个体抑郁水平的研究, 以被试的自我报告量表为

基础。测量抑郁的调查者可能不想仅用只有一个题目涉及抑郁的问卷。而如果有多个不同的题目来调查人们当前心境的话, 这样的测量将会更加可靠。

将重复测量设计与研究者为了确定或者为了增强一个测验的信度而对被试进行重复测量是两回事。在重复测量设计中, 研究者操纵自变量以比较被试在一个实验的两个或更多条件下的行为。重复测量设计中最关键的重复测量, 是在实验的所有条件下对每一个被试进行测量。

力进行评定,就必须使用重复测量设计。许多研究领域,如心理物理学(光照刺激实验)等对重复测量设计有很重的依赖。《知觉和心理物理学》(Perception & Psychophysics)、《实验心理学杂志:人类知觉和行为》(Journal of Experimental Psychology: Human Perception and Performance)常常刊登一些使用重复测量设计的实验。

练习效应在重复测量设计中的作用

- 因为参与自变量的每一条件(水平)的被试是相同的,重复测量设计不会被个体差异变量所混淆。
- 在重复测量设计中,由于重复测量的原因(不是因为自变量的原因),被试在不同条件下的表现可能会发生变化,这些变化可称作练习效应。
- 当自变量的不同条件以同样的顺序呈现给被试的时候,练习效应可能会威胁到重复测量实验的内部效度。
- 有两种类型的重复测量设计(完全的和不完全的),它们分别采用特定方式来控制练习效应。

练习效应的定义

除了我们所描述的优点外,重复测量设计还有另外一个重要的优点。在一个重复测量设计中,被试的特征不会和实验操纵的自变量混淆。重复测量设计中,所有条件下都是相同的被试接受测试,所以不可能出现在一种条件下的被试与另一种条件下的被试相比,变得更加聪明、健康,或者动机更强的情况。说得更正式一点就是,在重复测量设计中,个体差异变量不会导致混淆。不存在由个体差异变量导致的混淆,这是重复测量设计的一个非常大的优点。然而这并不意味着使用重复测量设计就不存在对内部效度的威胁。

被试状态可能随着时间而发生变化,这可能会对内部效度产生威胁。重复测量设计中对被试的重复测试,使被试能够练习实验任务。作为练习的结果,被试可能会因为对任务更加了解而能够越来越好地完成任务;也可能会因为疲劳和枯燥感等因素而使得他们的表现越来越差。被试在重复测量设计中发生的这些变化,被统称为练习效应(practice effects)。

卡尼曼、弗里德锐森、施赖伯和雷德梅尔(Kahneman, Frederickson, Schreiber, & Redelmeier, 1993)在一系列实验中使用了重复测量设计来研究人们对于疼痛的感知。他们的研究结果显示,人们对于疼痛的评价受感受强度和感受时刻的影响。换句话说,人们是通过疼痛最剧烈和最后时刻的疼痛感受来判断有多疼的。这两个因素对于疼痛的感受比疼痛持续性所起的作用更重要。我们将使用卡尼曼等人做过的其中一个实验为例,说明练习效应的特性和在重复测量设计中对它们进行控制的必要性。

图 8.1 不断练习一个新技能既有消极作用又有积极作用。重复同样的过程能够获得绩效改善,但也会导致疲劳、动机减弱甚至厌倦



这个实验的目的是为了确定在经历最后时刻的疼痛时,疼痛强度的影响。实验有两个条件:第一个条件(短实验)下,被试将一只手浸入冷水中(14 摄氏度)持续 60 秒;在第二个条件(长实验)下,被试先把手放入 14 摄氏度的水中持续 60 秒,但是随后伴随着水温逐渐变暖至 15 摄氏度,继续把手浸入里面 30 秒。每一个被试都经历一次短实验和长实验,两次实验间隔时间为 7 分钟。因变量是人们在做完每个实验后,对疼痛程度的评定。实验中最重要测量是,当学生被问及更愿意重复哪个实验时他们所做的选择。

研究者发现,学生对疼痛的等级评定显示,水温在 15 摄氏度时会有疼痛感,但是比水温在 14 摄氏度时的疼痛感明显要轻。学生们还认为,总体来说长实验比短实验更加令人感觉到疼痛。然而,当学生们被问及他们更愿意选择哪种实验进行重复时,60%的人选择了更加疼痛的长实验。卡尼曼等人(1993)认为,学生选择长实验,因为长实验在最后时刻的疼痛感要小。相对于学生在长实验的最后时刻体验到的小一点的疼痛感而言,长实验的持续时间在学生做出选择中所起的作用要小一些。为了完成这个有趣的观察,研究者需要在他们的实验中控制练习效应。

假定卡尼曼等人(1993)对所有被试都是先实施短实验,然后实施长实验。如果仅仅采用了这一种顺序,则短实验总是在长实验之前进行。在从一个实验到另一个实验过程中,学生可能会对浸入冷水中的感觉更加熟悉;或许他们可能在经历过一个实验后,对可能的疼痛变得不再那么担心。在做这两个实验的过程中,学生可能会在很多方面发生改变。这些在两个实验中发生的变化反映的都是练习效应。如果只采用一种实验顺序,练习效应和实验的两个条件可能会发生混

淆。如果学生对两种条件做出了不同的评定,我们就不能够解释这种差异。也就是说,我们不会知道这种差异到底是由于自变量导致的还是由于实验顺序导致的。

卡尼曼等人(1993)通过对两个条件下的练习效应进行平衡以对其进行控制。他们用两种不同的顺序对条件进行操作。一半学生先进行短实验后进行长实验,另一半学生则相反。由于两种顺序得到平衡,如果学生更加喜欢重复长实验,就不能归因于实验顺序的因素。第一个实验对第二个实验的练习效应产生的影响在两种条件下可以被平衡,因为有一半的被试是在长实验中经历练习效应,另一半的被试是在短实验中经历练习效应。因此,两种条件下的任何差异表现,都是由短实验和长实验的特性决定的,而不是练习效应所致。

通常,在重复测量设计中,应该平衡各个实验条件间的练习效应。使用重复测量设计法进行实验的关键是要学会使用合适的方法来平衡练习效应。在介绍如何使用特殊的平衡方法之前,我们将简要介绍重复测量设计的两种类型。

这两种重复测量设计是完全设计和不完全设计。两种设计中用来平衡练习效应的特殊方法是不同的,但是用来指代这些平衡方法的通用术语叫抵消平衡法(counterbalancing)。在完全设计中,每一个被试的练习效应都得到了平衡,这种平衡是让每一个被试经历几种条件下的操作且每次所使用的顺序不同。在不完全设计中,每一条件对每个被试只执行一次,且执行条件的顺序在被试间有所不同。这样结合所有被试的结果时,不完全设计中的练习效应就得到了平衡。这点看起来可能有些令人迷惑,但是在我们更加充分地介绍这些类型的设计后会变得明白。应谨记的是,使用重复测量设计的一个主要目标是控制练习效应。

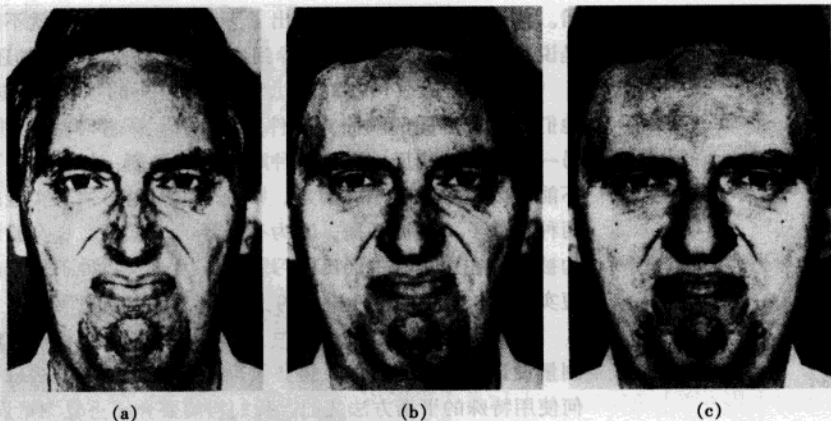
在完全设计中平衡练习效应

- 完全设计是通过在被试内使用随机区组分配或 ABBA 抵消平衡法来平衡练习效应的。
- 在随机区组分配中,实验条件(一个区组内的)以随机的方式排序。
- 在 ABBA 抵消平衡法中,实验条件先以一种随机顺序呈现,然后再呈现相反的顺序。
- 当练习效应呈非线性时,或者当被试成绩会被期望效应影响时,随机区组分配比 ABBA 抵消平衡法更好。

研究显示,通过观看描绘人类六种基本情绪(开心、惊奇、恐惧、悲伤、愤怒和厌恶)的面部表情照片,人们很容易准确辨别出照片所表达的情绪类型。萨克海姆、格和索西(Sackheim, Gur, & Saucy, 1978)使用了重复测量设计来确定我们面部的一侧是否比另一侧表现的情绪更加强烈。他们使用了一张面孔及其镜像图,将两幅图像从正中间分成两部分,其中一幅组合图是两个版本中的左侧组成的,另一幅组合图是由两个版本中的右侧组成的。参见图 8.2,中间是一张表现厌恶表情的原初面孔图,呈现在原初图两侧的是面孔的组合图。在图 8.2 中的两张组合图,其中一张厌恶的表情是否比另一张看起来更强烈呢?

将图 8.2 中的面孔以幻灯片的形式呈现给被试,并要求被试用 7 级评分来评价每一张面部表情所表达情绪的强烈程度。幻灯片呈现给每个被试的时间为 10 秒,

图 8.2 由原初面孔的左侧组合成一张面孔 (a), 原初面孔 (b), 和由原初面孔的右侧组合成一张面孔 (c)。面孔表现的是厌恶表情 (来自萨克海姆等, 1978)



然后给予被试 35 秒作出等级评定。实验中的自变量是描述一个人情绪的图片 (左侧组合、原初图、右侧组合)。每一个被试要评定 54 张幻灯片, 其中分别有 18 张左侧组合、18 张原初图和 18 张右侧组合面孔。

被试对左侧组合面孔的情绪强烈度评价一致高于右侧。萨克海姆等使用大脑半球特殊化来解释这种结果。通常, 大脑左半球控制身体的右侧, 右半球控制身体的左侧。因此, 面孔的左侧组合反映了右半球的控制, 右侧组合反映了左半球的控制。左侧组合图获得的情绪强烈度越高, 那么就显示右半球可能比左半球更多地参与到情绪表情的产生上。

对结果的解释关键取决于幻灯片呈现给被试的顺序。设想如果首先呈现所有的原初面孔, 然后呈现所有的右侧组合面孔, 最后是所有的左侧组合, 那么将会发生什么? 如果你能想像到自己在这么长的时间内 (40 分钟) 来对每张面孔进行评定, 你就能体会到什么是练习效应。在对这些面孔进行情绪评定时, 你的注意力、动机和感受肯定会发生变化。如果你对最后呈现的面孔评定等级较高, 那么你的评定等级反映的可能是你本身拥有的疲惫和厌倦情绪, 而不是图像中实际描绘的情绪强烈程度。而通过使用平衡技术, 就可确保三个版本的面孔在各个位置上出现的概率相同。

在完全设计中, 每种条件下被试都将进行若干次实验, 以便在被试内平衡练习效应。当任务很简单而且不太费时的情况下 (如判断面孔的情绪强烈度), 有可能对每个被试进行若干次同样的处理。事实上, 在一些完全设计中, 仅仅只测试了一到两个被试, 而且每个被试要逐个参与上百个试验 (或上百次回答)。可是, 研究者更加普遍的是使用类似萨克海姆等人使用的程序, 那就是要测试多个被试, 每个被试都被给予次数相对较少的处理。研究者在决定如何安排完全设计的处理顺序时有两种选择: 随机区组分配法和 ABBA 抵消平衡法。

随机区组分配 我们在第 7 章介绍了随机区组分配技术在随机组设计中的应用。随机区组分配法还可以在完全设计中用来对每一个被试的实验条件进行排序。例如, 萨克海姆等人将三种面孔图片 (左侧组合、原初图、右侧组合) 中的每一种都对每一个被试呈现 18 次。表 8.1 中显示的试验顺序说明了怎样利用随机区组分

配对他们的3种实验条件进行排序。54个试验序列被分解成3种试验的18个区组。每一个试验区组包括以随机顺序呈现的3种实验条件。通常,一个随机表的区组的数量是与每种条件被执行的次数相同,而且每个区组的大小是与实验中条件的数量等同的。

如果一个被试是按照表8.1中呈现的区组随机分配序列来对面孔情绪的等级进行评定的话,那么被试评定时的注意力、动机或者体验方面的变化对所有条件产

表8.1 在一个具有3个条件,每一个条件被执行18次的实验中,全部54个试验的区组随机序列

试 验	条 件	试 验	条 件
1	O	28	O
2	L	29	L
3	R	30	R
4	R	31	R
5	O	32	L
6	L	33	O
7	R	34	O
8	O	35	R
9	L	36	L
10	L	37	L
11	O	38	R
12	R	39	O
13	O	40	O
14	L	41	R
15	R	42	L
16	R	43	R
17	L	44	O
18	O	45	L
19	R	46	R
20	L	47	L
21	O	48	O
22	L	49	O
23	R	50	R
24	O	51	L
25	R	52	R
26	L	53	O
27	O	54	L

注意:条件是萨克海姆等人(1978)使用的照片的3个版本:L=左侧组合面孔,O=原初面孔,R=右侧组合面孔。

生的影响将是相同的。在 3 种实验条件之间的练习效应可以认为已经被平衡掉了。对每一条件在随机区组分配序列列表中的平均位置进行确定, 可以粗略显示对练习效应的平衡情况。这可以通过计算每一条件出现的试验数量总和, 然后除以 18 来计算。例如, 面孔的原初版本出现在实验 1、5、8、11、13、18、21、24、27、28、33、34、39、40、44、48、49 和 53 中。因此, 原初面孔的平均位置为 27.6。左侧和右侧组合面孔的对应值分别为 27.7 和 27.2。这些平均数如此接近, 说明任何版本在 54 个试验序列中出现的先后概率差不多是相同的。

随机区组分配法在平衡练习效应时是有效的, 但是在平衡掉练习效应前, 每一种条件必须重复好几次。我们不能期望在进行完两三个区组后就能平衡掉练习效应, 就如同我们在随机组设计中不能期望只用两三个区组的样本规模就可形成比较组一样。为了使随机区组有效, 通常需要对每个条件进行次数充分的试验, 当这样做不可能时, 幸运的是, 还有一种技术可以用来平衡练习效应。

ABBA 抵消平衡法 ABBA 抵消平衡法采用最简单的形式, 仅仅需要对每一种条件进行两次操作, 就可以在完全设计中平衡练习效应。ABBA 抵消平衡法首先以一种序列 (先 A 后 B) 呈现条件, 然后再用相反的顺序呈现 (先 B 后 A)。这一名称指的是当实验中仅有两种条件 (A 和 B) 时的序列, 但是 ABBA 抵消平衡法并不限于只有两种条件的实验。萨克海姆等人已经根据表 8.2 中呈现的 ABBA 序列安排了他们的面孔刺激实验。在这种情况下有三种条件, 应该采取 ABCCBA 的序列。前三个实验的条件顺序在试验 4 到试验 6 中进行了简单的反转。

只有当练习效应是线性的时候, 才适合使用 ABBA 抵消平衡法。如果练习效应是线性的, 那么每个连续实验的练习效应就具有同样的大小。表 8.2 中标为“练习效应 (线性)”的一行说明 ABBA 抵消平衡法是如何平衡练习效应的。在这个例子中, 每个试验成绩都加了 1 个“单元”的假定练习效应。因为第一个实验没有练习效应, 所以加在实验 1 上的练习效应量是 0。实验 2 加了 1 个单元的练习效应, 因为被试经历了第一个实验。在第三个实验中, 加了 2 个单元, 因为被试经历了两个实验, 依此类推。

通过合计每种条件的练习效应值, 可以对练习效应的影响有所了解。例如, 左侧组合条件下得到的练习效应最小 (0) 和最大 (5); 右侧组合条件得到的是中间值 (2 和 3)。对于两个条件的假定练习效应和都为 5。ABBA 循环能够应用于任何数量的条件下, 但是每一种条件重复的次数必须为偶数。循环的重复次数越多, ABBA 抵消平衡法就越能有效平衡练习效应。然而, 通常是在条件数量和对每种条件进行重复的次数较少时才使用 ABBA 抵消平衡法。

虽然 ABBA 抵消平衡法提供了一种平衡练习效应的简单方法, 但是它也有局

表 8.2 在 3 个条件 (左侧组合, 原初的, 右侧组合) 的实验中各试验的 ABBA 抵消平衡序列

条件:	实验 1	实验 2	实验 3	实验 4	实验 5	实验 6
	左侧	原初	右侧	右侧	原初	左侧
练习效应 (线性)	+0	+1	+2	+3	+4	+5
练习效应 (非线性)	0	+6	+6	+6	+6	+6

限性。例如, 当一个任务的练习效应是非线性的时候, ABBA 抵消平衡法是无效的。这在表 8.2 中标示为“练习效应(非线性)”一行中清晰可见。在这个例子中, 左侧组合面孔一共收到了 6 个单元的假想练习效应, 另外两个条件每一个都收到了 12 个单元的练习效应。当练习效应最初发生急剧变化随后变化很小的时候, 研究者常常忽略早期试验中的表现, 而一直等待练习效应达到“稳定状态”。为了达到稳定状态, 可能需要对每种条件进行几次重复, 所以研究者倾向于在这类情况下使用随机区组分配来平衡练习效应。

当期望效应发生时, ABBA 抵消平衡法也是无效的。当一个被试能预期序列中下一个刺激条件是什么时, 期望效应就发生了。被试对那个条件的反应可能会更多地被这种期望所影响, 而不是被与这种条件相关的实际体验所影响。例如, 设想一个时间知觉实验, 被试在这个实验中的任务是估计时间间隔——从计算机屏幕上呈现一个“开始”信号到屏幕上再呈现一个“停止”信号之间的时间长度。(当然, 被试不能通过计算或者有节奏的敲击来计算时间间隔的长度。)如果这个实验中的时间间隔是 12、24、36 秒, 那么一种可能的 ABBA 序列为 12-24-36-36-24-12。如果这个循环被重复好几次, 被试可能就能识别出这种模式, 并且会对时间间隔的增加和减少产生一种期望。他们的时间估测值可能就会迅速地体现出这种模式的影响, 而不是反映出对每个独立时间间隔的感知。如果可能存在期望效应, 就应该使用随机区组技术而非 ABBA 抵消平衡法。

在不完全设计中平衡练习效应

- 在不完全设计中, 练习效应在多个被试间进行平衡, 而不是像完全设计那样在每个被试内进行平衡。
- 在不完全设计中平衡练习效应的规则是, 实验的每一个条件在每个序列位置上(第一、第二, 等等)呈现的可能性必须相等。
- 当拥有四个或四个以下条件时, 在不完全设计中用来平衡练习效应的最好方法是使用所有可能的条件顺序。
- 在不完全设计中, 两种特别的顺序呈现方法分别是“拉丁方”和“轮转的随机开始顺序”。
- 无论是使用所有可能顺序法, 还是使用选择顺序法, 被试都应该被随机分配到不同的序列中。

在不完全设计中, 每一个被试都只被处理一次。否则, 每一个被试的自变量水平都会和这些水平被呈现的顺序完全混淆在一起, 致使任何一个被试的结果都难以得到解释。例如, 在一个不完全设计中的第一个被试, 可能先在实验条件(E)下被测试, 然后再在控制条件(C)下被测试。实验条件和控制条件下被试表现的任何差异既可以归因于自变量效应, 也可以归因为由 EC 顺序导致的练习效应。为了区分条件顺序和自变量之间的混淆, 我们可以对不同的被试执行不同的条件顺序。例如, 我们可以采用 CE 顺序对第二个被试执行我们的不完全设计实验条件, 即首先在控制条件下进行测试, 而后在实验条件下进行测试。通过这种方

法,我们可以用两个被试来平衡两个条件的顺序效应。

我们用埃伯(Erber, 1991)的一个实验说明不完全设计中平衡练习效应的方
法。这个实验研究的是人们的情绪对他们感知他人的影响。该实验分为两个阶段。
在第一阶段,有三个关于年轻女作家的故事,被试阅读其中的一个。以前的研究
显示,这些故事在引发人们的积极、中性的和消极的情绪方面很有效。积极情绪
条件的故事包括几个事件,最后以女作家获得大学奖学金而结束。中性故事描述
了女作家如何选择进入哪所大学。消极情绪故事描述了女作家在她大学的第一年
时,如何被一种罕见的疾病所击倒。因此,第一阶段后,被试被假定处在三种情
绪中的一种之中:积极的、中性的或者消极的(注意,在实验的第一阶段,每一
一个被试都只处在一种情绪组,所采用的是独立组设计。)

接着进入实验的第二阶段,并且被试认为第二阶段是一个独立的实验。要求
被试对四个“目标”个体采取特定行为的可能性(用拥抱来欢迎一个朋友,对天
气感到沮丧等等)进行等级评定。用某种程度上相矛盾的特质来对四个目标分别
进行描述:(1)情绪低落和热情;(2)悲观厌世和善解人意;(3)不自私和不
善交际;(4)值得信赖和有占有欲。采用重复测量设计以操作目标个体的自变
量。特别是,每个被试仅对目标个体变量进行一次等级评定,因此采用的是不完
全设计。

在介绍不完全设计中用来平衡练习效应的技术之前,我们先简单了解一下埃
伯的研究结果。这个研究中的因变量是被试对目标个体实施某种行为的可能性所
做的等级评定。研究者假设这些等级会受被试情绪的影响(积极的、中性的和消
极的)。因为对目标人物的描述既使用了积极的(如热情)特质,也使用了消极的
(如情绪低落)特质,研究者想要看看被试的情绪会决定两种特质中的哪一种。例
如,被试的任务是评定一个被描述为“情绪低落和热情的”人会有多大可能采取
积极的行为,如“用拥抱来欢迎一个朋友”,或者是采取一种消极的行为,如“对
天气感到沮丧”。用拥抱欢迎一个人是一个“热情”的人可能会做的事,对天气感
到沮丧是一个“情绪低落”的人可能拥有的行为。也许你会猜到在这个实验中
将发生什么。

结果显示,如果被试处在积极的情绪中,那么一个“情绪低落和热情”的人
更可能会被评定为采取积极行为。如果被试处在消极的情绪中,同样是“情绪低
落和热情”的人可能会被评定为会采取消极行为。情绪的影响在评定消极行为时
要大于在评定积极行为之时。这些研究结果显示,情绪会影响我们知觉其他人的
方式。埃伯(1991)论道:消极情绪可能会促使被试更容易感知他人的消极信
息。也就是说,如果你处于一个坏环境中,你可能会更加强调别人的消极特质。

现在我们把注意力转向不完全设计中使用的平衡技术。在一个不完全设计中,
通过改变目标的呈现顺序来平衡练习效应是非常重要的。在不完全设计中通常采
用的平衡规则是比较简单的:实验的每一个条件在每一个序列位置上出现的可能
性必须相等。有几项技术可以满足这个规则。使用这些平衡技术的效果是不同的,
但是合理使用这些技术是能满足基本要求的,而且结果也将更容易解释。也就
是说,如果采取了合理的平衡技术,那么我们将能够区分到底是自变量还是练习效
应影响了被试的行为。

所有可能顺序法 在不完全设计中平衡练习效应最好的方法是把所有可能的顺序都考虑到, 每个被试被随机分配到一种顺序中。两个条件就只有两种可能的顺序 (AB 和 BA); 3 个条件的话就有了 6 种可能的顺序 (ABC, ACB, BAC, BCA, CAB, CBA)。总体上, 如有 N 个条件的话, 就有 $N!$ 种可能的顺序, 这里 $N!$ 等于 $N(N-1)(N-2)(N-3)\cdots(N-[N-1])$ 。如我们看到的, 3 种条件就有 6 种可能的顺序, $3! (3 \times 2 \times 1 = 6)$ 。随着条件数量的增加, 需要的顺序数量大大增加。例如, 仅仅 5 个条件就有 120 种顺序可能, 6 种条件就有 720 种可能的顺序。由于这个原因, 通常在包含 4 个或 4 个以下条件的实验中, 才使用所有可能的顺序。

因为在埃伯 (1991) 的实验中, 有 4 个目标人物, 所以为了获得实验条件的所有可能顺序, 就需要 24 个序列, 序列在表 8.3 的左侧呈现。使用所有可能顺序法当然要满足一般规则——所有条件在每一个序列位置上出现的可能性必须相等。第一个序列位置最清楚地显示了这种平衡: 前面的 6 个序列都以情绪低落—热情 (M) 目标开始, 后面的每一套六个序列都是以其他三个目标中的一个开始的。4 种顺序都采用了相同的模式。例如, “M” 在第二个、第三个和第四个序列位置上也都出现了 6 次。

在决定使用所有可能顺序时, 还要注意到另外一个问题。为了使这项技术有效, 每个被试必须在所有可能的条件顺序中被测试。也就是说, 至少一个被试应该接受 MPUT 顺序, 至少一个被试应该接受 MPTU 顺序, 以此类推。因此, 使用所有可能顺序法要求至少要具有和可能顺序一样多的被试数, 也就是说, 如果实验有 4 个条件, 至少必须测试 24 个被试 (或者 48 个或者 72 个, 或者是 24 的其

表 8.3 在一个有 4 种条件的不完全重复测量设计中, 可用于平衡练习效应的几种技术

所有可能顺序法								选择顺序							
								拉丁方				轮转的随机开始顺序			
序列位置				序列位置				序列位置				序列位置			
第 1	第 2	第 3	第 4	第 1	第 2	第 3	第 4	第 1	第 2	第 3	第 4	第 1	第 2	第 3	第 4
M	P	U	T	U	M	P	T	M	P	U	T	P	U	T	M
M	P	T	U	U	M	T	P	P	T	M	U	U	T	M	P
M	U	P	T	U	P	M	T	T	U	P	M	T	M	P	U
M	U	T	P	U	P	T	M	U	M	Y	P	M	P	U	T
M	T	P	U	U	T	M	P								
M	T	U	P	U	T	P	M								
P	M	U	T	T	M	P	U								
P	M	T	U	T	M	U	P								
P	U	M	T	T	P	M	U								
P	U	T	M	T	P	U	M								
P	T	M	U	T	U	M	P								
P	T	U	M	T	U	P	M								

注: 4 种条件是以埃伯 (1991) 实验中的描述四个目标人群的第一个形容词的第一个字母来表示的: M 代表情绪低落—热情; P 代表悲观厌世—善解人意; U 代表不自私—不善交际; T 代表值得信赖—有占有欲。

他倍数)。这种约束条件要求在测试前就搞清楚参与实验的被试数。

选择顺序法 我们刚才描述了不完全设计中的平衡练习效应的一种方法——所有可能顺序法。然而，有时使用所有的顺序是不切实际的。例如，如果我们想要使用不完全设计来研究具有 7 个水平的自变量，若使用所有可能的顺序——每个被试参与 7 种条件的一种可能顺序，我们就需要有 5 040 个被试（即 $7!$ 个顺序）。显然，如果我们在 5 个或者更多条件下还想要使用不完全设计的话，我们就必须使用一种能够代替所有可能顺序法的方法。

仅仅使用所有可能顺序中的一部分顺序，也能平衡练习效应。所选择顺序的数量应该都是实验条件数量的倍数。例如，要做一个自变量有 7 个水平的实验，我们需要选择 7 个、14 个、21 个、28 个或者 7 的其他倍数来平衡练习效应。使用选择顺序的两个主要变式在表 8.3 中给予了说明，从中可以更加直接地比较平衡的种类，表中描述了自变量为 4 个水平情况下选择顺序的技术，这也是我们前面介绍的埃伯（1991）所做实验中使用的方法。

第一种使用选择顺序法来进行平衡的方法叫做拉丁方。在一个拉丁方中满足了对练习效应平衡的一般规则，那就是，每一条件在每个不同序列位置上出现一次。例如，表 8.3 的右半部分，我们可以在拉丁方中看到，“M”在第一个、第二

专栏 8.2

如何构建一个拉丁方

一个实验可以构建一个以上的有效拉丁方，构建一个具有偶数 (N) 条件的拉丁方的简单程序如下：

1. 对实验条件随机排序。
2. 为随机排列的条件分配从 1 到 N 的号码。

因此，如果你有 $N=4$ 个条件 (A,B,C,D)，而且随机顺序是 BADC 的话，那么 $B=1$ ， $A=2$ ， $D=3$ ， $C=4$ 。

3. 使用下面的规则产生第一行（条件的第一个顺序）：

1, 2, N , 3, $N-1$, 4, $N-2$, 5, $N-3$, 6 等。

在我们的例子中，这将产生 1, 2, 4, 3。

4. 给第一行的每一个数字加 1，来得到第二行（条件的第二个顺序），但是要记住 $N+1=1$ 。那么我们将获得 2, 3, 1, 4。

5. 通过给第二行的每一个数字加 1，得到第三行，仍要注意 $N+1=1$ 。

第三行将是 3, 4, 2, 1。

6. 对于每一个后续的行，都采取相似的程序。

7. 如在第二步中所确定的，将每个条件分配给它们所对应的数字。

这个例子的拉丁方将是：

```
B A C D
A D B C
D C A B
C B D A
```

如果有奇数个条件，那么必须构建两个拉丁方。第一个可以根据上面给出的针对偶数拉丁方的规则构建。第二个拉丁方是通过将第一个拉丁方的行进行反转产生的。例如，假定 $N=5$ ，并且第一个拉丁方的第一行是 BAECD。那么，第二个拉丁方的第一行将是 DCEAB。两个拉丁方结合起来将形成一个 $N \times 2N$ 方。无论在偶数还是奇数情况下，被试都应该被随机地分配到拉丁方的行中。因此，你必须具有至少和行列乘积数一样多的被试。（选择或构建拉丁方的程序也可参见 Winer, Brown, 和 Michels [1991, pp.674-679]。）

你能构建这个 4×4 拉丁方的第 4 行吗？

个、第三个、第四个序列位置上分别仅仅出现了一次，每个条件都是如此。另外，在一个拉丁方中，每个条件先于和后于其他条件的次数也仅有一次。对表 8.3 中拉丁方进行研究发现，“MP”顺序仅仅出现一次，“PM”也出现一次，“PU”顺序仅仅出现一次，“UP”也出现一次，等等。所有条件的组合都只出现一次。（专栏 8.2 描述了建立拉丁方的程序。）

使用选择顺序来进行平衡的第二种技术要求以一种随机顺序开始，并且系统地轮转这种序列，使每一种条件每次向左移一个位置（见表 8.3 右边的例子）。使用经过轮转的随机顺序能有效地平衡练习效应，因为每种条件在每个序列位置上同样都出现一次，如同拉丁方一样。尽管如此，序列系统地轮转意味着每个条件总是出现在相同的另一个条件之前或之后，这与拉丁方技术是不同的。运用轮转的随机顺序非常简单，而且它能够应用到多于四种条件的实验中去，这是这种轮转技术的主要优点。

使用所有可能顺序、拉丁方、轮转的随机开始顺序这三者在平衡练习效应时同样有效，因为所有的技术都保证了每个条件在每个序列位置上出现的可能性相等。无论一个人使用哪种技术来平衡练习效应，都应该在测试被试前充分准备好条件出现的序列，而且保证被试被随机分配到这些序列中去。

重复测量设计的数据分析

描述结果

- 对于一个完全设计，数据分析首先需要计算每一个被试的分数（如平均数、中数）。
- 描述统计用于总结所有被试在自变量的每个条件下的表现。

实验分析的第一步是准备一个数据矩阵，其中概括了被试在每一种实验条件下的表现。在随机组设计中，这意味着简单地列出在每种实验条件下被试的测试分数，然后用描述统计量如平均数和标准差来概括这些分数。在一个不完全重复测量设计中，每个被试在每种条件下只获得一个分数，但是计算每个条件的分数仍然相对比较直接。这时需要特别注意，要保证被试的分数是列在其对应的正确的条件下的。一旦每个条件的所有分数都被列了出来，就可以计算平均数和标准差来描述每个条件下被试的表现。

分析完全重复测量设计时，还有一个任务需要完成。在描述和总结结果前，必须先计算每个被试在每种条件下的分数。这一额外的步骤是必需的，因为在完全设计中，每一个被试在每一条件下被测试不止一次。例如前面讲的完全重复测量设计例子中，5 个被试参加了时间知觉测验。实验的目的不是为了测试被试估计时间的准确性，而是为了确定被试对时间的估计值是否会随着时间间隔长度的增加而系统增加。换句话说，被试能否区别不同长度的时间间隔？

实验中，每个被试在 4 种长度时间间隔（12、24、36、48 秒）的每个上都被

表 8.4 一个重复测量设计实验的数据矩阵表

被 试	数据矩阵			
	间隔长度 (秒)			
	12	24	36	48
1	13	21	30	38
2	10	15	38	35
3	12	23	31	32
4	12	15	22	32
5	16	36	69	60
平均数 (标准差)	12.6 (2.0)	22.0 (7.7)	38.0 (16.3)	39.4 (10.5)

注：表中的每个值代表了在每个时间间隔长度条件下，被试六个反应的中数。

测试了6次。用随机区组分配方法确定时间间隔呈现的顺序。因此，每个被试都回答出24个时间估计值，4种时间间隔中的每一种都提供了6个估计值。由于每个时间间隔的6个估计值都被练习效应污染了，所以需要有一个能够整体反映6个估计值的量。通常，需要计算出每个时间间隔的6个估计值的平均数，作为被试在每种条件下的单独估计值。但我们知道，极端分数能够影响平均数，且很有可能每个时间间隔的6个估计值中，被试至少给出了一个极端估计值。因此，6个估计值的中数可能能够最好地反映被试对时间间隔的估计。这些中数估计值列示在表8.4中。（你可能习惯上将平均数和中数作为一组被试的描述统计量，但是这些统计量也能用于个体在完成多个试验后的成绩描述，如本例。）

一旦获得了被试在每种条件下的个体分数，下一步就该通过适当的统计方法来总结被试的结果。表8.4中的“平均数 (标准差)”行，列出了每个时间间隔的平均数估计值和标准差。虽然表中只包含5个被试的数据资料，但是这些平均数估计值却显示被试能够区分不同长度的时间间隔，至少能够区分36秒以内的时间间隔。

如我们以前提到的（在第7章中），在描述一个实验结果时，应该包含效应大小。重复测量设计的效应大小通常是测量间的联系强度，叫做eta方 (η^2)。本实验的 η^2 值为0.80。这个值显示出被试的时间估计值的变异在很大程度上都可以归结于自变量——时间间隔长度的影响。在第12和13章中，还可以找到更多的有关效应大小的信息。在第13章中，我们将会举例说明如何使用表8.4中的数据计算 η^2 。

确定结果所揭示的内容

- 重复测量设计中虚无假设检验及置信区间的一般程序和逻辑与随机组设计中是相似的。

重复测量设计的实验数据分析程序与我们在第7章中所介绍的随机组设计的

拓展练习

对于这个练习，你需要计算在这个不完全重复测量设计中自变量的每一水平的平均数。

表中显示了一个实验中实施给6个被试三种剂量的药（0、50、100）。圆括号中的值代表每个被试在每个条件下完成问题解决任务所犯错误的数量。表中的数据可用来描述本实验中药品剂量这个变量的效果。

被 试	序列位置		
	1	2	3
1	0 (2)	50 (9)	100 (9)
2	0 (7)	100 (5)	50 (3)
3	50 (5)	0 (3)	100 (4)
4	50 (8)	100 (10)	0 (6)
5	100 (6)	0 (8)	50 (7)
6	100 (8)	50 (4)	0 (4)

实验数据分析程序相似。研究者使用虚无假设检验和置信区间来判断自变量是否对行为产生了影响。我们将使用时间知觉实验为例，说明研究者使用重复测量设计时怎样确认数据所揭示的内容。

时间知觉实验分析的焦点在于判断被试是否能够区别不同长度的时间间隔。除非我们知道表 8.4 中的平均数差异比误差变异大，否则我们不能说被试能够分辨出不同长度的时间间隔。也就是说，即使被试能区分出不同的时间间隔，但是我们不知道他们的表现是否仅仅由偶然因素引起的。我们应该考虑使用虚无假设检验和构建置信区间来帮助我们弄清自变量的效应。

重复测量设计数据分析的一个显著特征是误差变异的估计方式。我们在第 7 章中说过，在随机组设计中，组内被试间的个体差异单独提供了误差变异的一个估计值。可是在重复测量设计中，被试间的差异不仅仅被平衡了——它们实际上被从分析中排除了。由于重复测量中排除了由被试导致的系统变异，因此重复测量设计比随机组设计更加灵敏。在重复测量设计中，误差变异的来源主要是实验条件对不同被试的影响。



重复测量设计和独立组设计对误差变异的估计是有差别的，这意味着在虚无假设检验中 t 检验和 F 检验的计算也是不同的。与此类似，计算置信区间的方法也有所不同。在第 13 章中，作为重复测量设计的一部分，我们使用了表 8.4 中的数据，显示如何同时利用 F 检验和置信区间来做出判断。对这些数据进行了变异分析后（参见第 13 章），我们可以告诉你，对时间间隔所进行的 F 检验得出的概率显著性水平 $p = .0004$ ，因为得到的这个概率比常规的显著性水平（.05）小，所以时间间隔这个变量的作用就具有统计显著性。以这个结果为基础，我们可以说，被试对时间的估计确实由于时间间隔长度的影响而发生了变化。我们已经通过计算效应大小（ $\eta^2 = .80$ ）得知，它代表的是一个较大的影响。

在第 13 章中，我们使用了同样的数据来计算表 8.4 中的平均数的 .95 置信区间。4 个条件的置信区间（以秒为单位）分别为：（12）5.4–19.8；（24）14.8–29.2；（36）30.8–45.2；（48）32.2–46.6。如你在第 7 章中了解到的（也参见专栏 12.5），当区间不重叠的时候，我们能够说由样本平均数估计的群体平均数是不同的。通过观察这几个区间，

你是否能够判断出哪些平均数不一样？一种检查置信区间之间关系的简单方法就是用图对它们进行标绘。可以参考第13章中的图13.2，那里标绘出的区间都是围绕在时间估计实验中获得的样本平均数周围的。

差异转换问题

- 当一个条件的效果持续发生作用，并且影响到随后条件中的表现时，差异转换就发生了。
- 因为差异转换会对重复测量设计的内部效度产生威胁，应该使用随机组设计对可能导致差异转换的变量进行测试。
- 可以通过相同自变量在重复测量设计和随机组设计中的结果比较，来鉴别是否发生了差异转换。

在重复测量设计中，研究者可以通过使用适当的技术来平衡练习效应，以克服练习效应可能产生的问题。但在重复测量设计中可能会出现一个更为严重的问题，那就是所谓的差异转换（Poulton, 1973, 1975, 1982; Poulton & Freeman, 1966）。当一个条件中的行为表现会因前面的条件而发生变化时，就会产生差异转换（difference transfer）。

有这样一个问题解决实验，用重复测量设计来比较两种类型指导语的作用。一套指导语（A）期望能够促进问题解决，另一套指导语（B）作为中性控制条件。被试按照AB顺序接受测试，有理由认为当他们遵从B的指导语时，将难以排除A指导语的影响。让被试不按照指导语A中所说的方式行事，可能反而会使被试不由自主地按照指导语A行事。比如若指导语是“不要去想一头粉红色的大象”，你肯定会对粉红色大象的印象更加挥之不去。当被试被要求遵从指导语B时，他们偏偏不能够摒除条件A的影响，那么这时两个条件之间的差异就被减少了。对于这些被试，条件B甚至一点都没有起到作用。实验中，被试更可能是在“AA”条件下测试，而不是在“AB”条件下测试。

差异转换通常会威胁内部效度，因为无法确定不同条件间是否真正存在差异，同时这种情形下也容易低估条件间的差异，从而降低研究结果的外部效度。因此，当差异转换可能发生时，研究者应该选择使用独立组设计。差异转换通常都是和指导语变量在一起的，对于这些变量建议不要使用重复测量设计（Underwood & Shaughnessy, 1975）。不幸的是，差异转换在任何重复测量设计中都可能出现。例如，被试在已经接受200单位的大麻后再接受50单位的大麻所产生的效应，与被试接受安慰剂后再接受50单位的大麻所产生的效应可能会不同（即被试在接受200剂量的大麻后，对大麻的耐受性会提高）。尽管如此，我们还是想办法确定差异转换是否已经发生。

确定差异转换是否发生的最好办法是分别做两个实验（Poulton, 1982）。这两

个实验研究的自变量相同,但是一个实验使用的是随机组设计,另一个实验使用的是重复测量设计。随机组设计不可能具有差异转换,因为每个被试在每个条件下只被测试一次。如果使用重复测量设计的实验所得结果与独立组设计中的一样,那么可能就没有差异转换。如果同样的自变量在两个设计中显示出了不同的作用,那么很可能是差异转换导致了重复测量设计出现了不同结果。如果确实发生了差异转换,就应该使用随机组设计的实验结果来判断自变量的作用。

小结

通过对每个被试执行所有的实验条件,重复测量设计提供了一种有效的实验方法。当仅有很少量的被试,或者只有通过少数被试进行多次测试才能对自变量进行最好的研究时,重复测量设计才是有用的。重复测量设计通常更加灵敏,另外,心理学研究的特殊领域(如心理物理学)可能更需要使用重复测量设计。

然而,要使重复测量设计实验具有可解释性,就必须平衡练习效应。练习效应是由于重复测试导致的被试行为上的变化。在一个完全重复测量设计中,练习效应的平衡是在每个被试内进行的。随机区组和 ABBA 抵消平衡法能够平衡练习效应。然而,当练习效应是非线性或者可能存在期望效应时,不能使用 ABBA 抵消平衡法。

在不完全重复测量设计中,每个被试只接受一次处理,练习效应的平衡是在被试间进行的。在不完全重复测量设计中,平衡练习效应可以使用所有可能顺序法,或者使用选择顺序法(拉丁方和轮转的随机开始顺序)。

重复测量设计与随机组设计数据的分析过程实质上是相同的。首先核对数据错误,然后使用描述统计量,如平均数、标准差和效应大小指标等对数据进行总结概括。虚无假设检验和置信区间可用来推论自变量是否对行为产生了影响。

重复测量设计中可能出现的最严重问题是差异转换,即一种条件下的结果依赖于其前面的条件。虽然有办法检查实验是否发生了差异转换,但是研究一旦发生了差异转换,我们几乎是无法挽救这个研究的。

重点概念

重复测量设计

灵敏度

练习效应

抵消平衡法

差异转换

复习思考题

1. 指出随机组设计中的平衡设计平衡了什么,在重复测量设计中又平衡了什么。

- 简述研究者选择重复测量设计的四个原因。
- 什么是灵敏度？解释一下为何重复测量设计常常比随机组设计更加灵敏。
- 区分一下完全重复测量设计和不完全重复测量设计。
- 在一个完全重复测量实验中，为了平衡练习效应，研究者有哪些选择？
- 在一个完全重复测量实验中，哪两种情况下不适宜使用 ABBA 抵消平衡法来平衡练习效应？
- 简述不完全重复测量实验中平衡练习效应的一般规则。
- 简述不完全重复测量实验中平衡练习效应的几种技术。鉴别一下这些技术中的哪一种更受欢迎，并且解释原因。
- 在完全重复测量设计的实验中，为什么首先要对数据进行归类总结？
- 在完全重复测量设计的实验中，研究者如何才能确定差异转换发生了？

挑战性问题

- 下面的问题代表了重复测量设计中平衡练习效应的不同情形。

A. 在一个使用完全设计的重复测量实验中，自变量是任务难度，包括三种水平（低、中、高）。请为这个实验的不同条件排序以平衡自变量的练习效应。请首先使用随机区组法平衡练习效应，然后使用 ABBA 抵消平衡法平衡练习效应。每一种条件要在顺序中出现两次。（可以使用附录中的随机数字表 A.1 中第一行的数字，来确定随机区组分配的两个随机序列）

B. 在一个使用不完全设计的重复测量实验中，自变量是一个段落的字体大小，有 6 个水平（7、8、9、10、11 和 12）。现有 6 个被试，为了平衡练习效应，请确定针对 6 个被试的各实验条件的执行顺序，列表说明。

- 追踪转轮是一种运动知觉测试。测试中使用一个转盘，转盘里嵌有一个硬币大小的圆板。给予被试一个指示器，当转盘旋转的时候，要求被试将指示器保持在圆板上。因变量是在每次试验中被试将指示器保持在圆板上的时间。该任务的学习与练习期间的试验次数有线性关系，且通常需要很长时间才能掌握。一个研究者想要研究一天中不同的时间点对该任务成绩的影响，自变量有 4 个不同水平（即 4 种条件，分别是上午 10 点、下午 2 点、下午 6 点和晚上 10 点）。在每种条件下，被试进行相同数量的试验，而且被试每

天只在一个条件下接受测试，连续测量 4 天。现有 96 个被试可参加实验。

- 在这个实验中，对时间变量采用的是何设计？
- 对于给予每个被试的第一个条件，可采取什么样的设计？如何利用每个被试第一种条件下的实验结果来检测是否发生了差异转换？

- 下表给出了一个不完全设计的重复测量实验中被试参与实验的顺序。在这个实验中，自变量是当被试专注于一项任务时，觉察到的一个声音的响度。三种声音分别为极度轻（ES）、非常轻（VS）和轻（S）。圆括号里的值代表了被试在每个条件中觉察到声音的次数。必要的话，请使用这个表，来回答下面的问题。

被 试		条件顺序		
1	ES (2)	VS (9)	S (9)	
2	VS (3)	S (5)	ES (7)	
3	S (4)	ES (3)	VS (5)	
4	ES (6)	S (10)	VS (8)	
5	VS (7)	ES (8)	S (6)	
6	S (8)	VS (4)	ES (4)	

- 在这个实验中，使用了什么方法来平衡练习效应？
- 用数值描述响度变量的总体作用。在描述统计的基础上对自变量的作用给予适当的文字描述。
- 如果对响度变量作用进行 F 检验，其概率显著性水平为 $p = .04$ ，那么关于响度变量能得出什么结论？

拓展练习的答案

将每个条件下的6个值(从每个被试那里得到一个)求平均,计算出每个条件的平均错误量。例如,对于0剂量条件来说,6个值是2、7、3、6、8和4。0剂量条件的平均数就是5.0,50剂量条件下的平均数就是6.0,100剂量条件下的平均数就是7.0。这些平均数显示了在问题解决任务上的错误随着药品剂量的增加而增加。使用虚无假设检验或者置信区间的推论统计检验,可以确认药品剂量这个变量是否产生了一个稳定的效果。

挑战性问题1的答案

- A. 分别将1、2、3分配给低、中、高条件,并使用附录中的随机数字表(表A.1)的第一行,以这一行中的第一个数字打头的话,那么随机区组序列就是:低—高—中—低—中—高。一种可能的ABBA抵消平衡序列是:低—中—高—高—中—低。
- B. 因为有6个条件,采取所有可能顺序是不可行的。因此,需要采用拉丁方或者轮转随机开始顺序来平衡练习效应。一种可能的轮转序列组是:

被 试	序列位置					
	第一	第二	第三	第四	第五	第六
1	8	10	11	9	7	12
2	10	11	9	7	12	8
3	11	9	7	12	8	10
4	9	7	12	8	10	11
5	7	12	8	10	11	9
6	12	8	10	11	9	7

第9章

复合设计

本章大纲

概 述

确定实验设计的指导原则

描述复合设计中的实验效应

2x2 设计举例

主效应和交互作用

描述交互作用

复合设计的数据分析

存在交互作用的数据分析

不存在交互作用的数据分析

解释交互作用

交互作用和理论检验

交互作用与外部效度

交互作用与天花板效应和地板效应

交互作用和自然组设计

小 结

概述

在第7和第8章中，我们主要介绍了单自变量效应的实验设计。研究者可以通过一组被试参加一种实验条件（独立组设计）或者每一被试均参与所有实验条件（重复测量设计）来研究自变量效应。¹由于我们希望读者先了解实验研究的基础知识，所以在上述两章中只是讨论了单自变量实验设计。然而，只有一个自变量的实验研究在当今的心理学的研究中并不普遍。取而代之的是，在一个实验中研究者更常采用两个或两个以上自变量的复合设计（complex design）。

我们这里以范巴尔恩及其同事（van Baaren, 2004）所设计的三个实验为例来说明复合设计。该研究的目的是探讨模仿对亲社会行为（如帮助他人）的促进作用。在每一个实验中，被试均单独接受一位主试的面谈。当被试描述其对10则广告的反应时，主试会模仿被试的姿势、身体朝向（如身体向前靠或朝后仰）以及手臂和腿的摆放位置。主试并没有告诉被试这个实验的真实目的，因为他们担心如果被试知道主试模仿其动作是想调查他们的帮助行为时，他们可能会改变他们自然表现出来的行为。为了了解模仿对亲社会行为的影响，你可以在与某人对话时小心地模仿这个人的行为。你注意到他们对你的行为或者态度有什么变化吗？

在第一个实验中，范巴尔恩等人通过随机组设计（参见第7章）只操纵了一个自变量。研究者想知道那些被模仿的被试是否在下一阶段的研究中会比那些没有被模仿的被试表现出更多的帮助行为。因此，自变量（模仿）具有两个水平：模仿与不模仿。一组被试被随机分配到模仿组而另外的则被分配到不模仿组。我们再看因变量的情况。主试假装不小心把六支笔掉到了地上并注意被试是否会帮忙把这些笔捡起来。结果表明，模仿组所有被试都弯腰把笔捡起来了，而不模仿组则只有33%的人这样做了（模仿效应达到统计显著性水平）。¹这个研究结果证实了模仿可以促进帮助行为的理论假设。

在第二个实验中，范巴尔恩等人想知道被模仿的被试是否会帮助其他人。采用相同的程序，还是以模仿作为自变量，他们进一步观察被试是否会捡起一位陌生女主试在进入实验室时掉下的笔。陌生女主试出现并不是一个操纵变量。在被试结束对广告的描述任务后，这位陌生主试进入实验室面对所有的被试。那些被模仿过的被试帮助捡起陌生女主试所掉下笔的可能性（84%）高于那些没有被模仿的被试（48%）。范巴尔恩等人得出结论，模仿可以增加人们的亲社会倾向。也就是说，那些被模仿过的人一般会变得更乐于帮助别人，而不只是帮助模仿过他/她的人。

在第三个实验中，范巴尔恩等人通过复合设计提高了研究的效率。他们同时考查了两个自变量。在他们的前两个实验中，他们间接地检验了被模仿之后提供帮助行为是否会在原来的主试和陌生主试之间表现出差异。第三个实验直接在一

¹ 第5章的“统计小贴士”指出， χ^2 检验（chi-square test of contingency）适合于检验称名变量（类别变量）。在范巴尔恩等人（2004）的前两个实验中，两个称名变量分别是模仿自变量（模仿与否）以及是否把笔捡起来。他们的结果表明，模仿和捡笔与不模仿和不捡笔之间存在强烈的联系。

个实验检验了这两个变量的作用。第一个自变量与实验 1 和 2 相同,就是模仿。被试被随机分配到模仿与不模仿条件下参加实验。第二个自变量是范巴尔恩等人通过操纵寻求帮助的主试。模仿条件与不模仿条件下的被试分别被随机分成两组,分别听原来的主试和陌生的主试叙述其请求。在完成模仿阶段的实验后,向被试提供一个机会,被试可以把他们参加实验所获报酬捐给一个儿童基金会。模仿组和不模仿组各有一半被试由原主试要求捐出报酬,而另一半被试则由一个陌生主试要求捐出报酬。因变量是所捐钱的数量,它代表被试愿意提供帮助的程度。

正如下表所示,这个实验有 4 个条件:(1)被模仿被试加原来的主试;(2)被模仿被试加陌生主试;(3)未被模仿被试加原来的主试;(4)未被模仿被试加陌生主试。

组别	提出捐赠请求的主试类型	
	原来的	陌生的
模仿组	1	2
不模仿组	3	4

在这个复合设计中,研究者针对全部自变量均采用了随机设计。也就是说,被试被随机分配到模仿与主试结合起来的各条件之中。复合设计因涉及到自变量的不同因子的组合也被称之为因子设计(factorial design)。因子组合是指把一个自变量的每一水平与另一自变量的每一水平进行配对。模仿组被试的帮助行为在原主试或陌生主试条件下经受检验;同样,不模仿组被试的帮助行为也在这两种条件下经受检验。在一个复合设计中,实验条件的数量等于实验中所有自变量水平数的乘积。在范巴尔恩等人(2004)的第三个实验中,组别具有两个水平(模仿组和不模仿组),而主试变量也有两个水平(原主试和陌生主试),因此,实验共有 4 个实验条件($2 \times 2 = 4$)。如果主试变量(原来和陌生)需要通过第三个模仿条件(如模仿面部表情)来测试,那么实验就会有 6 个条件($2 \times 3 = 6$)。

复合设计中因子组合可使我们确定每一自变量的主效应和各自变量水平相结合的文交互作用。例如,范巴尔恩等根据前两个实验的结果预期,那些被模仿的被试会比那些不被模仿的被试捐出更多的钱给儿童基金会。这代表的是模仿的主效应。也就是说,不考虑主试变量的两个水平(原来和陌生),被模仿的被试会比不被模仿的被试更愿意帮助别人。更为重要的是,研究者还可以直接检验这种帮助效应只适用于曾模仿过被试的主试还是可以迁移到陌生主试。比较所捐钱的数量可知,在原主试请求捐赠时,模仿所引起的帮助行为会比陌生主试要求捐赠时所引起的更大一些。也就是说,模仿会令被试提供给模仿者更多的帮助。²这种效应(即交互作用)只有在两个或更多自变量出现于同一实验中时才会出现。一般来说,当一个自变量的效应在另一自变量的不同水平上存在差异时,就表示出现了交互作用(interaction effect)。

² 尽管范巴尔恩等人(2004)的第三个实验中,组别和主试自变量之间存在一个交互作用,但他们的分析表明这个效应没有达到显著性水平。然而,这可能是因为他们所选取的样本数不足以检测到一个显著的交互作用。从第 7 章可知,这种错误是第二类错误。

至此,你可能会觉得复合设计看起来有些复杂。但是如果你往下看,概念会变得更清晰的。我们在一开始综述了复合实验设计的有关特点。接下来我们将介绍复合设计的有关步骤、分析手段以及怎样解释主效应和交互作用。我们还将介绍针对复合设计的数据分析。最后,我们将特别关注复合设计中交互作用的解释问题。

确定实验设计的指导原则

- 研究者运用复合设计研究一个实验中两个或更多自变量的效应。
- 在复合设计中,每个自变量都可以通过独立组设计或者重复测量设计而被研究。

一个复合设计实验具有两个或两个以上的自变量。在复合设计中,每一个自变量都可以通过独立组设计或重复测量设计(参见第7和第8章)研究。当一个复合设计既包含一个独立组变量也包含一个重复测量变量时,这种设计被称为混合设计(mixed design)。在讨论复合设计之前,我们需要简单回顾一下用于研究单一自变量的那些实验设计的特点。如果你能迅速而且准确地鉴别出自变量、它们的水平以及用来研究每个自变量的设计,那么你就能更好地理解对复合设计的这种阐释。同样,如果你能非常自信地选定针对你感兴趣的每一自变量最恰当的实验设计,那么你运用复合设计来完成研究的能力也会提高。

图9.1以一连串提问的形式总结了不同实验设计之间的区别。当你阅读课本或者论文中的某一实验时,你便会发现这个图是很有用的。当你在一个实验中确定每一自变量时,你可以根据图9.1的指引,决定哪种实验设计适合于研究你关注的自变量。在一个复合设计中,你应该穿过图9.1流程图,考量每一个自变量。图中最上面的一个问题是最关键的。问题“每个被试接受自变量的多少个水平?”的答案,将决定究竟采用独立组设计还是重复测量设计。接下来需要回答独立组设计中的有关问题,通过考查形成被试组的程序,你将能够确定一个特定的实验设计类型。对重复测量设计,我们提出的问题将帮助你区分完全和不完全重复测量设计,并且找出适用于每种类型的平衡技术。

描述复合设计中的实验效应

- 最简单的复合设计是 2×2 设计,即包含两个自变量,且每个自变量有两个水平。
- 在复合设计中,实验条件的数目等于每个自变量的水平数之积。
- 增加自变量的水平数或者包含更多的自变量个数可以使复合设计检验力更大,效率更高。

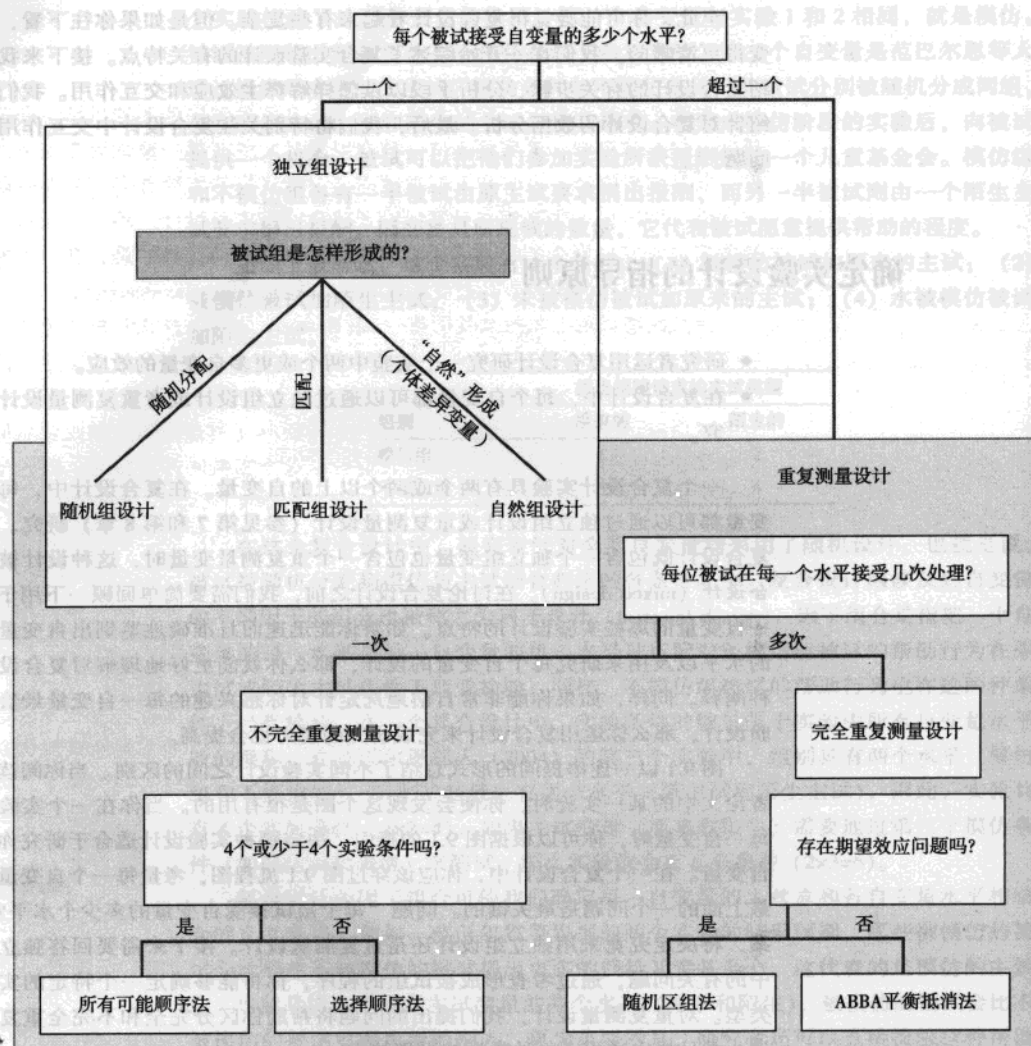


图 9.1 确定实验设计的流程图

一个最简单的实验设计必须具有一个两水平的自变量。同样，最简单的复合设计应该有两个自变量且每个自变量具有两个水平。通过确定实验中每一自变量的水平，你就可以知道复合设计的结构。 2×2 实验设计（读成“2 乘 2 实验设计”）就是一个最基本的复合设计。理论上，一个复合设计的总实验条件数可以是无穷大的，因为一个研究可以包含任意多个自变量，而每一自变量也可以取任意多个水平。然而，实际上一个实验包含 4 或 5 个以上自变量就已经不常见了。实验条件的总数目等于每一自变量的水平数之积。例如，范巴尔恩等人（2004）的第三个实验包含两个自变量，且每个自变量具有两个水平，因此这个 2×2 设计包含了 4

个实验条件。一个 3×3 实验设计包含两个自变量且每个自变量具有 3 个水平，因此共包含 9 个实验条件。一个 $3 \times 4 \times 2$ 实验设计包含三个自变量且分别具有 3、4 和 2 个水平，因此共包含 24 个实验条件。复合设计的首要优势就是它们可以使研究者观察自变量之间的交互作用。

2×2 设计是我们进一步了解复合设计的基础。然而， 2×2 设计还只是复合设计的一个最基本实验设计。复合设计可以通过两种方式在 2×2 设计之基础上扩充。研究者可以增加一个或者两个自变量的水平而产生 3×2 、 3×3 、 4×2 、 4×3 等复合设计。研究者也可以在同一实验中增加自变量的数目来扩充实验。每一自变量的水平数可从 2 扩展至某一不确定的上限。增加第 3 或者第 4 个自变量可以使实验变为 $2 \times 2 \times 2$ 、 $3 \times 3 \times 3$ 、 $2 \times 2 \times 4$ 、 $2 \times 3 \times 3 \times 2$ 等实验设计。

当实验中自变量的个数从 2 变为 3 时，复合设计的检验力和复杂性会有实质性的增加。两因素设计只能有一个交互作用，但在三因素设计中每一个自变量均可以与其他自变量出现交互作用以及三个自变量一起也可以出现交互作用。因此，把一个两因素设计变为三因素设计可以产生 4 种不同的交互作用。如果三个自变量以符号 A、B、C 表示，那么三因素实验设计可允许我们检验 A、B、C 的主效应； $A \times B$ 、 $A \times C$ 、 $B \times C$ 的二元交互作用；以及 $A \times B \times C$ 的三元交互作用。三个自变量实验的效率是很高的。复合设计在心理学研究中是非常有力的研究设计。专栏 9.1 给出了一个复合设计的例子。

2x2 设计举例

主效应和交互作用的特性在所有复合设计中都是一样的，但在一个 2×2 实验设计中，我们更容易看明白。这里以巴兹尼和谢弗 (Bazzini & Shaffer, 1999) 的研究为例。他们采用一个 2×2 实验设计来研究潜在约会对象的吸引力。很清楚，与我们生活在一起的人，特别是情人，其亲密关系可以说是我们生活中的重要部分。我们努力维持这种关系，而且当这种关系结束时，我们会感到悲伤与失落。

巴兹尼和谢弗研究了维持恋爱关系的一个重要方面，即处于恋爱中的个体对颇具吸引力的其他异性的反应。过去一些研究已经证明，恋爱中的人对具吸引力的其他异性的评分会低于那些不在恋爱中人的评分。这一结果被解释为恋爱中的个体比较负面地评价具吸引力的其他异性，这样做的目的是想维持与当前伴侣的关系。然而，巴兹尼和谢弗认为，另一种可能性是，不在恋爱中的个体会对一个具有吸引力的异性提高评分以寻求发展恋爱关系。巴兹尼和谢弗的目的就是研究关系维持（通过处于恋爱中的个体实现）和关系追求（通过目前未处于恋爱中的个体实现）两者中的哪一个能更好地解释对于有吸引力的潜在约会对象的评分。

在巴兹尼和谢弗 (1999) 的实验中，大学生被试首先完成一份关于当前恋爱状况的问卷。根据对这一问卷的反应，被试被分为“恋爱中的个体”（正处于恋爱中的个体）和“非恋爱中的个体”（未处于恋爱中的个体）。因此，第一个自变量是恋爱状况，具有两个水平，即恋爱中和非恋爱中。

另外，巴兹尼和谢弗要求被试阅读一个假想的情节，想像他们自己正与他/她最好的一位同性朋友坐在一家餐厅里，正好遇到“一位极富吸引力的异性”也坐

专栏 9.1

求职与肥胖

平吉托、杜戈尼、廷代尔和斯平 (Pingitore, Dugoni, Tindale, & Sping, 1994) 通过模拟招聘面试中对轻度肥胖人群的歧视现象。实验中, 被试先观看一段求职面试录像。在其中一个实验中, 他们采用了一个 $2 \times 2 \times 2$ 实验设计。第一个自变量是申请者的体重状况 (正常或者超重)。录像中工作申请者的角色由体重正常的专业演员扮演。在轻度肥胖条件下, 演员经过化妆并且穿着宽大的衣服, 他们看起来比平时要胖 20%。第二个自变量是申请者的性别 (男性或女性)。第三个自变量是被试对自己身体的关心程度 (高或低)。一个人对自己身体的关心程度被称为身体-图式变量 (body-schema variable)。身体图式是指被试对自己身体的满意度以及身体意识对自我概念的重要性。此实验的因变量是被试对是否聘用申请者所做的 7 点评分 (1 分: 绝对不会聘用; 7 分: 绝对会聘用)。实验采用一个自然组设计来研究身体-图式变量。被试被随机安排去评估体重正常或轻度超重的男性或女性申请者 (随机实验设计)。

作为描述平吉托等人 (1994) 实验结果的第一步, 我们将先考虑与申请者有关的两个自变量的交互作用: 申请者体重和申请者性别。对于正常体重的申请者来说, 被试给男性和女性的评分是非常接近的 (分别为 6.0 和 5.5)。然而, 对于超重的申请者, 被试的平均评分均降低了, 而且女性还会更低一些 (男女分别为 4.8 和 3.6)。这一交互作用模式表明女性超重者在求职面试过程中可能遭遇比男性超重者更多的歧视。

平吉托等人 (1994) 通过被试的身体-图式变量来进一步考查面试中的歧视现象。他们发现, 申请者体重和申请者性别的交互作用只有当被试高度关注他们自己的身体时才会出现。也就是说, 那些高度关注自己身体的被试给女性申请者尤其低的分数。当然, 那些不怎么关注自己身体的被试也会给超重申请者较低的分数, 但是他们对男女申请者的评分是接近的。

图 9.2 列出了平吉托等人 (1994) 采用这三个自变量的实验结果。正如你所看到的, 显示一个三变量实验的结果需要用两个图来表示; 其中一个图显示第三个变量某水平上其他两个自变量的结果, 而另一个图则显示在第三个变量另一个水平上其他两个自变量的结果。

我们可以这样总结图 9.2 中平吉托等人 (1994) 的实验结果, 即申请者体重和申请者性别之间的交互作用取决于被试的身体图式。我们把这个发现称之为三元交互作用。你可以看到, 当出现三元交互时, 我们在解释结果时必须考虑所有三个自变量的情况。总的来说, 当有两个自变量时, 交互作用表现为一个自变量的效应在另一自变量的不同水平上表现出差异。当在一个复合设计中有三个自变量时, 三元交互作用表现为两个自变量的交互作用在第三个自变量的不同水平上表现出差异。图 9.2 的结果很好地显示了这种关系, 即前两个自变量 (申请者的体重和性别) 的作用模式受第三个自变量 (被试的身体图式) 的影响。

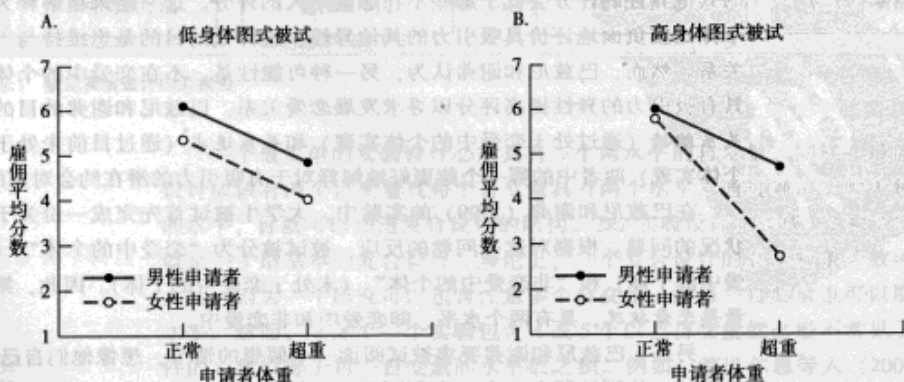


图 9.2 $2 \times 2 \times 2$ 复合设计的交互作用

表 9.1 包含交互作用的 2x2 设计举例

恋爱状况	情节类型		恋爱状况的平均数
	陌生异性对朋友表示好感	陌生异性对被试表示好感	
恋爱中的个体	9.77	9.25	9.52
非恋爱中的个体	10.00	11.13	10.54
情节类型的平均数	9.88	10.17	

引自 Bazzini 和 Shaffer (1999)

在那家餐厅里。被试被随机分配去阅读两个情节中的一个：这位异性对被试（所有被试均是异性恋倾向者）或者对他/她的朋友表示出好感。因此，巴兹尼和谢弗实验中的第二个自变量就是情节类型，也有两个水平，即该异性对被试表示好感或对他/她的朋友表示好感。

巴兹尼和谢弗推测，处于恋爱中的被试将会经历如下情节：这位异性对其表示好感会威胁到当前的恋爱关系。但是，如果这位异性对其朋友表示好感，则这位处于恋爱中的被试不会觉得其当前的恋爱关系受到了威胁。因此，处于恋爱中的被试在异性对其表示好感时，相对于异性对其朋友表示好感时，会对给予好感的异性给出更低的赞许性评分，以保护其当前的恋爱关系。相反地，巴兹尼和谢弗相信，那些非恋爱中的被试将会经历如下情节：异性对其表示好感实际上是发展一段新恋情的机会。当异性对其朋友表示好感时，被试就很难有这样的机会。因此，非恋爱中的被试在异性对其表示好感时，相对于异性对其朋友表示好感时，对给予好感的异性给出更高的赞许性评分。

巴兹尼和谢弗 (1999) 的研究采用了多个因变量，但我们将只重点介绍一个：被试对陌生异性所表现出的吸引力的评分。巴兹尼和谢弗把被试对三个项目（均是 5 点评分）的评分加起来，因此总分处于 3 至 15 分之间。表 9.1 给出了被试在 4 个条件下的平均评分。我们将利用此表的数据来阐述复合设计的两个关键概念：主效应和交互作用。

主效应和交互作用

- 在复合设计中，每一自变量的总效应被称为主效应。它表示一个自变量的每一水平在第二个变量水平上平均成绩的差异。
- 自变量之间的交互作用是指一个自变量的效应在第二个自变量的某些或全部水平上存在差异。

在任何复合设计实验中，我们可以在不考虑其他自变量效应的情况下，就每一自变量的总效应检验各假设。在复合设计中，一个自变量的总效应被称为该变量的主效应 (main effect)。例如，恋爱中的个体被试在实验中的平均数 (9.52) 是通过把异性对朋友表示好感以及异性对被试表示好感两种情节下被试对异性吸引力的评分求平均而得出的： $(9.77+9.25)/2=9.52$ 。同样地，非恋爱中的个体被试

拓展训练

在本练习中, 请你仅用描述统计来识别一个 2×2 实验设计中出现的主效应和交互作用。

本着熟能生巧的精神, 让我们把注意转移到练习上, 以帮助从 6 个复合实验设计中识别主效应和交互作用 (从 A 到 F)。在下面的每一表格和图中, 你需要确定每一自变量的效应是否在另一自变量的各水平上存在差异。也就是说, 变量之间是否存在交互作用? 在检查

完交互作用后, 你可以看看每一自变量是否产生了跨越另一自变量各水平的效应。也就是说, 一个或者两个自变量产生了主效应吗? 请你把表格 (图 9.3) 中的数据转化为图, 或者把图 (图 9.4 和 9.5) 中的数据转变成表格形式, 这种练习是很有帮助的。这个练习的目的就是让你能够懂得各种数据呈现形式。

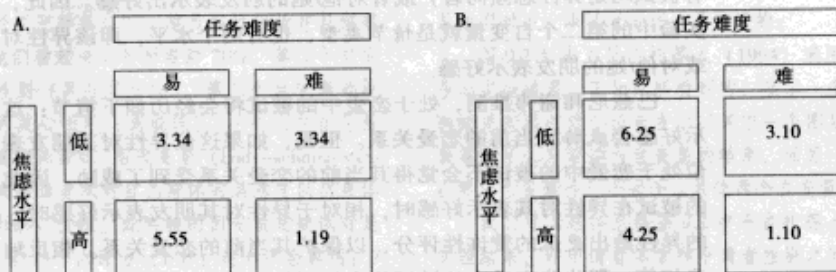


图 9.3 任务难度和焦虑水平共同影响正确反应平均数

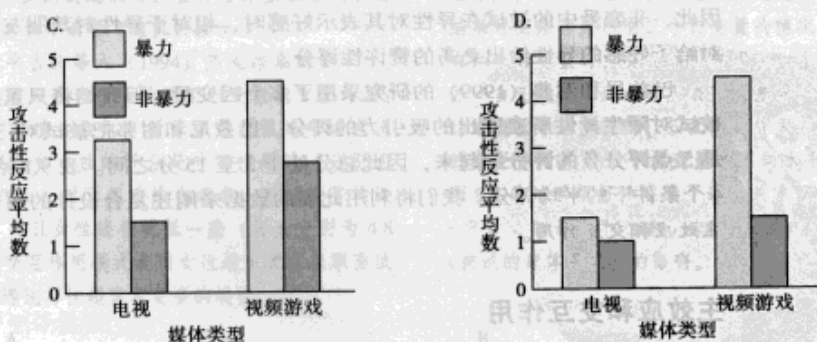


图 9.4 媒体类型和内容共同影响攻击性反应

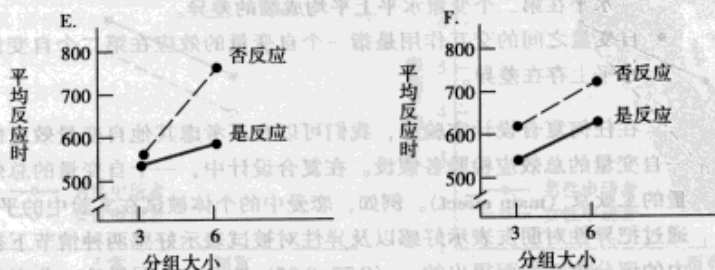


图 9.5 分组大小和反应类型共同影响平均反应时

在实验中的平均数是10.54。³某一主效应的均值表示某一自变量在每一水平上的总成绩（可通过自变量在这一水平下对第二个变量所有水平上各均数进行平均而得到）。在本例中，我们对情节变量两个水平上的成绩平均就可以得到恋爱状况主效应的均数。恋爱状况主效应等于恋爱状况变量两个水平上的均数之差（ $10.54 - 9.52 = 1.02$ ）。在巴兹尼和谢弗的实验中，恋爱状况变量的主效应表明，恋爱中的人（9.52）给陌生异性的吸引力评分会低于非恋爱中的人（10.54），且恋爱状况的主效应达到了显著性水平。但是，就像以前研究所发现的一样，我们可以把恋爱主效应解释为恋爱中的人降低对异性的评分以维持当前恋爱关系，或者解释为非恋爱中的人提高他们的评分以寻求发展恋爱关系。但是，我们并不能仅通过考查恋爱状况这一变量的主效应而确定支持哪种假说。

表 9.1 的一个最重要特征是这些数据反映了恋爱状况和情节类型这两个自变量之间存在交互作用。如果实验中只检验恋爱中的人这种情况（表 9.1 中第一行），那么我们会得出结论，当陌生异性对被试表示好感时获得的评分会比陌生异性对朋友表示好感时获得的评分低。另一方面，如果实验中只检验非恋爱中的人这种情况（表 9.1 中第二行），那么我们也可得出结论，当陌生异性对被试表示好感时获得的评分会比陌生异性对朋友表示好感时获得的评分高。这些结果表明，被试对陌生异性评分高低取决于情节类型。我们将在下一节“复合设计的数据分析”中介绍交互作用的统计分析。你现在只要能理解交互作用的意义就可以了，即交互作用是指一个自变量的效应大小在第二个自变量的不同水平上存在差异。

当第一个自变量与第二个自变量存在交互作用时，第二个自变量也一定与第一个自变量存在交互作用（也就是说，自变量的顺序并不重要）。例如，我们指出恋爱状况的主效应取决于情节类型而描述了表 9.1 中的交互作用现象。我们也可以反过来说：情节类型的主效应取决于被试的恋爱状况。当陌生异性对被试表示好感时，恋爱中的与非恋爱中的被试评分之差要比陌生异性对朋友表示好感时更大一些。

描述交互作用

- 用图（非平行线）或表（相减法）表示的描述统计可以帮助发现交互作用，但交互作用需要统计推论来确认。

你怎样描述一个交互作用取决于你要强调它的哪一方面。例如，巴兹尼和谢弗（Bazzini & Shaffer, 1999）强调情节类型对恋爱中的人和非恋爱中的人的不同影响，以检验基于关系维持和关系追求的理论假设。他们发现，恋爱中的被试在两个情节中对陌生异性的评分没有差异，但是非恋爱中的被试在陌生异性对被试表示好感时会比在陌生异性对朋友表示好感时的评分更高一些。巴兹尼和谢弗把这些发现解释为支持关系追求假说。这些发现也表明，复合设计出现交互作用可使研究者发现更多通过单变量实验不能发现的结果。

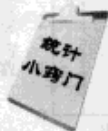
³ 只有当表中每一条件的被试数相同时，我们才能这样进行简单的平均以求出总均数。当各条件下被试数不同时，需要采用加权平均的计算方法（可参阅 Keppel, 1991）。

通常有三种方式来报告一个复合设计的描述统计结果：表、条形图和线形图。我们将在第14章详细介绍制作这些图和表的步骤以及标准的结果呈现形式。总的来说，表可用于任何形式的复合设计，而且在研究者知道实验中每一条件的具体数值时尤其有用。另一方面，在不强调具体数值的情况下，条形图和线形图则特别有助于显示结果的变化模式。线形图在描述复合设计的结果时特别有用，因为我们能非常容易地从线形图上看出交互作用。在线形图中，非平行线表明存在交互作用，而平行线表明不存在交互作用。

当一个 2×2 实验设计的结果用表格呈现出来后，通过相减法你就可以非常容易地评估是否存在交互作用。所谓相减法就是把表格中每一行（或每一列）的平均数相减。如果差值不同，那么数据中就可能存在交互作用。在运用相减法时，需要特别注意的是沿同一方向做减法运算。例如，在图9.1中，运用相减法，你可以把恋爱中的约会者在两个情节中的评分相减（ $9.77-9.25=.52$ ），然后按相同方法去计算非恋爱中的约会者的得分（ $10.00-11.13=-1.13$ ）。所得差值的正负符号也需要特别引起你的注意。相减法告诉你，这些差值是不同的，因而两个变量之间可能存在交互作用。

复合设计的数据分析

- 在两因素复合设计中，推论统计用于检验每一个自变量的主效应和两个自变量之间的交互作用。
- 描述统计对解释推论统计的结果是必要的。
- 研究者如何解释一个复合设计的结果取决于数据是否存在显著的交互作用。



复合设计的数据分析是在单自变量实验数据分析的基础上建立起来的（参见第7、12和13章）。数据分析的第一步就是用描述统计（如平均数、标准差和效应大小）来表述结果。虚无假设检验和置信区间这样的推论统计可以验证这些数据可以揭示什么现象。根据描述和推论统计的结果，研究者对他们的发现做出结论。复合设计的数据分析中，研究者常常采用方差分析（ F 检验）检验自变量的主效应以及它们之间的交互作用。在两因素实验设计中，我们可以检验三个效应：每一自变量可以各产生一个主效应，同时两个自变量结合可产生一个交互作用。

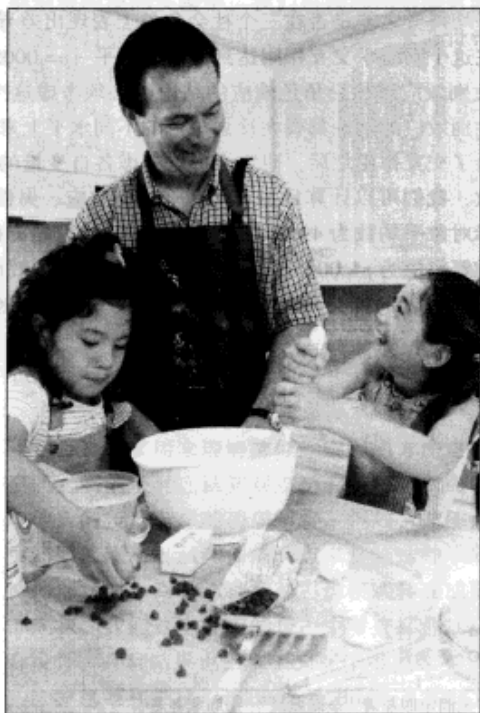
在本章接下来的内容里，读者的任务就是要掌握怎样对复合设计的数据进行分析处理，特别是研究者如何解释主效应和交互作用。你最好先阅读一下本节标题“复合设计的数据分析”之下的概要并且看一下第13章有关这一问题的讨论。这些章节所强调的均是关于数据分析的理论原则和逻辑，而不是关于数值计算的。幸运的是，计算机可以帮助我们对复合设计的数据结果进行充分的计算。但是，计算机并不能解释这些计算的结果。这就是你要留意之处。仔细阅读实验材料并观察所列举的图和表。

我们这里以关于他人的刻板印象研究为例来演示数据分析过程。伊格和斯

蒂芬 (Eagly & Steffen, 1984, p. 735) 描述了这样一种分别针对男人和女人的刻板印象: “女人比男人更强调共享 (即无私和关心别人), 但不强调果断和追求成功 (即自信和控制欲)。”在对大学生所进行的一系列实验研究中, 伊格尔和斯蒂芬检验了男女所扮演的社会角色是否会引起这种刻板印象。也就是说, 如果女性更有可能在家照看孩子而男性则更可能外出工作谋生, 那么男性就被认为更果断、更加追求成功, 而女性则更具共享的特征。伊格尔和斯蒂芬的实验结果支持他们的社会角色理论。

在本书一位作者的研究方法课上, 他试图重复伊格尔和斯蒂芬 (1984) 的实验。他推测, 伊格尔和斯蒂芬的结果可能不能重复, 因为时间过了几十年, 男女的社会角色已经发生了很大的改变。实验采用 2×3 的实验设计, 包括 6 组被试。实验者要求大学生对一个人 (目标对象) 进行简单描述, 并对其在 18 个特质上进行评分。第一个自变量是被描述者的性别: 男性或女性。第二个自变量是被描述者的社会角色: 无固定职业者、全职雇员和全职家务者。例如, 对一位男性雇员的描述是, “一位被公司全职雇佣的普通男性。”被试用来评估目标对象的测验是 18 个双极特质的形容词 (如 “完全没有攻击性” 作为量表的一极而 “非常具有攻击性” 作为另一极)。被试根据这两极在 5 点量表上对目标对象给出合适的评分。这样一来, 对目标对象的最低评分应该是 18 分 (即所有 18 个形容词均为 1 分) 而最高是 90 分 (即所有 18 个形容词均为 5 分)。得分高的目标对象表现为更具果

图 9.6 心理学研究显示, 关于男女社会角色的刻板印象影响我们对各行各业人员的看法



断和追求成功型的特征,而得分低的则表现为更具共享型特征。每一实验条件(共6个)下随机安排15名被试参加实验。

这个复合实验设计的结果见图9.7。图中的非平行线表明实验结果出现了交互作用。也就是说,男性和女性目标对象得分的差异会随社会角色的变化而变化。对于无固定职业者,女性目标对象被认为更不果断和更不追求成功。这一结果与伊格尔和斯蒂芬的研究结果类似,支持了人们对性别具有刻板印象的观点,即男性会被认为比女性更加果断和追求成功。然而,对于全职雇员,女性与男性目标得分相当,而且也均与男性在无固定职业者角色中的得分差不多(均比较果断、追求成功)。类似地,对于全职家务者,女性和男性的得分也差不多,而且均与女性无固定职业者角色中的得分差不多。从这三个结果可知,如果你被雇佣了你就被认为更具果断和追求成功(且更像一个男性),而如果你是做家务的,你就会被认为更不果断和更不追求成功(且更像一个女性)。

在这个实验中,我们可以发现三个可能的系统变异来源:目标对象的性别主效应、社会角色主效应以及目标对象性别和社会角色的交互作用。我们将在第13章分别采用虚无假设检验(和 F 检验)以及置信区间来分析复合设计的结果。我们这里重点介绍如何根据显著性检验的结果来解释主效应和交互作用。在复合设计以及其他设计中,一个统计显著性效应表明虚无假设为真的概率低于5%。

我们可把推论统计与描述统计结合起来来确定一个实验中是否出现了交互作用。检验交互作用是否存在实际上就是要确定一个自变量的效应是否因另一自变量的不同水平而不同。在性别刻板印象实验中,对交互作用进行检验将被用来确定男女的平均得分之差是否在三个社会角色上表现出差异。在进行显著性检验后,我们发现在这个实验中交互作用达到显著性水平($p=.0005$)。这样一来,任何关于目标对象性别效应和社会角色效应的结论都必须考虑这个交互作用。例如,任何关于对象性别效应的结论都要在社会角色的不同水平上来考虑。

在检验了交互作用之后,研究者可以考虑各自变量的主效应。根据图9.7所报告的平均数,我们可以计算目标对象性别的主效应。男性目标对象平均数为54.0,而女性目标对象平均数为49.0(不考虑社会角色)。经 F 检验,针对目标对象性别的主效应的概率值为 $p=.0005$ 。因此,目标对象性别主效应在统计上具有显著性意义。无固定职业者的社会角色主效应的平均数为52.5、全职雇员为57.5,而全职

图9.7 2×3复合设计的交互作用示例

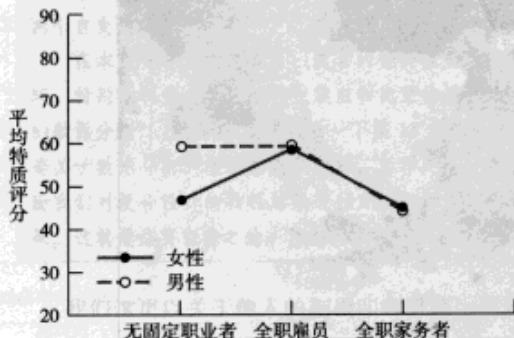


表 9.2 两因素实验数据分析方法

A×B 的交互作用显著吗?			
不显著		显著	
A 和 B 的主效应显著吗?		简单主效应显著吗?	
不显著	显著	不显著	显著
停止		停止	比较两均数差异
比较两均数差异		比较两均数差异	

家务者为44.5 (不考虑性别)。社会角色主效应在统计上具有显著性意义 (经 F 检验, $p=.0005$)。

根据对两个主效应的推论统计, 我们可知在这个实验中每个自变量均对被试的行为产生了一个显著性效应。然而, 我们在解释一个复合设计的主效应时需要特别谨慎。例如, 根据目标对象性别主效应具有统计显著性意义这一结果, 我们可以说男性被认为比女性更果断、更乐于追求成功。然而, 看一下图 9.7, 我们就知道这种情况只有在无固定职业者社会角色条件下是符合的。对于全职雇员和全职家务者, 男性和女性之间没有差别。总之, 当实验结果出现交互作用时, 我们在解释主效应时应谨慎。

在一个复合设计中, 我们也可采用其他方法来分析并解释数据。例如, 在性别刻板印象的实验中, 我们可以采用置信区间来检验社会角色的每一个水平上针对男女目标对象的平均数差异。针对复合设计的数据分析会因实验中出现显著交互作用而有所不同。

表 9.2 给出了变量间存在交互作用以及不存在交互作用时, 解释一个复合设计实验结果的有关步骤。我们将分别通过一个交互作用显著和交互作用不显著的实验来演示表 9.2 中所示的两条途径。

存在交互作用的数据分析

- 如果复合设计的结果存在一个显著的交互作用, 那么交互作用的意义可通过简单主效应分析和平均数两两比较来进一步确认。
- 简单主效应是指一个自变量在另一自变量某一个水平上的效应。

心理学家用来理解抑郁症成因的方法之一就是比较抑郁症患者和正常人思维模式之间的差异。认知观点研究抑郁症所关注的一个方面是归因, 即我们对自己或他人的行为所做的因果解释。例如, 抑郁症患者会表现出一种悲观、泛化和自责归因的倾向。从认知的观点来看, 这种不乐观的思维模式可能会引发抑郁并维持抑郁状态, 而将这种抑郁思维模式转变成更加乐观的思维模式, 有助于减轻抑郁症状。我们这里介绍的是一个考查抑郁症认知基础的典型研究。该研究采用复合设计, 并且结果也出现了交互作用。

罗德曼和伯格 (Rodman & Burger, 1985) 研究了一种特殊的现象——防御性

归因效应。在以前的实验研究中,要求不抑郁被试阅读一段关于事故的描述,每个被试阅读的事故中,伤者受的或是轻伤或是重伤。研究结果表明,相比于轻伤事故,不抑郁的个体在重伤事故条件下会把更多的事故责任归到肇事者身上。针对这种现象的一个解释是,人们不想把严重事故的原因归结为偶然性。如果人们把事故归结为偶然性,那么这些人就认为他们自己牵涉到一场严重事故之中的可能性会增加。人们把严重事故的责任更多归因于肇事者是一种防御性的归因,这可以使人觉得自己在将来被卷入严重事故中的概率很小。罗德曼和伯格推测,抑郁症患者可能会较少表现出这种防御性归因效应,因为他们消极的思维模式会使他们拥有更少的自我保护意识。罗德曼和伯格要检验如下假设:当一个人的抑郁水平增加时,其防御性归因效应将会减少。

罗德曼和伯格采用一个2×3实验设计测试了56位大学生。第一个自变量是所描述事故的严重性(严重与不严重),这个自变量采用随机组设计;第二个自变量是大学学生的抑郁水平(采用自然组设计)。根据大学生在一份抑郁量表上的得分,研究者选择了三种类型的被试:不抑郁、轻微抑郁和轻度抑郁。因变量是一个事故调查大问卷中的一个题目:要求大学生用百分数对四个可能的责任源分配责任比率,即事故中的三个涉事司机以及不可控的环境因素。在这个研究中,防御性归因效应表现为:被试在不严重事故条件下比严重事故条件下更多地把责任归因到不可控的环境因素(也就是偶然性)上。即被试防御性地避免将严重事故归因到不可控的环境因素上。表9.3列出了六种条件下这种不可控因素的平均百分比。

正如罗德曼和伯格所预期的那样,两个自变量之间存在交互作用:随着抑郁程度的增加,严重与不严重事故类型的归因结果出现了变化。(你可以用相减法来验证这一点。)不抑郁的大学生表现出防御性归因效应,而轻度抑郁的大学生则没有。研究者对这些数值进行虚无假设检验,表明存在一个显著性的交互作用。

一旦我们确定了两个自变量之间存在显著性的交互作用,那么我们还有必要更精确地分析交互作用的来源。正如表9.2所概括的那样,我们需要采用一些特定的统计方法来分析交互作用的来源。这些检验方法就是简单主效应和平均数的两两比较(参见第13章)。

简单主效应(simple main effect)是指一个自变量在另一自变量某一水平上的效应。我们可用罗德曼和伯格的实验结果来演示简单主效应的分析过程。表9.3显示了五种简单主效应:事故类型在三种抑郁水平上的效应和抑郁水平在两种事故类型上的效应。罗德曼和伯格预测,防御性归因效应(严重和不严重事故的平均

表 9.3 把责任归因为不可控的环境因素的百分比

事故类型	抑郁水平		
	不抑郁	轻微抑郁	轻度抑郁
严重	7.00 (9.2)	14.00 (16.1)	16.90 (16.0)
不严重	30.50 (22.2)	16.50 (12.7)	3.75 (3.5)

注:括号内数据为标准差。引自 Rodman 和 Burger(1985)。

数之差)会随抑郁水平的提高而减小。因此,他们决定检验事故类型在不同抑郁水平上的简单主效应。正如所预测的那样,他们发现事故类型在不抑郁水平上的简单主效应具有统计显著性意义,但在轻微和轻度抑郁水平上的简单主效应没有统计显著性意义。当在一个简单主效应中要比较三个或三个以上平均数两两之间的差异(如在每一事故水平上比较抑郁效应)时,平均数的两两比较可帮助分析简单主效应的来源(参见第13章)。

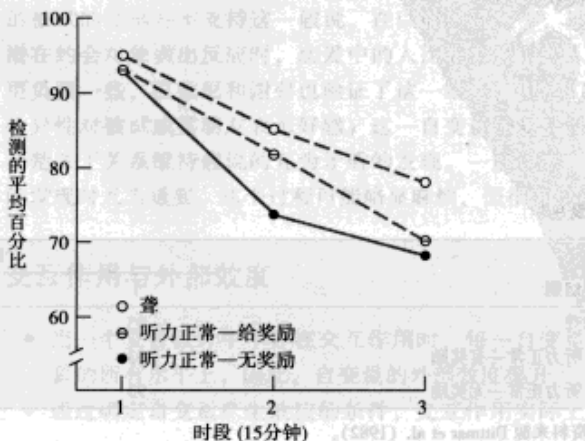
交互作用得到了充分的分析,研究者就可以考查每一自变量的主效应。然而,正如在本章开始时所提到的,当交互作用显著时,主效应就显得不那么重要了。例如,在上述实验中,显著的交互作用可告诉我们事故类型的主效应受到抑郁水平的影响。当我们了解到不严重事故导致把责任更多地归因到不可控的环境上时,实际上我们不会从这个主效应的分析中获得更多的东西。但是,在实际研究中也确实存在同时关注交互作用和主效应的情况。

不存在交互作用的数据分析

- 如果对复合设计结果的分析表明交互作用不显著,那么下一步就是确定主效应是否显著。
- 主效应的来源可通过平均数的两两比较,也可以运用置信区间比较两个平均数。

迪特马等人(Dittmar, Berch, & Warm, 1982)采用一个 3×3 实验设计来检验聋人在完成视警觉任务时是否比听力正常者更好一些。在一个视警觉任务中,被试需要在一段较长的时间里检测雷达屏幕上间断呈现的视觉信号(如亮点)。迪特马等预期,聋人在完成这种任务时具有一定的优势,因为耳聋可以帮助他们排除外界干扰。聋人组被试的成绩如果在组内是最好的就会予以奖励,研究者另外安排了两个听力正常组来确定奖励是否产生了激励作用。其中一个听力正常组被给

图 9.8 一个 3×3 复合实验设计的结果(无交互作用但存在两个主效应,引自迪特马等,1982)



予奖励而另一个组则没有。这三个被试组代表了第一个自变量的三个水平。所有三组被试均参加三段每段 15 分钟共 45 分钟的测试,而这三段时间是第二个自变量的三个水平。需要指出的是,这是一个混合设计,其中第一个自变量是一个独立组变量而第二个自变量是一个重复测量变量。

图 9.8 列出了三组被试在不同时间段的正确检测的百分比。我们可从中查看交互作用。尽管图中三条线段不是完全平行的,但三组被试的平均正确检测百分比似乎是以大致相同的斜率下降。推论统计分析的结果也进一步证实实验中不存在显著的交互作用。图 9.8 的数据说明了一个更为普遍的数据分析原则。描述统计所显示的数据模式不足以决定是否存在一个显著的交互作用。我们必须通过推论统计来确认数据是否存在显著的交互作用。

当一个实验中不存在显著的交互作用时,那么下一步就是检验每一自变量的主效应(参见表 9.2)。表 9.4 把迪特马等人(1982)的实验数据再呈现一遍以便更好地确定主效应。把三个时间段的数据合并起来求平均数,我们就能得到每一组的平均检测率(即组变量的主效应)。聋人、听力正常有奖励和听力正常无奖励三组的均数分别是 86.0、81.7 和 78.3,各组的主效应具有统计显著性意义。正如表 9.2 所表明的,统计显著的主效应可通过平均数的两两比较而进一步确定差异来源。这些比较可通过 t 检验或者置信区间来完成。例如,听力正常有奖励组和听力正常无奖励组成绩就没有显著性差异,这表明给予奖励并不会导致两组在成绩上出现差异。然而,两组的平均成绩(80.0)则显著低于聋人组(86.0)。这些比较可使我们得出两个结论:一是聋人与听力正常人相比,在视觉警示任务中表现更好;另一是给予奖励的方式并不会影响听力正常被试的视警觉任务成绩。

我们也可以利用表 9.4 中的数据检验时间变量的主效应。把三组被试的数据合并起来平均,我们就能得到时间变量的主效应,即 93.7(第一时段)、80.3(第二时段)和 72.0(第三时段)。时间变量的主效应具有统计显著性意义,并且直观可知平均成绩随着时间变化而呈下降趋势。平均数两两比较可用于确定这种连续下降(93.7 到 80.3 以及 80.3 到 72.0)是否具有统计显著性意义。对迪特马等人视警觉实验的数据分析表明,即使没有出现显著的交互作用,我们仍然可从复合设计中收获良多。

表 9.4



组别	时间段		
	1	2	3
聋	95	85	78
听力正常—有奖励	93	82	70
听力正常—无奖励	93	74	68

资料来源 Dittmar et al. (1982)。

解释交互作用

交互作用和理论检验

- 有些理论常常预期行为会受到两个或两个以上自变量的交互影响。因此，常常需要用复合设计来检验这样的理论。
- 理论检验过程中常会产生一些互相矛盾的发现，交互作用在解决这些矛盾时起了重要的作用。

理论在科学方法中起到非常关键的作用。复合设计也因可以同时检验主效应和交互作用而大大提升了研究者检验理论的能力。例如，巴兹尼和谢弗 (Bazzini & Shaffer, 1999) 关于潜在约会对象吸引力的假说就来自于一个人际关系理论。在他们的研究之前，关于潜在约会对象吸引力的研究倾向于支持关系维持假说。由于在研究中采用了复合设计，巴兹尼和谢弗就可以同时验证关系维持假说和关系追求假说。被试对陌生异性的吸引力评分为关系维持假说 (对恋爱中的人) 和关系追求假说 (对非恋爱中的人) 均提供了证据支持。巴兹尼和谢弗预期，恋爱中的被试在看待一位对其表示好感的有吸引力的异性时会更负面一些，以维持和保护其当前的恋爱关系。然而，与此预期相反，恋爱中的被试在陌生异性对自己和对其朋友表示好感时均给出了相近的评分。根据关系维持假说，恋爱中的人应该通过贬低一位有吸引力且可能带来威胁的异性以维持和保护其当前的恋爱关系。可是，他们的实验结果并不是这样，即不支持关系维持假说。

相反，从非恋爱中的被试获得的数据则支持关系追求假说。非恋爱中的被试在有吸引力的陌生异性对其表示好感时会比对其朋友表示好感时给出更高的评分。正如巴兹尼和谢弗 (1999) 所指出的，“被试倾向于提高对潜在约会对象的评价，似乎他们已被一段可能的恋爱关系迷住 (p.168)。”巴兹尼和谢弗的实验演示了怎样运用一个复合设计中出现的交互作用来更深刻地检验心理学理论。

巴兹尼和谢弗的实验也告诉我们复合设计中的交互作用可以帮助解决一些矛盾的发现。以前的研究支持关系维持假说，但巴兹尼和谢弗的实验中来自恋爱中的被试的证据并不支持这一假说。在以前的研究中，被试对一位向其表示好感的潜在约会对象做出反应时，恋爱中的人比非恋爱中的人对潜在约会对象的评分会更负面一些，巴兹尼和谢弗也验证了这一结果；但他们又通过一个情节类型 (陌生异性对被试或其朋友表示好感) 这一自变量的复合设计，从而使他们能够了解清楚关于关系维持假说的互为矛盾的发现。一般说来，复合设计在检验互为矛盾的发现时尤为重要。这个过程可能略显麻烦，但很有价值。

交互作用与外部效度

- 当一个复合设计中不存在交互作用时，每一自变量的效应可适用于另一自变量的所有水平上，因此，自变量的外部效度提升了。
- 通过确定自变量产生效应的条件，交互作用实际上为一个实验的外部效度界

定了范围。

在第7章中,我们花了一定的篇幅讨论了当实验只有一个自变量的情况下应如何去建立一个研究结论的外部效度。我们讲到怎样通过部分重复而建立外部效度(即研究结论的推广程度),我们还讨论了现场实验是怎样在现实中考查自变量的。我们现在再看看复合设计在外部效度的建立过程中所起的作用。正如你可能已经想到的,交互作用出现与否在确定一个复合设计的外部效度上起到关键作用。

当复合设计中没有出现交互作用时,我们知道每一自变量的效应可适用于另一自变量的所有水平上。例如,刚刚介绍过的迪特马的研究表明,检测成绩在三个时间段上逐步下降,但下降速度在聋人和听力正常人之间大致相同。也就是说,听力变量(聋和听力正常)和时间变量之间不存在交互作用。因此,在视知觉任务中被试检测成绩随时间推移而下降的结论同时适用于听力正常人和聋人。

当然,我们不能把我们的发现推广到实验条件之外去。例如,在视知觉任务中,听力变量和时间变量之间不存在交互作用并不能使我们得出结论说其他组(如儿童或训练有素的观察者)的检测成绩也将随时间推移而下降。同样地,我们也不能确定当三个时间段之间如果有一个短暂的休息时,这种成绩下降现象是否还会出现。需要注意的是,该实验没有发现一个显著的交互作用并不一定表明这个交互作用真的不存在,也许我们可能只是没有设计出一个足够灵敏的实验来发现这个效应而已。

正如我们所看到的,不出现交互作用会提升每一个自变量效应的外部效度。也许更重要的是,出现交互作用可以界定一个研究发现的外部效度的范围。例如,巴兹尼和谢弗发现,情节类型(陌生异性对被试表示好感或对其朋友表示好感)对恋爱中的被试在对潜在的约会对象评分时并没有什么影响。然而,非恋爱中的被试却在两个情节之中表现出差异。这一交互作用非常清楚地为情节类型评分这一效应设定了适用范围。根据这一结果,当人们问你情节类型对潜在约会对象评分的普遍结论是什么时,你应该回答,“它取决于其他条件”。在这个例子中,它取决于评分者是否正在谈恋爱。交互作用的出现为外部效度界定了范围,而且也告诉你这些范围是什么。

自变量之间存在交互作用使我们在判断某一变量对行为有没有影响上需要更为谨慎。那些对行为产生影响的自变量被称为**相关自变量**(relevant independent variable)。总的来说,相关自变量就是那些直接影响行为(产生主效应)或者当其与另一变量结合时产生交互作用的变量。区分那些影响和不影响行为的因素对建立充分的理论去解释行为,以及设计有效的干预措施去处理学校、医院和工厂等面临的问题都是十分重要的(参见第10章和第11章)。

几个原因促使我们在确认一个自变量为无关因素时应该慎重。第一,如果一个自变量在实验中主效应不显著,我们并不能认为如果另一个变量的不同水平被考虑的话,这个主效应还是不存在。第二,如果一个自变量在单自变量实验中主效应不显著,那么这并不意味着在一个复合设计中它将不会与另一个自变量产生交互作用。第三,一个效应没有达到统计显著性水平并不一定表明这个效应是不存在的。至少,在决定一个无关变量之前,我们都应该考虑我们所设计实验的灵

敏感度和统计分析的效力。(参见第13章关于统计分析效力的讨论。)就现在的情况来看,你最好不要轻易判定任何自变量是无关自变量。

交互作用与天花板效应和地板效应

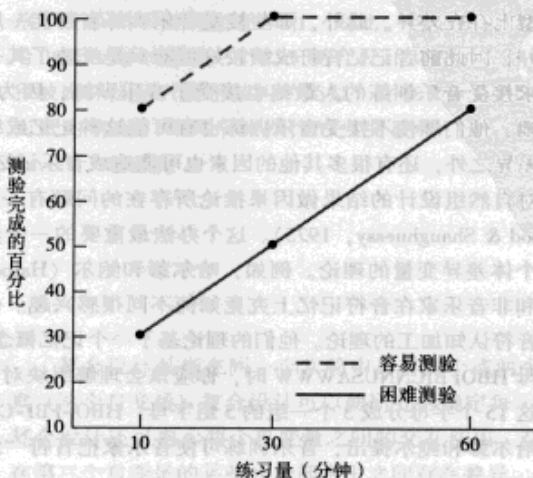
- 当被试成绩在实验的一个或多个条件达到最大值(天花板)或者最小值(地板)时,即使结果出现交互作用也是很难解释的。

这里以一个健身测验中练习量对成绩影响的 3×2 实验为例。这个可行但是假想的实验共包括6组被试,被试先用10、30或60分钟的时间来练习,练习任务分容易或困难两种,然后他们接受一个困难或容易的健康测试(任务与练习时的相同)。因变量是被试在15分钟时间内能够完成健康测试的百分比。实验结果见图9.9。

图9.9中的数据模式看起来像一个典型的交互作用;针对困难和容易两种测试,练习量效应存在显著差异。增加练习时间可以改善困难测验的成绩,但是在练习30分钟之后,容易测验的成绩就不增加了。如果对这些数据进行一个标准的统计分析,那么我们很有可能会发现存在一个显著的交互作用。不幸的是,这个交互作用却是特别难以解释的。对那些给予容易测验的组来说,30分钟练习后被试的成绩已经达到了最高点,因此在60分钟组被试的成绩不再显示出改善。即使那些给予60分钟练习的被试会从额外的练习中进一步获益,但就现有的因变量来看实验者也不能测量到这种改善。

上述实验所反映的这个常见的测量问题被称为天花板效应(ceiling effect)。只要一个实验的任一条件上的得分达到最大值,我们就要注意天花板效应。与此相对应,当得分达到最小值(如完成一个测验时正确率为零)时,我们就要注意地板效应(floor effect)。研究者在选择因变量时应使其值有足够变化的空间,以避免

图9.9 天花板效应示例



出现天花板效应和地板效应。例如，在上面的健身实验中，研究者把完成测验的次数作为因变量就会好些，平均完成次数可用来评估两个自变量的效应而且不用担心出现天花板效应了。需要指出的是，非复合设计的实验也有可能出现天花板效应。如果上面的健身实验只采用容易任务来测验，则还是会出现天花板效应。

交互作用和自然组设计

- 当研究者想知道为什么自然组之间得分存在差异时，他们会运用复合设计来对自然组变量做出因果推论。
- 对于自然组变量来说，做出因果推论需要三个步骤：解释为什么存在组间差异，操纵一个可以演示此推论过程的自变量，检验被操纵自变量和自然组变量之间是否存在交互作用。

在第7章中，我们简单介绍的自然组设计是心理学中最为常见的设计之一。研究者往往根据某一特征（如性别、年龄、内向、外向或攻击性）来区分不同被试，然后寻找这些个体差异变量与其他行为特征之间的系统性关系。自然组设计对建立个体特征与他们的成绩之间的关系非常有效。但是，正如第7章所提及的，当根据自然组设计的结果推论行为的原因时，这种设计实际上面临极大的挑战。

当我们试图作出成绩差异是由人们的某种特征所引起的结论时，这里实际上是有问题的。例如，某实验的被试是根据音乐训练情况来挑选的。其中一组被试的音乐训练时间为10年以上，而另一组被试则没有接受过正规的音乐训练。然后两组被试均被要求记忆一段含有10音符的旋律。结果表明，接受过音乐训练被试的成绩比没有接受过音乐训练组被试的成绩更好一些。

我们根据这些结果可以得出结论，对简单旋律的记忆随音乐训练量的变化而变化。但我们不能得出结论，音乐训练导致了更好的记忆成绩。为什么呢？接受过10年音乐训练者可能在很多其他的方面与没有接受音乐训练者不同。这两组被试可能在接受普通教育的类型和时间、家庭背景、社会经济地位以及听音乐的数量和类型上存在差异。此外，那些接受音乐训练者可能一般记忆力就比没有接受训练者好，因此前者记忆音符成绩较好可能只是反映了其一般记忆力好。最后，那些寻求接受音乐训练的人最终也接受了音乐训练，因为他们具有这方面的天赋。相应地，他们即使不接受音乐训练也有可能这种记忆成绩更好。总的来说，除了个体差异之外，还有很多其他的因素也可能造成音乐记忆成绩的差异。

对自然组设计的结果做因果推论所存在的问题有一个可能的解决办法（Underwood & Shaughnessy, 1975）。这个办法最重要的一个步骤是建立一个你所感兴趣的个体差异变量的理论。例如，哈尔彭和鲍尔（Halpern & Bower, 1982）对音乐家和非音乐家在音符记忆上究竟如何不同很感兴趣。他们建立了一个音乐训练影响音符认知加工的理论。他们的理论基于一个记忆概念即“组块”。在记下面15个字母HBOFBI CNNUSAWWW时，你应该会理解组块对记忆的帮助。组块可帮助你把这15个字母分成3个一组的5组字母：HBO-FBI-CNN-USA-WWW。

哈尔彭和鲍尔提出，音乐训练可使音乐家把音符“组块”成具有意义的音乐

单位,因而减少了重构一段简单旋律所需记忆的信息量。更进一步,如果这个过程可以解释音乐家和非音乐家在记忆音符上的差异,那么二者之间的差异在面对具有良好音乐结构的旋律时比面对结构不良的旋律时更大一些。哈尔彭和鲍尔操纵了音乐结构自变量来检验他们的理论。他们选用了三种不同旋律来测验音乐家和非音乐家。这三种不同旋律由非常相似的音符组成,但其音乐结构分别是好的、坏的以及随机分布的。

哈尔彭和鲍尔实验中的一个关键检验是,是否两个自变量(音乐训练和旋律类型)之间会出现交互作用。他们预期,音乐家和非音乐家在音符记忆上面对音乐结构好的旋律时差异会最大,其次是音乐结构坏的,在面对随机分布音乐结构旋律时的记忆成绩差异最小。实验结果与他们的预期完全一致。

哈尔彭和鲍尔根据交互作用的结果排除了许多其他的关于音乐家和非音乐家在记忆音符上存在差异的假说。接受普通教育的类型和时间、社会经济地位和家庭背景这些因素不太可能解释为什么旋律结构与音乐家和非音乐家在音符记忆上的差异之间会存在系统性关系。这些假说也不能解释为什么对于随机分布的旋律来说,音乐家和非音乐家在音符记忆上没有差异。这一交互作用使得这些简单的相关性解释不太合乎逻辑。

下面是对自然组设计做因果推论的有关步骤。

第一步:建立一个理论 第一步是建立一个理论来解释为什么组与组(根据个体差异变量来区分各个组)之间会出现得分上的差异。例如,哈尔彭和鲍尔就提出,音乐家和非音乐家在音符记忆上存在差异是由于这两种人在大脑中旋律组织(组块)的方式不同而引起的。

第二步:选定一个相关变量 第二步是选定一个可以操纵的而且被认为会影响这一理论过程的自变量。例如,哈尔彭和鲍尔认为音乐结构就是一个与组块效率有关的变量。

第三步:检验交互作用 这里所推荐的方法最关键的就是所操纵的自变量和个体差异变量之间出现交互作用。因此,这个相关自变量被应用到两个自然组之中。哈尔彭和鲍尔通过一个 2×3 复合设计发现了个体差异变量(音乐家和非音乐家)和被操纵变量(音乐结构的类型)存在交互作用。这个方法甚至也可用于三个自变量的情况:即两个被操作的自变量和一个个体差异变量(如参见 Anderson & Revelle, 1982)。

小结

复合设计是指在同一个实验中具有两个或两个以上自变量的设计。一个两因素(两个自变量)复合设计可以使研究者确定每一自变量的主效应。更重要的是,复合设计还可揭示两个自变量之间的交互作用。交互作用是指一个自变量的效应在第二个自变量的某些或全部水平之间存在差异。

最简单的复合设计是包含两个自变量且每个变量具有两个水平的 2×2 实验设计。在复合设计(因子设计)中,实验的条件数等于自变量的水平数之积(如 $2 \times 3 = 6$)。比 2×2 更为复杂一些的复合设计对了解行为可能更为有用。研究者通过增加一个或者两个自变量的水平数可产生 3×2 、 3×3 、 4×2 、 4×3 等复合设计。研究者也可以在同一实验中增加自变量而产生 $2 \times 2 \times 2$ 和 $2 \times 3 \times 3$ 等设计。包含 3 个自变量的实验,其效率会有实质性的增加。三因素实验设计可允许检验 3 个自变量各自的主效应、3 个二元交互作用和 1 个三元交互作用。

一个包含两个自变量的复合设计可以解释三种潜在的变异来源。每一自变量可产生一个显著性的主效应以及两个变量一起产生一个显著性的交互作用。交互作用可以通过相减法对数据表中的平均数进行处理,或者通过观察线形图中的线段是否平行而初步确定。如果交互作用被证明达到了显著性水平,那么我们可以进一步分析简单主效应,并在必要的时候再两两比较均数的大小。如果没有出现交互作用,我们就考查每一自变量的主效应,必要时我们也可以对平均数进行两两比较或者使用置信区间来分析数据。

复合设计在检验心理学理论上起到关键的作用。复合设计在解决理论矛盾时尤显重要。当采用复合设计并且结果没有出现交互作用时,我们知道每一自变量的效应可以推广到其他自变量的所有水平上。然而,当出现交互作用时,它可以帮助我们界定研究结论外部效度的适用范围。自变量之间存在交互作用时就要求我们扩展相关自变量的定义,以包括那些直接影响行为的变量(产生主效应)和那些与别的变量结合产生交互作用的变量。因实验结果中出现了天花板效应或地板效应等测量问题而引起的交互作用,不应与那些真正反映两个自变量的交互作用相混淆。

重要概念

复合设计	简单主效应
交互作用	相关自变量
主效应	天花板效应
地板效应	

复习思考题

1. 请列出下面复合设计中自变量的数目、每一自变量的水平数以及总的实验条件数: (1) 2×3 ; (2) 3×3 ; (3) $2 \times 2 \times 3$; (4) 4×3 。
2. 请列出包含下面自变量的复合设计的总实验条件数: (1) 任务类型具有 3 个水平(视觉、听觉和触觉)以及 (2) 儿童组别且具有两个水平(正常和发育迟滞)。
3. 运用巴兹尼和谢弗关于潜在约会对象吸引力的实验研究来演示下述问题: 在一个 2×2 的设计中存在一个交互作用,但可有两种方式描述这个交互作用。
4. 试述怎样用相减法从一个 2×2 设计的数据表中确定是否存在一个交互作用。

5. 试述从一个线形图中怎样判断存在交互作用。
6. 概述一个两因素复合设计中当交互作用存在时或者不存在时的数据处理步骤。
7. 请举例说明怎样运用一个复合设计来检验一个心理学理论的有关假说。
8. 在一个复合设计中, 实验结果的外部效度怎样受到交互作用的影响?
9. 试解释为什么研究者宣称一个变量对行为没有影响时应该慎重。
10. 试述天花板效应或地板效应的描述统计模式, 并解释为什么这种数据模式可能会影响推论统计对数据的解释。
11. 试解释在一个复合设计中怎样运用交互作用来对一个自然组设计的实验结果作出因果推论。

挑战性问题

1. 请设计一个双因素实验。其中变量A具有3个水平而变量B具有2个水平。
 - A. 请画图表示变量B存在主效应、变量A不存在主效应以及两个变量之间不存在交互作用。
 - B. 请画图表示变量A和B均不存在主效应, 但两个变量之间存在交互作用。
 - C. 请画图表示变量A和B均存在主效应, 但两个变量之间不存在交互作用。
2. 一个研究者采用复合设计研究训练(训练和不训练)和问题难度(难和易)对被试解决问题的影响。研究者共测试了80位被试, 共有4种实验条件, 每一条件随机分配20名被试。下表是被试解决问题的百分比。

问题难度	训练情况	
	训练	不练习
易	90	95
难	30	60

- A. 有证据表明该实验中存在交互作用吗?
 - B. 此实验所出现的什么情况会使你对待一个交互作用时持谨慎的态度?
 - C. 研究者通过对实验进行怎样的调整才能解释一个交互作用?
3. 一个心理学家想研究老年人在加工复杂视觉模式时反应时是否会变慢。50个65岁以上的老年人和50个大

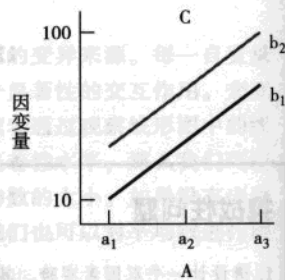
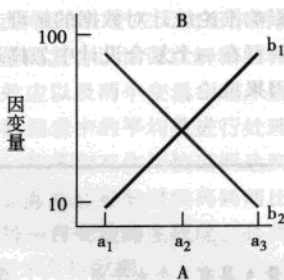
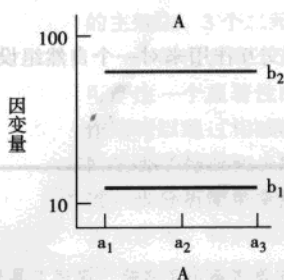
学生志愿者参加了实验。测试工具为镶嵌图形测验。实验者向被试呈现一个简单图形, 接着立即呈现一个包含这个简单图形的复杂图形。被试必须尽可能迅速地从这个复杂图形中指出这个简单图形来。被试的反应时从复杂图形出现开始记录直到找到简单图形为止。正如这位心理学家所预期的那样, 老年人的平均反应时比大学生明显要长(从任何标准来看, 结果均达到了显著性水平)。

- A. 这位心理学家根据这些结果宣称, 该实验中两组被试的反应时差异是因老年人加工复杂信息的能力出现了缺陷所引起的。你认为在得出这样的结论前, 这位心理学家应该采用一个复合设计的实验来证明这一点。你认为应该增加什么样的反应时测验来使他的实验变为复合设计? 请分别描述一个支持或不支持老年人加工复杂信息能力存在缺陷的结果。
 - B. 这位心理学家了解到其原来的设计存在问题, 因此他试图采用事后配对比法使各组被试之间尽量相等。他决定匹配健康状况(即越健康反应也越快)。尽管他在各组被试之间不能完全匹配, 但他确实发现最健康的15位老年人的数据反映出他们的平均反应时只比年轻人的稍微慢一点。请解释这位心理学家怎样根据这一结果改变年龄影响反应时的假说? 这位心理学家能不能做出这样的结论: 老年人在反应时上并不存在缺陷? 为什么?

拓展练习的答案

实验 A、D、E 中出现了交互作用, 实验 B、C、F 中没出现交互作用。

挑战性问题 1 的答案



概 述 第 10 章

我们在本章中，首先介绍应用研究的定义和范围。

然后，我们介绍应用研究的几种主要方法，包括实验法、调查法、访谈法、案例研究法等。最后，我们介绍应用研究的几个主要领域，包括心理学、教育学、医学、社会学等。

应用研究的定义和范围

应用研究是指将基础研究的成果应用于实际问题的研究。

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

应用研究的范围

第 四 编

应用研究

第 10 章
个案研究法和单被试设计

第 11 章
准实验设计和项目评估

第 10 章

个案研究法和单被试设计

本章大纲

概 述

个案研究法

个案研究法的特点

个案研究法的优点

个案研究法的缺陷

对个案研究证据的批判性思考

单被试 (小样本) 实验设计

单被试实验的主要特点

具体的实验设计

所有单被试设计共同的问题和局限性

小 结

概 述

我们在本书的前面几章一直强调基于群体的研究方法——用于检查一组或多组被试平均分数的方法，这在第7、8、9章中我们讨论实验方法时尤为明显。本章我们将介绍两种只使用一个被试的研究方法。我们称这类心理学研究设计为单个案研究设计。

自科学心理学诞生的19世纪起，单个案设计就已经被研究者们所使用了。心理物理法源自于费希纳的研究，在其1860年的著作《心理物理学原理》中就对该方法进行了介绍。费希纳及其后无数的心理物理学家都依赖于从一个或两个个案的实验中获得的數據来进行研究。在心理学的早期历史上，艾宾浩斯是使用单被试设计的另一个主要人物。事实上，艾宾浩斯所使用的单个案就是他自己。在他1885年发表的关于记忆的研究论文中，他本人既是被试又是主试。在长达几个月的时间里，他学习、复习并且尝试记忆了数百个无意义音节序列。他的实验数据为心理学家们提供了关于遗忘随时间变化的第一个系统性证据（参见 Zechmeister & Nyberg, 1982）。

在接下来的岁月里，认知心理学家们继续使用单个案实验设计开展有关研究。比如，林顿（Linton, 1978）在研究日常事件的记忆时，与艾宾浩斯一样，6年中一直用自己作为惟一的被试。那些关于专家行为的研究，无论是关于芭蕾舞演员、国际象棋选手，还是音乐家，也都大量采用小样本研究设计（如 Ericsson & Charness, 1994）。单个案研究的论文经常发表在各类心理学刊物上，一般来说，可以使用这种研究方法的领域跨度很广，从参与越战的退伍老兵的认知治疗（Kavany, 1997）到遗忘症患者的大脑机制研究（Gabrieli, Fleischman, Keane, Reminger, & Morrell, 1995）。

在这一章中，我们将讨论两类单个案研究方法：个案研究法和单被试实验设计。个案研究法经常与临床心理学家联系在一起，但其他领域的研究者也在使用这种重要的研究方法。例如，神经病学家奥利弗·萨克斯（Oliver Sacks, 1985, 1995）的个案研究就涉及一些罹患各种特殊神经疾病的个体。这些有趣的个案研究吸引了数以百万的人。萨克斯（1985）的代表作之一是《把妻子错当成帽子的男人》，正像书名所暗示的那样，书中描绘了这样一个男人，当他想要戴帽子的时候，他会抓住他妻子的脑袋并努力想拿起它，将其戴到自己头上。这些被萨克斯称为“临床故事”的个案，不仅可以使我们更深入地了解心理与大脑之间的关系，而且也揭示了遭遇严重的神经疾病时，个体是如何适应环境、应对困难以及成功生活的。

个案研究法常能够使研究者提出一些颇具价值的行为假说，此研究方法是对一些控制更为严格的行为研究方法的重要补充。在下文中，我们将综述个案研究法的优缺点。

我们在本章将要讨论的第二种单个案研究方法是单被试实验设计。我们使用单被试实验设计这个术语来标识这类心理学研究，是因为这种方法强调的是针对单个被试变量的操纵与解释。即使观察的是几个被试或者是一个单一“样组”，我

们也还是用同样的名称来称呼这类研究。¹ 单被试实验设计也常被称作“ $N=1$ 的实验设计”或“小样本实验设计”。这些设计具有行为的实验分析和应用行为分析这样的特点。你将会看到，这两种方法分别代表了小样本实验设计在基础和应用研究上的特点。单被试设计比个案研究法更具系统性而且控制更为严格。我们将探讨这些设计背后的原理，并具体介绍一些运用单被试实验设计的例子。这些实验设计其实是在第 8 章已经介绍过的重复测量设计的特例。

个案研究法

个案研究法的特点

- 个案研究是对个体行为进行详尽的描述和分析，没有小样本实验设计那么严格的控制。
- 个案研究是有正常和异常行为的理论假设的源泉。

个案研究 (case study) 是指对单一个体的行为进行详尽地描述与分析。个案研究常常使用定性的数据，但是也不尽然 (如 Smith, Harré, & Van Langenhove, 1995)。个案研究获得数据的方法可以多种多样，包括自然观察 (第 4 章)、访谈和心理测验 (第 5 章) 以及档案记录 (第 6 章) 等。个案研究常常会用来描述某一特别疗法的疗效与应用。比如，一份关于某一个案研究的临床报告可能会包括对某一症状的描述、诊断和治疗以及证明该疗法有效的证据。因此，个案研究能够提供关于该个体的大量信息。

然而，个案研究很少系统地控制实验变量。相反，不同的实验处理常被同时运用到同一个被试身上，而且心理学家对无关变量可能也极少控制 (如可能影响被试症状的家庭和工作环境等)。因此，个案研究的一个基本特征就是它们缺乏高度的控制。缺少了控制，研究者就很难对那些影响被试行为的变量 (包括任何处理手段) 做出有效的推论。控制程度是区分个案研究法与单被试实验设计的一个明显特征，单被试实验设计的控制程度更高 (参见 Kazdin, 1998)。

个案研究法已经被儿童心理学家、记忆研究者、动物行为研究者，以及人类学、犯罪学、神经病学、社会学等领域的研究者所广泛运用。在实际运用中，个案研究的形式与内容千差万别。已经发表的个案研究资料有的只有几页，有的可以写成一本书。例如，基尔希 (Kirsch, 1978) 报告了一个个案研究。该研究试图通过“自我管理训练”治疗一位不自信且存在社会性抑制的妇女。这一个案研究描述了 9 个疗程中的治疗情况以及长达 5 个月的事后追踪，研究结果发表在《心

¹ 在本书中，我们使用单被试研究设计而不是单个案研究设计 (single-case research design) 这一术语。这是因为前者更常用于描述单被试、少数几个被试或者单一组别的实验研究。此外，“单个案研究”这一术语也容易与个案研究法 (case study approach) 混淆。需要指出的是小样本研究 (small-n research) 强调的是样本小 (在统计学中小写 n 一般表示样本数)，而不只是一个小写字母 n 。

患者能成为他们自己的治疗师吗？一项个案研究实例

本文报告了自我管理训练 (self-management training, SMT) 的使用过程。该疗法的最大优点就是疗程简短, 同时又可避免有太多任务不能完成的情况……该方法的关键在于教会患者如何成为自己的治疗师。患者学会如何从不同维度评估其存在的问题, 并运用已有的治疗技术建立有针对性的策略来解决自己的问题。在这一过程中, 传统的患者与治疗师之间的关系模式就有了很大的改变。患者同时具有患者和治疗师的双重身份, 而治疗师则是以督导的身份出现了。

苏姗的个案

28 岁的已婚女士苏姗说她饱受记忆缺陷、智商不高和缺乏自信的折磨, 她认为这些缺陷“导致”她无法参加很多社会活动。她不能参与任何一个关于电影、戏剧、书籍或者杂志文章的讨论, 因为她不能很好地回忆起它们的内容。她常感到她理解别人的谈话, 并且认为这是由于自己智商低所造成的。她在这些社会活动中常采取一种消极的姿态, 以此来掩饰自己在理解能力上的缺陷。她不相信自己的意见, 而且有时甚至会怀疑自己能否产生任何见解。她感觉自己非常依赖他人, 总是采纳和执行他人的意见。

我用韦氏成人智力量表对她进行了测量, 发现她的言语智商大约有 120, 这绝不是一个低于正常的分数, 她的数字广度记忆分数 (量表分 5=12, 原始分 5=13) 表明至少她的短时记忆不会是有缺陷的。这些测验的结果证实了我在与她面谈时的推测: 她的智力与记忆都没有任何问题。在和她讨论了这个结论以后, 我建议我们一起思索当她感觉自己的智商、记忆力以及自信心都非常高之后, 她可以完成哪些目标。通过这种方法, 我们就能制订出一个行为目标列表, 包括“陈述自己的主张”、“要求对方再次解释, 以及承认自己忽略了某些事实”等。在整个治疗过程中, 我引导苏姗出声或不出声地复述那些令她焦虑的情境……同时安排她完成一些结构化的家庭作业使她逐步接近她的行为目标, 并且让她记录下她的进步。此外, 我与她还讨论了她的那些无事实根据的消极自我评价 (如“自己很愚蠢”等。) 我建议她一旦发觉自己产生这种消极的想法, 要马上有意识地对自己说一些更合适、更积极的自我评价 (如, “我一点也不蠢, 认为自己愚蠢是没有道理的”等)。

在第五个疗程中, 苏姗报告说她已经成功地完成了那些家庭作业。她不仅觉得很容易完成, 而且报告说在完成作业的过程中, 即使是第一次尝试, 她也没有任何的焦虑感……从这一刻开始, 治疗关系就已经完全改变了。在其后的疗程中, 苏姗自己评价她每周的进步, 决定她下一步应该做什么, 而且自己设计应该完成的家庭作业。我的角色就随之变成了一个学生治疗师的督导了, 一方面巩固她已取得的成绩, 另一方面时时提醒她不要忽略一些因素。

第九个疗程之后, 直接的治疗就停止了。在接下来的一个月里, 我通过电话联系过苏姗两次。她报告说她对自己的能力充满信心, 相信自己一定能达到预定目标。她特别提到自己有了一种能够掌握自己生活的全新感觉。我的感觉是, 她已经成功地采用了一种问题解决式的行为评价法, 并且已经可以相当熟练地设计策略来达到自己的目标。

追踪研究

在结束治疗的五个月后, 我联系苏姗并询问她进步的情况。她报告说她在社交场合中比以前说得更多了, 而且对自己独立处理事情 (如不依赖于她的丈夫等) 也更有信心了。总的说来, 她再也不觉得自己很愚蠢了。她总结道, “我感觉自己正处在一个比以往更高的境界上。”

我也问她, 在治疗过程中使用过的哪些技巧到现在仍然被她使用……最后, 她报告说, 在治疗结束后的五个月中至少在三个不同的场合下, 她告诉另一个人, “我不明白你刚才所说的, 你能否再解释一下?” 这样的提问是她以前感到无法做出的。在治疗前, 她认为这样做会让自己显得很蠢。

在追踪访谈结束三个月后, 我意外地收到了一封来自苏姗的信 (那时我已基本上不关心此事了)。信中她告诉我, 她曾使用过的一个假想练习是, 走进一个民族舞蹈培训班而且还感觉良好。“太好了, 它终于起作用了”她在信中这样写道。

资料来源: Kirsch, I. (1978). Teaching clients to be their own therapists: A case study illustration. *Psychotherapy: Theory, Research, and Practice*, 15, 302-305. (Reprinted by permission.)

理治疗；理论、研究与实践》杂志上，仅用了4页篇幅。在另一个个案研究中，人格心理学家戈登·奥尔波特通过分析一位母亲写给其朋友的301封信件，描述了这位母亲与儿子的关系。该文先刊登于连续两期《变态与社会心理学》杂志（1946年第41卷第3和第4期）上，而后又被作为经典个案研究于1963年写成了一本厚223页、名为《珍妮的来信》的书。

露丝·坎贝尔（Ruth Campbell, 1992）编写的一本名为《精神生活》的书，包含了对15个个体的个案研究。这些个体在思维、知觉、阅读、语言以及记忆方面存在异常（p. 2）。然而，这本书所关注的是如何通过这些个案来揭示正常认知功能的机制。例如，里多克和汉弗莱斯（Riddoch & Humphreys, 1992）报告了一位47岁的男性丹尼斯，他从爬梯上摔了下来，脑部严重创伤。从那以后，尽管丹尼斯有足够好的视力看电视，但他看到熟悉的人（他的家人）或熟悉的物体时，却无法辨认。然而，当丹尼斯听到熟人说话时，却能辨认出他们；当他触摸熟悉的物体时，也能辨认出这些物体。该个案的研究想通过丹尼斯的经历告诉我们：视觉辨认依赖于贮存的知识。坎贝尔强调说，《精神生活》这本书里的个案研究有别于其他的个案研究的地方，就是我们可以通过那些罕见的异常行为个案来了解人类正常的行为。而奥利弗·萨克斯的个案研究就只对那些患有罕见大脑疾病的个体进行细致生动的描述。由此可见，个案研究可以服务于许多的重要目的。

个案研究法的许多特点使它成为一种独特的行为研究方法。它在研究目的、所运用的方法以及所获信息类型上都与实验方法有所区别（Kazdin, 1998）。例如，个案研究法常被认为在本质上是“探索性”的，是有关行为的理论假设和观点的源泉（Bolgar, 1965）。另一方面，实验法常被看作是检验具体假设的手段，而个案研究法有时被认为是避免做更多控制的一种调查研究方法。但是卡兹丁（Kazdin, 1998）提出了一个更为恰当的观点，即个案研究法与心理学的其他研究方法是紧密联系的，并且是其他方法的一种补充。

对于心理学研究者来说，个案研究法既有其优点，同时也会存在一些局限性（参见Bolgar, 1965；Hersen & Barlow, 1976；Kazdin, 1998）。在我们评论这些优缺点之前，我们将先以一个前文提到过的个案研究（Kirsch, 1978）为例来介绍这种方法。请你务必仔细阅读这篇缩减版的个案研究报告，因为我们在后面讨论个案研究法的优点与局限时还要不时地来回顾这篇文章。

个案研究法的优点

- 个案研究有利于研究者提出新的观念和假设，能够为新的治疗技术和罕见现象的研究提供机会。
- 当单个被试的行为与理论原则或观点相悖时，说明这些科学理论受到了挑战；同时，理论也可从来自个案研究的证据中获得一些初步的支持。
- 个人特质研究（研究个体的独特性）是对常规研究（研究群体的特性）的一个补充。

行为假设的来源 个案研究为理解个体和洞察人们行为的原因提供了丰富的信息。

这些内容转变为研究假设后就可以接受控制更为严格的实验方法的检验。基尔希(Kirsch, 1978)在讨论对苏珊进行的成功心理治疗这一案例时,也认识到了个案研究的这一特点。他写道:“(该个案研究的)结论……应该被看作是暂时性的。我希望(这种治疗技术的)效用将被控制更为严格的研究所证实(p.305)。”当研究者进入一个几乎全新的领域时,个案研究是一个很自然的开始。

临床上创新的机会 个案研究法为试验新的治疗方法以及发展现有方法的一些独特用途提供了机会。自我管理训练技术(SMT)在心理治疗中的运用代表了一种临床上的创新,因为基尔希由此改变了患者—治疗师之间的传统关系。SMT技术的核心是使患者成为他们自己的治疗师。也就是说,患者找出问题并自己设计解决问题的办法。患者既是患者又是治疗师,而治疗师则变成了一名督导。类似地,库班(Kubany, 1997)报告了持续1天的“马拉松”认知治疗的效果:对一名遭受战争罪恶感折磨的越战老兵进行治疗。这类疗法通常在经过了多次一小时左右的治疗之后才会有效,但是这一在一天内完成的紧凑治疗法也取得了一定成功。这为我们提供了一种针对这类问题的新的临床治疗途径。

研究罕见现象的方法 个案研究也有助于研究一些罕见的事件。一些事件在自然界中太少见以至于我们不得不通过对一个案例的大量研究来了解它们。例如,在《精神生活》以及奥利弗·萨克斯的有关著作中,大多数个案研究都是描述那些患有罕见大脑缺陷疾病的个体。有一个例子就是关于“野孩”的个案研究。这些孩子小时候被人遗弃,并在缺乏足够人类交流的野生环境中长大。其中最著名的是“亚维农野男孩”[参见沙特克(Shattuck, 1994)《法律不允许的实验》:亚维农野男孩的故事]。那个后来被叫做维克多的男孩是1800年在法国亚维农发现的。在他5岁时,有人可能想杀了他(从他脖子上的刀伤可以推测),然后把他丢弃在森林里(可能是想让他自己死去)。当他被发现时,他大概有11或12岁大了。关于他的描述是这样的:

他体态像人,并且能直立行走,但是所有其他方面都表现得像一个动物。除了身上的那件破烂不堪的衬衣,他几乎是全裸的,他一点也不感到羞耻。他丝毫没有意识到自己是一个人,没有意识到自己与那些发现他的人属于同类。他不会讲话,只会发出些毫无意义的怪叫。(Shattuck, 1994, p.5)

对于很多人来说,维克多的出现似乎对回答当时关于人性的一些热点问题(现在仍是)很有意义。我们与动物有什么区别?人类怎么学会语言的?什么是“本性”以及什么是“文化”?维克多在一位热心医生和保育员的监护下,重新走入人类社会并逐步“人性化”的故事非常引人注目。但是,这个故事并没有给那些目前仍是热点的问题提供任何确定的答案。正像一位作者所说的那样,这个故事“迫使我们反思怎样与这些尚未解决的问题共存”(Shattuck, 1994, p.182)。这也许就是个案研究的力量吧。

对理论假说的挑战 假如有人确实看到了长着两个脑袋的火星星人,那么“所有火星人都长着三个脑袋”的理论就会不攻自破了。个案研究常可以通过提供“反例”的方式来促使我们对有关科学问题进行思考。所谓反例是指违反一般命题和普遍

图 10.1 仅存的亚维农野男孩肖像



原则的事例 (Kazdin, 1998)。例如, 让我们考虑一下这样一个理论: 人类语言发展取决于从 2 岁到青春期这一关键期内是否暴露于正常的语言环境之中 (Lenneberg, 1967)。因此, 如果一个儿童在这一关键期被剥夺了语言环境, 那么这个孩子将无法掌握语言。当然, 为了验证这个理论而剥夺一个儿童的语言学习机会是很不人道的。然而, “天然实验”却是存在的。这样一个“实验”涉及一名叫“吉尼”的儿童。吉尼于 1970 年被人发现。从 2 岁到 13 岁间她被残忍地剥夺了绝大多数与人接触和进行语言交流的机会 (Curtiss, 1977)。赖默 (Rymer, 1993) 根据时间顺序记录了吉尼被监禁和虐待的环境以及心理学家和语言学家试图令她康复的有关尝试。吉尼的经历为科学家们验证语言发展的关键期理论提供了一次难得的机会:

吉尼也许顶多能为伦伯格的理论提供一个有瑕疵的证明, 但是她的例子却敲响了一个反驳的警钟。如果吉尼不能够学习语言, 则其失败的原因还是含糊不清的——也许是她错过了学习语言的关键期, 也许她有一些情绪问题。但是, 如果吉尼在经历了这所有的一切之后仍能够学习语言, 可想而知, 这一反驳将是多么有力! (Rymer, 1993, pp. 121-122)

根据参与吉尼语言训练的那些心理学家的报告, 吉尼在语言上显示了一些发展, 但是这种语言发展并不是完全正常的。这一结果挑战了伦伯格理论的“强”版本, 即如果缺少关键期的语言经历, 人类完全不能习得语言。但是关于吉尼的个案研究却无法反驳伦伯格理论的“弱”版本, 即如果没有关键期的语言经历,

青春期后人类将不可能有正常的语言发展 (Curtiss, 1977, p.209)。

为心理学理论提供初步的支持 尽管个案研究并不能为某一假说提供决定性的证据,但是其结果却可以为某一心理学理论提供初步的支持。吉尼的事例就是一个很好的说明。这个个案研究为伦伯格的**语言学习关键期理论**的“弱”版本提供了重要支持。

我们这里以记忆研究为例,介绍个案研究能够为理论提供支持证据的观点。阿特金森和希夫瑞 (Atkinson & Shiffrin, 1968) 提出了一个对记忆研究领域影响深远的人类记忆模型。这个基于信息加工原理的模型包括一个短时记忆系统和一个长时记忆系统。尽管大量实验结果都为这个二分的人类记忆系统模型提供了支持证据,但是阿特金森和希夫瑞还是认为几个个案研究的结果“可能是人类记忆系统具有二分性最令人信服的证据”(p.97)。这些个案研究涉及一些因治疗癫痫而被手术切除了包括皮下游海马在内的部分颞叶脑区的患者。对阿特金森和希夫瑞的理论有特别重要意义的个案研究来自对一个叫 H.M. 的患者的研究 (参见 Hilts, 1995; Scoville & Milner, 1957)。在进行脑部手术后, H.M. 被发觉患有记忆障碍。尽管他能够进行正常的交谈并在短时间内记住事情,但 H.M. 无法回忆每一天发生的事情。他可以一遍又一遍地阅读同一本杂志,但事后还是不记得其中的内容。H.M. 似乎有一个完好的短时记忆系统,但是无法将信息输入长时记忆系统。尽管随后对 H.M. 以及与他有类似记忆缺损的患者所进行的测验揭示出记忆问题的本质远比原先想像的要复杂得多,但是只要人们还在讨论有关人类记忆的理论, H.M. 的个案研究就依然是十分重要的 (参见 Schacter, 1996)。

为行为的常规研究方法提供补充 心理学 (与一般意义上的科学一样) 寻求建立更具普遍意义的理论,即能够适用于更广泛有机体的“普遍规律”。因此,心理学研究就常常是以常规研究为其主要特征。**常规研究法** (nomothetic approach) 要求有大量的被试参加,并且寻找一个群体的“平均”或典型的反应。这个平均反应可能代表也可能不代表群体中的每一个人。然而,研究者希望能够从这个平均反应中预测出有机体在一般情况下可能做出的反应。

一些心理学家,其中最著名的是戈登·奥尔波特,认为仅有常规研究是不够的,个体在很多方面都远不是平均值可以代表的。奥尔波特认为个体是既独特又富有规律性的:个体的运算是遵循一些内部一致的规则的。奥尔波特更进一步认为,对个体的研究即**个人特质研究法** (idiographic approach) 也同样是心理学研究的一个重要组成部分 (参见 Smith et al., 1995)。

奥尔波特通过描述临床心理学家所面临的**任务**,说明了心理学研究对个人特质研究法的需要。临床心理学家的目的“并不是对总体进行预测,而是预知‘某一个人将做什么’。为了达到这一目标,精确的预测有时是有益的。普遍的或总体平均虽然有用,但是它们并不适用于其中的每一个人” (p.21)。奥尔波特建议我们在研究中不能偏爱常规研究和个人特质研究的其中任何一个,而要平衡地使用这两种研究方法。由个案研究代表的个人特质研究法至少可以从细致的观察中揭示出行为间的细微差别,而这些则是“群体法”常规研究可能会忽视的地方。而且,正如上面的例子所揭示的那样,个案研究有时还可以通过对反常个体的细致

拓展练习

在这个练习中，要求你在阅读完以下简短的描述后回答问题。

你的一位朋友这学期选修了一门心理学导论课。她在午餐时向你描述了她对当天上午课堂上所发生事情的反应。那堂课的主题是成人发展。教授向同学们讲述了两个有关结婚和离婚的研究。教授强调说，两个研究均是由该领域的著名学者所完成的优秀研究。在第一个研究中，研究者从一个精确定义的总体中随机选出许多对已婚夫妇进行调查。该研究的结果表明，被试中大约超过一半的婚姻都以离婚而告终，其中配偶间长时间的冲突和家族离婚史等因素都是离婚的可靠预测源。教授着重指出，统计检验也证明了这些预测因子的可靠性。第二个研究描述了关于某一对夫妇婚姻和家庭咨询的过程。这个个案研究描述了这对夫妇怎样从刚开始打算离婚到经过一年的治疗，最终决定坚守婚姻的过程。教授介绍了婚姻和家庭咨询师在治疗过程中所使用的各种技术，这些技术成功地帮助了这对夫妇正确认识并处理各

种冲突和问题。

教授还没来得及告诉大家这两个研究之间有怎样的联系，以及从这两个研究中能够得出什么样的结论，就到下课时间了。你将如何回应你朋友关于这堂课的问题呢？

1. 你朋友的一个问题是：她应该相信哪个研究的结论？第一个研究似乎在说婚姻中的冲突和家族离婚史是导致离婚的主要原因，然而第二个研究表明这些因素并不一定导致离婚。你的朋友说她比较倾向于相信第二个研究的结果，因为她认为第二个研究中个人化的例子相比第一个研究中那些数字更引人注目一些。你如何看待这个问题？
2. 你朋友还问到这些研究对她自己的经历有什么样的意义。也就是说，她是否可以根据这些研究的结果来预测她的婚姻状况。你如何看待这个问题？

观察，能够告知我们典型的和一般的行为模式。

个案研究法的缺陷

- 研究者不能根据个案研究的结果进行有效的因果推论，因为个案研究没有对无关变量进行控制，并且可能有好几个“处理”被同时使用到同一个被试身上（会引起混淆效应）。
- 观察者偏差和在搜集数据时的偏差可能会导致对个案研究结果的不正确解释。
- 一个个案研究的结论能否被推广取决于个案所属总体的变异情况；一些特征（如人格）要比另外一些特征（如视敏度）在个体间的变异更大。

无法做因果推论 科学研究的目标之一就是探求现象产生的原因——确定引发某一事件的所有因素。个案研究法的一个缺陷就在于从它的结果里几乎不能得到有关因果关系的结论。这一缺陷主要是因为研究者不能控制个案研究中的无关变量。因此，在所观察个体身上出现的行为改变也可以有其他的解释。

这里以基尔希（1978）用自我管理训练（SMT）技术治疗苏姗为例来说明这个问题。尽管苏姗很明显地从 SMT 治疗中获益了，但我们就可以明确说是 SMT 导致了苏姗的进步吗？很多疾病和情绪紊乱没经过治疗也会有所好转。运用个案研究法的研究者必须时刻考虑到，即使不给予任何处理，患者情况也可能会有改善的可能性。此外，某些情境也可能使得苏姗情况好转。临床心理学家的身份可

能使苏珊感到放心,治疗师的洞察力和反馈以及心理测验的结果也可能使她改变对自己的态度。作为治疗的一部分,治疗师要求苏珊出声或不出声地复述令她焦虑的情景,这一技术与脱敏技术很相似,而脱敏本身也是一种有效的治疗技术(Rimm & Masters, 1979)。

由于很多治疗方法都同时应用到了苏珊身上,所以我们无法肯定地说SMT是造成苏珊情况改善的惟一原因。正如我们所看到的,基尔希本人也意识到了个案研究法的局限性。他建议在采用更为严格的检验之前,从这个案例结果中得出的推论只能被看作是初步的和暂时性的。

那些关于与维克多一样的野孩以及吉尼一样被虐待儿童的个案研究,同样存在无法认定因果关系的问题。我们无法想像出一个与世隔绝或被患有情绪障碍的家长虐待的儿童有过多么痛苦的经历。无法正常地习得语言在多大程度上是情绪障碍造成的呢?关于这些孩子还有什么是我们所不知道的呢?吉尼在14个月大的时候罹患重病,其主治医生认为她那时就表现出了一些智力滞后的征兆(Curtiss, 1977)。吉尼即使没有经历过那段可怕的生活经历,是否仍会发育迟滞呢?从个案研究中得到的数据常常会让我们继续思考导致行为产生的真正原因。

潜在的偏差来源 个案研究常常依靠一个既是参与者又是观察者的研究者所获得的结果来得出结论(Bolgar, 1965)。也就是说,一个治疗师既观察患者的行为同时又参与到整个治疗的过程中。治疗师很有可能会特别相信治疗能够帮助患者。因此,作为治疗师,即使目的很明确,也难以客观地观察患者的行为。偏差现象并不只出现在个案研究法中,观察者偏差(参见第4章)和实验者偏差(参见第7章)也存在于其他方法之中。

一个个案研究的结果可能主要是由观察者的“印象”所决定的(Hersen & Barlow, 1976)。例如,基尔希(1978)描述苏珊“感觉到”她有能力实现她的目标,以及报告说她对自己的生活有了一种“控制感”。他说他“感觉到苏珊已经成功地采用了一种问题解决式的行为评价方法,并且已经能够熟练地自己设计策略以达成目标了”(p.304)。个案研究方法最大的缺点就在于它的结果仅仅依赖于观察者的主观报告。

当个案研究从诸如个人档案、诊疗记录、心理测试等来源获得信息时,偏差也会出现。像我们在第6章提到的档案记录,就存在很多偏差。进一步说,当被试提供他们自己的信息(自我报告)时,他们可能会歪曲或伪造信息以让自己“看起来很不错”。这种可能性就存在于苏珊的治疗之中。我们无法知道她是否自我报告中夸大了自己的进步。个体根据回忆而做的报告也可能引起偏差。认知心理学家已经反复证明了记忆是不准确的,特别是对那些发生在多年前的事实的回忆。考虑一下下面这个例子。这是著名的人格心理学家阿德勒自我分析的节选(Alfred. Adler, 1973, pp. 179-180):

记得到寄宿学校后不久,我就发现学校的路边有一座公墓。我每次都非常害怕经过那里,和别的孩子一起经过墓地时,我一眼也不敢向那边看。我每走一步都惊恐万分。除了因恐惧引起的极度不适外,我也对于自己比他人更怯懦感到十分恼火。有一天我决定结束我的这种恐惧感。这一次,我决定摆脱对死亡的恐惧。我站在离他人一

定距离的地方，把书包放在靠近公墓墙边的地上，然后多次来回跑过公墓，直到我觉得我已控制住了恐惧感。从那以后，我相信我经过此处公墓时不再觉得恐惧了。

三十年后，一次我碰到一位老同学，我们共同回忆起了一些童年的生活，我的那位老同学在那一带居住的时间比我长，我就问他那个曾让我不愉快的公墓到哪儿去了，然而他听到这个问题时表现得异常吃惊，坚持说我们学校的路边从来没有一座公墓。这时我才意识到那个关于公墓的故事其实是我希望克服死亡恐惧的一种象征。

由单一个体推而广之的问题 个案研究法的最重要局限之一就是其结果的外部效度不高。我们可以在什么程度上将单一个体的结果推广到更大的群体上呢？也许来自单一个体的结果完全不能推广。一个个案研究的结论能否被推广取决于这个个案来自总体的变异情况。比如，那些研究视觉的心理学家常常可以将他们从一个被试得到的结论推而广之，因为视觉研究者假设所有人的视觉系统都是非常相似的。因此，了解人类视觉系统的生理机制只需一个或几个被试就够了。与之相反的是，其他的一些心理过程（如学习、记忆、情绪、人格和心理健康等）则变异很大。当研究那些总体变异很大的心理过程时，我们不太可能说从一个人身上观察到的结果会适用于所有人。

因此，即使我们接受基尔希（1978）的结论，认为 SMT 作为一种心理治疗方法是有效的，我们仍不能确定这种疗法是否可以成功地用于智商、年龄、家庭背景以及性别等很多方面都和苏珊不一样的其他个体。就像从群体研究方法得出的结论一样，关于个案研究我们要做的是在其他个体身上重复原来的研究。

对个案研究证据的批判性思考

- 当根据来自个体的证据来评估某一特殊处理的有效性时，我们应该注意个案研究的局限性。

个案研究有时能为“新”发现提供富有戏剧性的证明，或是为一个“成功”的治疗方案提供支持证据。想一想你在媒体（如商业电视节目）上见到的那些为产品做的广告。有多少担心自己肥胖的人能够拒绝这样一个例子：一个以前很肥胖的人在用过了某产品后瘦了许多。来自个案研究的证据往往容易说服别人，而这对于科学研究来说既是一件好事又是一件坏事。个案研究展现的全新或不寻常的发现可以引发科学家们对他们的理论重新进行思考，或者引领他们走向新的富有成效的研究方向。从这个角度看，个案研究可以促进科学的发展。

然而，个案研究的缺陷就在于它的发现经常被不假思索地接受。那些急于减肥或治愈疾病的人，可能并不会去考虑源自个案研究的证据存在局限性这个问题。相反，人们会认为那些证据给治愈疾患带来了一线希望。对那些没有多少选择的人来说，这种抓住救命稻草的想法也不是完全不合理。但是，大多数情况下，人们不去想（也许是不愿意去想）某一特殊处理可能对他们自己并不起作用。

当我们没能考虑到个案研究提供的证据存在局限性时，我们就可能面临不幸的后果了。例如，20 世纪 80 年代早期，号称可以治疗癌症的药物苦杏仁苷就引起了很大的争议（Sun, 1981）。没有几个令人尊敬的科学家和医学研究者认为这种

药物可以治疗癌症。然而，苦杏仁苷的支持者们却报告了具有积极疗效的个案研究结果。在很大程度上迫于公众的压力，政府对此组织了严格控制条件、系统性的和花费不菲的药物实验。结果在控制严格的实验中没有发现该药物具有任何效用。正如一些研究者所评论的那样，许多用苦杏仁苷代替传统治疗的患者实际上延缓或中断了有效的疗程，造成了癌细胞在体内更多的扩散。

单被试（小样本）实验设计

- 在应用行为分析中，从行为的实验分析中建立起来的方法被应用来解决与社会相关的问题。

在本章剩下的内容中，我们将讨论单被试（小样本）实验设计。这些实验设计源自于斯金纳在20世纪30年代所建立的行为研究方法。斯金纳的这种方法叫做行为的实验分析（experimental analysis of behavior），它向我们展示了一个关于人类本性的独特的行为主义观点：它不仅可以指示心理学家如何做研究，还可以加深我们对社会组织等问题的理解。斯金纳的一些著作，包括《桃源二村》和《超越自由与尊严》，描述了行为实验分析所得出的原理是如何被用来促进社会进步的。

在行为实验分析中，样本常常是单个的被试或者是小样本被试（而不像前面几章提到的群体研究方法）。实验过程中对实验条件进行严格的控制，使得个体的行

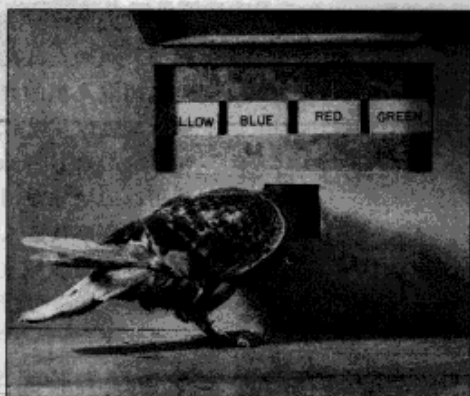
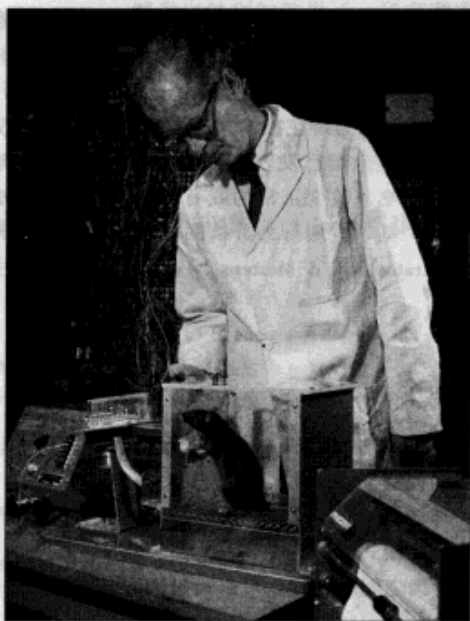


图 10.2 应用行为分析是斯金纳关于动物行为基础研究的扩展

为随着对自变量的操纵而发生系统的改变。正如斯金纳 (1966, p.21) 所评论的:

相对于对 1 000 只大鼠研究 1 小时, 或是对 100 只大鼠研究 10 小时, 研究者也可以对一只大鼠研究 1 000 个小时。这样一种程序不仅考虑到了个体性, 而且至少在仪器使用的效率以及研究者花费的时间和精力上, 与传统方法一样具有效率。对实验的一致性或者可重复性要求来说, 不在于使用的是哪种方法, 而在于所达到的控制程度, 而这样的要求是行为实验分析方法可以轻松达到的。



通常, 单被试实验设计所需要的统计分析是最少的。关于某一实验变量 (或实验处理) 效应的结论一般是通过肉眼观察后的行为记录而获得的。我们需要观察行为是否随实验处理的施加与撤除而出现了系统性的变化。因此, 恰当地定义、观察和记录行为就变得十分重要了。行为已被足够清楚和客观定义以便能够进行有效的观察和记录了吗? 行为的观察和记录是以连续的方式还是以某种固定间隔的方式来进行的呢? 虽然反应频率是行为的常用测量指标, 但是行为的持续时间或者行为的特征有时也被用作测量指标。而且, 正如你将在本章稍后所见到的, 研究者有时也会需要通过统计分析 (在行为记录中过大的变异) 做适当处理。除了这个简单介绍外, 我们将不再介绍与单被试实验设计有关的其他统计分析技术。

在应用行为分析 (applied behavior analysis) 中, 通过行为实验分析建立起来的方法被应用于相关的问题上。这些应用常常指的就是行为矫正 (behavior modification), 但当应用于临床上时, 行为疗法 (behavior therapy) 的说法则更受欢迎一些 (Wilson, 1978)。许多人认为行为疗法是比基于精神动力学模型的疗法更为有效的临床治疗方法。行为疗法关注可观察到的行为, 而不要求洞察问题的无意识根源。例如, 自我刺激行为 (如持续地摇晃身体、凝视光源、旋转等) 常被看作是自闭症儿童的典型特征, 而这些行为可能是由于偶然的强化所导致的。这样, 临床医生以及教师就可以运用行为矫正技术来控制行为出现的次数 (参见 Lovaas, Newsom, & Hickman, 1987)。大量已发表的研究论文表明, 行为矫正以及行为疗法可以成功地被用来治疗口吃, 改变正常的或有认知损伤儿童及成人、精神病患者以及许多其他人的行为。基于应用行为分析的方法还被学校心理学家们成功地应用于教育领域 (参见 Kratochwill & Martens, 1994)。这些研究主要发表于《应用行为分析杂志》。

单被试实验的主要特点

- 在一个单被试实验中研究者只操纵一个自变量, 这样, 这些设计就可能实现比个案研究更严格的控制。
- 在单被试实验中, 首先记录的是基线行为观察, 用以描述在没有进行任何处理的情况下被试的行为模式。
- 基线行为与干预后行为之间的差异, 可以通过肉眼观察行为来进行比较。

不言而喻, 单被试实验 (single-subject experiment) 主要关注单一个体或顶多

几个有限个体的行为改变。然而，正如我们在本章其后的部分将要看到的，单个“组”的行为也是单被试实验所关注的。在单被试实验中，研究者比较一个被持续观察的个体在各种处理条件下的行为。也就是说，研究者对所感兴趣的自变量（通常是实验处理）在一个个体上进行系统性操纵。相比无控制的个案研究来说，单被试实验设计是研究者的一种重要选择（Kazdin, 1982）。单被试实验设计相比多组实验设计来说也有其优势（参见专栏10.1）。

单被试实验的第一步通常是观察阶段或基线阶段（baseline stage）。在这个阶段，研究者记录下被试在实验处理前的行为。临床研究者通常测量目标行为在单位时间（比如一天或一小时）内出现的次数。例如，一个研究者可能记录10分钟的谈话过程中一个非常害羞的孩子使用眼神交流的次数；一位偏头痛患者在一周内报告头疼的次数；一位慢性口吃患者在一分钟内口语停顿的次数。通过基线记录，研究者可以描述被试在实验处理前的行为。最重要的是，基线可以让研究者预测在无处理情况下未来的行为（Kazdin, 1998）。当然，除非行为真的被观察到，否则研究者不能确切知道未来的行为，但基线测量可以让他们预测未来的行为。图10.3演示了基线记录的这一功能。在这一假设的例子中，三个集中趋势指标（平均数、中位数和众数）在对未来行为的预测上是一致的。对复杂行为的真实观测可能并不会如此一致，而且研究者必须仔细权衡三个集中趋势指标的优缺点（参见第12章）。

当研究者观察到个体的行为已经相对稳定了（即在某一个时段，行为几乎不再波动时），就可以引入实验处理了。下一步是用与基线记录阶段相同的指标记录被试的行为。处理后立即观察到的行为与基线阶段的行为进行比较，研究者就能够看出处理的效果了。研究者也可以通过行为记录图很容易地看出处理的效果。换句话说，在实验处理后行为是怎样改变的呢？我们可以估计没有实验处理时的

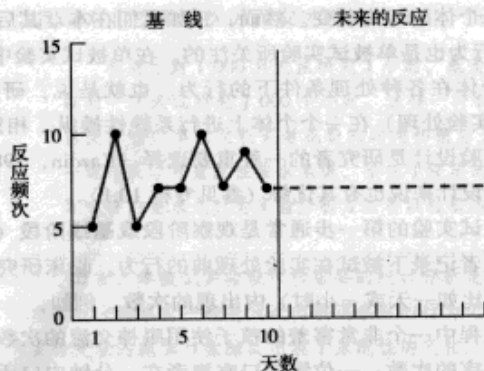
专栏 10.1

单被试设计相比多组设计的优势：少做多得

对于某些类型的应用研究来说，单被试实验设计可能比多组设计要更为合适一些（参见 Hersen & Barlow, 1976）。一种情况是研究目的是直接改变某一特定个体的行为。例如，一个多组实验研究可能做出这样的结论：“从总体上来说”，这种处理方法对矫正行为更加有效。然而，我们不可能根据一个组的平均结果而直接推测某种方法适用于某一特定个体。卡兹丁（Kazdin, 1982）对单被试实验的这一特点做出了很好的总结：“可能单被试实验设计最明显的优势就在于它可以对某一位患者进行细致的考察，并且有针对性地用实验方法评估某一治疗方法对患者的效用”（p.482）。

单被试实验相比多组实验的另一个优势涉及一个伦理问题，即在临床研究中可能出现的拒绝治疗的问题。在多组设计中，一种可能有效的治疗手段不能施予某些个体，因为考虑到实验的内部效度，这些个体必须作为控制组。而单被试实验设计仅用同一个体对比其在“无处理”和“有处理”条件下的行为结果，因此不会出现拒绝治疗的问题。此外，临床研究者常常很难找到足够的患者来做多个多组实验。比如说，一个临床医生可能只找到了很少的几个幽闭恐惧症患者（对封闭的空间异常恐惧）。当只有很少的几个满足条件的被试时，单被试实验可以为其研究提供现实的解决方案。

图 10.3 假定的“反应频次”的基线观测示例。基线阶段的数据（以实线表示）用来估计未来没有实验处理时的反应频次（以虚线表示）（资料来源：Adapted from Kazdin, 1998, p.209.）



行为模式（在图 10.3 中以虚线表示）。通过比较实验处理后的行为与没有实验处理时行为之间的差异，我们可以推测处理是否有效地改变了个体的行为。传统的单被试实验并不使用统计显著性检验，但在这个问题上还存在争论（Kratowill & Brody, 1978）。在本章稍后我们将讨论用目测来决定一个处理是否有效可能会带来的一些问题（也可参见 Kazdin, 1998）。

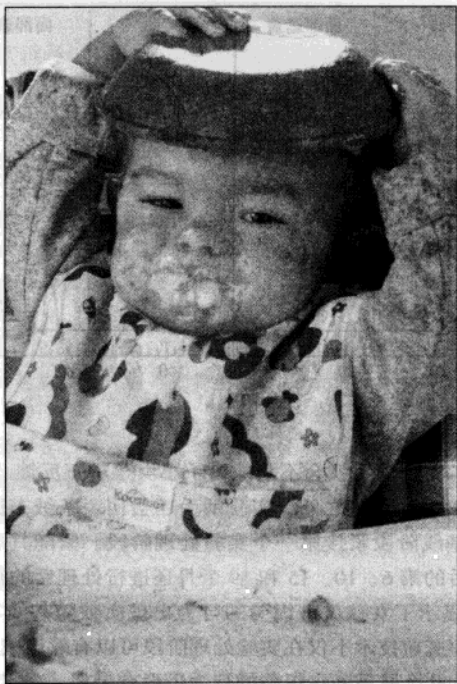
尽管研究者有很多可供的选择（Hersen & Barlow, 1976; Kazdin, 1980），但最常见的单被试设计是 ABAB 设计和多重基线设计（Kazdin, 1998）。

具体的实验设计

- 在 ABAB 设计中，基线（A）和处理（B）交替出现，以此来决定处理对行为的效应。
- 如果行为随着处理的引入与撤除而发生系统的变化，那么研究者就可以下结论说处理导致了行为的改变。
- 在撤去处理之后如果行为不能返回到基线水平，那么在 ABAB 设计中将实验处理解释为行为改变的原因就会出现问题。
- 出于伦理的考虑，心理学家有时候可能会放弃使用 ABAB 设计。
- 多重基线设计中，在处理被引入后，若在多于一条基线上出现了行为的改变，那么处理就是有效的。
- 多基线观察可以跨个体、行为或情境而进行。
- 当实验处理前的基线阶段出现了行为改变，将实验处理解释为行为改变的原因也会出现问题，这个问题经常出现在实验处理效应的推广过程中。

ABAB 设计 研究者运用 ABAB 设计 (ABAB design) 来证明当“无处理”和“有处理”交替出现时，行为会随之系统地发生改变。初始的基线阶段（A）紧接着处理阶段（B），然后又回到基线阶段（A），再进入另外一个处理阶段（B）。因为实验处理在第二个 A 阶段被撤除，所以行为上的任何改进此时都应是反向的。这种设计也被称作反向设计（reversal design）。研究者运用 ABAB 设计来观察行为是否在处理变量引入后马上发生改变（第一个 B），当处理撤除后行为是否返回到基线

图 10.4 应用行为分析已被用于控制儿童或成人的不良行为



水平(第二个A)。而且当处理变量再次引入后,行为是否仍会发生改变(第二个B)。如果行为随着处理变量的引入与撤除而相应地发生改变,那么研究者就有了充分的证据来说明是处理引起了行为的改变。

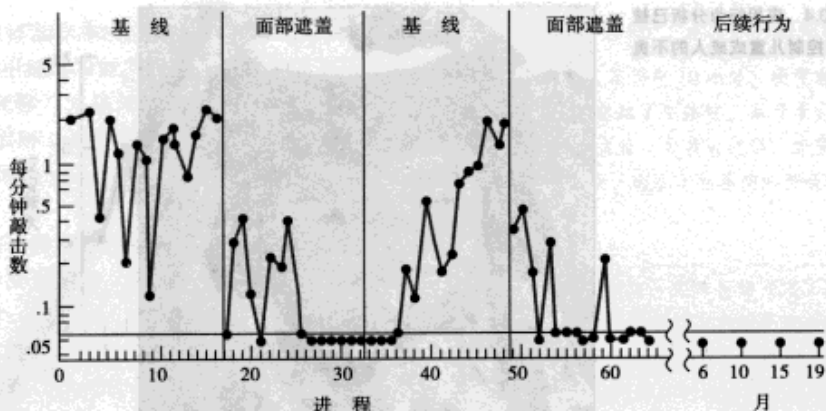
霍顿(Horton, 1987)用ABAB设计评估了面部遮盖技术对一名有严重心理损伤的8岁女孩的不良行为矫正效果。面部遮盖是一种比较轻微的厌恶治疗技术,当目标行为出现时就会用东西将脸遮住(如用一块柔软的布)。以往的研究证实了这一技术在减少自我伤害行为(如抽打自己的脸)上的有效性。霍顿想知道这种方法是否能够减少这名女孩在吃饭时敲击饭勺的行为。这种敲击饭勺的行为不仅产生噪音,而且使她把食物洒到地板上或把饭勺掉到地上,使得这个在特殊学校学习的女孩无法和她的同学一起就餐。

研究者先对这一敲击饭勺的行为进行明确定义,以将其与吃饭时正常的舀饭动作区分开。随后一位受过训练的实验助手观察并且对其施加实验处理。在吃饭时,每隔15分钟实验助手就会通过计算行为出现的次数来评估敲击饭勺行为出现的强度。在基线阶段,实验助手将目标行为出现的次数记录下来,并且每当目标行为出现时,就对她说:“不要敲了”,同时轻轻地抓住她的手腕,把她的手扳回到她的盘子边。整个过程用录像记录下来,并且由另外一名观察者独立地观看录像并记录目标行为出现的次数,以作为信度的检验。观察者间信度大约为96%。基线水平测量持续了16天。

第一个实验处理阶段从第17天开始,并一直持续了16天。每当敲击饭勺的

图 10.5 基线、处理以及处理后敲击饭勺行为的频次

(资料来源: Adapted from Horton, 1987.)



行为出现时, 实验助手就会不断地给予小女孩矫正反馈“不要敲了”, 并且将她的手扳回到盘子边。除此之外, 实验助手还会用一条绒布围兜罩住女孩的脸部 5 秒钟。第二个基线阶段紧接第一个实验处理阶段, 紧随其后的是第二个实验处理阶段。在实验后的第 6、10、15 和 19 个月还进行处理后的行为观察。

图 10.5 显示了女孩敲击饭勺的行为出现次数随基线阶段与处理阶段交替发生变化的过程。面部遮盖技术不仅在实验处理阶段可以有效地减少目标行为, 而且可以看出在实验后的几个月后敲击饭勺的行为仍没有出现。在实验的最后阶段, 无论是在家里还是在学校里, 那名女孩都不需要对她进行直接的指导了, 她因此被允许和其他同学一块吃饭。这一事实非常清楚地证明了面部遮盖技术对消除敲击饭勺行为的直接作用。这是因为, 面部遮盖技术是在实验过程中惟一被使用的实验处理, 而且对图 10.5 的目测可以看出, 行为随着实验处理的引入和撤除而相应地改变。当其他的干预程序不起作用时, 面部遮盖技术是控制儿童不良行为的有效方法。

ABAB 设计的一些方法学问题 ABAB 设计在方法学上的一个主要问题可以从图 10.5 中显示的研究结果中看出。在第二个基线阶段, 当面部遮盖被撤去后, 敲击饭勺的行为又增加了, 也就是说在实验处理阶段获得的进步消退了。如果在实验处理撤销后敲击饭勺的行为仍保持很低的出现频率会怎样呢? 当第二个基线阶段时行为没有返回到第一个基线时的行为水平, 那么研究者将如何对实验处理的有效性下结论呢? 在专栏 10.2 中我们讨论了行为没有返回到基线水平的一些原因。

不管是什么原因导致实验处理撤除后行为仍未返回到基线水平, 如果发生这种情况, 那么研究者就不能贸然下结论说是实验处理导致行为改变 (Kazdin, 1980, 1998)。研究者必须仔细检查实验环境, 努力找出可能导致混淆的变量, 或是对其他被试进行重复实验 (Hersen & Barlow, 1976)。

研究者在使用 ABAB 设计时还可能会面临伦理问题。假如实验处理与基线相比似乎说明其改善了个体的行为, 那么撤除一个可能对个体有效的治疗以检验其有效性是道德的吗? 你能想得到, 撤除一个可能对个体有效的治疗无论在什么情况下都是理由不充分的。一些行为有可能威胁到生命或者有损于健康, 因此, 观察到有效效应之后再撤除处理就可能是不道德的。例如, 一些孤独症儿童可能表现

专栏 10.2

为什么在反向设计中没有出现反向?

行为没有返回到基线水平的一个原因是:一旦实验处理产生了积极作用,它可能是持久的。这种情况在实验处理涉及教会个体一项新的技能时会出现。例如,某研究者的实验处理可能是教会一个残疾人学会坐通勤车上下班。一旦被试学会了这项技能,即使实验处理被撤除,这项技能也不大可能被“忘却”(返回到基线水平)。针对这种情况,直接的解决办法只能是放弃使用ABAB设计。

当行为没有返回到基线水平时,还有什么可能的原因呢?一种可能性是实验处理之外的变量是导致从基线到第一个处理阶段行为变化的原因。例如,个体在接受实验处理的阶段可能获得身边的人与朋友更多的关注。正是这种关注而非实验处理本身导致行为发生变化。如

果这种关注在实验处理被撤除后仍然存在,那么行为的改变也很可能持续下去。这样一种解释暗示了实验处理变量与其他未被控制的因素(如关注)混淆在一起。

也有可能尽管实验处理导致了行为的改善,但其他的变量在其后取而代之发挥着对新行为的影响。我们可以再一次认为关注对行为改变起了作用。当家人和朋友看到行为上的变化后,就对个体给予了更多的关注。想像一下当人们减轻了体重或是戒掉了吸烟时他们所得到的表扬。以关注形式出现的正强化可能会维持行为的改变,而这种改变最初是由实验处理导致的。在这种情况下,我们同样不可能期望实验处理被撤除后行为能够返回基线水平。

出撞头这样的自我伤害行为。如果一位临床研究者成功地减少了这种行为出现的次数,那么为了满足ABAB设计的要求而撤除处理就明显是不道德的。幸运的是,有一种单被试实验设计适合在上述情况下使用,它可以满足不撤除实验处理的要求,这种设计就是多重基线设计。

多重基线设计 多重基线设计也使用基线阶段和实验处理阶段相结合的方法,但是并不像ABAB设计那样需要撤除实验处理。不言而喻,使用多重基线设计时研究者需要建立多条基线,当引入实验处理后,如果行为在多条基线上都发生改变,那么多重基线设计就可以证明实验处理的有效性。

一个关于多重基线设计的例子是在不同的情境下对个体的行为施加处理。在这种情况下,第一步就是要选择通常发生行为(如儿童的攻击性行为)的情境(如:家里、教室和日托中心)来记录行为。研究者在基线条件下于各种情境中分别记录行为出现的次数。下一步是在某一个情境中(如家里)而不在其他情境中引入实验处理,研究者继续在所有情境下观察行为。多基线实验设计的一个关键特征就是实验处理一次只施加于某一个基线条件。在接受处理的情境下,行为应该有所改善,而在其他情境下不会有行为的改变。下一步是在第二个情境中施加实验处理(继续对第一个情境施加实验处理),但是在第三个情境下继续保持基线水平。行为应该只在有实验处理的情境下有所改变,而在基线条件下的情境中保持不变。最后一步就是在第三个情境下也施加实验处理,此时在第三个情境下的行为也应发生改变。在多重基线设计中证明处理有效性的关键证据就是:只有当处理被引入时,行为才会发生改变。

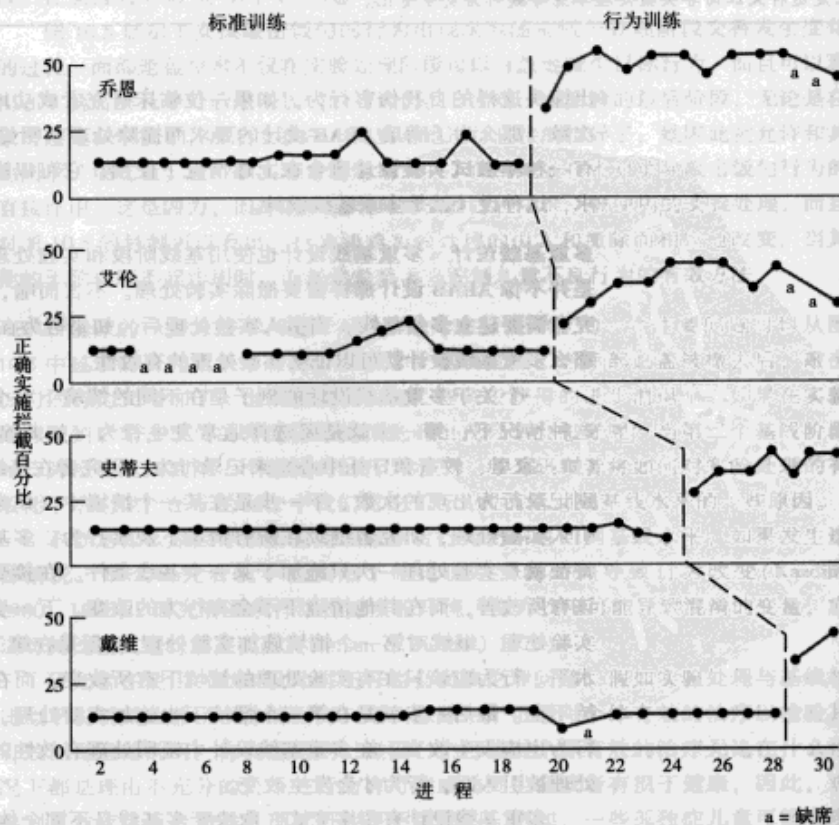
多重基线设计有很多变式,取决于多基线是不同个体、同一个体的不同行为,

还是同一个体处在不同的情境下。尽管听起来很复杂,但是多重基线设计却经常被使用,而且很容易理解。下面我们将分别用一个实例来讲解每一种多重基线设计。

在跨个体多重基线设计(multiple-baseline design across individuals)中,首先要建立不同个体的行为基线水平。当每一个体的行为都稳定了之后,一个实验处理就被施加到其中一个个体,然后是另外一个,再下一个,依此类推。与所有的多重基线设计一样,每一条基线(在此例中是每一个体)都在不同的时刻被施加一个实验处理。如果实验处理是有效的,那么对于每一个体来说,在实验处理施加之后行为马上就会发生变化。

一个跨个体多重基线设计例子来自运动心理学领域。艾利森等人(Allison & Ayllon, 1980)对一种包含多种行为技术的训练方法非常感兴趣。该方法涉及多种运动项目技能的掌握,包括橄榄球、网球和体操等。尽管他们发现这种方法对每种运动都是有效的,我们在这里只讨论这种行为训练方法对掌握橄榄球技能是否有效。参加实验的被试都是市级橄榄球项目的候补队员,选择他们是因为他们“完全缺乏基本的橄榄球技能”(p.299)。

图 10.6 多重基线示例: 4 位选手在标准训练和行为训练条件下的正确拦截百分比
(资料来源: From Allison & Ayllon, 1980.)



在艾利森和艾伦的研究中所要掌握的技能是拦截。根据8个成分对拦截技术进行操作性定义,从第一个站到并列争球线后到保持身体接触直到吹哨。行为训练方法的一些具体程序由橄榄球教练来完成。这些程序包括系统的口头反馈、正强化和负强化,以及其他的一些行为技术。实验者首先在“标准训练”条件下对橄榄球队中的几个成员分别建立基线行为。在标准训练程序下,教练会用到口头指导,提供偶尔的示范或口头表扬。当动作不正确时,会“大声地提醒球员,有时会骂他们愚蠢、不勇敢、没有意识,或者是更难听的话”(p.300)。简而言之,这是典型的消极行为训练实例。

实验者和第二个观察者记录球员们每10次尝试中正确拦截的次数。对4位球员在不同的时间开始进行行为训练。行为干预的结果如图10.6所示。对于这4位球员来说,行为训练法都有效地增加了他们正确拦截的次数。两位观察者之间的评分一致性从84%到94%,表明对行为的观察是可靠的。在行为训练被引入后,四位球员的拦截技能马上提高了。因此,这个多重基线设计的结果表明这种训练方法导致了球员成绩的提高。

第二种多重基线设计是建立同一个体的两种或两种以上的不同行为基线,称作跨行为多重基线设计(multiple-baseline design across behaviors)。一种实验处理首先施予一种行为,然后是另外一种,依次类推。如果在引入实验处理后每一种行为都立刻发生改变,则可以认为实验处理与行为改变之间存在因果关系。例如,吉纳、克兰兹、麦克兰纳汉和波尔森(Gena, Krantz, McClannahan, & Poulson, 1996)尝试向孤独症儿童教授各种适当的社会情感性行为。正如研究者所提到的,孤独症儿童经常会做出不合时宜的情感性行为,这些行为常常限制了他们与他人的有效交流以及人际关系的发展。实验处理包括口头表扬和代币奖赏(用于交换实物),用以此鼓励三或四类适当的情感性行为。目标行为是以下行为:表现出感激、谈论最喜欢的事物、对可笑的事情大笑、表现出同情以及表达厌恶。对行为记录的目测表明实验处理是有效的。当实验处理引入后,各种情感行为都立即出现了改变。

多重基线设计的第三种主要变式涉及对两种及两种以上的情境下的同一个体的行为建立基线,称为跨情境多重基线设计(multiple-baseline design across situations)。例如,前面我们在介绍多重基线设计的概念时提到,研究者可能为不同情境(如在家里、在教室或在日托中心)中儿童的攻击性行为出现的次数建立几条基线。与这种设计的其他变式一样,实验处理在不同时刻被施予不同基线条件,然后考察行为记录是否随着实验处理的引入而系统地发生改变。

范比尔列特、斯潘格勒和马歇尔(VanBierliet, Spangler, & Marshall, 1981)采用跨情境多重基线设计来研究食物的发放方式对精神病院病人之间口头交流的影响。研究者认为以往的对心理损伤病人发放食物的方式不利于病人之间的口头交流,而一种新的程序“家庭式服务模式”可能增进病人间的口头交流。在三种不同情境(早餐、午餐和晚餐)下,研究者测量五位住院病人之间的基线口头交流次数。研究者使用了两种方法评估行为改变:(1)考查五位病人言语交流的平均次数;(2)考查每位病人的行为记录。两种测量都表明在食物发放方式改变后,病人的言语交流行为有所增加。

多重基线设计在方法学上的问题

- 多重基线设计究竟需要多少条基线?

通常, 当要设计一个多重基线实验时人们首先会问: “我需要设立多少条基线?” 与单被试研究的许多其他方面一样, 这里并没有一个一成不变的规定。其最小值当然是两个, 但是两个通常被认为是不足的。有研究者推荐使用三或四条基线 (Hersen & Barlow, 1976)。

- 如果行为在干预之前就发生了变化该怎么办?

当实验处理被引入之前, 如果行为在基线阶段就已经发生变化, 那么任何一种多重基线设计都会出现问题。我们常常并不清楚这些尚未成熟的变化在基线阶段出现的具体原因。多重基线设计的逻辑特别依赖于行为在实验处理引入后立即发生变化。因此, 变化在处理引入前的基线阶段发生, 会使研究者难以以下结论说实验处理是有效的。如果实验处理前的行为变化只在几条基线的某一条上发生 (特别是有程序上或情境上的因素可以对行为改变做合理的解释时), 我们对多重基线设计结果的解释仍有几分信心。例如, 卡兹丁和埃里克森 (Kazdin & Erickson, 1975) 采用跨个体多重基线设计的方法帮助心理严重受损的患者对指示做出回应。他们用糖果和口头表扬对按照指示反应的病人给予强化, 而且对四个不同小组在不同的时刻点引入干预措施。在三个小组中, 引入正强化程序后被试的行为马上发生了改变。但在有最长基线阶段的第四个小组, 行为在干预引入前就逐渐变化了。研究者分析后认为, 这种情况的出现是第四组的患者对其他组的行为进行模仿造成的。

- 实验处理效应怎样推广到其他的行为和情境?

当要把从某一行为上得出的结论推广到其他的行为或情境中去时, 多重基线设计经常会碰到问题。例如, 研究者在晚餐时成功地增加了心理损伤病人间的口头交流次数。如果病人在其他的进餐时间口头交流次数也增加了, 那一点也不让人感到吃惊。口头交流的增加很有可能是由所提供的食物, 或是坐在一位熟人的身旁所引发的。事实上, 范比列特等人 (VanBiervliet et al., 1981) 并没有观察到结果可以推广到所有进餐时间, 因此证明了在用餐时间进行干预是有效的。

在将结论推广到其他方面时, 研究者应记住一条原则: “一盎司的预防抵得上一磅的事后治疗”。如果改变某一个个体的行为会影响其他人的行为, 如果某一情境下的行为会影响其他情境下的行为, 或者改变某一种行为可能会影响其他行为, 那么多重基线设计可能就需要进行改进, 或者可能要完全放弃 (Kazdin, 1998)。不幸的是, 预测变化何时会在多个基线上同时发生并不总是一件容易的事情, 但是这些问题与多重基线设计通常发现的效应相比似乎是不那么常见的 (Kazdin, 1998)。不管怎样, 有一点是明确的, 即只有每条基线都在引入实验处理后发生行为变化, 才得出多重基线设计是有效的结论。

所有单被试设计共同的问题和局限性

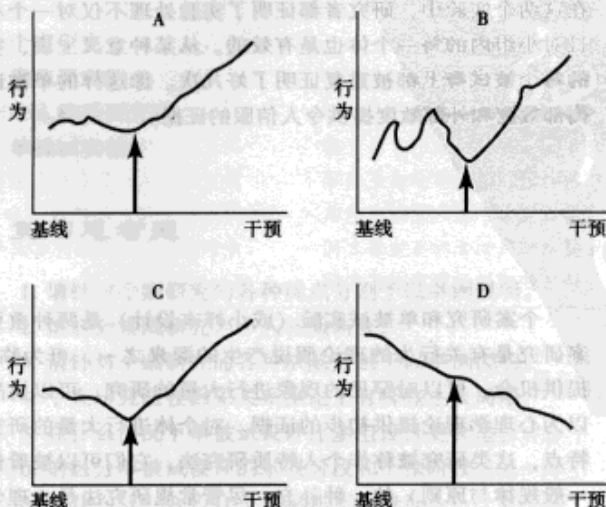
- 如果基线阶段表现出非常大的变异或有行为增多或减少的趋势时，研究者就很难对实验处理效应予以解释。
- 采用小样本组设计可以减少单被试实验所面临的外部效度低的问题。

基线记录存在的问题 理想的基线记录以及对干预的反应如图 10.7 的 A 栏所示。被试在基线阶段的行为非常稳定，当实验处理引入后行为立即发生了改变。如果这是 ABAB 设计或多重基线设计第一阶段的结果，那么我们就可以直接说处理引起了行为改变。但是，考虑一下图 10.7 中 B 栏所示的基线和处理阶段。尽管目标行为出现的次数在干预后表现出了增长，基线水平的变异却很大。因此很难判断究竟是处理产生的行为改变还是行为出现次数碰巧处于上升的趋势。总的来说，当基线阶段有很大变异时就很难确定干预是否有效。

有几种方法可以处理基线阶段的较大变异问题。一种方法是寻找情境中可能导致变异产生的因素，进而将其清除掉。比如，某一工作人员的在场可能导致精神病人的行为改变。另外一种方法是“等待它消失”——直到行为稳定后才结束基线测量。当然，我们无法预测行为的稳定状态什么时候会出现，无法预测这种稳定状态是否会出现。然而，在行为稳定之前引入实验处理将会造成结果难以被解释的问题。最后一种方法是平均数据点。通过使用多点的平均来记录行为，研究者有时可以降低变异性 (Kazdin, 1978)。

图 10.7 的 C 图显示了当基线表现出增长或下降的趋势时可能引发的另外一个问题。如果干预的目的是为了使行为出现的次数增加，那么 C 图干预之前显示的下降趋势不会给解释带来任何问题。干预逆转了下降趋势反而可以作为实验处理有效的证据。然而，如果干预的目的是为了减少行为出现的次数，问题就严重了，D 图表示了这种情况。这里我们看到行为在基线阶段有一个下降的趋势，在

图 10.7 行为记录示例：行为矫正中基线阶段和干预阶段的可能关系示意图（箭头表示干预开始）



实验处理阶段继续保持行为次数的减少。我们由此很难知道实验处理是否有效,因为干预后的行为减少有可能是干预的作用,也有可能只是延续了基线阶段的趋势。如果所期望的干预效应与基线趋势的方向一致,则干预引入后的行为变化应比D图中显示的更为明显,这样才能支持实验处理有效这一结论(Kazdin, 1978)。在单被试设计中实验处理效应经常是通过目测行为记录来判断的,这样就使得这个问题变得更加麻烦了。我们常常很难判断在行为记录中什么样的变化才算得上是一个更“显著的”变化(Parsonson & Baer, 1992)。这些情况下,在对目标行为的观察之外,配合使用一些其他的评价手段会有更好的效果(如,与“正常”个体进行对照或者获取与个体相熟人士的主观报告)。

外部效度的问题 对使用单被试研究设计常见的批评是其研究结果的外部效度很有限。换句话说,单被试实验似乎与个案研究法有着相同的局限性。因为每一个人都是独特的,所以有人会争辩说,我们无法知道某一特定的实验处理对某一个体的效应是否可以推广到其他个体身上。然而,有几点理由可以说明为什么单被试实验的外部效度并不像它看上去的那么低。

首先,用于单被试实验的干预类型常常是非常有效的,并且常常会产生巨大的和非常明显的行为改变(Kazdin, 1978)。所以,这些实验处理结果常常可以推广到其他个体。使用多重基线设计的研究也可以为单被试实验的效应推广提供证据。例如,一个跨个体多重基线设计常常可以表明一个特定的干预手段可以成功地改变几个个体的行为。同样,跨情境和跨行为的多重基线设计也可以为实验处理结果的外部效度提供证据。

要检验单被试实验中处理结果的外部效度,最好的办法是检验“单组”被试。有时单被试实验设计会使用一小组个体,也就是我们所说的小样本研究。我们已经介绍了两个这样的例子。范比尔列特等人(1981)发现,食物发放方式的改变能有效增加五位同时接受实验处理的心理损伤患者的口头交流行为。与此相似,卡兹丁和埃里克森发现正强化可以改善一个小组的心理损伤患者对指示的反应。在这两个实验中,研究者都证明了实验处理不仅对一个小组的被试是有效的,而且对小组内的每一个体也是有效的。从某种意义上说,实验处理的效应在一个组的每个被试身上都被重复证明了好几次。像这样的单被试实验可以同时为实验的内部效度和外部效度提供令人信服的证据。

小 结

个案研究和单被试实验(或小样本设计)是两种重要的单个案研究设计。个案研究是有关行为的理论假说产生的源泉之一,可为临床创新(如试验新疗法)提供机会,可以对罕见的现象进行大量的研究,可以挑战现有的理论假设,也可以为心理学理论提供初步的证据。对个体进行大量的研究是个案研究法的标志性特点,这类研究被称作个人特质研究法,它们可以被看作是对常规研究法(寻找一般规律与原则)的一种补充,尽管常规研究法是心理学的一个重要特征。但是

个案研究法不能被用来推出因果关系,在搜集和解释数据时也可能存在偏差。个案研究也不太能够使我们把单个被试的研究结果进行推广。此外,某些个案研究得出的一些“戏剧性”结果尽管可以为科学研究提供重要的启发,但它们常常被那些没有意识到这种方法存在局限性的人们不加评判地接受。

斯金纳建立了行为的实验分析法,又将行为实验分析中得到的原理应用到社会相关问题上。这些研究的主要方法就是单被试实验或小样本研究。尽管有很多种单被试设计,但最常见的是 ABAB 设计和多重基线设计。

ABAB 设计又称反向设计。它允许研究者通过行为随实验处理发生系统变化来证实某个处理效应。当行为在第一个实验处理 (B) 阶段发生改变,而在撤除实验处理的第二个基线阶段 (A) 行为又没有返回到基线水平时,运用这种设计就会带来方法学上的问题。当这种情况发生时,研究者就很难下结论说是实验处理而不是其他的因素导致了行为改变。此外,在 ABAB 设计中,一种处理表现出对个体有利后,在第二阶段仍撤除实验处理的做法会让研究者面临伦理上的批评。

多重基线设计通过实验处理的引入导致多条基线上行为发生改变来证明实验处理的有效性。首先研究者对不同的个体、不同行为或不同情境建立基线标准。当个体的行为并不是在实验处理引入后立即发生变化,或者处理效应需要推广到其他个体、行为或情境时,我们需要注意多重基线设计在方法学上的问题。

基线变异较大以及有上升或下降趋势的基线有时会使单被试设计的结果难以解释。基线变异较大的问题可以通过消除造成变异的来源、延长基线观察的时间或者平均数据点来解决。上升或下降的基线则需要研究者获得其他的证据来支持实验处理有效的结论。最后,单被试设计通常被批评缺少外部效度。然而,实验处理常常会造成行为的巨大改变,而且这些改变常常可以在其他个体身上重复。单组设计 (小样本研究) 可以为结果的跨被试推广提供便捷的证据。

重要概念

个案研究

常规研究法

个人特质研究法

单被试实验

基线阶段

ABAB 设计 (反向设计)

多重基线设计 (跨个体、跨行为、跨情境)

复习思考题

1. 请针对个案研究的各种优点分别予以举例说明。
2. 请区分常规研究与个人特质研究。
3. 请针对个案研究的各种缺陷分别予以举例说明。
4. 个案研究法在因果关系推论上有哪些主要局限?
5. 在什么情况下单被试设计比多组设计更合适?
6. 请区分单被试设计的基线阶段和干预阶段。
7. 为什么 ABAB 设计也被称作反向设计?

8. ABAB 设计有哪些特别的问题?

9. 请列出几种主要的多重基线设计的一般操作程序和原则。

10. 多重基线设计方法学上有什么问题?

11. 单被试实验方法学上有什么问题?

12. 哪些证据能够支持单被试实验的外部效度?

挑战性问题

1. 某一流行杂志发表了一项关于“泥浆疗法”如何成功治疗一位严重焦虑症患者的个案研究。患者的症状包括睡眠问题、厌食、与一群人交流时极度紧张以及感觉处于“边缘状态”和害怕。一位多次出现在电视谈话节目里的著名“泥浆疗法”治疗师负责治疗该患者。他首先教患者一种深度放松技术，并且要病人不停地念叨一个“神秘的词语”以阻止杂念。然后让患者在一个装满了泥浆的木桶中每天安静地呆上两个小时。在这段时间里，患者要练习他所教的放松技术，而且一旦感受到了任何焦虑感就要集中注意力重复那个神秘的词语。这种治疗非常昂贵，但是六个星期后患者向治疗师报告说他再也没有和以前一样的焦虑感了。因此，治疗师宣称患者被治愈了，并且把治疗的成功归结于病人能够沉浸于平静的泥浆之中。文章作者下结论说：“这是一种许多人都可以从中获益的治疗方法。”根据你对个案研究法局限性的了解，回答下面的问题：

A. 这个研究可能存在哪些问题?

B. 对治疗的成功你还能想到哪些其他的解释吗?

C. 只研究一个个体可能会产生哪些潜在的问题?

2. 一个5岁的儿童经常长皮疹。家庭医生告诉孩子的母亲，这种症状是因孩子吃了某种东西引起的。医生建议她仔细观察孩子平常吃的东西。这位母亲决定通过每天记录孩子是否长了皮疹以及孩子前一天所吃的东西来解决这个问题。她希望能够找到食物与皮疹之间的某种联系。尽管这种方法可能有助于找到某些食物

与皮疹之间的联系，但是根据单被试设计的原则和程序，我们应该能找到一个更好的办法。请解释一下这位母亲应如何运用其他更好的方法，并具体指出使用这种方法可能存在哪些问题。

3. 你找到了一份暑期工作，是在一个针对轻度心理障碍儿童的训练营中做咨询师，你负责指导一个小组的儿童，此外还要想办法让这些儿童更加关注训练营的各种室内活动（如手工制作和缝纫）。你决定试一下奖赏（使用某种糖果）的可行性。训练营主管需要知道你采用的干预手段是否有效，以便决定其是否可以用到训练营其他儿童身上。因此，你决定：

A. 根据强化原理制订一个干预计划，以增加儿童参与营地活动的时间。

B. 解释你需要记录哪些行为指标以及你如何判断你的干预手段对行为改变产生了作用。比如，你需要具体说明你要在什么时候测量，你怎样测量行为，以及你选择这一“实验”的理由。

C. 说说你将用什么样的证据来说服主管，使其相信你的干预手段对其他类似的儿童也是行之有效的。

4. 一位老师请你帮忙制定一个行为干预计划，以帮助其课堂上的一个问题儿童管理自己的行为。这名女孩在老师要求坐下时站起来，要求安静的时候大声说话，而且经常做出其他扰乱课堂秩序的行为。具体谈谈怎样运用正强化的手段（如糖果和小玩具），通过跨行为多重基线设计来改善该儿童的行为。

拓展练习的答案

1. 你可能倾向于赞成你的朋友。个人化的例子比数字统计更引人注目一些。然而，对比这两个研究，认识到他们采用了不同的研究方法是非常重要的。第一个研究采用的是依赖于大样本和用数字来进行描述的方法。第二个研究采用的是个案研究而不是数字描述的

概述

独特方法。认识到两种方法的不同后,就没有必要在两个研究结果中进行优劣的比较了。第一个研究的确证实了超过半数的婚姻以离婚告终,但是这也说明了少于半数的婚姻不是以离婚告终的。第二个研究说明尽管面临长时间的冲突和家庭离婚史等风险因素,但是婚姻并不一定以离婚告终。例如,一对夫妇从刚开始打算离婚到经过一年的治疗,最终决定坚守婚姻的。这个例子说明两种研究方法能够互相补充,而不是互相竞争的。

2. 你的朋友的第二个问题是心理系学生经常会问的一个问题。所有的研究结果对我有什么用?这两个研究的结果为你的朋友对自己未来的思考提供了有用的信息。第一个研究告诉我们离婚的确发生了,而某些因素也被认为是导致离婚的重要因素。第二个研究告诉我们尽管存在风险,婚姻也能渡过难关。系统和严格控制的研究所得出的结论可以补充我们由自身经验中获得的知识。事实上,你的朋友并不能依赖这些发现来决定自己的婚姻。更一般的问题是,这些心理学发现并不能够回答奥尔波特所提出的“某一个人将做什么”的问题。

挑战性问题的答案

- A. 这个个案研究的一个问题是同时作为治疗者和观察者的主试观察偏差。第二个问题是他做出的治疗效果仅仅依赖于患者的自我报告而做出的结论。
- B. 治疗的成功可能由于放松疗法本身、在焦虑时念叨的神秘的词语、患者所得到的关注,甚至是治疗的巨大开销。
- C. 采用一个被试进行研究的主要的潜在问题是缺乏外部效度。

真实验

真实验的特点

真实验是指,研究者随意地操纵自变量和各个条件,并对其他因素加以严格控制(高内效度)同时控制无关条件(高外效度)的实验。

第 11 章

准实验设计和项目评估

本章大纲

概 述

真实验

真实验的特点

在自然环境中进行真实验的困难

控制真实验的内部效度

真实验也可能控制不到的问题

准实验

非等控制组设计

非等控制组设计：兰格和罗丁的研究

影响非等控制组设计内部效度的因素

外部效度问题

间断时间系列设计

非等控制组的时间系列

项目评估

小 结

课堂练习的答案

1. 你可能倾向：赞成他的观点。个人化的例子比数字统计更引人注意一些。然而，对统计学的研究，认识到他们采用了不同的研究方法是非常重要的。第一个研究采用的是描述性方法，而第二个研究采用的是实验方法。第二个研究的结果比第一个研究的结果更可靠，因为它使用了数字统计进行描述的方法。第二个研究的结果比第一个研究的结果更可靠，因为它使用了数字统计进行描述的方法。

概述

总体而言,一个实验就是一个测试,是我们用来发现我们还不知道的东西而使用的一个程序。从这个角度来说,实验就如同向一个辣味配方中添加新的配料,来看看这些配料是否改善了食物的口味。改变钓鱼时使用的饵料、采用一个不同的上班路线以缩短上班路上的时间,这都是实验。可是你肯定已经认识到了,这些非正式的“实验”和那些在心理学研究中进行的典型实验有明显的不同。实验法不同于其他研究技术(比如观察和调查),它是确定因果关系最有效的方法。但是确定因果关系并不总是那么容易,前面几章已经告诉你研究者所面临任务的复杂性,他们试图通过发掘原因来理解现象。

本章我们继续讨论实验方法,但是会关注在自然环境中进行的实验,比如医院、学校或者企业。你会发现在这些环境下要想得到因果结论通常更加困难,并且当一个研究者离开实验室的限制,在自然环境中从事实验时就会产生新的问题。

研究者要在自然环境下进行实验的原因很多,其中之一是因为“现场实验”可以检测实验室研究结果的外部效度(见第7章)。也就是说,我们希望知道在实验室观察到的一种处理效果,在另一种环境中是否也能得到类似的重现。自然环境下的实验的其他原因则更加具有实践性。自然环境下的实验允许研究者达成一个重要目标,即它可以试着改善人们的生活和工作条件。政府可以试验一个新的税收系统,或者试验一种针对贫困人群的职业培训新方法。学校可以尝试改变午餐、课后管理或课程安排。一个企业可以试验新的产品设计、职工福利的发放方法或者弹性工作时间。这些情况与在实验室研究一样,都是为了确定“实验处理”是否带来了某种改变。允许病人进入医院急诊室会不会使病人得到更快、更有效的治疗?大学里的能源管理项目是否降低了能源的消耗?知道一个处理是否有效可以让我们做出重要决策:是否继续实施这个处理;是否花费这些额外的金钱;是否投入更多的时间和精力;或者是否根据实验结果改变现状。项目评估的目的之一就在于确定社会机构、政府部门和其他组织做出改变是否带来了预期的效果。

本章我们将描述自然环境下进行实验会遇到的困难,并且讨论如何克服这些困难以使实验顺利进行。然而,在实验室之外进行实验有时并不可行,此时就必须考虑实施近似于实验室条件的实验程序,我们会提供几种准实验技术。本章的最后,我们会简短地介绍一下项目评估的逻辑、流程以及可能的局限等内容。

真实验

真实验的特点

- 在真实验中,研究者随着实验进程操纵自变量和各个条件,并对其他因素进行严格的控制(常见的是通过随机分配各条件来实现)。

虽然很多日常活动(比如改变食谱的配料)也可以称为实验,但在本书所讨论的实验方法的背景下,我们不能认为它们是“真”实验。类似地,许多政府实施的“社会实验”以及那些由公司官员或教育机构进行的实验也不是真实验。一个真实验能够得出关于事件因果关系的确凿结论。

真实验有三个重要特点:

1. 在真实验中会实施某种干预或处理。
2. 真实验是需要进行严格控制的,一个实验者需要设计实验条件、分配被试,系统地操纵自变量和选择因变量。把被试随机分配到各个实验条件,通常被视为真实验最关键的特点(Judd et al., 1991)。
3. 最后,真实验还有一个特点就是恰当的比较。实际上,实验者努力进行控制以确立一个合适的比较对象来评价处理的效果。在最简单的实验情境中,这种比较在两个组之间进行,这两个组除了被关注的变量以外其他处理都一致。

在真实验中各个实验条件确定之后,因变量结果的任何不同,都能够逻辑性地归因于自变量水平的不同。在自然环境下进行的实验和实验室实验之间存在着不同,专栏 11.1 中列出了它们之间的一些重要区别。

在自然环境中进行真实验的困难

- 对于自然环境下的真实验,研究者在获得实验许可和接近被试时可能会遇到困难。
- 虽然有些人认为随机分配不公平,因为这种方法可能会使某些个体失去接受一种新处理的机会,但随机分配仍然是确定一种新的处理是否有效的最好和最合理的方法。

实验研究对解决和回答实践问题而言是一个有用的工具。然而,当我们试图在自然环境中进行实验时常常会遇到两个主要的困难。第一个问题是从当事人那里获得做研究的许可。除非他们认为研究是有用的,否则诸如学校理事会会长以及政府和企业领导等一些人就不太可能在经济上或其他方面给予支持。第二个更加重要的问题是获得被试,如果被试要随机分配到实验处理组或对照组,那么这个问题就更加麻烦。

把被试随机分配到不同实验条件表面上看首先是不公平的——毕竟这个做法要求至少要对一些被试不施加一种潜在有利的处理。假设你的大学要试验一种新的外语教学方法,当你去注册下学期的课程,却被告知会把自己随机分配到两个教学班的其中一个——一个班采用老的教学方法,而另一个则采用新的教学方法,你会如何反应?从研究方法的角度来看,两种教学方法必须在可比较的学生组之间实行,而随机分配是保证这种可比较性的最好的方法。虽然如此,但你很可能会感觉到随机分配是不公平的,特别是对那些被分到使用老方法(过时的?)教学的班级的被试。让我们来进一步看看随机分配的公平性。

如果那些负责选择外语教学方法的人已经知道新的方法比旧的方法更加有效,

专栏 11.1

实验室实验和自然环境中实验的区别

实验室之外进行的实验在很多重要的方面区别于实验室中进行的实验。当然不是每一个自然环境中进行的实验的所有方面都与实验室的实验不同,但如果你正在考虑在自然情境中做研究,我们强烈要求你能够考虑以下这些关键问题。

控制

科学家最为关心的就是控制。只有通过控制那些可能的影响因素,我们才能得出究竟是什么造成了某个结果。比如,在各个实验条件下随机分配被试就是一种控制,通过这种方法可以平衡个体之间的差异。研究者也可以通过控制让其他可能造成影响的因素保持恒定。在自然环境中,研究者也许不能总是像在实验室一样对被试的分配或实验条件进行那么严格的控制。他甚至需要对计划或实施“实验”时没有考虑到的因素是否有效进

行评价。这种“事后”评价特别困难,因为那些进行研究的人可能在计划和实施处理的过程中根本没有考虑到那些因素。

外部效度

在实验室的“人工”环境中进行严格的控制会增加研究的内部效度,但是通常也会降低研究的外部效度。所以自然环境中进行的实验有时可以用来确定实验室结果的外部效度。如果主要目的是为了检验一个特定的心理学理论而进行实验时,实验室结果的外部效度就没有那么重要了(Mook, 1983)。反之,在自然环境中所做研究的外部效度则通常是非常重要的,特别是要将实验结果作为大规模社会变革的理论基础时,比如推行控制酒后驾车的新规则或者改变投票人登记的新流程。一个有利于控制中西部各州酒后驾车的实验结果能够推广到



图 11.1 作为一个社会实验,“芝麻街”节目是为了改进成千上万儿童的教育而设计的

(续)

这个国家的其他地区吗？某所大学中一门改善学生批判性思维能力的课程也能够其他大学中发挥作用吗？当然，这些问题都与研究的外部效度相关。

目的

自然环境中的实验和实验室实验通常有不同的目的（见第2章）。实验室实验常常意味着单一目的的基础研究，为的是理解一种现象——确定“自然”怎样运转，实验目的在于获得知识。应用研究也是为了发掘现象产生的原因，但了解事件原因的目的在于改善现状。所以自然环境中的实验可能带有更多实践性的目的。

后果

有的实验会对社会产生深远影响，涉及到许许多多的人。针对智障儿童的“从头开始项目”和“芝麻街”电视节目，就是为了改进全国成千上万儿童的教育而设

计的社会实验。有的社会实验也会在自然环境中小规模地进行，比如在当地的学校或企业。企业进行实验可能是为了减少雇员的数量或者通过其他方法提高雇员的生活质量。与那些实验室研究相比，这些社会“实验”产生的后果也许更具有即刻的影响。当然，我们无法预期所有的后果，对于所有实验而言都是这样。

为了得到有意义的实验结论，我们必须考虑实验设计和这些“处理”的实施，以及尽我们所能去确定实验处理是否有效，这尤其重要。正如我们所看到的那样，这些实验常常花费很多金钱并且影响许多人的生活。相比而言，实验室实验的即刻效果可以是实质性的，但很可能非常小，可能只是直接影响一些研究人员和那些相对较少的被试的生活。实验室实验和自然环境中的实验对于研究者来说都是非常有用的方法，可是与其他方法一样，最好在我们理解了它们的优点和局限之后再使用。

那么就没有必要再检测这个方法了。在这种情况下，我们也认为不让控制组的学生采用新方法是不公平的。如果我们不了解新的方法是否更好，而且只有进行一个真实验，才能让我们确定新方法的效果，那么随机分配被试——你想叫它“抓阄法”也行——可能是最公平的方法。毕竟在新方法被认可和进一步完善之前，老的教学方式被认为是有效的。如果证明新方法不那么有效，那么随机分配实际上就“保护”了控制组的被试，使他们没有接受一种无效的处理。

其实保留对照组时也有很多种方法可以给所有被试都提供一种潜在有效的处理。一种方法是使处理轮流进行。比如，阿特金森（Atkinson, 1968）随机分配学生接受英语或数学的计算机辅助教学（“处理”），然后对两个组进行英语和数学的测验。每个组在测验中都同时作为另一个的控制组，因为它的成员并没有接受另一项计算机辅助教学。完成实验之后，可以给两个组都再进行他们先前未接触过的那项计算机辅助教学。这样所有的被试都接受了所有潜在有利的处理。

如果实验机构暂时无法向所有的人提供“处理”服务，那么设置一个合适的控制组也是行得通的。等待接受服务的人们可以成为一个“等候名单”控制组。然而必须把人们随机分配到这个“等候名单”。那些等待的人和先接受处理的人在重要的维度上应该没有差别（比如，更渴望得到处理）。要在实验组和对照组之间通过一种没有偏差的方法来分布这些特征，随机分配是必要的。

总有一些时候不能简单地使用随机分配方法。比如，在检测新药物的临床试验中，让病人同意随机分配到实验组或控制组（不施加处理的组）可能就会相当困难。正如你将要看到的那样，这些情况就只能使用准实验设计。本章将在后面介绍准实验设计的逻辑和程序。

控制真实验的内部效度

- 内部效度的威胁大多是那些对研究结果看似合理的混淆解释。
- 内部效度威胁的主要类型包括历史、成熟、测验、测试手段、回归、选择、被试缺损以及伴随选择的附加效应。

做实验之前，我们要考虑一下通过实验程序能够排除哪些可能的混淆解释。只有控制了这些混淆解释，我们才能得到一个确定的因果推论。前面的章节已经提到了许多未控制的因素，它们会作为混淆因素影响实验的内部效度（它们也被称为混淆）。前几章确认了几种类型的混淆（特别见第7章）。坎贝尔和斯坦利（Campbell & Stanley, 1966; Cook & Campbell, 1979; 也见 Shadish, Cook, & Campbell, 2002）¹已确认了八类混淆，他们称之为对内部效度的威胁（threats to internal validity）。前面已经讨论过其中的一些了。在分析完这些主要的内部效度威胁后，我们将能够评价众多实验程序对这些混淆因素的控制程度。

历史 如果在研究中某事件的发生而非实验处理造成了被试行为的改变，就会损害内部效度。一个真实验要求除了实验处理之外，对实验组和控制组的被试要同等对待（在实验中有相同的体验历史）。在实验室，通常可以通过平衡或保持实验条件恒定来实现。而在自然环境中做实验，研究者也许并不总是能够达到与实验室情境相当的控制程度，所以历史造成的混淆可能成为一个对内部效度的威胁。比如，假设你要测试一个批判性思维课程是否确实改变了学生们的思维，并且假设你只是考查学生们在课程开始和结束时批判性思维考试的成绩。在没有合适的对照组的情况下，可能是某些事件的发生而非实验处理（即批判性思维课程）提高了学生批判性思维的能力，所以历史（history）会成为对内在效度的威胁。举例来说，假设很多参加该课程的学生还浏览一个关于批判性思维的网页，而这并不属于课程。现在学生们的历史就包括浏览网页的经历，这会混淆实验处理并造成对该研究内部效度的威胁。

成熟 一个实验的被试必然会随着时间的变化而发生变化，如他们变得更年老、更有经验等等。随着时间推移的改变叫做成熟。比如，假设一个研究者想要评价采用新教学方法一年后儿童的学习情况，如果没有合适的对照，研究者可能把一组学龄儿童从学年初到学年末期间的成绩变化归因于教学干预的结果，可实际上这些变化也许只是因为成熟（maturation）而导致，这就威胁到实验的内部效度。

测验 前面的测验通常对随后的测验有影响。举例而言，很多学生经常会从一

¹ 在修订库克和坎贝尔（Cook & Campbell, 1979）关于准实验的经典论述中，沙迪什等（Shadish et al., 2002）加入“不确定的时间优先性”作为第九个对内部效度的威胁因素。虽然原因必然先于结果，但A在B之前这个事实本身并不真正表明A导致了B；另外，一些因果关系可能是双向的。我们没有在内部效度威胁因素中包含这一因素是因为我们认为：（a）介绍实验方法的逻辑时已经对这个问题进行了部分的讨论，并且（b）讨论将超出学生对大多数实验中简单的单向效应的理解。最后，在讨论选择以及工具、成熟等威胁的合并效应时，沙迪什等人说明了“附加的”比“交互的”在表述上更好的原因。我们也相应地改变了先前版本中讨论这些合并效应时使用的术语。

门课程的首次测验到第二次测验间取得进步。经过第一次测验后，学生对测验流程和教师的期望变得熟悉起来，这种熟悉感就在第二次测验时影响了他们的成绩。在心理学实验的背景下实施多个测验时（比如，前测—后测设计），如果处理效果不能和测验导致的效果分离开来，那么测验（testing）同样也是对内部效度的威胁。你能看出如果只是基于学生前测和后测的分数，就试图去评价一门批判性思维课程的效果会是多么有问题吗？

测试手段 随着时间改变的并不只是实验的被试（比如，成熟或者随着测验增加的熟悉度），还有用来测量被试表现的仪器。由观察者去评价行为时就更容易发生这种情况，例如，疲劳、期望和观察者的其他特点都可能造成观察者偏差。除非进行了控制，否则这些因素都可能造成**测试手段**（instrumentation）对内部效度的威胁。机械设备同样也可能随着重复使用而发生改变，作者认识的某研究者曾发现，学习实验中用来呈现材料的机器在实验开始和结束时并不是以同样的方式运作，在实验结束时进行的测量和实验开始时并不相同。所以表面上的学习效应有可能只是用来测量学习的设备发生改变所造成的。

回归 当个体被选择参加实验时，由于他们的“极端”分数，统计回归（regression）总是个问题。一个测验的极端值并不一定就像另一个测验那样极端。换句话说，一个非常差或者非常好的成绩后面（两者我们都遇到过），可能分别接着一个不那么坏或不那么好的成绩。比如，仔细想想以前你在课堂测验中得过的最好成绩，是什么导致了这次测验呢？可能是学习努力，但是也可能有一些运气，也许所有的因素加起来，恰好导致了一个极其优秀的成绩。如果我们需要做一个测验，可能考试材料恰好是你学得最用心的部分，或者考试形式是你特别喜欢的，或者考试时间是你感觉特别自信的时候，又或者所有这些情况都有。特别好的成绩是“极端的”，因为它们被偶然地夸大（超出我们通常的或典型的成绩）了。一个特别坏的测验成绩同样也可能因为坏运气而产生。当再次测验时（接着非常好或非常坏的成绩），几乎不可能所有的偶然因素又以相同方式共同产生影响，使我们得到那样的好分数或差成绩，这时我们很可能会看到一个成绩接近于所有分数的平均值，这种现象通常叫做**向平均数回归**。

比如有一组大学生在第一学期取得了很差的成绩（“前测”），现在想采取一些办法来提高他们的学业成绩，挑选一些获得极端值的被试（这个例子中是极不好的成绩），让我们假设随后实施了一个处理（比如，一个10小时的学习技能培训班），这时统计回归就会是一个对内效度的威胁，因为我们预期即使没有任何处理，由于统计回归的原因，这些学生的成绩在第二学期也会稍微好一点（“后测”）。一个无知的研究者可能会错误地把这种“回归效应”和“处理效应”混淆。

被试缺损 实验中的被试流失会造成对内部效度的威胁，例如实验中有些被试会在实验中途放弃继续参与。**被试缺损**（subject attrition）会因为被试的流失改变了群体的性质，使之不同于实验处理之前，比如破坏了通过随机分配建立起来的实验组和对照组的同等性。如果一个实验任务难度很大，且一些实验被试因不能完成任务而变得沮丧从而退出实验时，实验组中留下的被试就会和那些退出的被试有所不同（还可能与控制组被试也不同）。也许没有什么根本的不同，只是因为他

们能够完成实验任务（或者至少是坚持下来了）。

选择 研究开始的时候，一个群体与另一个群体之间就存在着差异，就会有因为选择（selection）而造成的对内部效度的威胁。除了组别而外，实验组的被试还可能在很多方面都不同于对照组。在实验室中常通过随机分配和平衡被试特征来处理这个问题。当在自然环境中做实验时，经常难以把被试随机分配到处理条件和对照条件。这些困难就使我们不能进行一个真实实验，从而可能因选择而威胁到内部效度。

伴随选择的附加效应 诸如成熟等影响内部效度的因素，会与选择因素结合起来，共同影响到内部效度，这可能需要得到额外关注。尤其当对照组不是随机分配形成的时候，更需要注意：（1）选择和成熟的附加效应，（2）选择和历史的附加效应，以及（3）选择和测试手段的附加效应。比如，如果比较大一学生组成的实验组和大二学生组成的控制组，就会产生选择和成熟的附加效应。第一学年学生的变化（随着他们更熟悉大学环境）可能比在第二学年时发生的变化更大。这些在成熟上的差异也许可以解释任何观察到的实验组和控制组之间的差异，实验处理所造成的结果就会被混淆。

当某个事件对两组被试产生了不同的效应，就造成了选择和历史的附加效应。两个整组之间比较时，也可能会存在这个问题，因为某种情况也许是某个组所特有的，这样某些因素对那个组的影响就会大于对另一个组的影响。举例来看，调查一个艾滋病宣传活动效果的研究中，涉及两个学校（一个有宣传活动，一个没有）。假设全国性媒体对艾滋病问题的关注同样程度地影响了两个学校的学生，可如果在此时期，一个学校有个患艾滋病的学生去世了，并且把消息刊登在学校的报纸上，我们认为该校学生就可能会受到与另一大学的那些学生不同的影响。在评价艾滋病宣传活动的效果时，就会发生选择和历史的附加效应。

最后，如果一个测试手段对一个组的行为改变比对另一个组的影响大，那么就可能造成选择和测试手段的附加效应。比如出现天花板效应或地板效应的时候，当仪器最初测得的一组分数太低（地板效应）以至于分数上任何下降都不能可靠地测量；或者太高（天花板效应）以至于任何升高都不能测量。可以想像到，如果实验组相对没有表现出变化（因为地板或天花板效应），而控制组则因其最初的平均值接近测量范围的中间值而表现出变化，那么就会对内部效度产生威胁。

真实实验的重大优点之一就是控制了所有这些对内部效度的威胁。正如坎贝尔（Campbell, 1969）强调的那样，如果可能就应该进行真实实验，但如果真实实验不可行再考虑进行准实验。“在我们能够做到的范围内，我们必须尽我们所能”（p. 411）。一方面需要关注实验处理效果以获得正确知识，另外一方面真实实验并不总是可行，准实验是这两个认识之间最好的妥协。

真实实验也可能控制不到的问题

● 如果研究中发生了实验污染、实验者期望效应和新奇效应，就会对内部效度造成损害。

专栏 11.2

实验的污染

实验中组间被试的相互交流会产生几种可能的效应,包括:(1)个体对接受不符合自己意愿的处理所产生的怨恨;(2)组间接受不同处理所产生的敌对;(3)实验组间处理的扩散(见 Cook & Campbell, 1979; Shadish et al., 2002)。

● 怨恨。假设已经把被试随机分配到控制组,假设控制组被试知道“其他”被试正在接受一个有益的处理。你认为控制组被试会做出什么反应?一个可能是控制组被试会感到怨恨和意志消沉。如库克和坎贝尔解释的那样,在工厂中,接受不符合自己意愿的处理的工人可能通过降低生产率来进行报复。在教育情境中,老师和学生也许会“丧失信心”或生气。“泄露”实验处理信息的这种效应可能使一个处理看上去比平常更好,因为控制组带着怨恨可能会降低其成绩。

- 敌对。另一种可能的效应是竞争或者敌对情绪,发生于控制组了解到另一组的好运气时。也就是说,控制组可能想降低自己和处理组之间的差异。如库克和坎贝尔指出的那样,当把原本一致的群体(如部门、工作小组、分支机构之类)分配到多种条件下时就可能发生这种情况。认识到另一组会因为与控制组的区别而使其看上去更好,于是控制组的被试就会“更加努力”以至于不会在比较中显得很差。
- 处理的扩散。还有另外一个实验污染就是处理的扩散。按照库克和坎贝尔的观点,这发生于控制组被试接受了给予处理组的信息并有意改变自己行为的时候。比如,控制组被试可能会运用那些呈现给实验组被试的信息来模仿他们的行为,这当然就减少了两组之间的差异,并且影响了实验的内部效度。

- 当组间被试交流有关实验的信息时,就发生了实验污染,这可能导致怨恨、敌对或处理的扩散。
- 新奇效应是因为一个新方法(比如实验处理)引发了兴奋、活力和积极性,从而改变了被试的行为。
- 当处理效应不能在特定的人群、情境、时间间进行推广时,就会对外部效度产生威胁。

在讨论准实验程序之前,我们应该指出,真实验也许不能控制好所有对实验结果解释造成损害的因素。虽然真实验消除了主要的对内部效度的威胁,但在自然环境中工作的研究者仍然必须提防另外一些额外影响。我们使用污染这个词来概括一大类威胁内部效度的因素。污染(contamination)发生于组间被试交流有关实验的信息之时。污染能导致一些有害的后果,如专栏 11.2 中所述。

真实验还可能会受到实验者期望效应的影响。实验者期望效应是指一个实验者无意中影响了实验结果。该效应会导致解释中的系统误差或数据记录的错误。当研究者偏差和期望导致观察、鉴别、记录 and 解释行为发生错误时,就发生了观察者偏差。控制观察者偏差或实验者效应的一些方法见第 4 章和第 7 章(比如采用双盲程序)。

新奇效应(novelty effects)发生在引进一个新方法或实验处理的时候(Shadish et al., 2002)。比如,有时工作场地发生了一点小变化或出现一些新事物,工作人员可能会因为他们工作环境的新奇变得兴奋和活跃。工作人员新得到的这种热情,

而不是处理本身,就可以导致行为发生变化。新奇效应的反面是中断效应,指的是一个创新或者一个新的工作流程中断了职工们的工作,使得他们不能保持通常的绩效。

一个特殊的新奇效应是霍桑效应,指的是工作群体因发现有“重要人物”对自己的工作感兴趣,而使自己的行为发生改变。事情发生于1924年到1932年的西部电气公司霍桑工厂,工厂位于美国伊利诺斯州靠近芝加哥的西塞罗市(Roethlisberger, 1977)。研究的目的是为了考察生产率和工作场所条件之间的关系。在一个实验中,改变车间的照明情况后考察了工人的行为表现。结果显示,研究期间实验组和控制组的生产率都提高了。虽然对这个效应的解释还存在一些争议(比如 Prasons, 1974),但霍桑效应一般是指被试觉察到有人对他们感兴趣从而改变了他们的行为。

作为霍桑效应的一个例子,一些囚犯被选来参与一项关于牢房条件改变和犯人对监狱生活态度两者之间关系的研究。如果囚犯们的态度发生了改变,那么这个结果可以归因于牢房条件的实际改变,也可以归因于囚犯们看见监狱管理部门关心他们,因而在品行上有所提升。在自然环境中工作的研究者必须意识到,被试行为的改变可能部分地归因于他们觉察到他人对自己感兴趣。因而,霍桑效应代表着一种特定类型的反应(即对自己正在被观察的知觉),我们已经在前面的章节讨论过了(第4章)。

除了对内部效度的损害所造成的问题之外,对外部效度的损害也可能削弱真实实验的效果。外部效度主要取决于我们的样本在人、情境和时间上与想要推广到的总体相比的代表性如何。通常采用随机分配来实现样本的代表性,可是因为随机样本的情况很少(见 Shadish et al., 2002),就很难说我们的被试样本、进行观察的地点或者测验被试的时间是所有人、情境、处理或结果的代表性样本。所以,研究者必须认识到实验自变量与被试类型或者实验涉及到的情境之间可能的交互作用。比如,来自于市区内学校志愿者组成的实验组和控制组在冬天观察到的情况,与当年春天郊区学校非志愿者得到的测试结果会是相同的吗?

库克和坎贝尔描述了几种方法以避免威胁外部效度,其中最重要的是要确定样本的代表性。可是他们指出,重复是对外部效度的最好检验。也就是说,他们认为确定外部效度是一个经验问题。因而在不同的时间和情境下,对不同类型的被试采用不同的实验处理来重复实验,就能很好地解决外部效度的问题。有时部分的重复能够成为一个实验的“组成部分”——如选择多于一个群体参加实验。举例而言,有个实验为了确定新教育方法的效果,既选择了来自于低收入群体家庭的孩子,又选择了来自高收入群体家庭的孩子来做测试,这样就可以使实验结果在两类群体之间推广,提高了实验的外部效度。

准实验

- 当真实实验不可行的时候,准实验提供了一个重要的备选方法。

- 准实验不像真实实验控制得那么严格，最为明显和代表性的是准试验没有随机分配。
- 当研究者进行准实验时，必须寻找额外的证据来消除对内部效度的损害。
- 单组的前测—后测设计叫做前实验设计，这是一种糟糕的实验设计，因为它几乎没有内部效度。

词典会告诉你前缀 *quasi-* 的一个定义是“类似的”，准实验程序类似于真实实验程序。总体来讲，准实验 (quasi-experiments) 包括某种干预或处理并且可以提供比较，但是不像真实实验控制得那么严格。正如随机是真实实验的特点一样，缺少随机也是准实验的特点。坎贝尔和斯坦利 (Campbell & Stanley, 1966) 解释说，当实验者没有办法做到一个真实实验所必要的控制时，就出现了准实验。

如果真实实验不可行的时候就可以考虑运用准实验，在此之前需要了解一些关于处理效果的知识。我们先前回顾过的坎贝尔和斯坦利 (Campbell & Stanley, 1966; Shadish et al., 2002) 关于对内部效度威胁的因素，可以帮助你巩固这方面的知识。另外，研究者还须准备寻找额外的证据，以排除准实验中没有进行控制造成的对内部效度的威胁。假设一个准实验没有消除历史因素对内部效度的威胁，研究者也许会说明在对情况进行了逻辑分析后，历史造成的威胁是不足信的。如果研究者能做到这些，那么就可以对这个准实验的内部效度做出一个更有力的辩论。研究者必须认识到准实验程序的特定缺陷，并且像侦探一样去提供任何能够克服这些缺陷的证据。当我们开始考虑恰当地使用准实验时，我们需要承认，真实实验和准实验的效力之间有很大的差异。在面对准实验程序所带来的解释上的问题之前，研究者应该竭尽所能使之接近一个真实实验的条件。

也许研究者在自然环境下进行实验时所面临的最严重的局限是，他们常常不能把被试随机分配到各个条件中去。如，实验对象是一个单独的组以及管理决策或实际操作中不允许随机分配被试的时候。举例而言，同班或同校的儿童或者某个特殊车间的工人就是这样一个能够接受处理干预，但不能随机分配被试的例子。

拓展练习

这个练习要求你思考一个单组前测—后测设计中可能存在的对内部效度的威胁：

一个心理学家对抑郁新疗法的效果感兴趣，他招募了一个样本，包括想减轻抑郁的 20 人。在研究的开始，他让所有被试完成一个关于他们抑郁症状的问卷。该测试的平均值是 42.0 (最高的可能分数是 63.0)，表明有严重的抑郁症状。(此测量中没有抑郁的个体一般分数范围是 0 到 10。)在接下来的 16 周里，这位心理学家对被试实施一种新的疗法。处理结束时被试再次完成抑郁问卷，后测的平均分数是 12.0，表明被试的抑郁症状显著降低，而且只有轻微的抑郁。于是这位心理学家得出

结论，认为这个新疗法是有效的；也就是说，处理改善了被试的抑郁症状。

采用单组前测—后测设计时，本来是不可能做出因果陈述的，不可能得出该心理学家的这个结论。为了理解为什么会这样，我们要求你思考一下这个研究中有哪一些潜在的对内部效度的威胁。

1. 历史效应可能对该研究的内部效度造成怎样的损害？
2. 成熟在这个研究中有可能起的作用？
3. 该研究是否可能存在测验和测验工具造成的威胁？
4. 统计回归可能对实验结果产生什么影响？

如果我们测量了这个群体处理前以及处理后的行为, 这样的—个“实验”可以如下表示:

$$O_1 \quad X \quad O_2$$

O_1 指第一次观察或前测, X 表示—个处理, 而 O_2 指的是第二次观察或后测。

如我们第 7 章所说的那样, 这种单组前测—后测设计是一种前实验设计或者可以更简单地称作—个糟糕的实验设计。这种前测和后测分数之间得到的差异可以归因于处理, 也可以归因于几种对内部效度的威胁因素中的任意—种, 包括历史、成熟、测验和测验工具所造成的损害 (还有实验者期望效应和新奇效应)。糟糕的实验并不能得出处理有效的结论。幸运的是, 还有准实验可以改进这种前实验设计。

非等控制组设计

- 在非等控制组设计中, 采用前测和后测方法来比较处理组 and 对照组。
- 如果这两组在处理前的前测分数上相似, 但是处理后的后测分数不同, 研究者就能够更加确信地判断处理产生了效果。
- 因历史、成熟、测验、测验工具和回归造成的对内部效度的威胁, 能够在非等控制组设计中得到控制。

如果满足两个条件: (1) 有一个“像”处理组的组能作为对照组; (2) 有机会获得处理组和对照组个体的前测和后测结果, 单组前测—后测设计就可以通过修正得到一个内部效度更好的准实验设计。坎贝尔和斯坦利 (Campbell & Stanley, 1966) 把满足这两个条件的准实验程序叫做非等控制组设计 (nonequivalent control group design)。我们不能说处理组和控制组的被试在所有重要的特征上都是等同的 (出现了—个选择造成的对内部效度的威胁), 因为对照组是基于前提条件挑选的而不是随机分配的。所以对两个组都进行前测来评价他们在因变量测量标准上的相似性就显得很必要了。非等控制组设计可以如下表示:

$$\begin{array}{c} O_1 \quad X \quad O_2 \\ \hline O_1 \quad O_2 \end{array}$$

虚线表示处理组和对照组并不是通过把被试随机分配到各个实验条件中而形成的。

添加—个对照组, 就能控制因历史、成熟、测验、测验工具和回归造成的对内部效度的威胁。简短回顾—下实验设计的逻辑就会更好地理解这个说法。我们希望以两个相似的组来开始实验, 然后—组接受处理而另—组不接受。如果之后两组的后测分数不同, 我们首先必须排除其他的解释, 然后才能得出处理导致了差异—结论。如果两组真的具有可比性并且都有相似的经历, 那么我们就能够认为历史、成熟、测验、测验工具和回归效应对两组产生的影响相当, 因而可以认为两组以相同的速率自然地发生着改变 (成熟), 受到相同的多次测验带来的效

应的影响,或者经历了相同的外部事件(历史)。如果对两个组而言这些效应都以相同的方式发生,就不可能用它们来解释后测结果中的组间差异了。所以,这些效应不再是对内部效度的威胁,研究者仅仅通过添加一个对照组就能提高因果推论的能力。但是这些因果推断的前提是研究开始时形成对照组,除了处理之外,两组经历有可比较性。在实践中这些都是难以实现的,如我们将看到的那样,伴随选择的附加效应所造成的对内部效度的威胁通常不能在设计中消除。

当你接近心理学研究方法课程的尾声时,你也许会感谢学习了非等控制组设计,有个实验考察了研究方法课程的学习对理性思考日常生活事件能力的影响(VanderStoep & Shaughnessy, 1997)。实验选择了学习研究方法课程进行比较的两个班级(恰好正在使用本书的人!),并将他们与学习发展心理学课程的两个班级进行比较。两类学生都在学期初和学期末接受测验,结果显示,研究方法课的学生比控制组的学生在理性思考方面有更大的进步。实验证明,学习研究方法课提高了学生对现实事件进行批判性思考的能力。

伴随着这个令人鼓舞的消息,让我们现在来详细地考察另一个采用非等控制组设计的研究。这会让我们更好地理解准实验设计的优点和局限。

非等控制组设计:兰格和罗丁的研究

- 准实验常常评价一个处理的整体效力,而处理常包含很多组成成分,接下来的研究就可以确定哪些成分对处理而言是关键的。

兰格和罗丁(Langer & Rodin, 1976)认为环境改变在一定程度上会给老年人带来失落、无助和低自尊的感觉,尤其当老人搬进疗养院时这个作用更加明显。虽然疗养院通常在生活上给予老人相当充分的照料,但是也经常提供了一个被兰格和罗丁称为“事实上决策功能缺失”的环境。即便是最简单的决策,比如什么时间起床,拜访什么人,看什么电影等等,也不需要老人们自己决定。在疗养院,很多或者大多数日常决策都已经为老人们决定好了,使他们几乎没有个人的责任和选择。

为了检验“缺少个人决策导致老年人在心理、甚至生理上衰弱”这样一个假设,兰格和罗丁在美国康涅狄格州的一个疗养院进行了准实验。自变量是向两组疗养院居住者传达不同类型的信息。其中一组强调让被试做出很多决策,决定如何安排他们的房间、探访、植物摆放以及电影的选择等等。这些居住者还会得到一株小植物作为礼物(如果他们接受的话),并且告知如果愿意就由他们自己照料这株小植物——这是责任引发条件;第二组是对照组,也把他们召集起来开个会,但传达的信息强调工作人员对他们的职责,这些居住者也会收到一株小植物作为礼物(不管他们选择要还是不要),并且告知护士会帮他们给植物浇水。

疗养院的居住者已经根据先前的安排被分配到不同的楼层和房间,而一些人已经在那里很长时间了。所以随机分配是不切实际的——而且可能也不是管理部门所希望的。因此研究者就选择了疗养院里两个不同楼层的居住者分别传达两种信息。用研究者的话说,选择楼层是“因为疗养院的主任、护士长和社工评估了这些居住者在生理上和心理健康状况以及先前的社会经济地位,认为这些方

面他们都是相似的”(Langer & Rodin, p. 193)。另外,在传达信息前1周和传达信息后3周给疗养人员发放问卷。问卷中有些项目是关于“他们对生活中的事件在总体上觉得有多少控制感,感觉有多幸福和多主动”(p. 194)。此外在传达信息之前和之后,每层楼的工作人员要对居住者在机敏性、社交性和活跃性这些特质上进行评定。研究者还通过举办一个竞赛,对社交兴趣进行巧妙的后测,竞赛要求被试猜一个大罐子里软糖的数量。如果疗养人员愿意参加这个竞赛,只需要在一张纸上写下他们的估计数量和自己的名字就可以了。因此这个实验有很多因变量来评定疗养人员的控制感、幸福感、积极性和兴趣水平等。

兰格和罗丁的研究很好地说明了非等控制组设计的程序。前测和后测结果之间的差异显示出,总体上责任引发组的疗养人员在处理后比对照组更幸福、更积极和更机敏。行为测量结果,如看电影的频率,也是责任引发组更高,虽然该组只有10个疗养人员参加了软糖竞赛,但对照组却只有一个人参与!研究者认为这些结果可能有实际意义,他们认为给予老人一些个人决策和感受竞争的机会将能够减少或者反转衰老等一些负性后果。

在讨论这个设计的局限性之前,你要注意兰格和罗丁研究的另一个特点,这也是很多自然环境下研究的特点。兰格和罗丁研究中的处理事实上有几个成分。如,工作人员鼓励疗养人员对很多不同的事情做出决策(如电影、房间等),并且给他们提供需要照料的小植物。但实验评定的是处理“包”,也就是说,评价的是处理整体的效力而非处理的单个成分。我们只知道(或者至少我们基于证据认为)包含所有这些成分的处理产生了作用,但不确切知道如果少了一些成分处理还会不会产生作用,或者其中某个成分是否比其他成分的作用更为重要。

自然环境下的研究通常以多个成分的处理为特点。此外,这种研究最初的目的常常是评定处理“包”的整体效应。所以,如果我们想要确定一个处理的关键成分,那么找到整体处理效应的证据可能就只是研究计划的第一步,这样既有实际意义又有理论意义。从现实角度来说,如果研究表明处理中只有一些成分对效应很重要,那么也许不太重要的成分就能够去掉,这可以使处理更加经济并且更易采纳和实施。从理论角度而言,确定处理的关键成分是否确实起关键作用是很重要的。当你听到研究显示出一个整体的处理效应时,你可以想想怎样用一些附加的研究才能揭示出处理效应中特定的关键成分。

影响非等控制组设计内部效度的因素

- 为了解释准实验设计得到的结果,研究者有必要确定研究中有哪些因素会对内部效度形成威胁。
- 当采用非等控制组设计的实验中包含伴随选择的附加效应、回归、观察者偏差、实验污染和新奇效应时,必须考虑其对内部效度的威胁。
- 虽然前测组是可比较的,但是这却不能保证组间在与研究结果相关的所有可能的方面都可以进行比较。

按照库克和坎贝尔(Cook & Campbell, 1979)的观点,除了那些因为:(1)

图 11.2 兰格和罗丁 (1976)

采用非等控制组设计来研究两种不同类型的信息对疗养人员行为的影响。因为没有进行“真实验”，所以研究者检查这个研究的特征以确定是否出现了对内外部效度的威胁



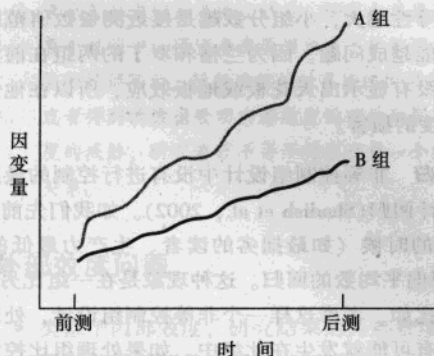
选择和成熟，(2) 选择和历史，(3) 选择和测试手段，以及 (4) 差异的统计回归这些因素造成的附加效应，非等控制组设计通常还需要对威胁内部效度的几种主要因素进行控制。我们会分析每个潜在的、威胁内部效度的因素如何给兰格和罗丁的结果解释制造麻烦，然后我们会解释兰格和罗丁又是怎样提出逻辑上的辩论和经验上的证据来反驳。我们还会检查实验者偏差和实验污染问题是怎样被控制的。最后，我们会对非等控制组设计中固有的确定外部效度时所面临的挑战做出简要的评论。

兰格和罗丁的研究中起初有一个重要发现，就是两组疗养院的居住者在前测结果中并没有显著差异。如果发现两个组在实验处理之前就有差异也不会让人惊讶，因为这些居住者并没有随机分配。即使前测分数显示组间不存在差异时，我们也不能认为两组是“等同的” (Campbell & Stanley, 1966)。我们会在接下来的讨论中解释为什么不能得出两组等同的结论。

选择—成熟效应 选择和成熟的附加效应发生于—组被试比另一组以更快的速度变得更有经验、更疲劳或更无聊时 (Shadish et al., 2002)。当处理组是自我选择的 (成员故意想要接受处理) 以及当对照组来自于一个与处理组不同的总体时，选择—成熟效应更可能损害内部效度 (Campbell & Stanley, 1966)。兰格和罗丁从同一个总体中随机地选择出他们的处理组和对照组 (但并非随机选择个体)，所以他们的设计就更加接近于一个真实验，但如果两组的个体来自于不同的总体就不会这样了 (Campbell & Stanley, 1966)。例如，如果把疗养院的居住者和针对老人的收容所居住者进行比较，或者如果疗养院不同楼层的居住者需要不同水平的看护，选择—成熟效应就更加可能发生。

(1) 即使处理组和对照组的前测分数均值相等，选择—成熟效应仍旧可能发生，因此我们还是不能得出两组等同 (具有可比性) 的结论。来自于不同总体的两

图 11.3 没有实验处理的情况下, 两组 (A 和 B) 可能的不同变化速度



组被试可能有不同的自然变化速度, 但是前测也许是在两组情况恰好差不多相同的时候进行的。图 11.3 举例说明了这个问题。组 A 的正常改变速度比组 B 更快, 但是前测却可能显示出两组没有差异。由于不同的变化速度, 两组可能在后测中表现出差异, 这可能会被误认为是处理的效应。不能仅仅基于前测中没有组间差异就得出两组具有可比性, 还有一个更为普遍的原因: 前测可能只在一个指标上或最多几个指标上对被试进行了测量。个体在一个指标上没有差异并不意味着他们在与该情境下的行为相关的其他指标上也不存在差异。

有任何理由去怀疑兰格和罗丁的研究中存在选择—成熟的效应吗? 也就是说, 接受处理的疗养者的自然变化速度比不接受处理的疗养者更快, 这样的判断合理吗? 几类证据表明这方面不存在问题。首先, 疗养院安排疗养者住在哪个楼层基本上是随机的, 而且这两个楼层也是随机分配到处理条件和非处理条件的。兰格和罗丁还报告说两个楼层居住者在社会经济地位和疗养院居住时间等指标上平均而言是相等的。最后, 虽然本身不是什么充分的证据, 但是这两个楼层的居住者在前测指标上没有差异。所有这些证据有力地表明, 在兰格和罗丁的研究中不存在因选择和成熟的附加效应而造成的对内效度的威胁。

选择—历史效应 另一个在非等控制组设计中没有控制的对内效度的损害, 就是选择和历史的附加效应。库克和坎贝尔 (Cook & Campbell, 1979) 把这个问题叫做局部历史效应, 它发生于当一个非实验处理的事件影响了一个组却没有影响到另一个组的时候。比如, 如果一件事只影响了疗养院一个楼层居住者的幸福感和机敏性, 而没有影响另一楼层, 那么局部历史就可能成为兰格和罗丁研究中的一个问题。你也许能想到很多的可能性, 例如一个楼层的护士换人可能导致居住者士气的升高或降低, 这取决于变化的性质以及新护士的行为与先前护士行为之间的差异。处理组和对照组个体所处的情境越是不同, 局部历史的问题也会变得越麻烦。兰格和罗丁并没有特别强调局部历史效应的问题。

选择—测试手段效应 如果在一个组中测试手段发生了改变而另一个组中则没有, 就容易造成选择和测试手段共同对效度的损害。比如, 天花板效应和地板效应可能使得对前测到后测行为变化的探测变得困难。如果这种情况只在一组中发生就更成问题, 即出现了选择—测试手段效应。沙迪什等人 (Shadish et al., 2002) 指

出,组间的不等性越大,小组分数越是接近测量数值范围的尾端,这种对内部效度的损害就可能越成问题。因为兰格和罗丁的两组在前测时没有差异,而且被试的行为表现也没有显示出天花板或地板效应,所以在他们的研究中似乎就不存在这种对内部效度的损害。

差异的统计回归 非等控制组设计中没有进行控制的最后一个对内部效度的损害就是差异的统计回归(Shadish et al., 2002)。如我们先前所说的那样,当个体是基于极端值选择的时候(如最拙劣的读者、生产力最低的工人、问题最严重的病人),就会出现向平均数的回归。这种现象是在一组比另一组更有可能出现回归的情况下发生。比如,考虑这样一个非等控制组设计,处理组由问题最严重的被试组成,回归很有可能就发生在此组中。如果处理组比控制组更可能发生回归,那么前测到后测的变化可能会被错误地解释为处理效应。因为兰格和罗丁的研究中两组来自于相同的总体,而且也没有证据表明一组的前测分数比另一组更极端,所以似乎也不存在因差异的统计回归所造成的对内部效度的损害。

期望效应、实验污染和新奇效应 兰格和罗丁的研究可能还受到另外三种对内部效度的损害——期望效应、实验污染和新奇效应的影响,它们甚至能够影响真实实验。如果实验中的观察者已经知道研究假设,他们就可能会无意识地把被试接受处理信息之后的分数评定得比之前更好。但因为所有观察者都不知道研究假设,所以基本上已经对这种观察者偏差或期望效应进行了控制。兰格和罗丁还意识到了可能的实验污染效应。控制组的居住者如果知道另一楼层的人有更多做决策的机会,他们可能会变得士气受挫。既然如此,使用疗养院不同的楼层就是有利的;兰格和罗丁指出,“楼层之间没有太多交流”(1976, p. 163)。因而差不多没有出现实验污染效应,至少没有达到会破坏研究内部效度的地步。

如果受处理楼层的居住者因为这种创新的责任——引发信息而变得热情起来,那么在兰格和罗丁的研究中就出现了新奇效应。因而,这种新的热情就可以解释任何的处理效应,而不是因接受处理的居住者责任感增加导致的结果。另外,对处理组给予特别关注可能会产生霍桑效应,即受处理楼层的居住者感觉自己更好。该研究很难完全排除新奇效应或霍桑效应。可是按照作者的说法,“两组得到的关注没有差异”(p. 194)。事实上,传达给两组的信息都强调工作人员照料他们并希望他们“快乐”。因而,在没有其他反面证据的情况下,我们可以下结论说兰格和罗丁观察到的行为改变是因为自变量效应而不是研究者没有控制好的无关变量效应。

在这一点上显而易见的是,研究者要确定一个自变量在特定的实验背景下是否“起作用”了,他或她必须有几分侦探的样子,去系统地收集和仔细地权衡有利于和不利于解释实验处理确实导致行为改变的证据。比如,像我们分析兰格和罗丁的研究那样,非等控制组设计中诸多对内部效度的损害必须根据前测和后测分数的实际结果来检查。通过深入细致地检查观测的结果,可以有很大把握来评估一个特定损害是否存在(见Shadish et al., 2002)。

正如库克和坎贝尔(1979, pp. 55~56)解释的那样:

评估内部效度是一个推理的过程,研究者必须系统地思考内部效度的每个威胁因

素是怎样影响结果的。然后,研究者要通过分析数据来决定哪个威胁可以被排除。在整个过程中,研究者都要当他自己的批评家,敏锐地检查所有可能的威胁。当所有威胁都被消除后,就能确信地对可能的因果关系做出结论。或许因为得不到合适的数据,或者得到的数据表明内部效度的确受到影响,总之,如果不能全部排除所有对内部效度的威胁,研究者就不得不得出这样一个结论:两个变量之间可能是也可能不是因果关系。

外部效度问题

- 类似于内部效度,研究结果的外部效度也必须得到批判性地检查。
- 研究结果外部效度的最好证明就是用不同的总体、情境和时间重复实验。

和内部效度同样,我们必须对一个准实验的外部效度进行系统的考察。有什么证据表明特定的结果只适用于特定的被试群体、情境和时间?比如,虽然兰格和罗丁建议看护老人的方式做出某些改变,我们还是可以质问责任引发信息的效力是否可以在不同的时间适用于所有的疗养人员和所有类型的疗养院。兰格和罗丁选择的那个疗养院是“由美国康涅狄格州评定的最舒适的看护单位之一”(1976, p. 193),表明其疗养者、工作人员和护士可能与其他疗养院的不同。如果这个疗养院的居住者比其他疗养院的居住者在进入疗养院之前相对更加独立(也许是因为社会经济地位的不同),那么搬进疗养院所体验到的改变可能就对他们有更大的影响,所以独立的机会对他们而言可能就更加重要。类似地,如果该疗养院的工作人员比其他疗养院的更能干,他们也许就能和居住者进行更有效的交流。

在最后的分析中,研究者必须准备用不同的总体、情境和时间重复实验结果以确定外部效度。应用于内部效度问题的推理过程也必须用来检查一个研究的外部效度。另外,我们必须了解一个事实,那就是一个研究不可能回答关于研究假设的所有问题。

间断时间系列设计

- 在简单间断时间系列设计中,研究者检查处理前后的一系列观察结果。
- 时间系列数据出现突然改变(不连续),就说明处理产生了效果。
- 在简单间断时间系列设计中,对内部效度的主要威胁是与实验处理同时发生的历史效应和测量(测试手段)的改变。

当研究者能在实施处理前后的一段时间内观察到因变量变化的时候,另一个准实验——简单间断时间系列设计(simple interrupted time-series design)就是可行的(Shadish et al., 2002)。该设计的关键是在实施处理前后进行间断性的测量。简单间断时间系列设计可以如下所示:

$$O_1O_2O_3O_4O_5 \quad X \quad O_6O_7O_8O_9O_{10}$$

在引进一个新产品、进行一个新的变革或者开始一个广告活动这样的情境下,

简单间断时间系列设计可用来评价一个处理的效应。坎贝尔 (Campbell, 1969) 在 20 世纪 50 年代中期考察了康涅狄格州一项社会政策产生的效应。政府制订了一个打击超速行驶的法令, 坎贝尔采用了一个间断时间系列设计来确定这个法令对车祸的影响。因为与车祸相关的统计数据通常由州行政部门来保存, 所以坎贝尔能够获得很多档案资料作为处理前和处理后的测量指标。除了车祸数量之外, 坎贝尔还关注超速违规的次数、驾驶员被吊销执照的次数以及其他与驾驶行为相关的数据。图 11.4 说明了打击法令前后因为超速行驶而被吊销执照的比例 (超速占有所有吊销执照的百分比)。图中有一个清晰的中断, 恰好是在处理出现的时候, 这种不连续性为处理效应提供了证据。事实上, 时间系列中的不连续性是处理效应的主要证据。

如坎贝尔所指出的那样, 因为渐变与随着时间的正常波动是区分不开的, 只有时间图中突然的改变能够解释法令的效应。不幸的是, 通常很少出现图 11.4 那样引人注目的变化。实际上, 坎贝尔于同一时间段对车祸的分析确实证明打击超速的法令起了作用, 但是车祸的变化并没有像吊销驾照的变化那样突然 (见 Campbell, 1969, 图 2)。

间断时间系列设计被用来评价 2001 年 9 月 11 日恐怖主义袭击和人们因飞行恐惧而避免乘坐飞机之间的关系 (Gigerenzer, 2004)。该研究的基本原理如下: 风险恐惧被定义为“可能性小但结果严重的事件, 比如 2001 年 9 月 11 日恐怖主义的袭击” (Gigerenzer, 2004, p. 286)。如果美国人为避免飞行的风险恐惧, 而改为开车到目的地, 那么就可能有更多的车祸。为了检验这个假设, 吉格仁则 (2004) 考察了 2001 年 9 月 11 日之后三个月美国交通部的数据, 还分析了在那之前五年中的数据, 比较车祸数量的平均值, 结果见图 11.5。

图中显示了 2001 年 (方框描述平均值) 以及之前五年中 (圆圈描述平均值) 12 个月内致命车祸的发生情况。另外, 前五年每个月的最高值和最低值也表示出来了 (每个均值周围的小棒)。10、11、12 月的车祸数据表明 2001 年致命车祸的

图 11.4 因超速行驶而被吊销的执照占有所有吊销执照的百分比
(摘自 Campbell, 1969)

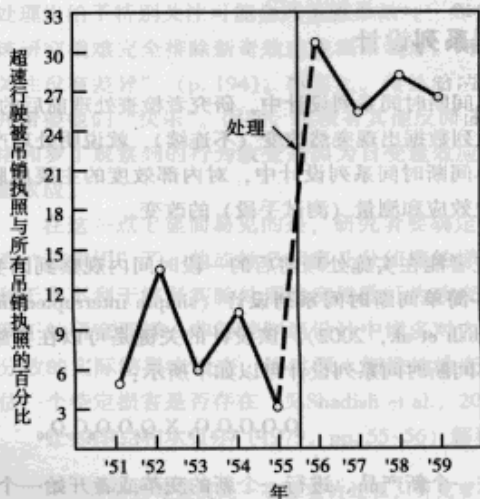
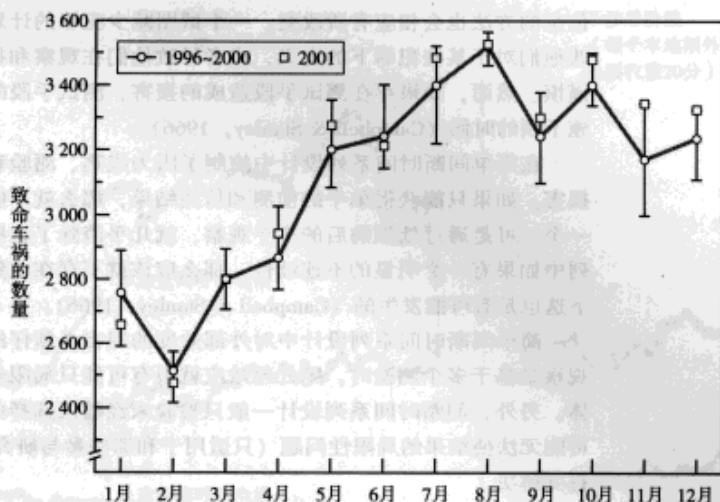


图 11.5 1996-2000 年和 2001 年在美国发生的致命的车祸数量。图中折线代表 1996-2000 年的平均值；均值周围的小棒表示那些年的最高值和最低值。小方框表示 2001 年各月的车祸数量，小圆圈表示 1996-2000 年各月的车祸数量（摘自 Gigerenzer, 2004）



数量与前五年最高值一样，甚至更高。基于这些数据（和统计分析），吉格仁则（2004）下结论说，2001 年 10、11 和 12 月由车祸造成的死亡比以往增加了 353 人。他把这种增加归因于 911 事件后美国人对于飞行的恐惧。吉格仁则还把车祸造成的 353 人死亡与 4 起飞机事故（当然地面的车祸事故更多）造成的 266 人死亡进行了比较，认为“如果能让公众更好地了解有关灾难事件的心理反应”，也许就能预防这种“心理引起的代价”（p. 287）。



虽然间断时间系列设计和其他准实验设计的结果有时能够通过观察数据图形的观察解释（比如见图 11.4），但通常需要进行统计分析。例如，吉格仁则（2004）采用了卡方检验说明 2001 年 9 月 11 日之后的车祸数量在统计上显著高于前五年的同期水平。我们先前在第 5 章和第 9 章提到了卡方检验，其他情况下可能需要用到更加复杂的分析。读者可以参考沙迪什（Shadish et al., 2002）的《一般因果推论的实验和准实验设计》以获得更多信息。

坎贝尔和斯坦利（Campbell & Stanley, 1966）总结了研究者采用简单间断时间系列设计时所面临的问题：“内部效度有可能出现问题，所发现的效应不是由处理引起，而是由其他因素导致了时间系列中的变化”（p. 39）。在这类设计中历史效应对内部效度的主要威胁（Shadish et al., 2002）。比如，是否可能因为其他因素而非“恐惧风险”造成图 11.5 描述的致命车祸的增加？

对时间系列设计内部效度特别有害的是自然循环的影响，包括季节性的变更（Cook & Campbell, 1979）。比如，分析康涅狄格州政府打击超速行驶的法令的效果时，坎贝尔（1969）就从邻州收集资料来排除因天气因素引起的可能变化，以巩固他对于该政策效果的研究结论。

测试手段也必须作为简单间断时间系列设计中对内部效度的损害来考虑（Shadish et al., 2002）。比如制定新的计划或社会政策时，档案的保存方式或收集



信息的方法也会相应有所改变。一个试图减少犯罪的计划可能导致权力机构去修改他们对于某些犯罪下的定义,或者导致他们在观察和报告犯罪活动时变得更加谨慎。然而,如果存在测试手段造成的损害,测试手段的改变必须的确发生在实施干预的时候(Campbell & Stanley, 1966)。

在简单间断时间系列设计中控制了因为成熟、测验和回归对内部效度造成的损害。如果只能获得单个的前测和后测结果,那么就不能排除这些损害中的任何一个。可是通过处理前后的多个观察,就几乎消除了这些损害。比如,在时间系列中如果有一个明显的不连续性,那么应该就不存在成熟效应,虽然在某些情况下这也是有可能发生的(Campbell & Stanley, 1966)。

简单间断时间系列设计中对外部效度的损害必须仔细检查。当处理前的行为观察是基于多个测验时,则处理效应就很有可能只局限于那些参与这些测验的个体。另外,间断时间系列设计一般只检验未经随机选择的单个的组,这个特点就可能无法使结果的局限性问题(只适用于和那些参与研究的人有相似特征的个体)得到解决。

非等控制组的时间系列

- 在非等控制组时间系列设计中,研究者在处理前后对处理组 and 对照组都进行一系列的观察。

可以通过在先前提到的非等控制组设计程序之后包含一个控制组来更好地提高间断时间系列设计的内部效度。对于非等控制组时间系列设计(time series with nonequivalent control group design),研究者必须找到一个能够与处理组匹配,并且也可以在实验组接受实施处理前后进行多次观察的组。这种设计如下表示:

$$\begin{array}{c} O_1 O_2 O_3 O_4 O_5 \quad X \quad O_6 O_7 O_8 O_9 O_{10} \\ \hline O_1 O_2 O_3 O_4 O_5 \quad O_6 O_7 O_8 O_9 O_{10} \end{array}$$

像以前一样,虚线用来表示控制组和实验组没有随机分配。非等控制组间断时间系列设计可以让研究者排除历史对内部效度的威胁。如先前提到的那样,坎贝尔(1969)把从邻州得到的车祸数据和康涅狄格州打击超速行驶法令后的车祸数据进行比较。康涅狄格州的车祸在打击措施之后显示出一个即刻的下降,而其他有可比性的州却没有显示出任何这样的下降。这个结果能够排除这样的观点:康涅狄格的车祸减少是由于有利的天气条件、改进的汽车设计或者其他任何康涅狄格州和邻州都有的因素所造成的。

如我们先前看到的,档案资料而非直接观察通常可以给时间系列设计提供基础。麦克斯威尼(McSweeney, 1978)关于“电话查询呼叫”问题的研究,就为非等控制组时间系列设计中档案资料的使用作了相当好的说明。电话查询呼叫花费电话公司的钱,他们只有通过对其他服务收取更多费用来支付这些花销。麦克斯威尼(1978)提到,有效降低电话查询呼叫的干预措施会使电话公司和消费者都受益。

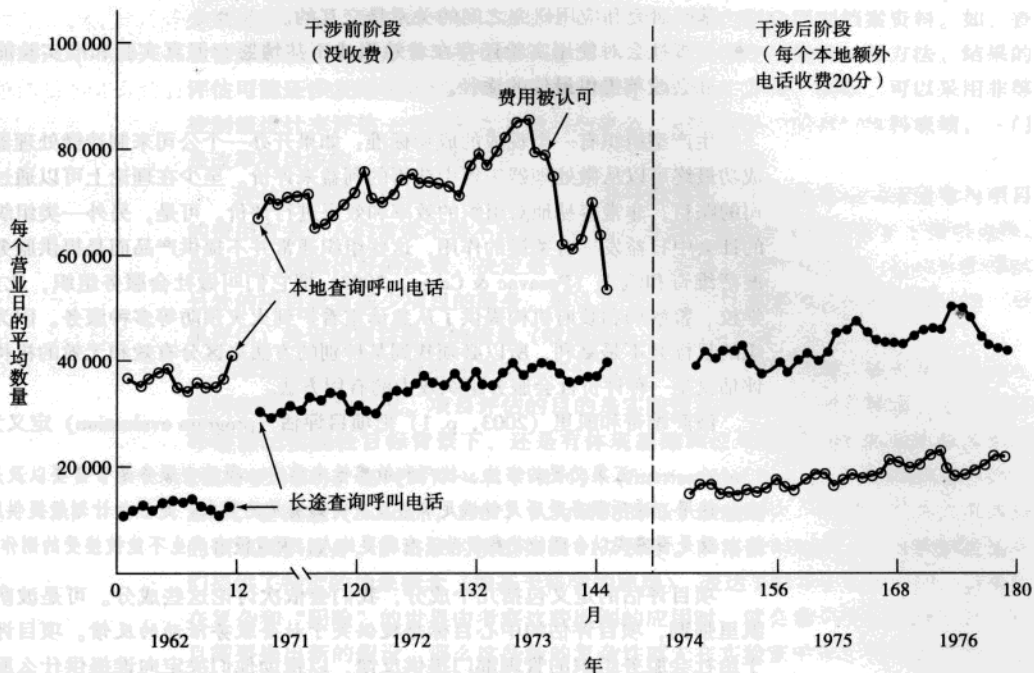


图 11.6 本地和长途查询呼叫电话在收费前后每个营业日的平均数量 (摘自 McSweeney, 1978)

1974 年, 美国辛辛那提贝尔电话公司开始对每月超出 3 个以上的电话查询呼叫进行收费, 每个额外的电话收取用户 20 美分。具体有多少人受到这项收费影响并不清楚, 但据估计它影响了超过一百万的用户。因为电话公司保留有大量的档案记录, 可以采用时间系列分析来评估该项干预措施的效果。另外, 简单时间系列研究可以通过使用非等控制组来扩大其效果。麦克斯威尼选择了从辛辛那提以外的区域打来的长途电话查询呼叫作为对照组 (这些电话不收费)。根据贝尔公司提供的记录资料, 麦克斯威尼作了时间系列图, 如图 11.6 所示。开始收费后本地电话的频率明显降低了, 而从该区域外打来的电话查询呼叫继续逐渐增加。时间系列中, 处理组表现出清晰的不连续性, 同时对对照组没有表现出任何的降低, 这就为“电话公司采取的干预处理带来了效果”提供了几乎无可争议的证据。随后的分析显示, 本地电话查询呼叫的减少使得每个消费者的电话账单平均节约了 65 美分。

项目评估

- 项目评估用来评定社会服务组织的效力并为管理机构提供关于服务的反馈。
- 项目评估者评定需要、过程、结果和社会服务的有效性。

- 基础研究和应用研究之间的关系是交互的。
- 尽管社会对使用实验还存在着勉强或不甚情愿，但真实验和准实验能为评价社会改革提供最佳的途径。

生产型组织有一个现成的成功标准：如果开办一个公司来制造微处理器，它的成功最终可以从微处理器销售中获得的利益来评价。至少在理论上可以通过检查公司的账目，非常容易地对组织的效率和效果进行评价。可是，另外一类组织在我们的社会中日益发挥着关键的作用，这些组织通常并不提供产品而是提供服务，所以波萨维奇和凯里 (Posavac & Carey, 2003) 把它们叫做社会服务组织，比如医院、学校、警察局和政府机构提供了从急诊室看护到火灾预防等多种服务。因为这些组织的目标并不是赢利，所以必须找到某种别的方法来区分有效和无效的机构。项目评估就是一种评价社会服务组织效力的有用方法。

波萨维奇和凯里 (2003, p. 1) 把项目评估 (program evaluation) 定义为：

……汇集必要的方法、技巧和敏感性来确定一项社会服务是否需要以及是否可能使用，这项服务是否足够满足那些没有得到满足的需求，是否如计划般提供服务，以及是否确实以合理的花费帮助了有需要的人，而且没有产生不能被接受的副作用。

项目评估的定义包括几个成分，我们会依次讨论这些成分。可是波萨维奇和凯里强调，项目评估的中心目标是提供关于社会服务活动的反馈。项目评估是用于给社会服务组织的管理部门提供反馈，以帮助他们决定向谁提供什么服务，以及怎样最高效率地提供服务。项目评估是一个综合性的学科，涉及到政治学、社会学、经济学、教育学和心理学。我们将在本章末讨论在自然环境中所做研究的项目评估，因为它代表了本书涉及的方法和原理在最大尺度范围内的应用。

波萨维奇和凯里 (2003) 确定了四个项目评估者需要提出的问题。这些问题是关于需要、过程、结果和效率的。对需要的评估试图确定某个机构也许可以向人们提供某项服务以满足他们没有得到满足的某种需求。比如，某城市政府收到一个针对社区老年人而制定的娱乐活动计划。这个城市首先会确定老年人是否确实需要这样的娱乐活动。如果他们的确想要，该城市就会希望进一步了解哪种娱乐活动对他们最有吸引力。调查法在评估需要的研究中被广泛使用。从需要评估中获得的信息可以帮助管理者来决定提供哪些娱乐活动。

一旦已经制定好一个项目，项目评估者就可以对已经确定的过程提出问题。在评估一个项目的过程时，观察法通常是很有用的。项目并不总是按照所计划的方式实施的，了解实施项目时正在确切地进行什么是很重要的。如果在专为老年人设计的娱乐活动项目中，那些计划好的活动没有被使用，这就表明可能项目没有适当地执行。与过程有关的问题（即关于一个项目是怎样在实际中执行的）可以让管理人员在实施服务的过程中做出调整，以便于加强现有的项目 (Posavac & Carey, 2003)。

项目评估者可能提出的下一阶段的问题涉及到结果。项目有效地达到其预定的目标了吗？比如，老年人现在有更多的娱乐活动而且对此感到愉快吗？这些特殊的活动是否比其他活动更受喜爱？一个为了抑制小区犯罪而设计的“社区—警戒”项目的结果，可以通过考查实施这个项目后入室盗窃和袭击次数是否确实减

少来进行评定。结果的评估可能会像第 6 章中所说的那样用到档案资料。如,查看警方的记录来确定各种犯罪率就是评估社区警戒项目效果的一种方法。结果的评估可能还涉及到在自然环境下进行的实验和准实验方法。例如,可以采用非等控制组设计来评估一个学校改革项目的效果,即比较学生的两门学科成绩,一门是改革项目中的,另一门则不是。

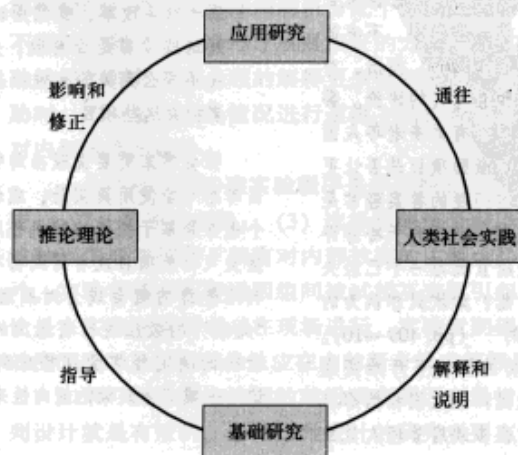
评估者最后可能提出的问题是关于项目效率的。关于效率的问题通常与项目的费用有关。人们常常不得不在那些政府或其他机构能够实现的服务中做出选择。如果我们想做出好的决策,决定是否继续某个项目,如何改进它,以及是否尝试另外的项目或者减少项目的服务,那么关于一个项目有多大成功的可能以及(结果评估)关于该项目费用的信息(效率评估)就是必要的。

本章前面和第 2 章描述了基础研究和应用研究之间的区别。项目评估也许是应用研究的极端情况。项目评估的目的是实践性而非理论性的。然而,即使在这种喧嚣的实践性目标背景下,还是有体现基础和应用研究之间交互关系的事例(见专栏 11.1)。在《国际心理学期刊》(*International Journal of Psychology*)的文章中,萨洛蒙(Salomon, 1987)认为研究的每个领域都以一种循环的方式为其他的领域提供服务。萨洛蒙的模型如图 11.7 所示。萨洛蒙特别提出,基础研究给我们提供了特定的抽象概念(如基于科学的原则),表达了特定的自然规律,当我们在复杂和“阴险”的世界中考察这些原则的应用时,就会意识到新的复杂性,而且需要提出新的假设。那么这些新的复杂性就先在实验室中进行测验和评估,然后再到真实世界中试验。

萨洛蒙把兰格的工作作为这种循环关系的一个具体例子。她发现一旦老年人进入疗养院,他们的健康状况就会下降(见 Langer, 1989; Langer & Rodin, 1976; 本章前面的例子)。这个观察结果使她提出一个理论,她在实验控制条件下对这个理论进行了检验,这对认知发展和教育理论而言是很有意义的(见 Langer, 1989, 1997; Langer & Piper, 1987)。该理论为她的应用工作提供了指导——设计新的疗养院模型。疗养院为了居住者的健康改变了看护方式,检验这种改变的实

图 11.7 基础和应用研究之间的交互关系

(摘自 Salomon, 1987, p. 444)



际效果无疑会修正她的理论。

根据坎贝尔 (Campbell, 1969) 的观点, 涉及到社会实验的政府官员们认为, 强调问题的重要性比解决问题更重要。他们必须清楚, 社会实验不是简单地付费购买一颗特定的万灵药 (这在大多数情况下几乎没有机会成功), 官员们必须允许尝试, 并且如果第一个方法失败的话, 还必须尝试不同的解决办法。换言之, 政府官员必须准备使用实验方法来确定社会问题及其有效的解决措施。

坎贝尔 (1969) 认为, 把社会改革和实验方法自然地结合在一起, 已经给社会政策的制定者带来了一些影响, 但这仍然是不够的 (见 Berk et al., 1987 和专栏 11.3), 其中的原因和他最初确定的一些原因是相同的。然而, 如果没有社会实验法, 特别是不能最大限度地使用随机现场实验, 那么社会政策制定者可能会普遍相信实际上不起作用的措施在发挥的作用, 反之亦然。这种不正确的决策会导致我们把经费和资源分配给无效的项目。

不久前, 电视上播出了一个叫“恐怖监狱 (Scared Straight)”的全国性节目。它描述了美国新泽西的罗韦州立监狱实施的一个青少年教育项目。该项目把青少年罪犯带到一个监狱, 让他们与监狱中挑选出来的罪犯见面, 目的是让这些青少年了解监狱生活的真实情况, 而项目的领头人也希望通过这个活动帮助青少年远离不法活动。没有确切证据表明该项目的效力如何, 但是有些人却认为成功率高达 80% 到 90% (见 Locke, Johnson, Kirigin-Ramp, Atwater, & Gerrard, 1986)。

专栏 11.3

从社会实验身上移走“无知的斗篷”

也许基础研究和项目评估之间最大的区别在于项目评估有政治和现实性。地方政府和国家不断地制定计划, 实施多种社会改革, 减税项目、工作激励项目、教育改革、警察部门改制和对老年人的医疗保障正是政府可能制定的社会改革项目中的一些类型。不幸的是, 曾任美国心理学会主席的坎贝尔 (1969) 指出, 对社会改革的结果通常不能进行有效和有意义的评估。警察技能的提高是否会减少犯罪率? 是否有更多老年人因收入降低而开始使用公共交通? 工作激励项目是否让更多失业者再次就业? 坎贝尔说, 这些问题的答案经常是找不到的, 因为大多数社会改革是在一个不利于冷静评价的政治气氛下制定的。比如, 政府官员与一个已经失败的项目联系在一起, 他们会怎么想? 如坎贝尔认为的那样, “在无知的斗篷下是安全的” (pp. 409-410)。另外, 许多社会改革是在一定成功的假设下开始实施的, 否则为什么要花费那么多税款呢? 让人们相信那样的假设而不必面对事情的真相, 对很多政府管理人员来

说是有利的。

坎贝尔 (p. 409) 进一步阐述了这一观点:

美国和其他现代化国家应该准备把实验方法应用于社会改革, 我们用这种方法试验新的项目, 了解这些项目是否有效, 并且在用多个标准检验后, 在不那么完美的有效性基础上保留、模拟、修正或者抛弃这些项目。

社会学家需要说服当局管理者, 在制定新的社会项目时尽可能使用真实验, 或者至少是准实验。比如, 一个也许是基于抓阄的随机程序, 可以用来决定哪一组人接受飞行员项目或者得到稀有资源, 没有接受项目或可利用资源的组会成为对照组, 然后就能够对一个社会“处理”的效应进行有意义的评估。现在, 关于谁得到什么的决定经常受到特殊群体的影响——比如通过游说——或者由政策的倾向性来决定。

这个项目只是全国许多类似项目中的一个，但这些项目真的有用吗？

对这个项目的几个评估研究得到了不一致的结果，有的认为结果明显有效，有的发现控制组和实验组被试没有差异，而有的研究表明该项目可能实际上在某些方面还增加了青少年犯罪的可能性。不那么严重的青少年罪犯可能在和成年囚犯见面之后增加了他们的犯罪活动，因为这些人最近开始显现出一种因为粗暴而被同伴认可的生活方式，这种粗暴也被囚犯经常投射出来的粗暴印象所强化。另一方面，犯罪频率更高的青少年罪犯，尽管在一段时间内在同伴中取得了一定地位，但是因为监狱意味着丧失这种地位，所以他们可能更害怕监狱生活（见 Locke et al., 1986）。

试图对这样的重大社会项目的效果进行评估，生动、清晰地表明了评估研究中固有的一些困难：随机分配被试的困难，让管理人员配合实验程序的困难，以及处理评估过程中被试流失的困难。然而，在合理的实验方法学基础上进行的项目评估，为政策制定者提供了所有层次（机构、社区、市、州、联邦）的信息，这能帮助他们在诸多解决社会问题的措施中做出更有远的选择。因为资源不可避免地处于短缺状态，所以有效利用资源是很关键的。

我们的愿望是，研究方法的知识能让你明智地参与到正在进行的关于社会实验作用的争论中去，也许还可以帮助你做出一些建设性的贡献。

小 结

自然环境下的实验在很多方面区别于心理学实验室中的实验。在自然环境下进行实验，有些是为了检验实验室结果的外部效度，有些是为了评价“实验处理”来改善人们工作和生活的条件的效果。

坎贝尔和其他研究者认为，社会必须愿意把实验方法应用于社会改革——用这种实验法对新项目的效果进行最明确的评估。很多情况下（比如，当可用的资源稀缺时）是不可能使用随机分配个体到处理和非处理条件中去的，如果真实验不可行，那么准实验程序就是最好的方法。准实验不同于真实验的地方在于：实验结果的其他可能合理的解释更少受到控制。当没有控制好对实验内部效度的威胁时，实验者就对其情况进行逻辑上的检查以及收集其他证据，来试着排除这些对内部效度的威胁。

一个特别理想的准实验程序是非等控制组设计。除去那些由于（1）选择和历史；（2）选择和成熟；（3）选择和测试手段以及（4）差异统计回归造成威胁，这种程序通常控制了所有对内部效度的主要威胁。除了对内部效度的主要威胁之外，实验者还必须警惕因组间被试相互交流引起的实验污染。在所有实验中，不论是在实验室警惕还是在现场进行，实验者期望效应（观察者偏差）问题、外部效度问题以及包括霍桑效应在内的新奇效应等都是潜在的问题。

如果可能在实施处理的前后观察到因变量指标的变化，那么简单间断时间系列设计就是有效的。采用这种设计的研究者要在时间系列中寻找一个与实验处理同时发生的突然改变（不连续性）。这种设计对内部效度的主要威胁是历史，即某

一个事件可能是时间造成的变化,而不是处理所引起。测试手段也可能成为一个问题,特别当处理是一种可能导致记录保留和数据收集方式发生改变的社会改革的时候。通过设置一个尽可能和实验组相似的控制组,就能提高简单时间系列设计的内部效度。比如非等控制组的时间系列就对很多可能的历史影响进行了控制。

在自然环境下进行研究的一个特别重要的目的就是项目评估。一些非心理学家的专业人员(如教育工作者、政治学家和社会学家)通常会参与到这个过程中来。各种各样的项目评估涉及到对需要、过程、结果和效率的评估。也许对项目评估最强的约束在于其周围的政治和社会现实。政府官员不愿意对社会改革进行评估,这经常成为难以逾越的障碍。然而,社会学家已经号召项目评估者去接触社会服务组织,我们应该积极响应这个号召,把最有效的服务带给最需要的人,以改善我们的社会。

重要概念

对内部效度的威胁

历史

成熟

测验

测试手段

回归

被试缺损

选择

实验污染

新奇效应

准实验

非等控制组设计

简单间断时间系列设计

非等控制组时间系列设计

项目评估

复习思考题

1. 举出两个理由说明为什么在自然环境下进行实验特别重要。
2. 解释实验室实验和自然环境中进行的实验在控制、外部效度、目的和结果上的区别。
3. 描述真实验的三个区别性特点,并且指出怎样按照这些特点定义自变量。
4. 当研究者试图在自然环境中进行实验时,他们不得不克服怎样的困难?
5. 列出两个可以让研究者随机分配被试而又能够让所有被试有机会接受实验处理的程序。
6. 描述并解释当实验污染发生时,控制组被试的三种可能反应所带来的后果。
7. 解释包括霍桑效应在内的新奇效应可能对研究者解释实验处理效应带来怎样的影响。
8. 库克和坎贝尔(1979)认为对外部效度的最好检验是什么?
9. 解释为什么在非等控制组设计中进行前测非常重要。
10. 解释对内部效度的威胁是怎样在非等控制组设计中得到控制的,并且描述该设计没有控制好的一个对内部效度的威胁。
11. 列出两个理由说明为什么我们不能下结论说非等控制组设计中的处理组和控制

组是等同的，甚至在两组的前测分数相等的情况下也不可以。

12. 解释在非等控制组设计中，历史对内部效度的威胁和“局部历史效应”之间的区别。
13. 在简单间断时间系列设计中，处理效应的主要证据是什么？该设计对内部效度的主要威胁又是什么？
14. 解释非等控制组设计和简单间断时间系列设计的结合是怎样减少对内部效度的威胁的？
15. 描述当评估者询问项目评估中四个典型问题的时候，针对任一问题各要搜寻哪些信息。

挑战性问题的

1. 有一个准实验用来确定多媒体教学是否有效。学习心理学导论的两个班都在下午由相同的讲师授课，其中一个班（处理组）讲师采用多媒体教学。而另一个班讲师讲授相同的内容但不使用多媒体。学生们注册这门课程的时候并不知道是否会采用多媒体教学，但两个班的学生并不是随机分配的。用一个包含两种形式的全面测验来评估学生学习课程后掌握的知识。这个全面测验能够作为一个可信和有效的测验来比较两个班的教学效果。在上课的第二天（前测）和期末（后测）对两个班进行测验。前测和后测中使用测验的不同形式。

- A. 在这个研究中使用了哪种准实验设计？
 - B. 讲师开始时想要进行一个真实验而不是准实验。如果你支持进行真实验来检验多媒体教学的效果，那么对随机分配的公平性做出批判性评价。
 - C. 解释讲师所采用的这种准实验设计为什么比只检验接受多媒体教学的学生的设计更有效。指出一个在仅检验接受多媒体教学的学生时不能得到控制而在本研究中得到控制的一个对内部效度的威胁。
2. 一个心理学家出版了一本关于离婚对男性、女性和儿童造成影响的书籍。她对离婚 10 年后的这种影响很感兴趣。她发现即使已经离婚 10 年，一半的女性和三分之一的男性仍然对离婚非常愤怒。虽然一半的男性和女性声称自己是快乐的，但是 25% 的女性和 20% 的男性仍然不能“让他们的生活重回正轨”。只有 10% 的离异家庭中前夫和前妻在 10 年后拥有快乐、满意的生活。最后，有一半以上父母离异的儿童，在他们长大成人后成就低并有自卑感。这些结果是基于一个在加利福尼亚州迈瑞县（一个富裕的郊区，居民大多教育良好）历时 15 年，涉及 60 对离异夫妻和他们的 131 个孩子的研究。解释如何用准实验设计来说明报告中的结果是因为离婚造成的。

3. 某大城市的警察局为了让警官及时了解法律的改变，需要在两种不同的培训方法中做出选择。一位聪明的警局管理人员决定进行一个研究来检验这两种方法。她决定做一个真实验，并且随机分配了 60 名警官到两个历时 6 个月的培训项目中去。最后对在两种培训方法下成功完成训练的警官实施一个法律知识的测验。完成项目 A 的 20 名警官比完成 B 的 28 名警官在测验中表现出更高的平均分数。该管理人员很聪明，并没有把这些结果作为两个项目效果的决定性证据来接受。就这个问题中仅有的资料，解释为什么她做出这样的决定。如果只有各 20 名警官（来自于最初每组随机分配的那 30 名）完成了这两个项目并且项目 A 仍然有很大的优势，那么她的决定会怎样改变。一定要指出她认为这些项目具有效果这一结论的局限性。
4. 一所小型大学有一个新的健身中心，决定引进一个加强教职工健康的项目。该项目计划在一个学期中进行每周 1 小时的训练。对下列每个关于项目评估的问题进行批判性评论。
 - A. 在项目的计划中，对需要的评估发挥了怎样的作用？
 - B. 一旦项目开始进行，关于项目过程的什么问题会有利于确保合理解释对项目结果的评估？
 - C. 如果不可能进行一个真实验的话，解释你会怎样检验此项目的效果。

拓展练习答案

1. 当一个事件能够解释被试的进步而不是处理导致进步时, 历史就成为一个威胁内部效度的因素。比如, 被试可能已经阅读了自助类的书籍, 尝试了草药, 向朋友或牧师交谈, 或者经历了任何潜在有利的“处理”。这些事件中的任何一个都可能带来抑郁的改善, 而不是心理学家的处理发挥了作用。
2. 成熟指被试随着时间而发生的自然的改变。我们所了解的可能是抑郁会随着时间好转。所以, 被试情况的改善可能反映的是抑郁随着时间的自然减轻, 而不是处理的效应。
3. 测验对内部效度的威胁发生于一个先前的测验效果影响了随后使用的检验。在这个研究中, 被试可能已经记住了他们先前对抑郁测量的反应, 并且也许想努力去表现出他们有所改善, 因而在后测中选择那些表示抑郁减轻的反应(即使他们并没有这么感觉)。测试手段对效度的威胁发生于用来评估想法、感觉和行为的测量随时间发生了改变。因为前测和后测中使用了相同的问卷, 这种威胁就不太可能发生了。
4. 当选择了前测分数很极端的被试时, 统计回归就有可能发生。在这个实验中, 被试是因为抑郁(抑郁测量的分数很高)而被选出来的。于是后测中表示情况有所改善的分数下降就可能是由于向平均数的统计回归, 而不是处理的效应。

挑战性问题1的答案

- A. 该研究采用了非等控制组设计。
- B. 学生们可能会觉得随机分配到两个班级是不公平的, 因为他们自己不能选择进入哪个班级。如果我们不知道多媒体教学是否有效, 那么随机分配就是最好和最公平的方法。
- C. 如果只是检测接受了多媒体教学的学生, 那么研究设计会成为一个单组前测-后测设计。前测-后测的单组设计中有几种因素会影响其内部效度, 这就是为什么把它叫做前实验设计或者一个糟糕的实验的原因。该研究中一个可能威胁内部效度的因素是成熟, 也就是说, 即使没有使用多媒体教学的班, 在课程结束时学生所掌握的心理学导论知识可能(我们希望)会有所增加。该研究采用的非等控制组设计控制了这种威胁, 因为任何由于成熟所造成的知识增加可能对两个班都是等同的。采用多媒体教学的班在前测到后测过程中的知识增加可以解释为教学的效果。

概 述

第12章

数据分析的目的在于帮助人们观察是否存在着有意义的联系和趋势

(Abelson, 1995)。各种心理多种多样。例如, 母亲对物品或人

存在表现出不同的反应。或者, 在一个研究中, 研究者可能发现

的曲线并不总是线性的。因此, 研究者需要能够识别和解释数据

中的模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则呈

现出另一种模式。因此, 研究者需要能够识别和解释数据中的

模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则

呈现出另一种模式。因此, 研究者需要能够识别和解释数据中

的模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则

呈现出另一种模式。因此, 研究者需要能够识别和解释数据中

的模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则

呈现出另一种模式。因此, 研究者需要能够识别和解释数据中

的模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则

呈现出另一种模式。因此, 研究者需要能够识别和解释数据中

的模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则

呈现出另一种模式。因此, 研究者需要能够识别和解释数据中

的模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则

呈现出另一种模式。因此, 研究者需要能够识别和解释数据中

的模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则

呈现出另一种模式。因此, 研究者需要能够识别和解释数据中

的模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则

呈现出另一种模式。因此, 研究者需要能够识别和解释数据中

的模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则

呈现出另一种模式。因此, 研究者需要能够识别和解释数据中

的模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则

呈现出另一种模式。因此, 研究者需要能够识别和解释数据中

的模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则

呈现出另一种模式。因此, 研究者需要能够识别和解释数据中

的模式, 并且能够解释这些模式。例如, 研究者可能发现, 在

某些情况下, 数据呈现出一种模式, 而在其他情况下, 数据则

呈现出另一种模式。因此, 研究者需要能够识别和解释数据中

第五编

分析和
报告研究成果

第12章

数据分析与解释 (I):
描述数据, 置信区间, 相关

第13章

数据分析与解释 (II):
统计显著性检验与分析

第14章

心理学的学术交流

第 12 章

数据分析与解释 (I): 描述数据, 置信区间, 相关

本章大纲

概 述

呈现研究结论

计算机辅助的数据分析

例证: 一个实验研究的数据分析: 均数比较

第一阶段: 了解数据

第二阶段: 概括数据

第三阶段: 用置信区间来判断数据所反映的意义

例证: 一个相关研究的数据分析

第一阶段: 了解数据

第二阶段: 概括数据

第三阶段: 为相关构建一个置信区间

小 结

概述

数据分析的首要目标在于确认我们的观察是否能够支持我们的观点和论断 (Abelson, 1995)。这些论断多种多样, 比如说, 母亲药物成瘾, 那么其孩子就更容易表现出学习困难; 或者, 某一个防止抑郁的实验项目是有效果的。不管这个论断是什么, 我们都需要进行仔细的准备, 以使我们的证据不仅在质量上有所保证, 并且能够以恰当的方式呈现。在进行定量研究的时候, 证据主要就是我们收集到的数据。为了进行有说服力的论证, 我们需要知道我们要从数据中寻找什么, 如何总结信息, 以及怎样才能最好地评估这些信息。

当然, 数据不可能凭空得来, 数据都是通过某种特定的研究方法获得的 (例如, 观察法, 调查法, 实验法等)。一旦在数据收集阶段出现严重错误, 我们就再也无法“挽救”这些数据了, 最好的办法就是重新收集。因此, 在收集数据并加以分析之前, 我们必须确保对以下几个方面进行细致的考虑: 研究假设 (关于行为的不确定性的论断) 的表述、正确的研究设计、适宜的测量方式以及统计的效力。当然, 我们还需要确保在数据收集的过程中将需求特征、实验者偏差、混淆变量或者其他由研究环境对实验的影响降到最低。简言之, 我们是要从一个“好”的研究中寻找数据, 这个研究既有较高的内部效度又有较高的外部效度, 同时兼备敏感性和可靠性 (见第 7~9 章)。

在一个完好的实验设计基础之上, 我们获得了可信的数据, 那么, 接下来需要做的是什么呢? 有三个相互独立且同时又相互联系的数据分析阶段: 了解数据, 概括数据, 以及证实数据所揭示的意义 (stage of data analysis: getting to know the data, summarizing the data 和 confirming what the data reveal) (参见专栏 12.1)。不管是进行一项观察研究 (见第 4 章), 还是一项基于量化数据的实验研究 (见第 7 至 9 章), 数据分析的开始两个阶段, 即了解数据和概括数据, 都是大致相同的。如果研究要分析的是两个变量之间的相关, 比如调查研究 (见第 5 章) 或者其他, 概括数据的过程会略有不同。我们将使用一些研究实例来说明数据分析的这些步骤, 包括那些一组或多组平均数分析的研究, 以及强调变量之间相关的研究。

对于第三阶段——证实数据所揭示的意义, 有两种不同且互为补充的方法, 一种方法是使用置信区间为总体参数的范围和做出估计的精确性提供证明。另一种方法基于虚无假设的显著性检验 (NHST)。两种方法都在第 7 章内进行了简要介绍, 而且, 正如我们所说, 这些方法是相互关联的, 但是, 他们之间也是有重要区别的。我们首先分别对它们进行介绍, 然后再阐述两者的信息如何能够得到整合。在这一章我们将讨论置信区间, 而第 13 章则讨论显著性检验, 同时, 在第 13 章我们还将讨论统计效力的概念及其与置信区间和显著性检验之间的关系。

例证

专栏 12.1

数据分析的三个阶段

对数据分析的三个主要阶段，我们可以进行如下描述：

1. **了解数据：**在第一阶段，我们希望能够熟悉数据。这是一个探索性的或称调查性的阶段 (Tukey, 1977)，我们仔细查阅这些数据，对它们形成一定的感觉，甚至如同有些人说的，与它们“交朋友” (Haglin, Mosteller, & Tukey, 1991, p.42)。我们要问的问题是：这些数字中有什么？这些数据中有错误吗？这些数据是否有意义？或者有没有理由怀疑其可靠性？(Abelson, 1995, p. 78) 在这一阶段，用图示方法初步分析数据的分布特点是非常重要的，即这些数据看上去是什么样子的？只有当我们熟悉了数据的大致特征，检验了其中的错误，并且确信这些数据是有意义的，我们才能进入数据分析的第二阶段。
2. **概括数据：**在第二阶段，我们对数据进行有意义的概括。在这一阶段，描述统计和分布图是非常重要的。应该怎样来组织数据？哪一种描述和概括的方法能够为我们提供最多的信息？我们感兴趣的因素在研究中起到了怎样的作用？我们能看到一些什么趋势吗？对数据进行初步的概括之后，我们就可以进入判断阶段了。
3. **证实数据所揭示的意义：**在第三个阶段，我们需要知道数据究竟告诉了我们什么。这些数据是不是证实了我们在研究开始时提出的研究假设？基于数据

证据我们能得出什么？有时我们需要的是一个类别判断，或者做出是与否判断，这时我们就像法官一样做出裁决。那么，我们有足够的证据进行宣判吗？这个效应是真实的吗？如果有人认为我们得到的结果只是一个偶然性的产物，那么在这个阶段我们就必须使用各种统计技术来反驳这一言论。在这一阶段的分析中，我们会在适当的情况下进行虚无假设检验。但是，我们评估数据并非总是对数据做类别判断 (如，Schmidt, 1996)，换句话说，我们并非一定要对结果的“真相”做一个确定性的陈述。我们的观点可能是基于概率，对变量效应在一定范围内能够起作用做出评估。换句话说，当某个变量存在的时候，某个结果发生的可能性会有多大？对于这类评估，我们特别推荐使用置信区间 (如，Cohen, 1995; Hunter, 1997; Loftus, 1996)。

判断的过程其实开始于数据分析的第一阶段或称探索阶段，即当我们首先对数据形成一种感觉的时候。在我们检验数据的大致特征时，我们就开始评估我们得到的东西。在概括阶段，我们对数据的趋势和分布获得了更多的了解，这为我们证实假设提供了基础。数据分析的最后一步称为证实阶段，特别用以强调我们是在这个阶段得出一个结论：数据到底表明了什么。但是，对这个过程而言，数据分析的每一个阶段获得的信息都是有用的 (Tukey, 1977)。

呈现研究结论

- 当数据分析完成后，我们必须用一个完整的结论来解释我们的发现，反驳对立的观点，并说明我们的结论是站得住脚的。

对行为研究结果做出一个有说服力结论需要的不仅仅是数据分析。完整的故事才有说服力。对一个律师来说，为了赢一个案子，不仅仅要让陪审团对案件的事实高度重视，还要将这些事实组织成一篇连贯的、有逻辑的陈述。如果所有证据都指向嫌疑犯是男管家，那么我们就需要让别人信服“为什么”是男管家 (而

不是厨师)做了某件事情。埃布尔斯 (Abelson, 1995, p.13) 对此也提出了相似的观点:

要使一个统计论证有最大的说服力, 高质量的证据、对整体效应清晰完整的表达是必要的, 但还远远不够, 同样重要的是一个完整的研究故事。

所以, 数据分析完成的时候, 我们需要用整体连贯的陈述来解释我们的发现, 反驳对立的观点, 并说明我们的结论是站得住脚的。在第13章和第14章, 我们将告诉你如何为你的研究完成一篇恰当的陈述报告。

计算机辅助的数据分析

- 研究者主要使用计算机来完成数据的统计分析。
- 使用计算机软件进行数据分析要求研究者对研究的设计和统计有清楚的认识。

大多数研究者的电脑都安装了统计软件。使用统计软件包完成统计并解释结果是一项基本技术, 研究者必须学会。一些较为流行的统计软件包大家都耳熟能详, 比如 BMDP, SAS, SPSS, 以及 STATA, 你应该有机会在心理系或者学校计算机中心的电脑上使用这些程序。

用计算机软件进行统计分析要求研究者对研究的设计和统计有清楚的认识。在第7、第8和第9章, 我们介绍了各种实验设计。如果你要用计算机辅助分析, 这些知识是必不可少的。电脑不知道你使用的是什么样的研究设计, 也不知道使用这种设计背后的逻辑根据 (尽管一些比较友好的程序会给出提示帮助你思考)。为了进行计算机分析, 你必须输入一些信息, 包括你使用的是哪一种类型的实验设计 (例如, 随机区组或者重复测量); 自变量的数目 (如, 单因素或者多因素); 每一个自变量有几个水平; 因变量的数目和对每个因变量的测量结果。你还必须能够清楚地阐释你的研究假设, 并且能够决定使用哪种统计方法来检验你的研究假设。

例证: 一个实验研究的数据分析: 均数比较

你认识多少单词? 或者说, 你的词汇量有多大? 在准备大学入学考试如 SAT 或者 ACT, 或者准备研究生入学考试 GRE 或 LSAT 的时候, 你也许会问自己这个问题, 因为这些考试都强调词汇量。然而, 测量一个人的词汇量水平是一个复杂的任务 (例, Anglin, 1993; Miller & Wakefield, 1993)。测量时会出现很多问题: 比如, 我们如何定义“单词”? “play, played, playing”应该算一个还是三个单词? 我们是否需要考虑那些技术性或者专业用词, 比如六个音节的化学合成物质的名字? 那些自造的合成词、宠物狗的名字、或者你对另一半的亲昵称呼在测量

时又需不需要考虑在内呢？一个比较直接的测量方法是问个体认识英语词典里面多少单词，但即使是使用这个方法我们也会遇到问题：词典的收词量和词汇范围各不相同，导致测量的结果会因为抽取词汇样本所使用的词典不同而出现差异。另外，测量词汇量的结果当然也会因测量方法的不同而不同，比如多项选择测验得到的词汇量大于要求写出单词定义的笔试得到的结果。

本书的作者之一对词汇量的问题很感兴趣，他进行了一项测量大学生和老年人的词汇量的研究（见 Zechmeister, Chronis, Cull, D'Anna, & Healy, 1995）。该研究从一本中等收词量的英语词典中分层随机取样得到 191 个单词，然后准备了五选一的多项选择测验。在测量题目中，单词的正确解释总是和四个诱饵或者分心选项一同出现，以使分辨出正确答案更困难。例如，回答者要在以下几个选项中确定“chivalry”的正确解释：a. warfare, b. herb, c. bravery, d. lewdness, e. courtesy。随机选出的单词测验题目被印在小册子上呈现给被试，被试由 26 名平均 18.5 岁的大学生和 26 名平均 76 岁的老年人组成。根据以往的研究，年长的被试组应该比年轻的被试组在词汇测验中取得更好的成绩。

我们将用这个词汇量研究的实验数据来说明数据分析的三个阶段。

第一阶段：了解数据

- 数据分析的第一步是检查数据的整体特征，对数据进行必要的编辑，或者说“清理”。
- 注意小心地检查数据中的错误，比如数据缺失、不可能的值（如超出了量表给定范围的值）以及极端值。
- 茎叶图可以有效地将一套数据的整体特征直观呈现出来，并有利于探测极端值。
- 数字的、图形的或者言语陈述的方式都可以有效总结数据，但好的数据描述通常会三种方法都用上。

清理数据 我们从检查数据的整体特征，并对数据进行必要的编辑或“清理”开始（Mosteller & Hoaglin, 1991）。我们需要认真检查数据错误，比如数据缺失、不可能的值（如超出了量表的给定范围的值）。数据出错可能是因为被试在测量时误用了量表（如搞反了重要性的排序），或者因为有人在将数据输入电脑的过程中遗漏或颠倒了数字的位置。在输入文稿的时候，很多人都是靠软件的拼写检查功能来发现排印错误和拼写错误的，然而，没有这样的工具来帮助我们检查输入电脑的数字错误（不过，关于减少错误的建议可参见 Kaschak & Moore, 2000）。研究者有责任在进行下一步的数据分析之前确保数据已经“干净”。

探测数据中的异常数据和错误特别重要。出现异常数据有时是因为数据记录出错，比如在 7 级评分量表的数据中出现数字 8，或是以大学生为样本的智力测验结果中出现某人的智商只有 10 分。异常数据的另一个来源是出现极端数据。极端数据是指那些看上去无法融入数据主体部分的数据，即使它仍属于可能数值范围之内。例如，在一个反应时研究中，我们预期大多数反应时都小于 1500ms，却出现了一个 4000ms 的数据。如果绝大多数的反应时都小于 1500ms，那么这个

4000ms 的数据当然可以被看作一个极端值。但是这个数值在反应时研究的一些情况下, 比如被试打了个喷嚏、因走神没有看到呈现的刺激, 或是以为数据收集已经结束正准备离开等情况, 仍然属于可能的数值范围之内。当被试填一份问卷的时候也可能因为读错了问题而做出极端的回答反应。探测极端值的时候, 研究者并不能仅仅根据定义来判断, 一些经验判断的原则也非常实用(可见 Zechmeister & Posavac, 2003)。

当异常数据集中出现时, 我们需要决定是否将这些数据剔除。当一个异常数据很明显是错误数据时, 研究者应该将它改正过来或者将其剔除。但是, 如果研究者剔除了数据, 就必须在实验结果中予以报告, 如果可能的话还应该解释数据为何出错。以哈登(Haden, 1998)进行的一个检验母亲与儿童对话风格的研究为例, 研究者选择了24位母亲, 她们都分别有一个70个月大的孩子和一个约40个月大的孩子。我们只需要知道研究中的这个细节: 每名被试都经历了四次家访, 每次家访之间间隔几天。由于其中一个家庭接受家访的时间安排出了问题, 对该家庭的第三次和第四次家访之间间隔了两个月。在初步数据分析的过程中, 研究者发现该家庭的数据出现了许多极端值。于是哈登决定将这个家庭的数据从数据集中剔除, 并在她发表的论文中对该家庭的情况和剔除数据的决定进行了合理的解释。

在数据分析的第一个阶段, 我们还需要对数据的分布进行有意义的描述。数据的离散程度如何? 他们的分布是正态还是偏态? 数据分析第一阶段的任务之一是, 确定在进一步分析之前是否需要转换数据。数据转换是一个“重新表达”的过程(Hoaglin, Mosteller, & Tukey, 1983)。简单的数据转换包括将英尺转换为英寸, 将摄氏度转换为华氏度, 将正确率转换为正确个数, 等等。有时也会用到一些比较复杂的数据转换。

认识一组数据最好的方法是对数据作图。用电脑进行数据分析的一大优势就是可以用多种图形呈现数据, 比如使用频数多边形或者柱状图等。作图的过程快速简便, 就像简单的度量衡转换一样, 通常只需要用不同类型的图操作几次我们就可以熟悉这个方法了。什么图可以最全面地呈现数据中的信息? 通过比较不同坐标的曲线, 我们可以发现数据中的什么信息呢? 多边形和柱状图之中, 哪一种可以表现出更多的信息呢? 一张图不仅比得上1000个词的表述, 更可以快速地总结出1000个数字所包含的信息。当我们逐渐熟悉了数据的各种图形, 就会发现哪些作图方法相对来说更好。

再以泽克迈斯特等人(1995)所做的词汇量研究的数据为例, 由于在做选择题时, 被试即使不知道单词的正确解释也可能碰巧选对了答案, 因此研究者需要修正个体测试中可能存在的猜测。但是后来发现老年组被试使用的单词测验中有两个印刷错误, 因此研究者剔除了这部分数据。因为这些问题, 两组被试回答的总数就不一样了, 因此需要将数据转换成正确率百分比来协调这个差异。

在清理数据之后, 研究者们在第一阶段的分析中得到了如下数据。这些数据以大学生和老年被试在选择题上的正确率来表示。

大学生组 ($n=26$): 59, 31, 47, 43, 54, 42, 38, 44, 48, 57, 42, 48, 30, 41, 59, 23, 62, 27, 53, 51, 39, 38, 50, 58, 56, 45。
 老年组 ($n=26$): 70, 59, 68, 68, 57, 66, 78, 78, 64, 43, 53, 83, 74, 69, 59, 44, 73, 65, 32, 60, 54, 64, 82, 62, 62, 78。

茎叶图 茎叶图 (stem-and-leaf displays) 对于将数据总体特征形象化以及探测极端值是非常有效的。(Turkey, 1977)。茎叶图的命名来源于人们习惯于将数值前部的数字 (此例中为十位上的数字) 称为“茎”, 并将拖尾部分的那些数字称为“叶”的习惯。

下面是我们在词汇量研究中大学生组数据的茎叶图:

2*	3
2	7
3*	01
3	889
4*	12234
4	5788
5*	0134
5	67899
6*	2

此例中前部的数字指的是第一位 (十位) 上的数字 (例, 2-, 3-, 4-), 尾部数字就是跟在前部这些重要数字后面的个位数字 (例, -5, -6, -8)。这种图将前部数字按照顺序垂直排列, 小的在最上面, 大的在最下面。尾部数字按照从小到大的升序排在整个分布中前部数字的后面。整个图中, 每一行都是一根茎和它的叶子 (Tukey, 1977)。比如说, 在上面这个图中, 3 这个茎上就有三个叶子, 分别为 8, 8, 9, 表明分布中有 38, 38, 39 这三个数值。根据惯例, 如果有许多数值呈现, 或者当整个数据集中只有极少量的前部数字时, 通常会在前部数字后面加一个星号 (*), 以表示这是整个区间的前一半 (Tukey, 1977)。例如, 5* 就是茎加上叶子 0, 1, 2, 3, 4 (即, 数值 50-54); 而前部数字 5 (不带 *) 就是茎加上叶子 5, 6, 7, 8, 9 (即, 数值 55-59)。在上图中, 茎 2* 有一片叶子, 3; 而茎 2 上有一片叶子, 7, 分别代表数值 23 和 27。

有时候, 茎部数字不止一个。例如, 如果数值的范围是从 50 到 150, 那么对于小于 100 的数值来说, 茎部数字只有一个 (8-, 9-, 等等), 而对于大于或者等于 100 的数值来说, 茎部数字就有两位 (10-, 11-, 12-, 等等)。

你可能已经发现了, 茎叶图横过来之后就是一个柱状图。但是, 它比柱状图有更多的优势, 因为在这里每一个数值都能够被看到; 因此, 茎叶图不会遗失特定信息, 而使用类别区间的柱状图则做不到这一点 (Howell, 2002)。而茎叶图最为重要的优点则是它能清晰地显示出整个分布的形状, 并且非常便于我们发现极端值。

仔细观察一下词汇研究中 26 个大学生数据的茎叶图。你看到了什么? 整体形态是正态分布的吗? (即, 对称且呈“钟形”), 还是偏态 (数据分布不对称, 偏向一个方向) 的? 数据是比较分散的, 还是集中围绕在某个特定数值周围? 有没

有反常的数值? 我们认为, 这个茎叶图显示数值大多数围绕在 40% 和 50% 这个部分, 而整个分布有些偏态。没有发现极端值(例如, 没有低于 10%, 也没有超过 60% 的数)。

在对两组数据进行比较时, 我们可以将两个茎叶图并排放在一起, 这样能够使很多问题显示出来。请看下面的这张图: 两组数据使用同一个茎, 一边的尾部数字从右向左递增(例如, 997 5), 而另一边的叶子(同一行中)从左向右递增(例如, 5 67899)。这表明在第一个分布中有 57, 59 和 59, 而在第二个分布中有 56, 57, 58, 59, 59。并排茎叶图可以被用于比较两组人群在问卷项目上的回答情况, 比如一个研究者想要比较两组人在社会经济地位、年龄、性别以及其他各个项目上的区别时, 他/她就可以使用茎叶图。

下图是词汇研究中两组被试的并排茎叶图:

老年人组	大学生组
	2*
	2
2	3*
	3
43	4*
	4
43	5*
997	5
44220	6*
98865	6
430	7*
888	7
32	8*

看一下图的左边部分, 这是老年人被试的数值。它有什么特点? 数据整体上来说, 是接近正态分布的, 虽然有一个极端值和一个异常值——“32”与其他数值看来差异较大(有些电脑程序能够自动地调整异常值)。在没有关于参与问卷者更多信息的情况下(比如, 那一天是不是服药过量, 或者有没有阅读障碍), 实验者没有理由将这样的数据从研究中删除。这个异常值的存在使得该组的变异性大大增加。尽管如此, 我们还是应该承认有些数据集本身就具有较大的可变性。比如说, 在这个研究中, 可能老年人组被试比大学生组被试样本更为不同质。(有一句格言: 在收集数据的同时尽可能多地收集一些关于被试的信息。要把极端值当作真实值来处理, 除非你能够确定这个极端值是由于错误或者与研究无关的环境因素造成的。)

现在来看一下并排茎叶图显示了两个分布的哪些信息。你应该能够立刻看到这两组数据有一定程度的重叠, 但是, 老年人组与大学生组相比, 获得更多 60 分以上的成绩。这个数据“图”已经开始显示出, 在词汇量的测试上老年人组被试的表现整体上要好于大学生被试。

小结 在数据分析的第一阶段——了解数据这一过程——我们应该确认:

- (a) 数据中错误出现的性质和频率如何。如果确实有错误存在的话, 考虑一下有没有办法纠正, 或者是否可以将这些数据丢弃;
- (b) 如果有反常数据, 或者不可能值, 需要分析一下是什么原因导致这些数值出现以及应该如何处理它们(保留或者丢弃);
- (c) 数据分布的整体特征和形状如何;
- (d) 有没有其他更好的表现数据分布特征的方法。

第二阶段: 概括数据

- 计算集中趋势, 包括平均数、中数和众数。
- 衡量离中趋势和变异的重要指标是全距和标准差。
- 平均数的标准误是理论抽样分布中平均数的标准差, 它用于测量我们对总体平均数估计得究竟如何。
- 效应大小是非常重要的测量指标, 因为它们能够排除样本容量大小的影响, 为自变量和因变量之间的关系强度提供信息。
- 在对两个均值进行比较时, 测量效应大小的一个重要指标是科恩 d 值。

可以用数字、图形或者文字来概括数据, 好的描述往往是三种方式并用的。在这一章里, 我们将集中讨论用数字进行概括的方法, 也就是描述统计, 不过我们同时也会使用到图形。在第 14 章中有关于如何用图形概括数据的内容, 同时用文字描述数据也是第 14 章的主要内容之一(主要请见研究报告中有关“结果”内容的写作指导)。

下面将通过集中趋势、离中趋势、平均数的标准差以及效应大小等指标概括词汇量研究例子中的数据。

集中趋势 集中趋势测量(measures of central tendency)包括平均数、中数和众数, 他们表明了数据向某一个值集中的趋势。**众数(mode)**是对集中趋势最为粗糙的测量: 它简单地指出了在所得数据中哪一个值出现的次数最多。如果在分布中有两个值的出现频率均高于其他的值, 而且这两个值出现于频率分布的不同位置, 这个分布就可以被称为双峰分布(即有两个众数)。

中数(median)是频率分布中的中间点。其计算方法为: 将所有的数值按照从高到最低排列, 然后找到那个将数据从中间平均分为两部分的数, 即为中数。比如在这样一个数集中: 4, 5, 6, 7, 8, 8, 中数是 6.5。当一共有偶数个数值时, 中数即为中间两个数值的平均值(在这个实例中, $(6+7)/2=6.5$)。当一共有奇数个数值时, 中数就是将全部数值按升序或者降序排列时中间的一个。对于数集 4, 5, 6, 17, 18, 中数就是 6。需要注意的是, 此例中即使最高值为 18 而非 18 时, 中数依然是 6。当整个分布中有极端值时, 中数是对集中趋势的最佳测量, 因为较之平均数而言, 中数较少受到极端值的影响。

平均数(mean)是集中趋势中最为常用的一种, 它是通过用所有数值的个数去除所有数值之和得到的。总体平均数用 μ 表示(希腊字母 mu); 在研究报告中, 样本的平均数是用 M 表示的。(标识 $\bar{X}$ 读作“ X 拔”, 是专门用于统计公式中的)。

一般都是通过平均数来报告集中趋势, 除非在分布中有极端值。一般大家所说的“平均值”指的都是算术平均数。比如, 对于数集 4, 5, 6, 7, 8, 8, 平均数就是: 6.33。

$$\bar{X} = \frac{\sum X}{N} = \frac{38}{6} = 6.33$$

其中:

Σ = 求和

X = 个体的 (原始) 分数

N = 个体的数目

对词汇研究两组被试集中趋势的测量:

	大学生组	老年人组
平均数 (M)	45.58	64.04
中数	46.00	64.50
众数	38, 42, 48, 59	78

正如你所看到的, 大学生组的平均表现要比老年人组的平均表现低很多, 它证实了我们在茎叶图中看到的: 老年人组的平均表现总体上要比大学生组好。需要注意的是, 各组的平均数和中数都是相似的; 因此, 虽然我们在茎叶图中看到老年人组中有一个极端值, 但是这个极端值的存在看来并没有太大地影响平均数对集中趋势的测量。大学生组数据中的众数不止一个, 每一个都出现了两次; 在老年人组中, 出现最高频率的分数是 78, 而它只出现了三次。正如你所看到的, 众数在小集合的数据中并不是特别有效。

离散程度或者变异量 在你报告集中趋势的时候, 一定要同时报告离中趋势 (变异) 测量 [measure of dispersion (variability)]。对集中趋势的测量表明的是在一个频率分布中数据“集中”趋向的值, 而离中趋势表明的是分布的广度或者变异。

对离中趋势最粗糙的测量 (与众数相对应) 是全距 (range)。全距通过用分布中的最高值减去最低值得到。比如, 在一个由 1, 3, 5, 7 组成的小分布中, 全距就是 $7-1$, 也就是 6。

离中趋势最常用的测量方法 (与平均数相对应) 是标准差 (standard deviation)。标准差能够近似地告诉你整体上来说数值离平均数有多远, 它是指每个值与平均数偏差的均方根。标准差的定义式如下:

$$\sqrt{\frac{\sum (X - \bar{X})^2}{N-1}}$$

其中:

X = 个体的原始分数

$\bar{X}$ = 样本平均数

N = 样本大小

Σ = 求和

这里离差平方的平均数除以的不是 N 而是 $N-1$, 这是为了得到基于样本的总体标准差的无偏估计值。总体标准差用 σ 表示 (希腊字母西格玛)。在研究报告中, 样本标准差用 SD 表示, 而在统计公式中它常用 s 表示。方差是标准差的平方 s^2 , 它是在各种推论统计中都相当重要的一个离散量。

对词汇量测试中两个组所得数据变异的测量如下:

	大学生组	老年人组
全距	23-62	32-83
方差 (s^2)	109.45	150.44
标准差 (SD 或 s)	10.46	12.27

需要注意的是, 茎叶图显示出老年人组有更大的变异; 而通过 SD 我们借助数据看到了这一分布的特征。

平均数的标准误 在推论统计中, 我们用样本平均数这个点来估计总体平均数。也就是说, 我们使用单个的值 ($\bar{X}$) 来估计 (推断) 总体的平均数 (μ), 如果我们能够确定用 $\bar{X}$ 来估计 μ 会有多少误差, 这将是非常有用的。数学中的中心极限定理告诉我们, 假如我们用无限个相同容量的样本, 计算每一个的样本平均数, 这些样本均数的平均数 ($\mu_{\bar{X}}$) 将与总体平均数 (μ) 相等, 而样本平均数的标准差将与总体标准差 (σ) 除以样本容量 (N) 平方根所得到的值相等。这个理论抽样平均数分布的标准差就被称为平均数标准误 (standard error of the mean) ($\sigma_{\bar{X}}$), 其定义式为:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

一般情况下, 我们不知道总体的标准差, 所以我们用样本标准差 s 来进行估计。然后, 通过下面这个公式我们就能够得到一个平均数标准误的估计量 (estimated standard error of the mean):

$$s_{\bar{X}} = \frac{s}{\sqrt{N}}$$

偏小的 $s_{\bar{X}}$ 表示我们对样本平均数的估计是比较接近的, 而较大的 $s_{\bar{X}}$ 则表示我们只是对样本平均数进行了一个粗糙的估计。平均数标准误的公式表明, 我们通过样本来估计总体平均数的能力大小依赖于样本容量 (大样本就能够进行更好的估计), 以及样本所属总体的变异性, 这是通过样本标准差来估计的 (总体中数值的变异越小, 我们对总体平均数的估计就越好)。我们在后面将会提到, 平均数的标准误在置信区间中扮演了非常重要的角色, 而且在研究中它经常与样本平均数一起呈现于结果概要中。在研究报告中它一般用 SE 表示。

效应大小的测量 当我们做实验时, 我们感兴趣的是自变量到底有没有效果, 如果有效果的话, 效果到底有多大。效应大小这一概念在第 7 章中已经进行了介绍。对效应大小的测量 (见 Kirk, 1996) 是非常重要的, 因为它使我们得以了解在排除了样本大小作用的情况下, 自变量与因变量之间的关系强度究竟如何。

在实验研究中, 常用的测量效应大小的方法被称为科恩 d 值, 它是将自变量

各水平平均数之间的差异除以组内标准差得到的一个比率。标准差告诉我们的近似地从平均上说数值与平均值之间的距离, 它测量的是数值平均值间的离散程度。而组内标准差告诉我们的则是由于个体差异 (即个体在反应上的不同) 而导致的“误差”。标准差是一个很好的测量平均数差异的方法。自变量效应的“大小” (各组平均数之间的差异) 则是衡量实验中所得数值的平均离散量的指标。

效应大小的测量 d 值, 我们将其定义为样本平均数之间的差异除以总体的标准差, 根据近代统计学家雅各布·科恩的名字将其命名为科恩 d 值 (更多关于 d 的信息请见 Cohen, 1988)。

$$\text{科恩 } d = \frac{\bar{X}_1 - \bar{X}_2}{\sigma}$$

总体标准差 (σ) 通过将各组组内变异相加, 再除以两组样本总数 (N) 得到。用样本方差估计总体标准差的公式如下:

$$\sigma = \sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{N}}$$

其中:

n_1 = 样本 1 的个数

n_2 = 样本 2 的个数

s_1^2 = 样本 1 的方差

s_2^2 = 样本 2 的方差

$N = n_1 + n_2$

如果有很大的组内变异 (即, 组内标准差很大), d 的分母就会很大。为了能够观察到自变量的效应, 在较大组内变异的情况下, 两组平均值之间的差异一定要大。当组内变异比较小 (d 的分母小), 同等大小的均数差异就会反应出很大的效应大小。因为效应大小是以标准差单位呈现的, 所以, 虽然不同的实验使用不同的因变量, 但是它们之间可以进行有意义的效应比较。比如, 一项强调对词汇意义辨别 (即, 多选项测试) 的测验, 对大学生和老年人进行了比较。而另一项研究则比较相似的两组被试在回忆词语定义上的表现, 这两个研究的效应大小就可以相互比较。正是这种比较形成了元分析的基础, 元分析研究对多个研究中某个特定自变量的效应进行概括总结 (见第 7 章)。

一些材料能够帮助我们更好地理解 d 比率的意义。科恩 (1992) 对效应大小进行了一个非常有用的分类, 分为小、中、大三类。科恩这样解释他对效应大小 (Effect Size, ES) 进行分类的基本理由:

我对效应大小进行分类的理由是: 中等的 ES 表示的效应仔细的观察者能够用肉眼观察到 (经调查, 在各个领域的研究中, 它近似于可观察效应的平均大小); 我设立小 ES, 它明显地比中 ES 要小, 但还不至于小到微不足道的程度; 而我设立大 ES, 它与中 ES 的距离和小 ES 与中 ES 的距离相同。尽管这些定义是比较主观的, 但是在经过一些微小的调整之后, 作为惯例……已经被广泛使用了 (p.156)。

效应大小的每一个等级都有相应的量化条件：比如，一个双组实验中，中等效应大小的 d 值为 .50；而小效应和大效应的 d 值则分别为 .20 和 .80。在对相似研究的结果进行比较时，这些效应程度的表示都非常有用。

需要着重注意的是，研究者们对平均数间标准化差异的定义略有不同（见，比如，Cohen, 1988; Kirk, 1996; Rosenthal, 1991）。使用哪种方法来测量效应大小是由研究者自行决定的。然而，既然心理学文献中出现的各种方法之间有差异，那么辨识研究报告中到底使用了哪一种方法来测量效应大小就尤为重要。

在词汇量研究中用科恩 d 值计算效应大小：

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sigma} = \frac{64.04 - 45.58}{\sqrt{\frac{(26-1)(150.04) + (26-1)(109.45)}{52}}} = 1.65$$

为了解释 1.65 这个值，我们可以使用科恩对效应大小的分类， $d=.20$ 是小效应， $d=.50$ 是中等的效应，而 $d=.80$ 对应的是大效应。因为我们的 d 值比 .80 大，所以能够得出结论：年龄大小对其所掌握的词汇量有着重要的作用。

结 论 在第二个阶段，也就是数据分析的概括阶段，我们需要确定：

- 研究中每一种条件或者每一组的集中趋势，例如平均数。
- 测量每种实验条件的离散趋势（变异性），诸如全距和标准差等。
- 每一个主要自变量的效应大小。
- 怎样才能最好地用图来展示数据的概貌（如，显示各条件下平均表现水平的数据）。

注意：虽然我们可以画一张词汇量研究中两组被试平均表现的图，但是，只有两组被试的情况一般并不需要画图，只有当研究结果包含两组以上的被试时，概括图才显得尤为重要。

第三阶段：用置信区间来判断数据所反映的意义

- 要确认数据到底告诉我们什么，一个重要的途径是建立总体参数的置信区间，这些总体参数可能是一个平均数，也可能是两组平均数之间的差异。

在数据分析的第三阶段，我们希望能够证实我们在熟悉以及概括数据时所获得的印象。这一阶段的主要任务是计算总体参数的置信区间（confidence interval for a population parameter）。置信区间（CI）可以针对单个总体平均数，也可以针对总体平均数之间的差异。我们首先介绍单个总体平均数置信区间，然后再介绍两个总体平均数差异的置信区间，并讨论当含有三个或者更多总体时对区间的解释。

对于置信区间的另一种称谓，你可能已经很熟悉了。你是不是听过媒体对某个样本的调查报告？在这些报告中你是不是听到过“误差范围”？在专栏 12.2 中，我们介绍误差范围的概念及其与置信区间的关系。

单个平均数的置信区间 从总体中抽取一个随机样本，其平均数就是对总体平均数的一个点估计。但是，我们可以预见，由于随机误差，不同样本的平均数之间

专栏 12.2

调查结果中的误差范围

如同你在第5章中学到的, 调查研究的有效性在很大程度上依赖于抽样。我们为了了解总体的特征(例如, 偏好、态度、人口统计特征)而开展调查研究, 但是调查整个总体往往是不现实的, 通常只能使用一个来自样本的信息来描述更大的总体。越有代表性的样本越可以准确描述总体, 但是, 样本的结果往往不太可能与总体一致。比如说, 如果一个教室里33名大学生的平均年龄为26.4, 那么从中抽取10个学生作为样本, 他们的平均年龄恰好为26.4的可能性就很小。同样, 如果65%的选民喜欢现任市长, 而35%的选民喜欢一个新市长, 假如我们随机抽取100个选民, 其比例恰好为65:35的可能性是很小的。我们知道, 抽样会带来一定的“偏差”, 来自样本的估计值与实际总体值之间会有一些“误差”。那么, 需要讨论的就是样本对总体的代表性究竟精确到了何种程度。

我们能够估计出样本结果和总体之间的误差范围, 误差范围并不给出总体值的精确估计(例如, “全体选民的65%喜欢现任市长”), 它给出的是一段间距, 这段间距中可能包含真实的总体值(例如, “全体选民中有60%至70%的人喜欢现任市长”)。这段间距有什么特殊性吗?

由于随机因素的作用, 样本结果与总体值之间会有差异, 而误差范围正是对这一差异的估计。误差范围使我们得以知晓由于抽样所造成的误差有多大——我们知道误差总会有的, 我们并不期望能够精确地描述总体。假设有许多选民接受了一项民意调查, 媒体发言人做了如下报告: “结果显示, 我们抽取的样本中63%的选民喜欢现任市长, 我们有95%的信心可以说, 这一民意测验的误差范围是5%。”这个以特定的信心水平(通常是

95%)来报告的误差范围表明, 实际总体中喜欢现任市长的市民比率估计应该落在58%到68%之间(从样本值63%减去或者加上5%)。记住, 通常我们并不知道真正的总体值。我们从样本中得到的信息以及误差范围只有如下这些: 样本中63%的选民喜欢现任市长, 并且我们有95%的信心可以说, 如果整个总体都被抽样, 那么58%到68%的总体会喜欢现任市长。这一结果可以在图上表示出来, 用点表示从样本中得到的数值(63%), 用误差条表示误差范围。图12.1展示的就是误差条及其中心的样本估计值。

媒体报告的全国范围的调查中, 一般都会包括误差范围。这些调查的目的是要告诉你真实总体值落于哪个误差范围之内。同样, 许多科学研究的目的也是告诉我们一个总体值的估计值, 也就是误差范围, 现在常常被称作“置信区间”。

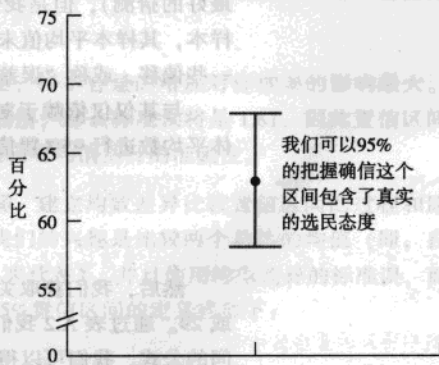


图 12.1 采用误差条来呈现选民态度调查的误差范围

一定会有差异。平均数标准误(s_x)提供了关于抽样误差“正常”范围的信息。通过计算置信区间, 我们确定了一个值域, 我们有一定的信心可以说总体平均值是落在这个值域范围之内的。你能够想得到, 我们定下的范围越大, 我们确信平均值落于其间的信心就会越高。但是, 置信区间越大, 我们能够得到的关于总体平均值的确切信息就越少。作为妥协, 研究者一致认同95%和99%的置信区间是估计总体平均值的最佳区间。

置信区间以我们对总体平均数的点估计值 ($\bar{X}$) 为中心, 95% 置信区间的边界值可以通过如下公式计算得到:

$$95\% \text{ 置信区间上限值: } \bar{X} + [t_{.05}][s_{\bar{X}}]$$

$$95\% \text{ 置信区间下限值: } \bar{X} - [t_{.05}][s_{\bar{X}}]$$

我们已经描述了计算样本平均数 ($\bar{X}$) 和平均数估计标准误 ($s_{\bar{X}}$) 的步骤。这两个 95% 区间等式中我们不熟悉的符号是 t 和 .05。

你可能通过其他心理学课程认识了 .05 的阿尔法 (α) 水平。它与推论统计的显著性检验 (NHST) 有着密切的联系, 在第 13 章中我们会讨论更多关于 α 水平的问题。在置信区间中, $\alpha = (1 - \text{置信水平})$ 是一个比率。所以, 对于 95% 的置信区间来说, $\alpha = (1 - .95) = .05$; 而对 99% 的置信区间来说, $\alpha = (1 - .99) = .01$ 。

等式中的 t 统计量由自由度来决定, t 的统计显著性可以通过查附录中的表 A.2 确定。对单一样本平均数, 自由度是 $N-1$ 。我们在第 13 章中讨论 NHST 时会继续讲到 t 统计量。在这里, 我们只关注如何用上述公式计算并正确解释置信区间。

我们将通过一个例子来阐释我们是如何得到单一样本平均数的置信区间的。如果你从你所在的大学中随机抽取了一些学生作为样本, 并且使用一个简单但是具有较高信效度的量表测量了他们的智商。假设一共测试了三十个学生 ($N=30$), 得到平均智商分数为 115, 样本标准差 14。学生总体是在你们学校就读的数千名学生。虽然样本平均值对总体平均值是一个不错的点估计 (即, 我们对总体平均值最好的猜测), 但是我们必须承认, 如果另外再抽取一个由 30 名学生组成的随机样本, 其样本平均值未必恰好会是 115。由于这个随机抽样的过程, 其中一定会有一些偏移, 或称“误差”。记住, 平均数的标准误就是用来测量估计中的误差的。

与其仅仅依赖于对总体平均数的点估计, 我们不如利用前面提到的公式对总体平均数进行 95% 置信区间的区间估计。我们首先计算平均数标准误:

$$s_{\bar{X}} = \frac{S}{\sqrt{N}} = \frac{14}{\sqrt{30}} = \frac{14}{5.48} = 2.55$$

然后, 我们获取关键的 t 值。因为有 30 个学生, 与 t 相关的自由度即为 30-1 或 29。通过表 A.2 我们找到 .05 阿尔法且自由度为 29 的 t 值为 2.04。通过置信区间的公式, 我们可以得到:

$$95\% \text{ 置信区间上限} = 115 + [2.04] [2.55]$$

$$95\% \text{ 置信区间下限} = 115 - [2.04] [2.55]$$

$$\text{上限值} = 115 + 5.20 = 120.20$$

$$\text{下限值} = 115 - 5.20 = 109.80$$

我们可以说, 109.80 到 120.20 这个区间有 95% 的可能包含了 (“捕获了”) 总体平均数 (参见专栏 12.3)。

区间越窄, 我们对总体均值的区间估计就越好。通过上下限的公式可以看到, 区间的宽度取决于 t 统计量和平均数的标准误。这两个值都与样本容量有关, 当样

专栏 12.3

单个均数置信区间的解释:套环和赌注

计算了总体均数 .95 的置信区间之后,我们可以说:

我们所得到的置信区间包含真实的总体均数的可能性是 95%。

正如穆雷克、拉竹和哈什曼 (Mulaik, Raju, & Harshman, 1997) 指出,置信区间可能包含也可能不包含真实的均数。均数置信区间 95% 的概率指的是,假如我们从总体中随机抽取相同容量的不同样本,并得到它们的置信区间,那么真实的总体均数有 95% 的概率会在这些区间中。也就是说,样本均数的置信区间告诉我们的是,如果我们在同样条件下重复这个实验的话,会得到怎样的结果 (例, Estes, 1997)。在 100 次的重复中,有 95 次真实均数会落在我们的置信区间内。

总体均数 95% 的置信区间所表示的不是如下这样的含义:

真实的均数落在这个区间内的可能性是 95%。

这个陈述看上去也许和前面的那个很相像,然而它们是不一样的。我们关注的是一个确定的值,一个常量;它是总体的一个特征,或者说参数。而区间则是不固定的,他们是样本数据的特征。区间是通过样本平均数得到的,它所测量的高散在各个研究之间都是不同的,置信区间亦是如此。

豪厄尔 (Howell, 2002) 作过一个很精彩的类比,可帮助理解上述这些事实与我们对置信区间的解释之间的关系。他指出,我们可以将参数 (比如总体平均数) 想像成一个赌注,而置信区间则是套环。根据样本数据,研究者们得到有特定宽度的套环,并且将其掷向赌注。当使用 95% 的置信区间时,套环有 95% 的机会可以套住赌注,剩下 5% 的概率套不中。“置信区间讲的是套环套中目标的概率,而不是目标 (参数) 落在套环中的概率。” (Howell, 2002, p. 208)

本容量增大时,两者都减小,但是,样本容量的增加对标准误的影响最大。在上面那个例子中,如果将样本容量翻倍,那么标准误将是 1.81,因此置信区间将会窄很多。要点:增加样本容量将提高对均值估计的准确性。

两个独立样本均数比较的置信区间 建立均数差异比较置信区间的过程和原理与单个均值时是相似的。因为现在我们的兴趣是比较两个总体的均值 (即,自变量的“效应”),所以,我们用 $\bar{X}_1 - \bar{X}_2$ 来代替 $\bar{X}$, 并且使用均值差异的标准误,而非均值的标准误。两个总体均值差异 95% 置信区间的定义式如下:

$$CI(95\%) = (\bar{X}_1 - \bar{X}_2) \pm (t_{.05})(s_{\bar{X}_1 - \bar{X}_2})$$

其中, t 可以在表 A.2 中查到,其自由度为 $[(n_1 + n_2) - 2]$, $\alpha = .05$ 。

均数之差的标准误计算公式如下:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right] \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}$$

为了更好地说明,我们可以计算一下在我们词汇量研究中两平均数差异的置信区间。关键的 t 值可以在表 A.2 中找到, α 设定为 .05, 自由度是 $26 + 26 - 2$ 即 50。

查得 t 值为 2.009。我们可以通过下面的计算得到两个平均数差异的标准误：

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\left[\frac{(26-1) 109.45 + (26-1) 150.44}{26+26-2} \right] \left[\frac{1}{26} + \frac{1}{26} \right]} = 3.16$$

因此，总体均数差异 95% 的置信区间为：

$$CI(95\%) = 18.46 \pm (2.009)(3.16) = 18.46 \pm 6.35$$

因此，上限为 $18.46 + 6.35 = 24.81$ ，下限为 $18.46 - 6.35 = 12.11$ 。所以，我们有 95% 的把握可以说 12.11 到 24.81 这个区间包含了词汇量测试中老年组和大学生组正确率的真实总体差异。

现在我们需要根据置信区间对两组平均表现的分析结果做出陈述，以确定词汇量研究中自变量的效应大小。

在词汇量测试中，大学生的平均成绩为 45.58 ($SD=10.46$)；老年组的成绩为 64.04 ($SD=12.27$)。在这项研究中，老年被试的词汇量比大学生被试大。我们所得到的置信区间 12.11 到 24.81，有 .95 的概率包含了真实总体均数差异 ($SE=3.16$)。

需要注意的是，0 这个值没有被包含在这个区间中。在解释两个均数差异的置信区间时，这一点是非常重要的（参见专栏 12.4）。

重复测量设计中两个均数差异的置信区间 目前我们已经讨论过了含有两组独立被试的实验。你可能已经意识到了，实验的操作也可以是每个被试都参与了实验的每种条件，或者通过一些因变量指标（例如 IQ 得分、体重等等）将被试进行

专栏 12.4

两个均数差异的置信区间的解释：寻找 0

在计算出了两组均数差异 95% 的置信区间之后，我们可以说：

我们得到的置信区间有 .95 的概率可能包含了真实的均数差异或者绝对的效应大小。

置信区间为效应大小提供了信息。通过置信区间我们得到自变量可能的效应大小。由于样本和程序等各种因素的变化，不同实验得到的效应大小是不一样的。置信区间“将效应大小的值范围具体化了”（Abelson, 1997, p.130）。也就是说，效应量大小可能是置信区间最小值到最大值之间的任何一个值。研究者们有时惊讶于一个区间需要很大，才能得到有较高信心的效应大小（例，Cohen, 1995）。因此，置信区间越狭窄，我们对

自变量真实效应大小的估计就越好。当然，置信区间的大小与样本容量是直接相关的。通过增大样本容量我们能够更加确切地知道效应大小究竟如何。

确定均数差异的置信区间中是否包含了 0，这一点是非常重要的。如果置信区间中包含了 0，那么我们必须接受有可能两个总体均数之间是没有差异的。因此，我们无法由此得出结论说自变量是有效应的。因为置信区间为我们提供的是效应大小的可能范围。如果 0 也在可能的取值范围之内，那么我们不得不承认我们不确定效应到底是否存在（例，Abelson, 1997）。在第 13 章，我们讨论使用 NHST 得到非显著性结果时，还会遇到类似的情况。

“匹配”。这种实验通常称为匹配组设计, 被试内设计, 或者重复测量设计(见第8章)。例如, 假设一个认知心理学家想要比较人们解答两种不同难题的成绩。她应该会选择一组被试两道题都做, 而不是找两组不同的被试, 每组做一题。(重复设计实验中展示实验材料的方式在第8章中已经提到。)因此, 所有被试在两道题目上都有得分。正如你将看到的, 两者分数的差异就是重复测量设计的测量标准。

在匹配组设计或重复测量设计的实验中, 估算效应大小的程序某种程度上要比独立组设计中更为复杂(关于这类情况下计算 d 的相关信息, 请见 Cohen, 1988; and Rosenthal & Rosnow, 1991)。有一种建议是将其当作独立组设计一样来计算效应大小, 像之前一样遵从科恩的方针(即, 20, .50, .80)(例, Zechmeister & Posavac, 2003)。

同样地, 重复测量设计中两个总体均数差异的置信区间的建构也包括两种条件。但是, 这样一来, 计算方法也要改变。具体地说, 当每个被试都参与了实验的两种条件时, t 是基于差异分数的(见第13章), 差异分数通过将每个被试的两个分数相减而得到。差异分数的均数($\bar{D}$ 拔)的定义式如下:

$$\bar{D} = \sum D/N$$

这里, $\bar{D}$ = 差异值, 而 N 是差异值的个数(即, 有几对数值。)注意, $\bar{D} = \bar{X}_1 - \bar{X}_2$ 。

差异值标准误(S_D)的定义式为:

$$S_D = \frac{S_D}{\sqrt{N}}$$

这里 S_D 是差异值的标准差。

关键值 t 可以通过查附录中的表 A.2 得到, 其自由度为 $N-1$ 。需要注意的是, N 是被试的个数, 即实验中共有几对数值。

重复测量设计中两个均数之间差异的置信区间定义式为:

$$CI = \bar{D} \pm (t_{\alpha/2})(S_D)$$

为了说明重复测量设计的置信区间, 我们还是用那个两道迷宫难题的认知心理学实验作为例子。假设完成测验的时间(秒)是因变量。如果有9个人接受了测试, 我们可以分析如下的假设结果:

参加者	(秒)		D	D^2
	难题 1	难题 2		
1	10	16	6	36
2	18	15	-3	9
3	9	17	8	64
4	14	15	1	1
5	18	15	-3	9
6	11	19	8	64
7	16	18	2	4
8	10	9	-1	1
9	13	19	6	36
			$\Sigma D=24$	$\Sigma D^2=224$

需要注意的是,为了获得差异分数(D),时间都是朝着同一个方向减的(例:完成第二题的时间减去完成第一题的时间)。差异分数的平均数为:

$$\bar{D} = \sum D/N = 24/9 = 2.67$$

差异分数标准误的定义式为:

$$s_D = \sqrt{\frac{\sum D^2 - (\sum D)^2}{N-1}} = \sqrt{\frac{224 - \frac{24^2}{9}}{8}} = \sqrt{\frac{160}{8}} = \sqrt{20} = 4.47$$

因此,这些数据平均差异的标准误为:

$$s_{\bar{D}} = \frac{s_D}{\sqrt{N}} = \frac{4.47}{\sqrt{9}} = \frac{4.47}{3} = 1.49$$

总体均数差异 95% 的置信区间可以通过如下计算得到:

$$CI = \bar{D} \pm (t_{.05})(S_{\bar{D}})$$

$t_{.05}$ 的值可以在表 A.2 中查到,自由度为 8, α 值为 .05。这个值是 2.31。

$$95\%CI = (2.67) \pm (2.31)(1.49) = 2.67 \pm 3.44, \text{即 } -.77 \text{ 到 } 6.11$$

注意, 0.0 在这个置信区间内。这样的话,我们就无法得出解答两道难题的(总体)平均时间是有差异的这一结论(参见专栏 12.4)。

多个独立组均数差异的置信区间 为了说明当有两个以上的均数时如何通过置信区间来分析和解释结果,我们不妨设想一项关于婴儿如何“把握图画本质”的研究(Deloache, Pierroutsakos, Uttal, Rosengren, & Gottlieb, 1998)。你有没有想过婴儿是如何理解画面上的物体与真实的物体之间的区别?德洛克和她的同事对此充满了兴趣,他们的研究证明,5个月大的婴儿不仅能辨识出物体与物体之间的相似之处,而且知道它们是不同的。但是,这些研究的发现却与我们平日的常识不同,我们常常听说和看到婴儿和小孩子尝试着去抓、去捡图片中的物体,甚至试图把脚伸进一张图片里的鞋子中!这些常识表明尽管图片中的物体是二维的表征,婴儿孩童还是将它们当作了真实物体。在其所做的四项研究中,德洛克等人检验了“在何种程度上婴儿会将描绘出来的物体当作真实物体”(p.205)。

我们将要集中探讨的是德洛克等人(1998)的第四个研究,不过先强调一下前三个研究的发现还是非常有帮助的。第一项研究是一个结构性的观察,在研究者的观察下,10名9个月大的婴儿自己翻看一个图画书,书中有“八张高度真实的彩色物体照片(普通塑料玩具)”(p.206)。结果显示,每个婴儿都翻着看了至少一幅图,其中8名婴儿至少有一次试着去抓图中的物体(平均3.7次)。在他们的第二项研究中,实验者试图确定婴儿能够对图片中的物体和现实中的对应物进行区分。研究中,实验者向8名9个月大的婴儿同时展示物体的图片和真实物体。令人吃惊的是,婴儿们都首先将手伸向物体而非图片(86%的次数),这表明在研

图 12.2 通过观察儿童探索
和指点图片物体的表现来测
验他们对于图画本质的把握
(来自 Deloache et al; 经允许
使用)



究一中, 婴儿去抓那个图片并不是因为他们不能区分二维和三维的物体。研究三
的目的在于要确定这些研究的外部效应, 研究者们选择的是“西非象牙海岸国家
的乡村地区, 观察了那里严重贫困而且文盲家庭的婴儿” (p.207)。(见图 12.2)
结果显示, 这些婴儿翻看和抓取物体(图片上是他们那里常见的物品)的表现与
美国婴儿是一样的。

第四项研究的目的是确定儿童在面对图片时的行为如何随年龄的变化而变化。
研究者测试了三个年龄组: 9 个月、15 个月以及 19 个月。每个年龄组都有 16 个
孩子, 其中 8 个女孩 8 个男孩。除了观察孩子们在面对图片时的手部探索行为
(抓握或者其他探索行为), 研究者还同时记录婴儿指点图片的行为。他们得到的
婴儿探索行为的结果见图 12.3。

研究中的自变量是孩子的年龄, 这是一个有三个水平的自然组设计: 9 个月、
15 个月和 19 个月, 这一变量在水平轴 (x 轴)。因变量是探索行为的次数, 在纵
轴 (y 轴) 上是这些行为的平均次数。正如从图 12.3 上看到的, 9 个月孩子的探索
行为平均次数是最高的, 15 个月大和 19 个月大的都要低很多。图中比较重要的信
息还有围绕在每个平均数周围的“统计棒”。我们用这些统计棒来确定自变量(年
龄)到底有没有效应。

图 12.3 中每个平均数周围的统计棒代表的是置信区间。如同你已经学过的,
置信区间告诉我们的是总体值可能的范围。由于抽样误差的存在, 我们无法精确
地估计总体值, 但是我们可以估计出一个大概的取值范围。置信区间的取值范围
越小, 我们对总体值的估计就越好。图 12.3 中每个统计棒都代表一个 95% 的置信
区间。但是, 多组研究中区间的计算方法与仅有一个平均数的研究略有不同。特
别地, 当计算平均数的估计标准误时, 我们可能会用到研究中各个组的合并方差。
让我们来具体说明。

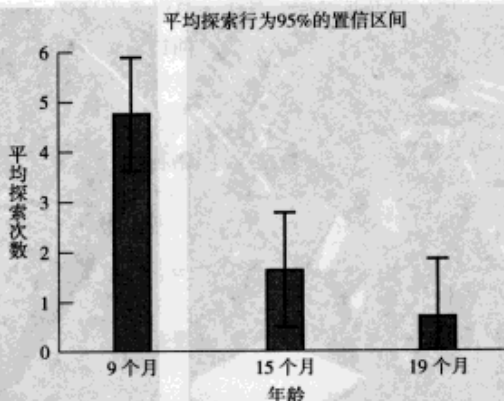
95% 置信区间的计算公式与只有一个平均数的情况是一致的:

$$95\% \text{ 置信区间的上限: } \bar{X} + [t_{.05}][s_{\bar{X}}]$$

$$95\% \text{ 置信区间的下限: } \bar{X} - [t_{.05}][s_{\bar{X}}]$$

图 12.3 9 个月、15 个月和 19 个月孩子探索行为平均数的 95% 置信区间

(来自 Delouche et al; 经允许使用)



但是, s_x 的计算方法与一个平均数的情况不同。同样, 关键值 t 的自由度的计算方法也不同了。为了估计平均数的标准误, 我们可以合并各组的方差来得到一个对变异的测量标准。在这里, 我们就把研究中每个组的信息都合并起来。当要比较两个或两个以上独立组的平均数时, 平均数标准误的计算方法如下。首先, 我们找到基于合并方差的标准差:¹

$$s_{\text{pooled}} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2 + \dots}{(n_1-1) + (n_2-1) + (n_3-1) + \dots}}$$

当样本容量一致时, 估计标准误的定义式为:

$$s_x = \frac{s_{\text{pooled}}}{\sqrt{n}} \quad \text{此时 } n = \text{每个组的样本容量}$$

自由度的计算公式为 $k(n-1)$, 这里 k 为独立组的个数。

再看图 12.3, 我们可以发现 9 个月婴儿组探索行为的样本平均数为 4.75。95% 置信区间等式中表达式 $[t_{.05}]$ 的值为 1.14。我们有 95% 的信心说 3.61 到 5.89 这个区间 (4.75 ± 1.14) 包含了 9 个月组的总体平均数。因此, 在这一研究中, 16 名 9 个月大的婴儿的样本被用来估计探索行为的平均次数, 如果我们用 9 个月大婴儿的总体来做实验的话, 这个估计就可以得到验证。对 15 个月年龄组, 探索行为的平均次数是 1.63, 我们有 95% 的信心说 .49 到 2.77 这个区间 (1.63 ± 1.14) 包含了总体平均数。19 个月组的样本平均数是 .69, 而且 95% 置信区间的下限低于 0.0 (受允许值范围的限制), 上限为 1.83 ($.69 \pm 1.14$)。

专栏 12.5 中所讲的内容是, 当有三个或三个以上的平均数时, 如何解释置信区间。

最后还有一点要提出的是, 当查看研究结果图上的统计棒时要谨慎小心。在

¹ 总体标准差的合并估计与组间方差分析 (ANOVA) 中均数方差的平方根相等。也就是说, $s_{\text{pooled}} = \sqrt{MS_{\text{error}}}$ 。对 ANOVA 的讨论见第 13 章。

专栏 12.5

当有三个或更多平均数时的置信区间解释: 区间重叠吗?

在许多研究情况中, 真正让我们感兴趣的并不是对总体特定平均值的估计。例如, 让我们真正感兴趣的不是9个月大的婴儿能够抓住图片的平均次数。而是总体的分布和总体平均数的比较 (Loftus & Masson, 1994)。也就是说, 我们希望能够比较不同组中的行为, 这可以通过使用置信区间来实现。

通过对总体平均数的估计, 我们可以问: 不同年龄组的婴幼儿在探索行为数量上有什么不同? 要回答这个问题, 我们可以研究图 12.3 中 95% 置信区间的重叠部分。需要记住, 置信区间与区间包含总体平均数的概率 (如 0.95) 相关; 区间的宽度决定了我们估计的精确程度。我们要谨记, 置信区间是为了提供关于我们估计的总体值——通常是平均数的好坏的信息。置信区间不是像 t 检验或 F 检验这样的统计检验, 不能通过直接比较两个或多个平均数, 来看出差异是否“统计上显著”。然而, 正如我们前面所说, 研究者通常对总体平均数的分布感兴趣, 而我们可以用置信区间帮助我们理解不同的分布模式。

当区间没有重叠, 我们可以相信总体不同。不相重叠的区间告诉我们, 通过样本平均数估计的总体很可能不同。例如, 在 95% 置信水平上, 9 个月大的婴儿与 15 个月大的幼儿的区间不相重叠。我们可以由此得出结论, 看图片时, 9 个月大的婴儿与 15 个月大的幼儿发生探索行为的数量不一样。(对样本平均的分析显示, 9 个月大的婴儿相比于 15 个月大的幼儿, 更多的翻看图片。) 比较 9 个月大的婴儿与 19 个月大的幼儿的区间可以得出相似的结论。比较 15 个月大的幼儿与 19 个月大的幼儿则会得到不同的结论, 图 12.3 表明这两组的区间重叠了。我们现在应该得到什么结论? 如果区间的一小部分重叠, 那么我们必须承认平均数的真实差异不确定, 并推迟判断。如果一个组的区间落在另一组区间内, 那么我们可以得出, 这两组总体平均没有差异 (见 Zechmeister & Posavac, 2003)。

根据这些指导方针, 观察图 12.3 中, 15 个月大的幼儿与 19 个月大的幼儿有哪些不同? 图中可以看到, 在 95% 的置信水平上, 15 个月大的幼儿的区间包含了 19 个月大的幼儿的区间。所以, 我们认为总体平均没有

不同。虽然样本均数不同 (分别为 1.63 和 0.69), 我们不能认定总体均数不同 (在心理学研究中, 我们更倾向于描述总体而非样本)。例如, 根据图 12.3 中的区间的重叠, 15 个月大的幼儿真实的总体均数可能是 0.69 (是 19 个月大幼儿的样本均数)。基于图 12.3 中呈现的平均数和置信区间, 我们可以得出结论: 9 个月大的婴儿相对于 15 个月大的幼儿和 19 个月大的幼儿, 更多地翻看图片; 这年龄较大的两组幼儿研究行为的数量没有差异。需要注意的是, 我们并没有说较大的两组 (总体) 之间不可能有差异。根据这些数据, 我们无法得出有差异的存在。但是, 我们同样不能基于这些数据就确切地得出差异是不存在的。只有当我们有了更多的研究结果, 也许使用更大的样本来对总体均数做出更为精确的估计之后, 才能进一步做结论。

图 12.4 显示了德洛克等人的婴幼儿指点图片实验的数据。根据图中显示的平均数和置信区间, 你能得出什么结论? 15 个月大的幼儿和 9 个月大的幼儿在指点行为上有区别吗? 为什么有或者为什么没有呢? 19 个月大的幼儿和 15 个月大的幼儿在指点行为上有区别吗? 为什么有或者为什么没有呢?

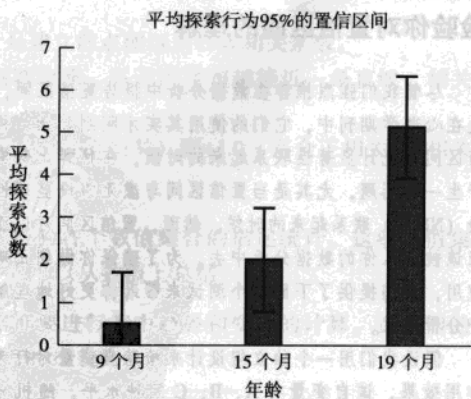


图 12.4 9 个月、15 个月和 19 个月孩子探索行为平均数的 95% 置信区间 (来自 Deloache et al; 经允许使用)

期刊杂志里，图中的统计棒有时候表示的是置信区间，但是也有可能代表均数的标准误或者标准差。（近似计算 95% 置信区间的快速技巧是将均数的标准误乘以 2。）而使问题变得更加复杂的是，作者常常会告诉读者呈现的到底是什么。所以，当呈现统计棒时，非常重要的一点是，告知读者它们表示的是什么，以及如何计算得到的（Estes, 1997）。

例证：一个相关研究的数据分析

- 当同一组人、事件、或物体的两个测量一起变化——也就是，当一个变量上的得分与另一个变量上的得分共变时，就存在相关。

正如你在第 2 章中看到的，预测是科学研究的一个重要目标，相关研究经常为预测提供基础。当同一组人、事件或者物体的两个测量一起变化——也就是，当一个变量上的得分与另一个变量上的得分共变时，相关就存在了。比如说，众所周知吸烟和肺部疾病之间的关系。一个人吸烟越多（如以吸烟年来测量），他感染肺部疾病的可能性也就越大，因此，吸烟和肺部疾病是共变的。这一共变也可以这样表示：人们吸烟越少，他们感染肺部疾病的可能性就越低。基于这一相关，我们可以对肺部疾病做一些预测。例如，如果我们知道一个人已经有多长的烟龄，我们就可以预测（一定程度上）他/她患肺部疾病的可能性。我们预测的性质以及我们在做预测时的信心取决于相关的方向和强度。

相关分析经常与调查研究联系在一起（参见第 5 章）。受访者填写问卷上关于

拓展练习

检验你对置信区间的理解

尽管我们强烈推荐在数据分析中报告置信区间，但是在心理学期刊中，它们的使用其实才刚刚开始。将置信区间与统计显著性联系起来的时候，在解释上确实会产生一些问题，尤其是当置信区间与虚无假设显著性检验（NHST）联系起来的时候。然而，置信区间可以而且应该被放入你的数据分析中去。为了确保你做到正确地使用，我们提供了下面这个测试来帮助你更好地理解这个分析技术。

假设我们用一个独立组设计来检验自变量对行为的作用效果，该自变量有 A、B、C 三种水平。随机分配 15 个被试到每种条件下，并且测量每一组的集中趋势和变异性。同时，调查者为每种条件下的平均数建立 95% 的置信区间。根据结果，研究者可能会做出如下结论，

请你判断它们是否正确：

1. 置信区间的范围表明了对总体均数的估计精确到了什么程度。
2. 如果两个区间是完全重叠的，我们就可以知道总体均数是一样的。
3. 真实的总体均数有 95% 的可能落于每个区间内。
4. 假如两个区间没有重叠，那么有 95% 的概率总体均数是不同的。
5. 如果两个区间没有重叠，那么我们有足够的证据可以说总体均数是不同的。

答案：（1）正确（2）错误（3）错误（4）错误（5）正确

人口统计学的一些变量(例如年龄、收入), 同时填写他们的态度、观点以及主观幸福感。然后, 研究者试图找出其中各种变量之间的联系, 即它们之间的相互关联。是不是低自尊的人同时也有饮食困难呢? 孩子在托儿所的时间是不是和他们母亲的依恋有关呢? SAT 的成绩能够预测进入大学之后的学业成就吗?

接下来我们将看看研究者们是如何分析和解释相关研究的。

第一阶段: 熟悉数据

因为相关研究中总是有两个分数集合, 而且分数之间的关系是我们感兴趣的。所以, 数据分析的步骤与集中探讨均数差异的研究有所不同。为了更好地说明问题, 假设一位研究者对两种主观幸福感测量之间的关系感兴趣, 他让大学生做了两种测量的自陈报告(对自陈报告数据的讨论见第5章)。两种测量使用的都是10级评分量表。一个测量的问题是这样的: 你对分数有多大的担心? (1=完全不担心, 10=非常担心)。另一个测量的问题是这样的: 你在班级测验中集中精力时感到困难吗? (1=完全没有, 10=非常严重)。

清理数据 每一位反应者都提供了两个分数, 两个集合的分数都应该仔细检查, 看看有没有出现错误, 诸如不可能的数值(如在量表全距范围之外的数)和极端值。同时, 可以使用茎叶图来检查每个集合里的数据。当使用量表的时候, 可能的回答是有限的, 与开放性的回答(如报告年收入)相比, 这种情况下出现极端值的可能性就会小一些。

结论 只有当调查者确信数据中已经不再含有错误或者是其他可能歪曲发现结果的因素, 才可以进行下一步。

第二阶段: 概括数据

- 对相关性数据的主要描述统计是建立散点图以及计算相关系数。
- 相关的程度可以从散点图上判断出来, 各点的排布越接近一条直线, 相关就越强。
- 相关系数的取值分布是从 -1.0 (完全负相关) 到 +1.0 (完全正相关); 相关系数为 0 表示两个变量之间没有关系。

在概括数据的时候, 首先应该检验各个数值集合的描述统计。这些数值集合之间的关系强度可以从图上总结, 也可以从数值上总结。

集中趋势和变异 两个集合的数据都要进行集中趋势和变异的计算。在我们的研究例子中, 两个集合反应的平均数和标准差为:

	担 心	注意困难
<i>M</i>	5.45	5.30
<i>SD</i>	1.93	1.98

在相关研究中,我们主要的兴趣不在于平均数之间的差异,而是分数集之间的关系。描述相关数据的主要方法是绘制散点图(scatterplot)和计算相关系数。散点图用来描述两个分数集合之间的关系,相关系数对散点图中观察到的这种关系提供了量上的概括。在试图解释一个相关系数前对散点图进行仔细的研究是很重要的。我们首先举例如何绘制散点图,而后再说明如何计算和解释相关系数。

绘制散点图 相关的本质可以通过散点图来表示出。两个变量的分数可以定位在 x 轴和 y 轴上,在每个变量上,每个个体都会有一个值(或分数)(例,担心和注意困难等级)。散点图显示了每对分数的交叉点,在图中,可以通过判断点与直线之间相对应的程度来观察相关的数值大小或程度。图12.5呈现了三种散点图。第一幅(a)和第三幅(c)图的相关要强于第二幅(b)图,这是因为(a)和(c)的点更近似于一条直线。

散点图中点的位置排列决定了相关的方向。当点的排列模式从左下角到右上角时(图a),相关为正(x 轴和 y 轴上的分数都很低或者 x 轴和 y 轴上的分数都很高)。当点的排列模式是从左上角到右下角时(图c),相关为负(x 轴上分数低而 y 轴上分数高或者 x 轴上分数高而 y 轴上分数低)。

假设有20个大学生对我们之前所提到的两个问题作答,再假设对错误和异常数据进行了检查并剔除。

我们希望发现第一个测量上的分数是否和第二个测量有关(即,“匹配”)。研究报告能否得出对成绩的担心和对测验的注意困难有关?我们可以描绘一幅散点图来表明这些分数之间的关系。我们首先要做的是确定每个人在两个10点量表上的反应。我们假想的数据可以如下:

被 试	担心 (X)	注意困难 (Y)
1	7	5
2	4	4
3	5	3
4	6	7
5	2	2
6	3	4
7	5	7
8	4	5
9	8	6
10	9	8
11	6	6
12	5	3
13	4	4
14	6	5
15	2	4
16	6	8
17	5	4
18	7	8
19	8	4
20	7	9

拓展练习

检验你对置信区间的理解

假设你收集了以下数据,并计算了相关系数 $r = 0.65$ 。请计算95%的置信区间,并解释其含义。如果数据不符合正态分布,请说明原因。

数据如下:

被 试	变量 X	变量 Y
1	12	15
2	8	10
3	15	18
4	10	12
5	18	20
6	14	16
7	9	11
8	16	19
9	11	13
10	17	21
11	13	17
12	10	14
13	19	22
14	12	15
15	16	18
16	11	13
17	18	20
18	14	16
19	9	11
20	17	21

请根据以上数据完成以下任务:

- 计算95%的置信区间。
- 解释置信区间的含义。
- 如果数据不符合正态分布,请说明原因。

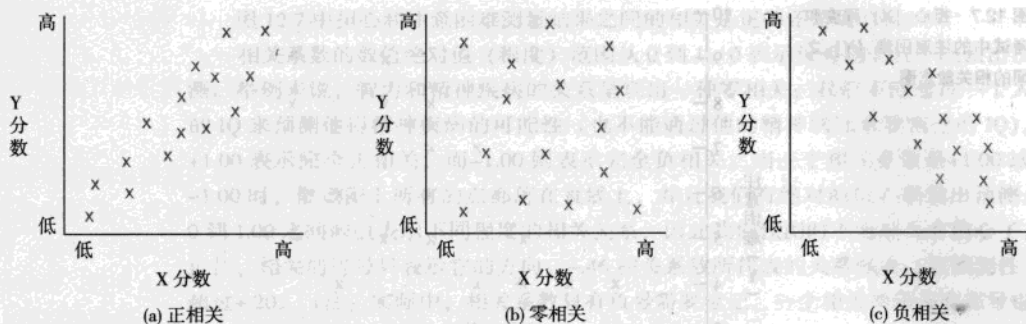


图 12.5 描述两个变量 X 和 Y 分数之间正相关 (a)、零相关 (b) 和负相关 (c) 的散点图

我们通过画出每个反应者测量结果的交叉点来构建散点图。图中的坐标轴代表两种测量水平。通常, 测量中“首先出现”的或者用来预测第二个行为的行为设置在水平轴即 x 轴上。第二个行为或者那些被第一个行为所预测的行为则设置在垂直轴即 y 轴上。在许多情况下这样的设置很简单。如果你将志愿者的血液酒精浓度和驾驶模拟器的行为绩效之间建立相关, 那么显而易见我们所测量的是酒精浓度和虚拟的驾驶模拟器的成绩, 而血液酒精浓度可以预测驾驶成绩。在另外的一些情况下要这样设置就没那么容易了。对成绩的担心是否就发生在测验注意困难前? 还是测验注意困难就会导致对成绩的担心? 我们相信这两种可能都存在。

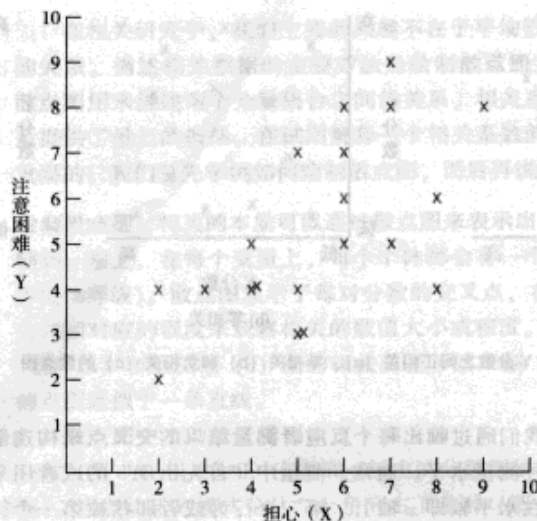
我们想要研究散点图可能的趋势, 更明确地说, 我们想要找出图中是否存在线性趋势 (linear trend) 的证据。简单地说, 线性趋势可以被概括为一条直线。正如你们所见, 图 12.5 中散点图 (a) 和 (c) 就展现出这种线性趋势, 当然散点图中也可能不存在这样的趋势。在这种情况下, 一个测量上的分数有可能和第二个测量上的低、中、高分任意一个匹配。如果在图中不存在可辨别的趋势, 如图 12.5 中间那幅, 那么我们得出结论, 这两个分数集之间没有关系。注意在这种情况下, 我们不能用我们在第一个测量上的分数来预测第二个测量上的分数。

最后, 我们也能在散点图中看出非线性的关系。图 12.6 提供了两个变量之间非线性关系的例子, 我们认为这些关系非常有趣, 而且值得进一步调查。然而, 在解释相关系数时非线性关系会遇到一些重要的问题。因此, 如果散点图中的趋势是非线性的, 就无法计算相关系数。在解释相关系数时, 对异常数据的解释也

图 12.6 两个变量 X 和 Y 的非线性关系的例子



图 12.7 担心 (X) 程度和考试中的注意困难 (Y) 之间的相关散点图



会遇到一些问题。

图 12.7 显示了一幅描述我们假想研究中担心 (X) 和注意困难 (Y) 测量分数关系的散点图。由于我们并不知道该例中哪个因素“首先发生”，因此在图 12.7 上我们随机地将担心设置在 x 轴上，注意困难设置在 y 轴上。也就是说，我们用对焦虑的测量来预测注意困难。你们在散点图中能看到什么趋势吗？它是一般的线性趋势吗？

计算相关系数 相关的方向和程度可以通过计算相关系数得出。相关系数是一种量化的指标，它说明了一个分数集（如担心等级）是否很好的预测了另一个分数集（如注意等级）。相关系数表示了两个变量在数量 and 方向上的关系。

相关系数的方向可正可负。**正相关 (positive correlation)** 表明当一个测量的值增加，另一个测量的值也增加（见图 12.5 中 a 图）。正如我们所见，吸烟和肺病的测量是正相关的（吸烟越多，肺病越多）。另一个可预测的关系关注 SAT 分数：SAT 分数和大学生第一学期 GPA 呈正相关。这样，我们可以预测高 SAT 分数的学生第一学期会获得高 GPA 分数。通过相关，我们也可以进行“逆向”预测。如果我们仅知道学生第一学期的 GPA 分数，我们可以“预测”他们在入学前的 SAT 分数。高 GPA 分数的学生更可能对应较高的 SAT 分数，而低 GPA 分数的学生更可能对应较低的 SAT 分数。

在一个**负相关 (negative correlation)** 中，一个测量的值增加，另一个测量的值反而减少（见图 12.5 中 c 图）。一个关于高校高年级学生的全国调查表明，在观看电视上所花费的时间和学业成就测验成绩是负相关（Keith, Reimers, Fehrmann, Pottebaum, & Aubrey, 1986）。学生看电视的时间越多，在学业成就测试中正确回答题目的数量就越少。相反的预测是什么呢？基于这个发现，如果你知道一个学生在学业成就测试中得分很高，那么你会预测那个学生看电视花了很多时间还是花了一点时间？

图 12.7 中担心和注意困难测量结果之间的相关是正的还是负的呢?

相关系数的数值绝对值(程度)范围从 0 到 1。0 表示没有相关且不能做出预测。举例来说, 智力和精神疾病的关系呈现出一种零相关; 我们不能通过一个人的 IQ 来预测他得精神疾病的可能性(也不能通过他的精神状态来预测他的 IQ)。+1.00 表示完全正相关, 而 -1.00 则表示完全负相关。当一个相关系数是 +1.00 或 -1.00 时, 散点图上所有的点都落在直线上, 并且我们有绝对的信心来做出预测。0 到 1.00 之间的值表示不同程度的相关关系, 因此我们预测时不是那么有信心了。记住, 相关的符号只表示它的方向; -.46 相关系数所代表的关系强度(可预测性)超过 +.20。(注: 实际中, 相关系数只有负号需要标记, 一个相关如果没有正号也没有负号, 那它就代表正的, 即, +.20=.20。)

最通常使用的计算相关系数的公式是皮尔逊积差相关系数 r , 公式为:

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$$

其中

X = 第一个变量的原始分数

Y = 第二个变量的原始分数

$\sum XY$ = X 和 Y 的积的和; 也就是每个 X 和 Y 相乘, 然后将这些积相加

N = 分数对的数量

使用上述公式, 在我们的例子中代表 20 个学生担心和注意困难关系的相关系数为 .62。正如图 12.7 散点图上所表现出来的情况, 担心获得低分数倾向于伴随着注意困难, 反之亦然。我们可以认为两个变量之间正相关。更明确地说, 我们可以认为学生越担心, 他们在测验期间就越难集中注意。但是我们能够说是担心导致了学生注意困难吗?

相关和因果关系 也许你还记得我们第 4 章和第 5 章所谈及的“相关并不意味着因果关系。”两个变量相关并不允许我们推断一个导致了另一个(即便是一个能够及时地预测另一个)。对于成绩的担心可能会导致测验时注意难以集中, 但也可能是测验时难以集中注意的体验导致了对成绩的担心。此外, 如果一个第三方的变量能够说明担心和注意困难之间的正相关, 那么它们之间就存在一种虚假相关或伪关系, 举例来说, 学习的时间长短可能是一种第三方变量, 它可以说明这种关系。当学习的时间不断增加, 学生可能会更加关注成绩, 同时在测验中更加难以集中注意。

结论 皮尔逊积差相关系数可以用于概括两个变量之间的关系。但是, 非常重要的一点是, 在计算皮尔逊积差相关系数之前, 要先看一下散点图, 确定这两个变量之间的关系用一条直线来模拟是最恰当的。也就是说, 变量之间确实存在线性趋势。相关系数越接近 1.00, 散点图就越接近一条直线, 而我们也越能够根据一个变量来预测另一个。

第三阶段：为相关构建一个置信区间

- 我们可以建立一个置信区间来估计总体相关系数 ρ ，就像估计总体平均数 μ 一样。

用一个样本的皮尔逊相关系数 r 来估计总体的相关系数如同用样本平均数来估计总体平均数。总体相关系数用希腊字母 ρ 来标记。此外，正如样本平均数受到样本误差和样本之间变异的影响，样本的相关系数也是如此。这样，在一些情况下我们可能希望用一个区间来估计总体值 ρ ，就像总体值 μ 那样。换句话说，我们可以计算 ρ 的一个置信区间。然而，我们把这个主题留给其他书讨论，那些书会提供更全面的统计过程（如，Zechmeister & Posavac, 2003）。

小 结

有三个不同但相关的数据分析阶段：了解数据，概括数据，确认数据揭示了什么。在第一阶段我们需要熟悉获得的数据，仔细观察它们，检查错误和反常的数据，对与其他数据不匹配的极端数据我们要特别敏感。茎叶图是使数据分布形象化和检测极端值的好方法。在第二阶段，我们要使用描述统计量和绘图来归纳数据。在这方面测量数据集中趋势的量（平均数、中位数、众数）、离散量和变异量（全距和标准差）都特别有用。当一个研究关注一个自变量对一个因变量上的效应时，描述自变量在因变量上的“效应有多大”就很重要。进行元分析时效应大小的测量很重要，元分析是用来归纳一个特定的变量在许多不同研究中的效应大小。比较两个平均数时一种重要的测量效应大小的方法是科恩 d 值。

在数据分析第三阶段，只有确认数据揭示了什么，我们才能说我们所做的结论是建立在测量数据证据上的。在这一分析阶段有两种方法：虚无假设显著性检验（NHST）和置信区间。这两种方法都是通过估计样本变异性来帮助研究者估计总体参数的真值。尽管随机样本的平均数是估计总体平均数的好指标，但样本间的一些偶然因素会导致估计中存在变异性（“误差”）。平均数的标准误能够衡量样本平均数是否很好的估计了总体平均数。NHST使研究者关注他们所获得的结果“归因于偶然”的可能性。置信区间决定了包含总体值（例，总体平均数）的某种可能性（通常是95%）的范围。置信区间非常像你在媒体报道研究结果中听到的“误差幅度”。区间越小，总体值的估计越准确。增加样本容量将会提高区间估计的准确度。

两个平均数差异的置信区间可以为两个样本平均数所代表的总体平均数间的差异提供根据。区间宽度表明自变量可能引起的效应大小。为两个平均数差异构造置信区间时，如果区间包含零值，那么我们就不能说存在着效应。换句话说，如果零在区间内，我们应该承认我们不确定变量有没有效应。置信区间可以通过独立组设计和重复测量设计来构造。

当有两个或三个平均数，可以给每个平均数构造一个置信区间。多组别研究

中可以通过判断区间是否重叠来判断平均数是否存在差异。当区间不重叠,我们就可以很有信心地认为这些样本平均数所估计的总体平均数确实存在差异。然而,当区间重叠,我们不能说总体平均数之间没有差异,我们只能承认尚无法确定是否真有差异,并且只能等到进一步研究后再去判断。

当研究者的目的是预测时通常采用相关研究,例如预测焦虑程度和纸笔测验成绩的关系。当对一同变化的人物、事件、事物进行两种不同的测量时,就有可能发现一种相关。正如当一个研究包含着比较平均数时需要仔细检查和归纳一样,相关研究也是如此。散点图描述了两个分数集之间的关系,相关系数对散点图中观察到的关系提供了量化的归纳。更明确地说,相关系数描述了散点图中的数据是否很好地拟合了直线。相关系数的值变化范围是-1.00到+1.00。相关系数的符号(=或+)表明两者之间关系的方向;系数的绝对值(0.0到1.00)表明关系的数量大小。相关系数越接近1.00(正或负),散点图中的点就越落在直线上,两者间的关系就越强。

当一个测量的值增加时,另一个测量的值也增加就说明两者存在正相关。在负相关中,一个测量的值增加,另一个测量的值则减少。我们知道两个测量之间的关系(相关)可以允许研究者通过一个测量上的分数来预测另一个测量上的分数。相关系数越紧接1.00,这种预测能力就越强。必须牢记的是相关本身是不能给变量间的因果关系提供根据,相关不意味着因果。

重要概念

数据分析阶段

了解数据

概括数据

确认数据揭示了什么

茎叶图

集中趋势测量

众数

中位数

平均数

离中趋势(变异性)测量

全距

标准差

平均数标准误

平均数标准误的估计量

总体参数的置信区间

散点图

线性趋势

正相关

负相关

复习思考题

1. 识别数据分析的三个阶段并说明在每个阶段研究者需要做些什么具体的事。
2. 构造一个“分析故事”时,研究者需要做什么来匹配研究结果。
3. 为什么一个研究者若使用计算机软件分析研究结果,必须首先对研究方法和统计过程非常熟悉?
4. 为下面数字集合构造一幅茎叶图;然后报告用这种方法检查数据时你学到了什么。36, 42, 25, 26, 26, 21, 22, 43, 40, 69, 21, 21, 23, 31, 32, 32,

- 34, 37, 37, 38, 43, 20, 21, 24, 23, 42, 24, 21, 27, 29, 34, 30, 41, 25, 28。
- 计算下面数据集的平均数、中位数和众数：7, 7, 2, 4, 2, 4, 5, 6, 4, 5。描述这三种测量集中趋势方法的优点和缺点。
 - 计算第5题中数据集的标准差。作为一种测量变异性的方法，标准差告诉你什么？
 - 平均数标准误差告诉你关于样本平均数的哪些信息？
 - 一项研究在调查新发明的药能不能提高记忆成绩，该项研究用老鼠做实验。研究中测量的是被注入记忆药物或盐水（控制）的老鼠在学习迷津过程中所犯错误的次数。老鼠被随机地分配到记忆增强药物组和控制组中。总共有30只老鼠进行测试，每组15只。药物组的平均数（标准差）是11.7（4.7）；控制组为15.1（5.1）。平均数越低成绩越好。这项研究的效应大小是多少呢？
 - 为什么置信区间也称为“误差幅度”呢？
 - 对25个学生组成的随机样本进行提问，询问他们关于大学食堂服务的意见。学生使用7点量表（1=糟糕，7=很棒）来表明他们的意见。25个学生的平均数为4.7，标准差为1.2。
 - 总体平均数的95%的置信区间是多少？
 - 用文字描述置信区间告诉你关于总体平均数的什么信息。
 - 问题8中两个平均数差异的95%的置信区间是多少？如何正确解释该区间？
 - 你如何使用置信区间在三个或三个以上的变量差异中求得结果？
 - 当检查散点图中所描述的数据时，为什么寻找数据的线性趋势很重要？
 - 一个研究者调查词汇量和阅读理解测试成绩间是否有关系。在词汇测试和阅读理解测试中分别安排15个六年级学生（所有分数都用正确百分率来表示），15个小学生的测试结果如下（词汇分数在前）：44, 67; 24, 33; 67, 45; 75, 54; 34, 45; 88, 79; 57, 67; 44, 32; 87, 95; 77, 67; 87, 78; 54, 67; 90, 78; 36, 55; 79, 91。画一幅散点图并计算出这些数据的相关系数。
 - 解释你是否能够用14题中计算出的相关系数来说明词汇量增加导致阅读理解水平的提高。

挑战性问题

- 一位认知心理学家研究四种不同环境下朗读有关盖茨堡战役的文章后，记忆保持程度如何。让我们简单地将四种条件表示为A, B, C和D。64 ($N=64$) 个大

试。因变量是立即进行回忆并写出有关文章信息的数量。四种表述条件下回忆的平均数和标准差如下：

	A	B	C	D
<i>M</i>	16.4	29.9	24.6	19.5
<i>SD</i>	4.6	7.1	5.9	6.3

(续)

- A. 计算四个样本平均数所估计的总体平均数在 95% 上的置信区间。
- B. 我们可以在各个总体平均数差异中得出什么样的结论? 解释置信区间的模式。
2. 一位发展心理学家研究母亲怀抱孩子的姿势对孩子睡眠的影响。调查者恳求 40 位刚做妈妈的母亲帮助他, 他训练其中的 20 位母亲抱小孩时把新生儿的头靠近她们的胸部; 另外 20 位母亲则没有什么特殊的怀抱方法。并且他还训练所有母亲记录 24 小时中孩子睡觉的时间。两个组都保持这样的记录 3 个月。结果发现, 指导组婴儿在 24 小时里平均睡眠时间为 12.6 小时 ($SD=5.1$); 在非指导组中, 平均睡眠时间为 10.1 小时 ($SD=6.3$)。
- A. 计算两个平均数差异在 95% 上的置信区间。
- B. 这个实验得出的置信区间能够说明什么?
- C. 这个实验效应的大小是多少? 解释基于科恩测量方法的效应大小, 包括小、中和大效应。
3. 一个研究者要求大学生在听古典音乐和摇滚音乐时玩一个指定的电子游戏。实验中所有 10 个学生在每个音乐环境下玩 15 分钟游戏。一半学生玩游戏时先听古典音乐再听摇滚音乐, 另一半则相反 (见第 8 章关于平衡重复测量设计的内容)。因变量是在 15 分钟内游戏时正确击中目标物的数量。10 个学生的成绩是:

学 生	古典音乐	摇滚音乐
1	46	76
2	67	69
3	55	51
4	63	78
5	49	66
6	76	67
7	58	63
8	75	75
9	69	78
10	77	85

- A. 计算每种条件的平均数。在平均数的对比中你看见了什么趋势?
- B. 计算差异分数的标准误。
- C. 找出在这个重复测量设计中两个平均数差异在 95% 上的置信区间。
- D. 通过对这些结果分析不同音乐类型对游戏成绩的效应。
4. 一个社会心理学家寻求纸笔偏见测试和种族刻板印象的关系。在学期初, 一个普通心理学班的学生被要求完成 6 个不同的问卷, 这些问卷用来测试偏见。在随后学期中, 学生被邀请参加一个实验, 用以调查他们对犯罪行为和法律强制措施的态度。作为实验的一部分, 学生们完成问卷时询问他们对于犯罪与种族关系的态度。研究者希望找出先前进行的偏见测量分数是否能够预测人们对于种族刻板印象的态度。在偏见测试中高分说明偏见比较大, 在种族刻板表中高分说明更加支持种族刻板印象。22 个学生的测量分数如下:

学 生	1	2	3	4	5	6	7	8	9	10	11
偏 见	19	15	22	12	9	19	16	21	24	13	10
种族刻板	7	6	9	6	4	7	8	9	5	5	7

学 生	12	13	14	15	16	17	18	19	20	21	22
偏 见	12	17	23	19	23	18	11	10	19	24	22
种族刻板	4	8	9	10	10	5	6	4	8	8	7

- A. 画一幅散点图来表明这两个测量间的关系。
- B. 研究散点图并评价这些数据有无线性趋势出现。
- C. 计算这些数据的相关系数, 并评价这种关系的方向和强度。
- D. 在相关分析的基础上, 研究者总结出偏见思维导致人们支持具有种族刻板认识。基于你所知道的相关证据的本质来评价这种结论。

挑战性问题 1 的答案

- A. 首先计算四个组的 S_{pooled} , 一定要注意问题提供了每个组的标准差和利用各种变量的公式。这样, 每个标准差必须再乘以 $n-1$ 。 s_{pooled} 的值为 6.04, 因此估计的平均数标准误是 $6.04/\sqrt{n}$ 或 1.51。根据表 A.2, t 在 .05 水平上的临界值是 2.00 (60df)。平均数的置信区间是:

第 13 章

数据分析与解释 (II):
统计显著性检验与分析

本章大纲

概 述

虚无假设显著性检验 (NHST)

实验敏感性和统计效力

NHST: 两个平均数的比较

独立组设计

重复测量设计

统计显著性和理论及实际意义

均数比较的建议

均数比较的结果报告

多重条件的数据分析

单因素独立组方差分析

三个以上独立组设计的效应大小计算

独立组设计的统计效力估计

多组实验设计的均数比较

单因素独立组实验设计的结果报告

重复测量的方差分析

独立组设计的双因素方差分析

含有交互作用的复合设计分析

无交互作用的方差分析

双因素独立组设计的效应大小计算

复合设计分析中置信区间的作用

双因素混合设计的方差分析

复合设计的结果报告

小 结

概 述

在第12章中,我们介绍了数据分析的三个主要阶段,即了解数据特点、概括数据结果以及确认数据的意义。在数据分析的最后阶段,我们将判断统计结果是否能为我们作出某种行为推论提供足够证据。在这些数据基础上,我们能否讨论这些行为?这一阶段有时被称作求证性数据分析(如,Tukey, 1977)。以此为出发点,我们将通过统计方法去求证数据的真实含义。在第12章中,我们强调用置信区间去分析数据。在本章中,我们将继续通过统计检验或虚无假设的显著性检验(NHST)探讨求证性数据分析。

NHST是求证性数据分析中最常用的方法。然而,统计显著性检验由于种种原因而备受批评(Cohen, 1995; Hunter, 1997; Loftus, 1991, 1996; Meehl, 1967; Schmidt, 1996)。几十年来,研究者们忽略了统计显著性检验中存在的问题,一直在误用(或错误解释)这些方法(Finch, Thomason, & Cumming, 2002)。甚至有批评者建议我们放弃这种方法(Hunter, 1997; Schmidt, 1996)。然而,大多数学者建议我们可以继续使用NHST,但应当谨慎使用(Abelson, 1995, 1997; Chow, 1988; Estes, 1997; Greenwald, Gonzalez, Harris, & Guthrie, 1996; Hagen, 1997; Krueger, 2001; Mulaik et al., 1997)。无论心理学界对这个问题的争论结果如何,大家的共识在于我们需要准确地了解NHST到底能做什么、不能做什么。有时,其他方法会与NHST势不两立,但有时候它们又是对NHST的一种补充。

在本章中,我们首先简要介绍NHST,然后再讨论实验敏感性和统计效力等概念。之后,我们将用NHST方法对第12章中计算出的两组平均数差值置信区间的数据进行分析。通过用NHST和置信区间对同一组数据进行分析,我们可以比较它们所得到的统计结果,并据此指出NHST到底可以做什么,从NHST分析中得到的结果如何对置信区间分析进行补充。最后,我们会提供一些建议,帮助你评价涉及均数比较研究的行为结论,并举例说明如何为你的研究构建一个完整的分析结论。

在对包含两个或两个以上组的研究进行数据分析时,最常用的方法是求证性数据分析,它是NHST的一种形式,也叫方差分析(ANOVA)。在本章的后半部分,我们将讨论使用ANOVA的原则、计算过程以及结果解释等。

虚无假设显著性检验(NHST)

- 虚无假设检验用来确定一个实验中的组间平均差是否比误差变异所引起的差异大。
- 虚无假设检验的第一步是假设两组之间没有差异,即自变量没有效应(虚无假设)。
- 假定虚无假设为真,就需要用概率论对实验结果的可能性做出估计。

- 如果虚无假设为真,那么,统计显著性结果出现的可能性就小。
- 由于对实验结果的判断建立在概率基础上,因此,可能会产生 I 类错误(拒绝了一个真的虚无假设)或 II 类错误(无法拒绝一个假的虚无假设)。

统计推论具有归纳性和间接性。之所以具有归纳性,原因在于我们是通过特定样本对总体进行一般推论,例如,计算置信区间。但是,与计算置信区间所用的方法不同,这种统计推论也具有间接性,因为它首先要建立虚无假设。虚无假设(null hypothesis) (H_0) 认为,自变量没有效应。一旦形成这种假设,我们就可以用概率论确定实验中由误差引起变异的可能性有多大,如果可能性低,我们就拒绝虚无假设,并得出自变量对因变量的确产生了效应的结论。拒绝虚无假设所依据的结果就是统计显著性。一个显著的统计结果仅仅意味着实验条件引起的变化大于只有误差(机遇水平)存在时所引起的变化(见专栏 13.1)。

统计结果显著说明虚无假设为真的可能性很小。但是,小到多大才可以接受呢?虽然对这一重要问题没有一个绝对的答案,但在科学研究领域,多数人认为,如果虚无假设为真的概率小于 5% (0.05),那么,就可以判断为在统计上是显著的。我们将在统计上显著的概率称作显著性水平(level of significance)。显著性水平用希腊字母 α 表示。因此,当我们说 0.05 水平、0.10 水平或 0.01 水平时,就分别用 $\alpha=0.05$ 、 $\alpha=0.10$ 或 $\alpha=0.01$ 表示。

显著的统计结果具有怎样的含义呢?我们获得的最有用的信息在于我们知道实验中出现了我们感兴趣的结果。更具体地说,观察结果出现的概率越小,则重复实验达到统计显著性的概率越高。但我们作出结论时必须很慎重。有时,研究者错误地认为,当在统计上 $p<0.05$ 时,那么,该结果重复出现的概率是 95%。实际情况并非如此,仅仅达到统计显著性 ($p<0.05$),并不能说明结果重复的概率。例如,某一结果出现的概率小于 0.05 (达到统计显著性),但也许重复该事件达到统计显著性的可能性仅有 50% (Greenwald et al., 1996)。当然,结果的准确概率

专栏 13.1

正面还是反面? 掷硬币与虚无假设

我们可以通过下面的两难问题来了解统计推论过程。一位朋友狡黠地笑着拿出一个硬币,要用它来决定晚餐由谁付钱。他已经准备好了要掷,你可以检查一下他的硬币是否有诈(要求检查)。然而,为了不破坏朋友之间的信任,你可以通过间接的方法进行检验:先假设硬币没有问题,然后,看扔的结果是否是正反面各占 50%。若干次后,也许你会觉得你的朋友试图通过某种手段最终让你付账。在实验中,我们也用类似方法对结果进行统计显著性检验。然而,我们所能做的只是将实

际结果和理想的正反面出现次数相等的结果进行比较。理解虚无假设的关键在于,我们能够认识到用概率规则去估计的是当只有机遇因素存在时结果产生的概率。这和你反复检查朋友的硬币然后再得出结论是一致的。众所周知,在机遇水平上,硬币正面和反面朝上的概率各占 50%。掷过若干次后,如果结果不是这样,那么你就有理由认为,有机遇以外的因素在起作用,即你的朋友可能作弊了。

确实可以对结果重复出现的可能性进行预测。概率越小,重复实验达到统计显著性的可能性越高(Posavac, 2002)。因此,美国心理学会强调,在进行 NHST 时,应当报告一个准确的概率。

不要在统计分析之后再选择显著性水平,应该在实验前就选择好。在进行统计分析之前确定显著性水平,可以避免我们将结果出现的概率当成显著性水平。严格来讲,在做推论统计时只有两种结果:拒绝虚无假设,或者是无法拒绝虚无假设。请注意,我们并没有说接受虚无假设。下面我们详细解释。

在实验中,当自变量的效应在统计上不显著时,我们并不能拒绝虚无假设。因此,在没有差异时,我们也不能接受虚无假设。实验中可能存在某种因素混淆了自变量效应(例如指导语模糊,自变量的操作性定义不够精确等)。我们在后面还会提到,样本小常常是无法拒绝虚无假设的主要原因。尽管从逻辑上无法证明虚无假设为真,但我们必须设法确定哪些自变量是无效的。NHST 可以帮助我们做出判断。统计上不显著的结果说明,自变量对行为的影响方式非常复杂。在这种情况下,我们希望获得更多信息,如样本大小、效应大小等(见下一节,实验敏感性和统计效力)。

在进行推论统计和以概率为决策依据的过程中仍然存在一个问题。不论做什么决策,也不论如何做这种决策,犯错误总是难免的。表 13.1 中列举了实验者可以获得的两种结果和两种正确决策。两种结果是自变量有或无可能影响行为。研究者可能的正确决策位于表中的左上角和右下角。如果自变量产生了效应,那么,研究者就应当拒绝虚无假设;如果自变量没有产生效应,那么,研究者就不能拒绝虚无假设。

表 13.1 中的另外两部分是两类错误(I类错误和II类错误)。这些错误的产生是由统计推论的概率性质决定的。当我们判断结果在统计上显著时,就意味着虚无假设条件下结果出现的概率小于 0.05,即 100 次试验中出现 5 次。而即使虚无假设为真,结果也有可能发生。因此,显著性水平代表犯 I 类错误(type I error)的概率,即拒绝了真实的虚无假设。可以通过提高显著性水平的方法来降低 I 类错误,如 $\alpha=0.01$ 。但这种做法的问题是它增加了 II 类错误(type II error)的可能性,即虚无假设为假时却没有拒绝。

虽然 I 类错误和 II 类错误问题并不能束缚我们,但它让我们知道,为什么研究者在描述实验的统计显著性结果时很少用“证实”,而是用“与假设一致”、“验证假设”或“支持假设”等表述方法,这些表述方式中暗含着 I 类错误和 II 类错误总是存在的意思。0.05 的显著性水平是一个折衷的取值,它在两类错误中达到平衡,避免某一类错误太大。I 类错误和 II 类错误问题也提醒我们,统计推论永远无法替代作为检验实验结果信度的最佳手段——重复检验。

表 13.1 推论统计中的几种可能的决策结果

	两种结果	
	虚无假设为假	虚无假设为真
拒绝虚无假设	正确决策	I 类错误
未拒绝虚无假设	II 类错误	正确决策

实验敏感性和统计效力

- 敏感性指的是当自变量确实存在效应时,实验能够探测出这种效应的程度。
- 统计效力是指研究者进行统计检验时,正确拒绝虚无假设的可能性。
- 统计检验效力受到统计显著性水平、处理效应大小以及样本大小等因素的影响。
- 提高统计效力的主要方法是增加样本容量。
- 重复测量设计比独立组设计更敏感并具有更高的统计效力。因为重复测量设计中的误差变异相对较小。
- 心理学研究中,使用NHST时犯II类错误的可能性比犯I类错误可能性高。
- 当结果统计上不显著时($p>0.05$),得出虚无假设为真的结论是错误的。

实验的敏感性是指在自变量确实存在效应的情况下,这种效应能够被探测的可能性(见第8章)。对实验而言,它是敏感性,而对统计检验而言,它是具有效力(power)的。统计检验的效力是指当虚无假设为假时,它被拒绝的概率。虚无假设是“没有差异”,当自变量产生了效应,即虚无假设为假,那么,就应该拒绝。请注意,我们将II类错误定义为当虚无假设为假时没有拒绝它的概率,统计效力就可以定义为“1减去II类错误的概率”。

统计效力说明了我们在多大程度上能够“看见”实验效应,在这里它是对研究的可重复性的估计。因为效力是拒绝假的虚无假设的概率,所以我们知道我们错过一个真实效应的可能性有多大。如果统计结果不显著,并且统计效力只有0.3,那么,这说明在10次重复研究中只有3次能够观察到实验效应,而其他7次则观察不到效应。这种情况下,我们必须通过重复实验获得更强的统计效力。否则,我们无法做出存在实验效应的判断。

统计效力由三个因素共同决定,它们分别是统计显著性水平、处理的效应大小以及样本大小(Keppel, 1991)。然而,在实际运用中,样本大小是控制统计检验效力的主要因素,不同样本大小所带来的统计效力的差异非常大。例如,科恩(Cohen, 1988)发现,单因素独立样本设计实验中,探测不同效应所需的样本容量有3个层次。样本容量是30时只能探测大的处理效应,样本容量是76时能探测中等处理效应,而只有当样本容量为464时才能探测出较小的实验处理效应。要探测小的处理效应所使用的样本容量是探测大处理效应所需样本容量的15倍。

重复测量的实验设计也会影响统计分析的效力。如第8章所述,重复测量实验要比独立组设计更加敏感。这是因为重复测量实验中的误差变异比较小。误差变异小可以增强探测实验处理效应的能力。这正是统计效力的含义,即探测已知的较小处理效应的能力。

在介绍NHST时,我们把I类错误和 α (0.05)等同起来。从逻辑上讲,犯这类错误时,虚无假设必须是假的。然而,有批评者认为,定义为零差别的虚无假设“总是假的”(Cohen, 1995, p.1000),或者保守点说“极少是真的”(Hunter, 1997, p.5)。如果一个效应总是或几乎总是出现(平均数之间的差异不是0),那么,我们不可能(至少是几乎不可能)将不存在的效应错误地说成存在。按照这个逻辑推理,我们只能犯II类错误(Hunter, 1997; Schmidt & Hunter,

专栏 13.2

你曾接受过虚无假设吗？

到现在为止，不论我们介绍了多少内容，总会有一些接受虚无假设（不是简单地拒绝它）的例子。伊汤和塞科雷斯特（Yeaton & Sechrest, 1986, pp. 836-837）令人信服地说明，应用研究中没有差异的结果非同小可。下面节选了他们用来描述自己观点的一些问题：被送进托儿所的儿童在智力、社会交往和情绪等方面是否比待在家里的儿童更出色？副作用小、价格便宜的新药

与现有药品在治疗心脏病方面是否具有等同的疗效？

这些重要的问题清楚地表明，接受虚无假设（无效应）不仅仅是一个理论问题，做出正确决策还涉及到一个像生与死这样的重要结果。弗里克（Frick, 1995）认为，对心理学而言，总是不断地接受虚无假设既不是我们的愿望，而且也失去了实践性。我们往往很难理直气壮地说：它们之间没有差异（Shadish et al., 2002）。

1997），也就是说实际的效应不存在。在许多心理学研究中，这类错误主要是由于统计效力低造成的，通常情况下，远远大于 0.05（Cohen, 1990; Hunter, 1997; Schmidt & Hunter, 1997）。如果只是在字面上接受虚无假设，即总体平均数之间的差异为零，或者如果我们相信，在某些情况下值得对效应进行检验，以反对无差异假定，那么，我们可以说肯定会产生 I 类错误（Abelson, 1997; Mulaik et al., 1997）。作为一个研究者，我们必须注意到，在有些情况下，当不存在效应时，就不要得出有效应的结论（见专栏 13.2）。

当统计效力低时很容易产生 II 类错误。文献中有许多研究的统计效力很低，心理学研究中使用 NHST 最常见的错误是 II 类错误。统计上没有得到显著性结果，并不意味着不存在实验效应（Schmidt, 1996）。事实上，计算效应值的重要原因在于我们可以将获得的效应与其他研究进行比较，而不管效应是否具有统计显著性。这就是元分析的目的（见第 7 章）。

要想在研究之前确定统计效力，就必须对实验中期望的效应值进行估计。对以往研究的效应值进行分析将有助于你做出合理估计。一旦确定了效应值，就可以从“效力表”中获得要“看见”效应所需的样本容量。效力分析步骤在均数比较的效力表附录中有详细介绍。要想获得效应值的准确估计，最好在研究之前进行效力分析。

在研究报告中也可使用效力表。当研究完成后而结果不显著，APA《出版手册》（2001）建议在研究结果中应该报告统计效力，通过这种方式可以和其他研究者交换探测已知效应的可能性。如果这个概率比较低，那么，研究者们可能会继续等待更多的重复研究，以便获得更加准确的结果。另一方面，如果统计结论不显著而统计效力足够大，那么，其他研究者就不必要再为此花费时间了。附录中也详细介绍了这方面内容。

NHST：两个平均数的比较

- 比较两组被试的平均数最恰当的推论统计检验方法是独立样本 t 检验。

- 使用 NHST 时, 应该同时报告效应大小。
- 比较同一组被试(或匹配组)的两个平均数的最恰当的推论统计检验方法是重复测量(被试内设计)的 t 检验。

下面我们介绍如何使用 NHST 比较两平均数。首先, 我们假设, 一项研究中有两组独立的平均数(数据曾在第 12 章中使用过, 来自于泽克迈斯特等(1995)关于词汇学习的研究)。然后, 我们假设有两组取自重复测量设计的相互关联的平均数。

独立组设计

大家还记得, 在词汇研究中, 研究者测量了大学生和老年人的词汇量。该组数据的适当分析方法是独立组 t 检验 (t -test for independent groups)。我们可以用这种方法检验大学生和老年人词汇量测验中的成绩的差异。我们把 t 定义为两个独立组平均数差值除以该差值的标准误, 即:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}, \text{ 其中}$$

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right] \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}$$

用上述公式计算得 $t=5.84$, 显著性水平选择 0.05。附录 A.2 中可以查出相应自由度 (df) 的 t 的临界值。当 $df=50$ 时 ($df=N-2$, $N=n_1+n_2$, 是两个样本容量的总和), t 的临界值小于 5.84。因此, t 检验结果显著。¹

一般的统计软件计算的结果中都会给出 t 值所对应的精确的概率值。APA 的《出版手册》(2001)中建议将精确的概率值也报告出来, 当精确的概率值小于 0.001 (如 $p=0.0004$) 时, 统计软件输出的结果将写成 0.000。当然, 精确的概率值不是 0.000, 而是小于 0.001 的某一个数。在这种情况下, 研究者不知道精确的概率值是多少, 一种解决办法就是将结果写成 $p \leq 0.0005$, 因为大于 0.0005 的数将被四舍五入成 0.001。

因此, 在词汇量实验中, 推论统计结果如下:

$$t(50) = 5.84, p \leq 0.0005$$

在第 12 章中, 我们介绍了如何计算两个均数比较的效应值 d 。在使用 NHST 时也应该给出效应值的大小。第 12 章中我们计算的词汇量研究中的效应值为 1.65。科恩 d 值也可以从独立组 t 检验中计算出来, 公式如下:

¹ 在附录 A.2 中并没有自由度为 50 的 t 值。在这重情况下, 可以将 t 值与表中与 50 相邻且最小的自由度所对应的 t 值进行比较。附录中, $df=40$ 所对应的临界值是 2.02。而计算的 $t=5.84$, 显然大于表中的临界值。正如文中所述, 使用统计软件进行统计分析时, 精确的概率值将自动给出, 一般也将被报告出来。

$$d = \frac{2t}{\sqrt{df}} \quad (\text{见 Rosenthal \& Rosnow, 1991})$$

即

$$d = \frac{2(5.84)}{\sqrt{50}} = \frac{11.68}{7.07} = 1.65$$

重复测量设计

我们已经讨论了两个独立组的实验设计。众所周知,在实验中也可以让每一个被试接受所有实验处理条件,或者按照某一变量(如 IQ 分数、重量等)将被试进行匹配。这种实验设计叫做匹配组设计(见第 7 章)或重复测量设计(见第 8 章)。重复测量设计的 NHST 与独立组实验设计相同,但两者的 t 检验则有所区别。这种 t 检验被称作直接差异 t 检验或重复测量(被试内) t 检验(repeated measures (within-subjects) t -test)。 t 值计算公式如下:

$$t = \frac{\bar{D}}{s_{\bar{D}}}$$

其中

$\bar{D}$ = 平均差或 $(\sum D) / N$

$s_{\bar{D}}$ = 差数的标准误,即 $s_{\bar{D}} = s_D / \sqrt{N}$, s_D 是差异分数的标准差, N 是差异分数的数目(成对分数数目)

重复测量 t 值的分子是得分差数的平均数($\bar{D}$),在数值上等于样本均数之差($\bar{X}_1 - \bar{X}_2$)。统计显著性可以通过查表中自由度为 $N-1$ 所对应的临界值(见附录表 A.2)。在这里 N 是实验中的被试人数或成对数据数目。 t 值的含义与独立组设计中相同。

在第 12 章中我们已经提到,对匹配组和重复测量设计的效应分析要比对独立组设计更复杂(Cohen, 1988; Rosenthal & Rosnow, 1991。在这种情况下主要用 d 值表示)。

统计显著性和理论及实际意义

- 我们必须认识到统计显著性不同于科学研究的显著性。
- 我们还必须承认,统计显著性也不是实践中或临床上的显著性。

统计显著性检验是科学研究结果分析的重要工具。然而,我们必须慎重地正确解释统计显著性的结果(参见专栏 13.3)。我们还必须注意,不要将统计显著性与科学意义上的显著性混淆。研究结果对科学的重要性取决于研究变量的性质

专栏 13.3

当统计结果显著时 ($p < 0.05$), 我们不能做哪些结论

- 我们不能确定平均数间实际差异的准确概率。例如, 我们不能说总体中真实差异的概率是 0.95。

NHST 的结果指的是假定虚无假设为真, 在机遇水平上存在差异的概率的大小。它并不是指现实世界中概率的大小 (Mulaik, 1997)。如果某一结果产生的概率小于我们设定的 α 值 (0.05), 那么, 我们能从中得到的结论只是: 在这种条件下, 该结果不可能是随机发生的。

- 统计显著性结果不能证明研究假设是正确的。(例如, 词汇量研究的结果并不能证明老年人的词汇知识比大学生的丰富。)

NHST (包括置信区间) 不能证明研究假设的正确性。一个显著的统计结果通常只是支持某一假设或为某一假设提供证据, 它无法单独证明研究假设的正确性。原因有以下几个方面: 首先, NHST 是一个概率问题, 它给出的是一个概率值 (p 大于或小于 0.05), 出错的可

能性总是存在, 如果它是一个证据的话, 那也只能算是有条件的证据。如前所述, 研究假设只能在假定虚无假设为真的基础上, 参照结果出现的概率进行间接检验。如果结果在机遇水平上出现的概率很低 (假定虚无假设为真), 那么, 虚无假设就不是真的。正如施密特和亨特 (Schmidt & Hunter, 1997) 所述, 研究者在使用 NHST 时 “并不关心真实的科学假设本身”。其次, 自变量效应的效力仅与产生这种效应的方法密切相关。NHST 中使用的数据说不定来自存在变量混淆或实验误差的研究中。因此, 有可能是其他因素产生了实验效应。(例如, 假设词汇研究中的老年被试是一些拼字游戏高手。) 我们前面已经提到, 一个拙劣的实验也可以产生巨大的实验效应。证实研究假设的证据不仅来源于因变量的效应, 同时也与研究方法密切相关。无论是 NHST、置信区间, 还是效应值都不能说明研究方法是无懈可击的。

(某些变量的效应比其他变量更强)、研究本身的可信性 (具有统计显著性的结果可能来自有缺陷的研究) 以及效应大小等因素 (Abelson, 1995)。

处理效应在实践或临床上的显著性也同样不仅仅取决于统计显著性, 它还包括研究的外部效度、效应大小以及与治疗有关的各种因素 (包括经费投入)。在实验室中获得的巨大实验处理效应未必能够在现实世界中出现 (即外部效度低)。因此, 某些研究结果对应用心理学家来说可能没有多少价值。当然, 有些具有巨大效应但无法推广到现实情景中的结果可能永远无法应用到实践中, 因为这些结果可能过于昂贵、难以实现、与实践背道而驰或与现有的处理太相似。

当统计效力足够强大时, 效应值小也可能产生统计显著性。如第 7 章所述, 外部效度是一个典型的问题。研究中设定的情景与研究结果所要应用的实际情景的相似性非常重要。因此, 如果一个研究的效应值小, 我们不可能去开展应用推广 (Rosenthal, 1990)。

均数比较的建议

在比较两个平均数之间差异时, 我们提出以下建议: 首先, 要记住数据分析

的最终目的是要从行为观察中获得的数据中得出结论。为了做出最好的结论，你可能去尝试各种不同的数据分析方法。请不要固执地认为，有且仅有一种方法可以分析获得的数据。如果有不同的选择（通常情况下是有的），就像APA的统计推论使用建议（Wilkinson et al., 1999）所推荐的那样，使用最简单的分析方法。其次，在使用NHST时，必须清楚地了解它的局限性及其结果的含义。应该报告效应大小、统计效力等指标，尤其当统计结果不显著的时候。然而，有些情况下效应大小信息得不到保证。例如，在对理论的方向性预测进行检验时（Chow, 1988），但这种情况很少见。在很多研究情景中和几乎所有的应用情景中，效应值信息是对NHST的一个非常重要和必要的补充。最后，研究者应当改变仅仅依赖NHST的习惯，还应当考虑报告置信区间，甚至包括与推论检验结果有关的 p 值。APA《出版手册》（2001）强烈推荐在报告中应该包括置信区间的结果。

均数比较的结果报告

下面我们开始以数据分析的三个阶段：来自置信区间的补充证据（第12章）、NHST分析以及《APA出版手册》的推荐方法，来讨论研究结果的报告模式。

词汇研究的结果报告如下：

大学生词汇多项选择测验平均分数为45.58 ($SD=10.46$)，老年组的平均成绩为64.04 ($SD=12.27$)。在0.05显著性水平上两者差异显著， $t(50)=5.85$ ， $p \leq 0.0005$ 。在本研究中老年人的词汇量大于大学生。效应大小科恩 d 值为1.65。置信区间为12.11至24.81内，有95%的可能性包含了总体平均差。

说明 用平均数、标准差等描述统计量对自变量在实验中引起因变量变化关系进行描述。按照《APA出版手册》的推荐，NHST的 α 水平应在 p 值之前给出。因为精确的概率小于0.001，而结果报告成 $p \leq 0.0005$ ；但是，当精确的概率为0.001或更大时才会将它写出来。精确概率包含了结果重复出现的精确概率信息（Posavac, 2002）。即精确 p 值越大，结果可靠性越低。如果只报告置信区间，就无法获得这些信息。“在本研究中老年人的……”这句话总结了统计分析结果。它直接告诉读者统计分析的结果。这一点随着研究中采用的统计分析的数量和复杂性的增加而显得越来越重要。《APA出版手册》建议，在研究报告中也应给出效应大小。这些信息对于做元分析的研究者和使用同样变量比较不同研究结果的研究者都非常重要。另外，置信区间给出实际平均差可能的效应区间，而不是单个数值，如科恩 d 值。由于0并不在区间内，而统计结果在0.05水平上显著（见第12章）。因此，《APA出版手册》强调指出，置信区间给出了效应位置的精确估计，而NHST并不包含这方面的信息。第12章中曾经指出，置信区间越小，估计越精确。

注意：在第12章已经指出，在涉及到两组平均数时，不需要用图表表示它们。只有在报告多于2组的数据时图表才显得更加重要。如果用图表描述结果，文字叙述必须指向这些图表（见第14章）。

拓展练习

测验你对虚无假设的理解

现在应该明白,理解、应用和解释 NHST 不是一件容易的事。即使有经验的研究者也会犯错。为了帮助你避免犯错,我们提供了一个关于 NHST 的测验,请根据给定信息做出判断。

假设用一个独立组设计来测验实验组和控制组的成绩。每种情形有 12 个被试, NHST 显著性为 .05, $t(22) = 4.52, p = .006$ 。研究者的结论如下,请判断正误。

1. 应该拒绝虚无假设。
2. 研究假设被证明是正确的。
3. 结果具有重要的科学意义。
4. 虚无假设正确的可能性只有 .006。
5. 如果重复实验, .05 水平统计显著性比 .02 的显著性发生的概率大。

答案: (1) T, (2) F, (3) F, (4) F, (5) T。

统计效力分析 当知道效应大小时,我们就可以进行统计效力分析。前面已经说过,统计效力是获得统计显著性效应的可能性。例如,老年人和大学生的词汇比较研究的效应大小为 0.5,按照科恩(Cohen, 1988)的经验定律,这是一个中等效应。我们可以从科恩的效力表中查到在平均差的检验中,要在 0.05 水平上“看见”0.5 的效应需要多少被试(见附录 A.4)。该表中标出了不同效应大小下统计效力与样本量之间的关系。在统计效力表中,就双组的对比研究而言,要使统计效力达到 0.8(双尾检验)时每组所需被试人数为 64。要达到中等效应大小,我们总共需要 128 (64×2) 个被试才能够在 10 次尝试中获得 8 次显著的结果。尽管泽克迈斯特等人希望获得中等的效应,但他们的词汇研究的统计效力偏低。虽然期望获得大的效应值,但从 26 个被试的结果中获得的统计效力为 0.80。

如果统计检验结果不显著,就必须给出统计效力的估计值。例如,如果使用独立组设计的统计结果为 $t(28) = 1.96, p > 0.05$,效应值为 0.5。那么,我们可以在事后决定其统计效力。假设两等组的人数都为 15 ($df = n_1 + n_2 - 2$ 或 $28 = 15 + 15 - 2$)。从附录表 A.4 中我们可以发现,该研究的统计效力为 0.26。在这样的样本容量条件下,获得显著统计结果的概率仅为 1/4,只有中等效应(0.50)。这种情形下,研究者必须决定,这样的结果是否具有实际和理论意义,或者是否还需要更多研究才能说明问题。

多重条件的数据分析

到目前为止,我们讨论了仅有两种实验条件或自变量只有两个水平的数据分析过程。当自变量包含两个以上水平(或条件),或者在心理学研究中通常包含了两个以上的自变量时,我们将如何进行数据分析呢?这种情况下的数据分析方法通常是方差分析(ANOVA)。

我们将介绍在以下 4 种研究中如何用 ANOVA 进行虚无假设的检验分析。它

们是单因素独立组设计、单因素重复测量设计、双因素独立组设计和双因素混合设计等。建议在学习这些之前先复习一下第7、8、9章关于实验设计的内容。

单因素独立组方差分析

- 方差分析是检验自变量是否对因变量产生显著统计效应的一种方法。
- 方差分析的基础是区分数据中的误差变异和系统变异。
- F 值是数据的组间变异和组内变异的比值。
- 方差分析总表包含了方差分析的各种结果，通过比较两个平均数，可以确定系统变异的特定来源。
- 尽管方差分析可以用来确定是否在统计上具有显著效应，但研究者仍然需要按照描述统计结果去解释实验结果。
- 独立组设计的效应大小测量包括 η^2 和科恩 f 值。
- 为了确定统计显著性的概率，应在研究实施前进行独立组设计统计效力分析；在进行 NHST 分析时，如果结果不显著，必须给出统计效力值。
- 两个平均数的比较可以识别在统计显著性中起作用的特定系统变异。

基本概念 统计推论要求检验实验结果是否在统计上显著。在心理学研究中，最常用的推论统计检验方法是方差分析（ANOVA）。顾名思义，方差分析是以实验中不同变异来源的分析为基础的。在本节中，我们将简要介绍单因素独立组设计（single-factor independent group design）的方差分析。尽管 ANOVA 既可以用来分析随机组也可以分析自然组设计的结果，但 ANOVA 的假设只是被严格限定在对随机组设计的数据分析。

在随机组实验中存在两个变异源。首先是由于被试中的个体差异造成的组内变异，这种变异在实验中是无法消除的，但可以通过随机分配被试加以平衡。在良好的实验中，同一组内不同被试间的差异是惟一的误差来源。因此，每组中所有被试的指导语应该以相同的方式呈现，被试接受处理条件的方式也应当相同（见第7章）。

随机组设计的第二个变异来源是组间变异。如果虚无假设为真（组间无差异），组间平均数的任何差异都应该是误差变异（即在各种被试的不同特点）。然而，我们已经知道，样本平均数不可能完全相等。由于存在取样误差，样本平均数可能会产生波动，即误差变异。因此，当虚无假设成立时，组间平均数的变异为实验的误差变异提供了第二个估计。如果虚无假设为真，那么，组间误差变异和组内误差变异应该相等。因此，随机组设计提供两种误差变异，一个是组间变异，一个是组内变异。

现在我们假设虚无假设为假，即实验中自变量产生了效应。如果自变量存在效应，那么，不同组的平均数是相等还是不同呢？你肯定认为它们是不同的。自变量存在效应时，它肯定使不同组的平均数之间产生系统差异。也就是说，自变量将产生一个组间变异源，它将使得不同组的平均数发生变化。这种系统的变异

将被叠加到由误差变异所引起的不同组平均数之间的差异上,即组间变异增大。

F 检验 现在我们来讨论用何种统计量判断自变量变异和误差变异哪个更大,这个统计量就是 F 值,它是根据发明这种统计方法的统计学家罗纳德·费舍(Ronald Fisher)的名字命名的。 F 检验(f -test)的定义如下:

$$F = \frac{\text{组间变异}}{\text{组内变异}} = \frac{\text{误差变异} + \text{系统变异}}{\text{误差变异}}$$

如果虚无假设为真,组间的系统变异为 0(自变量没有效应), F 值为 1.00(误差变异除以误差变异等于 1.00)。然而,随着系统变异的增加,则 F 值越来越大,大于 1.00。

如果我们能够把自变量的系统变异独立出来,那么,实验结果分析就会变得非常容易。但遗憾的是,组间的系统变异与误差变异共生在一起。因此,有时 F 值大于 1 只是因为对组间误差变异的估计刚好大于组内误差变异的估计(两个误差变异估计应该相等,但也可能因为随机因素而不等)。当 F 值比 1 大多少时我们才能确信真实的系统变异是由于自变量引起的呢?我们在前面讨论统计显著性时已经回答了这个问题。为了达到统计显著性, F 值必须足够大,以便在虚无假设为真时事件出现的概率小于我们所设定的显著性水平,通常是 0.05。

下面我们可以应用 NHST 原则和 ANOVA 过程分析不同类型实验数据。

单因素独立组设计方差分析 用 F 检验进行推论检验的第一步是确定统计分析所要回答的研究问题。通常问题的形式是“自变量对因变量是否存在效应”。一旦确定了要研究的问题,下一步就是要生成虚无假设。我们以不同记忆训练方法对记忆保持量的影响为例。自变量有 4 个水平(条件),因此,就选取了 4 组被试,每个样本或组代表一个总体。开始对实验的整体分析叫综合 F 检验(Omnibus F -test)。它的虚无假设是所有 4 组的总体平均数相等。虚无假设假定自变量没有效应。正式的虚无假设(H_0)按总体特性表述。总体平均数用 μ (“mu”)来表示。我们可以用不同表达式代表自变量的不同水平。这里的虚无假设是:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

与虚无假设相对立的应该是不同总体的平均数不相等。换言之,备择假设(H_1)认为 H_0 是错误的,平均数之间存在差异。因此,备择假设应该是:

$$H_1: H_0 \text{ 为假}$$

如果记忆训练类型确实影响记忆保持量(如果自变量产生了系统变异),那么,我们将拒绝虚无假设。

表 13.2 中是不同训练条件下,记忆测验中正确回忆单词的个数(共有 20 个单词)。按照记忆任务不同将被试随机分成 4 组,每组有 5 名被试。控制组没有任何指导语,其他 3 组的任务分别是用编一个故事(编故事法)、用视觉想像(想像法)或用押韵(韵律法)等方法记住给定的词。自变量是指定的记忆方式,用字母 A 表示。自变量的不同水平分别用 a_1 、 a_2 、 a_3 和 a_4 表示。每组的被试数量用 n

表 13.2 记忆实验中正确回忆单词的数量

被 试	指定的记忆方式 (A)				被 试	押韵组	
	控制组 (a ₁)	编故事组 (a ₂)	想像组 (a ₃)	押韵组 (a ₄)			
1	12	6	15	11	16	16	14
2	10	7	14	12	16	17	14
3	9	8	13	13	13	18	15
4	11	9	12	14	12	19	12
5	8	10	12	15	15	20	12
平均数	10.0	13.2	14.4	13.4			
标准差	1.6	1.3	1.8	1.3			
全距	8-12	12-15	12-16	12-15			

表示, 在这里 $n=5$ 。整个实验的被试数量用 N 表示, $N=20$ 。

分析中的一个重要步骤是构建类似表 13.2 的数据表。每个被试在各自条件下的正确反应次数都列在表中。为了理解实验结果, 在进行 ANOVA 之前对数据进行整理非常重要。将各组的平均数、标准差和全距等统计量列在原始数据下方。

在进行任何推论统计的显著性检验之前, 都应该对大致的统计结果有所了解。看一下自变量是否存在明显的效应, 即平均数之间是否存在变异。可以通过检查全距和标准差, 以便对每个组的变异有一个大致了解。(记住: 分数在样本平均数周围波动越小, 越容易看见存在的效应。) 全距或最大值和最小值之间的差值对于鉴别地板效应和天花板效应非常有用。组内的变异性相似吗? 我们希望变异相对同质, 因为在进行 ANOVA 时, 组内变异性太大将会使结果解释产生问题。

我们在对汇总的统计结果进行检查时发现, 平均数之间存在系统变异; 控制组 (10.0) 和想像组 (14.4) 的差异最大。所有实验组的平均数都大于控制组的平均数。并且所有组的全距和标准差都相似。这说明组间方差具有同质性 (相似)。许多计算机程序在进行方差分析时都会给出方差同质性检验。因此, 通过对该数据中每组的最大分数的检查表明, 不存在天花板效应。

方差分析的下一步是计算用于估计 F 值的分子和分母。最好使用计算机计算 F 检验的结果。因此, 我们将重点介绍如何解释计算出的结果。方差分析结果见表 13.3。

ANOVA 总表的解释 检验记忆训练效应的独立组设计的方差分析结果见表 13.3。注意: 4 个独立组中都有 5 个被试, $n=5$, 共计 $N=20$ 。了解 ANOVA 表中内容含义

表 13.3 记忆实验的方差分析总表

来源	平方和	自由度	均方	F 值	p
组间	54.55	3	18.18	7.80	0.002
误差	37.20	16	2.33		
总和	91.75	19			

非常重要。因此,在看实验效应之前,先熟悉一下表中各项的含义。

表的左栏中列出两个变异源,这两个变异源我们在前面已经介绍过。在本例中,自变量训练组是一个组间变异源,而组内变异可以作为对误差变异的估计。总变异是组间和组内变异之和。第3列是自由度(df)。通常情况下,自由度被定义为数据总数减1。因为自变量记忆训练有4个水平,所以,组间自由度为3;每组中有5名被试,因此,每组中的自由度为4。因为4组人数相同,所以,组内自由度就等于每个组内的自由度乘以组数,即 $4 \times 4 = 16$ 。总自由度等于被试总数减1($N-1$),或者等于组间自由度与组内自由度之和($3+16=19$)。

平方和(SS)和均方(MS)是计算 F 值所必需的。组间均方(第一行)是系统变异与误差变异之和,它等于组间平方和除以组间自由度($54.55/3=18.18$)。组内均方(第二行)等于组内平方和除以组内自由度($37.20/16=2.33$)。 F 值就等于组间均方除以组内均方($18.18/2.33=7.80$)。

下面我们就可以用方差分析表中的信息进行记忆训练实验的统计显著性检验了。也许你已经迫不及待地想得出结论,想知道虚无假设在什么情况下成立(没有效应),即系统变异与误差变异之和(计算 F 值的分子)应该与误差变异(计算 F 值的分母)相接近。在本例中,系统变异与误差变异之和(18.18)远远大于误差变异(2.33)。

表中的倒数第二列就是计算的 F 值(7.80)。表中的最后一列中是虚无假设为真的概率(0.002)。此概率小于显著性水平($\alpha=0.05$),因此,我们拒绝虚无假设,并且得出记忆训练效应在统计上是显著的。使用ANOVA进行虚无假设显著性检验的最终结果为:

$$F(3, 16) = 7.80, p = 0.002$$

F 值的统计结果是否显著决定于它的自由度。在本例中,组间和组内自由度分别为3和16。同时,给出了精确的概率,因为它能够告诉我们有关结果重复出现概率的信息。

当方差分析在统计上显著时,我们能够从中获得怎样的信息呢?从整体上看,我们了解了很重要的信息。我们可以说对自变量的操纵引起了因变量(词汇记忆成绩)的变化。另一方面,仅有统计显著性并不能说明自变量效应的性质。描述统计结果(在本例中见表13.2的词汇回忆平均数)则可以告诉我们效应的性质。只有通过了解各组平均数的变化模式,我们才能够知道实验中自变量与因变量之间的函数关系。一定要结合描述统计结果去解释相应的统计显著性。

尽管我们知道 F 检验结果显著,但我们并不清楚自变量和因变量之间的关系程度。因此,我们需要计算自变量的效应大小。仅从整体 F 检验中,我们也无法确定哪两组之间差异显著。值得庆幸的是有一种分析方法能够帮助我们确定实验中系统变异的特定来源。一种值得推荐的方法就是使用置信区间(见第12章)。置信区间是通过样本估计总体平均数的一种方法(见专栏12.5)。另一种方法是均数比较。我们首先讨论独立组ANOVA的效应大小计算和这种设计的统计效力分析,然后,再讨论两个平均数的比较。

三个以上独立组设计的效应大小计算

心理学文献中效应大小有许多不同的计算方法, 选用哪一种方法取决于不同的研究设计、不同的统计检验方法以及研究情景的不同特点等 (Cohen, 1992; Kirk, 1996; Rosenthal & Rosnow, 1991)。当我们知道其中一种方法计算出来的效应大小时, 也可以将它换算成另一种。因此, 比较不同方法计算出来的效应大小并不困难。两个以上独立组设计实验中使用的效应大小计算方法基于“联系强度” (Kirk, 1996)。这种方法的一个共通点是可以估计出总变异中有多大比例的变异来自自变量对因变量的效应, 因此, 它被广泛使用。最常用的联系强度指标是联系性测量 η^2 , 它可以利用方差分析表 (表 13.3) 中的信息非常容易地计算出来 (多数计算机程序中提供 η^2 结果作为效应大小指标)。 η^2 的定义如下:

$$\frac{\text{组间平方和}}{\text{总平方和}}$$

在本例中 (见表 13.3):

$$\eta^2 = \frac{54.55}{[(54.55) + (37.20)]} = 0.59$$

当没有方差分析表时, 也可以直接从 F 值中计算 η^2 (见 Rosenthal & Rosnow, 1991, p.441):

$$\eta^2 = \frac{(F)(df_{\text{效应}})}{[(F)(df_{\text{效应}})] + (df_{\text{误差}})}$$

在本例中:

$$\eta^2 = \frac{(7.80)(3)}{[(7.80)(3)] + 16} = 0.59$$

科恩发明了另一种效应大小指标 f , 它是针对三个及以上独立组实验设计的 (Cohen, 1988), 它和 d 一样, 也是一个标准化的效应值, f 值对估计两组实验设计的效应大小也很有用。但 d 是根据两组平均数的差值计算而来, 科恩 f 值 (Cohen's f) 则是通过不同组的平均数的离散程度计算的。 d 和 f 所表示的效应都是相对于总体内标准差而言的 (标准化)。科恩提供了 f 值的解释, 他认为, 小、中等和大的效应值 f 分别对应于 0.10、0.25 和 0.40。 f 值很难从 ANOVA 表 (表 13.3) 中计算出来, 但可以通过 η^2 来计算出 f (Cohen, 1988)。公式如下:

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}}$$

在本例中:

$$f = \sqrt{\frac{0.59}{1-0.59}} = 1.20$$

我们可以得出以下结论:记忆训练可以解释因变量中的59%,标准化的效应大小 f 值为1.20。根据科恩解释 f 值的原则(0.10、0.25、0.40),记忆训练在回忆分数上的效应很大。

独立组设计的统计效力估计

获得效应大小后,我们就可以对特定大小样本和自由度的统计效力进行估计。在本例中,我们设定显著性水平为0.05,每组有5名被试,组间效应自由度 $df=3$ 。效应值非常大, $f=1.20$ 。实验效应在统计上显著且效应值如此大的情况下,效力分析并没有多少必要。因此,让我们看另一个不同的结果。

假设ANOVA不显著,效应值 $f=0.4$,按照科恩的原则,这仍然是一个大的效应。这里要回答的一个重要问题是:实验的统计效力如何?在显著性水平为0.05、样本大小为5、 $df=3$ 时,效应出现的可能性有多大?从附录的效力表中我们发现,在这种情况下,统计效力为0.26。换言之,在这种情况下,获得统计显著性的概率仅为0.26。即仅有四分之一次尝试能够达到统计显著性结果。显然,实验的效力比较低,因此,我们就无法强调NHST所得到的不显著的结果。如果这样做,我们可能就忽略一个重要的事实,即自变量效应实际上是很大的。

尽管事后分析效力很重要,尤其在NHST检验不显著的情况下。但为了得到从一开始就预设的一个效应统计显著的概率,我们最好还是实验实施之前进行统计效力分析。如果一个研究者实验前知道研究的统计效力仅为0.26,或显著效应不出现的可能性是0.74,那么,就没有必要浪费时间和资源继续进行实验了。因此,在实验实施之前,研究者通过文献检索就可以发现,以往这方面的研究都能够获得多大的效应。由于可以通过增加样本容量提高统计效力,如果研究者希望达到的统计效力是0.80,那么,就可以算出达到0.80效力时所需的样本大小。

从效力表中可以找到当效应值为0.40、 $df=3$ 时,达到0.80统计效力所需要的样本容量。在附录的表A.5中可以查出,要达到0.80的统计效力,每组需要18名被试。研究者在实验前应当考虑这些因素。

多组实验设计的均数比较

单因素多组实验设计并不仅仅关心“存在某种效应”。在研究过程中,我们通常会有一个非常具体的假设,而不是笼统地说“这个变量在其他变量上存在效应”。 F 检验和效应大小都不能说明哪些平均数显著地不同于其他一些平均数。例如,在记忆实验例子中,我们无法确定“想像组”的平均数是否与“编故事组”的平均数之间有显著差异。 F 检验的结果只表明,所有组间的变异大于误差的变异。

我们可以通过另外两个方法了解单因素多组实验设计。一种方法是通过计算95%的置信区间检验总体平均数的模式。在第12章中,当我们讨论如何用置信区间比较多组实验设计时已经介绍过。置信区间可以用来确定从实验组平均数估算出来的总体平均数之间的差异。具体做法是检验置信区间是否重叠:如果重叠,

那么,重叠程度有多大(见专栏 12.5)。注意,置信区间的宽度说明估计的精确性。

第二种方法是使用 NHST 进行两组比较,以确定自变量的总体效应。研究者通过两均数比较(comparison of two means)可以检验特定的差异。这种比较可能非常复杂,例如,将实验中的两组或多组的平均数与其他组的平均数或多个组的平均数进行比较。然而,多数情况下我们感兴趣的是两个单组平均数之间的比较。这种比较通常是在确定 F 检验显著之后进行。

两均数比较的一种方法是 t 检验。但在多组实验设计中进行两组比较时,检验方法就需要进行一些修改。我们将使用 F 检验中的组间变异($MS_{\text{误差}}$)计算出合并方差。即从所有实验组数据中计算方差,而不是仅从两组数据中计算。这样的 t 检验公式如下:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{[MS_{\text{误差}}] \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}}$$

$MS_{\text{误差}}$ 值可以从 ANOVA 表中找到, t 检验的自由度与 $MS_{\text{误差}}$ (或 $k(n-1)$, k 是组数)有关。例如,表 13.3 中的组内 MS (误差)为 2.33,自由度为 16 ($4(5-1)=16$)。

在记忆训练实验中,我们要进行的就是记忆训练组的记忆成绩和控制组记忆成绩之间的比较。三个记忆训练组记忆保持的平均成绩是 13.67 ($n=15$),控制组是 10.00 ($n=5$)。我们的问题是,在不考虑记忆训练类型(编故事、想像和押韵)的情况下,训练比不训练的 memory 成绩好吗?虚无假设是这两个总体的平均数之间没有差异(样本平均数之间的差异是单纯的随机因素造成的)。把相应的值代入上述 t 检验公式中,我们就可以得到一个统计显著性效应, $t(16)=4.66$, $p=0.0003$ 。² 因此,在这个实验中,如果不考虑训练类型,训练后记忆成绩好于没有训练的记忆成绩。这个结论要比 F 检验更加具体,在 F 检验中,我们只能说实验的 4 种条件产生的变异大于随机条件下的变异。下一节中我们将讨论这些平均数的置信区间,它将帮助我们确定三种记忆训练条件下成绩的两两比较是否有意义。

两均数比较的科恩 d 值也可以用 t 检验的结果计算。计算公式如下:

$$d = \frac{2(t)}{\sqrt{df_{\text{误差}}}}$$

将记忆训练组和控制组 t 检验 ($t=4.66$, $df=16$) 数据带入上述公式,则 $d=2.33$ 。按照科恩的效应大小评判标准,训练后的记忆比没有训练的记忆产生了很大效果。

² 两个均数的 t 检验可以从 F 检验结果中的数据用上述公式进行手算。与 t 值有关的概率 (p 值)可以通过两种方法来确定。(1) 将计算的 t 值与附录表 A.2 中的临界值比较;(2) 用计算机程序计算,在其中输入 df 和 t 值便可以得到精确的单尾或双尾的概率值(见 <http://math.uc.edu/~brycw/classes/148/tables.htm>)。

本节中我们讨论了通过 NHST 和置信区间进行多重实验组的平均数之间的比较。由于每种方法都包含了独特的信息,我们可以对在这种条件下 NHST 和置信区间的使用加以总结。置信区间让我们可以判断所有实验条件下总体平均数的模式。置信区间宽度说明总体平均数估计的精确性。在使用 t 检验时,我们可以通过概率大小 ($p=0.001$) 决定是否拒绝虚无假设。如前所述,具体的概率值在解释结果时非常重要 (Posavac, 2002)。概率越低,重复实验时拒绝虚无假设的可能性越高 ($p<0.05$) (Zechmeister & Posavac, 2003)。在报告统计显著性结果时,我们一般报告最小的概率值 (计算机计算出来的结果中会自动给出精确的概率值)。

用 t 检验进行均数比较的结果也可以和以往研究结果进行对照。比较方法有两种:首先,看当前研究结果的显著性是否与以往研究相似,即是否重复了统计显著性结果。其次,计算与以往研究对应的两个均数比较的效应值 (科恩 d 值),这类似于元分析。上述两种比较都无法使用置信区间。与 NHST 不同,置信区间不能提供差异显著性的概率,效应大小可以在 t 检验后直接计算出来 (见第 12 章)。

总之,在数据分析中应该尽量使用不同统计方法,以便尽可能多地从数据中获得信息。

单因素独立组实验设计的结果报告

现在我们可以按照 APA《出版手册》(2001) (20~26 页) 报告数据分析结果。报告包括置信区间和 NHST 两部分。记忆训练效果研究结果报告如下:

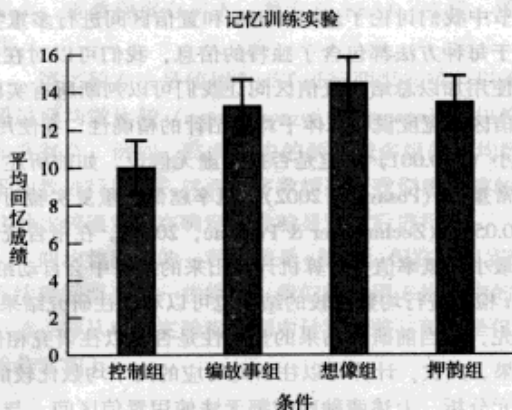
本研究探讨了记忆训练对词汇记忆效果的影响,计算了四种指导语条件下的平均回忆成绩 (共有 20 个词)。四种条件分别是用指定的词汇编故事、形成想像、押韵的方法记忆和控制组 (没有任何要求)。每组有 5 名被试 ($N=20$)。每组成绩的平均数和标准差 (括号内) 分别为编故事组 13.2 (1.3)、想像组 14.4 (1.8)、押韵组 13.4 (1.3) 以及控制组 10.0 (1.6)。四组记忆成绩之间存在显著差异, $F(3, 16) = 7.80, p=0.002, \eta^2=0.59$ 。记忆训练组与控制组的平均成绩之间进行比较:记忆训练组的平均保持成绩 (13.7) 显著高于控制组, $t(16) = 4.66, p<0.01$ 。两者比较的效应大小 d 值为 2.33, 这说明,相对于无训练组,记忆训练存在很大的效应。每组的 95% 置信区间见图 13.1。控制组的置信区间与记忆训练组的置信区间没有重叠。然而,3 种记忆训练条件下的置信区间却存在重叠 (包括样本平均数),这表明它们的总体平均数之间不会有差异。这些结果显示,无论哪种记忆训练都比没有训练明显地提高了记忆成绩。

注 释 在报告的开头,叙述了实验目的、被试人数 (20)、自变量水平数、自变量水平如何定义 (名称) 以及每个自变量水平中的测试人数 (5)。然后,报告每组的平均数的描述统计结果。置信区间宽度说明了对各组的总体平均数的估计的精确性以及总体平均数的模式。建立置信区间的步骤见第 12 章。由于 ANOVA 表中的 $MS_{\text{误差}}$ 的平方根等于 $s_{\text{合并}}$, 因此,95% 的置信区间可以定义为:

$$95\%CI = \bar{X} \pm [\sqrt{(MS_{\text{误差}}/n)}](t_{\text{临界}})$$

其中 $t_{\text{临界}}$ 是在 $MS_{\text{误差}}$ 自由度时的 t 检验的临界值。

图 13.1 记忆训练实验的平均数和 95%置信区间



在上述例子中, $MS_{误差}$ 自由度为 16 (见 ANOVA 表), 0.05 水平 (双尾) 上 t 的临界值为 2.12。因此:

$$95\%CI = \bar{X} \pm [\sqrt{(2.32/5)}](2.12) = \bar{X} \pm \sqrt{(0.464)}(2.12) \\ = \bar{X} \pm (0.68)(2.12) = \bar{X} \pm 1.44$$

如图 13.1 所示, 3 种记忆训练条件下的置信区间有重叠, 而它们与控制组的置信区间没有重叠。 F 检验表中包括精确的概率和效应大小测量 (η^2 值)。所有这些信息都无法从置信区间计算中获得。这些信息在记忆训练组和控制组的两组均数比较中也有所反映。通常情况下, 我们只选择一种效应大小测量, 如 d 值, 当然, 也有人喜欢用 η^2 。最后, 用一句话为读者总结实验结果。

在有些情况下, 也需要进行两组均数比较。例如, 将每组或多组的平均数和控制组的平均数进行比较, 但是, 在上述例子中, 从图 13.1 中置信区间可以看出, 三个训练组之间存在重叠, 而它们与控制组之间没有重叠。因此, 每个训练组可能和控制组之间存在差异, 而三个记忆训练组之间则没有显著差异 (见第 12 章)。

重复测量的方差分析

- 重复测量的方差分析的过程及虚无假设检验与独立组设计的方差分析相似。
- 在进行重复测量方差分析之前, 必须计算每个被试在每种条件下的平均分数 (平均数、中数等)。
- 计算所有被试在所有自变量水平上的描述统计数据。
- 重复测量方差分析的主要不同点在于对误差变异或残差变异的估计。残差变异是指把由自变量和被试引起的系统变异从总变异估计中剔除后的剩余部分。

重复测量实验设计的数据分析与独立组实验设计的分析方法相同。运用 NHST

原理确定实验所引起的差异是否大于仅由误差变异而引起的差异。在数据分析的开始阶段,通过方差分析确定自变量是否在不同水平之间产生了系统变异。如果统计结果显著,那么,就可以运用置信区间和两组均数比较确定系统变异的具体来源,即哪些水平之间存在差异。我们已经介绍了独立组实验设计的方差分析过程,本节将重点讨论重复测量实验设计的分析过程。我们将以第8章中的时间知觉实验为例。

数据整理 在重复测量的实验设计中,每个被试接受所有的实验处理条件。在一个完整的设计中,将对每个被试进行多次测量;而在不完整的设计中,每个被试只接受某一种处理条件。在第8章的实验中,每个被试对4种时间间隔(12、24、36和48秒)进行估计,这是一个不完全的重复测量实验设计。例如,在一次试验中,被试只是随机接受一种时间间隔(如36秒),然后估计这段时间的长度。

数据分析的第一步是建立一个数据矩阵,用来描述被试在不同实验条件下的平均分数。在计算这些结果时,应考虑如何平衡不同条件下的练习效应。在时间估计实验中,被试对每种时间间隔进行6次估计。因此,被试在整个实验中要进行24次时间估计。在这里使用被试在4种条件下的时间估计的中数进行数据分析,通常情况下应该使用每种条件下的平均数,但平均数容易受到极端分数的影响。被试有时(注意的影响)在估计时间间隔时会出现一些极端结果。5名被试的时间估计中数见表13.4的上半部分。例如,第1个被试对12秒间隔的6次估计成绩是13秒。下一步是计算不同被试在各个条件下的描述统计结果。每种条件下的均数和标准差(括号内)见表13.4。

分析的重点是被试能否区分开不同长度的时间间隔。也许你已经发现,只有知道表13.4中的均数差异大于由误差变异引起的差异时,我们才能确定被试是否

表13.4 重复测量实验设计的数据及方差分析结果

被 试	数据矩阵			
	间隔长度			
	12	24	36	48
1	13	21	30	38
2	10	15	38	35
3	12	23	31	32
4	12	15	22	32
5	16	36	69	60
均数 (SD)	12.6 (2.0)	22.0 (7.7)	38 (16.3)	39.4 (10.5)

注:表中的数是每种条件下被试6次反应的中数

变异来源	df	SS	MS	F	p
被试	4	1553.5	—	—	—
间隔长度	3	2515.6	838.5	15.6	0.000
残差(误差变异)	12	646.9	53.9		
总体	19	4716.0			

有能力将不同长度的时间区分开。也就是说,即使在比较不同条件下的均数时发现它们之间存在差异,我们也无法知道被试的时间判断结果与随机因素引起的差异是否相同。表 13.4 中数据分析的虚无假设为每种间隔时间的总体均数相同。为了对虚无假设进行 F 检验,我们需要对误差变异和系统变异(F 检验的分子)进行估计。不同被试在四种条件下的变异给我们的计算提供了信息。尽管 5 名被试对每种间隔时间进行了多次判断,但他们的估计均数不可能相同。因此,不同时间间隔下的均数也反映了误差变异。如果不同时间间隔长度系统地影响到被试判断,那么不同时间长度的估计均数就反映了系统变异。为了完成 F 检验,我们还需要对误差变异进行估计(F 值的分母)。见专栏 13.4。

重复测量方差分析的误差变异是实验条件对被试造成的差异,因此,对它的估计是必不可少的。首先,我们要计算实验的总变异。然后,从中减去两个潜在的变异源,即自变量和被试,剩余部分是残差,它就是我们误差变异的估计。在随机组设计的方差分析中,我们用组内变异作为误差变异的估计。残差变异就是重复测量 F 检验的分母(对误差变异的估计)。

ANOVA 表的解释 方差分析结果见表 13.4 中的下半部分。重复测量的方差分析可以通过计算机软件完成。我们的重点在于如何解释方差分析表中的数据,而不在于这些结果是如何计算出来的。表 13.4 列出了单因素重复测量设计中四个变异来源的结果。表中自下而上的四个变异源分别是(1)总体变异,(2)残差变异,(3)时间间隔长度引起的变异(自变量),(4)被试引起的变异。

在所有的方差分析表中,最重要的部分是反映自变量效应的 F 值和 F 检验中假定虚无假设成立的概率。在表 13.4 中重要的 F 检验是时间间隔长度。 F 值的分子是时间间隔长度的均方(MS),分母是残差的 MS 。因为有 4 种时间间隔,所以,分子的自由度为 3。残差变异的 df 为 12。它是总自由度减去被试自由度和时间间隔长度自由度($19-4-3=12$), F 值为 15.6,所对应的虚无假设成立的概率为 0.000,小于显著性水平 0.05。因此,我们拒绝虚无假设,并得出时间间隔长度是系统变异来源的结论,即被试对时间的估计受到时间间隔长度的影响。

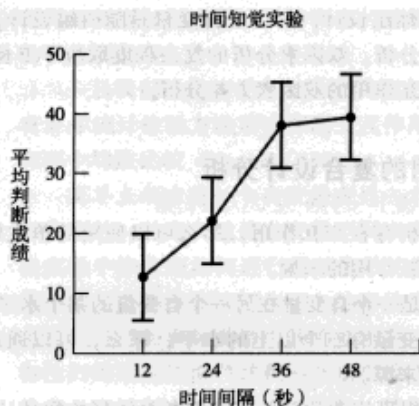
专栏 13.4

重复测量设计的误差与敏感性

重复测量设计方差分析的一个区别性特征是误差变异的估计。在随机组实验设计中,被试间差异在组间得到平衡,误差变异的估计作为计算 F 值的分母。由于这种设计中,被试只接受一种处理条件,被试之间的差异不能被完全剔除,只能被平衡。而在重复测量设计中,被试间存在系统变异,有的被试在某些条件下的成绩好于其他条件,而有的被试则相反。然而,由于每个被试

参与了所有实验条件,因此,被试之间的差异对每种条件下平均成绩的影响是等同的。所以,重复测量设计中不同条件下平均成绩之间的差异并不是被试间的系统差异。然而,在重复测量设计中,被试间的差异不是被平衡,而是在分析中被消除了。重复测量设计具有消除系统变异的能力,它比随机组设计更具敏感性。

图 13.2 时间知觉实验的均数及其 95% 置信区间



建立置信区间的方法与独立组实验设计相同。可以用方差分析表中的 $MS_{\text{误差}}$ 建立置信区间 (Loftus & Masson, 1994)。

$$95\%CI = \bar{X} \pm [\sqrt{(MS_{\text{误差}}/n)}](t_{\text{临界}})$$

其中 $t_{\text{临界}}$ 是在 $MS_{\text{误差}}$ (残差) 自由度时的 t 检验的临界值。对重复测量设计的解释也与独立组实验设计的解释相同。

效应大小测量 测量效应大小对数据分析也很重要。常用的效应大小指标是表示变量间联系强度的 η^2 。它是通过将残差平方和除以残差平方和与误差平方和之和。在我们的例子中:

$$\eta^2 = \frac{SS_{\text{效应}}}{SS_{\text{效应}} + SS_{\text{误差}}} = \frac{2515.6}{2515.6 + 646.9} = 0.795$$

它说明了自变量效应占总变异的比例。

有些情况下, 方差分析后还要进行两个平均数之间的比较 (例如, 比较两种相邻时间间隔的平均估计结果), 以便更加精确地表示出随时间间隔长度增加, 时间估计发生变化的特点。这里的两个平均数之间的比较过程与随机组设计的相似。是否进行两个均数比较取决于研究假设和置信区间 (见 Keppel, 1991)。

独立组设计的双因素方差分析

独立组设计的双因素方差分析主要用于实验中的两个自变量都含有两个或两个以上水平。在第 9 章中已对含有两个自变量的实验设计方法和理论基础进行了介绍。其中介绍了主效应与交互作用的概念。本章中我们将重点介绍因素设计的计算机辅助分析过程, 其中包括 A、B 的主效应以及 A×B 交互作用。在如下这些实验中: 两个自变量被操纵的一个随机组设计; 两个自变量的自然组设计; 或者

一个自变量是自然组设计, 另一个自变量是随机组设计, 上述三种情况都可使用独立组的双因素分析。双因素分析的复杂程度取决于 F 检验中是否存在交互作用。我们先介绍有交互作用的双因素方差分析。

含有交互作用的复合设计分析

- 如果方差分析存在交互作用, 那么可以使用简单主效应分析和两个均数的比较来确定交互作用的来源。
- 简单主效应是一个自变量在另一个自变量的某个水平上的效应。
- 如果一个自变量有三个以上的水平, 那么, 可以通过两个均数的比较确定简单主效应的来源。
- 可以用不同组平均数的置信区间判断总体平均数估计的精确性。

我们以第 9 章中罗德曼和伯格 (Rodman & Burger, 1985) 关于大学生阅读汽车事故材料的实验为例。其中一个自变量是事故的严重性 (严重和不严重), 学生被随机指派阅读两种材料中的一种。第二个自变量采用自然组设计。学生按照抑郁的纸笔测验分数分为无抑郁组、轻度抑郁组和中度抑郁组。这是一个 2×3 的实验设计。

在第 9 章中我们介绍了罗德曼和伯格所研究的现象叫防御性归因效应, 归因是对自己或他人行为的原因的解释。在他们的实验中, 当无抑郁被试阅读了有人在交通事故中受到了中度或重度伤害时, 他们在严重后果的条件下比在中等后果条件下更多地把责任归结到行人身上。这说明无抑郁被试不想把事故原因归因为偶然, 这种归因方式增加了他们对自己遇到严重交通事故知觉的可能性。将严重交通事故主要原因归因于行人的现象叫防御性归因。无抑郁被试用自己将来可能遇到严重交通事故的想法进行自我防御或保护。由于抑郁个体倾向于负性思维, 所以, 罗德曼和伯格推测, 他们不大可能出现防御性归因。即他们不会用自己也会遇到严重的交通事故的想法进行自我防御。

表 13.5 中是 6 种条件下的平均百分比。在第 9 章中我们已经对这些数据做过简单分析。在含有两个自变量的复合设计方差分析表中有 4 个变异源。它们是每个自变量的主效应、两个自变量之间的交互作用以及组内误差。例如, 在罗德曼和伯格实验中, 有事故类型和抑郁水平两个主效应。而在该实验中, 研究者感兴趣的是这两个因素间的交互作用。ANOVA 结果显示, 两者之间的交互作用显著。在表 13.5 中可以看出交互作用的存在, 因为严重与不严重的百分比之间的差异随

表 13.5 不可控情景下事故责任归因的平均百分比

事故类型	抑郁水平		
	无抑郁	轻度抑郁	中度抑郁
严重	7.00 (9.2)	14.00 (16.1)	16.90 (16.0)
不严重	30.50 (22.2)	16.50 (12.7)	3.75 (3.5)

注: 括号内是标准差。引自 Rodman & Burger(1985)。

不同抑郁增加而改变。具体而言,无抑郁学生表现出防御归因效应,而中等抑郁的学生则没有。

一旦我们确定两个自变量之间存在交互作用,我们就必须确定它的确切来源。有许多统计检验方法可以探测交互作用效应的来源。这些方法是简单主效应分析和两个均数比较(Keppel, 1991)。

简单主效应是一个自变量在另一个自变量的一个水平上的效应。事实上,交互作用的一种定义就是自变量不同水平上的效应不同。我们可以回到罗德曼和伯格实验中来介绍如何进行简单主效应分析。在表13.5中有5种简单主效应分析。其中的三个是不同抑郁水平上的事故类型(严重、不严重)效应,另外两个简单主效应是两种事故类型中的抑郁水平(无抑郁、轻度抑郁和中等抑郁)效应。防御性归因效应是指严重和不严重事故之间的差异,罗德曼和伯格预测,事故类型效应会随抑郁水平的增加而降低。罗德曼和伯格为了验证他们的预测,分析了不同抑郁水平的事故类型的简单主效应。正如他们所预测的那样,无抑郁学生的事故类型效应在统计上显著。即无抑郁学生对不可控情景中严重和不严重事故的归因是不同的。而轻度和中度抑郁学生的事故类型简单主效应都不显著。因此,对于这两组学生而言,他们对严重和不严重事故的归因没有差异。

罗德曼和伯格(1985)研究中的两个简单主效应与三个平均数有关。简单主效应可以检验无抑郁学生、轻度抑郁学生和中度抑郁学生对严重事故的归因的差异以及三组学生对不严重事故的归因差异。如果简单主效应显著,那么这些平均数(三组学生对严重事故的归因)之间差异显著。下一步将通过两组平均数的比较对简单主效应进行更加全面地分析(Keppel, 1991)。当显著的简单主效应中涉及两个以上变量水平时,为了确定这些水平之间的差异,就需要进行两个平均数之间的差异检验。在此过程中,同时对简单主效应中的平均数进行两两比较,从而寻找不同水平之间的差异。只有当三个以上水平的自变量存在简单主效应时,两个平均数之间的比较才有意义。如果只有两个水平,那么,简单主效应中已经包含了两个平均数之间的比较。因此,无须再做两个平均数间的比较。

分析了交互作用之后,研究者还可以分析自变量的主效应。但是,通常情况下,如果交互作用显著,主效应就没有什么意义了。例如,在罗德曼和伯格(1985)的实验中,被试对不同类型事故的归因取决于他们的抑郁水平。通过主效应分析,我们并不能更多地了解被试为什么将不严重事故责任归因于不可控环境的比例高于把严重事故做同样归因的比例。当然,在有些实验中,交互作用和主效应都很重要。

无交互作用的方差分析

- 如果方差分析中的交互作用不显著,那么下一步就是去确定主效应是否显著。
- 主效应显著性的来源可以通过均数的两两比较和构建置信区间来确定。

当交互作用不显著时,就需要确定每个自变量的主效应是否显著。如果自变量的主效应不显著,就不需要做进一步分析了。但是,如果主效应显著,那么还可以做很多分析。例如,如果自变量有三个以上的水平,那么,主效应显著性来

源可以通过不同水平两两均数之间的比较来确定。(当然,如果只有两个水平,那么,就不需要进一步分析了。)另一种方法是建立各组平均数的置信区间。这与单因素独立组设计的分析类似。复合设计的不同之处在于一个自变量的数据在另一个自变量的不同水平上分开计算。

双因素独立组设计的效应大小计算

通常用 η^2 (或方差所占比例) 作为复合设计 ANOVA 的效应大小指标,我们在单因素设计中已经讨论过。在计算 η^2 时,我们只去计算那些产生影响的效应。 η^2 的定义如下:

$$\eta^2 = \frac{SS_{\text{产生影响的效应}}}{SS_{\text{产生影响的效应}} + SS_{\text{组内}}} \quad (\text{见 Rosenthal \& Rosnow, 1991})$$

因此,在 $A \times B$ 设计中,三个效应都能够计算出 η^2 。

如前所述 (Rosenthal & Rosnow, 1991), 如果没有效应的平方和,那么, η^2 的计算公式为:

$$\eta^2 = \frac{(F)(df_{\text{效应}})}{(F)(df_{\text{效应}}) + (df_{\text{误差}})}$$

复合设计分析中置信区间的作用

对复合设计的分析还可以建立置信区间。例如,在 2×3 设计中可以将每个平均数的置信区间列出来。置信区间构建的方法和过程见第 12 章及本章的开头部分。公式为:

$$95\% \text{ 置信区间的上限: } \bar{X} + [t_{.05}][s_{\bar{X}}]$$

$$95\% \text{ 置信区间的下限: } \bar{X} - [t_{.05}][s_{\bar{X}}]$$

当样本大小相等时,标准误为:

$$s_{\bar{X}} = \frac{s_{\text{合并}}}{\sqrt{n}} \quad \text{其中 } n = \text{各组样本的大小。}$$

因为 ANOVA 中的 $MS_{\text{误差}}$ 的平方根等于 $s_{\text{合并}}$, 所以,我们就可以用下列公式计算 95% 的置信区间:

$$95\% CI = \bar{X} \pm (t_{.05})[\sqrt{MS_{\text{误差}}}/\sqrt{n}]$$

其中 $t_{.05}$ 值由与 $MS_{\text{误差}}$ 有关的自由度决定。

我们将以一个 2×3 独立组设计的结果为例,说明复合设计分析中如何使用置信

信区间。假定在实验中要求被试用优势手或非优势手完成一项运动操作任务。被试被随机分成三组,其中一组完成两次间隔 60 秒的操作任务,另一组完成间隔 30 秒的任务,还有一组完成间隔为 0 秒的任务。因变量是操作中的正确反应数目。因素 A 是手的优势性,包括两个水平(优势手和非优势手);因素 B 是两次操作之间的时间间隔,有 3 个水平(0、30 和 60 秒)。下面列出了 6 种实验条件下的反应平均数。每组有 5 人 ($n=5$)。将数据用图画出来更有助于我们理解交互作用的特性。 F 检验表明,交互作用是显著的, $F(2, 24) = 12.34, p \leq 0.0005$ 。(两个主效应也显著)

		延迟 (B)		
		0 秒 (b_1)	30 秒 (b_2)	60 秒 (b_3)
手 (A)	优势 (a_1)	19.0	19.0	20.0
	非优势 (a_2)	10.6	15.8	18.2

这个实验的 $MS_{\text{误差}}$ 是 2.45 ($df=24$)。所以,平均数的合并标准误为:

$$(\sqrt{MS_{\text{误差}}}/\sqrt{n}) = (\sqrt{2.45}/\sqrt{5}) = \sqrt{0.49} = 0.70$$

查附录表 A.2 得, df 为 24 时, t 的临界值为 2.06。所以,实验中 6 组的置信区间分别是:

$$a_1b_1 = \bar{X} \pm (2.06)(0.70) = 19.0 \pm 1.44 = 17.56-20.44$$

$$a_1b_2 = \bar{X} \pm (2.06)(0.70) = 19.0 \pm 1.44 = 17.56-20.44$$

$$a_1b_3 = \bar{X} \pm (2.06)(0.70) = 20.0 \pm 1.44 = 18.56-21.44$$

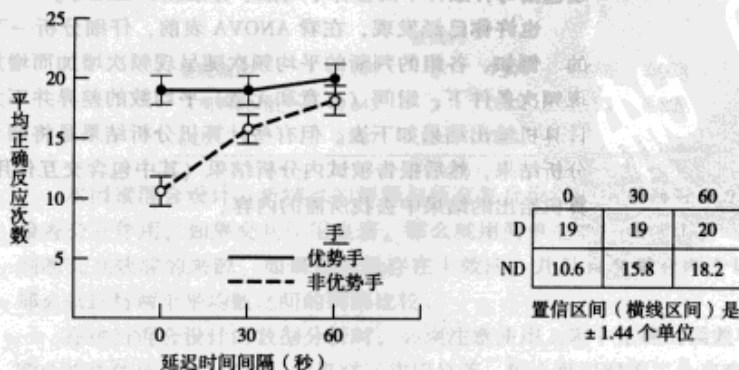
$$a_2b_1 = \bar{X} \pm (2.06)(0.70) = 10.6 \pm 1.44 = 9.16-12.04$$

$$a_2b_2 = \bar{X} \pm (2.06)(0.70) = 15.8 \pm 1.44 = 14.36-17.24$$

$$a_2b_3 = \bar{X} \pm (2.06)(0.70) = 18.2 \pm 1.44 = 16.76-19.64$$

图 13.3 是假定的实验中 6 个平均数的置信区间。置信区间说明了对总体平均数估计的准确性。我们要检验置信区间的宽度,并且通过检查样本平均数的置信区间是否重叠以及重叠的程度来确定总体平均数的模式。对置信区间的经验性解

图 13.3 不同时间间隔(秒)和不同手(优势手与非优势手)完成运动任务时的正确反应平均数,其中也包含了平均数的 95% 的置信区间



释方法是,如果两个置信区间没有重叠,那么,用 NHST 检验这两个平均数的结果可能就是显著的。(见第 12 章中的专栏 12.5)

在 95% 的置信区间上,交互作用表明优势手变量产生的效应随间隔时间的变化而改变。在间隔 0 秒和 30 秒的情况下,优势手都比非优势手的正确反应次数明显地多。这种差异在间隔 0 秒时最大;间隔 30 秒时次之;而在 60 秒时可能就没有差异(即置信区间重叠)。要下一个确定的结论,还需要其他数据。

解释交互作用的第二种方法是给出优势手变量的每一个水平上的延迟(时间间隔)效应。在优势手条件下,95% 的置信区间在 3 个时间间隔水平上都是重叠的。三个样本的平均数都被包含在代表三个延迟水平的置信区间内。因此,我们有理由推论,时间间隔变量并不影响优势手被试完成运动操作任务。相反,非优势手被试的成绩随时间间隔的不同而不同。95% 的置信区间表明,与间隔 30 秒和 60 秒相比,间隔 0 秒的反应正确性很低,而间隔 30 秒和 60 秒之间的差异并不大。(根据此结果尚不能做出结论。)

双因素混合设计的方差分析

当一个自变量是随机组或自然组,而第二个变量是重复测量的变量时,就应该采用双因素混合设计的方差分析。上述第一个自变量是被试间因素,用 A 表示;第二个自变量是被试内因素,用 B 表示。双因素混合设计分析是单因素独立组设计和重复测量设计分析的混合体。在第 9 章我们没有介绍这种混合设计的分析方法,在这里我们将深入讨论。

下表中的数据是被试对 4 段流行歌曲出现频次的平均判断结果。被试听一盘含有许多歌曲的磁带。其中一半被试不知道将要进行测试,称为无意组;另一半被试知道将要进行测试,称为有意组。所有被试听的歌曲中,有的播 1 遍,有的播 2 遍,有的播 3 遍。因此,实验是一个 2×3 设计,被试分成两组,每组 5 名被试;两组的指导语不同。歌曲播放次数是被试内变量,包括播放 1 遍、2 遍和 3 遍三个水平。下列数据是两组被试在 3 种播放次数条件下的平均判断次数;此外,还包括每种条件下的整体平均数、标准差和全距等。

也许你已经发现,在看 ANOVA 表前,仔细分析一下数据的趋势是非常重要的。例如,各组的判断的平均频次随呈现频次增加而增加。有趣的是,在每种呈现频次条件下,组间(有意和无意)平均数的差异并不大。双因素混合设计的计算机输出结果如下表。但有些计算机分析结果是将两个因素分开,先报告组间分析结果,然后报告被试内分析结果(其中包含交互作用效应)。因此,需要在计算机给出的结果中去找所需的内容。

组别	1 遍	2 遍	3 遍
有意组	2.81	3.24	3.91
无意组	2.81	3.24	3.91

组别 (A)	被试	播放次数 (B)		
		1 (b_1)	2 (b_2)	3 (b_3)
无意图组 (a_1)	1	1.2	2.2	3.0
	2	0.8	2.0	3.2
	3	1.2	1.8	2.8
	4	1.5	2.0	2.5
	5	1.2	2.5	3.2
	M	1.18	2.1	2.94
	SD	0.25	0.26	0.30
	全距	0.8-1.5	1.8-2.5	2.5-3.2
有意组 (a_2)	6	1.2	1.5	2.0
	7	1.0	2.0	2.5
	8	1.0	1.2	3.2
	9	1.2	2.2	2.8
	10	1.0	2.2	3.5
	M	1.08	1.82	2.80
	SD	0.11	0.45	0.59
	全距	1.0-1.2	1.2-2.2	2.0-3.5

上表包括两部分。“被试间”部分包括主效应的 F 值。这部分内容类似于单因素独立组设计的分析结果。其中的误差项是组内变异。概率值为 0.226, 大于显著性水平 0.05。所以, 该组效应在统计上不显著。第 2 部分是“被试内”部分。它包括被试内变量呈现频次以及呈现频次与组别的交互作用。通常情况下, 包含被试内变量(主效应或交互作用)的效应必须通过被试内设计中的残差进行检验。交互作用的 F 值小于 1, 因此交互效应不显著, 而呈现频次的主效应却十分显著。(在有些单因素被试内设计的计算机分析软件中, 计算结果可能比上表中的信息更多。)

被试间					
来源	SS	df	MS	F	p
组别	0.225	1	0.225	1.718	0.226
误差	1.049	8	0.131		
被试内					
呈现频次	15.149	2	7.574	58.640	0.000
呈现频次 $\times$ 组别	0.045	2	0.022	0.173	0.843
误差	2.067	16	0.129		

双因素混合设计分析结果的解释与所有复合设计的结果解释方法相同。即先检查交互作用, 如果交互作用显著, 那么就用简单主效应或两个平均数的比较来判断交互效应的来源。如果自变量存在主效应, 并且自变量有两个以上的水平, 那么就进行两个平均数之间的两两比较。

在进行混合设计的数据分析时, 必须注意使用上表中正确的误差项(主效应、两个均数的比较)。例如, 如果交互作用显著, 那么可以对第二个自变量的每个水

平上进行单因素方差分析,从而进行简单主效应分析。在上例中,如果组和呈现频次的交互作用显著,有意组的主效应可以通过该组的重复测量 ANOVA 进行检验(详见 Keppel, 1991)。

通常用 η^2 作为混合设计的效应大小估计,即自变量效应占总变异比例的估计。如上所述, η^2 是效应的 SS 除以效应的 SS 与误差的 SS 之和。

复合设计的结果报告

复合设计的结果报告与单因素 ANOVA 结果形式类似,但如果存在交互作用就有所不同了。下面是复合设计结果中应该包含的主要内容:

- 描述每个变量及其水平(条件);
- 在用文字或图表对每种条件下的统计结果进行描述时,应该包括置信区间;
- 给出主效应和交互作用的 F 值、 p 值;
- 给出每个效应的效应大小指标;
- 给出不显著效应的统计效力;
- 当交互作用显著时,要进行简单主效应分析;
- 在口头报告显著的交互作用时,应该给出自变量不同水平的平均数之间的差异;
- 在口头报告显著的主效应时,应该给出自变量不同水平的合并的平均数之间的差异;
- 通过比较两个平均数,确定引起主效应的系统变异源;
- 从分析结果中得出研究结论。

参见第 14 章中关于结果报告的内容。

小 结

虚无假设显著性检验(NHST)方法通常用于心理学实证研究的数据分析。NHST 主要是用来确定实验中的自变量所产生的差异是否大于仅有误差时所产生的差异。显著的统计结果是指,虚无假设为真的概率很小。在使用 NHST 时存在两种错误:当虚无假设为真时却被拒绝了,此时就会犯 I 类错误;I 类错误的概率就是 α 或显著性水平,通常为 0.05。而 II 类错误是虚无假设为假时却没有被拒绝,当研究没有足够大的统计效力而正确拒绝虚无假设时就会犯 II 类错误。增加统计效力的主要方法是增加样本容量。在研究实施之前,研究者可以通过统计效力表确定拒绝虚无假设所需的统计效力的大小,研究实施之后就可以从数据中获得探测效应。研究结果中应该给出与统计结果有关的精确的概率。

t 检验适合于比较两个平均数之间的差异。在对两个平均数进行差异检验的时候, 也应当给出效应大小指标, 如科恩 d 值。在报告 NHST 结果时, 必须注意: 统计显著性(或置信区间没有重叠)与科学上或实际中的显著意义是不同的。因此, 无论是 NHST、置信区间还是效应大小都不能说明研究方法是完美无缺的。单独使用其中任何一个指标都无法确定备择假设(自变量对因变量产生了效应)是正确的。只有我们仔细检查了数据分析方法是否正确后, 我们才能得出哪些因素影响行为的结论。

方差分析(ANOVA)是适合于比较三个或三个以上平均数的统计方法。ANOVA 的基础是估计数据中的误差变异和系统变异来源。 F 值等于误差变异与系统变异之和(如果存在系统变异)除以误差变异的结果。所有的分析结果都包含在方差分析表中。 F 值大, 说明自变量存在效应。单因素独立组设计的效应大小指标有科恩 f 值和 η^2 。为了更清楚地确定引起显著效应的系统变异源, F 检验后还可以对两个平均数进行比较。置信区间也可以作为多重组设计的 ANOVA 的一个补充, 在报告 NHST 结果时, 也应该给出相应的值。

在复合设计研究中, 如果要同时检验两个或两个以上的自变量对行为的效应, 那么应该采用多因素方差分析。如果一个变量是独立组变量(随机组或自然组), 另一个是重复测量的被试内变量, 那么, 这就是一个混合设计。用 F 检验可以计算两个主效应和一个交互作用。如果主效应显著, 那么, 就可以通过简单主效应分析判断交互作用效应的来源。简单主效应是一个自变量在另一个自变量的一个水平上的效应。同时, 也可以用置信区间来帮助我们理解复合设计中自变量的效应。复合设计的效应大小指标通常用 η^2 表示。

重要概念

虚无假设 (H_0)

显著性水平

I 类错误

II 类错误

效力

独立组 t 检验

重复测量(被试内) t 检验

方差分析

单因素独立组设计

F 检验

综合 F 检验

联系性测量 η^2

科恩 f 值

两均数比较

复习思考题

1. 统计检验结果显著的含义是什么?
2. 在进行 NHST 时, I 类错误和 II 类错误有什么区别?
3. 决定统计效力的三个因素是什么? 控制统计效力的主要方法是什么?
4. 重复测量的实验设计为什么会比随机组设计更敏感?
5. 分别举出一个效应大小指标的优缺点和局限。
6. 为什么统计上显著的结果可能在理论上和实际应用中都没有意义?

7. 简要描述 F 检验的原理。

8. F 检验和两个平均数比较有什么差异。

9. 重复测量 ANOVA 和独立组 ANOVA 的主要差异是什么？

10. 简单主效应和总体主效应之间有什么差异？

挑战性问题的答案

1. 一位研究者用实验来比较训练儿童阅读的两种教学方法。用两种方法教学后做测验，新方法的成绩好于老方法的成绩。统计结果是 $t(120) = 2.10, p = 0.04 (d=0.34)$ 。

A. 这个结果在统计上显著吗？

B. 实验中有多少被试？

C. 效应大小 d 值在这个研究中说明了什么？

D. 研究者的结论是：根据上述结果，在教儿童阅读时新方法具有明显的现实意义，它可以应用在实际情景中。你如何看待这样的结论？

E. 如果报告置信区间，对我们理解统计结果有什么帮助？

2. 社会心理学家要比较三种宣传方式如何影响大学生对反恐战争的态度。随机选取 90 名学生 ($N=90$)，并将他们分成三组。在对他们进行三种不同宣传后，纸笔测验他们对待战争的态度。用 ANOVA 检验三种宣传方式对学生态度的影响。方差分析表如下：

来源	平方和	df	均方	F	p
宣传方式	180.10	2	90.05	17.87	0.000
误差	438.50	87	5.04		

A. 上述统计结果是否显著？为什么？

B. 从上述结果中可以很容易地计算出哪种效应大小指标？其大小是多少？

C. 如何进行两个均数的比较，以便进一步解释这些结果？

D. 虽然没有提供各组的平均数，但仍然可以通过合并

方差计算置信区间。该研究中平均数的置信区间是多少？

3. 发展心理学家对四、六和八年级学生的思维类型进行了测量。每个年级分别有 28 名学生参加两种 (A、B) 测试，每种测试各有 14 人。因变量是测验结果正确的百分比。每组测验的平均成绩见下表：

测验	四年级	六年级	八年级
测验 A	38.14	63.64	80.21
测验 B	52.29	68.64	80.93

该实验的 ANOVA 结果见下表：

来源	平方和	df	均方	F	p
年级	17698.95	2	8849.48	96.72	0.000
测验	920.05	1	920.05	10.06	0.002
年级×测验	658.67	2	329.33	3.60	0.032
误差	7136.29	78	91.49		

A. 将实验结果的平均数用图表示出来。根据图上数据的走势，你对两个自变量之间存在交互作用表示怀疑吗？为什么？

B. 哪些效应显著？用自己的话将统计显著效应描述出来。

C. 年级和测验主效应的 η^2 是多少？

D. 用什么方法对交互作用效应的来源进行分析？

E. 对不同年级来说，测验的简单主效应为何？

F. 计算 6 个平均数的置信区间，并在图上画出来。

挑战性问题的答案

A. 是的。结果表明，虚无假设为真的概率小于 0.05。

B. 自由度是 120。对于独立组 t 检验而言， $df = n_1 + n_2 - 2$ ，因此，应该有 122 名被试。

- C. 按照科恩规则,效应大小 d 值为 0.20 说明效应较小,0.50 为中等效应,0.8 为大效应。 d 值为 0.34 时属于小效应。
- D. NHST 的结果并不直接说明其具有实际意义。如果使用一个新方法的费用高并且费时,或所需的新工具(如新的阅读材料)无法立即获得,那么,研究结果的实际应用价值可能就比较有限(至少在短期内)。尤其在效应值较小的情况下更是如此。另外, $p=0.04$ 说明,在 0.05 水平上重复出现统计显著性的概率并不太高。最后,我们需要仔细检查研究方法,以确定研究结果是正确的,并没有出现变量混淆和实验者误差。
- E. 两个总体平均数差异的置信区间说明不同方法之间差异的大小和准确性(检查置信区间的宽度)。

第 14 章

心理学的学术交流

本章大纲

概 述

互联网与心理学研究

研究报告的写作指导

研究报告的结构

标题页

摘 要

引 言

方 法

结 果

讨 论

参考文献

附 录

作者注释

脚 注

文稿各页的排序

口头报告

研究计划

研究报告写作范文

概述

科学研究是一种公开的活动。只有将机智的假设、精巧的研究设计、精确的数据收集程序、可靠的结果和深刻的理论解释公布于众,科学交流才有意义。巴塞洛缪(Bartholomew, 1982)曾强调指出:“出版发行是科学研究的必要组成部分,只有经历过出版过程的痛苦,最好在高质量的相关期刊上发表论文,科学研究才是有意义的。”巴塞洛缪认为作者在投稿时首先要选择“相关”的期刊,其原因是相关期刊会有一个同行评审过程。其他研究者(“同行”)可能会被授权评审你的论文文稿,这些研究者都是某些具体研究领域的专家,他们会给出关于论文的处理意见。你的研究在方法上是否有创新之处,你的研究是否真正能够为心理学的学科发展做出贡献,这些都由同行评审者来评定。然后,他们会把评审意见提交给一位资深的研究者,即期刊的主编。主编的工作就是决定哪篇论文适合发表。同行评审是控制发表的心理学研究质量的主要方法。

心理学者可以在很多期刊上发表他们的研究。例如,《心理科学》(*Psychological Science*),《记忆与认知》(*Memory & Cognition*),《儿童发展》(*Child Development*),《人格与社会心理学期刊》(*Journal of Personality and Social Psychology*),《公众兴趣心理科学》(*Psychological Science in the Public Interest*)和《临床与咨询心理学》(*Journal of Clinical and Consulting Psychology*),这些只是众多心理学期刊中的一小部分。如同我们前面所介绍的,最终发表哪篇文稿是由期刊的主编决定的,而他们做决定的依据是:(1)研究的质量;(2)编辑和同行评审者认为这篇文稿在写作上是否符合规范。因此,在出版过程中,论文的写作格式和内容同样重要。主编往往会选择那些内容新颖、表述清楚的文稿来发表。期刊的编辑们都有严格的审阅标准。2003年,美国心理学会的26个主要期刊的编辑们共审阅了将近6000(5978)篇文稿,最终大约只有三分之一的文稿通过审核发表(68%的退稿率)(美国心理学会,2003)。

论文评审和发表这两个过程都需要很长时间。一篇论文从提交到最终在期刊上发表至少需要一年时间(有时会更长)。文稿评审过程需要几个月,在此期间决定这篇论文是否能被采用。从论文被录用到它真正在期刊上发表,这个过程同样需要几个月。如果想及时发表自己的研究成果,研究者们可以参加由一些专业协会发起的学术会议,这些专业协会包括美国心理学会(American Psychological Association),美国心理学协会(American Psychological Society),心理环境协会(Psychnomic Society),儿童发展研究会(Society for Research in Child Development)。研究者也可以出席一些地区学会举办的学术会议,例如东方、中东、东南亚或者西方心理学学会组织的学术会议。在这些会议上,研究者可以通过简短的口头报告或墙报来介绍他们的最新研究。这种会议为同一领域的研究者们提供了一次及时讨论和交流的机会。处于“排版”(例如,正在等待出版)阶段中的研究报告也可以拿到会议上讨论,这样,与会者可以提前了解许多重要的尚未公开发表的研究成果。

研究者们一般都要向政府或者私营机构申请研究经费,以确保其研究项目顺

利进行。而政府或者私营机构通常会根据研究者们所提交的研究计划并对之进行评审以决定资助哪些项目。研究生在准备硕士研究或博士学位论文时同样要提交研究计划,院系委员会会在论文开始进行之前审核这些研究计划。此外,本科生在上心理学研究方法课或实验课时通常也需要提交一个研究计划。不同学术水平的研究者最终都会发现,提交研究计划是机构审查委员会(IRB)提出的要求,其目的是请一个相关机构对这些研究计划的伦理问题进行评审(见第3章)。无论在写作风格上还是写作格式上,研究计划与期刊上的研究报告都不一样。我们将在本章给出一些写作研究计划的建议和方法。

如何使用期刊文章、口头报告和研究计划?如果是参加心理学专业的毕业论文答辩,你很可能要同时利用这三种科学交流形式来阐述你的研究。即使以后你不想在心理学领域中有进一步的专业发展,标准的写作原则和口头报告原则同样适用于各种职业场合。例如,参加工作之后部门经理可能会要求你写一份关于最近一段时间销售业绩的备忘录,其在内容和形式上就类似于一篇简短的期刊文章。与你更加直接相关的就是,在上完研究方法课之后,你可能需要提交一份研究计划或研究报告,并且还要进行口头报告。本章将帮助你做好这些事情。

美国心理学会发行的《出版手册》第五版(Publication Manual, 2001)是心理学科学写作的主要参考资料。为保证各种心理学期刊风格一致,编辑和作者们都以这本手册作为评审和写作标准。《出版手册》几乎回答了关于文稿写作风格和格式的所有问题:包括各章的写作内容和文稿的组织结构,如何阐述观点和消除偏见,美国心理学会的编辑风格,参考文献的排序格式,以往期刊和其他出版物上投稿时的文稿写作以及包括文稿电子版授权指南在内的各方面内容。因此,《出版手册》是很珍贵的学习资源。同时,这本手册中还介绍了美国心理学会杂志的管理政策和对编辑们的管理,以及科学写作中的伦理道德问题(详见第3章中关于这个问题的介绍)。美国心理学会承诺,他们的编辑风格和出版技术不会做重大调整。美国心理学会网站(www.apastyle.org)也会及时公布最新版的《出版手册》,以及论文写作风格、学会管理政策和出版程序的最新变动。

本章的很多内容直接引自2001年版的《出版手册》,¹但我们并不是对该手册的简单重复。本章的主要目的是帮助你学会写作,这也是你在研究方法课上要学习的内容。如果你正计划进行高级心理学研究,除了《出版手册》之外,我们还建议你读2001年版的《美国心理学会出版手册》(Publication Manual of the American Psychological Association)作为书架上的常备书目。当然,美国心理学会还为我们提供了一些其他实用资源,例如,盖尔芬德和沃克(H. Gelfand & C.J. Walker)于2001年出版的《掌握美国心理学会风格:学生的工具书与训练指南》(Mastering APA Style: Student's Workbook and Training Guide)。(你可以登陆美国心理学会网站[www.apastyle.org]订购美国心理学会出版的书,美国心理学会购书的通讯地址是:American Psychological Association, Book Order Department, P.O.

¹ 2001年版《出版手册》的版权归美国心理学会所有。经允许后引用。未经发行者事先的书面同意,原始版本和2001年版本均不允许再版,或者以任何形式或途径发布,任何数据库或数据检索系统不得收录。

Box 92984, Washington, DC20090-2984)。

研究者之间仅通过学术期刊和学术会议这两种正式形式进行交流还无法完全满足科学交流与合作的需要。研究者之间的非正式交流对科学研究同样重要。许多研究观点都是在科研小组举行的会议上或者是同事之间的非正式交流过程中形成和提出的。不同地区的研究者之间习惯于利用互联网进行合作交流,通过电子邮件讨论他们的研究。研究者也通过互联网使用电子数据库。互联网使研究者们有能力快速而广泛地获取和发布与他们研究相关的信息。

互联网与心理学研究

互联网,尤其是电子邮件已经成为心理学研究者之间交流的一种必不可少的途径。学生和心理学者可以利用互联网所提供的各种方式进行研究活动。例如,一些研究者以互联网作为他们原始研究的指导工具(例如, Azar, 1994a; 1994b; Birnbaum, 2000; Kardas & Milford, 1996; Kelley-Milburn & Milburn, 1995)。本章将介绍心理学家利用互联网进行研究的几种重要形式。

对于大多数研究者来说,电子邮件是他们与同事、期刊编辑、合作者、资助单位的领导以及其他专家交流的主要工具。若你在阅读一篇期刊文章时有一个疑问,可以给文章作者发电子邮件请教一下这个问题。使用电子邮件非常简单、有效和方便。例如,你可以利用电子邮件 [shaughnessy@hope.edu] 与本书的第一作者约翰·肖纳西交流。

人们可以上网登陆这本教科书的主页,上面设有学习资源(例如,练习测验)、作者简介、版本变化信息,以及心理研究的一些其他资源。还有关于当前版本中的一些错误或疏漏,本书的发行地址,订购等信息。欢迎登陆我们的网站 [www.mhhe.com/shaughnessy7]。

讨论小组,又称“问题讨论表”(Listservs),可供个人自愿参与,讨论他们都感兴趣的心理学问题。小组中有一“群”“问题贡献者”,他们会提出一些可供讨论的主题,而小组成员会立即“提供”一些关于这些主题的信息。最后,这些信息由一位问题贡献者负责发送给各位小组成员。(如果你提出的是一个大型问题讨论,你很可能需要经常编辑你的电子邮箱,或者你会发现你的邮箱里全是大家积极响应的信息。)目前,互联网上有许多问题讨论列表,这些问题讨论表让全世界的研究者可以就许多问题进行讨论,其中也包括吸毒、宗教和女性研究等。有些问题讨论表是对外开放的,允许任何有兴趣的人来参加讨论,包括那些被动参与者(“潜水者”)。有些问题讨论表只对特定的人群开放(例如,APA分会的会员)。美国心理学会也为学生设立了讨论小组(psycSTUDENTS),他们可以登陆美国心理学会网站 [www.apastyle.org],参加上面的小组讨论。

网上数据库是贮存在互联网上的电子数据文件,人们可以上网查阅这些文件。心理学数据库中往往还涉及一些关于医学、酒精中毒和民意测验方面的文件。这些数据库对档案研究(详见第6章)和时间序列分析(详见第11章)很重要。大

型数据库会贮存数百个变量和大量可用于研究的参与者的数据, 这些数据库会帮助研究者解答心理学(例如, 临床心理学、社会心理学和发展心理学)研究中的问题。这些数据库是免费的, 研究者随时可以进入数据库检索数据, 这样也可以避免他们重复已有的研究。

电子期刊现在很常见, 目前, 很多杂志和会议要求提交文稿的电子版。网络和电子邮件的普及促进了评审过程的发展, 提交文稿、同行评审和编辑向作者反馈意见这些过程完全可以通过互联网实现。另外, 一些杂志只提供电子期刊。订阅者通过电子邮箱接收文章, 读者通过网络提交他们对每篇文章的评论。Psychology、Psyche 和 Prevention and Treatment 都是电子期刊。无论是往电子期刊还是印刷期刊上投稿, 想在相关杂志上发表文章的作者都期望同行来评审他们的研究。

我们在前面提过, 原始研究可以通过互联网完成。这里重复本书第1章的一个观点: 只要是以计算机为实验仪器, 以人为参与者的心理学研究, 都可以通过网络完成(Krantz & Dalal, 2000; 详见 Kraut 等人 2004 年发表的关于在线研究的信息)。在计划和实施研究时, 你会发现互联网对于研究的作用取决于你所做研究的具体需要和你网络的熟悉程度。如果你是新手, 我们再次建议你阅读弗雷利(Fraley)的著作《如何在互联网上进行行为研究》(How to Conduct Behavioral Research Over the Internet. 2004; New York: Guilford Press)。

研究报告的写作指导

学习写作类似于学习游泳、开车或是弹钢琴, 只靠阅读这一种活动是不可能取得进步的。专家的书面对建议也只能帮助你入门。要实现有效的写作, 首先必须有一个有价值的研究让你去写, 然后, 尽可能以最好的方式将其表述出来。学好写作的一个关键就是能够从“教练”——老师、朋友、编辑甚至是自己那里获取批评意见。克龙巴赫(Cronbach, 1992)的几篇文章是《心理学公报》(Psychological Bulletin)中被引用次数最多的, 关于写作, 他给出以下建议:

我的建议类似于著名的“野兔烹调法”, 其中的第一句就是“首先你要抓住兔子”。论文写作是这样, 首先你要有一个值得传递的信息, 然后, 注意写作细节……改写那些语句不流畅或者缺乏韵律、没有突出重点、让多数读者甚至作者本人理解起来很困难的句子。最佳写作效果是能够将专业文章写得充满文学气息, 将论文摘要写成难忘的故事, 将忙碌的速度写成舒适的节奏, 以及将事实写成诗歌。(p.391)

职业作家杰克·里德(Jack Ridl)提出了有效写作的第一定律: “写作, 不是奢望让人理解你, 而是尝试不要让人误解你。”写好文章如同把车开好, 最重要的是做好防御工作, 首先就假设任何地方都有可能产生误解! 要想避免出现问题的话, 你应该在开始写作之前考虑以下几个问题。

- 确定读者对象。如果认为自己的读者懂得的知识很多, 但实际上他们并不懂

得,那么你的写作就会让他们感到很困惑。如果低估你的读者,你的写作中就会出现一些令人厌烦的不必要的细节。无论是高估还是低估你的读者,都很可能导致你的文章少有人读。但是,如果一定要选择,你最好还是低估自己的读者。例如,在上心理学课时,你要准备一个研究报告,合理的假设读者是你的老师。以老师为假设读者来准备报告,会让你的报告忽略很多内容,原因是你觉得无论怎么写老师都能明白和理解。在下次研究方法课上,你最好以同学为假设读者,这样你的写作效果可能会更好。你会写出所有细节,甚至包括一些不必要的内容,但这只会让你的讲师简单地帮你“删除”一些不必要的内容,而不是“加入”一些漏掉的必要材料。无论假设读者是谁,在开始写作前一定要确定读者对象,并且在写作过程中要时刻牢记你的读者。

- **明确目的。**期刊文章写作一般属于解释性写作类型。韦氏字典中,将其定义为“传递信息或是解答疑难。”期刊文章的基本写作目的就是表述清楚并使人信服。首先,你要将你所做的研究工作和所发现的结果表述出来。其次,你要让读者相信你对研究结果的解释是合理的。
- **表述清楚。**思路清晰、表述清楚是写好解释性文章的基础。2001年版的《出版手册》(2001)介绍了清楚表述的方法:

有次序地阐述观点,同时语言要流畅和准确。做到这两点你就实现了有效的交流,而有效的交流是进行科学报告的基本目标。清楚而有逻辑地阐述每个观点,引发读者不断地思考,读者就会怀着愉快的心情读完你的文章。

《出版手册》列举了表述清楚的三条原则:

1. 表述要简洁,
 2. 准确,
 3. 遵守语法规则。
- **简洁。**只写必须写的内容,你就做到了表述简洁。为了让读者理解起来更容易,你在写作过程中要尽量使用常见的词汇和简短的句子。解决言语罗嗦的最好方法就是多次修改草稿,同时请人帮忙修改。
 - **准确。**准确就是选择正确的词汇组成语句,将你想说的内容表述出来。《出版手册》提出了一些关于表述准确的好建议:

确保每个单词都能准确表达你想要表达的意思。很多作者经常发现他们想用一个术语表达的意思与这个术语在字典中的定义有时会不一样。例如,在非正式情况下,“感觉”可以广泛代替“认为”或“相信”,但在正式的科学背景背景下,这种替代是不允许的。(pp.36-37)

- **遵守语法规则。**遵守语法规则是成功写作的一个必要条件。因为如果不遵守语法规则,读者就无法理解你所要表达的意思,同时会让读者觉得写作者水平很差,从而降低对你(和你的观点)的信任。表述简洁、准确和遵守语法规则并不能保证实现有效的写作,然而,它们的有效写作的前提。

- **写作公正。**作为一位论文作者，你应该尽量保证所选择的词汇和使用的句子结构对所有人都是公正和没有偏见的。美国心理学会列出了有关作者使用带有偏见语言的政策（《出版手册》（2001））。

作为发行者，只要作者的语言准确、清楚和符合语法规则，美国心理学会就会接受他们的语言选择。作为一个组织，美国心理学会要对科学负责，同时也有责任公平地对待每个个体与群体。这条政策要求在美国心理学会出版文章的作者在写作中要避免出现贬低他人的态度和带有偏见的假设。关于性别、性取向、种族、残疾或年龄的描述中通常会隐含对他人的偏见，作者在撰写这些方面内容时要避免出现偏见。科学写作应该尽量避免对他人或正在研究的群体做带有暗示性或无关的评价。

《出版手册》（2001）提出了实现无偏见写作的几条原则：

1. 具体恰当地描述个体或群体。例如，当研究涉及所有成年人时，使用短语“男人和女人”要比概括性术语“人们”更准确。在描述研究被试时，使用“华裔美国人”或“墨西哥裔美国人”可能要比亚裔美国人或西班牙裔美国人更准确。
 2. 在谈论人的时候，要注意自己所使用的标签。例如，我们在描述人的种族身份时所使用的术语。避免这个问题的最好方法就是任何时候都不要给人贴标签，更要杜绝使用让对方感觉到带有轻蔑意思的语言。想要实现这条原则就必须记住，某个群体所偏爱的标签会随时间而改变，同时，群体中的不同个体所偏爱的标签也不一样。
 3. “在研究报告中，描述和介绍人们时突出他们的参与性。”（p.65）。可以使用一些描述性语言介绍研究中的参与者，例如，用“大学生或儿童”，最好不要用“被试”。在答谢参与者时使用主动语态要好于使用被动语态——“完成这个调查的学生”要好于“被调查的学生”。《出版手册》（2001）中介绍了关于用标签标示人们的性别、性取向、种族、残疾和年龄时的指导原则，以及这些原则的运用与举例。
- **报告要有趣。**《出版手册》（2001）就科学写作的整体语气给出了一些有价值的建议：

尽管科学写作与文学创作在形式上有所不同，但没有必要也不应该将科学写作写得没特点或者写得很沉闷。在描述你的研究结果时，你可以直接呈现观点和发现，但最好以有趣的和引人入胜的方式进行描述，这种写作方式能反映出你对这个问题很投入。（p.10）

要想把握住研究报告的写作语气，一个恰当的方法就是尽量像讲故事一样描述你的研究。好的研究类似于动听的故事；动听的故事能够推动研究。

研究报告的结构

作者和读者都需要了解研究报告的结构。结构是关于研究的组织，研究者根据研究报告的组织结构写作，才能清楚地阐述研究并对研究结果进行合理的解释。在某种意义上，研究报告的结构类似于莎士比亚笔下的戏剧结构。在莎士比亚的戏剧中，剧作家和观众对戏剧中每一幕应该发生的情节抱有一定的期望。例如，第一幕中会出现“舞台情景”，高潮有望出现在第三幕中。相似地，在一个研究报告中，作者和读者对报告的每部分内容都有相同的期望。如同戏剧观众一样，研究报告的读者会期望在每部分中都能看到一些具体信息。如果想了解一个实验是如何做的，你应该阅读第二部分；如果想了解这个研究中关于数据分析的内容，你可能会阅读结果部分。结构不是用于束缚剧作家和报告作者的，而是为观众或读者提供一个便于找到戏剧或研究具体内容的工具。

一个研究报告包括以下几部分：

标题页

摘要

引言

方法

结果

讨论

参考文献

附录

作者注释

脚注

报告的主体

研究报告的主体由四部分组成：引言、方法、结果和讨论（“戏剧”的四幕）。标题页和摘要类似于戏剧正式开演前你看到的戏剧海报；参考文献、附录和注释类似于片尾字幕。在本章中，我们将介绍每一部分的写作内容和格式，详细内容请阅读《出版手册》（2001）。要想学会写研究报告，仅靠阅读本章内容和《出版手册》（2001）是不够的，还需要阅读你感兴趣的心理学领域所发表的期刊文章。当然，能否掌握研究报告写作这门技术最终在于你的实际练习。

标题页

研究报告的第一页是标题页。标题页的内容包括告诉读者这是一个关于什么的研究（例如，题目），这个研究是由谁做的（例如，作者），这个研究是在哪里完成的（例如，作者的单位），以及一个简短的开头，使读者明白这篇文章的主要内容（例如，“页首短题”）。在本章最后，我们给出了一个研究报告的写作范文，列举了标题页的正确写作格式和后续研究报告的写作内容。这个范文中的一些注释是用于强调研究报告最终打印时需要注意的一些问题。

题目可能是一篇论文中最关键的部分，因为它是被读者阅读的内容。题目应当清楚地显示出你的论文所研究的核心问题。《出版手册》（2001）中对题目写作的表述是：

题目应当是对一篇论文主要观点的简单概括，如果可能，这个概括要有特点。题目是对研究主题的简洁描述，同时，这个描述要能反映出正在进行的研究所包括的实际变量或理论问题以及二者之间的关系。

每篇文章的参考文献部分一般都会索引和汇编许多相关研究的题目。因此，题目中避免使用无意义的和没有明确目的的词汇。无意义的词汇只会增加题目的长度，误导人。例如，题目中一般不会使用方法和结果的词汇，也不会以“一个研究”或“一项实验研究”作为题目的开头。题目中避免使用缩写词，拼写出所有术语可以确保对文章准确而完整的索引。题目的长度最好在10~12个单词之内。（pp.10-11）

研究报告题目之下的署名一行包括研究者的姓名和他们所属的单位。我们在第3章中已经介绍过确定作者身份的标准，只有满足这些标准的人才可以列为研究报告的作者。作者要在注释中感谢其他对本研究有贡献的人。作者姓名的写作顺序是：先写作者的第一个名字，然后写中间名，最后是作者的姓。

《出版手册》（2001）用简短的语言描述了标题页的页头标题：“页头标题是一个简缩的题目，位于期刊文章页面的最上边，其作用是帮助读者识别这篇文章”（p.12）。页头标题的长度一般不超过50个字母，并且所有字母都要大写，位于左页边，与标题页第1行之间间隔2倍行距。标题页第1行的写作内容是该页的页码1和文稿页眉（具体见研究报告写作范文）。注意，页头标题只出现在文稿的标题页，它与每页文稿上都会出现的文稿页眉是不一样的。文稿页眉由论文题目的前几个单词组成，与文稿每页的页码处于同一行。

许多期刊通常采用匿名评审，或者出于作者的要求采取匿名评审。匿名评审就是评审人不知道作者的身份，因此，他们在评审过程中不会受作者的身份和地位的影响，这在一定程度上提高了评审过程的公正性。即使采用匿名评审，文稿的标题页也必须写出所有关于作者身份的信息。《出版手册》（2001）中介绍了匿名评审的程序：

匿名评审要求在评审过程中要隐去文稿中作者的身份信息。接受匿名评审的文稿其作者有责任为在文稿中隐去作者的身份信息提供便利：例如，作者必须将作者注释打印在文稿的标题页，编辑会在文稿送去评审前留下标题页。（p.361）

题目的写作技巧 研究报告的常见题目写作形式是“[因变量]与[自变量]之间的关系。”例如，“字谜问题的解答时间与问题难度之间的关系”就是一个好题目。题目不仅要能提供信息，而且要简洁（10~12个单词）。最重要的是，你的标题要尽可能清晰地阐述研究的内容。

摘要

摘要是一段关于研究内容和目的的总结，其长度在100~200个单词之间（大约10行）。摘要始于研究报告的第2页，以“摘要”为标题，并且居中（详见研究报告写作范文）。摘要中应该包括你的研究的四大成分：

- (a) 研究的问题；
- (b) 方法，包括本研究所使用的测验和仪器，数据收集过程和参与者的有关特征；
- (c) 结果；
- (d) 结论与结果的含义。

也就是说，摘要应该介绍研究报告引言、方法、结果和讨论四部分中的关键内容。

一个好的摘要会对读者阅读文章的其他部分产生重大影响。摘要还提供了用于查阅这篇文章的索引信息和文章检索信息。一篇文章的命运取决于读者对该篇文章摘要的反应。关于摘要的重要性，《出版手册》描述如下：

大多数人与一篇文章的第一次接触就是阅读这篇文章的摘要。通常的情况是，人们在写文献综述时，他们会利用电子摘要检索系统查阅相关文献，并且是在显示器上同时浏览几篇文章的摘要。无论是坐在计算机前看资料还是翻阅期刊杂志，读者都是根据摘要决定是否阅读全文。摘要写作不仅要用简短的话阐述大量信息，而且还要做到通俗易懂、结构清晰、语言简洁和内容完整。同时，在摘要中插入关键词会提高读者发现这篇摘要的可能性。（p.12）

摘要的写作技巧 写好摘要是一个很大的挑战。迎接这个挑战的最佳方式就是在最后写摘要。在写完研究报告的所有其他部分之后再写摘要，只有其他部分都写完你才能较容易地用自己的话对文章进行概括或者解释。

引言

引言的目的 在第3页的页面上边居中写上研究报告的题目。与论文的其他部分不同，引言部分没有标题，这是因为引言是论文的起始部分。因此，引言部分的第一段始于论文题目之后（详见研究报告写作范文）。引言部分的写作目的包括：

1. 介绍要研究的问题，说明研究这个问题的重要性；
2. 阐述本研究的理论意义，简单介绍与本研究相关的文献背景；
3. 阐明研究目的、基本原理和研究设计，以及通过逻辑推理得出的预期结果或指导本研究的假设。

上述三个目的的写作顺序是可以变化的，这里介绍的顺序是常用的写作顺序。

《出版手册》(2001)认为这三个目的都是为了:“通过总结相关研究结论和数据,一个成功的引言能够让读者清楚地了解已经进行了哪些研究和为什么要进行这些研究”(p.16)。

引言的第二个目的就是总结和综述相关研究。综述的目的不是要详尽无遗地罗列相关文献,相反,你应该仔细选择那些与你的研究直接相关的研究。在总结这些早期研究时,你应该强调前人研究中究竟是用哪些细节最有助于读者理解你做了什么以及你进行这些研究的原因。同时,你还要承认帮助你理解这个问题的早期研究者的贡献。当然,如果直接引用他们的研究成果,你必须用引号标明所引用的内容(见第3章引用他人成果的建议)。

参考文献是注明引用他人研究成果的常用形式。引用他人研究成果的表述形式有两种:一种是你你在论文中写出所引用文章作者的姓,并且在姓后用括号注明这篇文章的发表年份;另一种是概括介绍他人的研究,然后在括号内注明这个研究的作者和发表年份。例如,如果你在论文中引用了Lorna Hernandez Jarvis和Patricia V. Roehling在2003年发表的一项研究,引用形式是“Jarvis和Roehling(2003)研究发现……”,或者是“最新研究表明(Jarvis和Roehling,2003)……”。同时,你还要在参考文献部分写出Jarvis和Roehling这篇文章的完整索引信息,包括期刊名、卷和起止页码。心理学研究报告不使用脚注来注释参考文献。

只有真正阅读过的文章才可以作为你的论文的参考文献。如果你阅读了巴尼(Barney,2002)的一篇文章,文章里介绍了Ludwig(2000)的研究,在这种情况下,你不可以直接引用Ludwig的研究,除非你真正读过他的那篇文章。然而,你可以采用这种引用形式,“如同Barney(2002)所描述的,路德维奇(Ludwig,2000)的研究发现……”。采用这种引用形式的原因有两个:最重要的原因是你应该如实报告你所阅读过的内容。如果不具备诚实的学术态度,你就应该认识到自己所冒的风险,如果巴尼和路德维奇研究的报告是错误的,那么你就是在重复错误,同样要受到人们的批评。实现这条规则很简单:只引用你阅读过的研究。(见第3章)

引言的写作技巧 要想写出有效的引言,在开始写作之前你首先要回答以下四个问题:

- 这个问题为什么很重要?
- 如何将假设和实验设计与所研究的问题联系起来?
- 本研究的理论意义是什么,如何将本研究与以前该领域中的研究结合起来?
- 本研究的理论命题要检验什么,这些命题是怎么得来的? 《出版手册》(p.16)

心理学文献的检索 从长远的观点来看,要想提出研究的问题,掌握相关文献,最好的方法就是有规律地阅读你感兴趣领域中的期刊杂志。在刚开始接触心理学研究时,你可能没有机会去广泛地阅读文献,或者是不能确定一个你感兴趣的心理学研究领域。即使找到了感兴趣的研究领域,你也需要从一些新文献中寻找研究问题。例如,你可能想做某个实验,这时首先你要确定这个实验是否已经有人做过了。或者是你看过一个实验,想以它为基础再做另外一个实验,这时你就要

搜集与这个主题相关的其他研究。你也可以阅读这篇文章参考文献中引用的与你的主题相关的研究,这也是一个重要的文献检索来源。无论研究什么主题或问题,你都需要搜集心理学文献。心理学文献检索资料会帮助你解答以上这些问题。

目前,心理学文献的检索基础是《心理学摘要》(Psychological Abstracts),它创刊于1927年,是由美国心理学会组织出版的。《心理学摘要》属于月刊,其中的摘要来自于1000种国内和国际期刊杂志。这些摘要按照常规的分类进行分类,例如,“生理与心理失调”。通过印刷版的《心理学摘要》检索文献是一件既耗时又费力的事,利用电子数据库检索会更容易一些。

目前,PsycINFO是主要的在线心理学文献检索数据库。人们可以利用在线数据库First-Search和InfoTrac进入到PsycINFO中检索文章。可以向当地图书馆管理人员咨询一下他们都提供哪些在线服务项目。人们可以浏览电子数据库中所有文章的题目和摘要,也可以浏览包含特定关键词的文章题目和摘要。在这种类型数据库中检索文献的最有效方法就是使用交叉关键词法进行检索,所谓交叉关键词法就是同时使用两个关键词进行检索。例如,某学生想调查强奸和约会时其他性侵犯的发生率[例如,约会强奸(date rapes)]。交叉关键词法就是同时以RAPE(强奸)和字符串DAT为关键词检索相关文献。选择以字符串DAT为检索词的目的是,这种方法可以检索到包括DATE、DATES和DATING等在内的文献。利用这种交叉检索法可以检索到75篇文献,其中73篇是英文文献。

利用电子数据库查阅文献的优势就是人们可以快速地进入到一个大型信息库,并对其中的信息进行检索。当然,这种检索方式也存在一些潜在的弊端。问题之一就是你能搜索到多少相关文献直接取决于你的检索能力。在利用电子数据库进行检索时,你能得到的就是你所寻找的。使用不同的关键词多次检索如此大的数据库之后,我们常常会自负的认为“这些就是关于这个主题的所有文献。”然而,很可能出现这种情况——无论怎样检索电子数据库都会遗漏一些信息。关键词检索法有时也是不可靠的。字符串DAT可以识别所有带有DATA的研究,很多学生的参考文献中也都会包括与强奸有关的数据——但是并不是全部在约会情境中。这些潜在的问题不是电子数据库造成的,而是由电子数据库的使用者引起的。如果能够熟练使用电子数据库,利用PsycINFO检索心理学文献的优势要远大于其弊端。除了一些基本要素外,约斯维克(Joswick, 1994)还提供了使用电子数据库查阅文献的一些建议。

方 法

研究报告主体的第二大部分就是方法。方法部分紧随引言结束而出现。方法部分的标题(方法)居中,并且标题与引言部分和方法正文之间各间隔两倍行距(详见研究报告写作范文)。引言部分概括介绍你已经完成的研究,方法部分则是详细描述你的研究。《出版手册》直接阐述了方法部分的写作目的:

方法部分就是详细描述你是如何进行研究的,要能够让读者根据你的描述评价你所使用的方法是否得当,结果是否可信和有效。同时,你的描述能够让有经验的研究者重复该研究。

方法部分的写作听起来很简单，实际上写好这部分是很难的，这是因为你要准确描述你所做的每一步。如果想体验一下这个挑战究竟有多大，你不妨试着用清楚而有趣的语言描述一下你是如何系鞋带的。

方法部分的写作技巧 写好方法部分的关键就是组织。幸运的是，基本上所有研究报告方法部分的组织结构都是一样的，都是由一些基本成分组成的，而大多数研究报告都给出了你所需要的这些基本成分的组织模式。在写此部分内容之前，我们必须回答一个问题，这也是我们经常向第一次写研究报告的学生提问的问题：“我应该写得多么详细？”写得太概括或者是太详细都会影响论文的质量。《出版手册》中关于这条写作规则的描述非常简单：“在这些成分中只写那些有助于理解和重复这个研究的必要信息”（p.18）。正如我们以前所说的，体会这条规则的最佳方式就是阅读期刊文章的相应部分，然后进行实践练习。同时，可以向你的老师请教你的研究报告需要写到何种详细程度。

研究方法部分的三个常见成分是参与者、实验材料（仪器）和实验程序。各部分标题位于左页边顶格写，字体为斜体（详见研究报告写作范文）。《出版手册》（p.18）总结了“参与者”这一部分的写作目的和内容。

在科学和心理学的研究中选择合适的参与者或对象非常重要，尤其是在结果评价（进行各組间比较）、结果推广、在重复研究中进行比较、文献综述或者是二次数据分析比较中都特别重要。

当以人为研究被试时，要报告选择和分配被试的程序，以及与被试之间的协议和报酬问题。报告被试的主要人口学特征，例如，性别、年龄、种族。如果可能，还要报告参与者的社会经济地位、健康状况和性取向。如果某些人口学特征是实验变量，或者它对结果解释很重要，研究报告要具体描述这个群体——例如，国籍、受教育程度、健康状况以及语言的选择和使用。

当以动物为被试时，研究报告要介绍动物的属、种类、品种号或其他具体身份特征，例如，名子、供应地区和血缘关系。

研究还要报告被试总数以及每种实验条件下的被试数量。如果有被试没有完成实验，还需要报告有多少被试未完成实验并说明原因。（pp.18-19）

实验仪器或实验材料有时可以不报告。如果所使用的实验工具仅仅就是纸、铅笔和一台计算机，你最好是在实验程序的单元中报告这些信息，而不用单独列出实验仪器部分。但是，如果你所使用的仪器或材料在研究中很重要，这时应该单独列出一个部分。如果使用的仪器很复杂或者是定制仪器，无论对于读者还是作者，采用图表报告实验仪器效果会更好。一般来说，如果实验采用的是机械仪器，标题常使用“仪器”；如果实验使用的是一些非机械的工具，例如纸笔问卷，标题常使用“材料”。如果你所使用的仪器或材料是由他人提供的，你应该注明该研究者所做的工作，同时，你还要在报告中介绍所使用材料的基本特征。

研究程序是“方法”部分最重要的组成单元。在这个部分中，你要完整描述对参与者开始测验到结束实验整个过程。这部分写作通常是按时间顺序进行的。

在写“程序”部分时，首先要列出每组参与者测验过程中的重要步骤。然后，你可以依次介绍各组参与者的实验程序，也可以同时介绍所有参与者都参与的实验程序，然后再指出各组之间不同的实验程序。无论采用哪种组织方式（只有在尝试了两种组织方式之后你才能确定哪种方式更好），你最好先列出程序中的一些重要特征，然后再开始写作。在实验过程中，应当以解释性语言向参与者呈现指导语，但是，如果指导语是对实验操作的精确定义，这时不要作任何解释和修改，逐字呈现指导语。《出版手册》认为“方法”部分，尤其是“程序”单元，“应该向读者详细介绍你都做了什么，如何做的，只有这样读者才可能重复你的研究”（p.20）。

结果

结果是研究报告主体的第三大部分，标题“结果”居中。结果部分紧随方法部分结束出现，标题与方法部分和结果正文之间各间隔两倍行距（详见研究报告写作范文）。像戏剧表演的第三幕一样，结果部分是研究报告的高潮——本研究的实际发现。然而，大多数学生在结果部分只报告一些必要的统计信息，对高潮的描写令人感觉很平淡。解决这个问题的最好办法就是根据统一格式报告出统计概念。此外，你在开始写结果部分之前最好先阅读一下第12章和第13章的内容，或者是先列出一个简单的组织结构，然后根据这个组织结构写。

结果部分是对你在引言部分提出问题的回答。结果部分的写作原则是“坚持事实，只写事实。”而讨论部分还要写事实以外的内容。

统计结果写作 《出版手册》精辟地总结了结果部分的写作目的：

结果部分的内容包括描述收集到的数据和对数据进行的统计或分析处理方法。详

表 14.1 结果部分典型段落的写作结构

1. 陈述进行该项统计分析的目的。
2. 明确用哪些描述性统计指标来概括数据。
3. 以文本、表格或图示形式呈现各个实验条件下的描述性统计结果。
4. 如果采用表格或图示，要指出读者应该关注的主要结果。
5. 列出置信区间、效应大小和推论统计检验的结果，并说明进行这些统计检验的原因。
6. 概括各项统计检验可得出的结论，但不要对结论进行讨论。结论讨论属于讨论部分的写作内容。

写作范文

为考察回忆与指导类型之间的关系，实验记录了不同指导条件下每个参与者回忆出的单词数量。只有当参与者回忆出的单词与目标项目中的某个单词正确匹配时，这个测验才能得分。错误拼写的单词如果与某个目标项目的拼写很相似，这个拼写错误的单词也按照正确处理。奇异想像条件、一般想像条件和控制条件下各组参与者回忆出的平均单词数量（标准差）分别是：15.6（1.44）、15.2（1.15）和10.1（1.00）。95%置信区间值分别是：奇异想像，13.18-18.02；一般想像，12.78-17.62；控制条件，7.68-12.52。由此可见，三组均值之间的差异达到统计显著性， $F(2, 72) = 162.84, p \leq .0005, MSE = 1.47, \eta^2 = .82$ 。置信区间之间的比较表明，想像条件与控制条件之间存在显著差异，但两种想像条件之间没有显著差异。这表明，给予想像指导的参与者的记忆成绩要高于没有进行任何指导的参与者的成绩，但两种想像指导条件下参与者的保持成绩之间没有显著差异。

细而适度地报告数据的目的是用于证明结论。你要报告所有相关结果,包括与假设相违背的结果,不要报告某个具体参与者的成绩或原始数据,除非你采用的是个案设计或者是想就样本举例。同时,在这部分不要对结果的意义进行讨论。(p.20)

实现有效写作结果部分的方法就是利用段落写作结构。表 14.1 给出了结果部分段落写作的结构,并且列举了一篇期刊文章结果部分中的一个段落。

结果部分的写作技巧 我们建议你按照下列步骤进行结果部分写作。

- 第 1 步。结果部分各段落写作的第一步是用简洁的语言介绍进行该项分析的目的。一般不超过一句话。例如,在上面的段落写作范文中,分析的目的是“考察回忆与指导类型之间的关系。”清楚阐述数据分析的目的有两个原因:一个是帮助读者理解你的数据分析逻辑,另一个更重要的原因是确保你自己理解进行这项数据分析的目的。
- 第 2 步。结果部分各段落写作的第二步是明确因变量的描述统计量。例如,在上面的例子中,你可能会以平均回忆单词数量作为描述统计量。另外,你还可以用其他描述统计量来概括每种实验条件下的结果,例如,反应时中数或者是每分钟的反应累积值。
- 第 3 步。第三步是总结各种实验条件下的描述统计结果。在报告集中趋势测量值时还要报告其他相关变量的测量值,例如,报告均值时还要报告这个均值的标准差,同时,最好也报告一下效应大小。如果只有两种或三种实验条件,使用文本描述实验结果即可。例如,在只有两种条件的研究中,你可以这样描述,“实验组平均正确反应值是 10.5 ($SD=2.1$),控制组是 8.2 ($SD=1.8$)。效应大小 (d) 是 1.20。”一般情况下,用于总结结果的数据会很多,这时你可能需要利用表格或图示(曲线图)来呈现它们。我们会在后面介绍构建表格或图示的程序。《出版手册》就表格和图示的使用给出了一些好的建议:

以清楚而简洁的方式报告数据。可以利用表格呈现真实的数据,如果准备充分,还可以以读者熟悉的表格形式(例如,ANOVA 表格)呈现复杂的数据和分析结果。专业的图示能够吸引读者的注意力,而读者也能够快速地对图示结果产生一个视觉印象。图示能够很好地反映出各种复杂的关系和一些基本的比较,但是在精确性上它们往往不如表格。

与文本形式相比,采用表格或图示呈现数据和分析结果效果可能更好;例如,用表格呈现多元方差分析结果,读者阅读起来更容易一些。但是,在结果写作过程中要避免在几个地方重复呈现相同的数据,当数据可以用几句话概括时就不必使用表格。(p.21)

- 第 4 步。我们不要错误地认为给出表格或图示就结束了。读者需要在你的帮助下从表格或图示中获取更多的信息。你是最了解研究结果的人,因此,你是提供这种帮助的最佳人选。你应当直接引导读者注意表格或图示中重要的数据,特别是那些与你在引言中提出的假设一致(或存在差异)的结

果数据。不要既用表格又用图形呈现相同的数据结果。无论选择哪种数据呈现形式，一定要用文本强调表格或图示中的重要结果。

- 第5步。结果部分各段落写作的第五步是呈现推论统计检验的结果。每种推论统计检验的报告内容包括：检验名称（通常使用符号 t 、 r 或 F 来表示）；自由度（在检验名称之后用括号括起来）；检验结果。除了效应大小之外，还必须报告检验结果的发生概率。《出版手册》中通过一个很有说服力的例子来说明报告置信区间的重要性。（见第12章和第13章。）

置信区间（用于估计参数，参数之间的函数关系如均值之间的差异，以及效应大小）是结果报告的最有效方式。因为置信区间中包含了位置和精确度信息，根据这些信息可以直接推论出显著性水平。总之，报告置信区间是最好的报告策略。因此，强烈推荐报告置信区间。（p.22）

例如，你可能看过，“控制条件下95%置信区间值是4.5~6.4，实验条件下是7.2~9.1。药品变量的效应达到差异显著， $F(1, 64) = 7.15$ ， $p \leq .0005$ 。效应大小（ d ）是.85。”

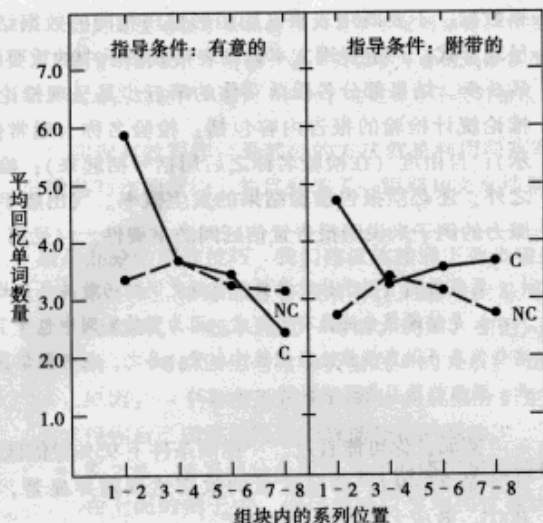
- 结论。结果部分写作的最后一步就是简单概括每种检验可得出的结论。例如，假设在一个研究中实验组答对的平均分是10，控制组是5，两个均值的置信区间之间没有交叉重叠。那么，你就可以得出一个这样的结论：“控制组的成绩要低于实验组。”这个例子比较简单，结论也很明显，但是结论表述一定要恰当，特别是概括复杂的统计分析结果的结论。

结果部分各段落的写作结构见表14.1。根据这种方法进行写作会减轻读者在统计方面的阅读负担。结果部分写作的最大挑战是确定哪些结果最重要，确定所有与引言中提出的问题相关的数据。在对结果部分写作进行总结之前，我们还要简短地介绍一下构建表格和图形的基本过程，因为，表格和图示都是有效报告结果的工具。

采用表格呈现数据 表格是能够以简洁的形式呈现大量数据的有效方式。表格的作用是对论文文本内容的补充，而不是对信息的复制。一个研究报告中的表格要按顺序连续编号。在文本内容写作中，如果涉及到某个表格，作者只需写这个表格的编号即可。因此，对表格进行编号可以方便文本内容写作。每个表格都要有一个简短的说明性标题。同时，表格的行与列要标记清楚。表格中的数据在文本报告中要保持同样的精确度（例如，所有数值都要保留相同的小数位数），所有数值都要与相应行和列的变量名对应。我们在本章结尾的研究报告写作范文中列举了一个表格。

采用图示呈现数据 与表格一样，图示也是呈现大量信息的一种简洁方式。一个图示有两个基本轴：水平轴或 x 轴与垂直轴或 y 轴。一般情况下，自变量的各个水平分布在 x 轴上，因变量分布在 y 轴上。当研究中包含两个或多个自变量时，第二个和其他自变量的各水平作为数据标签分布在图内或者是用于图示说明。在图14.1中，因变量的数值（平均回忆单词数量）分布在 y 轴上， x 轴上是自变量

图 14.1 平均回忆单词数量 (可能是 10) 与组块内的系列位置、线索提示 (C=有提示; NC=无提示) 和指导条件之间的关系

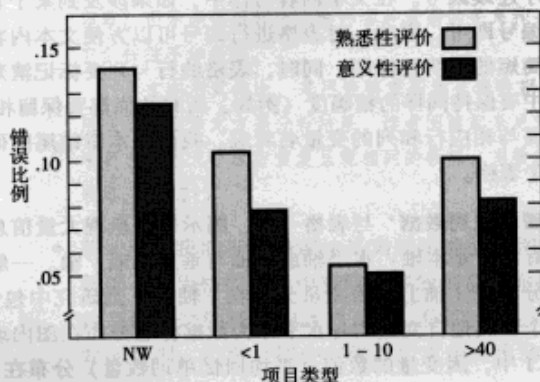


的各个水平 (系列位置)。第二个自变量的两个水平 (有提示或无提示) 作为数据标签分布在图内, 第三个自变量 (指导) 的两个水平分别用作图示的两个版面上的标题。

心理学中常用的图示主要包括两种类型: 曲线图和条形图。常用的图示类型是如图 14.1 之类的曲线图。当分布在 x 轴上的自变量是一个称名量表变量时, 常使用条形图。例如, 主修不同专业课 (自变量) 的学生的平均 GPA (因变量), 你可能会利用条形图构建这个图示。图 14.2 是一个条形图的例子。

还有其他类型图示。每种图示都有其具体的特征。在建立图示时, 一定要标记清楚 x 轴和 y 轴, 轴标题要紧贴对应的轴。同时, 要用标签式坐标标记标明每个轴上的选定点 (又称为格点), 坐标标签通常横向排列。建立图示时还要确定 x 轴和 y 轴的坐标刻度, 这样读者才能看懂分布在整个图示中的数据。在提交的文稿中, 图示与图示说明可以分开呈现 (详见研究报告写作范文)。

图 14.2 在对词汇项目进行熟悉性评价或意义性评价后两组大学生的再认错误比例。词汇项目分为: 非词汇 (NW), 桑代克-洛奇频数表中出现概率低于百万分之一, 位于百万分之一至百万分之四十之间, 高于百万分之四十的词汇



讨 论

研究报告主体的第四大部分是讨论。讨论部分（与方法和结果部分一样）紧随结果部分结束而出现，以“讨论”为居中标题（详见研究报告写作范文示例）。与结果部分不同，讨论部分“不仅仅写事实”。《出版手册》如此阐述：

结果写完之后，你就要开始评价结果和解释它们的意思，特别是要讨论与最初假设有关的那些结果。你可以核查、解释和评价研究结果，也可以对结果进行推论。（p.26）

（注意：如果结果比较少，讨论也很直接并且篇幅较短，这时可将结果和讨论两部分合并为“结果与讨论”部分。）

讨论部分的第一项写作内容就是清楚地介绍你的研究的一些重要发现。你不要重复叙述结果部分的一些描述性统计，也不要介绍得出这些发现所使用的统计分析方法。你应该着重讨论你的发现是如何证明或反驳了你最初的假设。在对结果进行简单介绍之后该写什么，不该写什么，《出版手册》给出了一些建议：

首先要澄清你的结果与其他人的研究结果之间的相同和不同之处，从而进一步证实你的结论。然而，不要简单的描述和重复已有的观点。要重新论述，并且这些新的论述要能够支持你的观点，同时还要能够帮助读者理解问题。在讨论部分还要写出本研究的不足之处，以及对结果进行的其他解释。（p.26）

讨论部分的写作语气和风格应与引言部分一致。但是，需要注意的是讨论部分的论述要与结果部分所报告的数据一致。例如，如果两组均值之间的差异是不显著的，你不可说一组好于另一组，因为两组之间的差异并没有达到“显著”的程度。

《出版手册》中提到，在讨论时要写出你的发现与你在引言中所用的文献之间的关系。如果结果与最初假设不一致，你要解释产生不一致的原因。这种事（事实之后）解释最好是尝试性的思考。如果对产生不一致结果的原因不清楚，你不可以随便解释。在研究报告中，利用一个段落来阐述本研究的局限性或存在的问题也很必要并有意义。如何找出你的研究的局限性或存在的问题，最好的办法就是尝试像其他人一样批评你的研究。

如果可能，提出后续研究作为讨论部分的最后总结。后续指的是你本该研究但却没有研究的问题。后续研究要研究什么以及为什么要进行这些研究，要尽量写得具体些。同时，你还要阐述通过后续研究应该能够得出哪些我们不知道的结论。“以年龄小些的人为参与者进行这个研究可能会更有意义”，如果你这样解释后续研究，读者很难理解。如果你能详细解释为什么采用年龄小一些的人为参与者可能会得到不同的结果，通过后续实验预期会得出什么结论，这样读者理解起来会更容易。记住，在写后续研究时不要空喊口号，一定要写得具体些。在写研究报告时，你应该注重质量，而非数量。

讨论部分的写作技巧 《出版手册》给出了讨论部分写作的一些好建议：“讨论部分的写作内容可能包括对以下几类问题的回答：

- 问题选择：为什么这个问题很重要？
- 分析水平：在更复杂或更简单的分析水平上，这些发现是如何与现象联系起来的？
- 结果应用与综合：如果这些结果是有效的，并且可以重复，根据实验结果你可能解释或模拟哪些真实的心理现象？以本研究为依据指导实际问题可靠吗？

这些问题的答案就是你的研究的主要贡献，同时也向该领域以外的读者证明为什么他们应当关注你的研究发现。读者有权获得清楚、明确和直接的答案。” (p.27)

参考文献

我们在前面已经介绍了引用参考文献的程序，主要包括引用文章作者的姓名和出版年份。在讨论部分结束后，另起一页以“参考文献”为居中标题开始参考文献部分的写作（详见研究报告写作范文）。作者要完整地列出每一条文献。“论文中的数据是用于支持解释和结论的，而参考文献用于支持文献中的论点”《出版手册》(p.28)。

参考文献主要包括期刊文章、书籍、编著中的章节和互联网资源四种类型，这四种类型文献足以满足作者写研究报告的需要。表 14.2 列出了每种类型参考文献的写作格式。如表 14.2 所示，除了首行之外，每条参考文献的其他行都要缩进。

表 14.2 参考文献的写作格式

期刊文章
Loftus, E. F., & Burns, T. E. (1982). Mental shock can produce retrograde amnesia. <i>Memory & Cognition</i> , 10, 318-323.
书 籍
Posavac, E. J., & Carey, R. G. (2003). <i>Program evaluation</i> (6th ed.). Englewood Cliffs, NJ: Prentice Hall.
编著中的章节
Weiss, J. M. (1997). Psychological and behavioral influences on gastrointestinal lesions in animal models. In J. D. Maser & M. E. P. Seligman (Eds.), <i>Psychopathology: Experimental models</i> (pp.232-269). San Francisco: W. H. Freeman.
电子期刊文章
Kirsch, E., & Sapirstein, G. (1988). Listening to prozac but hearing placebo: A meta-analysis of antidepressant medication. <i>Prevention and Treatment</i> [On-line serial], 1. Available: journals.apa.org/prevention/volume1/pre0010002a.html.

与文稿其他部分写作一样,参考文献题目与正文之间也是间隔两倍的行距。引用期刊文章,其参考文献的写作内容包括作者、出版日期、文章题目、杂志名称、卷数和起止页码。引用书籍,其写作内容包括作者、出版日期、书名、发行城市和发行单位。引用编著中的章节,其写作内容包括章节作者、日期、章节题目、本书编者的姓名、书名、本章的起止页码、本书的发行城市和发行单位。如果引用的文献来源于电子期刊,要写清网址。

如果引用的信息来源于互联网,特别是某个网站,一定要详细写明其来源,这有助于读者在网上查找这一信息。然而,需要注意的是网站内容会定期更新,所以,读者有时无法根据你提供的来源找到这些信息。因此,哈纳克和克莱蓬格(Harnack & Kleppinger, 1997)建议你在写参考文献时要标明进入这个网站的日期,例如,在引用文献之后用一个括号注明浏览该网站的日期。你可以登陆美国心理学会网站[www.apastyle.org],查看引用互联网资源的最新写作格式以及关于美国心理学会的写作风格、出版政策和出版程序方面的最新变动。

如果参考文献写作格式基本正确,并且各条文献的顺序经过认真编排,就会大大减轻读者的阅读烦恼。根据每篇文献第一作者的姓氏在字母表中的位置对所有文献排序,如果两篇文章的作者相同,根据两篇文章发表年份从早期到近期排序。

附录

学生提交的研究报告中经常会有附录,而公开发表的文章中一般不加附录。学生在投稿时,要在参考文献之后加上附录,每页上只放一个附录。(注意:讲师可能还要求你交其他附录,例如,原始数据,统计分析工作表,或者是打印出来的数据分析,也可能是一份口头指导语的打印稿,或者是一系列具体的实验材料。)用字母(例如,A, B, C等)编排所有附录,当文本中涉及某个附录时一定要加上这个附录的字母编号。例如,你可能会这样写:“指导语详见附录A。”

作者注释

《出版手册》中简单地描述了作者注释部分的写作内容:

作者注释部分的写作内容包括(1)注明各作者的所属单位,(2)标明资金来源,(3)向同事对本研究的贡献和一些私人帮助致谢,(4)写明这篇文章的通讯作者,以方便有兴趣的读者与他就这篇文章进行信息交流。

作者注释要另起一页呈现,以“作者注释”为居中标题,位于参考文献结束之后,如果有附录要把作者注释放在附录之后(详见研究报告写作范文)。

脚注

由于脚注不可用于注明参考文献,所以在期刊文章中很少有脚注,即使在学

生的研究报告也很少有脚注。如果文本中出现脚注,其类型有两种:内容脚注和版权许可脚注。内容脚注用于补充或详述文本材料。版权许可脚注用于注明大篇幅引用内容的出处。在文本中,所有脚注(包括两种类型脚注)按阿拉伯数字连续排序。在打印稿或电子文稿中,脚注要在参考文献部分结束后另起一页呈现,以“脚注”为居中标题(详见研究报告写作范文)。

文稿各页的排序

研究报告各页应该按下面的顺序连续编排:

标题页

摘要

文本(引言、方法、结果和讨论)

参考文献(另起一页)

*附录(另起一页)

作者注释(另起一页)

*脚注(一起排序,另起一页)

表格(另起一页)

图表说明(一起排序,另起一页)

图表(每页只呈现一个图表)

*表示期刊文章中一般不会显示。

口头报告

心理学家会定期参加一些专业性的学术会议,并且会针对自己的研究做一个简短的口头报告。同样,学生也可能在课堂上、院系研究论坛上(许多班级共同参与)或者是研究生会议上对他们的研究进行口头报告。所有这些场合都有一个共同要求——口头报告的时间通常不超过10~15分钟。想在有限的时间内详细介绍一篇期刊文章的所有内容是不可能的。一般来说,《出版手册》认为“口头报告的材料在详细程度、组织结构和呈现方式上都与书面材料不同”(p.329)。要想读者能够听懂你的研究,《出版手册》建议你:

省略研究过程的大部分细节,因为听者无法达到读者阅读文字时那么专注。听者想了解的是(1)你研究了什么,为什么要研究它;(2)你是如何研究的(大概介绍一下);(3)你发现了什么研究结果;(4)如何在实际中应用结果。做口头报告要有一个意识:让听众知道主题是什么,并激发他们的兴趣。如果想让同行研究者了解详细内容,你可以为他们提供书面论文的复本。(pp.329-330)

作者所在学校生物系的一位同事提出了准备口头报告的五条原则,学生可以根据这些原则准备课堂上的口头报告。如同所有精辟的格言一样,这五条原则听

起来很简单,但在实践中一些经验丰富的研究者也常会忽视它们。

口头报告的五条原则

- 不要奢望在10分钟内报告出所有内容。克制这种幻想的办法就是将你的报告内容限于一两个要点。哪些是“最重要的信息”?口头报告的时间很短,因此,你不可能将支持这些主要观点的证据都报告出来,更不可能有时间讨论一些次要的问题。
- 培养一种好的台风。要想培养出好的台风,首先需要具备良好的演讲技巧。很多人需要照稿读,如果能脱稿讲效果会更好。有些人一面对观众,讲话速度就会加快,特别是阅读书面材料时。在口头报告时应该以中等速率,用简单、直接的句子阐述研究结论,最重要的是做到声音洪亮,吐字清楚。
- 利用有效的视觉辅助工具。有效的视觉辅助工具可以帮助观众理解你所讲的内容。幻灯片、字幕和电子投影设备都能保证观众在一定距离内看清楚你所讲的内容。熟练应用视听辅助工具也是一种技能,经常使用就能掌握这种技能。无论采用哪种形式的视觉辅助工具,一定要保证呈现的内容与你解释的内容一致。
- 留出听众提问的时间。一定要牢记第四条原则,留出听众提问的时间。原因主要有两个:一是大多数学术会议和课堂报告都要求作者在口头报告时有时间与听众交流;另一个是给听众提问的机会,他们才会认真听讲和积极思考,尽管他们提的问题有时会很尖锐。
- 找一个人帮助你进行口头报告练习。最后一条原则可能是最重要的。在一位听众前练习报告要比你自己单独一遍又一遍的重复练习效果好得多。个人重复练习只能帮助你准备“彩排”,无法代替在他人面前进行的演习。

以上五条是准备口头报告的训练原则。然而,掌握口头报告技能的最好方式还是尽可能多次参加“游戏情境”下的口头报告。虽然,只通过练习可能无法达到最佳效果,但练习却是我们所知的提高口头报告能力的最好办法。

研究计划

在本章的最后,我们要再次讨论研究计划的写作。在本章开始时我们就提到,研究者必须要为他们的研究申请资金,他们可以向私人机构或政府部门提交申请计划。学生上研究方法课时通常也要提交研究计划,以此来阐述他们将要进行的研究。即使不用提交书面计划,聪明的研究者还是会提前认真思考与研究相关的文献、可能出现的实际困难、收集到的数据可以做哪些统计分析以及如何解释预期结果等问题。这个最初的周密思考可以帮助你形成一个可行的研究计划,并且能够对其进行合理的分析与解释。

写研究计划的目的是保证可行的实验设计在实施后能得到一个有显著科学价值的可解释的实证结果。无论准备多么充分,如果没有研究计划,就无法保证能

得出重要的结果。研究者在他们的学术研究之初就学会了墨菲定律。墨菲定律的本质就是“有可能犯错误的事情做了就一定会犯错误。”为了避免那些可以避免的问题出现,写研究计划还是很值得的。

书面研究计划与期刊文章的写作格式基本相似,只是某些部分的标题略有不同。研究计划包括以下几个主要部分:

引言

方法

预期结果与数据分析计划

结论

参考文献

附录

给研究评估委员会的材料

研究计划中没有摘要。与期刊文章相比,研究计划的引言部分可以更广泛地引用相关文献。研究计划的表述问题和提出假设的逻辑与期刊文章相同。同时,研究计划的方法部分尽可能与将要进行的研究过程一致。因此,研究计划与研究报告在方法部分的写作格式和内容上都一样。

研究计划有一部分是“预期结果和数据分析计划”,这部分需要对预期研究结果进行简单讨论。在大多数情况下,实际会得出什么样的结果还不一定。然而,你总是有一些关于将会得出什么结果的想法(其形式是假设或预期)。预期结果中可能还需要用表格或图示来表达你期望得到的结果。在写作研究计划时要着重强调重要的结果。预期结果的数据分析计划也属于研究计划的一部分内容。例如,如果你正在考虑进行一个复合设计,那么,你要测验哪些效应、使用什么统计方法来检验这些效应等都要写清楚。同时,你还要想到与预期结果不一致的一些合理偏差,如果结果与研究假设不一致,你要给出合理解释。研究计划的最后一部分是结论,在这部分你要简单阐述结论和预期结果的应用。

参考文献部分的写作形式与最终提交的研究报告的形式完全一样。附录用于补充研究计划,其写作内容包括在随后实验中使用的一系列材料。附录一般包括指导语、实验所使用的仪器类型、实验材料或材料说明等。例如,如果你正在进行的研究中要求学生记忆一系列单词,附录应包括:随机生成的单词序列、用于呈现材料的仪器类型、各种实验条件下的指导语,以及各种随机化条件。

最后,研究计划中还包括向研究评估委员会(IRB)等类似机构提交的材料,这些委员会负责评审研究中的伦理道德问题(详见第3章)。

研究报告写作范文

First few words of title appear before page number

Running head: FALSE MEMORIES

Type flush with left margin.

Type running head all caps.

5 spaces

False Recall 1

Number pages consecutively, beginning with title page.

False Recall and Recognition of List Items Following Delay

Jeanne S. Zechmeister and Eugene B. Zechmeister

Loyola University of Chicago

Use upper and lower case for title, name, and affiliation (centered).

Note:
Use one-inch margins at the top, bottom, right, and left of all pages.

No paragraph indentation for abstract.

Abstract should be a single paragraph not exceeding 120 words.

Center, do not italicize

↓
Abstract

False Recall 2

Whether false memories persist over a delay was investigated using H. L. Roediger and K. B. McDermott's (1995) false memory procedure. College students ($N = 19$) studied 12 15-item lists and either recalled items immediately or following a 90-s delay. A first recognition test immediately followed recall of lists; a second test occurred 2 days later. The proportion of correctly recalled list items decreased with the 90-s delay, whereas the proportion of falsely recalled critical items increased. The overall proportion of items recognized as "old" was greater for false critical items than for correct list items, but both false recognition and correct recognition decreased over the 2-day period. These findings are compared to other studies that examined false memories following delays.

Note:

Double-space between all lines of the manuscript. There should be no more than 27 lines of text on each page.

When using a word processor, use left justification only (not full). Use italics for statistical symbols and infrequently otherwise (underline if using a typewriter), and use one font size throughout the manuscript.

Leave one space at end of all sentences, and after all commas, colons, and semicolons.

Identify all authors (surnames only) the first time you cite a source (five or fewer authors).

List multiple sources in alphabetical order, separated by a semicolon.

Indent every paragraph using the tab key (set at five to seven spaces or 1/2 inch).

Abbreviated terms are first spelled out, followed by the abbreviation in parentheses.

Title appears centered on page 3, with first letter of major words in caps.

False Recall 3

False Recall and Recognition of List Items Following Delay

Roediger and McDermott (1995) captured the attention of memory researchers by reminding us of the importance of examining participants' errors in recall and that these errors can sometimes be quite predictable. Moreover, these "false memories" are relevant to ongoing debates about false memories that may occur in more natural settings (e.g., clinics, courtrooms). Roediger and McDermott, and many others, have demonstrated that false memories are easily created and that participants' phenomenological experience of these memories mimics their memories for actual events (e.g., Mather, Henkel, & Johnson, 1997; Norman & Schacter, 1997; Payne, Elie, Blackwell, & Neuschatz, 1996). The purpose of the present study was to replicate Roediger and McDermott's (1995) findings and to extend their findings by examining the persistence of false memories following delays. If false memories persist following delays, important implications exist regarding courtroom testimony and recovered memories in therapy. For example, memory retrieval in these cases follows varying periods of delay and the accuracy of these memories may thus be questioned.

In this false memory paradigm word lists are constructed such that each word is highly associated with a critical, not-presented word. For example, participants may hear the words *bed*, *rest*, *awake*, *tired*, *dream*, etc., but never the critical word, *sleep*. False recall and recognition for the word *sleep* is quite high. Underwood (1965) proposed that an implicit associative response (IAR) occurs when participants study an associated word. For example, when participants hear "bed" they may think of "sleep" (the IAR). With the next word, "rest," they again may experience "sleep" as an IAR, and so on. When their memory is tested participants may have difficulty discriminating the

Use Latin abbreviations (such as *e.g.* and *i.e.*) only in parentheses.

Use numbers to express amounts 10 and above, and when numbers are grouped for comparison with numbers 10 and above.

An "s" (with no period) is the abbreviation for second.

Use number for units of time or measure.

Note: Use past tense to describe the study.

Flush with left margin, italicized, only first letter in caps.

Always use words to express numbers that begin a sentence.

False Recall 4

IAR from the presented items, resulting in false recall or recognition. An alternative account relies on the extent to which the lists of associated words activate a schema for the words (e.g., a "sleep" schema). According to this account, recall after a delay may be based on memory for the schema rather than the actual list presentation.

In the present study participants listened to 12 lists comprised of 15 words that were associated to a critical, not-presented word. Participants were randomly assigned to recall 4 of the 12 lists in either an immediate-recall condition or following a 90-s delay in a single trial of free recall. When not recalling words (and during the 90-s delays) participants completed distractor tasks consisting of math problems. Additionally, two recognition tests were administered. The first recognition test followed the presentation and recall of the 15 lists; the second was 2 days later.

We hypothesized that the proportions of correct recall and false recall would not differ in the immediate-recall condition because of participants' inability to differentiate their memory for the presented words from the IARs generated for the critical items. In the delayed-recall condition, we expected the proportion of false items recalled to be greater than the proportion of correct items recalled because of participants' activation of a schema for the list to aid recall. This activation was assumed to lead to recall of the highly associated, critical item. Furthermore, we anticipated that correct recall would decrease over time but false recall would increase with the delay. Similar predictions were made for participants' recognition of correct list items and false critical items.

Method

Center, no italics, only the first letter in caps.

Participants

Nineteen students (16 women, 3 men) from a research methods in psychology

Note: Method, Results, and Discussion sections do not begin on new page (unless coincidentally).

Note: "Data" is plural.

Indented, and italicized paragraph heading, followed by a period. Only first letter of first word in caps.

Use words to express numbers less than 10.

The word "participants" is preferred over "subjects."

class at Loyola University of Chicago participated as part of a class project. Data for one student were omitted from all analyses because the student attempted recall for the wrong lists; recognition data for two students were omitted from analyses because they missed the second class session in which the experiment was conducted.

Materials

Word lists. We used 12 15-item word lists from Roediger and McDermott's (1995) study. They constructed 24 lists by selecting 15 words that were highly associated with a critical word that was not presented. Words were ordered such that the strongest associates were presented first, followed by more weakly associated words. For example, the list for the critical word, *needle*, was *thread, pin, eye, sewing, sharp, point, prick, thimble, haystack, thorn, hurt, injection, syringe, cloth, and knitting*.

Six lists in the present study were identified by Roediger and McDermott as having a high likelihood of creating false memories (identified by the critical item): *chair, mountain, needle, rough, sleep, sweet*. The other six lists used in the present study for recall were randomly selected from their remaining lists (identified by the critical item): *bread, cold, doctor, foot, music, slow*.

Participants recalled words from 4 of the 12 lists in a single trial of free recall. In order to rule out the possibility that proportions of correct and false recall would be determined by the particular lists used for the recall task, the four recall lists were counterbalanced. The 12 lists were randomly divided into three sets of four lists, with the provision that two lists in each set have high likelihoods for creating false memories. Participants were randomly assigned to recall words for four lists. Seven participants recalled the chair, bread, slow, and sleep lists; seven participants recalled the needle,

False Recall 6

foot, doctor, and rough lists, and four participants recalled the sweet, music, cold, and mountain lists.

Distractor task. Participants completed a distractor task consisting of math problems when not recalling words and during the 90-s delay. We generated 20 pages of math problems. Each page had six complex math problems involving addition, subtraction, multiplication, and division.

Recall packets. We created six different packets to manipulate the recall condition (immediate, delayed) and counterbalance the lists used for recall. Four pages presented recall instructions based on those used by Roediger and McDermott (1995): "Please write down as many words from the list that you remember. A good strategy is to write down the last few items first, then the rest of the words in any order." Twenty pages had math problems. Between each page was a colored sheet of paper, which masked the participants' next task (recall or math problems). The order of the 24 pages depended on whether participants were in the immediate- or delayed-recall condition and which list set participants were assigned to recall. For example, the "music" list was presented first. Following the presentation of the list, participants assigned to recall the music list in the immediate condition opened their packet to a page with recall instructions; all other participants opened their packet to a page of math problems. After 90 s, a tone signaled participants to turn to the next white page (following the colored sheet). If assigned to recall the music list in the delayed-recall condition, participants' next sheet had recall instructions; all other participants received math problems. This sequence was followed for the remaining 11 word lists.

Recognition tests. Participants completed two different recognition tests, the first

Place periods or commas within quotation marks.

False Recall 7

followed the presentation and recall of lists (immediate), the second was 2 days later (delayed). Because three list sets were used for recall (to counterbalance recall lists across participants), we created three versions of the first recognition test (session 1) and three versions of the second recognition test (session 2). Following Roediger and McDermott's terminology, recognition-test items that were presented during the recall portion of the study are referred to as "studied," and new distractor items are referred to as "not studied." Each 72-item recognition test consisted of six different types of items: 10 Studied/Recalled items (5 each from two of the four lists participants were asked to recall); 2 corresponding critical items for these lists; 20 Studied/Not Recalled items (5 each from four of the eight lists participants did not recall); 4 corresponding critical items for these lists; 30 Not Studied Distractor items (5 each from six lists that were not presented); and 6 corresponding critical items for the distractor lists (these lists also were from Roediger & McDermott, 1995). The five items selected from the word lists for each recognition test were the first, third, fifth, seventh, and ninth words. The 72 words on each list were randomly ordered. Participants made recognition judgments using a 4-point scale (1 = sure new, 2 = probably new, 3 = probably old, and 4 = sure old).

Procedure

A 2 (condition) X 2 (memory type) mixed factorial design was used for the recall portion of the experiment. Students were randomly assigned to either the immediate-recall condition or the delayed-recall condition; thus, the condition variable was manipulated as a random groups design. Memory type was a repeated measures variable with two levels: correct and false memory. The recognition portion of the experiment was a repeated measures independent variable. Participants completed the

A comma separates authors surnames and year of publication when citation is within parentheses.

Use ampersand to indicate "and" for references within parentheses.

Italicize the anchors of a scale.

Numbers are used to indicate the levels of a design. Include one space to left and right of multiplication sign.

"min" (without a period) is the abbreviation for minute.

first recognition test immediately following the presentation and recall of word lists and completed the delayed recognition test 2 days later. The dependent variables were proportion of correct and false recall and proportion of correct and false recognition.

The experiment was conducted during two scheduled class periods. In the first session, lasting 50 min, the experimenter gave students a consent form and a packet that corresponded to their condition when they entered the class (students were randomly assigned to conditions prior to their arrival). A participant number appeared on top of the packet. Instructions requested students not to open their packet. After students read and signed their consent form, they listened to tape-recorded instructions for the study. Following this, the experimenter stopped the tape recorder and asked if there were any questions. Tape-recorded word lists were then presented. The random order of the lists was as follows (identified by the critical item): music, needle, sleep, doctor, slow, mountain, rough, bread, sweet, cold, chair, and foot. Words were presented at the rate of approximately one word every 1½ s. Following the list presentation, participants heard a tone informing them to turn to the next white page. They then either recalled list items or completed math problems for 90 s. A tone then signaled participants to turn to the next white page. This page instructed them to recall the list items or complete math problems. This procedure was followed for all 12 lists.

After the study and recall phase of the experiment, the experimenter distributed the first recognition test and asked participants to record their participant number on the top of the recognition test. After completing the recognition test, participants were asked to write their participant number in their notebook; they were then dismissed. At the beginning of the next class session (2 days later), participants reported their number

Note: Words are not hyphenated at the end of a line.

Note: Always double-space. Do not insert extra spaces between sections.

Center, no italics.

False Recall 9

and they were given the corresponding second recognition test. Each recognition test took approximately 5 min to complete. The instructor then fully debriefed participants regarding the purpose of the experiment.

→ Results

Results are presented first for participants' recall and then for their recognition of list items. The dependent variables were proportion of correct recall/recognition and proportion of false recall/recognition. Proportion of correct recall was determined by counting the number of words participants correctly recalled and dividing by 60 (the total number of list items they were asked to recall). Proportion of false recall was determined by counting the number of critical items falsely recalled and dividing by 4 (the total number of critical items for the four lists participants recalled).

Similar proportions were computed for recognition data. For ease of comparisons, participants' recognition responses were collapsed such that ratings of 1 or 2 indicated judgments that the word was "new" (i.e., not presented during the recall portion of the study), and ratings of 3 or 4 indicated the word was "old" (i.e., presented during the recall portion). The number of items judged "old" for each type of word (e.g., studied, recalled; not studied distractor) was counted and divided by the total number of words of that type presented on the recognition test.

Recall

Mean proportions for correct and false recall were entered into a 2 (condition) X 2 (memory type) mixed-design analysis of variance (ANOVA). Across the two recall conditions the proportion of falsely recalled critical items ($M = .60$, $SD = .23$) tended to be greater than the proportion of correctly recalled list items ($M = .48$, $SD = .12$), $F(1,$

M is the abbreviation for arithmetic mean.

SD is the abbreviation for standard deviation.

Is the abbreviation
for Mean Square
Error.

Italicize all
statistical terms

Do not capitalize
names of conditions.

Footnotes appear on a
separate page near
end of manuscript
(not at bottom
of the page).

Do not abbreviate
"Figure." Figure
caption and figure
appear on separate
pages at end of
manuscript.

Degrees of freedom
for statistical tests
are reported in
parentheses (with no
space after statistical
term).

False Recall 10

16) = 4.38, $p = .053$ ($MSE = .03$, effect size $r = .47$).¹ The observed power to detect this large effect with the small sample size ($n = 9$ per condition), at $\alpha = .05$, was estimated to be only .50. The main effect of condition was not significant, $F(1, 16) = .10$, $p > .05$ ($MSE = .02$, $r = .10$, power = .06). As predicted, however, recall condition and memory type interacted to influence participants' recall, $F(1, 16) = 9.50$, $p < .01$ ($MSE = .03$, $r = .61$). Simple main effects analyses revealed that recall differed in the immediate- and delayed-recall conditions in the predicted direction (see Figure 1). Participants in the immediate-recall condition did not differ in their proportions of correct recall and false recall, $F(1, 8) = .41$, $p > .05$ ($MSE = .036$, $r = .22$, power = .09); however, participants in the delayed-recall condition were more likely to recall a greater proportion of false critical items than correct list items, $F(1, 8) = 16.57$, $p < .005$ ($MSE = .024$, $r = .82$).²

A second set of simple main effects analyses examined the effect of delay for each type of recall. As expected, the proportion of correctly recalled list items decreased following delay and the proportion of false recall increased following delay (see Figure 1). For correct list items, participants in the delayed-recall condition recalled fewer items than participants who recalled list items immediately, $F(1, 16) = 15.42$, $p < .005$ ($MSE = .008$). The effect size (r) was .70, indicating a very large effect of immediate- vs. delayed-recall on participants' correct recall (see Cohen, 1988). In contrast, the proportion of falsely recalled critical items increased following the 90-s delay relative to the immediate-recall condition, $F(1, 16) = 3.77$, $p = .07$. The effect size (r) was .44; however, the estimated power to detect this large effect was only .45 ($\alpha = .05$).

Recognition

Across two testing sessions separated by 2 days, we examined the proportion of items recognized as "old" as a function of whether items were correct list items or false critical items and whether the items were recalled (vs. not recalled) during the recall portion of the experiment (see Table 1).³ Distractor items were excluded from the analyses because the proportion of these items judged "old" was generally low across item type and sessions, $M = .18$ ($SD = .12$). Main effects were observed for the time, memory type, and recalled/not recalled variables. As expected, memory for the items decreased over the 2-day period, as the proportion of all items recognized as "old" decreased at Time 2 ($M = .71$, $SD = .14$) relative to Time 1 ($M = .87$, $SD = .11$), $F(1, 14) = 16.40$, $p < .005$ ($MSE = .04$, $r = .73$). Additionally, the main effect of memory type indicated that participants rated a higher proportion of false critical items "old" ($M = .85$, $SD = .11$) than correct list items ($M = .73$, $SD = .13$), $F(1, 14) = 14.72$, $p < .005$ ($MSE = .026$, $r = .71$). Thus, in terms of proportions, false recognition was greater than correct recognition at Time 1 and Time 2. Finally, the proportion of items judged "old" was greater for lists participants were asked to recall ($M = .87$, $SD = .09$) relative to lists they did not recall ($M = .71$, $SD = .14$), $F(1, 14) = 23.57$, $p < .001$ ($MSE = .032$, $r = .79$). This is the "testing effect" identified by Roediger and McDermott (1995), in which rehearsal during the recall portion of the experiment increases subsequent recognition.

There was also a Time X Recalled/Not Recalled interaction, $F(1, 14) = 7.49$, $p < .05$ ($MSE = .024$, $r = .59$). This interaction indicated that recognition of words for recalled lists was slightly higher at Time 1 ($M = .90$, $SD = .07$) relative to Time 2 ($M = .83$, $SD = .16$), $F(1, 15) = 3.22$, $p < .10$ ($MSE = .013$, $r = .42$, power = .39). In contrast,

Table appears on separate page near end of manuscript.

Note use of spaces and punctuation when typing statistics.

Variable names in caps when they appear with a multiplication sign.

Capitalize nouns that are followed by a number.

False Recall 12

recognition of words from not-recalled lists dropped significantly at Time 2 ($M = .60$, $SD = .18$) relative to Time 1 ($M = .83$, $SD = .18$), $F(1, 15) = 18.83$, $p < .005$ ($MSE = .023$, $r = .74$). Thus, rehearsal of the lists (through recall) increased the proportion of correct items and false critical items recognized as "old" even after a 2-day period, relative to lists that were not rehearsed.

As noted above, we observed differential effects of delay on the proportion recall of correct list items and false critical items. However, a similar effect was not observed for recognition data, as the Time X Memory Type interaction was not statistically significant, $F(1, 14) = .028$, $p > .05$ ($MSE = .014$). The effect size, r , for this interaction was quite small (.04, power = .05). Thus, the 2-day delay between the recognition tests resulted in similar decreases in the proportions of correct list items and false critical items that were recognized as "old."

Note: Make sure you use the "greater than" or "less than" sign correctly.

Center, no italics.

Discussion

Participants recalled list items in a single trial of free recall immediately after list presentation or following a 90-s filled delay. The proportion of correctly recalled list items decreased as a function of delay, but the proportion of false critical items recalled increased following the 90-s delay. Participants' recognition of correct list items and false critical items was tested following a 2-day delay. In terms of proportions, false recognition of critical items was greater than correct recognition of list items at Time 1 and Time 2, but recognition for both types of items decreased over the 2-day period. A testing effect occurred such that correct and false recognition were greater for lists participants had attempted recall relative to lists not recalled. The effects of rehearsal persisted over the 2-day period. In general, delay produced large effect sizes for both

Year of publication
in parentheses after
author's surname.

Do not use
contractions such as
"wasn't."

Year of publication
is not repeated
within a paragraph
if it is clear that
the same study is
being described.

Year of
publication appears
in parentheses
even when
reference has
appeared
previously in
manuscript.

Use "et al."
following surname
of first author for work with three or more authors.

Note: First citation to this work lists all authors (see p. 3).

False Recall 13

recall and recognition.

McDermott (1996) used a 30-s filled delay to examine the effects of delay on participants' recall of correct list items and false critical items. Although she observed a decrease in correct recall following delay, her participants' false recall of critical items was not affected by delay. In contrast, we observed an increase in the proportion of false critical items recalled following a 90-s delay. One reason for this discrepancy may be derived from our differing findings for correct list recall. In McDermott's study the proportion of correct list items recalled following a 30-s delay was .50; in our study, the proportion of correct recall following a 90-s delay was .39. To the extent that a longer delay increased participants' forgetting of correct list items in our study, they may have had to rely more on schema-based recall, which we hypothesized to increase the occurrence of false memories. Alternatively, discrimination between IARs and memory for list items may become more difficult with longer delays. As a result, greater proportions of false recall would be expected (as was observed). This explanation is corroborated by McDermott's findings that after a 2-day delay, the proportion of falsely recalled critical items was greater than the proportion of correctly recalled items (similar to our findings for recognition). Thus, in both studies, participants demonstrated high levels of false recall that persisted over time.

Payne et al. (1996) examined participants' recognition memory for correct list items and false critical items in immediate (2 min) and delayed (24 hr) conditions. Similar to our recognition data, Payne et al. observed that false critical items were recognized with higher likelihood at both testing sessions than correct list items. However, they also observed an interaction between item type (correct, false) and

"hr" (with no period) is the abbreviation for hour.

retention (immediate, delayed) for their recognition data that we did not. They observed more forgetting for correct list items than false critical items over the 24-hr period, whereas over our 2-day period we observed similar decreases for correct list items and false critical items. Differences in the retention interval (24 vs. 48 hr), the number of word lists (16 vs. 12), or other procedural differences may account for this discrepancy.

The findings of this study and those of other recent studies indicate that false memories are frequent and persist following delays. Moreover, there is accumulating evidence to suggest that false memories for highly associated information may exceed correct memory over time. These data are limited to individuals' recall for highly associated list items. Important areas for future research include false memories for natural categories such as people and events. Also, it is not clear to what extent individuals' emotions may affect correct and false recall. This question is particularly relevant to the issue of false memory in the emotionally charged situations of eyewitness testimony and therapy.

Note: References section begins on a new page immediately after the Discussion section.

Italicize titles of books. Only first letter of first word and word following a colon in caps.

Use hanging indents for each reference entry. Center, no italics.

Publisher of book.

Use ampersand before last author's name.

First letter of questionnaire names in caps.

Use initials for first and middle names.

Period after year of publication (in parentheses).

Page numbers.

Use lower case for titles of articles and book chapters (except first word and word after a colon).

Italicize journal name and volume number. Capitalize first letter of major words in journal name. Use comma after journal name and volume number.

False Recall 15

References

Cohen, J. (1988). *Statistical power analysis for behavioral sciences* (2nd ed.). New York: Academic Press.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Mather, M., Henkel, L. A., & Johnson, M. J. (1997). Evaluating characteristics of false memories: Remember/know judgments and Memory Characteristics Questionnaire compared. *Memory & Cognition*, 25, 826-837.

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language*, 35, 212-230.

Norman, K., & Schacter, D. L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition*, 25, 838-848.

Payne, D. G., Elie, C. J., Blackwell, J. M., & Neuschatz, J. S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language*, 35, 261-285.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 803-814.

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology*, 70, 122-129.

Note: List references in alphabetical order using first author's surname. Double-space references, no extra space between references.

Authors' department affiliations in first paragraph.

Acknowledge people who assisted with the research and manuscript preparation.

Last paragraph in Author Note specifies author's address for correspondence. Use this form exactly with your name and address.

Center, no italics

↓
Author Note

False Recall 16

Jeanne S. Zechmeister, Department of Psychology; Eugene B. Zechmeister,
Department of Psychology.

Portions of this research were presented at the Annual Meeting of the Psychonomic Society, November 1996. The authors thank the students in the first author's Research Methods in Psychology course at Loyola University of Chicago who participated in and discussed this study, and John J. Shaughnessy for his suggestions for this manuscript.

Send correspondence concerning this manuscript to Jeanne S. Zechmeister,
Department of Psychology, Loyola University of Chicago, 6525 N. Sheridan Road,
Chicago, IL 60626; e-mail: jzechme@luc.edu.

81	False Recall	Footnotes False Recall 17 ¹ The effect size, r, was derived from ANOVA output that presented η^2 as a measure of effect size. η^2 represents the proportion of variance in the dependent variable that is accounted for by the independent variable. η^2 is analogous to r^2 in analyses involving one df for the effect. The square root of η^2 represents r, the magnitude of linear relationship between two variables. Cohen (1988) offers the following guidelines for interpreting the size of an effect using r: small effect, $r = .10$; medium effect, $r = .25$; large effect, $r = .40$. ² Simple main effect analyses were calculated by treating the comparison as a one-way ANOVA, as recommended by Keppel (1991) for mixed designs. ³ We also examined whether immediate- vs. delayed-recall conditions influenced recognition in the ANOVA. Because this variable did not produce any statistically significant effects in the recognition analyses, it will not be discussed further.	
71	(01.)	(AS.) .105) Note: Use footnotes sparingly in manuscript	

Type flush with left margin.

Line separates title from body of table.

Center column headings over appropriate columns.

Line separates headings from table data.

Line at end of table data.

Table "notes" are typed flush with left margin.

Table title, italicized, with first letters of major words in caps.

False Recall 18

Table 1

Mean Proportions for Recognizing Items as "Old" at Time 1 and Time 2 (48-hr Delay)

	Studied and recalled	Studied and not recalled	Distractor
Correct list items			
Time 1	.84 (.12)	.79 (.17)	.19 (.12)
Time 2	.76 (.22)	.56 (.16)	.18 (.16)
False critical items			
Time 1	.97 (.12)	.88 (.20)	.28 (.13)
Time 2	.91 (.20)	.64 (.24)	.17 (.10)

Note. Standard deviations are in parentheses. "Recalled" and "Not Recalled" refer to whether participants were asked to recall list items at Time 1, not whether these items were produced by participants during recall.

Note: Use double-space in table. Place only one table on a page.

Flush with left margin. Do not abbreviate "Figure."

Center, no italics.

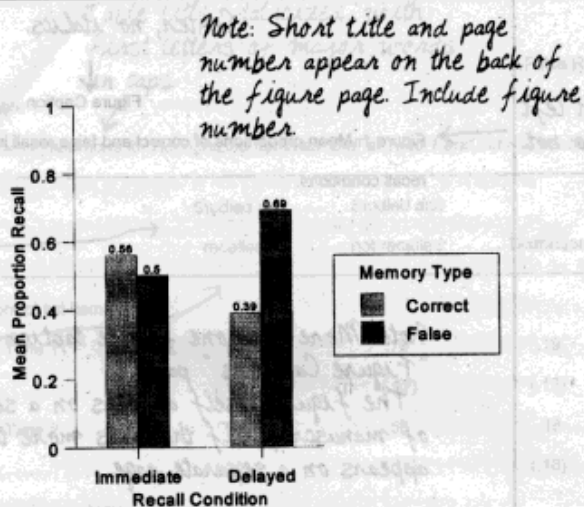
False Recall 19

Figure Caption

→ Figure 1. Mean proportions of correct and false recall in the immediate- and delayed-recall conditions.

Note: More than one figure caption may appear on a "Figure Captions" page.

The figure itself appears on a separate page at end of manuscript. If there is more than one figure, each appears on a separate page.



附录

部分t分布值表

部分t分布值表

统计表

附录大纲

表 A.1 随机数字表

表 A.2 部分t分布值表

表 A.3 F分布的临界值表

表 A.4 独立组均值比较检验力表 $\alpha=0.05$ (双侧检验)表 A.5 独立组变异分析检验力表 $\alpha=0.05$

0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10.0	10.1	10.2	10.3	10.4	10.5	10.6	10.7	10.8	10.9	11.0	11.1	11.2	11.3	11.4	11.5	11.6	11.7	11.8	11.9	12.0	12.1	12.2	12.3	12.4	12.5	12.6	12.7	12.8	12.9	13.0	13.1	13.2	13.3	13.4	13.5	13.6	13.7	13.8	13.9	14.0	14.1	14.2	14.3	14.4	14.5	14.6	14.7	14.8	14.9	15.0	15.1	15.2	15.3	15.4	15.5	15.6	15.7	15.8	15.9	16.0	16.1	16.2	16.3	16.4	16.5	16.6	16.7	16.8	16.9	17.0	17.1	17.2	17.3	17.4	17.5	17.6	17.7	17.8	17.9	18.0	18.1	18.2	18.3	18.4	18.5	18.6	18.7	18.8	18.9	19.0	19.1	19.2	19.3	19.4	19.5	19.6	19.7	19.8	19.9	20.0	20.1	20.2	20.3	20.4	20.5	20.6	20.7	20.8	20.9	21.0	21.1	21.2	21.3	21.4	21.5	21.6	21.7	21.8	21.9	22.0	22.1	22.2	22.3	22.4	22.5	22.6	22.7	22.8	22.9	23.0	23.1	23.2	23.3	23.4	23.5	23.6	23.7	23.8	23.9	24.0	24.1	24.2	24.3	24.4	24.5	24.6	24.7	24.8	24.9	25.0	25.1	25.2	25.3	25.4	25.5	25.6	25.7	25.8	25.9	26.0	26.1	26.2	26.3	26.4	26.5	26.6	26.7	26.8	26.9	27.0	27.1	27.2	27.3	27.4	27.5	27.6	27.7	27.8	27.9	28.0	28.1	28.2	28.3	28.4	28.5	28.6	28.7	28.8	28.9	29.0	29.1	29.2	29.3	29.4	29.5	29.6	29.7	29.8	29.9	30.0	30.1	30.2	30.3	30.4	30.5	30.6	30.7	30.8	30.9	31.0	31.1	31.2	31.3	31.4	31.5	31.6	31.7	31.8	31.9	32.0	32.1	32.2	32.3	32.4	32.5	32.6	32.7	32.8	32.9	33.0	33.1	33.2	33.3	33.4	33.5	33.6	33.7	33.8	33.9	34.0	34.1	34.2	34.3	34.4	34.5	34.6	34.7	34.8	34.9	35.0	35.1	35.2	35.3	35.4	35.5	35.6	35.7	35.8	35.9	36.0	36.1	36.2	36.3	36.4	36.5	36.6	36.7	36.8	36.9	37.0	37.1	37.2	37.3	37.4	37.5	37.6	37.7	37.8	37.9	38.0	38.1	38.2	38.3	38.4	38.5	38.6	38.7	38.8	38.9	39.0	39.1	39.2	39.3	39.4	39.5	39.6	39.7	39.8	39.9	40.0	40.1	40.2	40.3	40.4	40.5	40.6	40.7	40.8	40.9	41.0	41.1	41.2	41.3	41.4	41.5	41.6	41.7	41.8	41.9	42.0	42.1	42.2	42.3	42.4	42.5	42.6	42.7	42.8	42.9	43.0	43.1	43.2	43.3	43.4	43.5	43.6	43.7	43.8	43.9	44.0	44.1	44.2	44.3	44.4	44.5	44.6	44.7	44.8	44.9	45.0	45.1	45.2	45.3	45.4	45.5	45.6	45.7	45.8	45.9	46.0	46.1	46.2	46.3	46.4	46.5	46.6	46.7	46.8	46.9	47.0	47.1	47.2	47.3	47.4	47.5	47.6	47.7	47.8	47.9	48.0	48.1	48.2	48.3	48.4	48.5	48.6	48.7	48.8	48.9	49.0	49.1	49.2	49.3	49.4	49.5	49.6	49.7	49.8	49.9	50.0	50.1	50.2	50.3	50.4	50.5	50.6	50.7	50.8	50.9	51.0	51.1	51.2	51.3	51.4	51.5	51.6	51.7	51.8	51.9	52.0	52.1	52.2	52.3	52.4	52.5	52.6	52.7	52.8	52.9	53.0	53.1	53.2	53.3	53.4	53.5	53.6	53.7	53.8	53.9	54.0	54.1	54.2	54.3	54.4	54.5	54.6	54.7	54.8	54.9	55.0	55.1	55.2	55.3	55.4	55.5	55.6	55.7	55.8	55.9	56.0	56.1	56.2	56.3	56.4	56.5	56.6	56.7	56.8	56.9	57.0	57.1	57.2	57.3	57.4	57.5	57.6	57.7	57.8	57.9	58.0	58.1	58.2	58.3	58.4	58.5	58.6	58.7	58.8	58.9	59.0	59.1	59.2	59.3	59.4	59.5	59.6	59.7	59.8	59.9	60.0	60.1	60.2	60.3	60.4	60.5	60.6	60.7	60.8	60.9	61.0	61.1	61.2	61.3	61.4	61.5	61.6	61.7	61.8	61.9	62.0	62.1	62.2	62.3	62.4	62.5	62.6	62.7	62.8	62.9	63.0	63.1	63.2	63.3	63.4	63.5	63.6	63.7	63.8	63.9	64.0	64.1	64.2	64.3	64.4	64.5	64.6	64.7	64.8	64.9	65.0	65.1	65.2	65.3	65.4	65.5	65.6	65.7	65.8	65.9	66.0	66.1	66.2	66.3	66.4	66.5	66.6	66.7	66.8	66.9	67.0	67.1	67.2	67.3	67.4	67.5	67.6	67.7	67.8	67.9	68.0	68.1	68.2	68.3	68.4	68.5	68.6	68.7	68.8	68.9	69.0	69.1	69.2	69.3	69.4	69.5	69.6	69.7	69.8	69.9	70.0	70.1	70.2	70.3	70.4	70.5	70.6	70.7	70.8	70.9	71.0	71.1	71.2	71.3	71.4	71.5	71.6	71.7	71.8	71.9	72.0	72.1	72.2	72.3	72.4	72.5	72.6	72.7	72.8	72.9	73.0	73.1	73.2	73.3	73.4	73.5	73.6	73.7	73.8	73.9	74.0	74.1	74.2	74.3	74.4	74.5	74.6	74.7	74.8	74.9	75.0	75.1	75.2	75.3	75.4	75.5	75.6	75.7	75.8	75.9	76.0	76.1	76.2	76.3	76.4	76.5	76.6	76.7	76.8	76.9	77.0	77.1	77.2	77.3	77.4	77.5	77.6	77.7	77.8	77.9	78.0	78.1	78.2	78.3	78.4	78.5	78.6	78.7	78.8	78.9	79.0	79.1	79.2	79.3	79.4	79.5	79.6	79.7	79.8	79.9	80.0	80.1	80.2	80.3	80.4	80.5	80.6	80.7	80.8	80.9	81.0	81.1	81.2	81.3	81.4	81.5	81.6	81.7	81.8	81.9	82.0	82.1	82.2	82.3	82.4	82.5	82.6	82.7	82.8	82.9	83.0	83.1	83.2	83.3	83.4	83.5	83.6	83.7	83.8	83.9	84.0	84.1	84.2	84.3	84.4	84.5	84.6	84.7	84.8	84.9	85.0	85.1	85.2	85.3	85.4	85.5	85.6	85.7	85.8	85.9	86.0	86.1	86.2	86.3	86.4	86.5	86.6	86.7	86.8	86.9	87.0	87.1	87.2	87.3	87.4	87.5	87.6	87.7	87.8	87.9	88.0	88.1	88.2	88.3	88.4	88.5	88.6	88.7	88.8	88.9	89.0	89.1	89.2	89.3	89.4	89.5	89.6	89.7	89.8	89.9	90.0	90.1	90.2	90.3	90.4	90.5	90.6	90.7	90.8	90.9	91.0	91.1	91.2	91.3	91.4	91.5	91.6	91.7	91.8	91.9	92.0	92.1	92.2	92.3	92.4	92.5	92.6	92.7	92.8	92.9	93.0	93.1	93.2	93.3	93.4	93.5	93.6	93.7	93.8	93.9	94.0	94.1	94.2	94.3	94.4	94.5	94.6	94.7	94.8	94.9	95.0	95.1	95.2	95.3	95.4	95.5	95.6	95.7	95.8	95.9	96.0	96.1	96.2	96.3	96.4	96.5	96.6	96.7	96.8	96.9	97.0	97.1	97.2	97.3	97.4	97.5	97.6	97.7	97.8	97.9	98.0	98.1	98.2	98.3	98.4	98.5	98.6	98.7	98.8	98.9	99.0	99.1	99.2	99.3	99.4	99.5	99.6	99.7	99.8	99.9	100.0	100.1	100.2	100.3	100.4	100.5	100.6	100.7	100.8	100.9	101.0	101.1	101.2	101.3	101.4	101.5	101.6	101.7	101.8	101.9	102.0	102.1	102.2	102.3	102.4	102.5	102.6	102.7	102.8	102.9	103.0	103.1	103.2	103.3	103.4	103.5	103.6	103.7	103.8	103.9	104.0	104.1	104.2	104.3	104.4	104.5	104.6	104.7	104.8	104.9	105.0	105.1	105.2	105.3	105.4	105.5	105.6	105.7	105.8	105.9	106.0	106.1	106.2	106.3	106.4	106.5	106.6	106.7	106.8	106.9	107.0	107.1	107.2	107.3	107.4	107.5	107.6	107.7	107.8	107.9	108.0	108.1	108.2	108.3	108.4	108.5	108.6	108.7	108.8	108.9	109.0	109.1	109.2	109.3	109.4	109.5	109.6	109.7	109.8	109.9	110.0	110.1	110.2	110.3	110.4	110.5	110.6	110.7	110.8	110.9	111.0	111.1	111.2	111.3	111.4	111.5	111.6	111.7	111.8	111.9	112.0	112.1	112.2	112.3	112.4	112.5	112.6	112.7	112.8	112.9	113.0	113.1	113.2	113.3	113.4	113.5	113.6	113.7	113.8	113.9	114.0	114.1	114.2	114.3	114.4	114.5	114.6	114.7	114.8	114.9	115.0	115.1	115.2	115.3	115.4	115.5	115.6	115.7	115.8	115.9	116.0	116.1	116.2	116.3	116.4	116.5	116.6	116.7	116.8	116.9	117.0	117.1	117.2	117.3	117.4	117.5	117.6	117.7	117.8	117.9	118.0	118.1	118.2	118.3	118.4	118.5	118.6	118.7	118.8	118.9	119.0	119.1	119.2	119.3	119.4	119.5	119.6	119.7	119.8	119.9	120.0	120.1	120.2	120.3	120.4	120.5	120.6	120.7	120.8	120.9	121.0	121.1	121.2	121.3	121.4	121.5	121.6	121.7	121.8	121.9	122.0	122.1	122.2	122.3	122.4	122.5	122.6	122.7	122.8	122.9	123.0	123.1	123.2	123.3	123.4	123.5	123.6	123.7	123.8	123.9	124.0	124.1	124.2	124.3	124.4	124.5	124.6	124.7	124.8	124.9	125.0	125.1	125.2	125.3	125.4	125.5	125.6	125.7	125.8	125.9	126.0	126.1	126.2	126.3	126.4	126.5	126.6	126.7	126.8	126.9	127.0	127.1	127.2	127.3	127.4	127.5	127.6	127.7	127.8	127.9	128.0	128.1	128.2	128.3	128.4	128.5	128.6	128.7	128.8	128.9	129.0	129.1	129.2	129.3	129.4	129.5	129.6	129.7	129.8	129.9	130.0
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

表A.1 随机数字表

行	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	10480	15011	01536	02011	81647	91646	69179	14194	62590	36207	20969	99570	91291	90700
2	22368	46573	25595	85393	30995	89198	27982	53402	93965	34095	52666	19174	39615	99505
3	24130	48360	22527	97265	76393	64809	15179	24830	49340	32081	30680	19655	63348	58629
4	42167	93093	06243	61680	07856	16376	39440	53537	71341	57004	00849	74917	97758	16379
5	37570	39975	81837	16656	06121	91782	60468	81305	49684	60672	14110	06927	01263	54613
6	77921	06907	11008	42751	27756	53498	18602	70659	90655	15053	21916	81825	44394	42880
7	99562	72905	56420	69994	98872	31016	71194	18738	44013	48840	63213	21069	10634	12952
8	96301	91977	65463	07972	18876	20922	94595	56869	69014	60045	18425	84903	42508	32307
9	89579	14342	63661	10281	17453	18103	57740	84378	25331	12566	58678	44947	05585	56941
10	85475	36857	53342	53988	53060	59533	38867	62300	08158	17983	16439	11458	18593	64952
11	28918	69578	88231	33276	70997	79936	56865	05859	90106	31595	01547	85590	91610	78188
12	63553	40961	48235	03427	49626	69445	18663	72695	52180	20847	12234	90511	33703	90322
13	09429	93969	52636	92737	88974	33488	36320	17617	30015	08272	84115	27156	30613	74952
14	10365	61129	87529	85689	48237	52267	67689	93394	01511	26358	85104	20285	29975	89868
15	07119	97336	71048	08178	77233	13916	47564	81506	97735	85977	29372	74461	28551	90707
16	51085	12765	51821	51259	77452	16308	60756	92144	49442	53900	70960	63990	75601	40719
17	02368	21382	52404	60268	89368	19885	55322	44819	01188	65255	64835	44919	05944	55157
18	01011	54092	33362	94904	31273	04146	18594	29852	71585	85030	51132	01915	92747	64951
19	52162	53916	46369	58586	23216	14513	83149	98736	23495	64350	94738	17752	35156	35749
20	07056	97628	33787	09998	42698	06691	76988	13602	51851	46104	88916	19509	25625	58104
21	48663	91245	85828	14346	09172	30168	90229	04734	59193	22178	30421	61666	99904	32812
22	54164	58492	22421	74103	47070	25306	76468	26384	58151	06646	21524	15227	96909	44592
23	32639	32363	05597	24200	13363	38005	94342	28728	35806	06912	17012	64161	18296	22851
24	29334	27001	87637	87308	58731	00256	45834	15298	46557	41135	10367	07684	36188	18510
25	02488	33062	28834	07351	19731	92420	60952	61280	50001	67658	32586	86679	50720	94953
26	81525	72295	04839	96423	24878	82651	66566	14778	76797	14780	13300	87074	79666	95725
27	29676	20591	68086	26432	46901	20849	89768	81536	86645	12659	92259	57102	80428	25280
28	00742	57392	39064	66432	84673	40027	32832	61362	98947	96067	64760	64584	96096	98253
29	05366	04213	25669	26422	44407	44048	37937	63904	45766	66134	75470	66520	34693	90449
30	91921	26418	64117	94305	26766	25940	39972	22209	71500	64568	91402	42416	07844	69618
31	00582	04711	87917	77341	42206	35126	74087	99547	81817	42607	43808	76655	62028	76630
32	00725	69884	62797	56170	86324	88072	76222	36086	84637	93161	76038	65855	77919	88006
33	69011	65795	95876	55293	18988	27354	26575	08625	40801	59920	29841	80150	12777	48501
34	25976	57948	29888	88604	67917	48708	18912	82271	65424	69774	33611	54262	85963	03547
35	09763	83473	93577	12908	30883	18317	28290	35797	05998	41688	34952	37888	38917	88050
36	91567	42595	27958	30134	04024	86385	29880	99730	55536	84855	29080	09250	79656	73211
37	17955	56349	90999	49127	20044	59931	06115	20542	18059	02008	73708	83517	36103	42791
38	46503	18584	18845	49618	02304	51038	20655	58727	28168	15475	56942	53389	20562	87338
39	92157	89634	94824	78171	84610	82834	09922	25417	44137	48413	25555	21246	35509	20468
40	14577	62765	35605	81263	39667	47358	56873	56307	61607	49518	89696	20103	77490	18062
41	98427	07523	33362	64270	01638	92477	66969	98420	04880	45585	46565	04102	46880	45709
42	34914	63976	88720	82765	34476	17032	87589	40836	32427	70002	70663	88863	77775	69348
43	70060	28277	39475	46473	23219	53416	94970	25832	69975	94884	19661	72828	00102	66794
44	53976	54914	06990	67245	68350	82948	11398	42878	80287	88267	47363	46634	06541	97809
45	76072	29515	40980	07391	58745	25774	22987	80059	39911	96189	41151	14222	60697	59583
46	90725	52210	83974	29992	65831	38857	50490	83765	55657	14361	31720	57375	56228	41546
47	64364	67412	33339	31926	14883	24413	59744	92351	97473	89286	35931	04110	23726	51900
48	08962	00358	31662	25388	61642	34072	81249	35648	56891	69352	48373	45578	78547	81788
49	95012	68379	93526	70765	10592	04542	76463	54328	02349	17247	28865	14777	62730	92277
50	15664	10493	20492	38391	91132	21999	59516	81652	27195	48223	46751	22923	32261	85653

*资料来源: Table of 105,000 Random Decimal Digits, Statement no. 4914, File no. 261-A-1, Interstate Commerce Commission, Washington, D.C. May 1949.

表A.2 部分t-分布值表*

应用指南: 要找一个 t 值, 首先在左侧一栏找到与平均数标准误相关的自由度, 然后选择相应 α 水平 (双尾) 下的 t 值。

$\alpha=.05$ 的一栏表示95%的置信区间, $\alpha=.01$ 的一栏表示99%的置信区间。

df	$\alpha = .05$	$\alpha = .01$	df	$\alpha = .05$	$\alpha = .01$
1	12.71	63.66	18	2.10	2.88
2	4.30	9.92	19	2.09	2.86
3	3.18	5.84	20	2.09	2.84
4	2.78	4.60	21	2.08	2.83
5	2.57	4.03	22	2.07	2.82
6	2.45	3.71	23	2.07	2.81
7	2.36	3.50	24	2.06	2.80
8	2.31	3.36	25	2.06	2.79
9	2.26	3.25	26	2.06	2.78
10	2.23	3.17	27	2.05	2.77
11	2.20	3.11	28	2.05	2.76
12	2.18	3.06	29	2.04	2.76
13	2.16	3.01	30	2.04	2.75
14	2.14	2.98	40	2.02	2.70
15	2.13	2.95	60	2.00	2.66
16	2.12	2.92	120	1.98	2.62
17	2.11	2.90	无限的	1.96	2.58

*本表摘自 *Biometrika table for Statisticians* 中的表12, 卷1 (3d ed.), 纽约: 剑桥大学出版社, 1970, 作者: E.S. Pearson, H.O. Hartley, 经 *Biometrika* 管理者同意。

表A.3 F-分布的临界值表*

应用指南: 要找到一个F值, 需要在表中找行与列的交叉点。其中行值代表F比率的分子自由度, 列值表示F比率的母自由度。表中黑体字表示 $\alpha=.05$ 水平下的F值, 罗马字体表示 $\alpha=.01$ 水平下的F值。例如, 假定我们采取 $\alpha=.05$, 分别找分子自由度 (dfnum=2) 和分母自由度 (dfdénom=12), 从表中我们可查得 $\alpha=.05$ 水平的临界值 $F(2, 12)=3.89$ 。如果我们得到的F值大于或等于临界值, 我们可以拒绝虚无假设, 如果我们得到的F值比临界值小, 则不能拒绝虚无假设。

		分子自由度																	
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	无限的
分母自由度	1	161 4052	200 4999	216 5403	225 5625	230 5764	234 5859	237 5928	239 5981	241 6022	242 6056	244 6106	246 6157	248 6209	249 6325	250 6261	251 6287	252 6313	254 6366
	2	18.5 98.5	19.0 99.0	19.2 99.2	19.2 99.2	19.3 99.3	19.3 99.3	19.4 99.4	19.4 99.4	19.4 99.4	19.4 99.4	19.4 99.4	19.4 99.4	19.4 99.4	19.4 99.5	19.5 99.5	19.5 99.5	19.5 99.5	19.5 99.5
	3	10.1 34.1	9.55 30.8	9.28 29.5	9.12 28.7	9.01 28.2	8.94 27.9	8.89 27.7	8.85 27.5	8.81 27.4	8.79 27.2	8.74 27.0	8.70 26.9	8.66 26.7	8.64 26.6	8.62 26.5	8.59 26.4	8.57 26.3	8.53 26.1
	4	7.71 21.2	6.94 18.0	6.59 16.7	6.39 16.0	6.26 15.5	6.16 15.2	6.09 15.0	6.04 14.8	6.00 14.7	5.96 14.6	5.91 14.4	5.86 14.2	5.80 14.0	5.77 13.9	5.75 13.8	5.72 13.8	5.69 13.6	5.63 13.5
	5	6.61 16.3	5.79 13.3	5.41 12.1	5.19 11.4	5.05 11.0	4.95 10.7	4.88 10.5	4.82 10.3	4.77 10.2	4.74 10.0	4.68 9.89	4.62 9.72	4.56 9.55	4.53 9.47	4.50 9.38	4.46 9.29	4.43 9.20	4.26 9.02
	6	5.99 13.8	5.14 10.9	4.76 9.78	4.53 9.15	4.39 8.75	4.28 8.47	4.21 8.26	4.15 8.10	4.10 7.98	4.06 7.87	4.00 7.72	3.94 7.56	3.87 7.40	3.84 7.31	3.81 7.23	3.77 7.14	3.74 7.06	3.67 6.88
	7	5.59 12.2	4.74 9.55	4.35 8.45	4.12 7.85	3.97 7.46	3.87 7.19	3.79 6.99	3.73 6.84	3.68 6.72	3.64 6.62	3.57 6.47	3.51 6.31	3.44 6.16	3.41 6.07	3.38 5.99	3.34 5.91	3.30 5.82	3.23 5.65
	8	5.32 11.3	4.46 8.65	4.07 7.59	3.84 7.01	3.69 6.63	3.58 6.37	3.50 6.18	3.44 6.03	3.39 5.91	3.35 5.81	3.28 5.67	3.22 5.52	3.15 5.36	3.12 5.28	3.10 5.20	3.08 5.12	3.04 5.03	2.93 4.86
	9	5.12 10.6	4.26 8.02	3.86 6.99	3.63 6.42	3.48 6.06	3.37 5.80	3.29 5.61	3.23 5.47	3.18 5.35	3.14 5.26	3.07 5.11	3.01 4.96	2.94 4.81	2.90 4.73	2.86 4.65	2.83 4.57	2.79 4.48	2.71 4.31
	10	4.96 10.0	4.10 7.56	3.71 6.55	3.48 5.99	3.33 5.64	3.22 5.39	3.14 5.20	3.07 5.06	3.02 4.94	2.98 4.85	2.91 4.71	2.85 4.56	2.77 4.41	2.74 4.33	2.70 4.25	2.66 4.17	2.62 4.08	2.54 3.91
	11	4.84 9.65	3.98 7.21	3.59 6.22	3.36 5.67	3.20 5.32	3.09 5.07	3.01 4.89	2.95 4.74	2.90 4.63	2.85 4.54	2.79 4.40	2.72 4.25	2.65 4.10	2.61 4.02	2.57 3.94	2.53 3.86	2.49 3.78	2.40 3.60
	12	4.75 9.33	3.89 6.93	3.49 5.95	3.26 5.41	3.11 5.06	3.00 4.82	2.91 4.64	2.85 4.50	2.80 4.39	2.75 4.30	2.69 4.16	2.62 4.01	2.54 3.86	2.51 3.78	2.47 3.70	2.43 3.62	2.38 3.54	2.30 3.36
	13	4.67 9.07	3.81 6.70	3.41 5.74	3.18 5.21	3.03 4.86	2.92 4.62	2.83 4.44	2.77 4.30	2.71 4.19	2.67 4.10	2.60 3.96	2.53 3.82	2.46 3.66	2.42 3.59	2.38 3.51	2.34 3.43	2.30 3.34	2.21 3.17
	14	4.60 8.86	3.74 6.51	3.34 5.56	3.11 5.04	2.96 4.69	2.85 4.46	2.76 4.28	2.70 4.14	2.65 4.03	2.60 3.94	2.53 3.80	2.46 3.66	2.39 3.51	2.35 3.43	2.31 3.35	2.27 3.27	2.22 3.18	2.13 3.00
	15	4.54 8.68	3.68 6.36	3.29 5.42	3.06 4.89	2.90 4.56	2.79 4.32	2.71 4.14	2.64 4.00	2.59 3.89	2.54 3.80	2.48 3.67	2.40 3.52	2.33 3.37	2.29 3.29	2.25 3.21	2.20 3.13	2.16 3.05	2.07 2.87
	16	4.49 8.53	3.63 6.23	3.24 5.29	3.01 4.77	2.85 4.44	2.74 4.20	2.66 4.03	2.59 3.89	2.54 3.78	2.49 3.69	2.42 3.55	2.35 3.41	2.28 3.26	2.24 3.18	2.19 3.10	2.15 3.02	2.11 2.93	2.01 2.75
	17	4.45 8.40	3.59 6.11	3.20 5.18	2.96 4.67	2.81 4.34	2.70 4.10	2.61 3.93	2.55 3.79	2.49 3.68	2.45 3.59	2.38 3.46	2.31 3.31	2.23 3.16	2.19 3.08	2.15 3.00	2.10 2.92	2.06 2.83	1.96 2.65
	18	4.41 8.29	3.55 6.01	3.16 5.09	2.93 4.58	2.77 4.25	2.66 4.01	2.58 3.84	2.51 3.71	2.46 3.60	2.41 3.51	2.34 3.37	2.27 3.23	2.19 3.08	2.15 3.00	2.11 2.92	2.06 2.84	2.02 2.75	1.92 2.57

(续)

表A.3 F-分布的临界值表* (续)

		分子自由度																	
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	无限的
分母自由度	19	4.38 8.18	3.52 5.93	3.13 5.01	2.90 4.50	2.74 4.17	2.63 3.94	2.54 3.77	2.48 3.63	2.42 3.52	2.38 3.43	2.31 3.30	2.23 3.15	2.16 3.00	2.11 2.92	2.07 2.84	2.03 2.76	1.98 2.67	1.88 2.49
	20	4.35 8.10	3.49 5.85	3.10 4.94	2.87 4.43	2.71 4.10	2.60 3.87	2.51 3.70	2.45 3.56	2.39 3.46	2.35 3.37	2.28 3.23	2.20 3.09	2.12 2.94	2.08 2.86	2.04 2.78	1.99 2.69	1.95 2.61	1.84 2.42
	22	4.30 7.95	3.44 5.72	3.05 4.82	2.82 4.31	2.66 3.99	2.55 3.76	2.46 3.59	2.40 3.45	2.34 3.35	2.30 3.26	2.23 3.12	2.15 2.98	2.07 2.83	2.03 2.75	1.98 2.67	1.94 2.58	1.89 2.50	1.78 2.31
	24	4.26 7.82	3.40 5.61	3.01 4.72	2.78 4.22	2.62 3.90	2.51 3.67	2.42 3.50	2.36 3.36	2.30 3.26	2.25 3.17	2.18 3.03	2.11 2.89	2.03 2.74	1.98 2.66	1.94 2.58	1.89 2.49	1.84 2.40	1.73 2.21
	26	4.23 7.72	3.37 5.53	2.98 4.64	2.74 4.14	2.59 3.82	2.47 3.59	2.39 3.42	2.32 3.29	2.27 3.18	2.22 3.09	2.15 2.96	2.07 2.81	1.99 2.66	1.95 2.58	1.90 2.50	1.85 2.42	1.80 2.33	1.69 2.13
	28	4.20 7.64	3.34 5.45	2.95 4.57	2.71 4.07	2.56 3.75	2.45 3.53	2.36 3.36	2.29 3.23	2.24 3.12	2.19 3.03	2.12 2.90	2.04 2.75	1.96 2.60	1.91 2.52	1.87 2.44	1.82 2.35	1.77 2.26	1.65 2.06
	30	4.17 7.56	3.32 5.39	2.92 4.51	2.69 4.02	2.53 3.70	2.42 3.47	2.33 3.30	2.27 3.17	2.21 3.07	2.16 2.98	2.09 2.84	2.01 2.70	1.93 2.55	1.89 2.47	1.84 2.39	1.79 2.30	1.74 2.21	1.62 2.01
	40	4.08 7.31	3.23 5.18	2.84 4.31	2.61 3.83	2.45 3.51	2.34 3.29	2.25 3.12	2.18 2.99	2.12 2.89	2.08 2.80	2.00 2.66	1.92 2.52	1.84 2.37	1.79 2.29	1.74 2.20	1.69 2.11	1.64 2.02	1.51 1.80
	60	4.00 7.06	3.15 4.98	2.76 4.13	2.53 3.65	2.37 3.34	2.25 3.12	2.17 2.95	2.10 2.82	2.04 2.72	1.99 2.63	1.92 2.50	1.84 2.35	1.75 2.20	1.7 2.12	1.65 2.03	1.59 1.94	1.53 1.84	1.39 1.60
	120	3.92 6.85	3.07 4.79	2.68 3.95	2.45 3.48	2.29 3.17	2.17 2.96	2.09 2.79	2.02 2.66	1.96 2.56	1.91 2.47	1.83 2.34	1.75 2.19	1.66 2.03	1.61 1.95	1.55 1.86	1.50 1.76	1.43 1.66	1.25 1.38
无限的	3.84 6.63	3.00 4.61	2.60 3.78	2.37 3.32	2.21 3.02	2.10 2.80	2.01 2.64	1.94 2.51	1.88 2.41	1.83 2.32	1.75 2.18	1.67 2.04	1.57 1.88	1.52 1.79	1.46 1.70	1.39 1.59	1.32 1.47	1.00 1.00	

*本表摘自 *Biometrika table for Statisticians* 中的表18, 卷1 (3d ed.), 纽约: 剑桥大学出版社, 1970, 作者: E.S. Pearson, H.O. Hartley, 经 *Biometrika* 管理者同意。

表A.4 两独立组均值比较效力表 $\alpha=.05$ (双尾检验)

提示: 本表是独立组 t 检验 ($\alpha=.05$, 双尾检验条件下) 不同效应大小 (.10-1.00) 对样本大小的函数值。效应大小基于科恩 d 值。本表可用于评价事后检验的统计效力。例: 当每一组中样本大小 $n=20$, 效应大小中等 (.50), 统计效力是.33。相同的样本大小情况下, 效应量大 (.80) 时统计效力为.69。本表也可以用来进行实验计划。假设要求效应大小是.06, 根据科恩规则, 仅大于中等效应值, 当统计效力为.80的时候需要多大的样本? 看表中效应大小为.60栏, 你会发现随样本大小的增加, 效力也增加, 当样本大小为50时, 效力为.84, 要想效力为.80, 我们可以估计样本大小约为45 (内插法找到40和50的中间值)。当效力为.70时, 样本大小是多大? 通过表你可以发现每一组样本大小是35时效力为.70 (以Cohen为依据, 1988, pp.36-37)

样本大小	效应大小 (d)								
	(小)			(中等)			(大)		
	.10	.20	.30	.40	.50	.60	.70	.80	1.00
10	.06	.07	.10	.13	.18	.24	.31	.39	.56
15	.06	.08	.12	.18	.26	.35	.45	.56	.75
20	.06	.09	.15	.23	.33	.45	.58	.69	.87
25	.06	.11	.18	.28	.41	.55	.68	.79	.93
30	.07	.12	.21	.33	.47	.63	.76	.86	.97
35	.07	.13	.23	.38	.54	.70	.82	.91	.98
40	.07	.14	.26	.42	.60	.75	.87	.94	.99
50	.08	.17	.32	.50	.70	.84	.93	.98	*
60	.08	.19	.37	.58	.77	.90	.97	.99	*
80	.10	.24	.47	.71	.88	.96	.99	*	*
100	.11	.29	.56	.80	.94	.99	*	*	*

表A.5 独立组方差分析的效力表 $\alpha=.05$

提示: 本表是独立组设计不同效应量 (.05-.50) 对样本大小的函数值。效应量是用科恩 f 值进行评价的 (见 12 章)。统计效力分别按独立组数为 2, 3, 4, 5 进行归纳的。本表可以对事后统计效力进行评价。例如, 假定一个研究采用了 3 组 ($n=30$), 对数据进行方差分析。进一步假设效应大小 f 值为 .40, 你可以发现当样本大小为 30 时, 统计效力为 .93。本表也可以用来计划一个实验中特定数量效应下的样本大小。假设先前研究揭示兴趣变异的效应量平均数为 .30。当统计效力为 .80 的 4 个组实验中, 样本大小为多少? 查看表中 4 个组的独立组实验的统计效力, 你会发现随着样本大小的增加, 统计效力增加, 当样本大小为 30 时, 统计效力为 .84, n 为 25 时统计效力为 .76。如果我们的效应大小是一个小效应量 .10, 则当样本大小为 60 时统计效力仅仅为 .22 (以 Cohen 为依据, 1988, pp.311-318)

2组 ($df=1$)

效应大小 (f)									
样本大小	(小)			(中等)			(大)		
	.05	.10	.15	.20	.25	.30	.35	.40	.50
5	.05	.06	.07	.08	.11	.13	.16	.20	.29
10	.05	.07	.09	.13	.18	.25	.32	.40	.57
15	.06	.08	.12	.18	.26	.36	.47	.57	.76
20	.06	.09	.15	.23	.34	.46	.59	.70	.88
25	.06	.10	.18	.29	.42	.56	.69	.80	.94
30	.06	.11	.21	.34	.49	.64	.77	.87	.97
35	.07	.12	.24	.39	.55	.71	.83	.92	.99
40	.07	.14	.27	.43	.61	.77	.88	.95	.99
50	.07	.16	.32	.52	.71	.85	.94	.98	*
60	.08	.19	.38	.60	.79	.91	.97	.99	*

3组 ($df=2$)

效应大小 (f)									
样本 大小	(小)				(中等)			(大)	
	.05	.10	.15	.20	.25	.30	.35	.40	.50
5	.05	.06	.07	.09	.11	.14	.17	.22	.32
10	.05	.07	.10	.14	.20	.27	.35	.45	.64
15	.06	.08	.13	.20	.29	.40	.52	.64	.84
20	.06	.09	.16	.26	.38	.52	.66	.78	.93
25	.06	.10	.19	.32	.47	.63	.77	.87	.98
30	.06	.12	.22	.37	.55	.71	.85	.93	.99
35	.07	.13	.26	.43	.62	.79	.90	.96	*
40	.07	.15	.29	.48	.68	.84	.94	.98	*
50	.08	.18	.36	.58	.79	.92	.98	.99	*
60	.08	.21	.42	.67	.86	.96	.99	*	*

表A.5 独立组方差分析的效力表 $\alpha=.05^*$ (续)

4组 ($df=3$)									
样本大小	效应大小 (f)								
	.05	(小) .10	.15	.20	(中等) .25	.30	.35	(大) .40	.50
5	.05	.06	.07	.09	.12	.16	.21	.26	.40
10	.06	.07	.10	.16	.23	.33	.44	.56	.78
15	.06	.09	.14	.23	.36	.50	.65	.78	.94
20	.06	.10	.18	.31	.47	.65	.79	.90	.99
25	.06	.12	.23	.39	.58	.76	.89	.96	*
30	.07	.13	.27	.46	.67	.84	.94	.98	*
35	.07	.15	.31	.54	.75	.90	.97	.99	*
40	.07	.17	.36	.60	.81	.94	.99	*	*
50	.08	.19	.40	.65	.85	.96	.99	*	*
60	.09	.22	.47	.74	.91	.98	*	*	*
5组 ($df=4$)									
样本大小	效应大小 (f)								
	.05	(小) .10	.15	.20	(中等) .25	.30	.35	(大) .40	.50
5	.05	.06	.07	.10	.13	.17	.22	.29	.44
10	.06	.07	.11	.17	.25	.36	.48	.61	.83
15	.06	.09	.15	.25	.39	.55	.70	.82	.96
20	.06	.11	.20	.34	.52	.70	.84	.93	.99
25	.07	.12	.24	.43	.63	.81	.92	.98	*
30	.07	.14	.29	.51	.73	.88	.96	.99	*
35	.07	.16	.34	.65	.80	.93	.98	*	*
40	.08	.18	.39	.68	.86	.96	.99	*	*
50	.08	.21	.44	.71	.90	.98	*	*	*
60	.09	.24	.52	.80	.95	.99	*	*	*

* 表示统计效力大于.995

Glossary

术语表

ABAB 设计 (逆向设计)

ABAB design (reversal design) A single-subject experimental design in which an initial baseline stage (A) is followed by a treatment stage (B), a return to baseline (A), and then another treatment stage (B); the researcher observes whether behavior changes on introduction of the treatment, reverses when the treatment is withdrawn, and improves again when the treatment is reintroduced.

α

alpha See level of significance.

单因素方差分析

ANOVA The analysis of variance, or ANOVA, is the most commonly used inferential test for examining a null hypothesis when comparing more than two means in a single-factor study, or in studies with more than one factor (i.e., independent variable). The ANOVA test is based on analyzing different sources of variation in an experiment.

应用研究

applied research Research that seeks knowledge that will improve a situation. See also basic research.

档案数据

archival data Source of evidence based on records or documents relating the activities of individuals, institutions, governments, and other groups; used as an alternative to or in conjunction with other research methods.

缺 损

attrition See **subject attrition**.

基线阶段

baseline stage First stage of a single-subject experiment in which a record is made of the individual's behavior prior

to any intervention.

基础研究

basic research Research that seeks knowledge to increase understanding of behavior and mental processes and to test theories. See also applied research.

区组随机化

block randomization The most common technique for carrying out random assignment in the random groups design; each block includes a random order of the conditions and there are as many blocks as there are subjects in each condition of the experiment.

个案研究

case study An intensive description and analysis of a single individual.

因果推论

causal inference Identification of the cause or causes of a phenomenon, by establishing covariation of cause and effect, a time-order relationship with cause preceding effect, and the elimination of plausible alternative causes.

天花板 (地板) 效应

ceiling (and floor) effect Measurement problem whereby the researcher cannot measure the effects of an independent variable or a possible interaction effect because performance has reached a maximum (minimum) in any condition of the experiment.

集中趋势

central tendency See measures of central tendency.

编 码

coding The initial step in data reduction, especially with

narrative records, in which units of behavior or particular events are identified and classified according to specific criteria.

科恩 d 值

Cohen's d A frequently used measure of effect size in which the difference in means for two conditions is divided by the average variability of participants' scores (within-group standard deviation). Based on Cohen's guidelines, d values of .20, .50, and .80 represent small, medium, and large effects, respectively, of an independent variable.

科恩 f 值

Cohen's f A measure of effect size when there are more than two means that defines an effect relative to the degree of dispersal among group means. Based on Cohen's guidelines, an f value of .10, .25, and .40 defines a small, medium, and large effect size, respectively.

两个平均数的比较

comparison of two means A statistical technique that can be applied (usually after obtaining a statistically significant omnibus F -test) to locate the specific source of systematic variation in an experiment by comparing means two at a time.

复合设计

complex design Experiment in which two or more independent variables are studied simultaneously.

置信区间

confidence intervals Indicate the range of values which we can expect to contain a population value with a specified degree of confidence (e.g., 95%).

总体参数的置信区间

confidence interval for a population parameter A range of values around a sample statistic (e.g., sample mean) with specified probability (e.g., .95) that the population parameter (e.g., population mean) has been captured within that interval.

证实数据揭示内容

confirming what the data reveal In the third stage of data analysis the researcher determines what the data tell us about behavior. Statistical techniques are used to counter ar-

guments that the results are simply "due to chance."

混淆的

confounding Occurs when the independent variable of interest systematically covaries with a second, unintended independent variable.

构念

construct A concept or idea used in psychological theories to explain behavior or mental processes; examples include aggression, depression, intelligence, memory, and personality.

污染

contamination Occurs when there is communication of information about the experiment between groups of participants.

内容分析

content analysis Any of a variety of techniques for making inferences by objectively identifying specific characteristics of messages, usually written communications but may be any form of message; used extensively in the analysis of archival data.

控制

control Key component of the scientific method whereby the effect of various factors possibly responsible for a phenomenon are isolated; three basic types of control are manipulation, holding conditions constant, and balancing.

相关

correlation Exists when two different measures of the same people, events, or things vary together; the presence of a correlation makes it possible to predict values on one variable by knowing the values on the second variable.

相关系数

correlation coefficient Statistic indicating how well two measures vary together; absolute size ranges from 0.0 (no correlation) to 1.00 (perfect correlation); direction of covariation is indicated by the sign of the coefficient, a plus (+) indicating that both measures covary in the same direction and a minus (-) indicating that the variables vary in opposite directions.

相关研究

correlational research Research to identify predictive relationships among naturally occurring variables.

抵消平衡

counterbalancing A control technique for distributing (balancing) practice effects across the conditions of a repeated measures design. How counterbalancing is accomplished depends on whether a complete or an incomplete repeated measures design is used.

横断设计

cross-sectional design Survey research design in which one or more samples of the population are selected and information is collected from the samples at one time.

数据简化

data reduction Process in the analysis of behavioral data whereby results are meaningfully organized and statements summarizing important findings are prepared.

事后解释

debriefing Process following a research session through which participants are informed about the rationale for the research in which they participated, about the need for any deception, and about their specific contribution to the research. Important goals of debriefing are to clear up any misconceptions and to leave participants with a positive feeling toward psychological research.

欺骗

deception Intentionally withholding information from a participant about significant aspects of a research project or presenting misinformation about the research to participants.

需求特征

demand characteristics Cues and other information used by participants to guide their behavior in a psychological study, often leading participants to do what they believe the observer (experimenter) expects them to do.

因变量

dependent variable Measure of behavior used by the researcher to assess the effect (if any) of the independent variable.

差异转换

differential transfer Potential problem in repeated measures designs when performance in one condition differs depending on which of two other conditions precedes it.

双盲

double-blind Both the participant and the observer are kept unaware (blind) of what treatment is being administered.

效应大小

effect size Index of the strength of the relationship between the independent variable and dependent variable that is independent of sample size.

实证方法

empirical approach Approach to acquiring knowledge that emphasizes direct observation and experimentation as a way of answering questions.

平均数的标准误差估计

estimated standard error of the mean An estimate of the true standard error obtained by dividing the sample standard deviation by the square root of the sample size.

联系性测量 η^2

eta squared (η^2) A measure of the strength of association (or effect size) based on the proportion of variance accounted for by the effect of the independent variable on the dependent variable.

民族中心主义

ethnocentrism An attempt to understand the behavior of individuals in different cultures based solely on experiences in one's own culture.

实验

experiment A controlled research situation in which scientists manipulate one or more factors and observe the effects of this manipulation on behavior.

实验者效应

experimenter effects Experimenters' expectations that may lead them to treat subjects differently in different groups or to record data in a biased manner.

外部效度

external validity The extent to which the results of a research study can be generalized to different populations, settings, and conditions.

因子设计

factorial design See **complex design**.

现场实验

field experiment Procedure in which one or more independent variables is manipulated by an observer in a natural setting to determine the effect on behavior.

地板效应

floor effect See **ceiling effect**.

F检验

F-test In the analysis of variance, or ANOVA, the ratio of between-group variation and within-group or error variation.

了解数据

getting to know the data In this first stage of data analysis the researcher inspects the data for errors and outliers and generally becomes familiar with the general features of the data.

霍桑效应

Hawthorne effect See **novelty effects**.

历史

history The occurrence of an event other than the treatment that can threaten internal validity if it produces changes in the research participants' behavior.

假设, 假说

hypothesis A tentative explanation for a phenomenon.

个人特质研究法

idiographic approach Intensive study of an individual, with an emphasis on both individual uniqueness and lawfulness.

独立组设计

independent groups design Each separate group of subjects in the experiment represents a different condition as defined by the level of the independent variable.

自变量

independent variable Factor for which the researcher

manipulates at least two levels in order to determine its effect on behavior.

个体差异变量

individual differences variable A characteristic or trait that varies consistently across individuals, such as level of depression, age, intelligence, gender. Because this variable is formed from pre-existing groups (i.e., it occurs "naturally") an individual differences variable is sometimes called a natural groups variable. Another term sometimes used synonymously with individual differences variable is subject variable.

知情同意

informed consent Explicitly expressed willingness to participate in a research project based on clear understanding of the nature of the research, of the consequences of not participating, and of all factors that might be expected to influence willingness to participate.

仪器

instrumentation Changes over time can take place not only in the participants of an experiment, but also in the instruments used to measure the participants' performance. These changes due to instrumentation can threaten internal validity if they cannot be separated from the effect of the treatment.

交互作用

interaction effect When the effect of one independent variable differs depending on the level of a second independent variable.

内部效度

internal validity Degree to which differences in performance can be attributed unambiguously to an effect of an independent variable, as opposed to an effect of some other (uncontrolled) variable; an internally valid study is free of confounds.

时间系列设计

interobserver reliability Degree to which two independent observers are in agreement.

简单间断时间系列设计

interrupted time-series design See **simple interrupted time-series design** and **time series with**

nonequivalent control group design.

访谈者偏差

interviewer bias Occurs when the interviewer tries to adjust the wording of a question to fit the respondent or records only selected portions of the respondent's answers.

显著性水平

level of significance The probability when testing the null hypothesis that is used to indicate whether an outcome is statistically significant. Level of significance, or alpha, is equal to the probability of a Type I error.

线性趋势

linear trend A trend in the data that is appropriately summarized by a straight line.

纵向研究设计

longitudinal design Research design in which the same sample of respondents is interviewed (tested) more than once.

主效应

main effect Overall effect of an independent variable in a complex design.

匹配组设计

matched groups design Type of independent groups design in which the researcher forms comparable groups by matching subjects on a pretest task and then randomly assigning the members of these matched sets of subjects to the conditions of the experiment.

成熟

maturation Change associated with the passage of time per se is called maturation. Changes participants undergo in an experiment that are due to maturation and not due to the treatment can threaten internal validity.

均值

mean The arithmetic mean, or average, is determined by dividing the sum of the scores by the number of scores contributing to that sum. The mean is the most commonly used measure of central tendency.

测量量表

measurement scale One of four levels of physical and

psychological measurement: nominal (categorizing), ordinal (ranking), interval (specifying distance between stimuli), and ratio (having an absolute zero point).

集中趋势量度

measures of central tendency Measures such as the mean, median, and mode, that identify a score that the data tend to center around.

离中趋势量度 (变异)

measures of dispersion (variability) Measures such as the range and standard deviation that describe the degree of dispersion of numbers in a distribution.

技术性被试缺失

mechanical subject loss Occurs when a subject fails to complete the experiment because of equipment failure or because of experimenter error.

中位数

median The middle point in a distribution, above which half the scores fall and below which half fall.

元分析

meta-analysis Analysis of results of several (often, very many) independent experiments investigating the same research area; the measure used in a meta-analysis is typically effect size.

最小风险

minimal risk A research participant is said to experience minimal risk when probability and magnitude of harm or discomfort anticipated in the research are not greater than that ordinarily encountered in daily life or during the performance of routine tests.

众数

mode The score that appears most frequently in the distribution.

多种方法

multimethod approach Approach to hypothesis testing that seeks evidence by collecting data using several different research procedures and measures of behavior; a recognition of the fact that any single observation of behavior can result from some artifact of the measuring process.

多重基线设计 (跨个体, 跨行为, 跨情境)

multiple-baseline design (across individuals, across behaviors, across situations) A single-subject experimental design in which the effect of a treatment is demonstrated by showing that behaviors in more than one baseline change as a consequence of the introduction of a treatment; multiple baselines are established for different individuals, for different behaviors in the same individual, or for the same individual in different situations.

N=1 设计

N = 1 designs See **single-subject experiment**.

叙述性记录

narrative record Record intended to provide a more or less faithful reproduction of behavior as it originally occurred.

自然组设计

natural groups design Type of independent groups design in which the conditions represent the selected levels of a naturally occurring independent variable, for example, the individual differences variable age.

自然观察

naturalistic observation Observation of behavior in a more or less natural setting without any attempt by the observer to intervene.

负相关

negative correlation A relationship between two variables in which values for one measure increase as the values of the other measure decrease.

常规法则研究法

nomothetic approach Approach to research that seeks to establish broad generalizations or laws that apply to large groups (populations) of individuals; the average or typical performance of a group is emphasized.

非等控制组设计

nonequivalent control group design Quasiexperimental procedure in which a comparison is made between control and treatment groups that have been established on some basis other than through random assignment of participants to groups.

非随机抽样

nonprobability sampling A sampling procedure in which there is no way to estimate the probability of each element's being included in the sample; a common type is convenience sampling.

新异效应

novelty effects Threats to internal validity of a study that occur when people's behavior changes simply because an innovation (e.g., a treatment) produces excitement, energy, and enthusiasm; a Hawthorne effect is a special case of novelty effects.

虚无假设 (H_0)

null hypothesis (H_0) Assumption used as the first step in statistical inference whereby the independent variable is said to have had no effect.

虚无假设显著性检验

null hypothesis significance testing (NHST) A procedure for statistical inference used to decide whether a variable has produced an effect in a study. NHST begins with the assumption that the variable has no effect (see **null hypothesis**), and probability theory is used to determine the probability that the effect (e.g., a mean difference between conditions) observed in a study would occur simply by error variation ("chance"). If the likelihood of the observed effect is small (see **level of significance**), assuming the null hypothesis is true, we infer the variable produced a reliable effect (see **statistically significant**).

观察者偏差

observer bias Systematic errors in observation often resulting from the observer's expectancies regarding the outcome of a study (i.e., expectancy effects).

综合 F 检验

omnibus F-test The initial overall analysis based on ANOVA.

操作定义

operational definition Procedure whereby a concept is defined solely in terms of the observable procedures used to produce and measure it.

参与者观察

participant observation Observation of behavior by someone who also has an active and significant role in the situation or context in which behavior is recorded.

物理示踪

physical traces Source of evidence that is based on the remnants, fragments, and products of past behavior; used as an alternative to or in conjunction with other research methods.

安慰剂效应

placebo control Procedure by which a substance that resembles a drug or other active substance but that is actually an inert, or inactive, substance is given to participants.

学术剽窃

plagiarism Presentation of another's ideas or work without clearly identifying the source.

总体

population Set of all the cases of interest.

正相关

positive correlation A relationship between two variables in which values for one measure increase as the values of the other measure also increase.

检验力

power Probability in a statistical test that a false null hypothesis will be rejected; power is related to the level of significance selected, the size of the treatment effect, and the sample size.

练习效应

practice effects Changes that participants undergo with repeated testing. Practice effects are the summation of both positive (e.g., familiarity with a task) and negative (e.g., boredom) factors associated with repeated measurement.

隐私

privacy Right of individuals to decide how information about them is to be communicated to others.

概率抽样

probability sampling Sampling procedure in which the probability that each element of the population will be included in the sample can be specified.

程序评估

program evaluation Research that seeks to determine whether a change proposed by an institution, government agency, or other unit of society is needed and likely to have an effect as planned or, when implemented, to actually have an effect.

准实验

quasi-experiments Procedures that resemble characteristics of true experiments, for example, that some type of intervention or treatment is used and a comparison is provided, but are lacking in the degree of control that is found in true experiments.

问卷

questionnaire A set of predetermined questions for all respondents that serves as the primary research instrument in survey research.

随机分配

random assignment Most common technique for forming groups as part of an independent groups design; the goal is to establish equivalent groups by balancing individual differences.

随机组设计

random groups design Most common type of independent groups design in which subjects are randomly selected or randomly assigned to each group such that groups are considered comparable at the start of the experiment.

随机抽样

random sampling See simple random sampling.

全距

range The difference between the highest and lowest number in a distribution.

反应性

reactivity Influence that an observer has on the behavior under observation; behavior influenced by an observer may not be representative of behavior that occurs when an observer is not present.

回归 (向均值)

regression (to the mean) Statistical regression can occur when individuals have been selected to participate in

an experiment because of their "extreme" scores. Statistical regression is a threat to internal validity because individuals selected from extreme groups would be expected to have less extreme scores on a second test (the "posttest") *without any treatment* simply due to statistical regression.

相关自变量

relevant independent variable Independent variable that has been shown to influence behavior, either directly, by producing a main effect, or indirectly, by resulting in an interaction effect in combination with a second independent variable.

信度

reliability A measurement is reliable when it is consistent.

重复测量设计 (被试内)

repeated measures designs Research designs in which each subject participates in all conditions of the experiment (i.e., measurement is repeated on the same subject).

重复测量 (组内设计) 的 t 检验

repeated measures (within-subjects) t-test An inferential test for comparing two means from the same group of subjects or from two groups of subjects "matched" on some measure related to the dependent variable.

重复

replication Repeating the exact procedures used in an experiment to determine whether the same results are obtained.

代表性

representativeness A sample is representative to the extent that it has the same distribution of characteristics as the population from which it was selected; our ability to generalize from sample to population is critically dependent on representativeness.

反应偏差

response bias Threat to the representativeness of a sample that occurs when some participants selected to respond to a survey systematically fail to complete the survey (e.g., due to failure to complete a lengthy questionnaire or to reply with a request to participate in a phone survey).

逆向设计

reversal design See **ABAB design**.

风险/收益比

risk/benefit ratio Subjective evaluation of the risk to a research participant relative to the benefit both to the individual and to society of the results of the proposed research.

样本

sample Something less than all the cases of interest; in survey research, a subset of the population actually drawn from the sampling frame.

散点图

scatterplot A graph showing the relationship between two variables by indicating the intersection of two measures obtained from the same person, thing, or event.

科学方法

scientific method Approach to knowledge that emphasizes empirical rather than intuitive processes, testable hypotheses, systematic and controlled observation of operationally defined phenomena, data collection using accurate and precise instrumentation, valid and reliable measures, and objective reporting of results; scientists tend to be critical and, most importantly, skeptical.

选择/淘汰

selection Selection is a threat to internal validity when, from the outset of a study, differences exist between the kinds of individuals in one group and those in another group in the experiment.

选择偏差

selection bias Threat to the representativeness of a sample that occurs when the procedures used to select a sample result in the over- or underrepresentation of a significant segment of the population.

选择性保留

selective deposit Bias that results from the way physical traces are laid down and the way archival sources are produced, edited, or altered, as they are established; when present, the bias severely limits generality of research findings.

选择性被试缺失

selective subject loss Occurs when subjects are lost dif-

ferentially across the conditions of the experiment as the result of some characteristic of each subject that is related to the outcome of the study.

选择性残余

selective survival Bias that results from the way physical traces and archives survive over time; when present, the bias severely limits the external validity of research findings.

敏感性

sensitivity Refers to the likelihood in an experiment that the effect of an independent variable will be detected when that variable does, indeed, have an effect; sensitivity is increased to the extent that error variation is reduced (e.g., by holding variables constant rather than balancing them).

简单中断时间序列设计

simple interrupted time-series design Quasiexperimental procedure in which changes in a dependent variable are observed for some period of time both before and after a treatment is introduced.

简单主效应

simple main effect Effect of one independent variable at one level of a second independent variable in a complex design.

简单随机抽样 (随机选择)

simple random sampling (random selection) Type of probability sampling in which each possible sample of a specified size in the population has an equal chance of being selected.

单因素独立组设计

single-factor independent groups design An experiment that involves independent groups with one independent variable.

单一被试实验

single-subject experiment A procedure that focuses on behavior change in one individual by systematically contrasting conditions within that individual while continuously monitoring behavior.

情景抽样

situation sampling Random or systematic selection of situations in which observations are made with the goal of

representativeness across circumstances, locations, and conditions.

小 n 研究

small- n research See **single-subject experiment**.

社会需求

social desirability Pressures on survey respondents to answer as they think they should respond in accordance with what is most socially acceptable, and not in accordance with what they actually believe.

伪关系

spurious relationship What exists when evidence falsely indicates that two or more variables are associated.

数据分析阶段

stages of data analysis Three stages of data analysis are: getting to know the data, summarizing the data, and confirming what the data reveal.

标准差

standard deviation The most commonly used measure of dispersion that indicates approximately how far on the average scores differ from the mean.

均值的标准误

standard error of the mean The standard deviation of the sampling distribution of means.

统计显著

statistically significant When the probability of an obtained difference in an experiment is smaller than would be expected if error variation alone were assumed to be responsible for the difference, the difference is statistically significant.

茎叶图

stem-and-leaf display A technique for visualizing both the general features of a data set and specific item information by creating leading digits as "stems" and trailing digits as "leaves."

分层随机抽样

stratified random sampling Type of probability sampling in which the population is divided into subpopulations called strata and random samples are drawn from each of

these strata.

结构化观察

structured observation Variety of observational methods using intervention in which the degree of control is often less than in field experiments; frequently used by clinical and developmental psychologists when making behavioral assessments.

样本流失

subject attrition A threat to internal validity occurs when participants are lost from an experiment, for example, when participants drop out of the research project. The loss of participants changes the nature of a group from that established prior to the introduction of the treatment—for example, by destroying the equivalence of groups that had been established through random assignment.

连续独立样本设计

successive independent samples design Survey research design in which a series of cross-sectional surveys is done and the same questions are asked of each succeeding sample of respondents.

概括数据

summarizing the data In this second stage of data analysis the researcher uses descriptive statistics and graphical displays to summarize the information in a data set. Trends and patterns in the data set are described.

测试

testing Taking a test generally has an effect on subsequent testing. Testing can threaten internal validity if the effect of a treatment cannot be separated from the effect of testing.

理论

theory Logically organized set of propositions that serves to define events, describe relationships among events, and explain the occurrence of these events; scientific theories guide research and organize empirical knowledge.

对内部效度的威胁

threats to internal validity Possible causes of a phe-

nomenon that must be controlled so a clear cause-effect inference can be made.

时间抽样

time sampling Selection of observation intervals either systematically or randomly with the goal of obtaining a representative sample of behavior.

不等控制组时间序列设计 (也见简单间隔时间序列设计)

time series with nonequivalent control group design (See also **simple interrupted timeseries design**) Quasi-experimental procedure that improves on the validity of a simple timeseries design by including a nonequivalent control group; both treatment and comparison groups are observed for a period of time both before and after the treatment.

独立组 t 检验

t-test for independent groups An inferential test for comparing two means from different groups of subjects.

I 类错误

Type I error The probability of rejecting the null hypothesis when it is true, equal to the level of significance, or alpha.

II 类错误

Type II error The probability of failing to reject the null hypothesis when it is false.

隐藏 (不反应的) 测量

unobtrusive (nonreactive) measures Measures of behavior that eliminate the problem of reactivity because observations are made in such a way that the presence of the observer is not detected by those being observed.

效度

validity The “truthfulness” of a measure; a valid measure is one that measures what it claims to measure.

变异

variability See **measures of dispersion**.

References

参考文献

- Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.
- Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, 8, 12-15.
- Adair, J. G., Dushenko, T. W., & Lindsay, R. C. L. (1985). Ethical regulations and their impact on research practice. *American Psychologist*, 40, 59-72.
- Adjang, O. M. J. (1986). Exploring the social environment: A developmental study of teasing in chimpanzees. *Ethology*, 73, 136-160.
- Adler, A. (1973). *Practice and theory of individual psychology* (P. Radin, Trans.). Totowa, NJ: Littlefield, Adams.
- Adler, T. (1991, December). Outright fraud rare, but not poor science. *APA Monitor*, 11.
- Allison, M. G., & Ayllon, T. (1980). Behavioral coaching in the development of skills in football, gymnastics, and tennis. *Journal of Applied Behavior Analysis*, 13, 297-314.
- Allport, G. W. (1946). Letters from Jenny. *Journal of Abnormal and Social Psychology*, 41, 3-4.
- Allport, G. W. (1961). *Pattern in growth and personality*. New York: Holt, Rinehart and Winston.
- Allport, G. W. (1965). *Letters from Jenny*. New York: Harcourt, Brace & World.
- Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behavior*, 48, 1-41.
- Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, 64, 431-441.
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., Text Revision). Washington, DC: Author.
- American Psychological Association. (2001). *Publication manual* (5th ed). Washington, DC: Author.
- American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, 57, 1060-1073.
- Anderson, C. A. (2001). Heat and violence. *Current Directions in Psychological Science*, 10, 33-38.
- Anderson, C. A., Berkowitz, L., Donnerstein, E., Huesmann, L. R., Johnson, J. D., Linz, D., Malamuth, N. M., & Wartella, E. (2003). The influence of media violence on youth. *Psychological Science in the Public Interest*, 4, 81-110.
- Anderson, C. A., & Bushman, B. J. (1997). External validity of "trivial" experiments: The case of laboratory aggression. *Review of General Psychology*, 1, 19-41.
- Anderson, C. R. (1976). Coping behaviors as intervening mechanisms in the inverted-U stress-performance relationship. *Journal of Applied Psychology*, 61, 30-34.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., & Milson, J. R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96, 703-719.
- Anderson, K. J., & Revelle, W. (1982). Impulsivity, caffeine, and proofreading: A test of the Easterbrook hypothesis. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 614-624.
- Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*, 58 (10, Serial No. 238).
- Atkinson, R. C. (1968). Computerized instruction and the learning process. *American Psychologist*, 23, 225-239.
- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation*, Vol. 2 (pp. 89-195). New York: Academic Press.
- Azar, B. (1994a, August). Computers create global research lab. *APA Monitor*, 1, 16.
- Azar, B. (1994b, August). Research made easier by computer networks. *APA Monitor*, 16.
- Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist*, 44, 1185-1193.

- Barker, R. G., Wright, H. F., Schoggen, M. F., & Barker, L. S. (1978). Day in the life of Mary Ennis. In R. G. Barker et al. (Eds.), *Habitats, environments, and human behavior* (pp. 51-98). San Francisco: Jossey-Bass.
- Baron, J. N., & Reiss, P. C. (1985). Same time, next year: Aggregate analyses of the mass media and violent behavior. *American Sociological Review*, 50, 347-363.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-660.
- Bartholomew, G. A. (1982). Scientific innovation and creativity: A zoologist's point of view. *American Zoologist*, 22, 227-335.
- Baumeister, R. F. (1995). Disputing the effects of championship pressures and home audiences. *Journal of Personality and Social Psychology*, 68, 644-648.
- Baumeister, R. F., Campbell, J. D., Krueger, J. L., & Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *Psychological Science in the Public Interest*, 4, 1-44.
- Baumrind, D. (1985). Research using intentional deception: Ethical issues revisited. *American Psychologist*, 40, 165-174.
- Bazzini, D. G., & Shaffer, D. R. (1999). Resisting temptation revisited: Devaluation versus enhancement of an attractive suitor by exclusive and nonexclusive daters. *Personality and Social Psychology Bulletin*, 25, 162-176.
- Behnke, S. (2003). Academic and clinical training under APA's new ethics code. *Monitor on Psychology*, 34, 64.
- Berk, R. A., Boruch, R. F., Chambers, D. L., Rossi, P. H., & Witte, A. D. (1987). Social policy experimentation: A position paper. In D. S. Cordray & M. W. Lipsey (Eds.), *Evaluation Studies Review Annual*, Vol. 11 (pp. 630-672). Newbury Park, CA: Sage.
- Bickman, L. (1976). Observational methods. In C. Sellitz, L. S. Wrightsman, & S. W. Cook (Eds.), *Research methods in social relations* (pp. 251-290). New York: Holt, Rinehart and Winston.
- Birnbaum, M. H. (2000). Decision making in the lab and on the Web. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 3-34). San Diego, CA: Academic Press.
- Blanchard, F. A., Crandall, C. S., Brigham, J. C., & Vaughn, L. A. (1994). Condemning and condoning racism: A social context approach to interracial settings. *Journal of Applied Psychology*, 79, 993-997.
- Blanck, P. D., Bellack, A. S., Rosnow, R. L., Rotheram-Borus, M. J., & Schooler, N. R. (1992). Scientific rewards and conflicts of ethical choices in human subjects research. *American Psychologist*, 47, 959-965.
- Bolgar, H. (1965). The case study method. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 28-39). New York: McGraw-Hill.
- Bond, C. F., Jr., & Titus, L. J. (1983). Social facilitation: A meta-analysis of 241 studies. *Psychological Bulletin*, 94, 265-292.
- Boring, E. G. (1954). The nature and history of experimental control. *American Journal of Psychology*, 67, 573-589.
- Brainerd, C. J. (1978). *Piaget's theory of intelligence*. Englewood Cliffs, NJ: Prentice-Hall.
- Brandt, R. M. (1972). *Studying behavior in natural settings*. New York: Holt, Rinehart and Winston: University Press of America, 1981.
- Bröder, A. (1998). Deception can be acceptable. *American Psychologist*, 53, 805-806.
- Brown, R., & Kulik, J. (1977). Flashbulb memories. *Cognition*, 5, 73-99.
- Buchanan, T. (2000). Potential of the Internet for personality research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 121-139). San Diego, CA: Academic Press.
- Bushman, B. J., & Cantor, J. (2003). Media ratings for violence and sex: Implications for policymakers and parents. *American Psychologist*, 58, 130-141.
- Bushman, B. J., & Phillips, C. M. (2001). If the television program bleeds, memory for the advertisement recedes. *Current Directions in Psychological Science*, 10, 43-47.
- Campbell, A. (1981). *The sense of well-being in America*. New York: McGraw-Hill.
- Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, 24, 409-429.
- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Campbell, R. (Ed.). (1992). *Mental lives: Case studies in cognition*. Oxford: Blackwell.

- the nature of pictures. *Psychological Science*, 9, 205-210.
- Dickie, J. R. (1987). Interrelationships within the mother-father-infant triad. In P. W. Berman, & F. A. Pedersen (Eds.), *Men's transitions to parenthood: Longitudinal studies of early family experience* (pp. 113-143). Hillsdale, NJ: Erlbaum.
- Diener, E., & Crandall, R. (1978). *Ethics in social and behavioral research*. Chicago: The University of Chicago Press.
- Dittmar, M. L., Berch, D. B., & Warm, J. S. (1982). Sustained visual attention in deaf and hearing adults. *Bulletin of the Psychonomic Society*, 19, 339-342.
- Dolan, C. A., Sherwood, A., & Light, K. C. (1992). Cognitive coping strategies and blood pressure responses to real-life stress in healthy young men. *Health Psychology*, 11, 233-240.
- Eagly, A. H., & Steffen, V. J. (1984). Gender stereotypes stem from the distribution of women and men into social roles. *Journal of Personality and Social Psychology*, 46, 735-754.
- Eibl-Eibesfeldt, I. (1975). *Ethology: The biology of behavior*. New York: Holt, Rinehart and Winston.
- Endersby, J. W., & Towle, M. J. (1996). Tailgate partisanship: Political and social expression through bumper stickers. *The Social Science Journal*, 33, 307-319.
- Entwisle, D. R., & Astone, N. M. (1994). Some practical guidelines for measuring youth's race/ethnicity and socioeconomic status. *Child Development*, 65, 1521-1540.
- Epley, N., & Huff, C. (1998). Suspicion, affective response, and educational benefit as a result of deception in psychology research. *Personality and Social Psychology Bulletin*, 24, 759-768.
- Epstein, S. (1979). The stability of behavior: On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, 37, 1097-1126.
- Erber, R. (1991). Affective and semantic priming: Effects of mood on category accessibility and inference. *Journal of Experimental Social Psychology*, 27, 480-498.
- Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, 49, 725-747.
- Estes, W. K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review*, 4, 330-341.
- Evans, R., & Donnerstein, E. (1974). Some implications for psychological research of early versus late term participation by college students. *Journal of Research in Personality*, 8, 102-109.
- Feeney, D. M. (1987). Human rights and animal welfare. *American Psychologist*, 42, 593-599.
- Festinger, L., Riecken, H., & Schachter, S. (1956). *When prophecy fails*. Minneapolis: University of Minnesota Press.
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think. *Psychological Science*, 15, 119-126.
- Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory & Psychology*, 12, 825-853.
- Fine, M. A., & Kurdek, L. A. (1993). Reflections on determining authorship credit and authorship order on faculty-student collaborations. *American Psychologist*, 48, 1141-1147.
- Fisher, C. B., & Fryberg, D. (1994). Participant partners: College students weigh the costs and benefits of deceptive research. *American Psychologist*, 49, 417-427.
- Fossey, D. (1981). Imperiled giants of the forest. *National Geographic*, 159, 501-523.
- Fossey, D. (1983). *Gorillas in the mist*. Boston: Houghton-Mifflin.
- Fowler, R. D. (1992). Report of the chief executive officer: A year of building for the future. *American Psychologist*, 47, 876-883.
- Fraley, R. C. (2004). *How to conduct behavioral research over the Internet*. New York: Guilford Press.
- Frame, C. L., & Strauss, C. C. (1987). Parental informed consent and sample bias in grade-school children. *Journal of Social and Clinical Psychology*, 5, 227-236.
- Frank, M. G., & Gilovich, T. (1988). The dark side of self- and social perception: Black uniforms and aggression in professional sports. *Journal of Personality and Social Psychology*, 54, 74-85.
- Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition*, 23, 132-138.
- Friedman, H. S., Tucker, J. S., Schwartz, J. E., Tomlinson-Keasy, C., Martin, L. R., Wingard, D. L., & Criqui, M. H. (1995). Psychosocial and

[General Information]

书名=心理学研究方法 原书第7版

作者= (美) 约翰·肖内西, 尤金·泽克迈斯特, 珍妮·泽克迈斯特著

页数=436

ISBN=1978-7-115-22326-5

出版社=北京市:人民邮电出版社

出版日期=2010.04

SS号=12627593

DX号=000006874652

did=n25

PdgPath=6461615E616A60959AA49C74737D6074737D6267606163603131363132373031

RealLink=http://tta300.5read.com/300-09/diskCBL/CBL16/02/!00001.pdg

Purl=http://image1.5read.com/image/ss2jpg.dll?did=n25&pid=88060493745A2411AFF60EF0D69C657747D9465F8FC20D3C30C0B6AE22373AE233891DA662F5FC844CA37E4208EEC8D8D3058C6FD63CA2BDF13645E4B56AC91B517A3AB0B0A1508E71A3F91DD45CC8B7ABECBBAE2AEB908933C6FAD97B89F719FD8632E1382AB438262D838C02E7B18920E&jid=/000001.jpg

封面
书名
版权
前言
目录

第一编

一般问题

第1章

绪论

科学心理学

科学的背景

历史背景

社会和文化背景

道德背景

像研究者一样思考

评价媒体报道中的研究发现

开始做研究

小结

第2章

科学方法

科学方法和日常方法

一般方法

观察

报告

概念

仪器

测量

假设

科学方法的目标

描述

预测

解释

应用

科学理论的建构和检验

小结

第3章

心理学研究中的伦理问题

引言

研究前需要思考的伦理问题

A P A 伦理标准

注解

风险收益比

风险判定

最小风险

风险应对

知情同意

A P A 伦理标准

注解

心理学研究中的欺骗

A P A 伦理标准

注解

事后解释

A P A 伦理标准

注解

用动物做研究

A P A 伦理标准

注解

心理学研究的报告

A P A 伦理标准

注解

伦理决策的步骤

小结

第二编

描述性方法

第4章

观察法

概述

行为样本

时间抽样

情境抽样

观察法的分类

非干预观察
干预观察
参与性观察
结构化观察
现场实验法
行为记录
行为的定性记录
行为的定量测量
观察数据的分析
数据简化
观察者信度
批判性思考观察研究
观察者的影响
观察者偏差
小结

第 5 章 调查研究

概要
调查研究的使用
调查研究的特点
调查研究中的抽样
抽样的基本术语
抽样方法
调查方法
邮寄调查法
个人面谈法
电话访谈法
网络调查法
调查研究的设计
横断设计
连续独立样本设计
纵向设计
问卷
作为调查工具的问卷
自陈测量的信度和效度
编制问卷
关于调查研究的批判性思考
报告行为与实际行为的一致性问题
相关与因果关系
小结

第 6 章 行为的隐蔽测量

概述
物理痕迹
基本原理
物理痕迹的类型
问题与局限
档案数据
基本原理
档案数据的类型
内容分析
问题与局限
伦理问题与隐蔽测量
小结

第三编 实验方法

第 7 章 独立组设计

概述
心理学家为什么要做实验
实验研究的逻辑
随机组设计
随机分配
对内部效度的威胁
实验结果的解释
数据分析在实验中的作用
描述结果
确认结果所揭示的内容
数据分析所不能告诉我们的

确立实验结果的外部效度

匹配组设计

自然组设计

小结

第 8 章 重复测量设计

概述

研究者为什么使用重复测量设计

练习效应在重复测量设计中的作用

练习效应的定义

在完全设计中平衡练习效应

在不完全设计中平衡练习效应

重复测量设计的数据分析

描述结果

确定结果所揭示的内容

差异转换问题

小结

第 9 章 复合设计

概述

确定实验设计的指导原则

描述复合设计中的实验效应

2×2 设计举例

主效应和交互作用

描述交互作用

复合设计的数据分析

存在交互作用的数据分析

不存在交互作用的数据分析

解释交互作用

交互作用和理论检验

交互作用与外部效度

交互作用与天花板效应和地板效应

交互作用和自然组设计

小结

第四编 应用研究

第 10 章 个案研究法和单被试设计

概述

个案研究法

个案研究法的特点

个案研究法的优点

个案研究法的缺陷

对个案研究证据的批判性思考

单被试（小样本）实验设计

单被试实验的主要特点

具体的实验设计

所有单被试设计共同的问题和局限性

小结

第 11 章 准实验设计和项目评估

概述

真实验

真实验的特点

在自然环境中进行真实验的困难

控制真实验的内部效度

真实验也可能控制不到的问题

准实验

非等控制组设计

非等控制组设计：兰格和罗丁的研究

影响非等控制组设计内部效度的因素

外部效度问题

间断时间系列设计

非等控制组的时间系列

项目评估

小结

第五编 分析和报告研究结果

第 12 章 数据分析与解释（I）：描述数据，置信区间，相关

概述

呈现研究结论

计算机辅助的数据分析

例证：一个实验研究的数据分析：均数比较

第一阶段：了解数据

第二阶段：概括数据

第三阶段：用置信区间来判断数据所反映的意义

例证：一个相关研究的数据分析

第一阶段：熟悉数据

第二阶段：概括数据

第三阶段：为相关构建一个置信区间

小结

第13章 数据分析与解释（II）：统计显著性检验与分析

概述

虚无假设显著性检验（NHST）

实验敏感性和统计效力

NHST：两个平均数的比较

独立组设计

重复测量设计

统计显著性和理论及实际意义

均数比较的建议

均数比较的结果报告

多重条件的数据分析

单因素独立组方差分析

三个以上独立组设计的效应大小计算

独立组设计的统计效力估计

多组实验设计的均数比较

单因素独立组实验设计的结果报告

重复测量的方差分析

独立组设计的双因素方差分析

含有交互作用的复合设计分析

无交互作用的方差分析

双因素独立组设计的效应大小计算

复合设计分析中置信区间的作用

双因素混合设计的方差分析

复合设计的结果报告

小结

第14章 心理学的学术交流

概述

互联网与心理学研究

研究报告的写作指导

研究报告的结构

标题页

摘要

引言

方法

结果

讨论

参考文献

附录

作者注释

脚注

文稿各页的排序

口头报告

研究计划

研究报告写作范文

附录统计表

术语表

参考文献 433