

有关行为研究方法学的六点思考

王晓田

(Psychology Department, University of South Dakota, USA)

摘要 本文提供作者的研究经验和反思,就行为研究方法学的多样性和互补性,统计方法与实验想法的权衡,信-效度的关系,被试与测试者期望效应对数据的污染,避免方法学和统计检验的规定成为束缚研究的新八股,以及理论与实证的互动进行了探讨。

关键词 行为研究方法

分类号 B841

1 三足鼎立的行为研究方法

大约 8 年前,在一项由美国国家基金会(NSF)资助的项目中,我们在美国中西部的农业区发放了一批问卷。其中有一道情景题,要求参与者根据价钱的高低和对环境污染的程度选择施肥的种类。在回收的问卷中有一封来自一位快 70 岁的农场主的短信。在信中他对我们研究的真实性提出了质疑。他问道:“您以为 EPA (环境保护署)会允许我选用污染程度高的化肥吗?”他还提醒说,如果进去的是垃圾,出来还是垃圾 (garbage in, garbage out)。也就是说,问题的情景不真实,答案也不会真实。还有一个问题他可能不知道,知道了也许会更生气。就是那些情景题是“做了手脚的”。用实验心理学的术语说就是用了被试间设计 (between-subject design)。以化肥选择题为例,每位被试在其它条件随机化的情况下,所收到的选择题在化肥的价钱和对环境污染的程度上各有不同。结果表明,这两个实验变量在被试不知情的情况下,影响了他们对化肥种类的选择。如此看来,我们对真实性的定义可能有层次上的不同:它既可以定义为日常生活中具体的真实,也可以是心理层面的真实。而实验不是一种现实的复制而是一种简化了的现实。

然而,那位农场主提出的问题是很有道理的。心理学研究的终极对象是无法直观的心理机制,这使得研究者处于一种盲人摸象的状态。要避免片面性,

就需要不同的研究者和不同的研究方法不断进行交流和整合。

研究的方法各有侧重,但是没有所谓高下之分。如果我们把心理学或行为学的研究视作一个三足鼎立的三角结构的话,那它的三个端点可以分别是准确性、普适性和真实性。而一种研究方法往往无法同时兼顾到这三项指标,比如,实验研究的准确性高,可以验证因果关系,但真实性因实验控制而降低,普适性也因样本等原因而受到限制。案例分析 and 实地研究的真实性高,但因为揭示的是变量之间的相关关系而不一定是因果关系而降低了研究的准确性和普适性。理论研究和行为模型的建构则能够以高屋建瓴之势,收纲举目张之效。因为是纲,普适性则高。也因为纲,则不免抽象,因而使准确性和真实性降低。

如图 1 所示,心理学的研究方法还可以根据三个不同的维度进行分类 (Hendricks, Marvel, & Barrington, 1990)。这三个维度分别是实验环境、数据收集的方法和研究设计。其中,实验环境包括实验室(L)和实地(F)两类,数据方法包括自我报告(S)和观测(O)两种,而研究设计则分为实验(E),相关性研究(C)和描述性研究(D)三个类型。由此得出 12 种 (2×2×3) 研究类型。上述包括化肥选择题在内的实验性问卷研究则可归入 FSE 一类。对研究方法的分类既有助于我们认清自身研究的现状又有助于对进行互补性研究的规划。

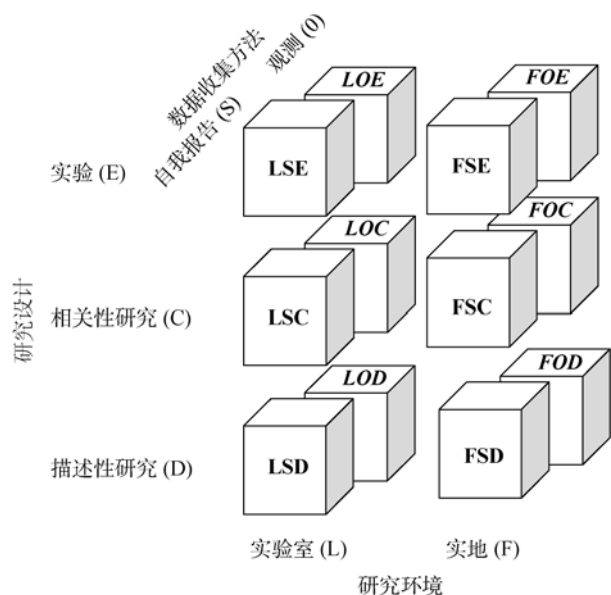


图 1 三维研究方法分类图 (参照 Hendricks, Marvel, & Barrington, 1990)

2 实验设计与统计方法：谁是本谁是末

一项研究就像一段探索的行程，有着自身的路径。这其中的过程固然重要，但结果才是衡量成败的标准。如何从起点到达目的地的关键在于既要充分了解起点所处的位置，也就是历史的和目前的研究状况，又要推断目的地的远近和可能的通路。研究方法也就成了爬山涉水，披荆斩棘所需的工具或是从 A 处到 B 处的桥梁。因此，研究的方法既不可能一成不变又不可能预先设定。所以，许多实验方法论中提示的先选好统计方法再因材施教地设计实验的要求是一种本末倒置。

我当助理教授的第一年，一次和系里两位教授聊到新的实验设想，其中一位在听了我的简单介绍后问道：“那你准备用什么统计方法分析数据？”我回答说：“还没想好。”他显得很吃惊地说：“不知道方法和工具怎能开始设计呢？”这也像是说没有练好手艺就不要去揽活计。尽管有道理，我还是觉得这是一种取舍。我宁愿让方法迁就想法，而不愿让想法迁就方法。

3 高信度低效度的危险：一个跨文化研究的例子

一项研究的信度(reliability)主要指它的方法的可靠性和结果的稳定性(可重复性)。而一项研究的效度(validity)则主要反映在它的内容、标准和概念

的正确性上。也就是研究观测的内容和指标是否真正反映了所想要探讨的问题。在心理学研究中，对信度和效度之间关系的一个常用的比喻是打靶(Trochim, 2004)：射击的效度由着弹点与靶心的距离决定，正中靶心的射击成绩效度最高。射击的信度则反映在弹点之间的距离，着弹点越集中，信度就越高。图 2 显示了三种信-效度之间的关系。因为效度比信度更重要，研究者要格外注意图 2 中的高信度低效度的危险。

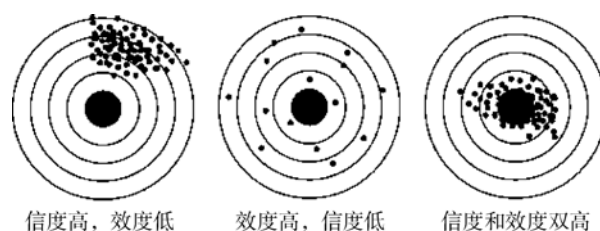


图 2 信度和效度的关系 (参照 Trochim, 2004)

对于高信度低效度的危险以及可能的防范方法，我自己的一项研究(Wang, 1996)可以作为一个旁注。这项研究利用 Tversky 和 Kahneman (1981)的框架效应(framing effect)问题对决策的机制进行了新的探讨。Kahneman 和 Tversky 用著名的“亚洲疾病”问题昭示了称为框架效应的一种非理性决策现象。假设一种可以致命的亚洲疾病感染了 600 个人，而只有两种可行的救治方案可供选择。如果选择方案 A，将有 200 人被救活；如果选择方案 B 将有 1/3 的几率救活 600 人，2/3 的几率无人获救。当用存活率对预期结果进行了上述的描述(框架)后，大多数的被试(72%)选择规避风险的方案 A。然而，当用死亡率对同样的预期结果进行框架(选择方案 A，将有 400 人死亡；选择方案 B，将有 1/3 的几率无人死亡，2/3 的几率 600 人死亡)，人们的风险偏好发生了反转，大多数的被试(78%)选择了冒险的方案 B。这种仅仅因为对同一问题和预期结果的不同的语言表述而发生的风险偏好反转成为人类决策的非理性的一个经典例子。从组织学的角度出发，我们发现在“亚洲疾病”问题中有一个尚未被注意到的变量，那就是危险群体的大小。基于一种进化论的考量，人类进化的绝大部分时间是在一种小群体的环境中度过的。这也提示我们，人的认知和决策的心理机制也应该是有选择性地适应于这种典型的小群体的生态和社会环境。当人们需要为众多陌生人的生命或财产作决定时，这种进化中非典型

的生态环境可能使决策者缺乏可供选择的适应性策略，转而求助于语言语气等其它相对次要的决策线索，因而产生了语言的框架效应（更详细的理论及机制探讨参见 Wang, 2008）。实验的结果支持了这一理论假说，当疾病问题涉及到的人数众多时（6000，或更多），被试的风险选择出现了框架效应。然而，当疾病问题涉及的人数在人类进化中常见的群体大小范围内时（60，6，或更少），人们的风险偏好不再受语言框架的左右，表现出了一种同舟共济的冒险倾向。美国的样本和中国的样本所得到的结果相当一致。唯一的文化差异表现在从有框架效应到没有框架效应的转折点上。中国被试的数据显示，框架效应发生在危险群体的大小为 6000 而非 600 时（Wang, 1996）。这一现象提示中国人的相关群体的概念涵盖的人数更多。有意思的是，如果我们没有从理论出发对同一疾病问题在不同的群体情境中进行检验，而只是简单地用中文版的“亚洲疾病”问题对框架效应进行信效度检验，所得出的结论将会大相径庭。首先，在人数为 600 的情况下没有发现框架效应，可能导致对框架效应构念效度（construct validity）的错误性的置疑。另外，这种结果的信度可以因为可重复而增高，但由此而得出的任何结论很可能是无效的。这就像只在黑屋子里研究颜色视觉，得出的结果可以是可靠的，但得出的结论（被试没有颜色视觉）却是不正确的。避免研究中出现高信度低效度问题的一个方法就是在理论的指导下对一种现象在多种环境或情境中进行验证，不仅要了解导致一种现象发生的前因后果还要探寻抑制此种现象发生的条件。

4 聪明的被试和无知的主试：期望效应的污染

我曾多次听说国内一家大型的 IT 企业对组织行为学的研究十分重视，不仅对员工的能力考核十分详尽，而且配备了专职的人员帮助或替代企业外的研究人员在专门的测试中心进行实验和访谈。听了之后，我既为研究工作在企业受到重视而感到高兴，却又禁不住为可能存在的隐患感到担忧。可以想象的是，当一群员工被请到企业的测试中心参加某项研究时，当员工们得知参加研究的时间工资照付时，当他（她）们见到企业内的专职人员将成为主试人员时，这些员工很可能会感到一种组织压力或期望，也可能产生一种群体思维（groupthink）的倾向。面对主试人员提出的问题时就可能自觉或不自觉

地“投其所好”，因而使得收集到的数据出现期望效应的污染。

被试不是被动的，主试也可能不自觉地为主试所控。这样的例子在现实中并不少见，一个世纪之前，心理学家 Oskar Pfungst (1911) 用实验的方法纠正了人们盛传的有关一匹叫聪明的汉斯（Clever Hans）的马的传言。汉斯的主人 von Osten 多次演示这匹马可以通过摇头或点头或使用前蹄连续敲地的方法准确回答用不同语言写出和说出的有关地理、历史、科学、文字和数字运算的多种问题。汉斯在主人不在场的情况下依然表现出色。许多人，包括一些知名的动物学家和心理学家均对这匹叫汉斯的马进行过测试，但没有人发现任何破绽。直到 1907 年 Pfungst 采用了实验控制的方法逐步分离出了汉斯回答正确的条件和导致回答不正确时的条件。在戴上眼罩的情况下，汉斯的表现变得十分糟糕。进一步的实验发现，在缺乏有关测试者的视觉信息时汉斯便无法正确回答问题。这匹神奇的马的确非常聪明，它通过观察测试它的人的眼神，头部动作和面部度表情的细微变化精准地猜出各种问题的预期答案。从这一案例中我们可以看到，在观察者和被观察者互有期望的情况下，实验的方法比问卷和访谈等方法更可能避免期望效应的误导，因而也更为可靠。

然而，就是在有实验控制的情况下，如果数据的记录是人为的，期望效应的污染仍然可以发生。Rosenthal (1976) 的研究告诉我们，主试（实验）人员对实验结果的预期可以明显干扰和降低实验结果的准确性和可靠性。在这样的情况下，“双盲”的实验设计可以使被试和实验记录者双方都不了解实验的目的，从而防止或降低期望效应对数据的污染。

5 避免方法学和统计检验的规定成为束缚研究的新八股

近年来，一些学者开始质疑心理学中制度化了的显著性检验。我们常常用否定无差别假设（null hypothesis）的检验结果作为支持与之相对立的工作假设（alternative hypothesis）的证据。但是，显著性检验并没有对工作假说进行任何具体的评估。另外，心理学中常用的一些数据分析的方法无视统计理论中的根本性争论，把有争议的甚至是对立的统计理念混为一体（Gigerenzer, 1998, 2004）。

心理学著名学术期刊大多采用 0.05 作为显著

性检验的门槛。这也造成了研究中的一些困境。我的一位硕士研究生的实验假说是 X 效应在传统的情景 A 中出现,而在情景 B 中减弱或消失。她在对数据做了初步分析后告诉我实验预期得到了支持。可是在她几天后交给我的报告中却没有对这一部分的数据作具体的描述,而只是简单地报告说两种情景下 X 效应均不显著,因此,工作假说被否定。询问之下我了解到,预测的 X 效应的 p 值为 0.053,于是她去向一位统计学教授讨教。听起来她得到的回答大约是, p 值超过了 0.05 就是无效应,也就意味着工作假说被否定。我建议她两种情景下的结果都要给出具体的数据,并且不要轻易放弃工作假说。显著性检验所检验的只是无差别假说;无差别假说没有被否定并不说明另一个假说就因此被否定了。在这个例子中,统计学教授扮演了宗教仪式中牧师的角色,我们虚心好学的二年级研究生的思路在 $p < 0.05$ 的铁律面前嘎然而止。这 0.05 就像是研究思路上的—道鬼门关,闯过去了便无需再回头,没闯过去则不再向前看。科学精神中所必需的批判性和怀疑性既不提倡轻易的否定也不提倡轻易的肯定。因此,作为科学研究工具的统计分析和实验设计要避免成为一种新的科举式的死板框定。

6 题目(Topic) 不是想法 (Idea), 方法不是思路

近年来国内行为学和管理学的研究在掌握和应用实验设计和统计方法方面取得了长足的进步。但是,不少研究计划的不足之处在过于庞大,不够集中。研究者交流很少,各自为政。许多研究计划的套路相似,多是重新造车。从深度访谈开始,到量表开发,再到信效度检验,最后用回归和结构方程从数据中找关系。这样既造成一定程度的重复和浪费又因为铺得很开而不易深入。造成这种循环性浪费的一个潜在原因是这些研究不是从想法和灵感入手而是在选题后直接进入到实验设计。在课题确定后便使用研究方法开路。这种数据发掘(Data

Mining)和撒大网捕鱼(Fishing Expedition) 的策略很难形成理论体系。

Gerd Gigerenzer (1998, p. 202) 说:“没有理论的数据就像没有爹娘的孤儿,它们的预期寿命也因此而缩短。”有想法再加上理论的指导才可能形成有价值的思路。在《改变心理学的 40 项研究》—书中, Hock (2002)介绍了 40 项著名的心理学研究的产生渊源和背景。这 40 项心理学实验都是有感而发,没有一项是淘金的结果。这些研究或起源于个人经历和经验的启发,或出于对—种普遍现象的好奇和置疑,或受到宏观理论或微观理论的启发,或得益于—种新的的研究手段,或基于对某种行为的功能性设想,抑或是跨学科思想和成果的融合。这些影响深远的心理学研究提示我们题目不能替代想法,方法不能替代思路。而有了想法和思路则可以因地制宜地寻找和发现有用的方法和工具。

参 考 文 献

- Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology*, 8, 195–204.
- Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, 33, 587–606.
- Hendricks, B., Marvel, M. K., & Barrington, B. L. (1990). The dimensions of psychological research. *Teaching of Psychology*, 17, 76–82.
- Hock, R. R. (2009). *Forty Studies that Changed Psychology: Explorations into the History of Psychological Research* (6th Ed.). Upper Saddle River, NJ: Prentice Hall.
- Pfungst, O. (1911; reprinted 1965). *Clever Hans: The horse of Mr. von Osten* (C. L. Rahn, Trans.). New York: Holt, Rinehart & Winston.
- Rosenthal, R. 1976. *Experimenter Effects in Behavioral Research*. New York: John Wiley.
- Trochim, W. M. K. (2004). *Research Methods: The Concise Knowledge Base*. Florence, Kentucky, Cengage Custom Publishing
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453–458.
- Wang, X. T. (1996). Domain-specific rationality in human choices: Violations of utility axioms and social contexts. *Cognition*, 60, 31–63.
- Wang, X. T. (2008). Risk communication and risky choice in context: Ambiguity and ambivalence hypothesis. *Annals of the New York Academy of Sciences*, 1128, 78–89.

Six Issues in Methodology of Behavioral Research

Xiao-Tian WANG

(Psychology Department, University of South Dakota, USA)

Abstract The paper provides reflections on taxonomy of complementary research methods, tradeoffs between statistical methods and research ideas, the relationship between reliability and validity, data contamination by expectancy of the subjects or experimenters, the need for avoiding methodological and statistical rituals that may hamper research development, and the synergy of theory and data.

Key words methodology in behavioral research