迷失的边界：心理学虚无假设检验方法探究

焦　璨　张敏强

摘　要：源于统计学的虚无假设检验自引入到心理学研究后，便得到了广泛的运用，成为科学主义心理学兴起以来运用最广泛、影响力最大的心理学技术，但国内外心理学界对虚无假设检验方法的使用皆存在诸多的问题和错误，突破了其应有的边界。究其历史及发展，对样本容量的依赖，显著性结论的不确定性，以及无法提供结果的可重复性等是虚无假设检验方法被集中诟病之处。效果量是影响统计功效的重要因素。引入效果量的概念，报告和解释效果量，将效果量作为研究结果不可或缺的部分，是对虚无假设检验理论与技术的完善。

关键词：虚无假设检验　科学主义心理学　效果量

作者焦璨，心理学博士，深圳大学师范学院心理学系副教授（深圳　518060）；张敏强，华南师范大学心理学院、心理应用研究中心教授（广州　510631）。

脱胎于哲学的心理学，在其最初的发展阶段，弥散着的是厚重的人文主义精神，但自从科学主义被引入进来后，人文主义色彩就逐渐淡去，科学主义色彩则逐渐增强，最终占据了心理学研究的统治地位。E. H. 韦伯首先将实验法应用于心理学，冯特在此基础上创建了实验心理学，而以华生为代表的行为主义则使实证研究方法日臻成熟和完善。科学主义心理学以经验自然主义为楷模，强调客观、量化的实证研究方法，统计方法等被尽可能地运用到心理学研究中来，成为实证研究佐证研究假设的主要手段，也成为心理测量、心理与教育评价等分支学科发展的基石。应该说，在促进心理学从哲学母体中分离出来，成为一门独立的科学，并使心理学朝客观化、精确化方向发展的历史进程中，科学主义心理学曾经产生过重要的作用。

20世纪40年代，源于统计学的虚无假设检验（null statistical hypothesis testing）方法被引入心理学研究，并在此后的70余年间得到了广泛的运用，成为科学主义心理学兴起以来运用最广泛、影响力最大的心理学技术。① 但问题随之而来，

① 以目前在中国最具影响力的《心理学报》、《心理科学》的虚无假设检验应用情况作元分析，

由于过分强调虚无假设检验方法的使用，而较少关注其使用的边界，导致了在心理学研究中，只有那些具备了虚无假设检验显著性结论的研究论文，才有更多机会在重要的心理学刊物上发表，而未得到显著性结论的研究论文以及未采用虚无假设检验方法的成果，则常常被束之高阁。

本文不准备把心理学中的科学主义和人文主义分出高下优劣，仅仅希望从统计方法的角度，以虚无假设检验方法为例，从其使用、发展的历史入手，反思虚无假设检验方法在心理学研究中的应用，以为心理学及相关学科的学者从事相关研究，提供另一种角度的思考。

一、虚无假设检验的历史发展

有学者认为，统计经常被滥用和误用，主要原因还是在于应用工作者的统计知识不足。① 不过，依笔者之见，这种知识似乎还应包括对统计方法的发展历史的了解。正因为如此，从历史的角度出发，探讨虚无假设检验的误用、滥用的根源，对于研究者而言，具有正本清源的重要意义。

在心理学量化研究中，运用虚无假设检验佐证心理学研究成果的历史悠久。目前，虚无假设检验已是心理学研究的重要研究工具和方法，也成为了研究论文接受、发表的重要标准之一。

作为一种正规的统计方法，虚无假设检验是在同一数据基础上对两个或者两个以上假设统计模型的检验，是在“虚无假设（null hypothesis）H_0 为真”的前提下计算以往观测到的结果所发生的概率，其理论依据是“小概率事件在一次试验中几乎不可能发生”的原理。虚无假设检验为我们提供了拒绝某一假设的工具。

现有的一些数理统计学以及心理统计学、教育统计学等应用统计学教材对于虚无假设检验的介绍、讨论，使读者产生错觉，认为虚无假设检验是一套固有的理论，将之视为统一的、没有争议的理论方法。实则不然，目前我们广泛使用的虚无假设检验是种不完备的结合体，即 20 世纪 20 年代、30 年代费希尔（R. A. Fisher）的显著性检验（significance testing）和奈曼—皮尔逊（Neyman-Pearson）的假设检验

① 从 1998 年到 2007 年，两个杂志总共刊发 4115 篇论文，其中使用虚无假设检验以佐证研究结论的有 2752 篇，占总数的 67%；而作为目前国内最权威的心理学专业刊物的《心理学报》，更是达到了 84.8%。（参见焦璨等：《心理研究中统计方法应用的元分析》，《心理科学》2010 年第 33 期）此外，2008 年至 2012 年仍然保持这一趋势，两个刊物发表论文中使用虚无假设检验的占 72.08%，其中《心理学报》占 75.13%，《心理科学》占 70.08%。

① 参见温忠麟：《屡遭误用和错批的心理统计》，《华南师范大学学报》2010 年第 1 期。

(hypothesis testing) 的结合体。① 这是现代统计学发展史上影响颇为重大且又相互矛盾的两大学派。尊为“现代统计学之父”的费希尔发展了从测量数据中产生显著性水平的方法，而奈曼和皮尔逊则提出了一套严格地拒绝某一事先确定的假设的决策程序。

(一) 费希尔的显著性检验模式

费希尔的显著性检验模式一般包括以下几个步骤：②

(1) 确定假设 H；(2) 决定合适的检验统计量及其在 H 为真的前提下的分布；(3) 从测量数据中计算检验统计量 T；(4) 使用 H 为真的前提下 T 的分布，确定与之相对应的显著性水平 P；(5) 如果获得的显著性水平非常小，则拒绝 H。值得注意的是：这是显著性检验可以得到的唯一的结论。

在费希尔的理论框架中，仅设置了一个关于检验统计量 T 的已知分布的假设 H。若该检验统计量和其条件期望值 E (T/H) 相差甚远，假设则不大可能偶然发生。

显著性、P 值都是其理论中重要的概念。显著性指“概率低到足以拒绝的程度”，更直观的解释是“计算结果出现的概率很低”。P 值是判定是否有显著性的概率，费希尔笃信显著性检验只有在连续实验的相互联系中才有意义，所有这些实验都是为了解释某一特定处理的作用。使用显著性检验可以得到以下三种可能的结论：(1) 若 P 值很小（通常小于 0.01），则可以断言某种结果已经显现出来；(2) 若 P 值很大（通常大于 0.2），即使某一结果真实存在，进行大规模的实验也会因为该结果发生的可能性太小而不大可能得到这一结果；(3) 若 P 值介于两者之间，应该改进实验设计以得到一个更好的结果。

(二) 奈曼—皮尔逊的假设检验模式

奈曼—皮尔逊的假设检验模式一般包括以下几个步骤：③

(1) 确定零假设 H_0 和备择假设 H_1；(2) 决定合适的检验统计量及其在 H_0 为真的前提下的分布；(3) 指定显著性水平 α，并决定与之相应的检验统计量在 H_0 为真的前提下的临界值 C_α；(4) 从测量数据中计算检验统计量 T；(5) 若 C_α 和 T 相差很大，则拒绝 H_0；反之，则不能拒绝 H_0。

奈曼和皮尔逊认为至少有两个可能的假设，显著性检验才有意义。他们把费希尔理论中被检验的假设称之为零假设 H_0 (nil hypothesis，亦称为原假设)，其他可能的假设为备择假设 H_1 (alternative hypothesis，亦称为研究假设)。检验零假设

① 参见 W. Hager，“The Statistical Theories of Fisher and of Neyman and Pearson: A Methodological Perspective,” *Theory & Psychology*, vol. 23, no. 2, 2013, pp. 251-270.

②③ 参见 J. Gill，“The Insignificance of Null Hypothesis Significance Testing,” *Political Research Quarterly*, vol. 52, no. 3, 1999, pp. 647-674.

H_0，必须要有一组定义明确的备择假设 H_1。当一个备择假设 H_1 为真时，则该备择假设被接受的概率称之为该检验的统计功效（statistical power），是衡量一个检验方法好坏的指标。

假设检验的目的是用来推翻零假设 H_0。零假设就是研究者所要攻击的“靶子”，应该被研究结果所推翻。因此，根据奈曼的思想，实验研究设计必须使最终数据有最大的检验效力，才能拒绝零假设 H_0，即表明差异存在、实验处理有效。

该模式中计算 P 值的目的是为了检验 H_0。奈曼认为 P 值是通过计算得到的与观测值有关的理论概率：从长期来看该观测值发生次数的比率，它与现实没有联系，是对似是而非问题的间接测量。在奈曼—皮尔逊的理论中，研究者需要事先设定一个固定的值 α（比如 0.05），当 $P \leqslant 0.05$ 时，就拒绝零假设，即意味着从长期来看，该研究者正好会有 5%的机会拒绝一个正确的零假设。这样就将 P 值与现实生活联系起来，把假设检验放进一个可以计算与检验决策相联系的概率的架构中，以决定哪一种检验方法比别的检验方法更好。

（三）虚无假设检验模式

费希尔流派和奈曼—皮尔逊流派坚持己见，相持不下，20 世纪 40 年代的教材编写者为了使统计检验模式成为一公认的理论范式，尝试将二者结合起来，形成了现在的虚无假设检验模式。大致包括以下几个步骤：①

（1）确定虚无假设 H_0 和备择假设 H_1；（2）决定合适的检验统计量及其在 H_0 为真的前提下的分布；（3）指定显著性水平 α；（4）从测量数据中计算 P 值；（5）若 $P \leqslant \alpha$，则拒绝 H_0；反之，则不能拒绝 H_0。

费希尔的显著性检验中，只确定了一个假设 H，并且通过测量数据计算 P 值，用以估计研究假设的证据强度。奈曼—皮尔逊的假设检验中确定了两个不对等的假设：H_0 和 H_1，并在事先确定的 α 水平基础上拒绝其中之一。费希尔将显著性水平定义为数据的函数，是后验概率；奈曼—皮尔逊将显著性水平定义为先验概率，其确定甚至先于查看数据。费希尔反对预先选择显著性水平，也反对两种强制性的决策结论。奈曼—皮尔逊不同意对 P 的解释，并认为 P 值是主观的、无用的，只有将 P 值与现实联系起来才有意义。虚无假设检验范式横跨这两个体系，选择一个先验概率 α，却使用 P 值或者用星号标识 P 值的范围，估计证据的强弱。这一范式虽然包含备择假设，却没有考虑统计功效。目前，心理学研究者常错误地将虚无假设检验视为接受某一假设并将之作为理论构建的依据。②

---

① 参见张敏强：《教育与心理统计学》，北京：人民教育出版社，2010 年，第 127—162 页。

② R. W. Frick, “The Appropriate Use of Null Hypothesis Testing,” *Psychological Methods*, vol. 1, no. 4, 1996, p. 379.

这一结合体也试图调和两种理论中对于如何定义假设的不同观点。采纳了奈曼—皮尔逊的两种假设，其中零假设 H_0 等同于费希尔的假设，却将之视为零差异、零效应，即不存在差异或效应；而费希尔仅将之定义为“无效的”，即当前实验对于证据的收集无效。这一结合体部分地使用奈曼—皮尔逊的决策过程，将没有拒绝零假设视为类似于对零假设的适度支持。在这一结合体中，混淆了费希尔中由测量数据得到的 P 值和奈曼—皮尔逊中事先确定的 α。P 值或者以星号标识的 P 值的范围不是事先确定的，不是犯Ⅰ型错误的长期概率，但是使用者常如此认为。

显然，这种结合是不完备的。Ⅰ型错误因其主观、简单，受到研究者过度关注，而Ⅱ型错误因其计算复杂几乎被研究者忽略。同样作为虚无假设检验的重要组成部分，统计功效却没有和显著性得到同等关注，以至于虚无假设检验在过去的几十年间，一直饱受批评，其源头应来自于此。

二、心理学研究中的虚无假设检验

国内外心理学研究在使用虚无假设检验时存在诸多的问题和错误，我们认为其原因主要有以下几点：

第一，心理学研究对象的特殊性。心理学的研究对象是人，是人的心理特质、心理现象，具有间接测量性、不可重复性。由于人具有学习能力及成长性，所以相对于其他自然科学实验，心理学研究的重复性、被试的选择等难度更大。对这种特殊的研究对象，虚无假设检验的使用亦应有其特殊性，并关注其可重复性，为后续研究提供借鉴。

第二，对心理学研究目的认识不足。心理研究工作者考虑了实验因素的可控性、取样的方便性，但研究目的仅停留在实验组和控制组之间、各因素及因素各水平之间是否存在差异，是否获得统计上的显著性差异。他们未意识到差异的大小所提供的信息才是真正的研究目的，① 导致研究成果与实际应用之间脱节。

第三，虚无假设检验方法应用中存在问题。尽管大部分研究者都意识到统计方法的重要性，认可统计分析方法在心理学研究中的地位和作用，并大量使用虚无假设检验。但是，学术刊物对心理统计学、心理测量学等基础方法学研究领域并不重视，未强调统计方法应用的科学性要求，加上对统计软件的依赖，使很多心理学研究者对心理统计学、心理测量学原理不求甚解，忽视各类虚无假设检验方法的使用前提，将统计分析方法直接引入各种研究过程。

第四，抽样缺乏科学性。抽样必定伴随着误差，误差有随机误差和系统误差之分。心理学研究者将样本异质性视为随机因素，将样本异质性产生的误差视为随机

① B. H. Biskin, “Comment on Significance Testing,” *Measurement and Evaluation in Counseling and Development*, vol. 31, no. 1, 1998, pp. 58-62.

误差。实际上，样本异质性产生的误差应包含随机（异质性）误差和系统（异质性）误差。其中，随机异质性是指被试之间能力、心理特质的异质性，包括被试的天赋水平和努力程度等影响因素。系统异质性是指不同被试群体之间的异质性，包括被试所处的地区、民族、家庭背景、学校背景以及教育经历等影响因素。随机异质性和系统异质性共同影响样本异质性，所以抽样时应同时考虑二者。目前，由于人、财、物等因素以及取样的困难，心理学研究常进行方便抽样，采用学生样本。发表的心理学研究论文常采用诸如“在某高校随机抽取学生××名”、“在某地区随机抽取被试××名”等被试选取方案，用以研究大学生或者其他群体的某一心理现象。这种做法将不同被试群体之间的异质性纳入到统计模型中。然而，统计方法处理的是随机误差。因此，无形之中就将样本异质性视为随机因素。

鉴于以上几点，在心理学研究中使用虚无假设检验等统计分析方法需慎之又慎。自20世纪70年代开始，一批心理统计学、数理统计学等领域的国外学者，对虚无假设检验展开了检讨并提出批评。科亨（J. Cohen）、尼克尔森（R. S. Nickerson）、汤普森（B. Thompson）和科克（R. E. Kirk）就是其中较为典型的代表。

心理统计学家科亨在20世纪90年代发表了一篇颇有影响力的文章，引起了心理学研究者的广泛关注。他指出，在经历了40多年的批评以后，虚无假设检验仍然坚持约定俗成的0.05的决策标准，其原因主要在于研究者普遍错误地将P值认作是虚无假设错误的概率，其余数是研究结果可以重复的概率。①

尼克尔森认为，虚无假设检验在心理学等社会科学中的使用如此广泛却遭致争议，其主要原因在于研究者混淆了绝对概率和条件概率，对虚无假设检验存在错误理解、错误使用，主要包括：② 拒绝虚无假设H_0，则意味着指导虚无假设的理论是错误的；小的P值是结果可重复的证据；统计显著性意味着理论上或实际应用中的显著性；某一实验设定的α值是解释实验结果时即将犯Ⅰ型错误的概率；未能拒绝虚无假设H_0等同于论证H_0为真。当然，尼克尔森认为，只要使用得当，虚无假设检验仍不失为解释心理学等实验数据的有效手段和方法。

汤普森则认为虚无假设检验存在不足，主要体现在：③（1）过分依赖样本；（2）一些比较（如P值总是和0.05这一显著性水平相比较）具有荒谬性；（3）一些无法避免的窘境，如拒绝虚无假设H_0并不等同于接受备择假设H_1、一分为二的决策标准、统计学

① 参见J. Cohen, “The Earth Is Round (p<0.05),” *American Psychologist*, vol. 49, 1994, pp. 997-1003.

② 参见R. S. Nickerson, “Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy,” *Psychological Methods*, vol. 5, no. 2, 2000, pp. 241-301.

③ 参见B. Thompson, “Statistical Significance Tests, Effect Size Reporting, and the Vain Pursuit of Pseudo-Objectivity,” *Theory & Psychology*, vol. 9, no. 2, 1999, pp. 192-196.

意义上的显著性和实际应用或临床显著性的区分等。科克亦认为虚无假设检验:① (1) 并没有告诉研究者所想要知道的结果。在科学推断中,研究者所想要知道的是在获得当前数据的前提下虚无假设 H_0 为真的概率,即 P (H_0/D) (D 表示当前数据 data);而虚无假设检验能够告诉研究者的是在总体中虚无假设 H_0 为真的前提下获得当前数据的概率 P (D/H_0);(2) 只能提供拒绝错误的虚无假设的统计功效,而拒绝虚无假设只是意味着没有找到明显地拒绝虚无假设的证据,并不意味着虚无假设就代表了世界真实的状态;(3) 把确定—不确定这一连续体变为二元的拒绝或接受的决定。在这种二元决策思想的指导下,可能仅因为实验设计上存在微小差异,而导致研究者会对同样的实验效应做出截然不同的结论。因而,二元决策标准使绝对 P 值这一连续的比率变量简化为二元称名变量,导致信息丧失,无法提供不确定程度的信息,直接影响对某一研究成果的正确、合理的解释,甚至会阻碍心理科学的进步。

总结心理统计学者对于虚无假设检验的批评与反思,至少在以下方面已达成一致。② 第一,虚无假设检验对样本容量的依赖性。同一检验,样本容量大的所提供的自由度也大。无论自变量的影响如何,相对于小样本,大样本更容易拒绝虚无假设,得到

① 参见 R. E. Kirk, "Practical Significance: A Concept Whose Time Has Come," *Educational and Psychological Measurement*, vol. 56, 1996, pp. 746-759.

② 对虚无假设检验的批评与反思参见: M. Wilkerson and M. R. Olson, "Misconceptions about Sample Size, Statistical Significance, and Treatment Effect," *The Journal of Psychology*, vol. 131, no. 6, 1997, pp. 627-631; T. Vacha-Haase and B. Thompson, "Further Comments on Statistical Significance Tests," *Measurement and Evaluation in Counseling and Development*, vol. 31, no. 1, 1998, pp. 63-67; B. Thompson, "Significance, Effect Size, Stepwise Methods, and Other Issues: Strong Arguments Move the Field," *The Journal of Experimental Education*, vol. 70, no. 1, 2001, pp. 80-93; B. Thompson, "'Statistical', 'Practical', and 'Clinical': How Many Kinds of Significance Do Counselors Need to Consider?" *Journal of Counseling & Development*, vol. 80, no. 1, 2002, pp. 64-71; F. Fidler, C. Geoff and B. Mark et al., "Statistical Reform in Medicine, Psychology and Ecology," *The Journal of Socio-Economics*, vol. 33, no. 5, 2004, pp. 615-630; H. C. Kraemer and D. J. Kupfer, "Size of Treatment Effects and Their Importance to Clinical Research and Practice," *Biological Psychiatry*, vol. 59, no. 11, 2006, pp. 990-996; E. J. Wagenmakers, "A Practical Solution to the Pervasive Problems of P Values," *Psychonomic Bulletin & Review*, vol. 14, no. 5, 2007, pp. 779-804; R. Hubbard and R. M. Lindsay, "Why P Values Are not a Useful Measure of Evidence in Statistical Significance Testing," *Theory & Psychology*, vol. 18, no. 1, 2008, pp. 69-88; M. Orlitzky, "How Can Significance Tests Be Deinstitutionalized?" *Organizational Research Methods*, vol. 15, no. 2, 2012, pp. 199-228; A. Fritz, T. Scherndl and A. Kühberger, "A Comprehensive Review of Reporting Practices in Psychological Journals: Are Effect Sizes Really Enough?" *Theory & Psychology*, vol. 23, no. 1, 2013, pp. 98-122.

统计显著性结论。由于世间万物或多或少地存在差异，所以“无差异”的虚无假设在现实世界中是不成立的。只要样本容量足够大，就会有足够的统计功效拒绝虚无假设，得到显著性结论。虚无假设检验也因此成为了“使研究者受累”的“体力劳动”。①

第二，显著性结论的不确定性。有学者认为有七个因素会影响虚无假设检验的结果，其中有两个尤为重要：效果量和样本容量。② 汤普森指出，在某一研究中计算出来的P值是许多研究特质的函数，但尤其受到样本容量和研究效果量的联合影响。③ 因此，检验统计量、效果量和样本容量的关系表达为下式：④

$$\text{检验统计量} = \text{效果量} \times \text{样本容量} \quad (1)$$

如公式（1）所示，统计显著性结果可能由大样本或者大效果量产生，无需同时满足。其他条件相等的情况下，实验设计或处理对因变量的效应越大，所产生的检验统计量越大。效应很小时，使用大样本，也极有可能获得统计显著性结论，反之亦然。因此，样本容量、效果量二者的角色在虚无假设检验中无法截然分开，实验效应和样本大小的交互关系难以理解，无法断定是否存在真实的效应。统计显著性结论也因此具有不确定性。

第三，统计显著性不等于结果的可重复性。⑤ 虚无假设检验所计算出来的P值，表达的是总体中虚无假设绝对为真的前提下，所获得当前样本数据的概率。统计推论的方向不是由样本推断总体，而是由总体推断样本。⑥ 这并不是研究者所期待的，

① **参见** R. S. Nickerson, “Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy,” pp. 241-301.

② **参见** A. L. Schneider and R. E. Darcy, “Policy Implications of Using Significance Tests in Evaluation Research,” *Evaluation Review*, vol. 8, no. 4, 1984, pp. 573-582.

③ **参见** B. Thompson, “Statistical Significance Tests, Effect Size Reporting, and the Vain Pursuit of Pseudo-Objectivity,” pp. 192-196.

④ **参见** S. Pedersen, “Effect Sizes and ‘What If’ Analyses as Supplements to Statistical Significance Tests,” *Journal of Early Intervention*, vol. 25, no. 4, 2003, pp. 310-319.

⑤ **参见** S. Scarr, “Rules of Evidence: A Larger Context for the Statistical Debate,” *Psychological Science*, vol. 8, no. 1, 1997, pp. 16-20; D. Sohn, “Statistical Significance and Replicability: Why the Former Does Not Presage the Latter,” *Theory & Psychology*, vol. 8, no. 3, 1998, pp. 291-311; A. G. Greenwald, R. Gonzalez, R. J. Harris and D. Guthrie, “Effect Sizes and P Values: What Should Be Reported and What Should Be Replicated?” *Psychophysiology*, vol. 33, no. 2, 1996, pp. 175-183.

⑥ B. Thompson and P. Snyder, “Statistical Significance Testing Practices in *The Journal of Experimental Education*,” *The Journal of Experimental Education*, vol. 66, no. 1, 1997, pp. 75-83.

研究者希冀虚无假设检验可以评价总体、推断总体。这就可以推测以后的研究者从同一总体中抽取样本所得的结果。唯有产生对总体的推断才可提供关于研究结果是否可以重复的信息。由于虚无假设检验的显著性结论不能检验总体，因而不能提供结果的可重复性。然而，很多研究者并没有真正理解虚无假设检验的这个逻辑。

由于显著性结论成为大部分心理学杂志接受或者发表论文的标准之一，并且研究者惯性地认为虚无假设检验探讨的是样本之间的差异，提供的是关于研究结果的重要性或可重复性，以致在虚无假设检验的应用上，长期完全依赖 P 值来做出判断。对 P 值的过分强调，实际上会使研究者偏离研究目标——决定数据是否支持所提出的科学假设，并且确定研究成果在实际应用中的重要性、有用性。这种将统计意义上的显著性等同于实际应用或临床应用上的显著性，是心理学研究者中存在的一个普遍错误认识。研究者认为小的 P 值比大的 P 值具有更强的实际效应，所以常使用不同个数的星号或者“显著”、“非常显著”、“极其显著”的字词来标识。P 值是用来检验样本统计量的概率，是虚无假设为真时结果发生的概率。P 值实质上是一随机变量，混淆了样本容量和效果量的效应，只有在样本容量一定的情况下，才可以得出结论：P 值越小，效应越大。总之，虚无假设检验直接估计的是样本的可能性而非总体的可能性，没有估计结果可重复性的概率，而可重复性对于心理学知识的积累以及研究成果的科学性、可推广性尤为重要。

三、效果量及其作用

对样本容量的依赖以及无法估计研究结论的可重复性、可推广性，是研究者对虚无假设检验提出的最为主要和尖锐的批评。虚无假设检验是运用反证法的思想，通过拒绝虚无假设 H_0 来验证备择假设 H_1 的真实性。一般认为在研究总体中虚无假设 H_0 要么为真，要么不为真。从逻辑上来讲，虚无假设检验，特别是在心理学研究中不能运用这种“全或无”的准则。虚无假设 H_0 不为真，只是在某种程度上不为真，研究结果是在某种程度上偏离虚无假设 H_0。效果量就是指虚无假设 H_0 不为真的程度，实际上就是偏离虚无假设 H_0 程度的一种指数。它反映自变量与因变量之间关系的强弱，是研究对象之间差异的大小、实验效应大小的真实程度、研究结果重要性指标。提出虚无假设检验新模式，将效果量作为虚无假设检验的重要补充，其作用和意义主要体现在：

第一，假设检验本身只能提供差异有无这种不确定信息，无法提供确定的差异大小，而心理学研究者不应仅为实验处理的实施提供统计学依据，更需关注某一实验处理相对于另一处理的优势有多大，从而为该人群的相关心理学问题提出解决办法和改进措施。虚无假设检验过分依赖样本，无法告诉研究者或读者实验效应的重要性、变量间关系的强弱、实验结果的实践意义等，显著性结论也并不等同于结论

可重复性的测度。这些对于心理学知识的积累尤为关键，研究者也越来越意识到提供有关研究结果跨样本稳定性证据的重要性。由于基于显著性水平的估计总是高估了效应的真实大小，因而很多研究者否定将 P 值作为效应大小的直接测度，并已达成削减虚无假设检验的重要性及其在心理学量化研究中的地位，寻求补充可比较的、可测度结果重要性和可重复性的指标的共识，因此，一些学者及研究机构都相继提出使用效果量作为假设检验的重要补充，报告 P 值时应该同时报告效果量。① 效果量被视为“科学研究的最终目标”，② 并且“在教育学和心理学研究中，没有什么比使用虚无假设检验时对结果的效果量进行估计更为重要”。③

第二，效果量表示总体中变量之间的关系，是对因变量和自变量关系强弱的测度，对样本结论远离虚无假设的期望程度的量化。效果量不受样本的影响，是一种真实的存在。效果量在研究结果报告中甚为关键，可使读者完全理解研究的重要性，为读者提供评估观察效应或关系强度的足够信息。心理学研究在结论部分应该报告某一形式的效应指标或关系强度是有必要的，没有报告效果量应该作为研究设计或者研究报告的过失之一。作为一个成熟的研究领域，心理学研究结果的效果量应该比统计显著性结论显得更为重要。在实际应用中，效果量是决定统计功效、所需样本量的一个重要因素。不同的实验处理效果量不同。当效果量提高时，偏离虚无假设 H_0 的程度越大，研究结果也就越接近备择假设 H_1 为真。效果量和统计功效之间的关系是：在样本容量和 α 水平等其他因素都一定的情况下，效果量增加或减少，统计功效值也随之增加或减少，反之亦然。

第三，研究者应该测度的是效应大小，而不是统计意义上的显著性。效果量是研究应用性的指标，而 P 值仅是研究统计学意义上的显著性指标。因为从心理学角度来讲，统计意义上的差异是否真正有差异，绝不仅是由 P 值决定的，而取决于诸多因素。仅仅报告 P 值，目前暂无有效的统计工具可以决定研究结果对于读者日后

① American Psychological Association, *Publication Manual of the American Psychological Association* (*4th ed.*), Washington, DC: Author, 1994; American Psychological Association, *Publication Manual of the American Psychological Association* (*5th ed.*), Washington, DC: Author, 2001; American Psychological Association, *Publication Manual of the American Psychological Association* (*6th ed.*), Washington, DC: Author, 2009; L. Wilkinson and the Task Force on Statistical Inference, “Statistical Methods in Psychology Journals,” *American Psychologist*, vol. 54, no. 8, 1999, pp. 594-604.

② **参见** L. Wilkinson and the Task Force on Statistical Inference, “Statistical Methods in Psychology Journals,” pp. 594-604.

③ **参见** B. Thompson, “Research News and Comment: AERA Editorial Policies Regarding Statistical Significance Testing: Three Suggested Reforms,” *Educational Researcher*, vol. 23, no. 2, 1996, pp. 26-30.

研究是否有用或者重要，无法判断是存在真正意义上差异还是仅仅意味着统计学差异。P 值过度依赖样本，不能估计研究结果的可重复性，只是对效应的一种混淆测度，相对小的 P 值并不能充分说明研究中自变量和因变量之间有很强的关系，对于心理学“理论发展毫无用处”，甚至“阻碍了心理学的发展”。① 同时，P 值是个随机变量，随样本不同而不同。比较两个不同实验或者同一实验中基于不同的变量所计算出来的 P 值，进而得出其一更为显著的结论是不妥的。

效果量因此显得重要，尤其是对于已得到显著性结论的研究而言。相对于 P 值，效果量可以提供研究结论更为确定的应用价值，若只报告 P 值而不报告效果量，会失去关键信息，无法获取研究结论的实际应用性价值的信息。虚无假设检验不能进行跨样本、跨研究的比较，其原因在于显著性结论中的 P 值只是个随机变量，随研究样本的变化而变化，不具有无标度（scale-free）的特性。而效果量指标具有无标度的特性，其大小通过计算标准化差异来估计，不管样本大小和变量的初始测度如何，它都可用来比较同一研究中不同变量的处理效应，也可跨研究地比较相同变量或者不同变量的处理效应，而效果量提供了效应大小的指标或者提供过去和现在研究的比较标准，跨研究的效果量比较可以提供研究结果的可重复性，可以确定效应是否稳定存在而非偶然发生，以帮助研究者确定后续研究的重要变量、特征。汤普森在解释报告效果量的原因时指出，报告效果量可以：（1）促进更高质量的元分析研究或者回顾；（2）可以促使后来的研究者设计更为明确的参数和结果期望；（3）有助于评估研究结果是否适宜于其他不同的研究背景，即研究结果和其他研究的相似之处以及研究中对于这种相似性或者差异性有所贡献的特质。②

对效果量的正确、合理的解释亦成为学者关注的问题。科亨提供了他所定义的“大”的、“中”的、“小”的效果量的标尺。③ 他期望这些标尺主要用于研究对象是没有探索性研究的领域，这些标尺只是一个广义上的指南。并且他强调：若人们使用某一严格的效果量标准就无异于将统计显著性水平刻板地设定为 0.05，二者是一样的愚蠢。至少对于已经进行过相关研究的领域而言，试图将效果量的相关区域用“大”的、“中”的、“小”的或者类似的描述性形容词来表达是不明智的。效果量的确定应该基于研究背景，研究的效果量在以后的类似研究中可以得到重复，在跨样

① 参见 R. S. Nickerson，“Null Hypothesis Significance Testing：A Review of an Old and Continuing Controversy，” pp. 241-301.

② 参见 B. Thompson，“In Praise of Brilliance：Where That Praise Really Belongs，” *American Psychologist*，vol. 53，no. 7，1998，pp. 799-800.

③ 参见 F. L. Schmidt，“Statistical Significance Testing and Cumulative Knowledge in Psychology：Implications for Training of Researchers，” *Psychological Methods*，vol. 1，no. 2，1996，pp. 115-129.

本研究中相对稳定。当研究者观察重要的结果变量时，即使是非常小的效果量也可以是显著的。因此，很多学者建议应和先前相关研究文献中的效果量进行直接、明了的比较，“应该在先前研究报告效果量的背景下报告和解释效果量，这对于一个好的研究而言是重要的”。①

四、效果量的常用指标

虽然国内心理学学者对于效果量这一概念较为陌生，国外心理学学者也是近年来对效果量有所关注，但效果量并非新生事物。早在70多年以前，费希尔就提出方差分析的研究结果应该包括相关比率或 η^2 以表示自变量和因变量之间关系的强弱。自此，不少研究者提出不同的效果量指标。科克在总结他人研究的基础上，于1996年就提出了41种不同的效果量指标。② 所以，在报告效果量时应清楚地指出所使用的效果量的类别和名称，以帮助读者正确估计其强度，并用以指导后续研究。

简言之，效果量的大小是来自总体1的随机样本的实验处理的强度大于来自总体2的随机样本的概率：$P(X_1 > X_2)$。依据获得效果量的不同方法可以将效果量分成不同的类别。

克莱恩（R. Kline）认为效果量一般可以分为广义的两族：（1）标准化平均差异或群组差异指标。当统计分析使用平均数来比较潜在的群组差异时使用该族效果量。（2）关系强度指标，即考虑方差的关系或者方差解释率的效果量指标。该族效果量是基于因变量相关联的某一特殊变异和总体变异的比率，常适用于使用广义线性模型（GLM）的跨研究设计，意味着自变量所能解释的因变量的变异，产生的变量间的关系强度或方差解释比率。如：表示双变量Pearson相关系数平方的 r^2，表示三个或更多变量的多元相关系数平方的 R^2。③ 科克呈现了三类效果量：标准化平均数差异；考虑方差的效果量；混合效果量。④ 汤普森依据是否有偏将效果量进一步划分为有偏（未调校）效果量和无偏（调校）效果量两类。⑤ 汤普森在2006年又将效果量分成三种：标准化平均数差异效果量；未调校的考虑方差的效果量；调校的

① 参见 L. Wilkinson and the Task Force on Statistical Inference，“Statistical Methods in Psychology Journals，” pp. 594-604.

②④ 参见 R. E. Kirk，“Practical Significance：A Concept Whose Time Has Come，” pp. 746-759.

③ 参见 R. Kline，*Beyond Significance Testing：Reforming Data Analysis Methods in Behavioral Research*，Washington，DC：American Psychological Association，2004.

⑤ 参见 B. Thompson，“The ‘Significance’ Crisis in Psychology and Education，” *Journal of Socio-Economics*，vol. 33，no. 5，2004，pp. 607-613.

考虑方差的效果量。① 目前，学者普遍接受汤普森的划分标准。

（一）标准化平均数差异效果量指标（Standarized Difference Effect Size）

采用事前事后设计、对照组实验设计，需要使用平均数比较潜在的群组差异时，可用该类指标。理论上，实验效应可以直接用某一统计量的群组差异来表示。但心理学不同于医学等学科，其结果变量在本质上并无有意义的量尺，即使是对于多元事后测量也可能会得到不同的标准差。所以不能直接使用组间平均数来计算、比较某一研究或不同研究中的效果量。所以，应采用标准化这一统计策略，用组间平均数差异除以特定尺度的标准差，使之在量尺上获得自由，从而可以比较不同研究中的效果量。最为常用的是赫奇斯（Hedges）提出的 g 指标和科亨提出的 d 指标。

$$g=\frac{M_{experiment}-M_{control}}{S_{pooled}} \tag{2}$$

其中 $S_{pooled}=\sqrt{\frac{(n_1-1)\ s_1^2+(n_2-1)\ s_2^2}{n_1+n_2-2}}$

$$d=\frac{M_1-M_2}{\sigma_{pooled}} \tag{3}$$

其中 $\sigma_{pooled}=\sqrt{\frac{(n_1-1)\ s_1^2+(n_2-1)\ s_2^2}{n_1+n_2}}$

标准化指标可比较两个或多个测度的个人分数，尽管这些分数是不同量尺的测度。这类指标的提出沿用标准分数的思想，可以直观地理解效果量是如何比较不同研究的结果，即使这些研究采用的是不同因变量，也可很好地利用已获得的统计量。这类指标使用两组结果变量的标准差的加权平均数来估计标准差，其主要考虑是实验处理对于事后分数的离散程度不会产生影响，使用联合标准差可以加大估计的样本容量，提供更为精确的效果量估计。这类指标完全依赖于实验情境，是较好的标准化差异效果量指标。

（二）考虑方差的效果量（Variance-Accounted-For Effect Size）

使用 GLM 的跨实验设计，用以评量变量间的关系时应该使用考虑方差的效果量指标。该类指标是基于与因变量相关联的某一特殊变异和总变异的比率，意味着自变量所能解释因变量的变异，反应变量间的关系强度或者方差解释率。可进一步将该类效果量指标分为有偏（未调校）效果量和无偏（调校）效果量。

有偏效果量和抽样过程、抽样误差相关联，一般会高于由原始样本直接获取的

① 参见 B. Thompson, *Foundations of Behavioral Statistics: An Insight-Based Approach*, New York: Guilford, 2006; "Role of Effect Sizes in Contemporary Research in Counseling," *Counseling and Values*, vol. 50, no. 3, 2006, pp. 176-186.

效果量。无偏效果量可更好地反映总体真实的效果量及以后研究样本中的效果量，其原因在于从统计学意义上来讲该类指标是正确的：计算关系强度的指标是正偏的，无偏效果量调校研究中的标准误对于将来研究样本效果量的估计会比总体效果量的估计有所收缩。由于未调校的效果量没有考虑样本容量或因素的水平数，所以调校的效果量倾向于小于（至少不大于）未调校的效果量。研究中有很多因素会影响效果量的有偏性，主要包括：（1）结果或测量的信度。一般而言，测验结果的信度越高，效果量的偏性越小；（2）样本容量，大样本产生的效果量偏性较小，样本量足够大时，有偏和无偏的效果量会相等。大样本倾向于产生更为稳定的效果量；（3）自变量的数目，自变量越少，且样本容量越大，效果量的偏性越小；（4）研究样本的同质性，样本越同质，效果量的偏性越小；（5）研究设计类型，实证性研究可以产生偏性较小的效果量，设计过程、实施过程缜密、严格的研究会减少效果量有偏的大小，效果量需要调校的可能性越小，若研究设计不够理想，应当报告调校的效果量。①

（1）未调校的考虑方差的效果量常用指标有以下几种：

1）决定系数 r^2 或 R^2，可用以下公式计算：

$$r^2 \text{ 或 } R^2 = \frac{SS_E}{SS_T} \tag{4}$$

其中，SS_E 表示处理平方和，SS_T 表示总平方和。r^2 或 R^2 描述了结果变量分数的变异中有多少可以由预测变量分数的变异来解释或者预测。由于所有统计量都相互关联，都可作为广义线性模型的一部分。所以，所有研究都可计算决定系数。

2）η^2，其计算公式如下：

$$\eta^2 = \frac{SS_E}{SS_T} \tag{5}$$

这一效果量适用于单变量方差分析，其中，SS_E 表示处理平方和，SS_T 表示总平方和。η^2 描述了结果分数变异中有多少可以由被试所属的群组解释或预测。

（2）调校的效果量倾于小于（至少不会大于）未调校的效果量，其原因主要在于前者考虑了抽样误差。个体特质决定了样本总不能很好地代表总体。通常统计分析过程无法区分样本所代表的总体中存在的变异和总体中不存在的变异，样本效果量常常正偏态地高估了总体中真实的效果量。所以必须考虑抽样误差带来的后果：样本大小对于抽样误差的变异有重要影响，小样本倾于产生更多的抽样误差，而大样本会高估效果量；测量变量的数目对于抽样误差有影响，研究中测量的变量越多，抽样误差越大；总体效果量对抽样误差有影响。研究者总是不可能知道总体的效果量，只能通过样本效果量进行估计、调校，所以，很难看到这种动态的变化、影响。

① 参见 B. Thompson, *Foundations of Behavioral Statistics: An Insight-Based Approach*.

调校的考虑方差的效果量常用指标有以下几种：

1）对于 r^2 或 R^2，*Ezekiel* 使用调校估计 R^{2*}。R^{2*}可以用以下公式计算：

$$R^{2*}=1-\left(\frac{n-1}{n-v-1}\right)(1-R^2) \tag{6}$$

其中，n 表示样本容量，v 表示预测变量数。这一公式也可以等同地表示为：

$$R^{2*}=R^2-(1-R^2)\left(\frac{v}{n-v-1}\right) \tag{7}$$

2）对于方差分析，调校的效果量为 ω^2，这是 Hays 提出的。其公式为：

$$\omega^2=\frac{[SS_b-(k-1)MS_w]}{(SS_T+MS_w)} \tag{8}$$

其中，k 表示方差分析中的水平数，SS_b、SS_T 是效应平方和和总平方和，MS_w 表示误差均方。

总之，以上类型涵括了现有的参数化的效果量指标。标准化平均数差异效果量易于计算；考虑方差的效果量是基于广义线性模型，由组间变异和总变异的比率计算得到的，解释为变量变异中由自变量所产生变异的比率。当样本量在很大时，各种效果量间的波动不大。①

五、结　　论

近年来，ERP、fMRI 等技术的引入，为心理学的发展提供了越来越强大的技术支持；而与此相伴的以结构方程模型、多层次线性模型、社会网络分析等为代表的统计分析方法不断向高阶发展，已经使心理学的研究日趋情境化、真实化。

但是，正如科亨所言，统计分析方法不是越复杂越好，而是越简单越好。② 虚无假设检验方法在科学主义心理学盛行的这百余年历史中，尽管不是最复杂、最高级的，但其在心理学研究中的主体地位不仅没有动摇，反而日渐重要。因此，适时地予以审视，不仅很有必要，也是确保心理学研究结论可靠性的重要方法。

目前虚无假设检验屡遭质疑不在于虚无假设检验本身，而在于不适当地应用和误用。从历史源头出发，厘清现有虚无假设检验理论体系的历史，辨析在发展过程中不同流派的学者对于这一问题的讨论，对于培养研究者关于虚无假设检验并非万能的统计方法意识，促使其在研究中合理地使用虚无假设检验是很有必要的。我们也希望能为社会学、管理学、经济学等社会科学正确应用统计方法提供借鉴意义。

① 参见 B. Thompson，*Foundations of Behavioral Statistics：An Insight-Based Approach*.

② 参见 J. Cohen，“Things I Have Learned (So Far)，” *American Psychologist*，vol. 45，1990，pp. 1304-1312.

在虚无假设检验的诸多应用问题中，最大的问题在于这种方法只能提供是否存在差异的信息，无法告知研究结论的可重复性。效果量可弥补这一缺陷。效果量是影响统计功效的重要因素，是“研究假设为真的程度，是研究结果在研究样本中具有实际显著性的程度，也就是研究结果的重要性程度”。① 不过分依赖样本、具有无标度特性是效果量指标的两大特点。使用将虚无假设检验和效果量相结合的虚无假设检验新模式，同时报告虚无假设检验的显著性结论和效果量，可以让读者同时了解研究的统计显著性和实际显著性。诚然，实际显著性比统计显著性在实际工作中来得重要。这样的报告形式还可以避免显著性结论对样本的依赖性、不能提供研究结果可重复性指标、不能进行跨研究或跨样本的比较等不足。因此，心理学研究结果同时报告、解释显著性结论和效果量，无疑是对虚无假设检验理论与技术的完善，也是心理学科不断科学发展的助推手。

近年来，国内学者开始关注效果量的问题。② 心理学权威刊物也相继提出研究结果要报告效果量，但实际报告者相对较少。③ 不少研究者只是知其然而不知其所以然，仅仅停留在知晓要报告效果量的层面。因此，探讨效果量的本质、内涵及其作用对于心理学研究者正确理解效果量有促进作用；而对于效果量常用指标的探讨也可以帮助、指导研究者在不同的情形下选用合适的指标。

效果量目前给我们展示的也许只是它合理的一面，但随着研究的深入，其局限和不合理也将会逐渐暴露出来。无论是效果量还是虚无假设检验，都是科学主义心理学在心理学研究中的一个侧面。现实的社会是丰富多彩的，科学主义心理学只是为我们提供了一个解释世界的理路，并不能解释世界的全部。我们所能做的就是要用真正符合社会现实的思维和逻辑，来驾驭科学主义心理学，使其朝着科学、严谨的方向不断发展。

〔责任编辑：莫　斌　责任编审：柯锦华〕

① 参见 M. Hojat and G. Xu，“A Visitor’s Guide to Effect Sizes：Statistical Significance versus Practical (Clinical) Importance of Research Findings，” *Advances in Health Sciences Education*，vol. 9，no. 3，2004，pp. 241-249.

② 参见温忠麟：《屡遭误用和错批的心理统计》，《华南师范大学学报》2010 年第 1 期；胡竹菁、戴海琦：《方差分析的统计检验力和效果大小的常用方法比较》，《心理学探新》2011 年第 3 期；郑昊敏、温忠麟、吴艳：《心理学常用效应量的选用与分析》，《心理科学进展》2011 年第 12 期，等等。另外，笔者于 2007 年 11 月、2008 年 11 月、2012 年 11 月先后在第十一届全国心理学学术会议、全国教育与心理统计与测量学术年会暨第八届海峡两岸心理与教育测验学术研讨会、第十五届全国心理学学术会议上都提出中国心理学界应将是否报告效果量作为心理学研究论文是否发表的重要准则之一。

③ 以《心理学报》2013 年第 10 期发表的学术论文为例，共有 8 篇文章使用了虚无假设检验，仅有 1 篇报告了效果量。

political identity contribute more to the formation of modern gender attitudes; for men, sharing the housework equally with their wives contributes more to the formation of modern gender attitudes.

(5) Core Concepts and Their Identification in Criminal Procedure Law

Wang Jiancheng • 130 •

The revision and effective implementation of the Criminal Procedure Law is inseparable from the guidance provided by several important concepts. At the heart of the law are the concepts of legal procedure, protection of human rights, evidentiary adjudication, due process, and efficacy of legal action. Among these concepts, legal procedure reflects the distinction between criminal procedure legislation and the administration of justice; protection of human rights embodies the protection of the rights and interests of the accused and all members of society; evidentiary adjudication determines the process of judicial decision and the grounds of its legitimacy; due process expresses the essentially legal procedural character of criminal procedure law; and the efficacy of legal action reveals the tension between procedural justification and judicial resources as well as a means of resolving this issue. Establishing and adhering to such concepts is a long-term process, but it is one that constitutes an indispensable stage in becoming a country ruled by law.

(6) The Lost Boundary: A Study of the Null Hypothesis Testing Method in Psychology

Jiao Can and Zhang Minqiang • 148 •

After being introduced into psychology from statistics, the null hypothesis testing method was widely used and became the most common and the most influential psychological technique since the emergence of scientistic psychology. However, psychological circles at home and abroad have found many problems and errors in its use, as the method has been used outside its proper boundaries. An examination of its history and development shows that it is most often criticized for dependence on sample capacity, uncertainty over the degree of significance of conclusions, and non-replicable findings. Effect size is the main factor that influences statistical power. Introducing the concept of effect size, reporting and explaining effect size, and making effect size an integral part of research findings will improve the theory and technique of null hypothesis testing.

(7) Book Circulation and the Poetics of East Asia—The Example of *Chongbirok*

Zhang Bowei • 164 •