Hedges Q检验的性能评估与标准制定

Performance Evaluation and Standard Setting for Hedges Q Test

纪凌开·著

中国社会科学出版社

Hedges Q检验的性能评估与标准制定

Performance Evaluation and Standard Setting for Hedges Q Test

纪凌开·著

中国社会科学出版社

图书在版编目(CIP)数据

Hedges Q 检验的性能评估与标准制定/纪凌开著. —北京:
中国社会科学出版社, 2017.3
ISBN 978-7-5161-9264-1

Ⅰ.①H… Ⅱ.①纪… Ⅲ.①心理学分析 Ⅳ.①B84

中国版本图书馆 CIP 数据核字(2016)第 266502 号

出 版 人 赵剑英
责任编辑 赵 丽
责任校对 闫 萃
责任印制 王 超

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 http://www.csspw.cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印刷装订 北京君升印刷有限公司
版 次 2017 年 3 月第 1 版
印 次 2017 年 3 月第 1 次印刷

开 本 710×1000 1/16
印 张 18.75
字 数 288 千字
定 价 79.00 元

凡购买中国社会科学出版社图书,如有质量问题请与本社营销中心联系调换
电话:010-84083683

目　录

引　言

冯特1879年在德国莱比锡大学建立世界上第一个心理学实验室，标志着心理学的研究从传统的哲学思辨道路转入实证研究道路的开始。自此，心理学各个具体领域的实证研究犹如雨后春笋，欣欣向荣。发展至今，可以说在心理学的几乎每个具体研究领域都积累了大量的实证研究成果。然而，不同或相同研究者就某个相同或相似研究主题所得出的研究结果常常并不一致，有些观察到的研究效应强，有些观察到的研究效应弱，有些没有或甚至出现了相反的研究效应。这种现象不独出现在心理学研究领域，在教育学、社会学、经济学、农学、生物学和医学等学科领域，也存在类似情况。比如医学领域，每年有超过两百万篇论文被发表，当不同的研究者或同一研究者试图对同一现象进行多次研究时，他们有时会发现并困惑于不同的研究有不同的研究结果（Rosenthal & DiMatteo，2001）。

这种现象的存在直接导致的可能后果有：其一，实践上，众多彼此不一致甚至矛盾的研究结果非常不利于研究结果使用者进行决策。比如政策制定，众所周知，科学的政策制定是基于事实以及对事物真理性认识的基础上进行的，而这又依赖于实证性的科学研究。可是多年来，政策制定者一直受困于这样一个事实，即对同一问题的不同研究报告常常出现结果相互矛盾或对立。举个简单的例子，比如有的研究表明注射白蛋白有利于提高高危病人的抢救成功率，而有的研究认为不利于高危病人的抢救，甚至会导致死亡，那么卫生部有关监管部门究竟是应该同意还是应该取消在高危病人抢救时将注射白蛋白作为辅助措施呢？其二，在理论发展上，对同样问题进行大量的研究是非常必要的，但是“对同一问题的不同研究，研究结果相互矛盾的情况很常见（Hunter &

Schmidt，1996）”，由此便导致后继研究者依据新的研究结果推翻先前研究结果并对前人进行“无情地开火”，这是科学研究领域中的公开现实。例如，自华莱士（Wallas，1926）提出问题解决的理论模型以来，研究者对于孕育期（incubation period）是否真的有利于问题解决进行了大量的研究。Olton 和 Johnson（1976）在研究中没有发现孕育期对问题解决具有任何作用，而 Smith 和 Blankenship（1989）的研究结果发现孕育期对于问题解决具有强效应。显然，前者否定孕育期有助于问题解决的假设，而后者却持肯定的立场。那么，究竟是前者接近事实的真相，还是后者接近事实的真相，抑或两者犹如“盲人摸象”？变量间关系的理解是理论赖以建构与发展的基石，而前面所举的例子所涉及的核心变量只有孕育期与问题解决，情况尚属简单；如果研究所涉及的变量有多个（两个以上），则情况就更加复杂。因此，实践上的挫折与理论发展上的困惑都促使人们进一步深入地思考这样一个问题，即单个实证研究的结果是否可靠？并进而思考单个实证研究在科学研究中的地位以及相同或相似课题中多个独立进行的实证研究间的关系等问题。是什么原因导致相同或相似的课题上多个独立进行的研究结果出现变异？随机误差因素与系统因素在这种变异中究竟起何作用？这些问题的回答比单个研究结果更为重要。至此，科学研究需要一种能够对迄今为止所积累起来的关于同一或相似研究课题中的多个实证研究进行系统分析、合成且能合理解释为什么研究效应在不同具体研究间会产生变异的这样一种方法。

纵观科学研究的发展历史，对同一研究主题下多个独立进行的原始研究结果进行评价和合成以得出综合性结论的方法整体上有两类：一类是质性合成法，另一类是基于统计原理的量化合成法。文献质性合成指的是传统的叙述性研究合成法（narrative methods of research integration），这种方法的主要作用体现在识别与描述某个领域的最新进展并对其发展状况与发展趋势进行讨论，或者引用实证性证据支持、丰富与重新评估某一具体理论或尝试性支持某一新提出的理论，或者将不同“轨迹”上的研究系统地组织成一个有机的知识体系。但是，涉及对同一领域的相同或相似问题进行合成时，这种合成方法的文献合成质量饱受批评。最令人恼火的地方是它似乎非常容易对特定综述者的偏差做出

反应（Glass，1976）。简单地说，就是对相同或相似课题下的研究文献集进行综述时，不同的综述者很可能得出不同的文献合成研究结果。在对文献进行叙述性综述时，大量有价值的信息被综述者忽视是很常见的现象，并且综述者可能还会对所掌握的信息进行不恰当的权重处理（Copper & Rosenthal，1980）。综述者还可能会无意识或有意识地对文献进行倾向性取舍，以利于支持自己的理论观点或自己对文献的理解。而且，有时这种取舍主要还是依据单个研究的显著性检验结果进行的。众所周知，一个不争的事实是显著性水平与研究的样本容量有关。一个实际上较弱的效应可以通过简单地增大被试样本容量而使检验结果变得更显著，一个较强的效应可能通过减少被试样本容量而变得检验不显著。因此，传统对文献的质性合成方法基本上不符合科学研究有效性与可靠性这两大质量评估标准。对于传统文献的质性合成法，Field（2005）极不客气地斥之为“东拉西扯式综述”（discursive reviews）。在此背景下，文献量化合成方法登上历史舞台是科学研究方法论适应现实需要的必然发展。

研究文献的量化分析与合成的实质是借助于统计方法对某一相同或相似研究主题下的文献集进行系统的综述与评价，以获得综合性的结论。在这个意义上，元分析（meta-analysis）实际上成了文献量化合成的代名词（DerSimonian & Laird，1986；Johnson Mullen & Salas，1995；Viechtbauer，2007）。“元分析”一词最初由 Glass（1976）提出，意指“一种关于分析的分析（the analysis of analysis），是为了获得综合性的研究结论而对众多独立的单个研究结果所进行的统计分析”。Glass 首先提出“元分析”一词并不意味着 1976 年是元分析实践的起点，现在公认的第一个元分析工作由 Pearson（1904）在研究天花预防接种与存活之间的关系时所完成的。当时，他对所搜集到的研究文献中的相关系数进行了平均化处理，求取加权与未加权的平均相关系数（Rosenthal & DiMatteo，2001）。在 20 世纪前四分之三的时间里，Pearson 的这种统计处理方法一直被各个科学领域的研究者所使用。同时在这个过程中也逐渐产生了多种对众多独立研究结果进行统计处理的其他方法。比如，基于单个研究统计检验显著性水平的概率合成法（Fisher，1932；Pearson，1933）、Cochran（1937）与 Yates 和 Cochran（1938）首次对平均效应

(the mean effect) 和处理效应的研究间变异的关注以及 Wilkinson (1951) 对单个研究结果进行二分处理的策略 (显著与不显著，由此应用二项分布作为统计推断的基础) 等等。其中，值得一提的是，Wilkinson 的论文是在心理学期刊 *Psychological Bulletin* 上正式发表的，这标志着心理学界对元分析技术的最早关注。在 20 世纪 50 年代前，基于统计检验显著性水平的概率合成被认为是一种对众多原始研究结果进行合成的合理方法，而 Cochran 与 Yates 等人更具现代元分析色彩的文献统计合成思想却没有获得正式发展。为什么后者这种思想的进一步发展会受到阻碍？Bangert-Drowns (1986) 认为当时有两大障碍没有得到克服：一是如何把不同研究结果中因变量的测量转换到相同的量尺上去；二是如何把研究特征有机地融入到元分析中去。在对第一个问题的解决中，Cohen (1962) 做出了杰出的贡献。Cohen 在有关检验力的研究中提出了“效应量” (effect size) 这个概念，并给出了总体效应量 (population effect size) 的计算公式。实际上，Cohen 的这一工作使得人们可以将不同研究结果中因变量的测量转换到相同的量尺上去；而对于第二个问题，Light 和 Smith (1971) 提出类分析 (cluster analysis) 技术 (可以看作一种特殊的元分析技术) 为解决这个问题提供了有价值的启示。自此以后，元分析技术的发展非常迅速，一方面研究内容不断拓宽，另一方面元分析所基于的统计基础日益丰富与深入。在此过程中，元分析技术逐渐发展出多种各具特色且相对成熟的分支。

现在，元分析由于较之于传统的、叙述性的文献合成方法或综述方法更加严格与精确 (Johnson, Mullen, & Salas, 1995) 且能提供令人信服与可靠的证据 (Higgins, Thompson, Deeks, & Altman, 2003) 而被普遍视作一种精确而客观的文献合成方法 (Hardy & Thompson, 1996)，这一点可以从当前元分析式文献综述研究论文在心理学主要期刊上的发表态势得到印证 (Field, 2003)。至今，在心理学领域内，几乎所有有影响力的期刊都接纳元分析的研究论文，并且数量迅速增多，其中 *Psychological Bulletin* 是主要的代表之一。元分析也被广泛应用于教育学、心理学、社会学、医学、生态学、生物学、管理学等领域 (Arnqvist, & Wooster, 1995; DerSimonian, & Laird, 1986; Mari′n-Marti′nez & Sa′nchez-Meca, 2009; Knoben & Oerlemans, 2006; Kisamore & Brannick, 2008)。

实际上，合理的元分析结果已经被视作制定政策、决策和促进理论发展的最可靠、最高等级（具有相对终极性）的研究证据。比如，在医学领域，关于前面所举的注射白蛋白是否有利于高危病人抢救的例子，元分析的结果是其整体上存在明显的致命负面效应（Cochrane Injuries Group Albumin Reviewers，1998），这个研究结果直接导致注射白蛋白作为高危病人抢救的重要辅助措施这一治疗方法的取消。既然元分析较之于文献的质性合成分析具有巨大的优越性，这是否就意味着只要把元分析技术简单套用到相同或相似研究主题下的众多独立研究文献上去就一定能够得到合理的综合性结论？实际上，这是不可能的。可以说，一个元分析的成功与否主要取决于两个方面：一是研究文献的搜集、纳入是否全面以及信息的提取是否科学；二是对从众多原始研究上所获得的信息如何统计分析。第一个方面不是本书研究的兴趣所在，本书的研究兴趣将集中在第二个方面。在元分析实践中，对相同的数据采用不同的元分析方法，结果并不见得一定是一致的（Sanchez-Meca & Marin-Martinez，1998）。即使研究文献的搜集、纳入与信息的提取均很客观与全面，但如果元分析时所采用的统计技术不合适、性能不够稳健（robustness）或者元分析技术没有得到正确应用，则元分析结论的有效性将会不理想甚至可能会出现很大的偏差，从而可能导致严重的后果。

由于元分析技术内容极为丰富，故对其所有技术的性能都进行系统的比较与评估不是这次研究所能完成的事情。因此，本次研究将研究兴趣界定在元分析某一重要组成技术的性能评估范围之内。在元分析的所有工作中，探测效应量（effect size）是否异质（heterogeneity）是其关键性的组成部分（Hardy & Thompson，1998），也是元分析三大主要任务中第一个必须解决的问题（Copper & Hedges，1994；Huedo-Medina，Sa′nchez-Meca，& Botella，2006）。如果元分析中原始研究（primary study）间的效应量同质（homogeneity），就应该用固定效应模型对效应量进行合成，获得平均效应量作为总体效应量的估计值。此时，这个平均效应量富有意义，且较之于采用随机效应模型所得到的结果更加精确，它是元分析中所有原始研究所关注的研究变量间作用（总体效应量）的高度概括；相反，如果原始研究的效应量是异质（heterogeneity）的，元分析者则应该采用随机效应模型或采用调节效应分析的统计策略

(Field, 2001; Hedges & Vevea, 1998; Overton, 1998; Raudenbush, 1994)。此时，若依然采用固定效应模型进行效应量合成则会产生不正确的分析结果，并产生误导作用。

正因为效应量的同质性或异质性检验是正确进行元分析的基本前提，它也引发了众多方法研究者对该领域产生浓厚的兴趣。但迄今为止，对现有同质性检验方法性能的评估仍需要进行系统、深入、艰苦的研究。为此，本文将以 Hedges d 作为效应量指标对效应量同质性检验的主要方法之一——Q 检验的性能进行系统而深入的评估（以 Hedges d 为效应量指标时的 Q 检验在后面被称作 Hedges Q 检验）。同时，由于元分析者在实践领域中常常并不关心元分析所基于的假设被违背可能给元分析研究结果所带来的后果（Wolf，1990），而且，常常会轻易地依据检验结果做出接受或拒绝效应量同质性假设的二分性决策。实际上，这些做法在元分析实践领域中很可能会导致不良的后果。因此，本论文一方面希望在前人研究的基础上对 Hedges Q 检验性能进行更系统、更深入的研究；另一方面，也期待在对 Hedges Q 检验性能系统、深入研究的基础上制定出具有实践指导价值的 Hedges Q 检验检验结果的质量评估标准。

第一章　Hedges Q 检验研究现状

第一节　基本概念

为便于表述与理解，有必要对一些基本概念作出清楚、一致的解释与界定：

一　原始研究与效应量

元分析研究中最为基础的工作是穷尽式地搜集前人在某一相同或相近研究主题下独立进行的全部研究文献，而不管这些研究是否正式发表。然后，在符合研究目的的前提下按照客观、可靠与有效的纳入标准对这些研究文献进行挑选，剔除无法满足要求的那些研究，确定最终得以进入元分析的研究，形成元分析研究文献集。这里，这些前人所独立进行的、被正式纳入元分析的单个研究就是原始研究（primary study）。在这个定义中，有一点需要指出的是有些原始研究由多个独立进行的子研究组成，这种情况在心理学研究与其他科学研究中很常见。倘若这些原始研究的子研究也满足元分析的条件，则每一个子研究实际上就是一个原始研究。元分析中，每个原始研究的具体研究结果都是元分析的研究对象，为了能够对它们进行统计处理，这些研究结果常常需要以效应量（effect size）的形式进行表征。

何谓效应量？不同的研究者对其含义在表述形式上存在着一些差异。Cohen（1988）将其表述为“简单地说，效应量是一个表示研究中观察到的关系大小的指标”。Snyder 和 Lawson（1993）则将其表述为“衡量实验效应强度或变量关系强度的指标”。Kelley 和 Kristopher（2012）将其表述为“（研究）现象强度的测量”。Nakagawa 和 Cuthill

（2007）则认为，效应量是一个估计某研究效应大小的统计量。实际上只要进行简单对比，就会发现这些定义所描述的内容是相同的，即效应量是一种反映研究效应或研究变量间关系强度大小的量化指标。效应量，在国内有时也被译作效果量（权朝鲁，2003），本人认为同一个概念的不同译法虽无伤大雅，但有时会对初阅者造成困惑，有必要统一名称。由于在实验研究中，自变量对因变量的影响常常被称作实验处理效应（主效应、交互效应与简单效应等）；在变量间关系建模研究时，这种关系要么称作预测效应（回归分析），要么称作直接效应或间接效应（结构方程或路径分析），而这些效应一般不被称作效果。基于此种理解，本人认为效应量译法更为合适。

元分析研究中对效应量进行描述时，常常有几个含义不同但彼此间又有内在联系的概念，它们分别是总体效应量（population effect size）、样本效应量（sample effect size）与全局效应量（overall effect size），正确理解这些概念间的异同是展开元分析研究的基础。总体效应量有时也称作真效应量（true effect size），它指的是某原始研究中自变量对因变量的真实效应大小，或者是该研究中研究变量间的真实关系的强度。总体效应量是一个参数，其值反映的是变量间作用或关系的大小。因此，其值是一个恒量，它不受研究过程中随机抽样因素的影响。样本效应量有时也称为观察效应量（observed effect size），指的是某原始研究所探索的总体效应量在实际研究中的一次具体实现，是在该研究中所观察到的总体效应量的估计值。由于样本效应量基于研究变量的样本数据计算而得，因而其值毋庸置疑会受随机抽样因素的影响。因此，观察效应量实质上是一个样本统计量。然而，不同于显著性检验，观察效应量会随着随机抽样样本容量的增大而一致性地趋近于总体效应量。这也就是说，在样本容量达到一定水平以后，观察效应量值受样本容量大小的影响甚微。但显著性检验则不一样，在其他条件不变的情况下统计检验的显著性会随着随机抽样的样本容量的增大而越来越显著。此外，与总体效应量不同，全局效应量在元分析中指的是总体效应量分布的平均数，它可以被理解为元分析所包含的全部研究总体效应量的平均数。

在元分析的发展历史上，正是由于“效应量”这一概念的提出才使元分析有了突破性的进展。前面已经指出，就是因为这一概念的提出

才使得人们能够将相同主题下的所有原始研究结果转换到相同的量尺上来（Cooper，1998；Hedges & Olkin，1985；Huedo-Medina，Sa′nchez-Meca，& Botella，2006；Hunter & Schmidt，2004）。只有将所有原始研究的结果以相同的量尺进行量化，才可能将不同的原始研究结果进行合成。正是由于这个道理，“效应量”这一概念的提出和采用为文献的量化合成奠定了坚实的测量学基础。

二 Hedges d 效应指标

如何对研究效应进行量化？研究者在不同的研究领域、解决不同的问题时会采用不同的效应量指标。如同在医学研究领域中的对数比值比（log OR）（Fleiss，1994）或在代际历史演变研究领域中的未标准化平均数或平均差（Twenge，1997）是最常采用的效应量指标一样，Hedges d 效应指标是心理学研究领域最常采用的两种效应指标之一（Rosenthal，1994）。该效应指标主要应用于实验研究领域或非实验性质的比较研究领域（如领导力是否存在性别差异等）中研究效应大小的量化。

1969 年，Cohen 将原始研究的总体效应量被定义为 $\theta_i = (\mu_{iE} - \mu_{iC}) / \sigma_i$。这个定义中，$\mu_{iE}$、$\mu_{iC}$ 与 σ_i 分别指的是第 i 个原始研究实验组与控制组的总体平均数与共同标准差。在此基础上，Hedges（1981）数理上证明了 d_i 是总体效应量的一个无偏、一致且有效的样本估计量。Hedges d_i 的计算公式如下：

$$d_i = c(m_i) \times (\bar{y}_{iE} - \bar{y}_{ic}) / s_i \tag{1}$$

其中：

$$s_i = \sqrt{[(n_{iE} - 1)s_{iE}^2 + (n_{iC} - 1)s_{iC}^2] / (n_{iE} + n_{iC} - 2)} \tag{2}$$

$$m_i = n_{iE} + n_{iC} - 2 \tag{3}$$

$$c(m_i) \approx 1 - 3/(4\,m_i - 1) \tag{4}$$

这里，n_{iE} 与 n_{iC} 分别指的是第 i 个原始研究实验组与控制组的样本容量，s_{iE} 与 s_{iC} 分别指的是第 i 个原始研究实验组与控制组数据的标准差，$\bar{y}_{iE}$ 与 $\bar{y}_{ic}$ 指的是第 i 个原始研究实验组与控制组数据的平均数。

就其本质而言，Hedges d 指标实质上是一种标准化平均差（stand-

ardized mean difference），但标准化平均差类的效应指标除 Hedges d 指标之外还有 Cohen Δ 指标（Cohen，1969）与 Glass g 指标（Glass，1976）。Hedges（1981）研究表明，后两个指标作为总体效应量的样本估计值是有偏的。为此，d 指标可以看作Δ 指标校正版。然而，g 指标与前两者有所不同，Glass 认为在研究中施加实验处理效应会导致实验组数据的方差增大，从而导致实验组与控制组数据的方差非齐。据此，他主张效应量计算公式的分母应该采用控制组的标准差而不是实验组与控制组的联合标准差。但是，至少目前在同质性检验领域中的研究结果似乎并不支持 Glass 的这种观点（Huedo-Medina，Sa′nchez-Meca，& Botella，2006）。当然，这还需要在其他方面展开更多研究。但无论如何，当实验组与控制组数据的方差齐性时，由于 g 指标在计算效应量分母标准差时没有考虑实验组所提供的信息，其有效性不如 d 指标。此外，当效应量被估算出来之后，研究者该如何评价研究效应的大小呢？为此，Cohen（1969）给出了标准化平均差类效应量大小的经验判断标准，即效应量小、中与大的临界标准依次为 0.2、0.5 与 0.8。

三　效应量同质检验或异质性检验

通过正确采集信息进而计算获得每个原始研究的观察效应量是元分析研究的前提。此后，元分析的第一步工作通常就是对效应量进行同质性检验（Sanchez-Meca，& Marin-Martinez，1997）。为清楚阐释效应量同质性或异质性检验的内容，了解效应量同质的含义是必需的，而这又依赖于对观察效应量变异原因的分析。在现实元分析研究中，原始研究观察效应量的值彼此间常常并不相等。那么，究竟是什么原因导致观察效应量在不同的原始研究中出现变异呢？一般而言，如果扣除数据造假（学术伦理不是本文讨论的问题）的缘故，则只有两种原因：

其一，随机抽样误差（sampling error）因素。因为每个原始研究所观察到的效应量是依据样本数据而不是总体数据计算而得到的，故其本身就是一个样本统计量。因此，观察效应量无疑会受随机抽样误差因素的影响。这一类由于随机抽样误差因素所引起的效应量变异在元分析中始终是存在的，我们这里将之称为研究内变异（within-study variability）；

其二，原始研究总体效应量间存在着真正的变异。这意味着研究变量间的关系强度在不同原始研究中确实是不相等的。在这里，这种原始研究总体效应量间的变异（总体效应量方差）被称为研究间变异（between-study variance）。总体效应量的研究间变异是由于不同原始研究具有各自独特的研究特征所导致。比如，不同原始研究的被试样本特征、研究设计、实验处理、实验程序控制等方面可能各不相同等（Brockwell & Gordon, 2001; Erez, Bloom, & Wells, 2006; Hunter & Schmidt, 2000; National Research Council, 1992）。

在前面论述基础之上，假设元分析包含 k 个原始研究，这些原始研究的总体效应量分别为 θ_1，θ_2，…，θ_k。如果元分析中的这 k 个总体效应量彼此间并无差异（即 θ_1，θ_2，$=\cdots=\theta_k=\theta$），则此时观察效应量间的变异仅仅由于随机抽样误差因素所导致，我们就称这些效应量是同质的；相反，如果原始研究总体效应量间存在着研究间变异，则我们就称这些效应量是异质的（heterogeneity）。显然，效应量异质时，仅仅依据抽样误差因素不足以全面地解释观察效应量间的变异。然而，在元分析实践中，由于无法从观察效应量的变化中直观地判断这些效应量是否同质，故需要借助于一定的统计方法来探测与识别这些观察效应量是否同质。而这种借以探测与识别观察效应量是否同质或异质的统计方法就被称作同质性检验（homogeneity test），它很多情况下也会被称作异质性检验（test for heterogeneity）。实际上，犹如硬币的正反面都表述同一个硬币一样，同质性检验与异质性检验实际上表述的是相同的事情。在这里尤其要提醒读者注意，效应量同质与根据同质性检验结果推断效应量同质以及效应量异质与根据效应量同质性检验结果推断效应量异质是两回事。因为观察效应量同质，但同质性检验未必能准确判断其同质；同样，观察效应量异质，同质性检验也未必一定能将这种异质性检测出来。概而言之，元分析中效应量同质性检验的基本逻辑就是如果观察效应量的变异超过了我们所预期的、由随机抽样误差因素所导致的变异，则我们就认为效应量是异质的，即存在总体效应量间的变异，否则，我们就认为效应量是同质的。

四 名义Ⅰ类错误率、Ⅰ类错误率与统计检验力

如同任何一个其他统计检验一样，效应量同质性检验的性能评估将围绕此类检验对Ⅰ类错误率（type Ⅰ error rate）的控制表现与统计检验力（statistical power）的实际表现这两个方面展开。因此，在探讨Hedges Q 检验的性能时就有必要对一些与此有关的基本概念进行澄清。

统计学假设检验（hypothesis test）领域中，名义Ⅰ类错误率、Ⅰ类错误率与显著性水平是三个不同但又有内在联系的概念。众所周知，统计推断与逻辑推断在科学研究中是两种截然不同的推断方式。如果基于逻辑推断基础之上，那么无论是接受研究假设还是拒绝研究假设均不会面临犯错误的风险。比如，初中几何中的反证法就是属于这种推断方式。然而，在假设检验中，统计推断与逻辑推断并不一样。在统计推断中，无论最终是接受研究假设还是拒绝研究假设均会面临犯错误的可能。并且，统计推断可能犯的错误有两类。不是面临犯Ⅰ类错误（type Ⅰ error）的风险，就是面临犯Ⅱ类错误（type Ⅱ error）的风险。其中：Ⅰ类错误指的是虚无假设（H_0）为真时，但根据检验结果拒绝 H_0 时所犯的错误。在假设检验领域，为了控制犯Ⅰ类错误可能给实践或科学研究带来不良后果的影响程度，人们在依据某次检验结果进行统计推断之前，人们会事先确定容许犯Ⅰ类错误的概率（通常以 α 表示）。传统上，这个概率值会被定得很小，国际普遍采用的 α 值有 0.05 或 0.01。这个在统计检验之前就事先被确定的、容许犯Ⅰ类错误的概率就是名义Ⅰ类错误率（nominal type Ⅰ error rate），它又通常被称作名义显著性水平（nominal significance level）。名义Ⅰ类错误率体现了错误拒绝虚无假设（H_0为真）时研究者事先确定的所愿意承担的犯错风险（概率）。在确定名义Ⅰ类错误率 α 的前提下，人们可以根据抽样分布理论确定某次统计检验的接受假设区间与拒绝假设区间。此时，如果 H_0 为真，然而依据检验统计量的实际值进而拒绝 H_0 的实际概率即为Ⅰ类错误率（type Ⅰ error rate）。如果Ⅰ类错误率与名义Ⅰ类错误率吻合良好，则表示该统计检验对Ⅰ类错误率的控制非常理想。但如果Ⅰ类错误率与名义Ⅰ类错误率相差较大，则表明该统计检验对Ⅰ类错误率的控制实际上并不理想。如果Ⅰ类错误率实质性地大于名义Ⅰ类错误率，则表明该检验