DOI: 10.13718/j.cnki.xdzk.2016.02.019

元分析中统计异质性的检验：一项 Monte Carlo 研究①

陈　维[1,2]，　韦　嘉[1]，　赵守盈[2]，　张进辅[1]

1. 西南大学 心理学部，重庆 400715；
2. 贵州师范大学 教育科学学院 贵州省普通高校基础心理与认知神经科学特色实验室，贵阳 550001

摘要：元分析中，异质性检验常为模型的选择和调节分析提供重要的参考依据. 有多种方法和途径对异质性进行评价，包括 Q，H，I^2 和似然比检验. 通过模拟研究比较，发现单一的 Q 统计量可能导致错误的模型选择，H 和 I^2 值没有因为估计方法(矩估计和似然估计)的不同而对异质性的检验结果存有较大差异，是一个有效可靠的辅助统计量；Q 统计量与似然比检验相比，在 I 类错误率上控制较严格，但统计功效不如似然比检验. 研究结果表明，可以选用 Q 统计量结合报告 H 或 I^2 值的方式作为异质性检验的最佳方案.

关　键　词：元分析；统计异质性；似然比；蒙特卡罗

中图分类号：B842　　**文献标志码**：A　　**文章编号**：1673-9868(2016)02-0120-06

在元分析中，常常通过对异质性的评价作为模型选择和调节分析的重要参考依据[1]，因为只有同质的研究才能进行合并分析，所以需要检验异质性. 所谓异质性就是指由于纳入元分析中的各研究样本容量、研究设计、研究对象以及评价指标等因素的不同，使得不同效应量间存在不同程度的差异[2]. 其检验对象主要是统计异质性，它是不同研究效应量的真实差异，即真实效应的随机变异(τ^2)，受临床异质性和方法异质性的直接影响[3]，常见的检验方法有 Q，H，I^2 统计量和似然比(Likelihood Ratio，LR)检验(延用传统习惯，将统计异质性简述为异质性).

Q 统计量是基于总变异的检验，在大样本情形下，假设效应量服从(或者近似服从)正态分布，Q 就服从自由度为($k-1$)的卡方分布[1]. 若 Q 显著，表明存在异质性(由于方差不能为负值，故为单侧检验，下同)，由于 Q 统计量与异质性的大小有关，不具备尺度不变性(经线性转换后，异质性不变)和大小不变性(不受纳入研究数量的影响)，加上它易受纳入研究数量的影响(当纳入研究数量较少时，没有足够的统计功效去检验到异质性；当纳入研究数量多时，易让不重要的效应呈现统计显著性)以及不能描述异质性的大小[4-5]，所以需要寻求新的统计量，其中最简单的方法就是校正自由度，即 $H=\sqrt{Q/(k-1)}$[6]，这是一个相对值，若研究间没有异质性，则 H 为 1；若 $H>1.5$，则表明肯定存在异质性；$H<1.2$ 时，则认为同质；$1.2\leqslant H\leqslant 1.5$ 时，则不确定是否具有异质性. 但鉴于 H 的解释性不高，又延伸出 $I^2=(\tau^2/(\tau^2+\sigma_{\varepsilon i}^2))=((H^2-1)/H^2)$，反映效应量置信区间的重叠程度，并且不依赖于真实效应的大小和分布，其值越大，异质性越大，因此若大于 50%，则肯定异质性很高[2,6]. 然而，H 和 I^2 均是基于 τ^2 的矩估计(即 DL 法)[7]，在统计量的性能(如一致性、有效性和均方误)上，不如似然估计[1,7].

元分析的统计模型实际上可作为一般线性混合效应模型的特例，模型为：

① 收稿日期：2014-10-16
基金项目：贵州省科学技术厅、贵州师范大学联合科技基金项目(黔科合 LH 字[2014]7069 号).
作者简介：陈　维(1984-)，男，湖北利川人，博士研究生，讲师，主要从事心理统计与测量的研究.
通信作者：张进辅，教授，博士研究生导师.

$$\theta_k = 1(\mu_\theta) + I(\tau_k) + (\varepsilon_k) \tag{1}$$

其中：1为以1为元素的$k\times1$的列向量；I为$k\times k$的单位矩阵；θ,τ和ε均为$k\times1$的列向量[8-9]. 为保证模型的可识别性，基于抽样理论用样本误差估计代替总体误差，然后通过限制性极大似然法(Restricted Maximum Likelihood Estimation，REML，这是因为标准的极大似然法在估计固定效应时，导致自由度减少，估计值常呈负偏，故需要修正)[9-11]进行迭代估计，即可求得异质性的似然估计值[8]，然后依据LR检验判断显著性，若$P(\chi_1^2 > LR) < 2\alpha$，需拒绝$H_0$，表明存在异质性.

当然，上述方法并不乏些许模拟研究关注，如有学者通过GUASS软件生成随机数，比较了Q和I^2两者关于异质性评价的稳定性，得出了I^2值稳定于Q统计量的结果，但这些研究没有比较不同异质性估计方法(DL和REML)下的区别，没有将H值纳入比较范围[4,12]. Viechtbauer[13-14]关于异质性检验的模拟研究，其设计虽比较全面，但亦存在一个局限，即单个研究中样本容量的选择均在100以下，这与心理学调查研究中的样本容量存有较大差异，故可调整该变量对结论进行验证. 此外，在一些元分析中未检验异质性[15]，可见异质性的重要性尚未引起注意. 基于上述分析，本文通过比较Q及校正统计量、比较2种不同估计方法(DL和REML)下的H和I^2值、比较Q检验与LR检验的I类错误率和统计功效，寻求能准确有效地检验异质性的最佳方案.

1 方 法

1.1 研究设计

以Pearson相关系数为效应量，具体因素包含：① 真实效应的随机变异，即τ^2值，0,0.04,0.08,0.16[8,14]；② 效应量大小，依据Cohen[16]关于相关系数的标准，取低、中和高3个值，即0.1,0.3和0.5；③ 研究数量，取5,10,20,40和80[4,10]；④ 比较Q统计量及校正值，DL与REML的H和I^2值以及Q检验与似然比检验，共3组. 另外，样本容量取常见的200,300,500,1 000，随机分配给每个研究，前3个因素为被试间设计，最后1个因素为被试内设计. 在计算功效时，若I^2值为0，所需要的非中心参数亦等于0，没有参考价值，故在似然比与Q统计量进行比较时，不考虑τ^2值为0的情形[2,17]. 所以，共计$4\times3\times5\times2+3\times3\times5\times1=165$种组合.

1.2 数据生成及分析

数据生成及分析采用Rx64 3.0.3，以RStudio作为语言实现平台，涉及的包有“metafor”[18]. 数据生成依据随机效应模型生成，即

$$\theta_i \sim N(\theta_\mu, \sigma_{\varepsilon i}^2 + \tau^2) \tag{2}$$

以Hedges-Olkin元分析范式为基础，需将相关系数转换为Fisher z值后进行元分析，针对每种组合均生成一批正态数据，并且要求shaprio检验的p值至少大于0.10，然后再随机抽取一个样本进行分析. 功效计算指后验功效分析，故可用G*power 3.1中通用检验菜单估计功效，主要涉及χ^2检验[19]，df依据实验条件而定，置信水平α均取0.05.

1.3 比较标准

参照Field和Viechtbauer[4,9,11,14]等学者的模拟研究方法：① 针对Q统计量及其校正值，将H和I^2的结果与Q的显著性结果进行比较，只要Q检验的结果有一项与其它两类统计量的结果不一致，即可判断Q统计量需要辅助统计量；② 针对DL和REML的H和I^2值，比较两者结果是否一致，判断两者是否可以作为一个好的辅助统计量；③ 针对Q与LR检验，对两者的I类错误率和统计功效进行比较，分析两者在不同条件下的效能与差异.

2 结 果

2.1 Q统计量及校正值

Q检验及校正值的检验结果见表1. 在设计的$4\times3\times5\times3=180$个处理条件中，有174个检验结果是一致的，有6处不一致，其中5处是Q检验显著，而H和I^2统计量没有达到具有异质性的标准，它们分别是：效应量为0.3、研究数为40、研究间变异为0.16时；效应量为0.5、研究数为5、研究间变异为0.16

时；效应量为 0.5、研究数为 10、研究间变异为 0.04 时；效应量为 0.5、研究数为 20、研究间变异为 0.08 时；效应量为 0.5、研究数为 40、研究间变异为 0.16 时. 也就是说，Q 统计量提示应选择随机效应模型，但实际上异质性在可接受范围内，因此选择固定效应模型进行分析，结果的推广性可能更好，可见 Q 统计量可以让不重要的异质性在统计上变得显著. 此外，还有一处不一致的表现是 Q 检验不显著，而 H 和 I^2 统计量达到了异质性的标准，即效应量为 0.5，研究数为 5，研究间变异为 0.08. 这与上面 5 处表现恰好相反，表明 Q 统计量对于异质性不敏感，易受研究数量影响. 因此，可以认为单一的 Q 统计量在为模型选择提供依据时，存有偏差，应选用其它检验统计量或者结合其它统计量(H 或 I^2)共同检验异质性.

表 1　Q 统计量及校正值的比较

研究数	τ^2	效应量	0.1				0.3				0.5			
		统计量	Q	I^2/%	H^2	结果	Q	I^2/%	H^2	结果	Q	I^2/%	H^2	结果
5	0		1.82	0	1	√	0.14	0	1	√	0.47	0	1	√
	0.04		0.88	0	1	√	18.92***	78.86	4.73	√	6.71	40.43	1.68	√
	0.08		86.65***	95.38	21.66	√	15.11**	73.52	3.78	√	9.49	57.84	2.37	×
	0.16		14.11**	71.65	3.53	√	13.37***	70.08	3.34	√	7.88*	49.30	1.97	×
10	0		2.60	0	1	√	0.87	0	1	√	0.94	0	1	√
	0.04		39.68***	77.32	4.41	√	29.99***	69.99	3.33	√	17.30*	47.98	1.92	×
	0.08		22.65**	60.27	2.52	√	24.49**	63.25	2.72	√	117.48***	92.34	13.05	√
	0.16		145.66***	93.82	16.18	√	131.06***	93.13	14.56	√	43.09***	79.11	4.79	√
20	0		1.10	0	1	√	0.82	0	1	√	0.99	0	1	√
	0.04		8.31	0	1	√	13.17	0	1	√	10.19	0	1	√
	0.08		53.82**	64.70	2.83	√	48.25***	60.62	2.54	√	31.22*	39.16	1.64	×
	0.16		92.38***	79.43	4.86	√	47.33***	59.86	2.49	√	69.19***	72.54	3.64	√
40	0		0.76	0	1	√	0.84	0	1	√	0.93	0	1	√
	0.04		22.16	0	1	√	28.81	0	1	√	19.34	0	1	√
	0.08		34.58	0	1	√	21.49	0	1	√	47.16	17.31	1.21	√
	0.16		49.16	20.68	1.26	√	85.16***	54.21	2.18	√	76.21***	48.83	1.95	×
80	0		1.09	0	1	√	1.07	0	1	√	0.92	0	1	√
	0.04		25.42	0	1	√	19.05	0	1	√	16.06	0	1	√
	0.08		41.35	0	1	√	32.87	0	1	√	38.11	0	1	√
	0.16		77.79	0	1	√	86.53	8.7	1.1	√	68.72	0	1	

注：* $p<0.05$，** $p<0.01$，*** $p<0.001$；“√”代表 3 类检验统计量的结果一致；“×”代表不一致.

2.2　矩估计(DL)与限制性极大似然估计(REML)的 H 和 I^2 比较

DL 与 REML 关于 H 和 I^2 的估计结果见表 2. 依据 H 的判断标准，可以发现 DL 与 REML 法的结果完全一致，并无较大差异. 同时以 50%作为 I^2 值判断异质性有无的标准时，无论是 DL 法还是 REML 法，两者关于异质性的评估完全一致，没有因估计方法不同而有较大的差异，表明稳定性好. 此外，在相同条件下，H 值高的估计方法，其 I^2 值也较高，并且均相差不大，可以认为 H 和 I^2 关于异质性的评价结果具有高度一致性，均可以作为检验异质性的有效统计量. 值得注意的是 I^2 不但易于解释(非抽样误差所引起的变异(异质性)占总变异的百分比)，而且提供了异质性的大小信息，可以考虑将其作为异质性检验的首选辅助统计量.

2.3　Q 检验与似然(LR)比检验的比较

Q 检验与 LR 检验的 I 类错误率和统计功效见表 3. 研究结果显示，大多数情况下，Q 检验在 I 类错误率和统计功效上均小于 LR 检验，表明 Q 控制 I 类错误率较保守，但功效不如 LR 检验. 此外，当效应量同质性很高时，两者在 I 类错误率和统计功效上的差异很小. 但 Q 检验在研究数量较少时，统计功效不高；在研究数量较多时，I 类错误率又会很低，可能会使不重要的效应呈现统计显著性. 可见 Q 检验易受研究数量的影响，需要结合其它统计量予以共同检验异质性.

表2 矩估计与似然估计的 H 和 I^2 比较 %

研究数量	τ^2	效应量	0.1				0.3				0.5			
		统计量	H		I^2		H		I^2		H		I^2	
		估计方法	DL	REML	DL	REML	DL	REML	DL	REML	DL	REML	DL	REML
5	0		1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0
	0.04		2.521 9	2.580 7	84.28	84.99	2.174 9	2.300 0	78.87	81.11	2.447 4	2.461 7	83.31	83.49
	0.08		2.449 5	2.416 6	83.32	82.88	1.774 8	1.752 1	68.27	67.44	1.878 8	1.822 1	71.69	69.89
	0.16		4.287 2	4.532 1	94.56	95.13	2.449 5	2.535 7	83.34	84.46	4.781 2	4.741 3	95.62	95.55
10	0		1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0
	0.04		1.249 0	1.280 6	36.02	38.97	1.396 4	1.400 0	48.82	49.04	1.292 3	1.307 7	40.04	41.40
	0.08		1.288 4	1.311 5	39.81	41.69	1.737 8	1.631 0	66.87	62.34	2.032 2	1.962 1	75.76	74.01
	0.16		2.336 7	2.100 0	81.70	77.32	1.780 4	1.830 3	68.48	70.15	3.046 3	3.106 4	89.22	89.64
20	0		1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0
	0.04		1.000 0	1.000 0	0	0.19	1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0
	0.08		1.261 0	1.284 5	37.21	39.45	1.873 5	1.732 1	71.54	66.66	1.363 8	1.374 8	46.29	47.06
	0.16		1.905 3	1.913 1	72.43	72.67	1.780 4	1.830 3	68.48	70.15	2.553 4	2.576 8	84.65	84.93
40	0		1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0
	0.04		1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0
	0.08		1.000 0	1.113 6	0	19.60	1.034 4	1.131 4	6.36	21.82	1.216 6	1.272 8	32.27	38.82
	0.16		1.326 7	1.345 4	43.21	44.87	2.002 5	1.9209	75.06	72.92	1.746 4	1.661 3	67.23	63.83
80	0		1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0
	0.04		1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0
	0.08		1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0	1.000 0	1.000 0	0	0
	0.16		1.000 0	1.029 6	0	6.10	1.000 0	1.058 3	0	10.52	1.144 6	1.191 6	23.93	29.42

表3 Q 检验与似然比检验的 I 类错误率和功效

研究数量	τ^2	0.1				0.3				0.5			
		I 类错误率		功 效		I 类错误率		功 效		I 类错误率		功 效	
		Q	LR	Q	LR	Q	LR	Q	LR	Q	LR	Q	LR
5	0.04	0.0127	0.050 4	0.285 4	0.498 7	0.654 8	1.000 0	0.050 0	0.050 0	0.417 7	0.948 0	0.050 0	0.050 5
	0.08	0.028 5	0.069 9	0.217 6	0.441 5	<0.000 1	<0.000 1	0.051 1	0.987 2	0.002 4	0.004 7	0.425 0	0.807 0
	0.16	0.005 6	0.008 4	0.354 9	0.751 0	<0.000 1	<0.000 1	0.966 7	0.999 9	0.008 9	0.043 8	0.315 9	0.522 3
10	0.04	0.015 9	0.014 6	0.228 7	0.685 0	0.149 8	0.170 3	0.085 9	0.278 4	<0.000 1	<0.000 1	0.743 1	0.996 9
	0.08	<0.000 1	<0.000 1	0.836 3	0.998 0	<0.000 1	<0.000 1	0.859 1	0.999 1	<0.000 1	<0.000 1	0.693 5	0.992 2
	0.16	<0.000 1	<0.000 1	0.999 7	1.000 0	<0.000 1	<0.000 1	0.955 8	1.000 0	<0.000 1	<0.000 1	0.968 8	1.000 0
20	0.04	0.987 5	1.000 0	0.050 0	0.050 0	0.361 8	0.191 8	0.052 2	0.256 9	0.756 0	1.000 0	0.050 0	0.050 0
	0.08	0.004 1	0.001 4	0.274 9	0.892 6	0.022 5	0.025 8	0.171 1	0.606 1	0.001 5	0.000 7	0.341 8	0.922 6
	0.16	0.006 5	0.005 8	0.245 8	0.787 6	<0.000 1	<0.000 1	0.956 2	1.000 0	<0.000 1	<0.000 1	0.8028	0.999 0
40	0.04	0.999 3	1.000 0	0.050 0	0.050 0	0.998 3	1.000 0	0.050 0	0.050 0	0.997 6	1.000 0	0.005 0	0.050 0
	0.08	0.023 9	0.004 8	0.138 8	0.804 3	0.026 9	0.008 1	0.133 6	0.754 7	0.623 4	0.412 1	0.005 0	0.129 9
	0.16	0.000 4	<0.000 1	0.351 7	0.974 6	0.053 5	0.113 5	0.106 0	0.353 2	0.038 6	0.032 4	0.118 7	0.571 3
80	0.04	1.000 0	1.000 0	0.050 0	0.050 0	1.000 0	1.000 0	0.050 0	0.050 0	1.000 0	1.000 0	0.050 0	0.050 0
	0.08	1.000 0	1.000 0	0.050 0	0.050 0	0.995 8	1.000 0	0.050 0	0.050 0	0.967 0	0.943 6	0.050 0	0.050 6
	0.16	0.419 5	0.051 4	0.103 3	0.495 2	0.189 2	0.029 8	0.061 2	0.584 1	0.971 5	1.000 0	0.050 0	0.050 0

3 讨　论

元分析中，异质性检验的结果对于模型选择和调节分析均有十分重要的参考作用. 本文通过 3 种比较途径去寻求用于检验异质性的最佳方案，结果发现，单一的 Q 检验并不能保证检验结果的准确性，有时甚至会产生相反的结论.

H 和 I^2 是在 Q 统计量的基础上演变而来的，它们并没有因为异质性的估计方法(DL 和 REML)不同，而对异质性的结果判断存有差异. 相反，它们的检验结果完全一致，表明两者均是有效可靠的统计量，但两者并不是严格的统计检验，故可考虑作为辅助统计量. 通过比较 Q 和 LR 检验的 I 类错误率和统计功效，发现 Q 检验在控制 I 类错误率上较严格和保守些，而统计功效则低于 LR 检验. 实际上，这与两者的理论假设不同有关，Q 检验是立足于总变异，而 LR 检验则立足于异质性的似然估计，就此点来说 LR 检验显然优于 Q 检验. 但就实际应用来说，由于 Q 检验在控制 I 类错误率上相对较好些，若一味地追求功效反而易发生相反的检验结果，也就失去检验的本质. 此外，Q 检验简单易算，且几乎现有的元分析软件(如：Comprehensive Meta-Analysis V2，Stata，RevMan 和 Meta-analyst 等)中都会报告该统计量，而 LR 则相对较复杂，需要一定的软件编程能力，故可考虑选择 Q 统计量作为异质性的统计检验. 当然，亦可视学者期望控制 I 类错误还是追求检验效能，在权衡两类错误的前提下，选择统计检验，但本文推荐 Q 统计量结合 H 或 I^2 值的方式，作为检验异质性的最佳方案.

值得注意的是，I^2 值不但提供了异质性大小的信息，还易于解释，可理解为非抽样误差引起的变异占总变异的比值[2,6]描述了异质性在总变异所占的比率，或者说是效应量不能解释的部分所占的比例，给出了异质性的大小信息且不受研究数量影响，但此结果与有些关于 I^2 的解释存有差异[20-21]，这些研究认为 I^2 值是观察变异中效应值的真实差异所占的比值，然而，本文认为在 τ^2 的估计值中，不但有真实效应的随机变异，还存在调节变量引起的变异，所以将 I^2 直接理解为真实效应引起的变异比值，存有一定的不当之处.

参考文献：

[1] HEDGES L V, OLKIN I. Statistical Methods for Meta-Analysis [M]. New York: Academic Press, 1985.

[2] BORENSTEIN M, HEDGES L V, HIGGINS J P, et al. Introduction to Meta-Analysis [M]. New York: John Wiley & Sons, 2009.

[3] HIGGINS J P T, GREEN S. Cochrane Handbook for Systematic Reviews of Interventions [M]. West Sussex: Wiley Online Library, 2008.

[4] HUEDO-MEDINA T B, SÁNCHE-MECA J, MARIN-MARTINEZ F, et al. Assessing Heterogeneity in Meta-Analysis: Q Statistic or I2 Index [J] Psychological Methods, 2006, 11(2): 193－206.

[5] HUNTER J E, SCHMIDT F L. Fixed Effects vs. Random Effects Meta-Analysis Models: Implications for Cumulative Research Knowledge [J]. International Journal of Selection and Assessment, 2000, 8(4): 275－292.

[6] HIGGINS J P T, THOMPSON S G. Quantifying Heterogeneity in a Meta-analysis [J]. Statistics in Medicine, 2002, 21(11): 1539－1558.

[7] DERSIMONIAN R, LAIRD N. Meta-Analysis in Clinical Trials [J]. Controlled Clinical Trials, 1986, 7(3): 177－188.

[8] BROCKWELL S E, GORDON I R. A Comparison of Statistical Methods for Meta-analysis [J]. Statistics in Medicine, 2001, 20(6): 825－840.

[9] VIECHTBAUER W. Hypothesis Tests for Population Heterogeneity in Meta-Analysis [J]. British Journal of Mathematical and Statistical Psychology, 2007, 60(1): 29－60.

[10] HARDY R J, THOMPSON S G. A likelihood Approach to Meta-Analysis with Random effects [J]. Statistics in Medicine, 1996, 15(6): 619－629.

[11] VIECHTBAUER W. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-effects Model [J]. Journal of Educational and Behavioral Statistics, 2005, 30(3): 261－293.

[12] KOETSE M J, FLORAX R J, de GROOT H L. Consequences of Effect Size Heterogeneity for Meta-Analysis: a Monte Carlo Study [J]. Statistical Methods and Applications, 2010, 19(2): 217－236.

[13] THOMPSON S G, SHARP S J. Explaining Heterogeneity in Meta-analysis: a Comparison of Methods [J]. Statistics in Medicine, 1999, 18(20): 2693－2708.
[14] FIELD A P. Meta-Analysis of Correlation Coefficients: a Monte Carlo Comparison of Fixed-and Random-Effects Methods [J]. Psychological Methods, 2001, 6(2): 161－180.
[15] 胡发军, 张庆林. 大学新生 SCL-90 调查结果的元分析 [J]. 西南大学学报(自然科学版), 2009, 31(2): 152－155.
[16] COHEN J. Statistical Power Analysis for the Behavioral Sciencies [M]. 1988: Routledge.
[17] HEDGES L V, PIGOTT T D. The Power of Statistical Tests in Meta-Analysis [J]. Psychological Methods, 2001, 6(3): 203－217.
[18] VIECHTBAUER W. Conducting Meta-Analyses in R with the Metafor Package [J]. Journal of Statistical Software, 2010, 36(3): 1－48.
[19] FAUL F, ERDFELDER E, LANG A G, et al. G Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences [J]. Behavior Research Methods, 2007, 39(2): 175－191.
[20] 张辉华, 王 辉. 个体情绪智力与工作场所绩效关系的元分析 [J]. 心理学报, 2011, 43(2): 188－202.
[21] 巩文冰, 张进辅. 个体知觉的压力与情绪智力关系的元分析 [J]. 西南师范大学学报(自然科学版), 2012, 37(10): 146－151.

Testing the Statistical Heterogeneity in Meta-Analysis: A Monte Carlo Study

CHEN Wei[1,2], WEI Jia[1], ZHAO Shou-ying[2], ZHANG Ji-fu[1]

1. *Faculty of Psychology, Southwest University, Chongqing 400715, China;*
2. *School of Educational Science, Guizhou Normal University, Guizhou General Colleges Key Laboratory of Funolamental Psychology and Cognitive Neuroscience, Guiyang 550001, China*

Abstract: The result of heterogeneity test is a very important evidence for model choice and moderator analysis in meta-analysis. Many methods and approaches are available for assessing heterogeneity, including Q, H, I^2 and likelihood ratio test. According to the simulation results, we found that single Q may be inefficient for model choice. Both H and I^2 had inconsistent results of the heterogeneity test. They had no big difference due to the estimators (DL and REML). The Q test controlled the I-type error more conservatively, but its power was lower than that of the likelihood ratio test. Finally, we recommended that Q, in combination with H or I^2, be used as the best method for heterogeneity assessment.

Key words: meta-analysis; statistical heterogeneity; likelihood ratio; Monte Carlo

责任编辑 胡 杨