教育研究中效应量的使用：问题与建议*

郭　衍　宋　爽　曹一鸣

[摘　要]随着教育研究的范式转型和量化研究的不断发展，关于教育研究中统计方法使用的讨论也越来越受重视。近年来不断有科研团体、学术期刊和专家学者提出弃用p值，并强调报告效应量的必要性，其目的是警示或避免研究者对统计方法的误用和滥用。借助原始文献，从解析背后原理的角度：阐述了“废除”p值背后的无奈原因，比较了效应量和p值的内涵、功能差异；分析了目前研究领域对效应量的迷信、误读和误用；对效应量的恰当选择和大小解释及统计结果报告提出建议。提醒教育研究工作者摒弃拿来主义和经验主义，形成严谨科学的量化研究思路与做法，谨防对效应量的简单套用，避免重蹈p值的覆辙。

[关键词]效应量；量化研究；p值；结果报告；元分析

[中图分类号]G40-034　[文献标识码]A　[文章编号]1009-718X(2022)01-0035-07

随着我国教育研究范式的转型，实证研究方法越来越受重视，统计推断作为量化研究中极为重要的组成部分，对量化的描述、分析、推断以探究教育问题起着不可忽视的作用。然而，统计推断中常见指标p值近年来却陷入一场风波：从要求准确报告，[1]到调整显著性水平，[2]甚至有期刊禁止报告。[3]事实上，西方学术界关于p值大规模争论可以追溯到1999年的心理学研究中，Wilkinson和美国心理协会（American Psychological Association，简称APA）对心理学期刊中统计方法的使用提出指导方针：研究者无论是否得到统计意义显著的结果，都需要对其主要结果的效应量进行报告[4]。此后，很多关于p值的讨论大多伴随着“效应量”一同出现，如，APA于1999年开始强制要求研究者报告主要结果的效应量，仿佛效应量已经成为p值的替代品。

本文使用原始文献揭示效应量的提出者的初衷，以及他们对后来研究者的告诫。本文还提出关于效应量选择和解释的方法，以及对统计结果报告的建议，消除效应量使用和解读的迷思。希冀能够引导教育研究工作者科学规范地使用效应量，更多地关注研究结果的实际效应，不被统计方法的错用、误用和滥用束缚住科学研究的脚步。

一、当前“废除”p值实属无奈

Ronald Fisher教授在20世纪初提出p值，其目的是为了用一种客观的方式来描述数据和原假设的

郭　衍　北京师范大学数学科学学院　副教授　博士生导师　100875
宋　爽　首都师范大学教师教育学院　讲师　100089
曹一鸣　北京师范大学数学科学学院　教授　博士生导师　100875

*本文为国家社会科学基金“十三五”规划教育学国家青年课题“教育神经科学视域下学生问题解决能力发展研究”（CHA180266）的阶段性成果之一。

相符程度，0.05、0.01等显著性水平也只是一种习惯性的经验参考。近年来教育研究中的样本量不断扩大，动辄几千、几万，甚至在大型项目中轻易达到上百万、千万的规模，这使得在效应未发生明显改变的情况下，统计检验量的值因样本量增加而大幅提升，同时自由度的大幅提高也使得检验量所对应的p值较之小样本变得更小，轻松达到显著水平；从而导致只要稍微挖掘一下这些大型项目的数据，似乎就很容易得到“显著”的结果。

某些专家学者也逐渐发现了这种问题，提出了调整显著性水平参考值的想法。如，Benjamin等人在权威期刊《自然-人类行为》（*Nature Human Behaviour*）撰文主张将显著性水平从0.05调整至0.005。这种做法虽然可以过滤掉一些“伪显著”的结果，但也在一定程度上扼杀了不少“低成本”的小样本研究，致使很多有价值的研究成果无缘发表。可以想象，盲目追求将p值作为研究是否成功或论文能否发表的依据是十分可怕的：一方面，研究者得到了某种有意义的发现，但由于p值未能足够的小，不得不放弃结果，或再想办法补充样本，直至结果“足够显著”；另一方面，期刊评审和研究生导师如果只以p值大小作为标准，也会导致研究结果偏差或发表偏倚（publication bias）。长此以往恶性循环，研究者只会更倾向于选择大样本数据进行挖掘而忽视研究问题本身，或是花费大量人力、物力、财力进行无意义的样本补充和模型调试，调查抽样不再看重代表性而是追求大数量，统计描述和推断沦为数字游戏。

为了避免研究者盲目追求p值大小而忘记研究的初衷，部分科研团体和学术期刊及时作出反应。2016年，美国统计协会（American Statistical Association，简称ASA）首次以官方名义解释，p值经常被错误地使用和理解，导致了某些学术期刊劝阻甚至放弃使用p值的论文。[5]2018年初，《政治分析》（*Political analysis*）主编Jeff Gill也表示，“p值本身不足以提供支持给定模型或假设的证据”，同时“很多社会科学研究者对p值存在误解”，因此决定禁止报告p值。2019年3月，Amrhein等人在顶级期刊《自然》（*Nature*）号召科研工作者放弃使用“统计显著性”。[6]

由此可见，p值的提出是希望能够量化描述“显著性”，本身并无问题。调整显著性水平的参考值是为了进一步优化“显著结果”，但造成研究者盲目追求甚至操纵样本以减小p值的行为却违背了这一调整的初衷，所以部分期刊只能禁止报告p值以遏制不良风气，这实属无奈之举。

二、“效应量”绝非p值的替代品

国内已有一些教育研究工作者，特别是数学教育研究工作者开始关注p值的争论，从解析假设检验的角度理性分析了误用p值的现状，提出了补充规范报告置信区间、效应量等的建议[7-9]。很多研究者也已经积极响应对效应量等报告的建议，但相当数量的研究者知其然不知其所以然，以为效应量的报告就可以解决p值误用所带来的问题。

事实上，早在APA要求研究者报告效应量的半个世纪之前，著名统计学家Jacob Cohen就已经在其很多著作中介绍并推荐了效应量。“效应量（effect size）”并不是为了解决p值问题而提出的新概念，其含义几乎就如同字面意思一样，即表示某种效应的大小。2019年，Bakker、Cai等几位数学教育研究国际权威期刊的主编、编委曾联合撰文，非常前瞻性地给解释研究结果的效应量提出了十二点考虑的建议[10]。

（一）效应量的含义

效应量并不是一种类似于t值或标准化回归系数等被赋予特殊含义的、专有的统计量，其外延非常广阔。Cohen在其1969年出版的《行为科学的统计检验力分析》（*Statistical Power Analysis for the Behavioral Sciences*）中以“某种现象存在的程度（the degree to which the phenomenon exists）”引入效应量[11]。在该书后续的章节中，Cohen用“the size of the effect of …”或“the effect size of …”等描述来指代效应量。这说明Cohen只是将effect size作为一个方便使用的短语，并没有将其当成是专有名词给出明确的定义。此外，Cohen也在该书第一章的

概述中提出，效应量可以是“各种量（varying values）”，并表示为了降低这种多样性而在之后章节中所介绍的是标准化后的效应量，包括d、r、q、g、h、ω、f、f^2等，更进一步具体地表明了效应量概念外延的广泛性。

既然效应量并没有明确的定义，而是包含一系列统计量的集合概念，那么下文就介绍一些常见的效应量。Ellis在2010年出版的关于效应量的手册达到了过千次的引用量，[12]而国内介绍效应量的论文中也有两篇在中国知网上达到了过千的下载量，有理由相信这些文献对国内研究者产生了广泛影响。Ellis按照关注的问题将效应量分为两个“家族”，分别是表示组之间差异的“d family”和表示关联性的“r family”；郑昊敏等人将效应量按其统计意义分成三类，包括进行两组或多组均值比较的差异类，衡量两个或多个变量共变程度的相关类，以及总体非正态或组之间样本容量不同时描述组之间差异的组重叠类；[13]卢谢峰等人则将最常用且易于理解的效应量分为标准差异型和关联强度型。[14]以上三种分类方式有一定重合，本文基于方便数据分析的角度考虑，将常用效应量分为三类，包括：（1）用于均值差异比较，和t检验关联紧密的d家族；（2）表示关联性及解释率，和相关、回归、方差分析关联紧密的r家族；（3）表示分类变量分布差异，和卡方检验关联紧密的OR家族。（见表1）

常见效应量分类汇总（并未包括所有效应量） 表1

常见效应量	Ellis，2010	郑昊敏等	卢谢峰等	本文
Cohen's d, Glass's δ, Hedges's g, …	d family	差异类	标准差异型	d家族
比值比 OR, 风险率 RR, …		--	--	OR家族
φ, Cramer's V, …	r family	相关类	关联强度型	
$r, \tau, r^2, \eta^2, f, \Omega, \omega^2$, …				r家族
Improvement-over-chance index, …	--	组重叠	--	--

尽管这些文章的作者都强调了效应量的多样性，指出效应量可以是任何我们感兴趣的量的大小，或声明文中所列举的只是常用且易于理解的效应量，但读者往往只关注了作者所列举的表格中包含的部分效应量，却没有意识到均值、中位数、相关系数、回归的斜率、均值差异等都是重要的效应量。[15]为了避免再次被忽略，本文特意在表格的标题中作了说明。正如Ellis在其书中所描述：“科研论文的作者在并不了解效应量的情况下报告效应量的现象并不少见（It is not uncommon for authors of research papers to report effect sizes without knowing it）”，如果作者仅为了迎合出版方或审稿人的要求在不了解效应量含义的情况下机械地报告，或者不考虑效应量对所说明问题的作用，而是偏好于一些对他们来说比较“新奇的指标”，无疑违背了之前部分学术期刊、科研团体和专家学者推崇报告效应量的初衷。

经过几十年的关注和讨论后，也有越来越多的研究者认同效应量等统计指标均是为研究问题服务的。心理学领域顶级期刊《心理学方法》（*Psychological Methods*）于2012年刊登的一篇论文《关于效应量》（On Effect Size），从多个角度对效应量进行了全面的分析与描述。该文作者Kelley和Preacher给出了他们关于效应量的定义，“a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest”，[16]也就是说，他们认为效应量是“一种为解决感兴趣的问题而使用的描述某种现象程度大小的数量表达”。该定义以关注解决问题为最终目的，这样才能突显出效应量与总体参数或样本统计量之间的差异。Kelley等人的强调将有助于提醒使用效应量的研究者应当更加关注其研究问题本身，而不是迷信于某种统计量的计算公式。

（二）效应量与p值的功能区别

总的来说，效应量和p值结果的报告出于不同的目的，它们在使用数据来说明问题的方式上具有根本性的差异。假设检验试图通过样本特征对总体特征进行统计推断，其本质是一种推测，并不会也无法确定原假设正确与否；但在样本有足够代表性的情况下，假设检验中的p值可以描绘出样本和总体（特征是确定但未知的）相符的程度，它为利用样本推断总体的过程提供了一种量化的指标，p值本身是非常有用的。p值确实不能测量或代表一个效应的大小或一个结果的重要性，这是p值的局限

性，但自始至终p值就从未承担过这一作用。

与假设检验及p值不同，效应量是无法进行统计推断的，当研究者报告利用样本数据信息得到的样本效应量时，它所反映的只是样本的某一种特征或者某几种特征的综合，即使据此对总体进行“推断”，也只能非常粗略地称总体的效应量可能类似于得到的样本效应量，但这种“类似”是没有量化指标可以描述的，只能通过报告效应量的置信区间来呈现对总体的估计，而这时就会用到p值；但效应量及其置信区间的报告却能很直观地呈现某种效应，标准化的效应量也有助于提高各研究结果间的可比性。

效应量可以描述效应的大小，这是对p值局限性的弥补，但效应量反过来也不能提供p值所能提供的信息，更加不能够代表一个结果的重要性。部分学术期刊“废除”p值而提倡报告效应量，起到了价值导向的作用，希望研究者由盲目追逐显著性是否达到某种水平转而关注研究发现的实际效应大小。效应量从含义和功能上都有异于p值，绝非是p值的替代品。事实上，即使面对越来越多针对p值的批评，ASA的声明中也从未赞同或提及效应量可以取代p值。

（三）效应量的大小描述

可能以Cohen名字命名的几个效应量太过著名（如Cohen's *d*），及其给出了几个效应量“小”“中”“大”的经验划分，目前很多流传甚广的“效应量达到中等水平”“效应量达到较大水平”等表述几乎都源于Cohen给出的这份表格。为避免再次引起误解，本文特意没有列出该表格。其大致内容包括若干常见效应量，并分别给出小于某值时为小，介于某区间时为中，大于某值时为大。这导致许多研究者对效应量一知半解、争相效仿，将效应量超过某一值、达到某一等级视为检验研究结果的最高准则，仿佛又走起了追求p值大小的老路。

效应量报告中的常见错误就是过度依赖关于效应量大小的经验划分。许多研究者奉行“拿来主义”，直接使用了Cohen的标准，却忽视了其反复强调的关于使用此标准的注意事项。Cohen在《行为科学的统计检验力分析》一书首章关于效应量的总述中便提及，他将对每一个所列举的效应量给出形容其大小程度的操作性定义，紧接着便毫不避忌地以“有很大风险（with many dangers）”“武断（arbitrary）”“存在被误解的风险（a risk of being misunderstood）”来形容自己的做法。而在之后的每个章节中，一旦提及“小”“中”“大”的操作性定义，Cohen更是不厌其烦地强调不应该依赖这个标准。例如，他指出“仅仅在没有更好的依据时才推荐使用”（该书2.2节）；Cohen还预见到“读者可能会发现此处定义的‘大’对其研究领域来说过大或过小”，此时他“竭力主张研究者给出更合适的定义”（该书3.2节）；如果研究者可以依据理论给出更恰当准确的标准，他则“强烈劝阻研究者使用该书中的操作性定义”（该书4.2节）。由此可见，效应量的提出者反复告诫不要依赖他给出的经验划分，生怕对后世造成不良影响。

事实上，不仅“小”“中”“大”本身是相对的，这种相对性还表现在研究领域的差异上，甚至可以说，对教育研究中的特定研究领域、研究问题和研究方法，这种“小”“中”“大”都是相对的。尽管Cohen已经不断重申自己的警告，但后来的研究者“拿来主义”的做法还是层出不穷。Glass等人在1981年就已经指出，将效应量数值的区间和“小”“中”“大”等描述性形容词相对应是“极不明智的（there is no wisdom whatsoever）”[17]。Thompson也曾在2001年撰文告诫，如果像对待p=0.05那样僵硬地依据某个固定的标准来解释效应量，那么就是“在另一种框架下重复做着同样的蠢事”。[18]2012年，范德堡大学Lipsey等人在为美国教育部下属的教育科学研究院起草的报告中，正面抨击了盲目使用效应量大小标准的行为[19]，其报告称，许多研究者完全忽略了Cohen的警告，而这些不加区分地在并不适用的领域采用Cohen的大小定义的行为是“不恰当且有误导性的”。

效应量的大小更多的应该是与同类研究进行比较，Lipsey等人的报告还给出了一个极好的例子。他们提醒研究者注意，对于学生学业成就发展的研

究来说，如以数学测试前后测差异的标准化指标作为效应量，大量的干预研究（教学实验）均鲜能达到0.30的效应量大小。那么，基于这些研究的效应量的分布，一个干预效果达到d=0.25的效应就可以称之为“大”了，而被Cohen普遍定义为“中等”的d=0.5简直可以称为“巨大（huge）”。

三、如何正确使用效应量

总结前文，效应量的误解及误用主要集中在三个方面：（1）对效应量的作用过于迷信；（2）对效应量的理解过于狭隘；（3）对效应量大小描述的使用过于刻板。在了解效应量的基本概念及其大小描述的正确观念后，前文已经说明了效应量无法替代p值的功能，下文提供的一些报告效应量的建议，也有助于研究者破解“神化”效应量的迷思。

（一）如何选择合适的效应量

前文已经说明了效应量具有多样性，正因为这种多样性，在研究报告中如何选择一个合适的效应量就变得尤为重要。Preacher和Kelley等人讨论了“合适”的效应量应该具备的特征：包括应当被恰当地测量并给出测量方法和用来说明的问题；应当同时报告置信区间；对总体效应量的点估计应当独立于样本量；测量方法应具有无偏性、一致性、高效性等特征。[20]

当然，并非所有报告的效应量都要完全满足以上要求，尤为关键和最为基本的要求是，应当选择最适合解决研究问题的效应量。很遗憾，众多国内（包括国外）教育研究的文献中关于效应量的选取和报告都未能满足基本要求，和“合适”的要求更相差甚远。例如：在用方差分析研究组间差异并且不关心事后检验结果时，不报告η^2描述方差解释的比例却报告Cohen' s d；在报告样本效应量之后试图以样本效应量指代总体效应量却不报告置信区间；在报告某些计算公式不统一的效应量，如d的时候未说明测量方法（未指出使用联合标准差还是控制组标准差）；等等。

（二）如何对效应量的大小进行解释

效应量的大小应当是针对所在研究领域及研究问题的，不区分研究领域及研究问题的“经验值”都是武断的，对效应量大小的判定不存在一个普遍适用的法则。合适的标准应该基于研究对象相似、实验方法相似、测量方式相似的研究结果的分布来确定。

Lipsey等人的报告提供了定义大小标准的一个正面示例，对教育研究者有很大的参考意义。如果想了解某种新的教学方法是否有效，可以和对应年龄学生学业成就自然发展的变化进行比较。如，Bloom等人根据CAT 5、SAT 9等标准化学业成就追踪测试得到每一年学生阅读、数学、科学、社会学科的平均提高程度，发现每年数学学业成就的标准化提高值存在差异，小学一年级到小学二年级的提高效应量为1.03，而初中一年级到初中二年级的提高效应量为0.32。[21]此时，如果某研究通过历时一个月的干预，结果的标准化效应量为0.30，对于初一年级学生来说这就应当被定性为“大”的干预效果，而对于小学一年级的学生来说，可能就应当定性为“中等”或“小”的干预效果了。

因此，在解读效应量的“定性”程度时，和前人的研究结果或大样本研究及元分析结果的效应量进行比较，将有助于对该数据结果进行更为客观及针对性的解读。

（三）警惕效应量的误用和滥用

要求报告效应量的本意源于许多研究者对假设检验结果的误读和过度解释，并且研究者可能根本没有提供关于原始结果的完整信息，让想要全面了解研究结果的审稿人或读者无法计算效应量。另外，标准化的效应量使得相似研究间的比较和元分析的综合概括成为可能。基于此，一些科研机构和学术期刊才特别提出了报告效应量的硬性要求，毕竟目前研究者已经会自觉报告假设检验的结果。

对假设检验的质疑和批评都源于研究者对它的误读和过度解释，试图让p值承担它不应该承担的作用，甚至引起了一些学者的错误批判[22]。同样，对效应量的误用和过度解释，将有可能使其沦为下一个p值。效应量虽然可以解释p值无法解释的内容，但它也无法起到p值刻画推断的作用，如果在

样本中得到了某个效应量，就“断定”总体也具有该性质，就同样犯了过度解读的错误。

四、总结与讨论

（一）规则应当先立后破

行文至此会产生一种感觉，似乎都是当年的统计学家提出的经验参考值导致了后来研究者对p值和效应量的误用和滥用。那么，统计学家为什么不出面解决问题呢？因为在统计学中，这根本就不是问题，统计学家只提供了一些他们认为好用的工具，而对于其他研究者使用这些工具后所作的推论和结论，统计学家认为自己并非该研究领域的专业人士，自然不会指手画脚，同时也认为自己所给出的经验划分会被其他领域的专业人士理智看待。现阶段，贸然破除统计学家所提供的一般性规则，在论文撰写中“废除”p值甚至效应量，无疑是不明智的，只会令教育研究者更加无所适从。目前，我国教育研究领域对量化研究方法的使用仍处于蓬勃发展阶段，很多研究者尚属于初学者，依靠研究者本人的经验或对同类研究的掌握从而对自身研究结果作出评判，可能还是要求过高。这种属于“后规则”时代的要求不能一蹴而就，只能希望有能力的研究者具有这种自觉。

作为统计方法的使用者，心理学领域发展更为成熟，值得我们借鉴参考。依据APA在2010年发布的《美国心理协会出版手册》，“假设检验是起点，在这之后增加报告效应量、置信区间和全面的描述才能表达出结果的完整含义”。所以，学位论文、期刊论文可以设置规则，要求教育研究者在报告统计结果时，p值、效应量、置信区间等都完整呈现。至于显著性到底如何，效应是否理想，研究生导师和论文评审则应更多地肩负起评判的职责，甚至可以成立专门的学术委员会接受咨询。待研究者对统计方法的运用、对统计结果的解读都日趋成熟时，再逐渐将这种“职责”交还给研究者本人，从而破除规则。

（二）鼓励元分析的研究

对效应量的解读更多需要依赖同类或近似研究的结果作为参考，而元分析（meta-analysis）就是提供这种参考的理想来源。简单来说，元分析就是对一系列论文中的统计指标进行再分析，从而得到基于更大群体或更多类型的综合结论，可以被视为针对量化研究结果的量化文献综述。在教育研究领域，近年来已成为畅销书的《可见的学习》就是一本综合了800多项元分析的综合报告[23]，分学生、家庭、学校、教师、课程和教学策略六个领域分析了诸多因素对学生学业成绩的影响，为同类影响因素研究结果的解释和比较提供了极为有力的参考。例如，元分析结果显示教学策略使两个平行班学生产生成绩差异的d值为0.4，如果某项教学实验只能达到0.3，则其实是一个低于“总体水平”的效应。当然，这种参考还可能受限于策略类型、学段、学科等差异，特别是文化、地区差异的影响，所以我国教育研究领域发展元分析研究是十分必要的。

随着元分析技术的不断进步与完善，调节变量的分组研究在元分析中也越来越受到重视。元分析的目的从累计样本，到探索不同类型群体、测量方式的效应量差异，从量变实现了质变。同样以教学策略对学生学业成绩的影响研究为例，现代元分析技术考虑到样本的异质性，对不同教学策略类型、学生的年龄、学业成绩的学科、研究所在国家等编码后进行调节变量分析，可以得到不同情况下的综合效应量及其差异，拓展了分析对象本身的研究问题，为同类研究提供了更为细致丰富的参考，赋予了元分析崭新的生命力。

开展元分析研究和合理准确使用效应量两者也是相辅相成的。元分析结果方面对个体研究结果的效应量能够提供解释和比较的参考；元分析的对象即是诸多个体研究中提取出来的效应量，规范报告效应量也是元分析得以开展的重要保障。

（三）对统计工具存敬畏之心

相较于统计工具，教育研究者对于自己使用的其他研究工具往往有很好的把握。例如使用问卷或测试前，总是认真划分维度、详细论证信效度、经过多轮试测和专家评议，才能使用相应的工具。同样，对于统计工具的使用，自然也绝不能在尚未真

正弄清其原理的状态下就“盲目依赖”。教育研究者使用统计工具时，应该在乎的是统计本身，而非统计软件的操作；就像我们使用问卷时更应在意的是题项内容和整体结构，而非选择在何种网络平台发布问卷。

事实上，对于有把握的工具，我们自然不会产生“神化”和“误用”。例如，将学生成绩划分为优、良、及格、不及格，教育研究者可能在部分场合也会使用这种表达方式，但决不会执着于这种粗糙的划分，认为59分和60分存在什么重大差异。这种划分可以帮助“外行人”大致理解分数的含义，“外行人”不知道怎么去看均值、标准差、分布，也不知道考试的命题蓝图、不同题目的考查目标等。

作为教育研究工作者，既然我们使用统计方法来解决教育问题，就不能以“外行人”自居，不能盲目追求经验划分，更不能将之视为评判研究结果的最高纲领。对统计工具应有相同的敬畏之心，关注其原理而非操作，至少应当对自己使用的统计方法的基本理念有所了解，有足够的信心对自己的统计结果给予解释。

最后，以ASA声明中的一段话作为结语：“良好的统计实践作为良好的科学实践的重要组成部分，它强调：良好的研究设计和实施，各种数值和图表的总结，对所研究现象的理解，对所得结果的解释，有完整的报告和适当的逻辑，对数据的含义有量化的理解。没有任何一种单一的指标可以取代科学推理”——无论是p值，还是效应量。

[参考文献]

[1] Association For Psychological Science.Submission Guidelines[EB/OL].(2017-10-30) http://www.psychologicalscience.org/publications/psychological_science/ps-submissions.

[2] Benjamin, D. J., Berger, J. O., Johannesson, M., et al. Redefine statistical significance[J].Nature Human Behaviour,2018,2(1):6-10.

[3] Gill, J. Comments from the New Editor[J].Political Analysis, 2018,6(1):1-2.

[4] Wilkinson, L. Statistical methods in psychology journals: Guidelines and explanations[J].American Psychologist, 1999, 54(8):594.

[5] Wasserstein,R. L.,& Lazar,N. A.The ASA's statement on p-values: context, process, and purpose[J].The American Statistician, 2016,70(2):129-133.

[6] Amrhein, V., Greenland, S., & McShane, B.Scientists rise up against statistical significance[J].Nature, 2019, 567(7748):306-307.

[7] 王光明，李健，张京顺.教育实证研究中的p值使用：问题、思考与建议[J].教育科学研究，2018(2):59-65.

[8] 宋爽，曹一鸣.如何正确解读假设检验结果——兼谈数学教育研究中p值误用问题[J].数学通报，2019，58(7):14-18.

[9] 沈光辉，范涌峰，陈婷. 教育研究中的P值使用：问题及对策——兼谈效应量的使用[J].数学教育学报，2019，28(4):92-98.

[10] Bakker, A., Cai, J. , English, L., et al.Beyond small, medium, or large: Points of consideration when interpreting effect sizes[J]. Educational Studies in Mathematics, 2019(102):1-8.

[11] Cohen, J.Statistical Power Analysis for the Behavioral Sciences (2nd Edition)[M].NJ: Lawrence Erlbaum Associates, 1988.

[12] Ellis, P. D. The Essential Guide to Effect Sizes: Statistical Power, Meta-analysis, and The Interpretation of Research Results[M]. Cambridge University Press, 2010.

[13] 郑昊敏，温忠麟，吴艳.心理学常用效应量的选用与分析[J].心理科学进展，2011，19(12):1868-1878.

[14] 卢谢峰，唐源鸿，曾凡梅.效应量：估计、报告和解释[J].心理学探新，2011，31(3):260-264.

[15] Lipsey, M. W., & Wilson, D. B. Practical meta-analysis[M].NY: Sage Publication Inc, 2001.

[16] Kelley, K., & Preacher, K. J. On effect size[J].Psychological Methods, 2012, 17(2):137-152.

[17] Glass, G.V., Smith, M. L., & McGaw, B. Meta-analysis in Social Research[M].Sage Publications, Incorporated, 1981.

[18] Thompson, B.Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field[J].The Journal of Experimental Education, 2001, 70(1):80-93.

[19] Lipsey, M. W., Puzio, K., Yun, C., et al.Translating the statistical representation of the effects of education interventions into more readily interpretable forms[J].National Center for Special Education Research, 2012:54.

[20] Preache, K. J., & Kelley, K.Effect size measures for mediation models:quantitative strategies for communicating indirect effects[J]. Psychological Methods, 2011, 16(2):93-115.

[21] Bloom, H. S., Hill, C. J., Black, A. R, et al. Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions[J].Journal of Research on Educational Effectiveness, 2008, 1(4):289-328.

[22] 温忠麟，吴艳.屡遭误用和错批的心理统计[J].华南师范大学学报：(社会科学版)，2010(1):47-54.

[23] Hattie, J. Visible learning: A Synthesis of over 800 Meta-analyses Relating to Achievement[M].Routledge, 2008.

（**责任编辑**：周　镭）