

·专题:文献证据检索的科学性评价·

编者按:在社会科学领域,由于研究对象的异质性、多元性和情境依赖性,常常使研究者针对同一研究问题却得到千差万别的研究结果,从而使其研究发现在社会现实中的应用受到局限,也致使科学界及社会对社会科学科学性存在质疑。循证社会科学领域的研究者运用循证研究的理念和方法,探索以社会科学领域的原始研究和社会实践中所获得的原始证据为起点,通过系统评价(systematic review)和元分析(meta-analysis)对结果加以研究整合(research synthesis)以获取更高层次的证据,进而实现社会科学领域证据的质量提升和转化应用,最终助推社会科学领域的科学化和标准化发展。文献证据检索作为循证社会科学领域展开证据整合的关键环节,如何获取高质量的原始证据就成为了一个至关重要的研究问题。

基于上述背景,本刊特邀西北师范大学周文杰教授团队基于兰州大学杨克虎教授担任首席专家的国家社会科学基金重大项目“循证社会科学的理论体系、国际经验与中国路径研究”(项目编号:19ZDA142)组织了本“文献证据检索的科学性评价”专题。作为该重大基金项目第一子课题“循证社会科学的理论体系研究”相关研究成果,专题文章以循证社会科学文献证据检索质量评价为研究对象,在引入信度、效度等测量工具质量评价指标的同时提出了证据检索的饱和度、冗余度、敏感度等文献证据评价指标,并通过全面系统的实证研究对文献数据库中各类检索途径的检索效率进行了对比及分析。

本专题论文所提出的饱和度、冗余度、敏感度和信度等评估指标,以及所得出的相关研究结论,既是循证社会科学学术话语体系中不可或缺的重要组成部分,也为构建循证社会科学理论体系奠定了基础。

循证视角下文献证据检索的科学性评价:缘起、指标与趋势*

周文杰^{1,3} 赵悦言¹ 魏志鹏^{2,3} 杨克虎^{2,3}

(1.西北师范大学商学院 甘肃兰州 730070)

(2.兰州大学基础医学院循证医学中心 甘肃兰州 730000)

(3.兰州大学循证社会科学研究中心 甘肃兰州 730000)

摘要:科学化、规范化的循证社会科学研究以全面、精准的文献证据检索为基本保障。对文献证据检索的科学性评价在立足于传统的查全率和查准率指标,以及引入信度、效度等测量工具质量评价指标的基础上,进一步提出了饱和度、冗余度、敏感度等指标。基于循证社会科学的未来发展,可以发现,文献证据检索的科学性评价呈现出检索过程趋于标准化、检索结果评价趋于可计量化、不依赖于文献数据全集的自动化评估方法、考虑检索成本的效益等发展趋势。

关键词:循证社会科学;文献证据检索;科学性评价;指标体系

中图分类号:G252.7

文献标识码:A

DOI:10.11968/tsyqb.1003-6938.2021089

Scientific Evaluation of Literature Evidence Retrieval Quality from the Perspective of Evidence-based Research: Initiation, Index and Trend

Abstract The retrieval quality of literature evidence is the basic guarantee of high-level evidence-based research. Based on the analysis of the initial problem of document evidence retrieval in the field of evidence-based research, this paper draws on the reasonable components of traditional recall and precision indicators and overcomes their shortcomings, puts forward new literature evidence retrieval quality assessment indicators such as saturation, redundancy and sensitivity, and introduces the indicators of the quality of measurement tools such as reliability and validity into this field. Furthermore, this paper shed light on the future trend of scientific evaluation of literature evidence retrieval quality in evidence-based social science research. The literature quality evaluation index involved in this paper lays a foundation

* 本文系国家社会科学基金重大项目“循证社会科学的理论体系、国际经验与中国路径研究”(项目编号:19ZDA142)研究成果之一。

收稿日期:2021-11-28;责任编辑:胡刚

for the follow-up empirical research.

Key words Evidence-based Social Science; literature evidence retrieval; scientific evaluation; index system

循证研究的基本目的,是为了对存在分歧甚至对立的原始研究证据加以有效整合,以获取更高层次、更具有普遍意义的科学证据。Jessica Gurevitch 等^[1]提出,对原始的科学研究结果加以综合,以达到全面理解和解决问题,并确定研究结果变化的来源是科学进程的基本组成部分。迄今为止,循证研究领域已发展了系统评价和元分析等一整套理论、方法与工具,Cochran、Campbell 等网络也为循证研究的规范化和更高层次研究证据的整合与交流提供了平台。近二三十年来,循证研究呈现出了由医学领域向社会科学扩展的趋向,系统评价(systematic review)、元分析(meta-analysis)及研究结果整合(research synthesis)的理论与方法也呈现出了蓬勃的发展态势^[2]。

在社会科学研究领域,由于研究对象具有多元化、异质性及依情境而变等特征,导致原始研究所获取的证据与自然科学相比存在更多的局限性和不稳定性,从而更迫切地需要社会科学领域开展科学的系统评价和元分析,以便获取更高层次的研究证据^[3]。无论是在自然科学领域的循证研究还是循证社会科学研究中,原始文献检索质量的高低都是系统评价和元分析能否消除偏倚,获得高质量证据的首要因素。为此,研究者亟待对循证社会科学研究中文献证据检索的质量展开深入评价,以确定相对科学的检索标准,从而保障基于这些原始研究证据而展开的系统评价和元分析更具科学意义。基于这一背景,本文旨在对文献证据检索科学性评价问题的缘起加以回顾,在汲取传统的查全率和查准率指标合理要素前提下,提出饱和度、冗余度、敏感度等指标,并引入信度和效度评估方法,以期发展出适合于循证社会科学自身特征的新的文献证据检索质量评价指标体系。在此基础上,本文还将对文献证据检索科学性评价的趋势做出判断。

1 文献证据检索科学性评价问题的缘起

科学文献的特征之一,是其中充斥着对某一科学问题的反复研究。研究者之所以对同一现象、问题或假设进行多次的重复分析,是为了获得更加概括、

更加接近于真实、具有更高质量的证据^[4]。然而,很多研究者都发现,即使针对同样的问题采用了类似的研究设计,研究者所获得的研究结果也常常存在差异,甚至存在相互矛盾和对立^[5-6]。《心理科学透视》(Journal Perspectives on Psychological Science)杂志曾出版了一期特刊,专门对重复研究得到不同发现的现象进行了系统评述^[7]。很多证据都表明,与自然科学相比,社会科学领域针对相同研究问题而得出不同研究结果的现象尤为突出。

自 1992 年,加拿大学者 Gordon Henry Guyatt 等首倡在医学教育领域应用循证方法以来^[8],研究证据整合的理论、方法和工具一直得到学界的广泛关注。近二三十年来,系统评价和元分析作为循证领域用以进行原始证据整合,获取更高层次证据的基本手段,尤其受到重视。1997 年,Lipsey 和 Wilson^[9]发表了基于 302 篇社会科学领域关于处理效应的元分析述评文章,标志着社会科学领域的研究证据整合进入了新的阶段。同年,Cochrane 合作网络正式成立,成为首个研究证据整合的全球性合作平台。1999 年,以促进社会科学领域研究整合为主要目标的 Campbell 合作网络建立,使社会科学领域的循证研究和循证实践具备了更加坚实的基础。

在研究证据整合的过程中,元分析的工具和方法扮演着极其重要的角色。迄今为止,元分析方法的发展和完善经历了若干重要阶段。Cochran^[10]指出,1954 年首个元分析中固定和随机效应计算方法的提出,1986 年研究间方差的累计计算方法的发展^[11],1997 年关于漏斗图(funnel plot)和 Egger 检验(Egger's test)在发表偏倚识别中的应用^[12],2002 年关于异质性检验指标 I^2 的提出^[13]可被视为元分析发展历程中的里程碑。经过多年的发展,当前元分析的方法和工具已越来越多样、丰富,其科学程度也越来越得到各领域研究者的认可。特别是在 1995 年“系统评价”这一术语提出以来^[14],循证领域的研究者进一步发展了 PRIMA 等一系列系统评价质量评估工具^[15],极大地提高了循证研究的规范性,使科学研究结果的整合在整

个科学发展的进程中发挥了重要的影响力。

尽管系统评价和元分析的科学化和规范化有效地促进了研究结果的整合,极大地提升了循证研究的质量,但如本文所述,系统评价和元分析的质量首先取决于原始证据获取是否全面。也就是说,如果对原始证据的检索存在着偏差,则无论系统评价和元分析的程序如何严谨、方法如何科学,其结果都可能存在偏倚。从这个意义上说,原始研究证据的检索是保障循证研究结果科学性的首要问题。然而,通过文献调查发现,迄今为止,学术界在文献证据检索的科学性评价方面尚无明确统一的评价标准,存在着明显的研究薄弱点。

着眼于促进社会科学领域研究证据的整合和高质量应用,循证社会科学领域尤其需要发展出科学规范的文献证据检索评判标准。在我国,2019年由杨克虎教授作为首席专家的国家社会科学基金重大项目“循证社会科学的理论体系、国际经验与中国路径研究”得以立项,标志着我国循证社会科学研究与应用已进入深化发展的新阶段。在我国循证社会科学蓬勃发展的背景下,发展一套统一的文献证据检索的质量评价规范和标准,对于促进循证社会科学理论的完善和实践的应用意义重大。

2 文献证据检索科学性评价的指标

围绕文献检索质量的评价,信息资源管理等领域都已展开了大量研究。这些研究表明,评价检索质量需要同时考虑两个相关关联的因素:在尽可能把相关的文献全部纳入进来的同时,把不相关的文献排除出去。按照这种逻辑,信息检索等领域已发展了查全率和查准率等指标,用于检索质量的评价。

2.1 查全率和查准率

查全率(Recall Ratio)主要是指从文献数据库内检出的相关文献数量在文献总体所占的比重。这一指标主要用于衡量在特定检索中检出相关文献的能力。查全率越高,意味着检索获得的相关文献越全面。彭奇志^[16]将影响查全率的因素总结为如下两个方面:首先,从文献数据库的角度来看,数据库收录文献信息不全,索引词汇缺乏控制和专指性,词表结构不完整,检索词间关系模糊或不正确,标引不详,标引前后不一致,标引人员遗漏了原文的重要概念

或用词不当等都可能影响查全率;其次,从检索者的检索方式来看,检索策略过于简单,检索词选择不当或检索词逻辑组配不当,检索途径和方法单一,检索者不够熟练或缺乏耐心,检索时不能全面地描述检索要求等也可能对查全率产生直接影响。向禹和付文韬^[17]分析发现,查全率存在如下局限性:首先,查全率描述是检索出的相关文献数量与存储在检索系统中的全部相关文献总量之比,但系统中相关文献问题究竟有多少一般是不可知的,只能估计;其次,查全率是一个建立在“假设”基础上的评价指标,这种“假设”是指检索出的相关信息对用户具有同等价值,但对于用户来说,所检出文献的相关程度可能比它的数量要重要得多。基于此,尹舒力^[18]指出,认为查全率“是一个不实际的概念”。

查准率(Precision Ratio)用以衡量特定检索中拒绝不相关文献的能力,主要指特定检索中,实际检索出来的文献中相关文献所占的比率。1956年,J.W.佩里、A.肯特等人首先提出了此项评价指标。1979年,F.W.兰开斯特在《情报检索系统——特性、试验与评价》(第二版)一书中对查准率的评估方法进行了进一步操作化,使之更容易被计算^[19]。查准率主要取决于检索语言的专指性和所拟定的检索策略能否准确表达用户真正的情报需求。若检索策略拟订的较宽泛,参与组配的检索词较少,主题词的概念比用户的信息需求宽泛,则查准率将降低。

查全率和查准率之间具有互逆的关系。如在极端情况下,如果研究者检索得到了文献数据库中所有文档,则获得了100%的查全率,但此时查准率却很低;如果研究者检索只能获得唯一的文档,查重率很低,但却可能有100%的查准率。如本文所述,鉴于文献证据检索质量之于循证社会科学研究结果整合的极端重要性,有必要在现有查全率与查准率指标的基础上加以进一步细化,发展出更加具有操作性的评价指标,以便保障系统评价和元分析开展之前文献证据获取的科学性。饱和度和冗余度是基于此种背景而提出。

2.2 饱和度和冗余度

饱和度和冗余度是一对在汲取查全率和查准率等传统检索评价指标合理成份的基础上,为发展更具有操作性的文献证据评价指标而提出的概念。所

谓饱和度,是指检索中不再有新的文献被纳入的状况。而冗余度则是一个与饱和度紧密关联的概念,具体指在特定检索过程中检索到不相关文献的情况。

在实际操作中,饱和度可以通过“滚雪球”的办法而得到。具体过程是:首先,研究者先通过特定的检索词,通过文献数据库所提出的各种检索入口进行检索,获得相应的文献,并对其去重;其次,对检索所获取的文献的参考文献进行再次梳理,将在前次检索中未获得的新文献纳入其中;第三,对参考文献的参考文献再进行梳理,并将新文献再次纳入。如此往复,直到不再有新的文献被纳入时,文献检索既达到饱和。构建饱和度指标的意义在于,这一指标为评价文献证据检索科学性提供了一个实质性的参照系。通过参照已实现饱和的文献数据集,研究者可以进一步发展出一系列文献证据检索科学性的评价指标来。

在形成饱和数据集的前提下,为进一步对文献证据检索的质量作出判断,研究者可以通过专家判断的方式,将实现饱和的文献数据集中的文献按照其与检索主题的相关度加以划分,如划分为高度相关、中度相关与低度相关三部分文献。这些文献中,高度相关和中度相关文献可用来评价文献检索的精确性(即查准率),而低相关文献则可用来评价文献检索的冗余度(从反面体现查准率)。可见,一个饱和检索的文献数据集可以同时作为检索精确性和冗余度评价的参考标准。

基于饱和度指标,可以对传统的查准率加以进一步完善和拓展。具体作法是,将饱和度区分为纯净饱和度和一般饱和度。其中纯净饱和度是指采用特定检索方式检索结果涵盖总文献数据集中高相关文献的程度,具体计算方法是:采用单项或者组合检索时与总数据集中高度相关文献的重合率,这一指标反映了特定检索途径是否能够准确定位高度相关研究证据的能力;而一般饱和度指特定检索结果涵盖整体数据集中中度相关文献的程度,这一指标反映了特定检索是否能够准确定位中度相关研究证据的能力。同理,基于冗余度指标可以从另外一个角度对传统的查全率指标加以完善。也就是说,通过计算特定检索途径所获得的文献在穷尽检索数据集低相关文献的比值,可以有效衡量特定检索途径获得无关文献的程度,从而对查准率作出反向

地评估。

总之,饱和度和冗余度是一对植根于传统的查全率和查准率指标但更具可操作性的评价指标。这对评价指标的提出,有助于为循证社会科学研究者提升文献证据检索质量提供重要参照。

2.3 敏感度

与传统的查全率和查准率指标相比较,饱和度和冗余度具有了更高地可操作性。由于饱和数据集的构建是一件极其繁琐的工作,在无法进行饱和检索的前提下,就有必要考虑特定检索方式在未达100%饱和时的检索质量,基于此,敏感度指标得以提出。

所谓敏感度,是指在不同样本覆盖度下,特定检索项目的查全率和查准率。这一指标的具体测度方法是,应用主题、题名、关键词、摘要和全文等单项检索与组合检索的不同抽样水平的数据与总数据中高相关组和中相关组进行匹配,分别计算得到的高相关组和中相关组匹配比例。在样本数目不同的前提下,如果检出的高、中相关文献匹配度均比较高,则表明相应的检索途径稳健而不敏感,从而具有相对更高的检索质量。

2.4 信度

检索的稳定性和可靠性是衡量检索质量的另一个重要指标。参照测量领域的一般作法,可以选用信度指标作为检索质量稳定性和可靠性的评价工具。

信度(Reliability),即可靠性,是指采用同样的方法对同一对象重复测量或者应用同种方法对同一现象在不同时点加以测量时所得结果的一致程度。这种一致性常常通过相关系数来表达,相关系数越高,则多次测量的结果越一致,测量结果就越稳定、可靠。

迄今为止,研究者已发展了重测信度、复本信度、折半信度、 α 信度系数等多种方法^[20],用以对信度作出科学评价。针对文献证据的检索质量评价,重测信度和复本信度的评价相对比较直观,具有更高的可操作性。具体而言,文献证据检索的重测信度指在不同时点上,针对相同的检索主题,在同一个检索途径下所获得文献的相关程度;而复本信度则指针对同一检索主题,但通过不同检索途径而获得的文献的相关程度。显然,无论是文献检索的重测信度还是复本信度,其相关系数越高,文献检索的稳定性和可靠

性就越有保障,因而检索质量越高。

2.5 效度

饱和度、冗余度、敏感度和信度从不同侧面评价了文献证据检索的质量。然而,需要注意的是,高质量的文献证据检索虽然为循证研究者开展系统评价和元分析提供了基本保障,但却并不能总是保证系统评价和元分析的科学性。这是因为,所检索到的文献从内容上是否足以涵盖循证研究的具体领域,也将对研究证据的整合产生重要影响。这种文献内容之于系统评价和元分析所需证据的覆盖程度,可以通过检索的效度加以评价。

效度(Validity)即有效性,它是指测量工具或手段能够准确测出所需测量的事物的程度。按照测量理论,效度主要用以衡量所测量到的结果反映所想要考察内容的程度。测量结果与要考察的内容越吻合,则效度越高;反之,则效度越低。目前,测量领域已发展了内容效度、效标关联效度、结构效度、表面效度等诸多具体评价方法。

在文献证据检索质量的评价中,内容效度、效标关联效度和结构效度都有着极其广泛的应用前景。如效标关联效度可以用来评价特定检索途径所获得的检索结果与整体结果之间的吻合程度,内容效度和结构效度则可以用来进行检索证据覆盖度与全面性的评价。

在文献证据检索质量评价中,不仅可以借鉴测量领域关于效度评价的方法,而且也可以从研究设计本身对效度的理念加以借鉴。从研究设计的角度看,效度可区分为内部效度和外部效度。内部效度涉及研究变量之间关系的确定程度衡量,主要反映了对研究结果解释的唯一性。也就是说,如果研究结果只有一种解释,那么研究的内部效度就高。外部效度则主要用来说明研究结果可外推的程度。即,研究结果在“脱离研究情境后”,仍然能够成立的程度^[19]。由于文献证据检索的目的是为循证领域的研究者展开系统评价和元分析提供保障,因此,对文献证据检索的内、外部效度的解析,事实上涉及了循证研究设计本身,因此,更具有理论价值和实践意义。

3 文献证据检索科学性评价的趋势

科学化、规范化的循证社会科学研究以全面、精

准的文献证据检索为基本保障。基于此,本文立足于传统的查全率和查准率指标,进一步提出或引入了饱和度、冗余度、敏感度、信度、效度等系列评价指标。这些指标提出后,本课题组已展开了一系列的实证研究工作,为循证社会科学领域的研究者展开文献证据检索质量评价提供了参照。展望循证科学的未来发展,文献证据检索的科学性评价表现出了如下几个明显趋向:

(1)检索过程趋于标准化。饱和度和敏感度等指标的提出及信度和效度等指标的应用,无疑会极大地提升文献证据检索的科学性。这些指标在循证社会科学进一步发展进程中重要的应用价值,在于促进系统评价和元分析之前文献证据检索的标准化和规范化。也就是说,如何基于上述评价指标的研究结果,发展一套规范、系统、全面的文献证据检索质量评价工具,并据此保障文献证据检索的质量,将是今后循证社会科学领域值得关注的一个重要问题。

(2)检索结果评价趋于可计量化。在本文所述的各种文献证据检索质量评价指标中,所依赖的主要是来自于测量领域的理论、方法和工具。未来,有必要对文献计量等领域的相关成果加以借鉴,使文献证据检索结果评价与文献计量相关研究最大程度地对接,如此,文献证据检索才能实现与图书情报学、文献计量学等相关领域的贯通,从而获得更加宽广的研究和应用前景。

(3)不依赖于文献数据全集的自动化评估方法。虽然从研究的角度看,上述系列指标比传统的检索质量评价指标更具可操作性,但是,这些指标仍然依赖于一个饱和的文献数据全集。由本文的描述可以看出,构建文献数据全集的过程通过人工“滚雪球”的方式完成,不仅费时费力,也与大数据时代自动化的处理趋向不相符合。为此,未来的文献证据检索质量评价需要尽可能实现自动化。在尚未实现自动化的前提下,也应尽量发展一些不依赖于文献数据全集的测度方法。

(4)检索成本效益的考虑。本文前述各项文献证据质量评价指标虽然各有侧重,各具独特性和合理性,然而,上述指标中并没有将用户检索的成本问题纳入考虑。为此,面向未来的循证社会科学研究者在对文献证据检索质量做出评价时,需要参考索引

擎等领域的最新研究趋向,将用户对文献证据检索的成本纳入考虑,以便使文献证据检索与大数据背景下的计算社会科学实现最大程度的协同发展。

4 结语

本文围绕循证研究中文献证据检索的质量评价问题,结合传统的检索评价指标,提出了一系列新的

评价指标。这些指标的提出,为后续实证研究的展开提供了前提,也为发展科学、规范的检索质量理念体系和评价工具奠定了基础。由于高质量的文献证据检索是系统评价和元分析科学性的基本保障,因此,关于文献证据检索质量评价标准的研究可被视为向循证社会科学构建其独特的学术话语体系和研究范式所迈出的第一步。

参考文献:

- [1] Gurevitch J, Koricheva J, Nakagawa S, et al. Meta-analysis and the science of research synthesis [J]. Nature, 2018, 555 (7695): 175-182.
- [2] 杨克虎,李秀霞,拜争刚.循证社会科学研究方法[M].兰州:兰州大学出版社,2018.
- [3] 胡晓玲,柳春艳.循证教育学概论[M].北京:中国社会科学出版社,2021.
- [4] Cooper H, Larry V Hedges, Jeffrey C. Valentine. The handbook of research synthesis and meta-analysis[M]. Russell Sage Foundation, 2019.
- [5] Valentine J C, Biglan A, Boruch R F, et al. Replication in prevention science[J]. Prevention Science, 2011, 12(2): 103.
- [6] Open Science Collaboration. Estimating the reproducibility of psychological science[J]. Science, 2015, 349(6251): acc4716.
- [7] Pashler H, Wagenmakers E J. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? [J]. Perspectives on psychological science, 2012, 7(6): 528-530.
- [8] 张鸣明,刘鸣.循证医学的概念和起源[J].华西医学, 1998, 13(3): 265.
- [9] Lipsey M W, Wilson D B. The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis[J]. American psychologist, 1993, 48(12): 1181.
- [10] Cochran W G. The combination of estimates from different experiments[J]. Biometrics, 1954, 10(1): 101-129.
- [11] DerSimonian R, Laird N. Meta-analysis in clinical trials[J]. Controlled clinical trials, 1986, 7(3): 177-188.
- [12] Egger M, Smith G D, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test [J]. Bmj, 1997, 315 (7109): 629-634.
- [13] Higgins J P T, Thompson S G. Quantifying heterogeneity in a meta-analysis[J]. Statistics in medicine, 2002, 21(11): 1539-1558.
- [14] Chalmers Iain, Douglas G Altman. Systematic reviews[M]. London: BMJ Publishing, 1995.
- [15] Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement[J]. PLoS medicine, 2009, 6(7): e1000097.
- [16] 彭奇志.信息检索与利用[M].北京:中国轻工业出版社,2013.
- [17] 向禹,付文韬.高校数字档案馆工程理论与实践[M].长沙:中南大学出版社,2015.
- [18] 尹舒力.对查全率、查准率及引得深度等概念的商榷[J].情报理论与实践, 1996(6): 33-35.
- [19] 刘奕群,马少平,洪涛,等.搜索引擎技术基础[M].北京:清华大学出版社,2010:34.
- [20] 屈芳,马旭玲,罗林明.调查问卷的信度分析及其影响因素研究[J].继续教育, 2015, 29(1): 32-34.
- [21] 李欣,石文典.内部效度、外部效度及其关系[J].心理研究, 2009, 2(1): 9-12.

作者简介:周文杰,男,西北师范大学商学院教授;赵悦言,女,西北师范大学商学院硕士研究生;魏志鹏,男,兰州大学基础医学院循证医学中心、兰州大学循证社会科学研究中心博士研究生;杨克虎,男,兰州大学基础医学院循证医学中心、兰州大学循证社会科学研究中心教授。