张力为,彭凡.体育科学如何应对可重复性危机?[J].体育学研究,2021,35(6):1-11.

专家荐语

科学研究的目的是尽可能准确建立事实,而判断是否是科学事实的标准之一就是可重复性。事实上,现代科学正在遭遇可重复性危机这样的棘手问题,在社会与行为科学领域尤为严重,大量经典假说和研究无法得到重复性验证。然而,体育科学领域对可重复性危机问题还没有给予足够关注,更没有像心理学领域那样开展大规模的重复实验。对于科学界而言,可重复性危机既是挑战又是机遇。文中系统地介绍了可重复性危机的由来、产生原因及应对方法,既有作者多年来在方法学上的思考与实践,也汇集了国内外对可重复性危机的最新认识。这种从学术机构引导和学者个人努力两个层面提出的解决之道,必将有助于体育科学研究质量不断提升。

——石岩,山西大学,教授

体育科学如何应对可重复性危机?

张力为,彭 凡
(北京体育大学 心理学院,北京 100084)

【摘 要】:可重复性是科学的重要特征。这一话题在最近10年成为科学界的热点问题,被称作可重复性危机,改变着某些学科的研究格局,成为当代科学的重大主题和重大挑战,引发了众多科学家的关注。体育科学领域也面临着可重复性危机的挑战,但此问题尚未引起体育科学研究者的足够重视。本文介绍了可重复性危机的由来,讨论了诱发可重复性危机的原因,包括科学理念原因,统计理解原因,以及研究实践原因。本文还从学术机构的引导和学者的个人努力两方面讨论了提高可重复性的方法,包括学术机构的方向引导,学术期刊的发表政策,提前注册的逐渐推广,关注元分析研究,做好样本量规划,重视并报告效果量和置信区间,以及开展多国多实验室的合作。笔者认为,关注并积极应对可重复性危机,有助于提高体育科学研究者的基本科学素养,并进而提高体育科学实证研究的成果质量和成果积累。

【关键词】:可重复性;科学道德;科学素养;元分析;假设检验;效果量;置信区间
【中图分类号】:G80-05 【文献标识码】:A 【文章编号】:2096-5656(2021)06-0001-11
DOI:10.15877/j.cnki.nsic.20211011.001

科学具有可控制性,可操作性,可证伪性,可重复性,可争辩性[1]。这些特征使其与艺术、宗教明显区分开来。经过重复检验的成果才能得到科学界的认可,教育界的传承,并指导对自然和社会的认识与改造。可重复性是科学的重要特征,这早已成为科学家的共识。但这一看似老生常谈的话题在最近10年成为科学界的热点问题,被称作可重复性危机[2-5],引发了许多科学家的关注。Nature最近发表的一份对1 576名科研人员所做的在线网络调查报告显示[6],在回答“是否存在可重复性危机”的问题时,52%的人认为存在严重危机,38%的人认为存在轻度危机,7%的人表示不知道,3%的人表示不存在危机。这一结果提示,可重复性危机正在成为当代科学的重大主题和重大挑战[3-4]。遗憾的是,可重复性危机对于大多数体育科学研究者而言还是一个相对陌生、未受重视的学术议题。笔者预测,它迟早(如果不是很快)会成为体育科学的重大议题,因为体育科学不能一直游离在科学大家庭之外。鉴于此,本文介绍可重复性危机的由来,分析可重复性危

收稿日期:2021-10-03
作者简介:张力为(1956—),男,四川成都人,博士,教授,博士生导师,研究方向:运动心理学。

机的原因，并讨论体育科学应采取哪些方法应对这一危机，以提高研究水平，积累研究成果，使体育科学成为真正意义上的科学。

1 可重复性危机的由来

1.1 关于“重复”的概念

周红霞[3]指出，可重复性（reproducibility）和可复制性（replicability）经常被混用，但实际上二者存在明显不同。美国国家科学基金会（NSF）社会、行为和经济（SBE）分部可复制科学小组将可重复性（reproducibility）定义为研究人员复制已有实验结果的能力，特别强调，需要使用原始科研人员的数据，遵循原始实验的过程。可复制性（replicability）指研究人员复制已有实验结果的能力，但是只需要遵循原始实验过程，数据需要重新收集。

Crandall、Sherman[7]则将重复研究分为两类，即精确重复（exact or direct replication）和概念重复（conceptual replications）。精确重复是指使用一种尽可能接近原研究的方式进行一项研究，即使用与原研究同样的材料、同样的操控、同样的因变量和同种类的被试。概念重复是指检测原始研究背后相同的基本观点或者假设，但是研究类型、研究设计、对自变量的操纵、对因变量的测量以及被试群体都可以与原研究显著不同。他们认为，对于科学进步而言，概念重复研究更具价值。

由此，我们或可将科学研究中的重复研究分为3级：第一级，精确复制，使用原始科研人员的数据，遵循原始实验的过程，不重新收集数据；第二级，精确重复，即使用与原研究同样的材料、同样的操控、同样的因变量和同种类的被试，重新收集数据；第三级，概念重复，即依据原始研究背后相同的基本观点或者假设，但是研究类型、研究设计、对自变量操纵、对因变量的测量以及被试群体都可以与原研究显著不同，重新收集数据。任思腾[2]则在更为宏观的意义上讨论了“重复”的概念，将科学议题中的重复概念分为3类并做了简明清晰的界定：重复实验（replication）属于实验活动，可重复性（replicability as a feature）是用于评价实验的一个特征或指标，可重复原则（replicability as a norm）是约束科学家行为的社会规范。

本文讨论的议题将分别涉及任思腾讨论的重复概念的3个不同方面，但聚焦在可重复性上。评价前人实验结果或研究结果是否得到重复的指标往往会包括统计结果的正负方向、显著性和效果量等3个方面。

1.2 可重复性危机的产生

尽管可重复性是科学界公认的科学特征，但这一“老生常谈”变成热议话题起源于近来科学界的重要学术期刊连续发文，对已有研究成果提出严重质疑。

21世纪初，一些学者对阳性结果提出质疑。Ioannidis[8]指出，大部分已发表论文可能是假阳性结果。之后，可重复性的问题成为多学科关注的焦点问题之一[9]。2010年，一些重要学术期刊开始高度警觉。Science曾用较大篇幅对科研成果的可重复性问题进行了深入探讨，结果显示，许多高端权威期刊中发表的研究结果均存在难以重复的问题[10]。Nature Reviews Drug Discovery曾发表过一项全球性调查研究，结果显示，衡量药品疗效的二期临床试验成功率在5年间由28%降至18%，其中，前期结果欠缺可重复性是导致二期试验成功率下降的一个重要因素[11-12]。

在心理学领域，2015年，多国研究者建立了一个开放性合作实验室（The Open Science Collaboration，OSC），在他们尝试重复的100项不同领域的心理学研究中，只有39%的研究得到了明确的可重复的结果[13]，并将这一结果发表在了*Science*上。同年，*Science*和*Nature*均对此发表了评论，提出可重复性危机可能是包括心理学领域在内的行为科学，甚至行为科学之外的领域（诸如医学、生物学等）共同面临的挑战。Ioannidis认为，在开展这一可重复性检验研究的过程中，由于重复者和原作者保持着密切联络，所有实验设置都在尽最大努力去提高可重复率，所以，实际的可重复率可能比39%还要低[14-15]。随后，*Science*继续发表了不同学者对此问题的讨论。少量学者针对OSC关于可重复性检验研究的结果仍保持乐观，认为是这一报告中的统计错误造成了这些研究的可重复率被低估[16]。但是，这一评论很快收到了来自9个国家35名研究者联合发表的不赞同回应，认为Gilbert等人过度乐观的评论是由于其分析中存在统计学的概念性错误，同时，有通过选择性报告进行推论的嫌疑[17]。

Camerer等[18]对实验经济学领域的18项研究进行重复,其中,只有11项研究发现了方向相同的显著性结果,平均效果量为原研究的66%。在癌症研究领域,两大科学机构Science Exchange和Center for Open Science精选并尝试重复发表于2010—2012年的10项癌症领域较高水平的研究,其中,只有6项研究得到了相同方向的显著性结果[19]。

Camerer等[20]最近又选取了2010—2015发表于Nature和Science上的21项研究尝试进行重复,每一项重复研究均保证足够的统计功效,并选取了原研究5倍样本量的参与者。结果显示,13篇(62%)研究发现了与原研究方向相同的显著效应,但效果量只有原研究的一半,假阳性结果和效果量虚高为可重复性低的两大重要原因。

上述这些在不同年代不同领域对可重复性问题有理有据的质疑引发了科学界对该问题的关注,加上一些学术不端行为的发酵(如美国杜克大学波蒂事件,韩国首尔大学黄禹锡事件),也引发了中国学者的关注和评论[2-5,21-24]。

2 可重复性危机的原因

2.1 科学理念原因

可重复性危机的科学理念原因是,长期以来学术界更多强调和认可的是创新性探索,不太重视重复性研究。

无论是对硕士论文、博士论文开题报告的评价还是对期刊投稿论文的审阅,创新性常常成为首要的评价指标。那些超出生活常识或专业常识的研究成果特别容易吸引读者关注,也更容易因其新异性而得到发表。如生活常识告诉我们,穿竖条的衣服显得瘦,但如果一项研究发现穿横条的衣服显得瘦,这种反常识的研究结果会更容易得到编辑的重视而得以发表,也更容易得到读者的青睐而津津乐道。但对这样的研究成果进行重复性检验,哪怕是实验做得很认真,很细致,很可靠,也只能因循前人思路而在创新性上得到低分,不会受到应有的重视。对2000—2002年在精神病领域发表的83篇高引用率文献进行分析,结果表明,其中40篇文献未得到任何的重复[21,25],可重复性仅为51.81%。对心理学文献的分析表明,重复研究仅占全部文献量的1.07%[26]。这种显性或隐性的学术诱导使得许多学生或研究者产生了刊发就相信、所发即所信的倾向。另外,也有一些研究,虽然在本质上是重复性研究,但却被作者表述为创新性研究。我们在硕士论文、博士论文的开题报告和研究项目申报书中最经常读到的一句话就是“前人对此没有研究”,申报人以此来佐证和提升研究的创新性。如前人曾邀请篮球运动员做被试,发现高手与新手有如下注意特征差别:高手的决策准确性高于新手,决策速度快于新手,注视目标的次数少于新手,但注视点固定在目标上的持续时间长于新手,静眼期(视觉固定在目标后到作出动作反应之间的时间,此时间可用于加工环境信息和筹划运动决策)长于新手。那么,如果一项博士论文的开题报告拟沿此思路采用这些指标来探讨排球运动员高手和新手的注意特征差别,这在本质是运动水平与注意特征关系的跨样本效度研究或概念性重复检验,但更容易被开题报告人描述成“在排球专项上前人对此没有研究”,并以此暗示研究的创新价值。其实,这项研究可能真的有价值,但不是开题报告人描述的“前人对此没有研究”的创新价值,而是在概念性重复检验过程中发现运动项目(同场对抗和隔网对抗)是调节变量的价值。

2.2 统计理解原因

可重复性危机的统计理解原因包括研究者对虚无假设检验的过度依赖、错误理解以及对样本量问题的忽视。

过度依赖虚无假设检验的表现之一就是,目前体育科学类期刊发表的实证研究所报告的推论性统计指标,多限于统计显著性,而对于效果量和置信区间的报告却还未形成传统。虚无假设检验达到了统计显著性标准并支持了研究假设的研究,发表的可能性更大,也是不争的事实[27-28]。这导致许多研究生和研究者将$p<0.05$视为生死线[1],继而导致后续的可重复性问题,即胡传鹏等[21]指出的,“重复失败无外乎两个原因:要么原研究是假阳性,要么重复研究是假阴性。近年的实证研究表明,心理学研究的失败重复,很大程度上是由于原研究的假阳性过高。”[21]

许多研究者还存在对虚无假设显著性检验(Null Hypothesis Significance Test, NHST)的错误理解。调查表明[29-31],当向调查者询问关于p值所代表的含义时,即使是心理统计学方向的老师,也只有

20%的人能够对关于p值的6个论断全部进行正确判断。本文第一作者在2019年12月19日对国内一所体育院校106名博士生的调查表明，在《体育科学研究方法》课程专门讲授了I型错误、II型错误以及检验功效之后一周，学生对p值的理解仍错误多多。只有0.9%的人（1人）能够对关于p值的6个论断全部进行正确的判断（图1）。

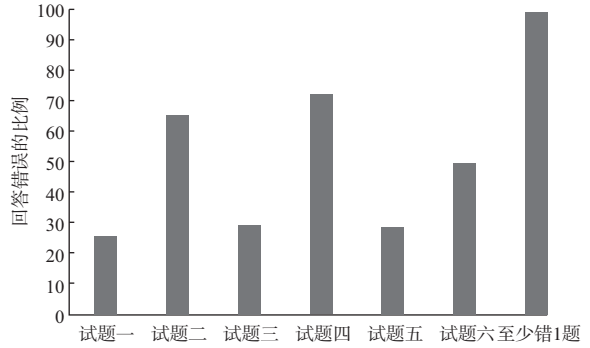

图1　国内某体育院校106名博士生在《体育科学研究方法》课程专门讲授了I型错误、II型错误以及检验功效之后1周（2019年12月19日），对p值理解错误的比例

Fig.1　The percentage of 106 Ph.D. students from a Chinese sport university misunderstanding p value one week after the course "Research Methods in Sport Science" that aims to introduce Type I error, Type II error, and Statistical Power

注：假定p=0.01情况下，参与者对6个关于p值的论断理解情况。6个关于p值的论断为（选项为正确/错误）：试题一：你完全证否了零假设；试题二：你发现了零假设为真的概率；试题三：你完全证明实验假设；试题四：你可推断出实验假设为真的概率；试题五：你可以得知，你拒绝零假设时犯错的概率；试题六：如果重复多次实验，99%实验结果显著。

样本量是否适宜对研究的可重复性具有重要影响。这些影响主要表现在两个方面：第一，样本量过小可能导致假阳性结果，即Ⅰ型错误的概率升高[32]，这一影响是众多研究中的显著结果无法得到重复的直接体现。当研究中预期的I型错误和Ⅱ型错误的概率确定时，便可以依据统计学公式计算出犯两类错误的可能性不超过既定概率所需的最小样本量，而样本量不足会提升犯两类错误的概率，加上发表偏倚对阳性结果的偏爱，研究者更容易非随机抽样（如随意或方便抽样）或选择性分析数据，从而报告出本不存在的效应。第二，被发表的小样本研究论文有更大可能高估了该效应的效果量，这一影响是小样本量降低研究可重复性的另一表现。受发表偏倚的影响，高效果量的小样本研究更容易被发表，这些研究的效果量往往比大样本量研究的效果量更大。因此，小样本导致的随机性会高估效果量，而这种情况正是大样本研究无法复制原小样本研究中大效果量的原因。再进一步，若研究者根据先前被高估的效果量来规划即将开展的研究所需的样本量，也会致使开展相应大小样本量的研究无法达到预期统计功效，降低研究的可重复性[32]。

2.3　研究实践原因

在一项重复性研究中，如果被试的选择、实验的操控、数据的处理等条件与前人研究存在些微差异，都可能导致无法得到相同或相似的结果。但本文关注的可重复性危机的研究实践原因是指研究者在实际研究过程中的可疑研究操作（questionable research practices）[33-34]。

可疑研究操作最重要的表现是"研究者在研究过程中，采用不合理的手段来达到统计上的显著（即通常所说的$p<0.05$），也称为p值操纵"（p–hacking）[28,33,35]。主要的可疑操作包括条件性选择样本量、选择性报告数据或采用多个小样本研究而避免进行一个大样本实验。采用多个小样本、低统计效力的研究，然后选择其中的阳性结果进行报告，而不是进行一个大样本量、统计效力高的研究，这种做法同样会造成效应量的"通胀"，从而降低研究的可重复性[21-36]。这是因为，受发表偏倚的影响，大效果量的小样本研究更容易被发表，这些研究的效果量往往比大样本量研究的效果量更大。也就是说，小样本导致的随机性会高估效果量，而这种情况正是大样本研究无法复制原小样本研究中大效果量的重要原因。

胡传鹏等[21]曾以研究过程为主线，简要清晰地列举了可疑研究操作的种类，值得体育科学研究者的关注。在研究开始前，可疑研究操作包括不公开研究假设，不根据统计功效确定样本量，不确定停止收集数据的规则，不确定数据分析方案，不确定排除或删除数据的规则。在收集数据的阶段，可疑研究操作包括根据结果的显著性确定是否停止收集数据，实施多个小样本的研究而非单个大样本的研究。在数据分析阶段，可疑研究操作包括根据结果的显著性使用分析策略，排除某些对结论不利的变量或数据；如果得到具有统计显著性的结果，则选择性地报告自变量和因变量，不报告效果量和效果量的置信区间；如果得到不具有统计显著性的结果，则不发表也不公开结果。最后，在研究完成后，可疑研究操作包括不进行重复检验，不鼓励其他人重复自

己的研究,不开放实验材料和分析代码,也不开放原始数据。

3 提高可重复性的方法

体育科学涵盖自然科学、社会科学和人文科学等3大领域,是非常繁复庞杂的学科体系,也因此包容着思路迥异的方法论视角、倾向和原则,以及差异很大的研究范式和采集数据进行分析的方法。不是所有体育学科都面临着可重复性的挑战。本文所讨论的提高可重复性的方法,只适用于需要通过实验、调查等实证研究手段采集数据进行分析的部分学科领域。由此,本文认为应对可重复性危机、解决可重复性问题可从学术机构和学者个人两方面入手。

3.1 学术机构的努力

3.1.1 学术机构的方向引导

各级各类体育科学学会可以通过专题研讨引发对可重复性问题的重视,也可以通过发表政策声明阐明对本学科领域可重复性问题的基本立场。

体育科学研究方法类课程应加强对可重复性这一科学特征的讨论,包括可重复性的概念以及提高可重复性的方法。这种讨论既可以在课程一开始讨论科学特征时进行,也可以在后续的研究设计、统计分析等章节中反复提及。体育科学研究方法课程还应将科研伦理作为大纲内容。目前,绝大多数体育科学研究方法教科书还未将研究伦理作为单章列出。这种明显缺失使得许多学生在本科阶段的教育中,就不了解研究伦理的重要、研究伦理与科学性质的关联、研究伦理的基本原则与具体原则,继而,当他们成为体育科学研究者时,也就难以在体育科学研究过程中严格遵守伦理规范和伦理原则。研究伦理体现了科学精神与人文精神的结合,体现了科学研究的终极目标,体现了研究者的基本学术素养。体育统计类课程在介绍推论性统计时,除了应介绍显著性检验的核心思想,还应说明显著性检验的局限,同时介绍效果量、置信区间以及统计功效的核心思想、计算方法和报告形式。

体育院校还可以通过对研究生毕业论文的中期进展检查和毕业论文终稿的抽查,检验是否存在可疑操作或学术不端行为。

3.1.2 学术期刊的发表政策

各类体育科学期刊是体育科学研究的重要引领平台,应通过主编致词、刊物声明、编辑部指南、投稿人自检报告模板等形式鼓励有助于提高可重复性的措施,包括鼓励提前注册研究,鼓励具有递进性和连续性的多项研究,要求承诺数据公开,要求研究方法细节的报告(尤其对于背景敏感性高的实验,更需要详尽报告实验环境、实验程序、实验控制的细节)等;也可通过专栏形式发表重复研究,引导研究者重视可重复性问题,并以此为导向,向夯实体育科学实证研究的基础这个方向迈进。

体育科学类期刊应该通过政策性的引导,向公开、透明、开放的方向迈进,降低研究者的自由度,减少研究中的可疑操作,从而提高期刊论文的学术质量。有学者呼吁,应当通过多种指标评价学术期刊质量[37]。影响因子至多不过是期刊质量的一个指标而已,而TOP指标完全可以成为期刊质量的补充性指标。我们希望看到能有体育科学类期刊率先借鉴促进科研公开与透明委员会(Transparency and Openness Promotion Committee, TOP)倡导的8标准3层级的TOP准则(表1)[3,38],以此来衡量和提高科学研究的透明度、开放性和可重复性[38],并引导作者和读者向此方向迈进。

表1 TOP准则的8个标准和3种层级

Tab.1 Summary of the Eight Standards and Three Levels of the TOP Guidelines

	层级 0	层级 1	层级 2	层级 3
引用标准	期刊鼓励数据、代码和材料引用,或者不作要求	在投稿指南中,期刊用明确的规则和清晰的范例描述数据引用	与期刊的投稿指南保持一致,要求文章提供数据和材料的适当引用	遵循期刊的作者指南,对数据和材料适当引用,否则不予发表
数据透明	期刊鼓励数据共享,或者不作要求	文章需要表明数据是否可用,如可用,在哪里能够访问	数据必须发布到一个可信的存储库,例外情况在文章提交时必须注明	数据必须发到一个可信的存储库,并且在文章发表前能重新独立生成报告中的分析结果
分析方法(代码)透明	期刊鼓励代码共享,或者不作要求	文章需要表明代码是否可用,如可用,在哪里能够访问	代码必须发布到一个可信的存储库,例外情况在文章提交时必须注明	代码必须发到一个可信的存储库,并且在文章发表前能重新独立生成报告中的分析结果

（续表1）

	层级0	层级1	层级2	层级3
研究材料透明	期刊鼓励材料共享，或者不作要求	文章需要表明材料是否可用，如可用，在哪里能够访问	材料必须发布到一个可信的存储库，例外情况在文章提交时必须注明	材料必须发到一个可信的存储库，并且在文章发表前能重新独立生成报告中的分析结果
设计和分析透明	期刊鼓励设计和分析透明，或者不作要求	期刊阐述设计透明标准	为审查和出版，期刊要求作者遵守设计透明标准	为审查和出版，期刊要求并强制作者遵守设计透明标准
研究预注册	期刊不作要求	期刊鼓励研究预注册，并在文章中提供预注册链接（如果存在预注册）	期刊鼓励研究预注册，在文章中提供链接，并提供满足预注册徽章要求的证明	期刊要求研究预注册，并在文章中提供链接和徽章
分析计划预注册	期刊不作要求	期刊鼓励采用事前分析计划，并在文章中提供注册分析计划的链接（如果存在注册分析计划）	期刊鼓励采用事前分析计划，在文章中提供链接，并提供满足注册分析计划徽章要求的证明	期刊要求采用带有分析计划的研究预注册，并在文章中提供链接和徽章
可复制	期刊不鼓励提交复制研究，或者不作要求	期刊鼓励提交复制研究	期刊鼓励提交复制研究，并对结果进行盲审	在获得研究结果之前，期刊使用“注册报告”作为复制研究的提交选项，对之进行同行评议

注：原文来自Nosek等人的研究成果，译文引自周红霞的研究成果。

自2015年TOP准则发布以来，已有5 000多种期刊开始执行TOP准则，所有主要学术出版商都签约该准则[3]。中国的一些学术期刊也在向此方向迈进，如为鼓励符合研究伦理和研究规范的研究实践，减少可疑操作，《心理学报》就要求投稿人在自检报告中明确回答以下问题：

“以治疗疾病为目的临床实验，建议在收集数据前预先备案（pre-register）；也鼓励其他实验研究预先备案。

“是否报告并分析了效果量？请写出计划的样本量，实际的样本量。如果二者有差别，请写出理由。

“为保证论文中数据报告的完备性，统计分析中如果剔除了部分数据，是否在文中报告？原因是什么？包含这部分数据统计结果如何变化？统计分析中是如何处理缺失数据的？使用量表时是否删除了其中的个别题目？原因是什么？如果包含这部分题目统计结果会如何变化？是否有测量的项目或者变量没有报告，原因是什么？

“研究用到的未经过同行评议和审查的实验材料、量表或问卷，是否附在文件的末尾以供审查？如果没有，请写出理由。如果该文发表，您是否愿意公开此程序与其他研究者共享？

“您的研究的原始数据是否可以上传备审稿人和编辑核查用？如果不能，请写出理由。如果该文发表，您是否愿意公开数据供其他研究者共享？”

当然，在推行TOP准则的过程中，不可能一刀切地在8个方面均立刻采用层级3的最严格标准。不同学科领域、不同发展水平的期刊可以根据具体情况选择执行TOP的时机和制订执行TOP的方案，但朝着TOP的方向尽快做出实质性的改变则是必需的，这是开展科学研究本身的需求，也是学术期刊应对竞争的需求。

3.1.3 提前注册的逐渐推广

提前注册指在研究开始前在学术研究网站提前注册即将执行的研究方案，包括研究假设，研究设计，样本规划，统计分析等，这是科学的开放性的具体体现。国内外一些学术期刊已开始明确鼓励学者提前注册。

提前注册的一种常见形式是预先备案网站的预注册。《心理学报》在自检报告中就要求投稿人回答以下问题：“您是否在本期刊的预先备案网站（https://osf.io/）上进行了预注册？如果您的研究有预先备案，会显著增加被录用的机会。”Yun等[39]最近开展了一项系列研究，试图回答“仪式化行为对运动员真的有帮助吗”的问题。在开展正式实验前，该研究者对即将开展的研究的预期假设、研究设计和变量设置以及统计分析策略等内容进行了提前注

册(https://doi.org/10.17605/OSF.IO/HMCG2),并在正式投稿的自检报告中以及论文的前言部分做出交代。预先备案网站预注册的特点是,在论文正式发表之后,读者可随时前往预先备案网站查看作者的预注册信息及后期添加的实验程序代码、实验数据、图表等内容。提前注册的另一种形式是注册式报告(Registered Reports,RRs)。这是当前一些学术期刊新推出的一种投稿类别。相比于传统的全文投稿,注册式报告投稿由两个阶段构成:①在进行正式实验之前,将研究的文献评述、问题提出、预期假设、研究设计以及统计分析策略等内容写成报告。期刊对这一报告进行初步审核与同行评审,并给出同行内接收(In-Principle Acceptance)、需要修改或拒稿的反馈。②在获得(或修改后获得)同行内接收的反馈后,作者按照注册式报告内描述的研究计划在一年内开展研究,完成研究报告全文并投出。只要作者严格执行了注册式报告中的研究计划,无论研究结果如何,期刊不会因为研究结果不符合假设而将其拒绝。研究表明,相比传统的全文投稿,注册式报告投稿在没有降低创新性的前提之下,在研究方法和统计分析的准确度及研究的总体质量等19个方面具有显著的提升[40]。

无论是预先备案网站的预注册,还是注册式报告投稿,其内容均包括:①研究信息:研究假设;②研究设计计划:研究类型、盲法、研究具体设计、随机化方法等;③样本规划:样本量确定依据、停止招募的规则等;④变量:操纵变量(自变量)、测量变量(因变量)及指标等;⑤分析计划:统计模型、推断标准、纳入排除标准、缺失值及额外分析等;⑥其他。此外,注册式报告投稿的第一阶段,即注册式报告中,还需完整撰写文献评述和问题提出部分,且作者不可在第二阶段的全文投稿中对其再进行修改。

提前公开自己的研究方案有两个优点[21],第一,可以防止研究者根据结果来修改自己的研究假设,将探索性的研究写成验证性研究[41-42];第二,可以减少研究者实验操作以及数据分析方面的自由度,公开报告主要研究结果,减少发表偏倚带来的问题[42]。提前注册意味着邀请所有感兴趣的人在研究方案执行过程中进行监督,防止研究者向有利于自己的方向修改研究计划和选择性地报告有利于自己的研究结果。

3.2 学者个人的努力

3.2.1 关注元分析研究

元分析也称荟萃分析,是通过统计方法对相似研究进行量化评价并得出综合性结论的方法。元分析有3个特点,第一,是以统计分析为基础的量化分析,因此,可被重复;第二,是对已有研究文献的客观评价,明显区别于主观评价,因此,可被重复;第三,注重对影响自变量与因变量关系的调节变量的探索,因此,有助于后续研究的具体推进。在积累科学证据的过程中,元分析被认为是获得最高层级证据的方法[43-44],因此,在医学、社会科学、教育科学、心理科学等领域得到普遍应用。最近30年,许多学科领域发表的元分析文献数量呈指数增长,甚至有研究者呼吁在科学研究领域应更多依靠“元分析思维”[45-46]。元分析在体育科学领域的应用也逐年增加,项明强等[47]最近撰文分析了体育科学领域元分析论文的质量问题。近年来,一些研究者还将元分析作为寻找研究切入点的方法或系列研究的总结。

Yun等[48]在“仪式化行为提升自我控制的心理机制”的研究中,开展正式实验前曾采用元分析对仪式化行为的不同心理效益进行定量评价,根据总体效应量和调节变量的相关结论,指导后续的系列实验研究。元分析发现,首先,仪式化行为对心理效益的影响达到中等效果量(d=0.53),该数值将作为后续实验设计中被试量选择(G*Power)的重要参数值。其次,对设置压力场景的研究进行亚组分析发现,压力条件起调节作用,高压情境下,仪式化行为提升心理效益的效应量更大。这一结论提示后续研究应设置高压力情境,以使仪式化行为发挥最大效能。最后,对任务指标的亚组分析发现,认知指标为小效应量,较不敏感,正式实验中应采用行为、生理等多项指标探讨后续研究的主题。

李丹阳等[49]在“自然环境改善认知和运动任务中的抑制性与坚持性自我控制”的研究中,在完成系列研究的4项实验之后,采用小型元分析(mini meta-analysis)对其结果进行定量评价。元分析发现,相比于简单休息,观看自然环境图片对个体在认知和运动任务中的抑制性与坚持性自我控制表现具有大效果量的影响(d=0.99)。这一发现表明,安全温和的自然环境具有强健的复愈能力,可以为恢复自我控制和改善操作表现提供便捷省力的干预方

式。这种小型元分析通常用于系列研究的内部元分析，可以帮助研究者对自己的系列研究进行更加精确的效果量估计[27]，其结果通常比单个实验的研究结果更具有说服力[50]，并且可以为研究问题的可重复性提供更有力的证据[51]。

总体而言，元分析所提供的定量评价和综合结论不但比单个研究更加可靠，而且让读者更容易在短时间内获取本领域的大量研究成果。因此，它往往成为研究者掌握本领域研究进展的快速通道和重要通道。

3.2.2 做好样本量规划

对样本量的规划是研究设计阶段重要的组成部分，不同的研究类型（如实验还是调查）、研究问题（如探索性问题还是证实性问题）、研究设计（如组间设计、组内设计还是混合设计）、被试群体（如人的实验还是动物实验以及被试的异质性）及数据处理方法（如多元回归、典型相关还是结构公式模型）对样本量要求的影响存在明显差异。研究者需要综合考虑以上诸多因素来最终确定样本量。

在实验研究中，从统计角度出发，样本量取决于显著性（α值）、统计功效和效果量。通常设定的显著性（α值）为0.05，统计功效为0.8（更好的标准是0.9或0.95）；尽管效果量的默认值为中等效果量，可作为没有特别依据时设定效果量的参考，但如果有前人实证研究或元分析的结果，则可据此通过G*Power[52]等软件计算得出。如关于自我控制损耗效应（先前的自我控制耗费能量，导致后续的自我控制变差），已有元分析的平均效果量从d=0.04，95% CI[-0.07, 0.15] [53] 到 d=0.16，95% CI[0.05, 0.26] [54]不等。对于研究者后续研究的样本量规划而言，这些结果既提供了效果量的统计依据，也提出了需要仔细分析小效果量背后的专业性问题。因此，样本量的规划不仅仅是研究领域的统计知识问题，更是专业领域的学术判断问题（与小效果量对应的120个试次的Stroop实验范式和实验操作）。

带有中介效应的实验研究，或者带有中介效应的调查研究，研究设计更为复杂，样本量的计算也会略微复杂一些。影响中介效应样本量的主要因素包括效果量（如路径系数的值）、统计功效、测量模型的特征（测量指标的数量和信度）、结构模型的特征（潜变量的个数，路径系数的值和数量）等。此类研究的样本量，可通过pwrSEM[55]（网址yilinandrewang.shinyapps.io/pwrSEM/）或Mplus[56]等软件计算得出。使用pwrSEM对带有潜变量的结构公式模型所需样本量进行计算，需分4步进行：界定模型，将所界定的模型可视化以检查是否符合自己原初的理论构想，设置测量模型和结构模型的参数值，估计统计功效直到统计功效达到预设标准（如0.80、0.90或0.95），如此，即可得到此种设定条件下的样本量。

3.2.3 重视并报告效果量和置信区间

前文曾特别提到，研究者对虚无假设检验的过度依赖和错误理解是产生可重复性危机的重要原因。对此问题的解决方案之一是加强对其他推论性统计指标的重视。

研究者对变量关系的数据进行统计时，如方差分析，回归分析，有3类问题需要关注，即变量关系的方向，变量关系的强度，以及变量关系的可靠程度。变量关系的方向可从描述性统计获得，或者从推论性统计中的正负号获得，变量关系的强度以效果量为指标，而变量关系的可靠程度因为多以显著性为指标而造成巨大的误解和误用，受到许多学者的批评。根据美国统计学会的界定，p值可以表明数据与某个特定统计模型之间不相容的程度，但p值本身并不能衡量模型或假设的可信度。在报告推论性统计结果时，显著性是个非黑即白的二分决定（$p<0.05$或$p>0.05$），很大程度上受研究设计和样本量影响，在重复性研究中波动大，不稳定[57]。相比而言，效果量及其置信区间可能是更合适的推论性统计指标。单个研究中的效果量是以样本数据为基础对总体效果量的点估计，效果量的置信区间则是对总体效果量的区间估计[21]。效果量及其置信区间无论从稳定性[58]还是就易于理解[59]的角度而言，均优于显著性；同时，效果量及其置信区间还是元分析的基础，对学科发展过程中的知识积累极其重要[21]。甚至有学者主张，可以用效果量及其置信区间来覆盖或取代显著性，并通过具体实例给出了如何进行操作的方法[57, 60]。他们将这种思路称为新统计，也可称作估计统计。估计统计的核心问题是“有多大效果?”，而回答这一核心问题的统计方法就是效果量及其置信区间的计算。

目前，在体育科学领域，报告效果量和置信区间的做法还不普遍，这不利于读者正确、全面地理解研

究结果，不利于研究成果的积累，也不利于解决可重复性问题。因此，笔者呼吁，体育科学领域的研究者在报告显著性时，应注意到该指标的明显局限性，同时重视并报告效果量和置信区间。

3.2.4 开展多国多实验室的合作

开展多国实验室合作，对重要的学术问题或有争议的学术问题开展研究，是应对可重复性危机的有效方法。如北京体育大学的一个研究团队最近参加了一项全世界不同地区12个实验室（含1 775名被试）的联合实验研究[61]，以回应是否存在自我损耗效应这个颇具争议的问题。所谓自我损耗效应，是指个体完成先前的自我控制任务时，会不断消耗自己的自我控制能量，从而导致完成后续的自我控制任务时效果下降[62]。自Baumeister等提出自我损耗效应之后，该问题一度成为社会心理学的热点话题，引发了众多研究，相关论文达5 000多篇。支持这一观点并发现这一效应的研究众多，但不支持的研究给出的理据难驳，包括元分析论文[63]和23个实验室（含2 141名被试）的大型重复性检验[53]，都得出了不支持的结果。面对这一争议，Dang等[61]再次发起了一项大型重复性研究。该项联合实验研究是提前注册研究，采用前人常用的自我损耗双任务范式，以Stroop任务为自我损耗诱发任务，反向眼跳任务为自我损耗检测任务，结果发现效果量d=0.16。笔者据此认为，的确存在自我损耗效应，但效果量较小。

从上述实例可见，如果在不同条件下（不同的地域、实验室、主试、被试、语言背景等），采用统一的实验设备、实验材料、实验程序、统计方法开展实验，如果观察到稳定的自变量与因变量之间的关系，则能更有把握地确定这种关系跨地域、跨实验室、跨主试、跨被试、跨语言背景的一致性和稳定性；如果没有观察到稳定的自变量与因变量之间的关系，则可以进一步溯源至地域、实验室、主试、被试、语言背景等条件。显然，多国实验室合作所提供的研究证据，包括重复性研究结果的证据，要比单一实验室提供的研究证据有力得多。

本文从可重复性危机的由来、可重复性危机的原因和提高可重复性的方法等3个方面讨论了可重复性危机对体育科学发展的现实意义。学术界面对可重复性危机采取的积极应对毫无疑问可以极大地提高学术研究质量和学术成果积累厚度。危机孕育着进步。如果体育科学研究者不趁此机会及时反省，采取行动，将会失去一次大幅度提高的机会。笔者希望本文的讨论能够引发体育科学研究者对可重复性问题的高度重视，在研究实践中采用合理的应对方法，跟上其他学科前进的步伐。

（致谢：本文撰写过程中，山西大学石岩教授、山西师范大学郑旗教授、福建师范大学方千华教授、北京体育大学周财亮副教授、北京体育大学王盼讲师对初稿提出了宝贵的意见，在此谨致以衷心的谢意！）

参考文献：

[1] 张力为.体育科学研究方法[M].北京：高等教育出版社，2002.

[2] 任思腾.科学实验中的可重复概念[J].自然辩证法通讯，2020，42(265)：50-56.

[3] 周红霞.科学研究的可重复性及其保障措施[J/OL].科学学研究：1-15[2021-11-26]. https：//doi.org/10.16192/j.cnki.1003-2053.20210531.002.

[4] 王阳，肖昆.可重复性危机与预注册新进路[J].科学学研究，2020，38(253)：14-21.

[5] 王阳，肖昆.论控制偏见的编辑制度革命——关于预注册遏制可重复性危机的机理研究[J/OL].科学学研究：1-10[2021-11-26]. https：//doi.org/10.16192/j.cnki.1003-2053.20210419.003.

[6] BAKER，M. 1，500 scientists lift the lid on reproducibility[J]. Nature，2016，533：452-454.

[7] CRANDALL C S，SHERMAN J W. On the scientific superiority of conceptual replications for scientific progress[J]. Journal of Experimental Social Psychology，2016，66：93-99.

[8] IOANNIDIS J P A. Why most published research findings are false[J]. PLOS Medicine，2005，2(8)：e124.

[9] PENG R D. Reproducible research and biostatistics[J]. Biostatistics，2009，10(3)：405-408.

[10] ENSERINK M. Final report on Stapel also blames field as a whole[J]. Science，2012，338(6112)：1270-1271.

[11] ARROWSMITH J. Phase II failures：2008–2010[J]. Nature Review Drug Discovery，2011，10：328-329.

[12] PRINZ F，Schlange T，Asadullah K. Believe it or not：How much can we rely on published data on potential drug targets ?[J]. Nature Review Drug Discovery，2011，10：712.

[13] OPEN Science Collaboration. Estimating the reproducibility of psychological science[J]. Science，2015，349(6251)：acc4716.

[14] BAKER，M.Over half of psychology studies fail reproducibility test[J/OL]. Nature，2015.

[15] BOHANNON J. Many psychology papers fail replication

test[J]. Science,2015,349(6251):910-911.
[16] GILBERT D T,KING G,PETTIGREW S,et al. Comment on "Estimating the reproducibility of psychological science" [J]. Science,2016,351(6277):1037.
[17] ANDERSON C J,BAHNÍK Š,BARNETT-COWAN M,et al. Response to Comment on "Estimating the reproducibility of psychological science"[J]. Science,2016,351(6244):1037.
[18] CAMERER C F,DREBER A,FORSELL E,et al. Evaluating replicability of laboratory experiments in economics[J]. Science,2016,351:1433-1436.
[19] ERRINGTON T M,IORNS E,GUNN W,et al. A open investigation of the reproductivity of cancer biology research[J]. eLIFE,2014,3:e04333.
[20] CAMERER C F,DREBER A,HOLZMEISTER F,et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015[J]. Nature Human Behavior,2018,2:637-644.
[21] 胡传鹏,王非,过继成思,等. 心理学研究中的可重复性问题:从危机到契机[J]. 心理科学进展,2016,24(9):1504-1518.
[22] 焦飞,王娟,李尊岭,等. 医学科研的可重复性与转化医学——从生命的复杂性谈起[J]. 医学与哲学:人文社会医学版,2012,9(33):21-23.
[23] 王炳顺. 杜克大学波蒂事件及研究的可重复性[J]. 中国医学伦理学,2013,26(6):683-686.
[24] 赵萌. 强制数据共享推进研究结果的可重复性:发表文章时你希望这样做吗?[J]. 中国组织工程研究,2016,36:5322.
[25] TAJIKA A,OGAWA Y,TAKESHIMA N,et al. Replication and contradiction of highly cited research papers in psychiatry:10-year follow-up[J]. The British Journal of Psychiatry,2015,207(4):357-362.
[26] MAKEL M C,PLUCKER J A,HEGARTY B. Replications in psychology research:How often do they really occur ?[J]. Perspectives on Psychological Science,2012,7(6):537-542.
[27] CUMMING G. The new statistics:Why and how[J]. Psychological science,2014,25(1):7-29.
[28] IOANNIDIS J P A. Why most discovered true associations are inflated[J]. Epidemiology,2008,19(5):640-648.
[29] GIGERENZER G. Mindless statistics[J]. The Journal of Socioeconomics,2004,33(5):587-606.
[30] HALLER H,KRAUSS S. Misinterpretations of significance:A problem students share with their teachers[J]. Methods of Psychological Research Online,2002,7(1):1-20.
[31] OAKES M. Statistical inference:A commentary for the social and behavioural sciences[M]. Chichester:Wiley,1986.
[32] BUTTON K S,IOANNIDIS J P A,MORKRYSZ C,et al. Power failure:Why small sample size undermines the reliability of neuroscience?[J]. Nature Reviews Neuroscience,2013,14(5):365-76.
[33] JOHN L K,LOEWENSTEIN G,PRELEC D. Measuring the prevalence of questionable research practices with incentives for truth telling[J]. Psychological Science,2012,23(5):524-532.
[34] SIMMONS J P,NELSON L P,NELSON L D,et al. False-positive psychology:undisclosed flexibility in data collection and analysis allows presenting anything as significant[J]. Psychological Science,2011,22:1359-1366.
[35] JOOBER R,SCHMITZ N,ANNABLE L,et al. Publication bias:What are the challenges,and can they be overcome?[J]. Journal of Psychiatry & Neuroscience,2012,37(3):149-152.
[36] BAKKER M,VAN DIJK A,WICHERTS J M. The rules of the game called psychological science[J]. Perspectives on Psychological Science,2012,7(6):543-554.
[37] KEPES S,BANKS G C,KEENER S K. The TOP factor:an indicator of quality to complement journal impact factor[J]. Industrial and Organizational Psychology,2020,13(3):328-333.
[38] NOSEK B A,ALTER G,BANKS G C,et al. Promoting an open research culture[J]. Science,2015,348(6242):1422-1425.
[39] YUN D T,ZHANG L W,QIU Y. Does pregame "compulsive action" really help athletes? Ritualized behavior enhances self-control[Manuscript submitted for publication]. School of Psychology,Beijing Sport University,2021.
[40] SODERBERG C K,ERRINGTON T M,SCHIAVONE S R,et al. Initial evidence of research quality of registered reports compared with the standard publishing model[J]. Nature Human Behaviour,2021,5:990-997.
[41] WAGENMAKERS E-J,WETZELS R,BORSBOOM D,et al. An agenda for purely confirmatory research[J]. Perspectives on Psychological Science,2012,7(6):632-638.
[42] KAPLAN R M,IRVIN V L. Likelihood of null effects of large NHLBI clinical trials has increased over time[J]. PLoS One,2015,10(8):e0132382.
[43] 张力为,孙国晓. 体育科学实证研究的逻辑流与证据链[J]. 体育科学,2017,37(4):3-10.
[44] SIDDAWAY A P,WOOD A M,HEDGES L V. How to Do a Systematic Review:A Best Practice Guide for Conducting and Reporting Narrative Reviews,Meta-Analyses,and Meta-Syntheses[J]. Annual Review of Psychology,2019,70(1):747-770.
[45] CUMMING G. The new statistics:why and how[J]. Psychological science,2014,25(1):7-29.
[46] GUREVITCH J,KORICHEVA J,NAKAGAWA S,et al. Meta-analysis,and the science of research synthesis[J]. Nature,2018,555(7695):175-182.
[47] 项明强,肖让,赵雪平,等. 70篇体育学元分析文献的质量评价与改进建议[J]. 上海体育学院学报,2021,45(4):1-9.
[48] YUN D T,ZHANG L W,QIU Y. Meta-Analysis on Psychological Benefits of Ritualized Behavior[Manuscript

submitted for publication]. School of Psychology, Beijing Sport University, 2021.
[49] 李丹阳, 张力为. 自然环境改善认知和运动任务中的抑制性与坚持性自我控制[J]. 中国体育科技, 2020, 56(1): 31-44.
[50] GOH J X, Hall J A, ROSENTHAL R. Mini meta-analysis of your own studies: Some arguments on why and a primer on how[J]. Social and Personality Psychology Compass, 2016, 10(10): 525-549.
[51] BRAVER S L, THOEMMES F J, ROSENTHAL R. Continuously cumulating meta-analysis and replicability[J]. Perspectives on Psychological Science, 2014, 9(3): 333-342.
[52] 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. Behavior Research Methods, 2007, 39: 175-191.
[53] HAGGER M S, CHATSISARANTIS N L D, ALBERTS, H, et al. A multilab preregistered replication of the ego-depletion effect[J]. Perspectives on Psychological Science, 2016, 11: 546-573.
[54] DANG J. Commentary: A Multilab Preregistered Replication of the Ego-Depletion Effect[J]. Frontiers in Psychology, 2016, 7: 1155.
[55] WANG Y A, RHEMTULLA M. Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial[J]. Advances in Methods and Practices in Psychological Science, 2021, 4(1): 1-17.
[56] THOEMMES F, MACKINNON D P, REISER M R. Power analysis for complex mediational designs using Monte Carlo methods[J]. Structural Equation Modeling: A Multidisciplinary Journal, 2010, 17(3), 510-534.
[57] CUMMING G. The new statistics: A how-to guide[J]. Australian Psychologist, 2013, 48: 161-170.
[58] CUMMING G, MAILLARDET R. Confidence intervals and replication: Where will the next mean fall ?[J]. Psychological Methods, 2006, 11(3): 217-227.
[59] COULSON M, HEALEY M, FIDLER F, et al. Confidence intervals permit, but do not guarantee, better inference than statistical significance testing[J]. Frontiers in Psychology, 2010, 1: 26.
[60] HOPKINS W G. Spreadsheets for analysis of validity and reliability[J]. Sportscience, 2017, 21(9): 36-44
[61] DANG J, BARKER P, BAUMERT A, et al. A multilab replication of the ego depletion effect[J]. Social Psychological and Personality Science, 2020, 12(1): 1-11.
[62] BAUMEISTER R F, BRATSLAVSKY E, MURAVEN M, et al. Ego depletion: Is the active self a limited resource ?[J]. Journal of Personality and Social Psychology, 1998, 74(5): 1252-1265.
[63] CARTER E C, KOFLER L M, FORSTER D E, et al. A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource[J]. Journal of Experimental Psychology: General, 2015, 144(4): 796-815.

作者贡献声明：

张力为：确定论文选题，撰写、修改、审核论文；彭凡：撰写、修改论文。

How Does Sport Science Cope with the Replicability Crisis?

ZHANG Liwei, PENG Fan

(*School of Psychology, Beijing Sport University, Beijing* 100084, *China*)

Abstract: Replicability is one of the defining hallmarks of science. This topic, which seems conventional to researchers, has however become a crucial issue as well as a grand challenge within the scientific community over the past decade, bringing scientists' attention to the "Replicability Crisis" which has been changing the general accepted practices in certain fields. Sport science also faces the challenges to replicability which have not yet raised the concern in this area. In this article, we discuss how the replicability crisis has evolved and how these causes have introduced challenges to replicability, including scientific concepts, statistical understandings, and research practices. This article also covers several approaches to improve the replicability of studies from both institutional and personal perspectives, including the guidance of the academic institutions, policies for publications, preregistration, emphasis on meta-analysis research, planning the sample size, reporting effect size and confidence intervals, and cooperation cross multinational laboratories. We propose that attention and positive strategies to cope with the replicability crisis will help improve researchers' basic scientific literacy in sport science, and consequently boost the quality and accumulation of empirical research in this area.

Key words: replicability; ethics of science; scientific literacy; meta analysis; hypothesis testing; effect size; confidence intervals