移动互联网时代的社会科学研究工具：众包的争议与发展

彭凯平[1,2]，刘世群[1]，倪士光[3]

（1. 清华-伯克利深圳学院，广东 深圳 518055；
2. 清华大学 心理学系，北京 100084；
3. 清华大学 深圳研究生院，广东 深圳 518055）

[摘 要] 社会科学实证研究长期忽视研究的外部效度，成为理论发展的一大隐忧。众包网站提供低成本、便利、快速和多样化的样本来源，带来了解决问题的契机。然而，众包的研究有效性遭受了样本偏差、回答者不用心、金钱动机、重复被试及互相影响等质疑。过去几年，研究者提出了一系列事实反驳各种争议，众包有效性得到了研究者的重视及重新认同。众包将成为移动互联网时代的社会科学研究工具，使用该平台收集及分析数据，将会提高国内研究在全球的可见度。

[关键词] 众包；社会科学；便利抽样；Amazon Mechanical Turk
[中图分类号] B 84 [文献标识码] A [文章编号] 1001-9162(2018)03-0113-11
[DOI] 10.16783/j.cnki.nwnus.2018.03.015

一、问题提出

研究者使用抽样方法获取数据，通过统计方法建立适用于目标总体的结论，是心理学实证研究的一般过程[1]。为了确保研究结论的总体效度(Population Validity)，选择代表性的研究样本是前提条件。然而，受限于研究成本、时间成本及可行性等因素，许多研究由于取样不足或样本偏差导致争议；科学研究长期忽视研究的外部效度，是心理知识及理论发展上的一大隐忧[2]。自 2010 年起，在社会科学研究中大量被采用的众包(Crowdsourcing)模式，为研究者提供了一个提升研究效度的契机。众包网站提供低成本、便利、快速和多样化的样本来源，提高了取样的数量、质量及代表性[3]。因此，本文目标是总结众包在社会科学研究中的应用现状，评价其工具属性的优劣，提出未来的研究方向，希望藉此带动国内众包网站及其研究的开展。

二、什么是众包

众包是发起者（requester）将一项工作(human intelligence task，HIT)在网络上分包给工作者（worker）共同完成的行为。发起者可以是企业、政府、研究人员或是网络项目执行人，工作具有明确的目标以及期限[4]。工作者基于获取金钱报酬、热情助人或打发时间等多元化动机，积极参与其中。工作者通常彼此不认识，工作通过网站进行沟通和管理。随着网络科技的发展，特别是进入移动互联网时代后，人们随时随地利用手机与世界联系，使得众包平台更具潜力[5]。

亚马逊公司（Amazon）开发的 Mechanical Turk（MTurk）[①]以简单可分割的小型分包工作为主（图 1），任务说明及执行的接口可以轻松地被使用者定制化成各种不同的功能，符合大部分社会科学问卷设计的需要[7]，因此成为社会科学研究者的代表性众包平台[8]。事实上，不仅大多数众

[收稿日期] 2018-03-29
[基金项目] 国家自然科学基金项目（31371017/31471001）
[第一作者简介] 彭凯平（1962—），男，湖南岳阳人，心理学博士，清华大学教授，博士生导师，从事文化和积极心理学研究

包研究以 MTurk 作为研究工具，而且针对众包优缺点及有效性的讨论，也多以 MTurk 作为主要的分析目标[9－10]。

在心理学研究上，众包已经被许多研究者采用。[②]截至 2017 年 3 月，以“Mechanical Turk”进行全文检索，在 PsycInfo 数据库中有 357 个结果，在 PsychArticle 中有 33 个结果，且都是 2010 年之后的研究成果；PsychArticle 的 33 篇文章中，有 28 篇直接使用了 MTurk 作为其收集样本的工具。这个现象显示 MTurk 已经成为心理研究的工具之一。例如，在决策行为的研究中，Rand，Greene 和 Nowak 利用 MTurk 招募了 1，955 名被试，发现人们在解决问题时倾向于与人合作，而非采取自利行为[15]。在认知研究中，Christine Ma-Kellams 和 Jennifer Lerner 招募了 314 名被试，发现一般人错误地以为直觉式思考比较能提高共情能力，实际上系统化思考反倒提升了共情能力[16]。在社会心理研究中，D. J. Hauser，Preston 和 Stansfield 雇用了 158 位 MTurk 的工作者，发现人会因为是否有与对方的互动而影响到是否愿意帮助别人[17]。在积极心理研究中，Diehl，Zauberman 和 Barasch 招募了 188 名被试，证实不停地拍照的确能提高当下的投入及正向体验[18]。Thomson 和 Siegel 以及 Siegel，Thomson 和 Navarro 分别招募了 485 及 1，133名被试，发现看见别人的美德的确会增加一个人的道德提升感，促使其更亲近社会也更愿意利他[19－20]。

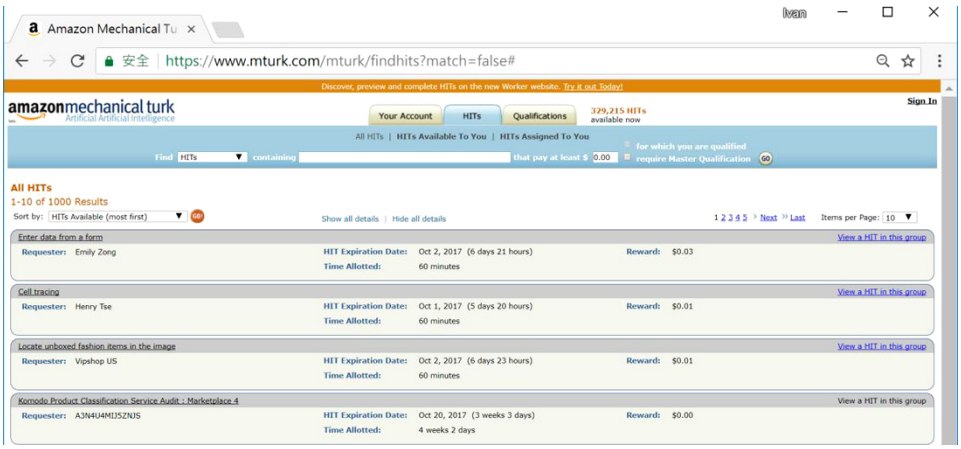

图 1　MTurk 截图

三、众包研究的优点

众包与其他常见的便利抽样方式，包括网络、校园或特定组织及社群来源相比，众包以较低成本找到更为有效或具代表性的样本，同时兼顾匿名性、质量及成本（详见表 1）。

（一）提高样本多元性

传统的社会科学研究因为过多使用大学生或是西方民主工业化国家高等教育的富有白人（WIERD）作为研究对象，导致研究成果常遭受攻击[21]。相反地，MTurk 提升了样本的多样性[22]，拥有来自美国及印度为主的 190 个国家大约 500，000 名工作者[23]。Huff 和 Tingley 比较 MTurk 与随机抽取的国会选举调查（Cooperative Congressional Election Survey，CCES），发现从几个不同的人口统计维度来看，MTurk 与 CCES 的数据差异不大[24]。Berinsky，Huber 和 Lenz 认为，众包样本的代表性虽然不及全国性的随机调查，但仍优于大部分面对面的传统方法[25]。

表 1　便利抽样方法比较

样本来源	众包	网络广告	校园	组织或社群
1 样本取得				
样本多元性	高	高	低	低
寻找特殊被试能力	高	普通	低	低
跨国研究能力	高	普通	低	低
2 研究成本				
招募被试成本	低	高	低	低
实验设计及管理成本	低	低	普通	普通
金流成本	低	高	低	低
被试交通及时间成本	低	低	普通	普通
3 研究质量				
回复质量控制能力	普通	低	普通	普通
匿名性	高	高	低	低
被试背景正确性	普通	低	高	高

（二）降低招募成本

众包网站建立了一个充满大量发起者与工作者的互动平台，因此研究人员不需要通过广告花费就能以较低成本让工作者看到其招募信息。Antoun等人发现，MTurk招募被试的成本只有Google及Facebook的5～15%[26]。与其他付费抽样服务相比，Wessling，Huber和Netzer则认为MTurk大约只要十分之一的成本[27]。

（三）降低研究成本

传统的实验室研究，除了现金报酬外，还有大量的管理及联络成本[28]，众包网站提供的工具满足了问卷需求，研究所需的金流作业、被试筛选、网络安全维护、网络管理、用户技术支持等能够通过众包网站的各种工具完成[7-9]。此外，基于网络平台，一些实验能够自动化进行，可以节省许多人力以及研究者时间[29]；被试可以随时随地参与实验，节省了时间及交通成本。

（四）提升回复质量

资料造假、回答时不用心、被试分享实验内容等研究常见问题，会危害数据质量。众包网站的许多功能可以帮助研究者减少这些潜在问题。首先，众包网站对于其工作者有实名要求。通过信用卡、住址及IP等信息，众包网站协助研究者确认工作者身份。再者，在网上进行实验，则可以减少实验者对被试的干扰。从动机上来说，Antoun等人发现众包平台招募的被试，是主动想要参与各种调研的被试，有较高的内部动机[26]。此外，众包网站的评分回馈机制提升了工作者的回复质量[30-31]。

（五）较好的匿名性

众包研究提供有效的匿名性，让研究者有机会接触到更多被试。许多心理研究的敏感性会让被试不愿意参与或回避回答，有些研究被试因为社交压力而影响作答。信息科技的使用减少了人与人直接接触的必要性[32]，让一些因为个人隐私而不愿意接触心理干预或治疗的个体愿意参与[33-34]，特别是较为敏感的治疗，例如在性行为及犯罪的治疗上效果特别明显[3,35]。虽然匿名对研究过程有很多帮助，但是真正的匿名研究实则不易，要维持匿名又要确认数据的真实性是一个两难挑战，而众包提供了很好的解决思路。在研究时，众包网站扮演中介者的角色，确认工作者的匿名性及数据正确性，也避免重复作答；对研究者来说，他们也无法实际接触隐私数据，所以维持了匿名性。

（六）能接触到特殊的被试族群

MTurk工作者背景多元数量众多，因此当研究者针对特定受试背景发出问卷需求，比较有机会邀请到足够多符合条件的被试。另外，许多研究者通过两阶段设计，先以低成本大规模地回收数万笔资料，再利用回收数据，辨识出合适对象进行下一步研究。例如Lynn使用MTurk收集了参加过两场特定战役的退伍军人资料[36]；Tran，Cabral，Patel和Cusack则利用MTurk针对五至七个月的婴儿进行了研究[37]。

（七）提高跨国研究的便利性

众包网站提供了跨国样本进行跨文化研究[38]。尤其对于非英语母语的研究者而言，MTurk为招募英语国家的被试提供便利的工具。许多非英语国家的研究者将英语系国家的研究成果，放在自己国家中进行验证，这个过程对于理论普适性有很大价值；然而，反过来将非英语国家的研究成果放在英语系国家中的验证研究，相对较少。主要的障碍及限制之一，就是跨国招募被试，取得信任，处理金流等研究困难，而这些问题通过MTurk可以有效解决[39-40]。

四、众包研究的有效性争议

基于上述优点，MTurk等众包网站作为被试招募工具近年来逐渐受到重视。但是，在研究方法选择上，更重要的是探讨这个方法能否达到研究的信效度要求，这正是众包研究的有效性程度面临的诸多挑战。

（一）众包的样本是否偏差？

样本代表性是众包的优点，但也是一个被质疑的重点。MTurk虽然号称有近两百个国家的工作者，实际上大部分的工作者都集中在美国及印度[38]。Steward等人的研究发现，虽然MTurk有大量登录的工作者，实际上经常在网上参与工作的人只有大约7,300人[41]。此外，众包对象主要是网络用户而且是属于主动愿意参与在线活动的人[42]，这个群体仅仅是一般人中的一小部分。与一般民众相比，MTurk的用户，普遍较年轻且拥有较高的教育程度；在美国人样本中，欧裔及亚裔过多，非裔及西裔太少；多数人没有工作，或没有固定工作，且收入较低；这些人未婚、租房生活的比例也较高。价值观上，他们较关心政治、追求自由、不重视宗教[9,24-25,43-45]。Holden，Dennie和Hicks在其研究中也提出质疑，MTurk的工作者必须使用信用卡进行注册，这个条件会限制经济上

弱势或是年龄未满18岁的人参与回答[46]。此外，从心理特质层面来分析，Goodman，Cryder和Cheema发现MTurk上的工作者相较于学生来说较为内向，情绪较不稳定，自信心较低[47]。Arditte等人则发现MTurk的工作者，表现出心理问题的比例相比一般样本显著偏高[48]。从众包组成变化来看，Ross等人发现MTurk的工作者背景在数周到数月的时间中持续地在变动[45]。Casey等人则发现，即使在同一天中不同时间就会呈现部分的差异[49]。最后，因为是否参与每一项工作是由工作者自由选择，而问卷选择的倾向也代表着不同背景与特质的群体。J. Chandler等人认为工作者会选择他们喜欢的工作类型，而且大部分工作者都会刻意记录他们喜欢的发起者，并且优先参与这些发起者工作[8]。

根据上述的讨论，众包样本的确存在代表性的风险，众包样本可能无法代表总体。然而，更值得我们探讨的是，这些问题是否导致研究的结果失去有效性。Berinsky等人发现，众包样本的背景与传统抽样方式或许有部分不同，但是对研究结论的实际影响却相当有限[43]。Paolacci等人，Horton等人及Berinsky等人都重做了许多经典实验，发现MTurk的被试也能够得到相似结论。这些实验与已有研究彼此验证理论有效性，似乎比过度讨论被试背景差异更有意义[7,38,43]。

与其他便利抽样相比，众包则更能代表一般人。Johnson和Borden在研究中比较了MTurk工作者与校园样本，发现MTurk用户平均年龄大十岁，更接近一般人的平均年龄[50]。Redmiles等人也发现，MTurk样本比一些市场调查公司的样本，更接近一般人[51]。事实上，在网络如此普及时代，我们更有理由相信网络人口比校园样本更接近真实人口分布，因此就网络用户来说，MTurk被试具有很好的代表性[38]。此外，许多研究不需要普适于所有人，因此即使众包样本的背景与总体有差异，并不足以让众包研究失去有效性。关键问题在于研究者如何设计实验，通过众包网站功能，找到与研究主题相关的被试[3]。Levay，Freese和Druckman发现，正确地用九个常用的人口统计变量进行条件控制，MTurk的结果就能非常具有代表性[52]。

（二）众包的回答者是否不用心？

不用心（inattentiveness）指被试没有专注于问题回答，包括单纯的不用心以及故意地随机乱填[53]。许多学者怀疑众包网站工作者的回答质量。J. Chandler等人发现，MTurk工作者有18%在填答问卷时同时在看电视、14%同时在听音乐、6%同时使用各种社交媒体[8]。Fleischer等人及Rouse也提到，在线问卷方法的不用心问题特别严重[54-55]。Huang等人进一步认为，分心或不用心会带来系统性偏差，提高随机分组的协方差以及一类误差几率，使得原本不应该产生的实证结论，错误地得到统计数据支持，而这个问题并不是随机分组就能改变[53]。

针对上述质疑，也有许多学者提出不同看法。首先，许多研究者相信，不用心回答问题可以通过良好的实验设计有效地控制并提升样本收集的质量[56]。包括置入指示性操作检查（instructional manipulation check，IMC）[57]、检查回复时间[50]、比较反义词的回复[58]，以及要求每一题的最少回复时间[59]等方法。此外，D. Chandler和Kapelner还发现，当MTurk被试被告知实验重要性，不仅参与实验的人增多，并且回答质量亦提高，证实了提升内部动机激励被试认真地作答[60]。Johnson和Borden使用六种人格测验比较了实验室与MTurk样本，发现MTurk样本能够提供合理的信度[50]。Holden等人，以及Buhrmester，Kwang和Gosling的研究也支持了上述结果[46,61]。反之，许多对回答质量的怀疑则未有实证支持。

研究者可以利用众包平台的评分机制来选择被试，提升应答率。评分机制是众包网站的特点，也是确保研究质量的重要流程。发起者能够对工作者评分，而这些评分纪录会成为其他发起者的参考，因此工作者会有谨慎回复问题的动机[54]。Peer等人发现，如果研究者筛选过去评分较好的MTurk工作者，这些工作者实验操控检验不通过的比例很低[56]。Berinsky等人进一步指出，MTurk的评分机制带来了很好的成效，激励工作者妥善地回答问题[43]。除了评分机制外，研究也能够从被试人口学背景上去筛选。Feitosa，Joseph和Newman发现，当研究者限制只有英语国家时，MTurk回答质量与其他问卷来源相当[62]。Litman，Robinson和Rosenzweig则发现研究者如果限制只招募美国本土的工作者时能进一步提升问卷质量[63]。

值得注意的是，实验操控检验等事后检验机制，一定程度上破坏了样本的普适性，并造成样本偏差[64-65]。当一部分问卷没有通过IMC时，删除这些样本一方面可能破坏外部有效性，留下这些样

本也可能伤害内部有效性[66]。此外，检验过程可能造成回复偏差。David J. Hauser 和 Schwarz，Mayo，Alfasi 和 Schwarz 都发现，被试的回答行为，会因为看到实验操控检验的题目，而产生心理变化进而影响后续作答[67-68]。

（三）金钱动机是否会破坏研究成果？

提供小额的金钱报酬一直是提高实验参与率及完成率的重要方法[69-70]，但是提供报酬可能会有负面的影响。这个问题在 MTurk 上受到了特别的重视，因为相对于传统研究，MTurk 的工作者参与实验的动机更多为金钱报酬[63]。Schmidt 的观察中提到，一个任务需要花多少时间以及会得到多少报酬是 MTurk 工作者最常讨论的问题之一[71]。Matthijsse，De Leeuw 和 Hox 也指出，确实有一类工作者，将回答问卷视为一种工作以获取报酬[72]。

金钱动机带来了 MTurk 研究的可能问题：首先，根据 Wessling 等人研究，为符合实验筛选要求，有相当比例的工作者会假造背景资料[27]。Cheung 等人也认为有一部分的工作者会为了得到报酬而假造账号[3]。J. J. Chandler 和 Paolacci 则发现，给的钱愈多，工作者说谎的比例愈高[73]。再者，以金钱为动机的工作者，对于问卷回复的内容质量可能较为不重视，而对于如何快速完成一份工作却更为在意，提高了问卷不用心及造假的风险。此外，金钱动机工作者，会偏向选择短时间与高报酬工作。而就研究伦理观点而言，研究关系不应是商业关系，如果被试以金钱报酬为目的，研究者还必须注重是否提供了符合最低薪资的报酬。而 MTurk 评鉴机制让发起者能够依照工作者的成果决定是否付给报酬，这并不符合研究伦理的标准[74]。

针对金钱动机这个议题，其他研究也提出了不同见解。首先就样本偏差来说，Kaufmann，Schulze 和 Veit 指出，虽然收入是工作者的一个重要动机，但是自主性等内部动机对工作者来说也很重要，因此金钱影响并非强烈[75]。Paolacci 等人也发现，只有约一成的人将 MTurk 当作主要的收入来源[38]，Horton 等人则认为 MTurk 工作者平均一个小时大约得到 1.4 美金，远低于最低工资。因此不能认定大多数被试有强烈的金钱动机[7]。不过，J. Chandler 和 Shapiro 则提醒我们必须进一步了解这些人是否因人格或背景特质无法找到有合理报酬的工作，才接受这些低于最低薪资的工作[9]。Matthijsse 等人在对各种在线作答的研究中指出，以报酬为目标的专业回答者，其人口统计特征与其他类型的回答者没有明显不同[72]。因此，虽然动机不同，但未必会造成样本偏差。

而针对回答质量的讨论，Hillygus，Jackson 和 Young 的研究发现以收入为主要动机的人，其讨好行为也未如预期般明显[76]。Aker 等人的研究也发现，给予较高的金钱报酬，非但不会降低样本的质量，相反会得到质量更好的回答[77]。反之，对不是以报酬为动机的研究工作者来说，由于他们常常是出于善意的帮忙，因此也很容易出现不用心及不投入回答问卷的状况[72]。

（四）众包的工作者可能重复参与？

当研究者设计实验时，会假设所有工作者是不重复而且对实验过程及置入刺激都不知情。MTurk 有 500,000 个用户，在每一次数百人抽样中会产生重复受试的机会似乎不高，然而这个假设并不正确[8]。首先，并非所有的 MTurk 工作者都是随时参与工作，因此如果研究者在短时间内重复收集上千笔数据，很容易发生重复。其次，用户参与任务的频率不同。MTurk 上会有一小部分以 MTurk 为主要收入来源的职业工作者，他们每天会参与许多小型任务，因此取样到他们的机会很高。J. Chandler 等人的调研发现大约有 10%的工作者回答了平台上 41%的问题。重复参与的另一种形式是参与类似的调查，也就是工作者虽然没有被同一个研究者重复取样，但是因为曾经参与过类似问卷，所以对问卷内容会产生预期，对实验操作会有心理预期[8]。

面对上述质疑，许多研究者也提出了不同见解。第一，虽然重复工作者在网络匿名问卷上是一个很难克服的问题，但 MTurk 本身从付款到评分等机制上，都在避免重复使用者。Berinsky 等人通过比对用户的 IP，发现重复用户低于 3%[43]。第二，在实验设计上，研究者在同一个工作中指派多个实验，在 MTurk 上，一个用户不能重复参与同一个任务，这些措施可以减少重复参与。在样本筛选上，研究者利用 MTurk 功能筛选没有经验的工作者，这也是一个管控职业工作者的好方法。同时，对于重复参与多个实验的研究者，有效做法是由研究者自己记录及管理所有参与过其研究工作者，再利用 MTurk 的设定排除曾经参与自己实验工作者[3]。第三，重复回答类似问题的严重程度与研究问题选择有关。如果研究者选择不是很常见

的研究方法及问题，这个问题就不会太严重[7]。第四，Schmidt 认为，大部分 MTurk 的工作者都不是以参与学术研究为其主要工作内容，因此研究者能接触到多元的工作者，无需担心重复作答的问题[71]。反之，职业工作者其实也有好处。J. Chandler 等人研究中发现，任务完成量大的职业工作者，通常在工作上比较专心，且对于配合长期追踪研究的回复率也较高。因此，MTurk 等众包网站的潜在缺点，实际上却也正是其优点[8]。

（五）被试之间是否会互相影响？

在研究设计上，我们都希望被试独立且没有互相影响。虽然众包使用者通常是彼此不认识的，但是却可能在网上互动而彼此影响。Schmidt 及 J. Chandler 等人在其研究中提到 MTurk 网站上有许多社群互动，部分工作者分享信息并相互讨论[8,71,78]。

然而，J. Chandler 等人观察发现，在论坛讨论实验内容的状况并不多[8]。事实上，当问卷的长度不长且报酬不高时，被试另外花时间分享及讨论内容的动机不强烈。同时，我们还是必须拿这个问题与其他研究方法相比，尤其是学校样本，MTurk 工作者的互动应会比同一个学校、课程或组织内的被试更少[3]。Edlund 等人在学生样本研究中发现，有少数学生的确会彼此告知研究内容，但他们也发现，如果研究者提醒这件事，并且得到学生口头承诺，这个问题可以得到有效的改善[79]。

五、讨论及建议

（一）结论

通过众包网站作为招募被试的来源，在过去几年大量地被社会科学与心理研究者所采纳。众包网站让研究人员能够以较低的成本接触更广泛或更特殊的被试，并且确保网络问卷的回复质量，因此成为一个广受关注的研究新方法。然而，在大受欢迎的同时，众包也遭受许多反对者的质疑。Chandler 和 Shapiro 认为 MTurk 是目前被最多人研究，也是最被清楚认识的非随机抽样工具[9]。而这个许多人质疑同时又有许多人支持的反复辩证过程，非但没有影响众包的学术研究应用，反而建立起了更扎实的实证基础。

我们也注意到因为争论，让许多研究者更重视其研究工具的外部效度，对整个学术发展来说非常正面。毕竟过去的研究人员普遍太过看重自己模型的变量关系，太过强调内部效度，而忽略了抽样是否代表目标群体[80]。因此，针对众包样本的种种批评，虽然可能是因为社会科学研究领域对一个新方法初期的不适应，但或许更能促使社会科学研究对普遍忽视其取样来源产生一种反省。通过这个“项庄舞剑”的过程，我们期望在众包的辩思中，重新回顾对取样方式的讨论，进而改善长期使用特定样本来源的普遍现象，提升心理理论的实证基础。

没有完美的研究工具，只有合适与否。从某种程度来说，所有样本都是便利性的样本，网络样本当然不是完全随机的[2,81]。不同的样本来源各有其优缺点，不论是校园、组织内或大规模的随机抽样，都有其值得讨论之处[82]。因此，挑选研究方法的重点并不在于方法本身的优缺点，也不应鼓励研究人员无止尽地追求样本的代表性，更重要的是选择一个适合研究目的的取样方法。同时，我们也认为更重要的是要求研究人员能清楚地说明获取样本的过程，其筛选条件为何，以及为何他们采用的样本能适合并满足其研究目的[8]。事实上，近年来许多针对 MTurk 研究的讨论方向，也从单纯 MTurk 的样本有效与否，转向为讨论 MTurk 适合什么样子的研究对象[51,83]。

就理论建构的有效性来说，通过不同的研究方法及样本来验证研究成果，是建立理论普适性的重要步骤[84]。从这个角度来看，众包存在的意义，不仅仅是提供一个低成本的样本来源，而是能由其样本的多样性，提升研究结果的普适性[85]，而在过去几年中，我们也愈来愈多地见到研究者利用 MTurk 平台重复验证过去已经发表过的心理研究[86-87]。因此，让许多有效的研究方法并存，通过大量的交互及反复验证，为心理学理论建构一个更加完整的实证基础，这或许是使用众包的更重要意义。

移动互联网时代开始，网络及行动装置普及、可穿戴传感器的应用，改变了心理研究收集、记录以及干预用户状态的方法[88]。随着在线心理研究的普及，众包被大量地应用在被试招募、数据收集甚至是网络心理干预的执行中[9]。同时，随着积极心理思潮的发展与自我帮助的普及，研究者愈来愈强调实验必须能在一般生活中完成[89]，使得网络心理实验成为一个逐渐普及的方式。相对于传统样本来源，众包更适合这个时代的需求以及研究范式改变。此外，科技的持续发展也在改善众包所受到的质疑。以网上身份造假及回答造假为例，近年

许多的研究都尝试透过各种人工智能的技术解决这些问题，也获得不错的成果[90－91]。本文研究者自己的研究经验也发现，尝试建立假身份希望骗过MTurk内部的审核机制并未如想象般容易。我们可以预见未来的众包平台将不仅仅是一个新的数据来源，而是提升数据正确性及有效性的重要帮手。

（二）建议

众包在国内仍是一个较为陌生的工具。目前众包与威客相关的研究，主要仍针对信息科技与商业应用为主，少数的研究曾经利用威客平台收集其所需要的问卷。但是对于使用众包作为科研平台的相关讨论仍然很少见[92]。参考MTurk在学术领域发展的历程，一个互联网工具必须历经反复验证，最后才可能成为一个新的研究范式。因此，我们期待有更多的国内社会科学研究者，尝试使用众包平台收集及分析数据，相信众包网站终将成为学术调研的一大助力。

其次，众包网站非常适合进行“中国问题”的跨文化及跨区域研究。国内研究者多数单向地将国外研究引到国内应用，却少见将国内的研究普适到其他国家。在传统的研究范式下，国内的心理研究者要在国外招募被试以及进行实验的成本比较高，而跨国的众包网站，则提供了不一样的机会。通过类似MTurk的众包工具，社会科学实验可以很容易招募美国等国家的被试，提高我们理论建构的国际化，也提高国内研究在全球的可见度。期望在未来看到愈来愈多的国内学者，利用众包网站拓展自己的研究成果，建立其研究成果的国际认同。

［注　释］

① Mechanical Turk，一般译为土耳其机器人，原指十九世纪时的一个骗人的把戏。发明人Wolfgang von Kempele，宣称设计出一个智能机械人能够下西洋棋，最后被发现是有人躲在其中[6]。

② 还有少数研究使用其他的众包平台，包括Crowd Flower，Prolific Academic[10]，Microworkers[11]，Samasource，ClickWorker[12]及Reddit[13]等等。此外，还包括一些公司通过网络广告提供的样本收集服务，例如Qualtrics，SSI[14]或是Google Analytics Service。

［参考文献］

[1] G. H. Bracht and G. V. Glass, “The external validity of experiments,” Am. Educ. Res. J., vol. 5, no. 4, pp. 437－474, 1968.

[2] R. N. Landers and T. S. Behrend, “An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples,” Ind. Organ. Psychol., vol. 8, no. 02, pp. 142－164, 2015.

[3] J. H. Cheung, D. K. Burns, R. R. Sinclair, and M. Sliter, “Amazon Mechanical Turk in Organizational Psychology: An Evaluation and Practical Recommendations,” J. Bus. Psychol., vol. 32, no. 4, pp. 347－361, 2017.

[4] E. Estellés-Arolas and F. González-Ladrón-De-Guevara, “Towards an integrated crowdsourcing definition,” J. Inf. Sci., vol. 38, no. 2, pp. 189－200, 2012.

[5] G. Chatzimilioudis, A. Konstantinidis, C. Laoudias and D. Zeinalipour-Yazti, “Crowdsourcing with smartphones,” IEEE Internet Comput., vol. 16, no. 5, pp. 36－44, 2012.

[6] J. Howe, “The rise of crowdsourcing,” Wired Mag., vol. 14, no. 6, pp. 1－4, 2006.

[7] J. J. Horton, D. G. Rand and R. J. Zeckhauser, “The online laboratory: Conducting experiments in a real labor market,” Exp. Econ., vol. 14, no. 3, pp. 399－425, 2011.

[8] J. Chandler, P. Mueller and G. Paolacci, “Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers,” Behav. Res. Methods, vol. 46, no. 1, pp. 112－130, 2014.

[9] J. Chandler and D. Shapiro, “Conducting clinical research using crowdsourced convenience samples,” Annu. Rev. Clin. Psychol., vol. 12, pp. 53－81, 2016.

[10] E. Peer, L. Brandimarte, S. Samat and A. Acquisti, “Beyond the Turk: Alternative platforms for crowdsourcing behavioral research,” J. Exp. Soc. Psychol., vol. 70, pp. 153－163, 2017.

[11] D. L. Crone and L. A. Williams, “Crowdsourcing participants for psychological research in Australia: A test of Microworkers,” Aust. J. Psychol., vol. 69, no. 1, pp. 39－47, 2017.

[12] G. B. Schmidt and W. M. Jettinghoff, “Using Amazon Mechanical Turk and other compensated crowdsourcing sites,” Bus. Horiz., vol. 59, no.

4, pp. 391－400, 2016.

[13] M. R. Jamnik and D. J. Lane, "The Use of Reddit as an Inexpensive Source for High-Quality Data.," Pract. Assess. Res. Eval., vol. 22, no. 5, 2017.

[14] K. S. Wessling, J. Huber and O. Netzer, "MTurk Character Misrepresentation: Assessment and Solutions," J. Consum. Res., vol. 44, no. 1, pp. 211－230, 2017.

[15] D. G. Rand, J. D. Greene and M. A. Nowak, "Spontaneous giving and calculated greed," Nature, vol. 489, no. 7416, pp. 427－430, 2012.

[16] Christine Ma-Kellams and Jennifer Lerner, "Trust Your Gut or Think Carefully Examining Whether an Intuitive, Versus a Systematic, Mode of Thought Produces Greater Empathic Accuracy," J. Pers. Soc. Psychol., vol. 111, no. 5, pp. 674－685, 2016.

[17] D. J. Hauser, S. D. Preston and R. B. Stansfield, "Altruism in the wild: when affiliative motives to help positive people overtake empathic motives to help the distressed.," J. Exp. Psychol. Gen., vol. 143, no. 3, pp. 1295－1305, 2014.

[18] K. Diehl, G. Zauberman and A. Barasch, "How taking photos increases enjoyment of experiences." J. Pers. Soc. Psychol., vol. 111, no. 2, pp. 119－140, 2016.

[19] A. L. Thomson and J. T. Siegel, "A moral act, elevation and prosocial behavior: Moderators of morality," J. Posit. Psychol., vol. 8, no. 1, pp. 50－64, 2013.

[20] J. T. Siegel, A. L. Thomson and M. A. Navarro, "Experimentally distinguishing elevation from gratitude: Oh, the morality," J. Posit. Psychol., vol. 9, no. 5, pp. 414－427, 2014.

[21] 彭凯平，刘钰，曹春梅，张伟. 虚拟社会心理学：现实，探索及意义[J]. 心理科学进展，2011，(7).

[22] S. D. Rhodes, D. A. Bowie and K. C. Hergenrather, "Collecting behavioural data using the world wide web: considerations for researchers," J. Epidemiol. Community Health, vol. 57, no. 1, pp. 68－73, 2003.

[23] G. Paolacci and J. Chandler, "Inside the Turk: Understanding Mechanical Turk as a participant pool," Curr. Dir. Psychol. Sci., vol. 23, no. 3, pp. 184－188, 2014.

[24] C. Huff and D. Tingley, "'Who are these people?' Evaluating the demographic characteristics and political preferences of MTurk survey respondents," Res. Polit., vol. 2, no. 3, 2015.

[25] A. J. Berinsky, G. A. Huber and G. S. Lenz, "Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk," Polit. Anal., vol. 20, no. 3, pp. 351－368, 2012.

[26] C. Antoun, C. Zhang, F. G. Conrad and M. F. Schober, "Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook and Amazon Mechanical Turk," Field Methods, vol. 28, no. 3, pp. 231－246, 2016.

[27] K. Sharpe Wessling, J. Huber and O. Netzer, "MTurk Character Misrepresentation: Assessment and Solutions," J. Consum. Res., vol. 44, no. 1, pp. 211－230, 2017.

[28] J. K. Goodman and G. Paolacci, "Crowdsourcing consumer research," J. Consum. Res., vol. 44, no. 1, pp. 196－210, 2017.

[29] G. Paolacci, J. Chandler and P. G. Ipeirotis, "Running experiments on amazon mechanical turk," 2010.

[30] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino and S. Dustdar, "Quality control in crowdsourcing systems: Issues and directions," IEEE Internet Comput., vol. 17, no. 2, pp. 76－81, 2013.

[31] L. De Alfaro, A. Kulshreshtha, I. Pye and B. T. Adler, "Reputation systems for open collaboration," Commun. ACM, vol. 54, no. 8, pp. 81－87, 2011.

[32] A. E. Kazdin and S. L. Blase, "Rebooting psychotherapy research and practice to reduce the burden of mental illness," Perspect. Psychol. Sci., vol. 6, no. 1, pp. 21－37, 2011.

[33] D. B. Hoch et al., "The feasibility and impact of delivering a mind-body intervention in a virtual world," Plos One, vol. 7, no. 3, p. e33843, 2012.

[34] J. Mitchell, D. Vella-Brodrick and B. Klein, "Positive psychology and the internet: A mental health opportunity," Sensoria J. Mind Brain Cult., vol. 6, no. 2, pp. 30－41, 2010.

[35] K. M. O'Neil and S. D. Penrod, "Methodological variables in Web-based research that may affect results: Sample type, monetary incentives and personal information," Behav. Res.

Methods, vol. 33, no. 2, pp. 226－233, 2001.

[36] B. M. －D. Lynn, Shared sense of purpose and well-being among veterans and non-veterans. North Carolina State University, 2014.

[37] M. Tran, L. Cabral, R. Patel and R. Cusack, "Online recruitment and testing of infants with Mechanical Turk," J. Exp. Child Psychol., vol. 156, pp. 168－178, 2017.

[38] G. Paolacci, J. Chandler and P. G. Ipeirotis, "Running experiments on Amazon Mechanical Turk.," Judgm. Decis. Mak., vol. 5, no. 5, pp. 411－419, 2010.

[39] K. A. Barchard, K. E. Grob and M. J. Roe, "Is sadness blue? The problem of using figurative language for emotions on psychological tests," Behav. Res. Methods, vol. 49, no. 2, pp. 443－456, 2017.

[40] A. Nishi, N. A. Christakis and D. G. Rand, "Cooperation, decision time and culture: Online experiments with American and Indian participants," PloS One, vol. 12, no. 2, p. e0171252, 2017.

[41] N. Stewart et al., "The average laboratory samples a population of 7, 300 Amazon Mechanical Turk workers," Judgm. Decis. Mak., vol. 10, no. 5, pp. 479－491, 2015.

[42] S. E. Woo, M. Keith and M. A. Thornton, "Amazon Mechanical Turk for industrial and organizational psychology: Advantages, challenges and practical recommendations," Ind. Organ. Psychol., vol. 8, no. 02, pp. 171－179, 2015.

[43] A. J. Berinsky, G. A. Huber and G. S. Lenz, "Using Mechanical Turk as a subject recruitment tool for experimental research," 2011.

[44] P. G. Ipeirotis, F. Provost and J. Wang, "Quality management on amazon mechanical turk," in Proceedings of the ACM SIGKDD workshop on human computation, 2010, pp. 64－67.

[45] J. Ross, L. Irani, M. Silberman, A. Zaldivar and B. Tomlinson, "Who are the crowdworkers: shifting demographics in mechanical turk," in CHI' 10 extended abstracts on Human factors in computing systems, 2010, pp. 2863－2872.

[46] C. J. Holden, T. Dennie and A. D. Hicks, "Assessing the reliability of the M5-120 on Amazon' s Mechanical Turk," Comput. Hum. Behav., vol. 29, no. 4, pp. 1749－1754, 2013.

[47] J. K. Goodman, C. E. Cryder and A. Cheema, "Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples," J. Behav. Decis. Mak., vol. 26, no. 3, pp. 213－224, 2013.

[48] K. A. Arditte, D. Çek, A. M. Shaw and K. R. Timpano, "The importance of assessing clinical phenomena in Mechanical Turk research.," Psychol. Assess., vol. 28, no. 6, pp. 684－691, 2016.

[49] L. S. Casey, J. Chandler, A. S. Levine, A. Proctor and D. Z. Strolovitch, "Intertemporal Differences Among MTurk Workers: Time-Based Sample Variations and Implications for Online Data Collection," SAGE Open, vol. 7, no. 2, 2017.

[50] D. R. Johnson and L. A. Borden, "Participants at your fingertips using Amazon' s Mechanical Turk to increase student-faculty collaborative research," Teach. Psychol., vol. 39, no. 4, pp. 245－251, 2012.

[51] E. M. Redmiles, S. Kross, A. Pradhan and M. L. Mazurek, "How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk and Web Panels to the US," 2017.

[52] K. E. Levay, J. Freese and J. N. Druckman, "The demographic and political composition of Mechanical Turk samples," Sage Open, vol. 6, no. 1, 2016.

[53] J. L. Huang, M. Liu and N. A. Bowling, "Insufficient effort responding: Examining an insidious confound in survey data.," J. Appl. Psychol., vol. 100, no. 3, pp. 828－845, 2015.

[54] A. Fleischer, A. D. Mead and J. Huang, "Inattentive Responding in MTurk and Other Online Samples," Ind. Organ. Psychol., vol. 8, no. 2, pp. 196－202, 2015.

[55] S. V. Rouse, "A reliability analysis of Mechanical Turk data," Comput. Hum. Behav., vol. 43, pp. 304－307, 2015.

[56] E. Peer, J. Vosgerau and A. Acquisti, "Reputation as a sufficient condition for data quality on Amazon Mechanical Turk," Behav. Res. Methods, vol. 46, no. 4, pp. 1023－1031, 2014.

[57] A. W. Meade and S. B. Craig, "Identifying careless responses in survey data.," Psychol. Methods, vol. 17, no. 3, pp. 437－455, 2012.

[58] J. L. Huang, P. G. Curran, J. Keeney, E. M. Poposki and R. P. DeShon, "Detecting and

deterring insufficient effort responding to surveys," J. Bus. Psychol., vol. 27, no. 1, pp. 99－114, 2012.

[59] A. Kapelner and D. Chandler, "Preventing Satisficing in Online Surveys," Proc. Crowdconf, 2010.

[60] D. Chandler and A. Kapelner, "Breaking monotony with meaning: Motivation in crowdsourcing markets," J. Econ. Behav. Organ., vol. 90, pp. 123－133, 2013.

[61] M. Buhrmester, T. Kwang and S. D. Gosling, "Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?," Perspect. Psychol. Sci., vol. 6, no. 1, pp. 3－5, 2011.

[62] J. Feitosa, D. L. Joseph and D. A. Newman, "Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English," Personal. Individ. Differ., vol. 75, pp. 47－52, 2015.

[63] L. Litman, J. Robinson and C. Rosenzweig, "The relationship between motivation, monetary compensation and data quality among US-and India-based workers on Mechanical Turk," Behav. Res. Methods, vol. 47, no. 2, pp. 519－528, 2015.

[64] D. M. Oppenheimer, T. Meyvis and N. Davidenko, "Instructional manipulation checks: Detecting satisficing to increase statistical power," J. Exp. Soc. Psychol., vol. 45, no. 4, pp. 867－872, 2009.

[65] J. S. Downs, M. B. Holbrook, S. Sheng and L. F. Cranor, "Are your participants gaming the system?: screening mechanical turk workers," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2010, pp. 2399－2402.

[66] A. J. Berinsky, M. F. Margolis and M. W. Sances, "Can we turn shirkers into workers?," J. Exp. Soc. Psychol., vol. 66, pp. 20 － 28, 2016.

[67] D. J. Hauser and N. Schwarz, "It's a trap! Instructional Manipulation checks prompt systematic thinking on 'tricky' tasks," SAGE Open, vol. 5, no. 2, p. 2158244015584617, 2015.

[68] R. Mayo, D. Alfasi and N. Schwarz, "Distrust and the positive test heuristic: Dispositional and situated social distrust improves performance on the Wason Rule Discovery Task.," J. Exp. Psychol. Gen., vol. 143, no. 3, pp. 985－990, 2014.

[69] J. S. Armstrong, "Monetary incentives in mail surveys," Public Opin. Q., vol. 39, no. 1, pp. 111－116, 1975.

[70] M. J. Shaw, T. J. Beebe, H. L. Jensen and S. A. Adlis, "The use of monetary incentives in a community survey: impact on response rates, data quality and cost.," Health Serv. Res., vol. 35, no. 6, pp. 1339－1346, 2001.

[71] G. B. Schmidt, "Fifty days an MTurk Worker: the social and motivational context for Amazon Mechanical Turk Workers," Ind. Organ. Psychol., vol. 8, no. 02, pp. 165－171, 2015.

[72] S. M. Matthijsse, E. D. De Leeuw and J. J. Hox, "Internet panels, professional respondents and data quality," Methodology, vol. 11, pp. 81－88, 2015.

[73] J. J. Chandler and G. Paolacci, "Lie for a Dime: When Most Prescreening Responses Are Honest but Most Study Participants Are Impostors," Soc. Psychol. Personal. Sci., 2017.

[74] I. H. Gleibs, "Are all 'research fields' equal? Rethinking practice for the use of data from crowdsourcing market places," Behav. Res. Methods, pp. 1－10, 2016.

[75] N. Kaufmann, T. Schulze and D. Veit, "More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk.," in AMCIS, 2011, vol. 11, pp. 1－11.

[76] D. S. Hillygus, N. Jackson and M. Young, "Professional respondents in non-probability online panels," Online Panel Res. Data Qual. Perspect., pp. 219－237, 2014.

[77] A. Aker, M. El-Haj, M. － D. Albakour and U. Kruschwitz, "Assessing Crowdsourcing Quality through Objective Tasks.," in LREC, 2012, pp. 1456－1461.

[78] J. W. Bentley, "Challenges with Amazon Mechanical Turk Research in Accounting," 2017.

[79] J. E. Edlund, B. J. Sagarin, J. J. Skowronski, S. J. Johnson and J. Kutter, "Whatever happens in the laboratory stays in the laboratory: The prevalence and prevention of participant crosstalk," Pers. Soc. Psychol. Bull., 2009.

[80] R. M. Groves, "Research on survey data quality," Public Opin. Q., vol. 51, pp. S156-S172, 1987.

[81] S. D. Gosling, S. Vazire, S. Srivastava and O. P. John, "Should we trust web-based studies? A comparative analysis of six preconceptions about

internet questionnaires.," Am. Psychol., vol. 59, no. 2, pp. 93－104, 2004.
[82] X. S. Zhu, J. L. Barnes-Farrell and D. K. Dalal, "Stop apologizing for your samples, start embracing them," Ind. Organ. Psychol., vol. 8, no. 02, pp. 228－232, 2015.
[83] M. G. Keith and P. D. Harms, "Is Mechanical Turk the Answer to Our Sampling Woes?," Ind. Organ. Psychol., vol. 9, no. 1, pp. 162－167, 2016.
[84] W. R. Shadish, "Revisiting field experimentation: field notes for the future.," Psychol. Methods, vol. 7, no. 1, pp. 3－18, 2002.
[85] D. J. Simons, "The value of direct replication," Perspect. Psychol. Sci., vol. 9, no. 1, pp. 76－80, 2014.
[86] J. R. Pauszek, P. Sztybel and B. S. Gibson, "Evaluating Amazon's Mechanical Turk for psychological research on the symbolic control of attention," Behav. Res. Methods, pp. 1－15, 2017.
[87] R. A. Zwaan et al., "Participant Nonnaiveté and the reproducibility of cognitive psychology," Psychon. Bull. Rev., pp. 1－5, 2017.
[88] M. E. Morris and A. Aguilera, "Mobile, social and wearable computing and the evolution of psychological practice.," Prof. Psychol. Res. Pract., vol. 43, no. 6, p. 622－626., 2012.
[89] A. S. Waterman, "The humanistic psychology-positive psychology divide: Contrasts in philosophical foundations.," Am. Psychol., vol. 68, no. 3, pp. 124－133, 2013.
[90] M. Hibbeln, J. Jenkins, C. Schneider, J. Valacich and M. Weinmann, "Investigating the Effect of Insurance Fraud on Mouse Usage in Human-Computer Interactions," 2014.
[91] M. Monaro, F. I. Fugazza, L. Gamberini and G. Sartori, "How Human-Mouse Interaction can Accurately Detect Faked Responses About Identity," in International Workshop on Symbiotic Interaction, 2016, pp. 115－124.
[92] 孟韬，张媛，董大海. 基于威客模式的众包参与行为影响因素研究 [J]. 中国软科学，2014，(12).

Amazon Mechanical Turk Crowdsourcing: A Research Tool in Mobile Internet Era

PENG Kai-ping[1,2], LIU Shi-qun[1], NI Shi-guang[3]
(1. Tsinghua-Berkeley Shenzhen Institute, Shenzhen, Guandong, 518055, PRC;
2. Department of Psychology, Tsinghua University, Beijing, 100084, PRC;
3. Graduate School at Shenzhen, Tsinghua University, Shenzhen, Guandong, 518055, PRC)

[**Abstract**] Representativeness, quality, and quantity of samples have critical implications for validating the conclusions in research. Despite their importance, textbooks and research reports have underemphasised the external validity. Recruiting participants from crowdsourcing websites, as exemplified by Amazon's Mechanical Turk (MTurk), has emerged to be a promising sampling strategy to mitigate the problem. However, the use of crowdsourcing samples gets suspicion. Many researchers questioned experiments using MTurk on their biased, inattentive, repeated, monetary motivated, and previously informed participants. Others, on the other hand, defended with empirical evidence. Despite the debate, more and more studies started to gather information from crowdsourcing samples. This paper presents both sides of the arguments with emphasis on the validity and reliability of sampling process, provides reasons to support the use of crowdsourcing samples in response to the technological advance in the mobile internet. This paper also urges researchers to investigate the potential of using crowdsourcing samples in cross-country research. It is an unprecedented opportunity to generalise research findings from local to global areas.

[**Key words**] crowdsourcing; Amazon Mechanical Turk; social sciences; convenience sampling

（责任编辑　胡晓春/校对　谷雨）