大规模开放在线学习学生互评效果实证研究

罗 恒[1] 左明章[1] 安东尼·鲁宾逊[2]

(1 华中师范大学 教育信息技术学院,武汉 430079;
2. 宾夕法尼亚州立大学 地理系,美国)

[摘要] 学生互评能够有效弥补教师评阅和机器评分的局限,是适用于大规模开放在线学习情境的重要评价模式。然而,现阶段对在线互评模式的准确性和有效性尚缺少基于实证的系统研究。基于此,本文通过对一门大规模开放在线课程(MOOC)的学生互评、自评和教师评分等数据的比较,得出在线互评模式的信度、效度、影响因素和学生认可度等初步结论。研究结果表明,尽管互评模式的评分者间信度并不理想,但综合考量多个评分结果的前提下,该模式能够为在线学习者提供较为一致可靠的最终得分。学生互评结果和教师评分结果的相关性系数高达 0.619 表明,该模式同时具有较好的聚合效度。此外,对课后问卷的统计分析表明,在线学习者对互评模式总体持积极态度,认可互评活动对反馈获得、课程投入和高阶思维培养等的有益影响。这些发现和结论对完善和改进面向大规模开放在线学习的评价模式有一定的指导意义。

[关键词] 学生互评;评分者间信度;聚合效度;大规模开放在线学习

[中图分类号] G436 [文献标识码] A [文章编号] 1007-2179(2017)01-0075-09

一、引言

《2015 年中国互联网年度热点洞察报告》显示,我国在线教育市场规模 2015 年达到 1192 亿元,用户达到 7227 万人,涵盖高等教育、中小学教育、职业培训及学前教育等领域。然而,在线教育迅速发展的背后亦伴随着对其教学质量的质疑。其中比较突出的一个问题是其“教-学-评”体系的不完善:现阶段尚缺少针对大规模在线学习活动的合理有效的评价模式和机制(Kauza, 2014; Piech et al., 2013)。以大规模开放在线课程为例,庞大的学生规模使授课教师没有足够时间和精力批改每份作业,而机器评分又不适用于评估复杂的学习成果,如项目设计、艺术作品及论文报告等(高地,2014;熊瑶等,2016)。在线评价机制的局限性导致大量在线课堂重教学内容而轻测评,学习活动被简化为单向的信息接收而缺乏双向实时的反馈与评价,严重影响了在线教育的教学效果和社会认可度(顾小清等,2013;康叶钦,2014;Mehaffy, 2012)。

为解决在线教学规模化所带来的评价难问题,Coursera 平台于 2012 年推出定标同伴评估(Calibrated Peer Review,简写为 CPR)模式。该模式将传统课堂内教师负责的评价活动众包给广大在线学习者,引导学生使用预先设定的评分标准为彼此的作业进行打分和提供反馈,旨在尽可能减少教师参与的同时最大程度地提高在线评价的准确性、有效性

[收稿日期]2016-10-04 [修回日期]2016-12-21 [DOI 编码]10.13966/j.cnki.kfjyyj.2017.01.009
[基金项目]2016 年度教育部人文社会科学青年基金项目“面向大规模在线教育的众包评测模型研究”(16YJC880054)。
[作者简介]罗恒,博士,讲师,华中师范大学教育信息技术学院(luoheng@mail.ccnu.edu.cn),研究方向:在线教育、技术整合教学、学习评价;左明章,博士,教授,华中师范大学教育信息技术学院,研究方向:教育数字媒体、教育技术理论;安东尼·鲁宾逊(Anthony Robinson),博士,助理教授,宾夕法尼亚州立大学主校区地理系,研究方向:地理信息系统、MOOC 教学。

和及时性。然而,当前针对这一互评模式的实证研究还比较匮乏,对开放在线教学环境中学生互评活动的信度、效度和优缺点尚缺少系统论证。因此,本研究依据在 Coursera 平台上的 MOOC 的教学经验和收集的学习数据,考察学生从设计到实施互评活动的全过程,并分析他们的评分数据和反馈结果,系统检验学生互评在大规模开放在线学习中的实际效果以及相关影响因素。具体来说,本研究旨在回答以下三个问题:

1)学生互评能否为大规模开放在线学习情境提供可靠有效的评价手段?

2)哪些潜在因素影响该情境下学生互评的信度与效度?

3)学生互评为大规模开放在线学习带来哪些潜在的益处和问题?

二、相关文献研究

(一)学生互评概述

学生互评,也称同伴互评,其核心是组织学习者对能力相当的其他学习者的学习作品或表现进行水平、价值或质量的考量和判定(Topping, 2009)。互评结果通常是量化的评定得分,有时也以文字评价的形式呈现。在很多情况下,互评结果包括上述两种形式,是它们的有机结合(Lu & Law, 2012; Strijbos et al., 2010)。学生互评作为一种学习评价方式有着悠久的历史,被广泛应用于自然科学(Billington, 1997; Butcher et al., 1995)、社会科学(Falchikov, 1994; Orpen, 1982)、医学(Hammond & Kern, 1959; Magin, 1993)、商学(Freeman, 1995; Kaimann, 1974)、二语习得(邓鄘鸣等,2010;韩冰,2009)和工程技术学科(Fry, 1990; Oldfield & Macalpine, 1995)等多个学科领域。

学生互评将教师从繁重的审阅任务中解放出来,极大地减轻了教学工作量。此外,文献研究表明互评活动本身也能促进学习的发生,为学生带来许多潜在的益处,如学习的主人翁精神和自治精神(Brown et al., 1995; Race, 1998),更高的学习动机(Vu & Dall'Alba, 2007),更强的社交存在感(Strijbos & Sluijsmans, 2010)以及高阶思维和元认知能力的发展等(Mok, 2011; Topping, 2009; Wen et al., 2006)。然而,这些潜在益处不能完全说服师生在教学过程中将互评作为主要评价方式(Cho et al., 2006; Magin, 2001; Stefani, 1994),对学生能力的不自信而导致的对互评结果信效度的质疑是该模式受阻的主要原因(Falchikov & Goldfinch, 2000; McGarr & Clifford, 2013)。

(二)学生互评的信度与效度

学生互评模式的信度与效度文献研究主要集中在传统面授课堂教学,鲜有针对自主在线学习情境的探索(Cho et al., 2006; Falchikov & Goldfinch, 2000; Zhang et al., 2008)。互评结果的信度一般由不同评分者对同一作业的评分一致性判定。互评结果的效度通常通过计算学生打分结果和专家打分结果的相关性系数得来,相关系数越高,证明互评结果的效度越高。一般认为,任课教师对授课内容有着深入了解,能够对学生的表现或作业给出准确、公正的分数和评价,因而文献中专家角色几乎都由任课老师担任。换言之,文献中讨论的学生互评信度和效度也可以看成学生评分者间信度以及"教师-学生"评分结果的聚合效度。

很多研究揭示了学生评分结果和教师评分结果之间具有较强的正相关性,由此可以证明学生的专业知识水平虽然不如教师,但是基于多个学生互评的最终分数具有较高的效度,因而有相当的参考价值。例如,法契科夫和戈德芬奇(Falchikov & Goldfinch,2000)曾对 1959 年至 1999 年间发表的 56 项关于学生互评的量化研究进行荟萃分析,发现学生评分结果和教师评分结果显著强相关($r = 0.69$)。也有学者针对在线教学(Bouzidi & Jaillet, 2009)和中学教育(Sadler & Good, 2006)情境下的学生互评效度进行研究并得出了相同结论:学生互评在以上两种教学情境中都有极高的效度,与教师评分的相关系数分别介于 $r = 0.88 - 0.91$ 和 $r = 0.91 - 0.94$ 之间。当然,我们也注意到少数文献报告了学生互评模式低效度的证据,在一些教学事件中学生评分结果和教师评分结果分歧较大(Cheng & Warren, 1999; Korman & Stubblefield, 1971; Mowl & Pain, 1995)。

与互评效度方面已有大量文献不同,互评信度方面的研究相对匮乏,研究者很少关注互评结果背后学生打分一致性问题。对互评信度的忽视将直接影响对其效度的判定,因为一个高效度的评测方法

也应该是稳定、一致与可靠的,必须同时满足较高的聚合效度和评分者间信度两个条件(Gay & Airasian, 2003)。值得注意的是,一些研究混淆了信度和效度概念,对统计分析结果做出了错误解读(Topping, 2009)。根据学生互评模式中评分者数目、评分者选择方式不同等,文献中给出了不同的计算互评信度的统计分析方法,如采用皮尔逊相关系数(Haaga, 1993)、比例方差(Marcoulides & Simkin, 1995)和组间相关系数(Cho et al., 2006; Miller, 2003)等。相关统计结果总体表明学生评分者在互评任务中能够给出较为一致和可靠的分数。但如果要对学生互评的信度做出更肯定和普遍的推论还需要更多来自不同教学情境的实证研究证据。

一些学者着重考察了影响学生互评信效度的因素。例如,法契科夫和戈德芬奇(2000)研究发现,学生使用复合分数按照预先指定标准对学术作品进行互评得到的评分结果更接近教师的评分结果,进而将"分数结构""作业类型""评分标准"确定为影响互评信度的重要因素。此外,"评分者的数目"也是影响互评分数信效度的重要因素。赵光洙等(Cho et al., 2006)发现,每增加一个学生参与作业评分都将显著提升评分结果的信效度。另一方面,一些通常被认为会影响学习评价的因素,如学科领域、课程难度和学生态度,则被证实对互评信效度的影响十分有限(Falchikov & Goldfinch, 2000; McGarr & Clifford, 2013)。

综上所述,相关文献研究从总体上支持学生互评的有效性和合理性,并指出一系列可能影响互评结果信效度的潜在因素。然而,我们应该看到这些研究大多基于大学面授学分制课程的教学情境,这种情境具有学生人数少、构成相对同质、教师能够全程监控指导等特征。相关研究发现是否适用于学生人数规模化、组成成分多元化的大规模开放在线学习情境尚不可知,亟待进一步检验与探索。

(三)来自 MOOC 的证据

基于众包概念的定标同伴评估(Calibrated Peer Review)在 Coursera 平台上一经推出,就吸引了不少教师、学生、学者和媒体的注意:不少人从教师或学生角度描述了 MOOC 课堂中使用学生互评的教学体验;在热门媒体网站和个人博客上关于学生互评的有效性、优越性和局限性的讨论也是持续不断、逐步升温(McEwen, 2013; Morrison, 2013; Neidlinger, 2013; Rees, 2013; Watters, 2012)。总之,关于学生互评这种评价模式,人们的看法分歧较大。例如,里斯(Rees, 2013)描述了她在一门世界历史 MOOC 的学习体验,认可为自己作业评分同学的专业与客观,并承认自己认真努力完成的作业往往能够获得更高分数。奈德林格(Neidlinger, 2013)则道出了很多 MOOC 学生对互评结果的不满,认为有相当一部分上课的学生并不具备评判作业质量的资格,且很多人评分只凭个人喜好而没有参考教师给出的评分标准。麦克尤恩(McEwen, 2013)和沃特斯(Watters, 2012)进一步指出在 MOOC 中使用学生互评的一些潜在问题,如反馈质量参差不齐、缺少互惠感和社区存在感,以及质量监控与调控的缺失等。当然,这些关于 MOOC 环境中学生互评效果的论断很多都是主观感受,没有经过实证研究验证。同时,基于实际 MOOC 评测数据而得出学生互评效果的研究十分匮乏。

三、研究方法

(一)研究情境

本研究收集和分析的数据来自于 Coursera 平台上的 MOOC"地图与地理空间革命"(www.coursera.org/course/maps)。该课程是美国宾夕法尼亚州立大学 2013 年开设的一门为期五周的地图绘制和地理空间分析入门课程。本文第三作者安东尼·鲁宾逊是该课程的主讲老师,第一作者罗恒参与了课程设计和实施过程。48984 名学生注册这门课程,但最后一周仍活跃的学生只有 8707 人。根据 7551 名学生在课程结束之后填报的人口特征数据显示:选修该课程的大部分是男生,女生只占 30% 左右;约 61% 的学生是全职工作时间之余学习课程;学生平均年龄是 36.5 岁;超过 80% 的学生有本科或以上学历,其中最高学历为本科的占 33.8%,研究生占 39.1%,博士生占 8%;30% 左右的学生来自美国,其余学生来自世界各国,以欧洲和东南亚地区居多。3064 名学生通过了该课程考核,其中 1211 人获优秀。

课程教师只在最后一周布置一次开放性作业,相应只有一次学生互评活动。期末作业要求每个学生自选一种地图绘制的工具或平台(如 ArcGIS On-

line、QGIS、和 GRASS),任选一个话题并设计和绘制一幅地图讲诉一个和生活切身相关的故事。主题可以是最近的一次旅行路线、家乡最棒餐馆的分布图或者某区域数年来经济文化的演变等。作业成绩占课程总分的20%,学生将根据教师撰写的评价量规从四个维度对上交的地图作品进行评分,包括展示清晰度、故事可信度、制图水平(如颜色、符号的使用和布局等)和设计美观度,每个维度得分从低到高为0分到3分。互评作业的总分为四个维度得分的总合,即在0分和12分之间。课程要求每位学生至少评价三份其他同学上交的作业。这些作业由Coursera平台随机分配给不同学生评分者。同时,每位学生也需对自己上交的作业进行自评。值得注意的是,Coursera平台为了应对极值分数,选择使用学生评分结果的中值(median)而不是平均值(mean)作为最终的互评分数。

(二)数据收集

本研究共收集了三类数据。第一类数据是MOOC学生针对开放性作业的互评和自评数据。该数据储存在Coursera平台后台数据库的submission_metadata部分:作业编号和最终互评分数存储在overall_evaluation_metadata中,单个学生评分者的评分结果储存在evaluation_metadata中,学生的自评分数存储在self_grading_metadata中。Coursera平台设置最多的评分者人数是5人。本研究中共1825份作业获得了5位评分者的评分,从而被选中进行后续的信效度分析。缺失评分数据的作业共有919份,这些作业被排除在数据分析之外。关于作业的最终互评分数,除了使用Coursera平台提供的基于中值的判定,本研究同时将各个学生评分的平均值作为最终结果。

第二类数据是教师对开放性作业的评分数据。考虑到教师评分工作量,本研究从1825份作业中随机选择了5%(N=93)并请课程主讲教师进行批改。通过作业的编号,教师可以在后台数据库中访问学生提交的作业原件,并按照相同的评价量规进行评分,包括作业的总分和四项标准的得分。因此,每份作业包含以下评价数据:5位学生评分者的评分结果、基于中值的最终互评分数、基于平均值的最终互评分数、教师评分结果和自评结果(见图1)。

第三类数据是学生对互评活动的态度。学生在课程结束后填写了MOOC学习体验自我评价问卷,其中有7道题是关于课程中的互评活动的,收集学生对互评活动的公正性、有效性和潜在益处的看法。考虑到从Coursera后台数据库提取问卷数据的复杂性,本研究没有采用平台自带的问卷工具,而是使用第三方问卷收集工具—Qualtrics,通过学生编号将Coursera平台上学生的学习数据和相应的学生问卷数据联结起来。

(三)数据分析

本研究中学生互评信度本质上是评分者间信度,测量的是不同学生评分者对同一评价任务评分结果的总体一致性。因为作业是随机分配给特定学生总体中的五位评分者,针对该评分机制,本研究选择了第一类组内相关系数(Case 1 Intraclass Correlation Coefficient,简称ICC[1])作为评分者间信度的

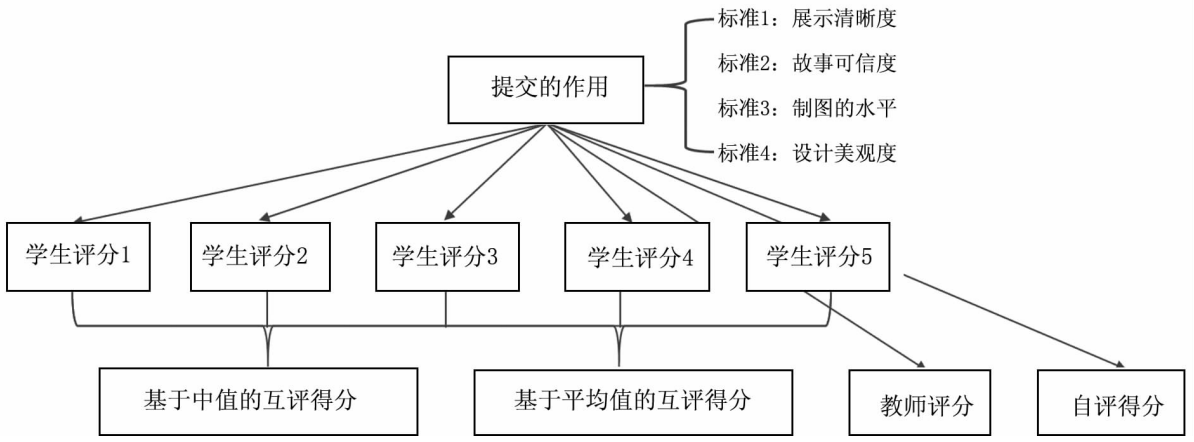

图1 互评作业的评价数据类别

估算方法。在该估算方法中,学生评分的差异及其交互作用被判定为测量误差。在 SPSS 统计分析软件中,ICC [1] 的计算通过可靠性分析模块中的单项随机组内相关系数计算功能实现。

学生互评效度的测量采用的是聚合效度,由学生互评得分和教师评分的相似度来判定,在统计分析中通过皮尔逊积矩相关系数(Pearson product-moment correlation coefficient)估算,通过 SPSS 统计分析软件中的皮尔逊双侧检验相关系数计算功能实现。本研究同时计算了基于中值和平均值的两种互评分数结果与教师评分结果的相关系数,由此比较不同互评分数统分计算方法的优劣。

本研究将学生的问卷数据从 Qualtrics 问卷平台下载后输入 SPSS 软件进行描述性分析,通过对 7 道问题得分的均值、频率和百分比的统计分析,可以从总体上把握学生对大规模开放在线学习环境中互评活动的态度,检验他们对互评活动潜在益处,如学习动机、社交存在感和高阶思维培养的认可度。

四、研究结果

(一)在线互评总分具备较高信度

本研究通过计算学生评分的第一类组内相关系数(ICC[1])测定在线互评得分的评分者间信度,相应的统计分析结果如表一所示。单个测量的 ICC[1] 系数反映了五个随机选择学生评分者对同一作业的评分一致性。该系数值为 0.262,表明单个学生对同一作业评分结果波动较大,评分者间信度较低,评分不可靠。但相对于单个测量,平均测量的 ICC[1]系数达到了 0.64,具备了较高的评测信度。该结果表明,如果互评分数不使用单个学生评分而是综合考量五个评分数据,如采用五个学生评分的均值,互评结果的可靠性将得到显著提升。通过针对 5 位评分者和 4 个评分标准的嵌套交叉随机情景分析,我们发现互评结果的概化系数(generalizability coefficient)维持 0.64 不变,测量标准误差仅发生了细微改变(从 0.272 变到 0.276)。该结果表明,互评误差来源主要来自评分者本身,通过修改评价量表和评分标准并不能进一步提高评分的信度。

表一 学生互评分数的第一类组内相关系数(N = 1825)

	组内相关性 ICC[1]	95%置信区间		使用真值 0 的 F 检验			
		下限	上限	F 值	df1	df2	Sig
单个测量	0.262	0.240	0.284	2.774	1824	7300	0
平均测量	0.640	0.613	0.665	2.774	1824	7300	0

本研究分别计算了使用 2 至 5 位学生评分者得到的互评结果的 ICC[1]系数(见表二),旨在探究评分者人数对互评分数信度的影响,并确定最佳评分者人数。这一结果与赵光洙等(Cho et al., 2006)的研究发现一致,评分者人数对互评结果的平均测量 ICC[1]系数有较大影响,评分者人数的增加能大幅提升互评结果的可靠性。ICC[1]系数在 0.4 – 0.7 之间通常被认为具备了中等评分者间信度(Dancey & Reidy, 2002),因此根据表二结果可以推断出,要使互评结果具备可接受的信度,至少需要三个学生评分者(ICC[1] > 0.4),而仅仅基于两个评分者判分的互评结果相对不可靠。

表二 学生评分者数目与(ICC[1])系数的关系

	5 个评分者	4 个评分者	3 个评分者	2 个评分者
单个测量	0.262	0.256	0.256	0.241
平均测量	0.640	0.580	0.508	0.389

(二)在线互评效度令人满意

Coursera 平台给出的基于中值的学生互评分数与教师评分有较强的正相关性(r = 0.619)(见表三)。该数据证明,Coursera 平台的学生互评模式能够提供接近于教师评阅准确度的评分结果。同时我们发现,如果不使用中值而使用平均值作为学生互评的最终得分,学生互评得分与教师得分的相关性反而会得到小幅提高,尽管提高的程度十分轻微(r = 0.669)。事实上,基于中值和基于平均值的互评分数本身十分相似,具有极高的相关性(r = 0.952)。

与学生互评相比,学生自评分数与教师评分的相关度较低(r = 0.341),表明学生对自己作业的评价与教师的专业评判出入较大。因此,我们认为学生自评结果聚合效度较低,不能成为合格的教师评分替代者。本研究同时考察了不同评分结果的均值,研究发现学生自评分数的均值最高(10.02),教师评分分数的均值最低(8.68),学生互评分数的均值居中(中值分为 9.19,平均值分为 9.1)。该结果表明学生倾向给自己的作业更高分数,而相比学生评分者,教师的评分标准总体上更加严苛。

表三 互评、自评与教师评分结果的皮尔逊相关系数

	教师评分	互评分数(中值)	互评分数(平均值)	自评分数
教师评分	1	0.619**	0.662**	0.341**
互评分数(中值)		1	0.952**	0.279**
互评分数(平均值)			1	0.464**
自评分数				1

** 相关值在 P=0.01(双侧检验)水平上显著

(三)参与互评活动能提升在线学习体验

表四总结了学生参与 MOOC 课程互评活动的态度和看法。总体来说,学生对于互评活动态度积极,63% 的学生认为参加互评活动有助于实现课程教学目标,即对空间思维能力的培养,70% 的学生推荐后续课程继续保留互评作业环节。学生总体上认可互评分数的公平性(占 62%)和他人反馈的价值(占 61%)。大部分学生认为参与互评活动提升了自己对课程的投入度(占 63%),并发展了自己的高阶思维能力,如审视与反思(占 72%),这一结果与已有研究发现一致。互评活动对社交存在感的影响在 7 个选项中得分最低,仅 57% 的学生认为参与互评活动增强了自己在线学习与其他同学的联系。当然,我们也注意到没有任何一个关于互评的问卷项获得了超过 80% 的积极评价,证明相当比例的学生对在线互评的效果持保留或否定态度。

五、讨论与反思

(一)学生互评能为大规模开放在线学习提供可靠有效的学习评价

尽管单个学生评分者的评分结果并不可靠,具有较低的评分者间信度(单个测量 ICC[1]=0.262),然而如果一项作业的最终互评分数是多个评分者评分的复合计算结果(如中值或均值),学生互评模式的信度较令人满意(平均测量 ICC[1]=0.64)。因此,Coursera 平台使用学生互评分数中值的做法值得借鉴,能够有效减少单个评分结果一致性低带来的影响,大幅提升最终评分结果的可靠性。学生互评分数与教师评分超过 0.6 的相关性系数证明了互评模式能够提供接近教师评阅准确度的评分结果,能为大规模开放在线学习活动提供相对准确有效的评价手段。相比之下,学生自评结果的效度不能令人满意,与教师评分的差别较大,分数相对偏高。考虑到学生在完成作业时往往已经有意或无意的检查过自己的作业,因此再次自评的环节并无必要。

当然,本研究对于学生互评信效度的支持证据基于"地图与地理空间革命"MOOC。该课程的学生受教育水平相对较高,80% 以上的学生都具备本科以上学历。考虑到互评作业在课程最后一周才发布,此时还活跃在课程学习中的学生已不到注册人数的 5%,属于学习投入度高、在线学习能力强的学习者(Waldrop, 2013;袁松鹤等,2014),基于该学生人群得出的关于互评信效度的结论有一定的局限性。相关结论能否适用于知识水平、学习能力和学习动机差异较大的大规模在线学生群体有待进一步研究验证。

(二)影响学生互评信效度的决定因素是评分者本身

和现有互评文献的结论一致,本研究也揭示了评分者人数是影响互评信度的重要因素,通过增加评分者人数就能够大幅提升互评结果的可靠性。而要使在线互评具备最基本的可靠性,至少要配备三名以上学生评分者。本研究同时也探索了统分方法

表四 学生对课后问卷中有关互评的相关问题回答情况

问卷项(1-5 从"完全不同意"到"完全同意")	N	1	2	3	4	5	积极评价(%)	均值
1. 参与学生互评的过程帮助我成长进步为一位具有空间思维的人	2121	83	137	562	950	389	63	3.67
2. 我从同伴那获得了关于我的地图作业比较公平的分数	1694	56	90	505	607	436	62	3.75
3. 同伴为我的地图作业提供的反馈十分有益	1719	59	106	509	677	368	61	3.69
4. 课程中参与互评的活动使我更加投入该课程学习	2027	97	165	489	780	496	63	3.70
5. 互评活动让我感觉和课上其他同学的联系更紧密了	2039	111	191	572	760	405	57	3.57
6. 互评活动为我提供了审视和反思课程内容的机会	2044	75	90	413	898	568	72	3.88
7. 我推荐后续课程继续保留使用学生互评的开放性作业	2185	111	126	420	732	796	70	3.90

对互评效度的影响,结果显示使用个体评分的中值和平均值作为最终分数对互评效度的影响不大。造成该现象的可能原因是学生评分的极值情况较少,即极少出现全部打最高分或最低分的评分者。从统计分析角度,可以使用中值的方法减少极值的不利影响,但更有效的方法可能还是从学生评分者本身入手,教育他们认真完成互评任务,同时建立相应的抽检、追责和奖惩制度。

本研究通过随机情景分析探索了评测误差的成因,结果显示误差来源主要是评分者本身而不是评价量规或标准。因此,对学生评分者进行互评培训和评分校验尤为重要。Coursera 平台的定标同行评估模式为学生提供了一种培训和校验的方式:学生在接受基本的培训后对数道样题进行评分,系统根据样题评分的准确性为每位学生设置权重,该权重决定了该学生在后续评分结果中所占的价值比重。然而,出于对时间、精力和可操作性的考虑,本研究没有在 MOOC 中安排相应的培训和校验步骤,该步骤对互评信效度的影响有待后续研究。

(三)学生总体上认可在线互评的评价模式

尽管不少学习者在新闻媒体和社交网站上表达了对大规模开放在线学习情境中使用学生互评手段的不满,并指出对其准确性、公平性和有效性的担忧(McEwen, 2013; Morrison, 2013; Neidlinger, 2013; Watters, 2012),本研究提供了与媒体舆论相悖的证据。问卷结果显示,60% 以上的 MOOC 学习者认为他们获得了公正的分数和有效的评价反馈,而持反对观点的学生不到 10%。事实上大部分学习者(70%)希望在后续课程中保留学生互评的任务,仅 11% 的学生希望将其移除。对于该现象的一个可能解释是获得了糟糕互评体验的学生更倾向在媒体上表达自己的不满,而他们的观点并不能代表整个在线学习者群体。我们也发现互评模式最被广泛认可的益处是对高阶思维能力,如审视和反思能力的培养。这一发现印证了布鲁姆(Bloom, 1956)对认知领域学习目标的分类理论。学生互评在认知目标分类中属于较高层次的认知活动,能够有效促进在线情境中有意义学习的发生。

综合来看,本研究基于一门 MOOC 中互评、自评和教师评分相关数据,对学生互评模式的信度、效度和相关影响因素和学生认可度进行探索与分析。研究结果表明传统面授课堂中广泛使用的学生互评模式也适用于大规模开放在线学习情境:在综合考量多名评分者评分结果的前提下,学生互评模式能够为学习者提供一个较为一致和可靠的最终得分。学生互评和教师评分结果的较高相关性也表明在线互评模式具备类似教师评阅的准确性。另一方面,学习者对参与互评活动总体上也持积极的态度,认可互评活动对获得反馈、课程投入度和高阶思维培养等方面的有益影响。值得注意的是,因为研究情境和评价数据的单一性,本研究对在线互评模式信效度的相关推论具有一定的局限性,对于影响互评效果因素的探索也不够深入,这些有待进一步探索。

[参考文献]

[1] Billington, H. L. (1997). Poster presentations and peer assessment: Novel forms of evaluation and assessment [J]. Journal of Biological Education, 31(3): 218-220.

[2] Bloom, B. S. (1956). Taxonomy of educational objectives, Handbook 1: Cognitive domain [M]. New York: David McKay.

[3] Bouzidi, L., & Jaillet, A. (2009). Can online peer assessment be trusted? [J]. Educational Technology & Society, 12 (4): 257-268.

[4] Brown, S., Race, P., & Rust, C. (1995). Using and experiencing assessment [A]. P. Knight (1995). Assessment for learning in higher education [C]. London: Kogan Page/SEDA: 75-85.

[5] Butcher, A. C., Stefani, L. A. J., & Tariq, V. N. (1995). Analysis of peer-, self- and staff-assessment in group project work [J]. Assessment in Education, 2(2): 165-185.

[6] Cheng, W., & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project [J]. Assessment & Evaluation in Higher Education, (24): 301-314.

[7] Cho, K., Schunn, C., & Wilson, R. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives [J]. Journal of Educational Psychology, 98 (4): 891-901.

[8] Dancey, C. P., & Reidy, J. (2002). Statistics without maths for psychology [M]. London: Prentice Hall.

[9] 邓鹂鸣, 岑粤(2010). 同伴互评反馈机制对中国学生二语写作能力发展的功效研究 [J]. 外语教学, 31(1): 59-63.

[10] Falchikov, N. (1994). Learning from peer feedback marking: student and teacher perspectives [A]. H. C. Foot, C. J. Howe, A. Anderson, A. K. Tolmie, & Warden D. A. (1994). Group and interactive learning [C]. Southampton and Boston: Computational Mechanics Publications: 411-416.

[11] Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teach-

er marks [J]. Review of Educational Research, 70 (3): 287-322.

[12] Freeman, M. (1995). Peer assessment by groups of group work [J]. Assessment and Evaluation in Higher Education, 20(3): 289-300.

[13] Fry, S. A. (1990). Implementation and evaluation of peer marking in higher education [J]. Assessment and Evaluation in Higher Education, 15(3): 177-189.

[14] 高地(2014). MOOC 热的冷思考——国际上对 MOOCs 课程教学六大问题的审思 [J]. 远程教育杂志, (2): 39-47.

[15] Gay, L. R., & Airasian, P. (2003). Educational research: Competencies for analysis and application [M]. Columbus, OH: Merrill, Prentice Hall.

[16] 顾小清, 胡艺龄, 蔡慧英(2013). MOOCs 的本土化诉求及其应对 [J]. 远程教育杂志, (5):3-11.

[17] Haaga, D. A. F. (1993). Peer review of term papers in graduate psychology courses [J]. Teaching of Psychology, 20 (1): 28-32.

[18] Hammond, K. R., & Kern, F. (1959). Teaching comprehensive medical care: A psychological study of a change in medical education [M]. Cambridge, MA: Harvard University Press.

[19] 韩冰(2009). 同伴互评在大学英语写作教学中的功效——基于写作流利性、复杂性及准确性的实证研究[J]. 教育理论与实践,(21): 40-42.

[20] Kaimann, R. A. (1974). The coincidence of student evaluation by professor and peer group using rank correlation [J]. The Journal of Educational Research, 68(4): 152-153.

[21] 康叶钦 (2014). 在线教育的"后 MOOC 时代"——SPOC 解析 [J]. 清华大学教育研究, 35(1): 85-93.

[22] Kauza, J. (2014). MOOC assigned [A]. S. D. Krause & C. Lowe (Eds.) (2014). More questions than answers: Scratching at the surface of MOOCs in higher education [C]. Anderson, SC: Parlor Press: 105-113.

[23] Korman, M., & Stubblefield, R. L. (1971). Medical school evaluation and internship performance [J]. Journal of Medical Education,(46): 670-673.

[24] Lu, J., & Law, N. (2012). Online peer assessment: Effects of cognitive and affective feedback [J]. Instructional Science, 40 (2): 257-275.

[25] Magin, D. (1993). Should student peer ratings be used as part of summative assessment? [J]. Higher Education Research and Development, (16): 537-542.

[26] Magin, D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work [J]. Studies in Higher Education, 26 (1): 53-63.

[27] Marcoulides, G. A., & Simkin, M. G. (1995). The consistency of peer review in student writing projects [J]. Journal of Education for Business, (70): 220-223.

[28] McEwen, K. (2013). Getting to know Coursera: Peer assessments [EB/OL]. [2016-7-22]. http://cft. vanderbilt. edu/2013/01/getting-to-know-coursera-peer-assessments/.

[29] McGarr, O., & Clifford, A. M. (2013). 'Just enough to make you take it seriously': Exploring students' attitudes towards peer assessment [J]. Higher education, 65(6): 677-693.

[30] Mehaffy, G. L. (2012). Challenge and change [J]. Educause Review, 47(7381): 25-42.

[31] Miller, P. J. (2003). The effect of scoring criteria specificity on peer and self-assessment [J]. Assessment & Evaluation in Higher Education, 28(4): 383-394.

[32] Mok, J. (2011). A case study of students´ perceptions of peer assessment in Hong Kong [J]. ELT journal, 65(3): 230-239.

[33] Morrison, D. (2013). Why and when peer grading is effective for open and online learning [EB/OL]. [2016-8-12]. http://onlinelearninginsights. wordpress. com/2013/03/09/why-and-when-peer-grading-is-effective-for-open-and-online-learning/.

[34] Mowl, G., & Pain, R. (1995). Using self and peer assessment to improve students' essay writing-a case study from geography [J]. Innovations in Education and Training International, 32: 324-335.

[35] Neidlinger, J. (2013). Does peer grading of essays really work in a Coursera online class? [EB/OL]. [2016-7-31]. http://lone-prairie. net/peer-grading-coursera/.

[36] Oldfield, K. A., & Macalpine, M. K. (1995). Peer and self-assessment at tertiary level - an experimental report [J]. Assessment and Evaluation in Higher Education, 20(1): 125-131.

[37] Orpen, C. (1982). Student versus lecturer assessment of learning: A research note [J]. Higher Education,(11): 567-572.

[38] Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned models of peer assessment in MOOCs [EB/OL]. [2016-8-15]. http://web. stanford. edu/ ~ cpiech/bio/papers/tuningPeerGrading. pdf.

[39] Race, P. (1998). Practical pointers on peer-assessment [A]. Brown S. (1998). Peer Assessment in Practice [C]. Birmingham, SEDA: 113-122.

[40] Rees, J. (2013). Peer Grading Can't Work [EB/OL]. [2016-7-31]. http://www. insidehighered. com/views/2013/03/05/essays-flaws-peer-grading-moocs.

[41] Sadler, P., & Good, E. (2006). The impact of self- and peer-grading on student learning [J]. Educational Assessment, 11 (1): 1-31.

[42] Stefani, L. A. J. (1994). Peer, self and tutor assessment: Relative reliabilities [J]. Studies in Higher Education, 19(1): 69-75.

[43] Strijbos, J. W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? [J]. Learning and instruction, 20(4): 291-303.

[44] Strijbos, J. W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments [J]. Learning and Instruction, 20(4): 265-269.

[45] Topping, K. J. (2009). Peer assessment [J]. Theory into

Practice, 48(1): 20-27.

[46] Vu, T. T., & Dall'Alba, G. (2007). Students' experience of peer assessment in a professional course [J]. Assessment & Evaluation in Higher Education, 32(5): 541-556.

[47] Waldrop, M. (2013). Campus 2.0 [J]. Nature, (495): 160-163.

[48] Watters, A. (2012). The problems with peer grading in coursera [EB/OL]. [2016-7-22]. http://www.insidehighered.com/blogs/hack-higher-education/problems-peer-grading-coursera.

[49] Wen, M. L., Tsai, C. C., & Chang, C. Y. (2006). Attitudes towards peer assessment: A comparison of the perspectives of pre-service and in-service teachers [J]. Innovations in Education and Teaching International, 43(1): 83-92.

[50] 熊瑶,孙开慧(2016). 慕课学生互评差纠方法及其比较[J]. 中国考试,(1):7-15.

[51] 袁松鹤, 刘选 (2014). 中国大学 MOOC 实践现状及共有问题——来自中国大学 MOOC 实践报告 [J]. 现代远程教育研究, (4): 3-12.

[52] Zhang, B., Johnston, L., & Kilic, G. B. (2008). Assessing the reliability of self- and peer rating in student group work [J]. Assessment & Evaluation in Higher Education, 33(3): 329-340.

(编辑:李学书)

An Empirical Study on the Effect of Peer Assessment in Massive Open Online Learning

LUO Heng[1], ZUO Mingzhang[1] & Anthony Robinson[2]

(1. *School of Educational Information and Technology, Central China Normal University, Wuhan* 430079, *China*;
2. *Department of Geology, Pennsylvania State University, University Park* 16802, *U. S. A.*)

Abstract: *Peer Assessment has the potential to overcome the limitations of teacher-grading and machine-grading, and therefore has become an important assessment method in the context of massive open online learning. However, there is a lack of empirical research that investigate the accuracy and effects of online peer assessment. To address such research need, this study analyzed the empirical data of peer assessment, teacher-grading and self-grading from a massive open online course (MOOC), and provided preliminary findings regarding the reliability, validity, influencing factors and perceived benefits of peer assessment for this specific context. The results show that although peer assessment suffers from inter-rater reliability, the composite score from all student graders can provide a reliable and consistent final result. The correlation coefficient of* 0.619 *indicates a high convergent validity for peer assessment. In addition, this study also explored the influence of factors such as training and calibration, number of graders and score calculation methods. Moreover, the post-course survey results reveal that online learners in general have positive attitude towards peer assessment, acknowledging benefits such as feedback acquisition, higher engagement and development of higher-order thinking. The findings of this study are expected to inform the revision and improvement of assessment models for massive open online learning.*

Key words: *peer assessment; inter-rater reliability; convergent validity; massive open online learning*