DOI:10.13811/j.cnki.eer.2019.06.008

MOOC同伴作业互评中反思意识与学习成效的关系研究

汪 琼， 欧阳嘉煜， 范逸洲

（北京大学 教育学院，北京 100871）

[摘 要] 同伴互评是MOOC中经常采用的一种教学活动。为了帮助学习者自信地对他人作业做出客观公正的评价并促进自我反思，课程团队会提供作业反思框架，即评价量规，希望学习者在使用评价量规撰写评语的过程中不断加深对课程教学目标的理解，通过观摩评议同伴作业促进自我反思。但教学用意未必所有学生都能够领会并落实，MOOC中有多少学习者在作业互评过程中具有反思意识、反思意识程度如何、反思意识与学业成效之间是否存在关联，是本研究感兴趣的问题。通过对“教师如何做研究”MOOC课程中79287条同伴互评数据进行深入分析后发现：在自然无干预的状态下，大部分学习者有一定的反思意识，但撰写的评语质量还有待提升；成绩合格者的评语长度和评语质量与学习成效呈现显著正相关，这表明同伴互评的评语长度及质量可以作为学习成效的预测指标。研究在使用定量方法对MOOC学习者评语数据进行教学意义分析方面作了一些方法尝试，可为后续研究提供参考。

[关键词] MOOC；同伴互评；反思意识；评语字数；评语质量；学习成效

[中图分类号] G434 [文献标志码] A

[作者简介] 汪琼(1965—)，女，安徽合肥人。教授，博士，主要从事教学设计研究。E-mail：wangqiong@pku.edu.cn。

一、问题的提出

同伴作业互评是MOOC中常用的一种社会建构学习活动，学习者需要根据课程团队提供的评价量规，对多个同伴的作业进行批改。在这个过程中，学习者不仅可以加深对教学要求的理解，促进自我反思[1]，还可以因观摩他人作业而取长补短，丰富观点和见识。

已有研究表明，学习者参与同伴互评能够提高他们的学习成绩、激发学习动机[2-4]。例如：“教师如何做研究”这门MOOC课程中就有学习者表示非常喜欢这个活动，认为“原以为自己已经做得很好了，却看到了别人更出色的作业”。对于被评价者来说，高质量且与被评价者相关的评语对其学习起到正向促进作用[5]，对于评价者来说，写评语的过程可以提升其批判性思考等高阶思维能力[6]。

虽然作业评语是“同伴作业互评”活动中提升学生学习质量的关键要素，但并不是所有学习者都能够意识到认真写评语的意义，所撰写的评语也并不都是有针对性的。在一项对3034条同伴互评评语进行文本编码分析的研究中，研究者发现：多数学习者撰写的评语内容都不涉及对作业的修改建议，即使提出了修改建议，这些建议通常也是不具体的或者难以操作[7]。

在“教师如何做研究”MOOC课程中也有学习者反映：同伴并没有根据评价量规对作业进行评价；评语内容单薄，只有几个字，对后续修改无指导价值；等等。我们随机抽取了300条“教师如何做研究”MOOC课程的同伴作业互评评语，按照字数由少至多的顺序排序后发现：学习者所写的评语长度不一，有些评语寥寥几字(如“很好”“厉害”“很有收获”)，有些评语长

基金项目：教育部—中国移动科研基金项目“慕课教学效果与慕课的教育资源质量评价体系及应用研究”(项目编号：MCM20170503)

篇大论(字数高达上千字);评语质量差异也很大,部分评语能够根据评价量规给出具有建设性的修改意见,但也有部分评语看不出与作业评价量规的关联。可见,并不是所有 MOOC 学习者都能够意识到评价量规实际提供了反思的框架,也不是所有学习者都能够根据评价量规撰写评语。那么,有多少 MOOC 学习者在参与同伴互评过程中具有反思意识或进行了审辩思考?由评语表达出的反思意识的强弱与学习成效之间是否有相关性?对上述问题的回答,将有助于增进我们对 MOOC 学习者学习特点的认识,相关发现也可以用于优化 MOOC 作业互评活动设计。

二、研究设计

本研究希望通过对 MOOC 作业互评活动中作业评语的研究,确定 MOOC 学习者在参与同伴作业互评过程中的反思意识强度,研究分析反思意识强弱与学习成效是否存在关系。

(一)研究数据来源

为保障本研究数据分析的有效性,本研究在样本选择上有以下几点考虑:一是同伴互评任务要具有一定复杂度,如果同伴互评任务过于简单,那么学习者撰写的评语内容和修改建议极易雷同;二是同伴互评的评价量规不能过于单一,否则学习者撰写的评语同质性也会很高,同质性会大大压缩评语的可分析空间;三是为了排除学习者对评价内容理解程度的差异对学习者所撰写的评语内容的影响,所选择分析的同伴互评活动最好是在课程后期进行的,一般来说,参与课程最后一周同伴互评的学习者基本都坚持学习了课程的所有内容,这部分学习者已经了解并掌握了课程的知识点,他们对于评价量规的熟悉和理解程度可视为同一水平。

根据上述考虑,本研究将研究样本确立为一门面向大中小学教师的研究方法培训 MOOC“教师如何做研究”,该课程共有五周教学内容。学习者在学完五周所有内容后需要根据模板撰写一份研究申请报告,作为作业提交并参与同伴作业互评,每个人至少要评 6 份作业。评价量规共有 6 个维度,分别是研究题目、研究关键词、研究背景、研究内容、研究过程和研究创新点和成果。表 1 列举了该作业同伴互评活动的评价量规。学习者根据评价量规给出每个维度的得分后,平台将自动计算总分(每个维度去掉最高分和最低分后取平均分,然后计算各个维度的平均分之和获得作业总分,满分 15 分),同时,学习者还需要在文本框内给出相应的评语,课程团队要求学习者能够根据评价量规进行评价,但平台本身对学习者提交的评语内容和评语字数无任何限制。

表 1　　同伴互评活动评价量规

评价维度	不合格	合　格	良　好	优　秀
	0 分	1 分	2 分	3 分
研究题目(2 分)	研究题目像口号,或者不是教师能够完成的研究,或者不是研究,是做事	需要修改	基本合格,简明、直接、清晰	
研究关键词(2 分)	没写关键词,或者 3 个关键词都不合适	只有 1 个关键词合适	有 2 个以上关键词合适	
研究背景(3 分)	没有说明为什么要做研究,或者这个研究不太可能取得什么价值	研究有些价值,但没有相关研究介绍	研究较有价值,且介绍了相关的理论和相关研究,但参考文献不够典型、不够多	研究问题很重要,相关研究分析清晰,参考文献典型且相关
研究内容(3 分)	没有提出研究问题,只说要做什么,没有说研究什么	提出了研究问题,对研究问题的分解不合理,逻辑性不强	提出了研究问题,且绘制了清晰的研究内容之间的关系图,或用文字步步论证了逻辑关系,没说重点难点,或说得不太对	提出了研究问题,且绘制了清晰的研究内容之间的关系图,或用文字步步论证了逻辑关系,对研究的重点和难点有清晰的认识和合理的判断
研究过程(3 分)	研究过程书写过于简单,看不出会怎么做,是否有条件做	部分研究问题选择的研究方法不合适,没有说明研究需要的条件	能够为每个研究问题选择合适的研究方法,说明了之前的研究基础以及具备本研究所需要的条件,没有说明信效度,或者说的不太对	能够为每个研究问题选择合适的研究方法,且说明了之前的研究基础以及具备本研究所需要的条件,研究过程完整,考虑了信效度
研究创新点和成果(2 分)	没有讨论研究可能的突破点	创新点阐述牵强,成果形态单一	创新点分析有道理,对成果可能的应用前景有多种设计	

本研究收集了该课程第一至第十一期的所有同伴作业互评的实际数据，即每位学员打出的分数，在此基础上得出每份作业的实际互评成绩作为本研究的分析数据，使用 Rstudio 工具进行数据分析。数据字段包括评价者 ID、被评价者 ID、课程 ID、学期 ID、作业 ID、每个评价维度的得分、评语、课程成绩等，共计 10184 人次参与了同伴互评环节，产生了 79287 条有效评语，相当于人均提供了 7.78 条评语，即人均批改了近 8 份作业，“超额”完成了互评作业任务。这也从一个侧面说明 MOOC 学习者是很愿意看同伴作业的。

(二)研究变量的操作化定义

本研究主要涉及两个研究变量：反思意识和学习成效。

1. 反思意识

“反思意识”是指学习者在进行作业互评的时候，是否有意识地运用课程所教的知识(分析视角)来分析评价他人作业。如果一个 MOOC 学习者在作业互评时能够看出被评作业的优点，就有可能学习这个优点，并将其应用到自己的作业改进中。由于作业互评量规就是课程团队希望学员采用的分析框架，浓缩了课程所教知识的维度，所以这里“反思意识”操作化定义为“学员在进行作业互评的时候，有无应用作业互评量规的意识”。

学习者反思意识的强度操作化定义为两部分：评语长度和评语质量。

“评语长度”是指评语的字数，包括所有中文字符、英文字符以及标点符号。如果评语太短，信息量太少，有可能是学习者没有仔细分析别人的作业，只是凭印象下了断言。而写得多的评语往往会基于评价量规对作业的一个或几个方面进行有针对性的分析，展示出评论者的思考。MOOC 平台中会记录所有的评语，很容易计算出评语长度。本研究的一个产出目标是：要发现自然状态下学习者愿意写且有意义的评语长度，以便在未来的互评活动中对评语字数提出合理的要求。

考虑到对上万条评语进行质量分析的操作可行性，这里将“评语质量”转化定义为“评语和评价量规之间的吻合度”。因为作业评价量规体现了教学团队对作业应达到水平的要求，它也是一种反思支架，帮助学习者结构化地运用课程所学知识对他人作业做出评价，因此，运用评价量规进行作业评价的过程，就是一个审辩和反思的过程。按照评价量规维度将思考的过程或结论写下来的评语往往针对性较强，也便于被被评论者理解和接受。如果学习者撰写的评语和评价量规之间毫无联系，那说明学习者可能并没有关注到评价量规，其评语质量难以保证。

本研究采用建立关键词词典与评语进行字符匹配的方式来确定评语和评价量规之间的吻合度。这里我们以评价量规中的维度 1(研究题目)为例说明词典的建立和匹配过程，如图 1 所示。首先，由两名研究者根据课程已有的评价量规内容进行关键词提取，这一步主要是提取评价量规中有意义的实词，对于一些副词、语气词等不纳入其中。尽管我们给出了非常详细的评价量规，但每个人对于量规的理解和表达仍然存在差异，不同的词汇可能表达了相同的含义。因此，为了提高关键词词典的全面性与精准性，第二步是从所有评语中随机抽取了 300 份评语，根据学习者撰写的真实评语对关键词词典进行补充，尽可能保证关键词词典囊括了量规中词汇的所有变体。最后，评价量规维度 1 建立的关键词词典如 “研究题目、题目、口号、不是研究、做事、简明、清晰选题、合格、过大、直接符合、要求、修改 (提出、问题、课题、立题、标题)”，其中“()”内的关键词是根据随机抽取的 300 条评语内容所补充的关键词。只要学习者撰写的评语中与词典中任何一个关键词匹配，我们就认为这条评语和维度 1 匹配成功，标识为 1，如果没有任何一个关键词匹配成功，则标识为 0。

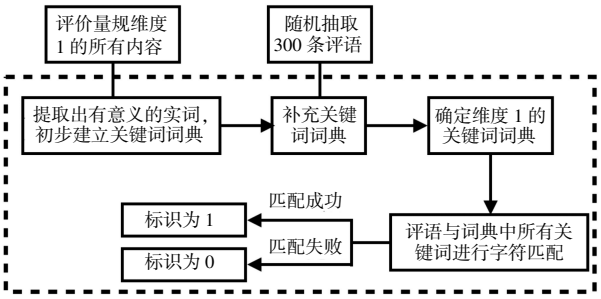


图 1 关键词词典的建立与字符匹配流程举例

根据上述匹配原则，我们按照评价维度 1 到维度 6的顺序逐一开始匹配，最终形成 79287 条二值序列表征的数据记录，共 64 种类型(如 000000、010000、001000、000100、000010、000001、100000、110000、111000、111100、111110、111111)，以此表示所有评语与评价量规之间的吻合度。例如：110000 就表示该评语内容和评价量规中的前两个维度匹配，与后四个维度不匹配。

2. 学习成效

本研究用学习者最后的课程成绩作为学习成效的操作性定义。这是因为本文研究的是自然无干预状态下 MOOC 学习者在进行作业同伴互评时的自然表现，我们的基本假设是：在自然无干预状态下，MOOC

学习者同伴互评活动中的表现是其一贯学习行为和态度的体现，与其学习成效之间会有一定的关联,课程成绩就是学习成效的结果体现。

本研究曾考虑过用学习者提交互评的作业成绩作为学习成效,但后来发现影响 MOOC 学习者作业成绩的因素很多，一方面 MOOC 学习者程度差异很大,互评成绩未必能够反映学生作业的真实水平，另一方面 MOOC 学习者有给同伴作业打高分的现象，图 2 为课程作业的成绩与给分人数的统计,可以看出:确实有给高分的倾向,所以单一的 MOOC 互评作业成绩作为学习者的 MOOC 学习成效不太具备解释性。

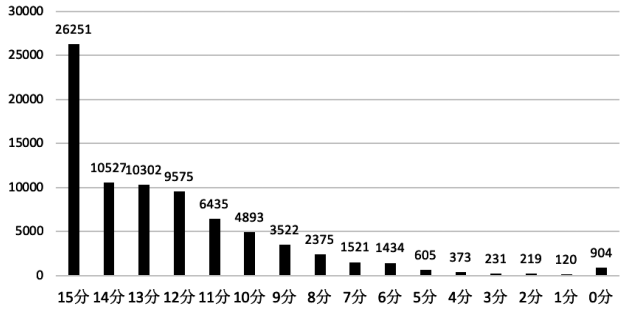


图 2 作业分数(横轴)与给分人数(纵轴)直方图

那么这是否意味着 MOOC 同伴互评活动不可用呢？对于“教师如何做研究”这门课程,互评作业的教学目的有两个:一是写作业的人要仔细思考如何写一份合格的项目申请书;二是评作业的人,要从至少六份申请书中获得写研究申请书的经验和教训。只要按照要求写作业,根据量规参与互评作业活动,就会有收获,也就达到了教学目标。这个作业相当于面授环境下的平时作业,占分只有 15%,作为可以帮助学习者开拓研究思路的教学活动,即使学习者给了不一定吻合作业水平的高分,也是在可接受范围内的。

“教师如何做研究”这门课程的成绩由四部分构成,分别是课程讨论(15%)、模块测验(50%)、课题申请书作业(15%)和期末考试(20%),总分 100 分,60 分以上为合格,80 分以上为优秀。

由于平台所提供的 11 期课程数据不完整的原因，我们只从这 10184 人次参与互评的学习者中匹配到 4407 位学习者的成绩,因此,本研究有关评语质量与学习成效的相关研究只是对这 4407 个样本数据的分析。

三、研究发现

(一)有关评语长度的发现

由于 99%的评语字数在 200 字以内，这里只对 200 字以内的评语进行统计分析。

1. 评语长度过短或过长都可能没有反思意识

将所收集的所有互评评语按长度归类统计,得到这门课程所有互评评语的评语长度分布图,如图 3 所示。

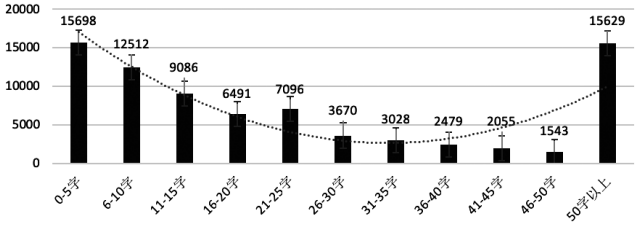


图 3　评语字数分布图

由图 3 可见,评语字数在五个字以内的记录条数达 15698,约占总数据的 20%,这类评语大多为态度的简单表达,缺乏具体的反馈信息,如“赞”“厉害”“非常好”等。这表明学习者在做同伴互评的时候很容易采用“评判”的姿态,而不是“评论”的方式。虽然这类评语初步反映出评价者对作业的认可程度,但当被评价者接收到这类评语时并不能获得对改进作业有帮助的建议,也就达不到同伴互评活动的真正目的。

这门课程 11 期中大于 500 字的评语有 28 条,其中 15 条来自同一个人，上传的是某篇文章，还有 5 条是不相干的文字摘录,只有 7 条是认真撰写的,占比 25%。这说明评语字数多不意味着学习者认真做互评了,还是有一些人在做“欺骗系统”的事情却未能被检举处罚。换句话说,作业互评的评语并不是字数越多就表示学习者互评越认真。这也提醒我们:课程团队在设计互评活动的时候，可以建议评语的最少字数和最多字数，这可能会有助于产生有内容价值的评语。

2. 成绩好的人所得到的长评语较多

将学习者的课程成绩和其作业所得到的评语字数做成图 4,可以看出:长评语基本集中在图的上方,即 80 分以上学习者获得了较多的长评语。

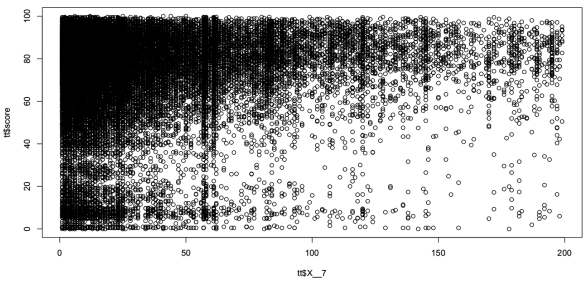


图 4　课程成绩(纵轴)与作业评语字数(横轴)

图 5 为作业分数与评语字数直方图,也是得分高的作业收获的长评语较多。从评阅者角度,对于不太合格的作业,往往可以简单明了地给出评价,而对于符合一定要求的作业,因为有值得互相学习的地方或

稍加精进之处，自然也就会多写一些评语。对 300 份抽样评语的阅读也证实了这一点，成绩好的学习者所撰写的作业（项目申请书）往往能够打动评阅人，引发共鸣和讨论。

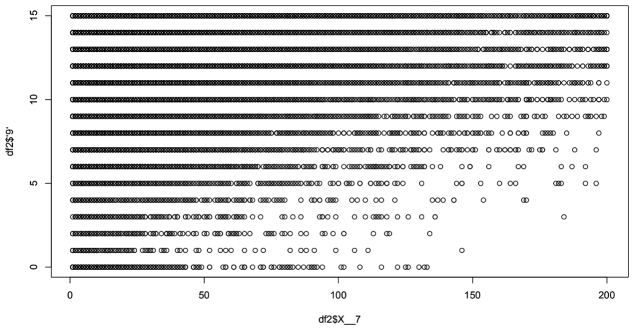


图 5 作业成绩（纵轴）与评语字数（横轴）

表 2 是三个得分区间分别对应的长评语条数及占比，从表中可以看出，得分越高的作业得到的长评语数更多，所占比例也越高。

表 2　不同得分区间的长评语（字数大于 50）条数

作业得分区间	>50 字的评语条数	各个区间评语条数	>50 字评语条数占比
0~5 分	340	2452	13.9%
6~10 分	2348	13745	17.1%
11~15 分	12941	63090	20.5%

这个发现说明在 MOOC 课程中，认真努力的人会得到更多的肯定，但这并不意味着成绩较低的学习者得不到应有的帮助。表 3 分析了每个作业分数对应的所有评语的长度均值和四分位值，可以看出各分值的作业得到的平均评语字数是差不多的。

表 3　不同作业分数的作业评语字数描述性统计

作业分数	评语字数均值	第一四分位数	第三四分位数
15	31.75	5	42
14	31.82	8	42
13	30.74	8	40
12	29.72	7	37
11	30.05	8	37
10	30.37	8	38
9	28.73	8	38
8	30.48	9	37
7	31.77	9	40
6	26.68	7	33
5	33.72	9	40.25
4	31.37	9	41
3	32.03	10	40
2	31.32	9	37
1	25.73	7	28.25
0	16.06	5	18.25

为了确定同一学习者所给评语长度之间有多大差异，绘制了图 6，其中横坐标为某学习者所给出的最长评语字数与最短评语字数之差，纵坐标为有这么多字数差异的学习者个数。从这张图可以看出，评语字数差异不大的人很少，并且我们进一步统计了撰写的所有评语均在 5 字、10 字和 20 字以内的学习者人数，结果分别为 4 人（0.04%）、28 人（0.3%）和 263 人（2.6%），相较于整体参与同伴互评的人数来说，这个数值所占比例极小。综上可以说明，参与作业互评的学习者在做这个活动的时候对每份作业还是有区别对待的，并不是敷衍的。

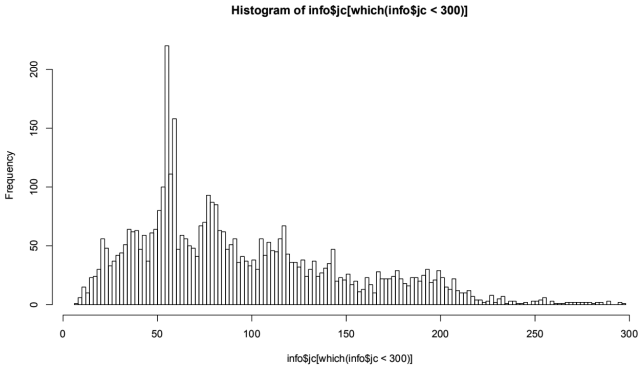


图 6　评语字数极差直方分布图

这几个数据在一起，证实了这门 MOOC 的作业互评活动是一个健康的学习生态，参与互评的学习者在这个过程中是认真参与的，不同长度的评语证实是有反思活动发生的，只是反思意识的意愿和强弱可能不同。这也在情理之中，因为并不是所批阅的作业都是认真完成的，也不是都能引发共鸣的，但只要互评者有认真批改的行为，我们都可以认为这门 MOOC 作业互评活动还是达到了教学目的的。

3. 评语长度有习惯规律

图 4 中有几条明显的直线，这意味着有些评语的长度不随所评作业的质量表现而变化，体现了学习者评语的自然长度，即写到多少字自认为说清楚了就行。从图上看这几条黑线对应的字数间距在 20~25 字左右。考虑到这个作业评分标准有 6 个维度，这可能是构成这些直线区间的原因。由于有些评语只评价了其中的若干维度，所以评语字数长短不一。与图 3 相参照，可以看出大多数的评语字数在 125 字以内，这也就是说，如果要求评语要有一定长度，最好不要要求字数超过 125 字，这是一般学习者书写评语的字数上线。阅读抽样的评语也发现，字数在 26 字左右的评语可以写出一点有针对性的文字了，即图 5 中第一条黑线的位置。这也就是说，如果要求评语要有实质性的内容，25 字可以设定为评语的最低字数要求。

(二)学习成效和评语长度呈现显著正相关

上面分析了这门课程作业互评评语长度的一些基本情况,可以看出很多评语字数并不多。假设写评语的过程就是运用评分量规进行批判性思考,也许成绩好的学生会写出更长的评语。

为了检验这个假设,我们对学习者的课程成绩与评语字数做了以下的处理:将所有学习者的课程成绩四舍五入取整后,将课程成绩相同且人数超过 10 个的学习者看作一个集合,并把课程成绩作为横坐标取值,将集合中所有学习者书写的评语字数的均值作为纵坐标取值,两个数值确立散点图中的一个点,代表获得该课程成绩的学习者撰写的评语字数平均值。例如:课程成绩为 76 分的学员共 15 个,那横坐标取值为 76(分),纵坐标取值则是 15 个学员所撰写的所有评语字数的平均值;而如果 78 分的学习者只有 1 个,由于样本数量太少,不具有一定的代表性,就舍弃这个样本。最后绘制出评语字数与学习者成绩对应的散点图,如图 7 所示。

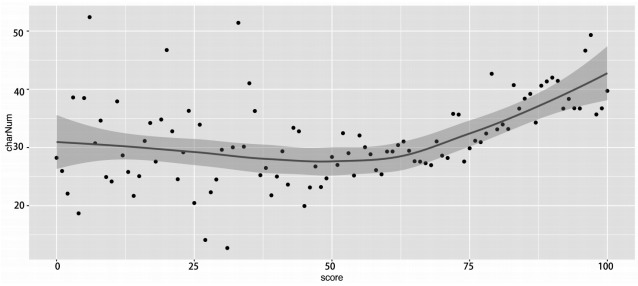


图 7　评语长度和所有学习者学业表现的散点图

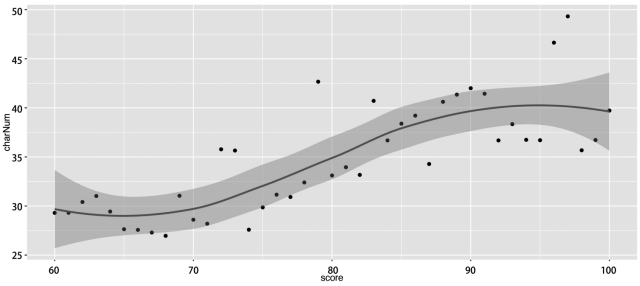


图 8　评语长度与合格学习者学业表现的散点图

如图 7 所示,成绩不合格的学习者评语字数分布离散,与拟和线之间的距离较大,数据不稳定,参考性较低。但成绩合格的学习者的评语字数则基本分布在拟合线两侧,稳定性较高,可进一步分析,因此,我们将横坐标取值区间设置为[60,100],即绘制评语长度与合格学习者学业表现对应的散点图(如图 8 所示),计算显示:评语字数与成绩之间呈现显著正相关(r=0.36,p=0.00016<0.001),也就是说,成绩越高的学习者所撰写的评语字数相对更多。这里成绩在 90~100 区间内的学习者撰写的平均评语字数趋于稳定(40 字),并没有持续上升,这表明在自然状态下,MOOC 学习者书写互评评语的长度确实有一个心理接受阈值。

(三)有关评语质量的发现

在研究变量的操作化定义中提到,本研究将“评语质量”转化定义为“评语和评价量规之间的吻合度”,并建立了作业六个评价维度的关键词词典,再采用与评语进行字符匹配的方式确定了每条评语和每条评价量规之间的吻合度编码(二值序列)。其基本假设是:如果学习者在写评语的时候能够借鉴作业评价量规的要求,就代表学习者在写评语的时候在思考作业(项目申请书)的某个方面应该达到的水准,即表现出一定的反思意识,且这样写出的评语会有一定的针对性,对被评者来说,这样的评语也会有参考价值,属于有质量的评语。作业评价量规共有 6 个维度,分别是研究题目、研究关键词、研究背景、研究内容、研究过程和研究创新点与成果。以下是对评语质量的研究发现:

1. 评价量规对学习者起到了一定的引导作用

基于前面对评语质量的分析思路,在获得每个评语的二值序列后,我们对所有的二值序列进行归并处理,根据“1”出现的次数划分出 7 个匹配维度,字段名为 match,如果一条二值序列为 000000,那么 match 的值标记为 0,如果某序列为 010000,则 match 的值标记为 1,以此类推,我们得到 79827 条取值范围是 0、1、2、3、4、5、6 的数据记录,也就是说,如果某一条评语的 match 值为 5,则说明该评语内容和评价量规中的 5 个维度是吻合的。将所有评语进行上述编码后分类统计得到图 9。

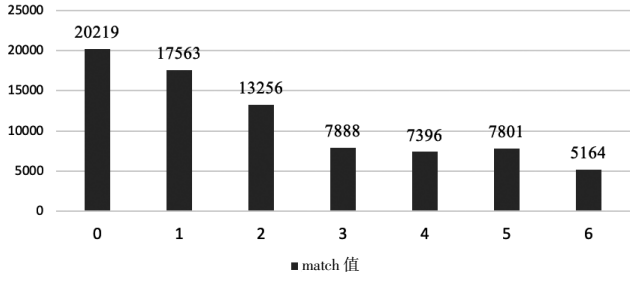


图 9　评语条数与所含评价量规个数

从图中可以看出,约 25.5%的评语和评价量规之间没有吻合之处,22.2%的评语只谈及了某一个评价维度。虽然我们不能说某条评语涉及的评价维度多,就一定比只涉及一个维度的评语质量高。但是,如果评语中没有出现六个评价维度的任一核心关键词,也确实不能算作是有质量的评语。因为,评价量规是课程团队提供给学习者参与同伴互评活动的反思框架,如果学习者撰写的评语和评价量规之间毫无关联,那么从一定程度上可以说明该学习者在写这条评语时

没有体现出反思意识。

对施予评价者的分析发现：只有一位学习者撰写的所有评语都是和评价量规无关的，可以忽略不计，其他学习者的评语还是与评价量规有关联的。所有学习者撰写的评语中平均 5.8 条（占比 74.6%）评语都是和评价量规相关，match 值的均值为 2.8。进一步分析每位学习者撰写的评语质量 match 值极差发现（如图 10 所示），大多数学习者撰写的多份评语质量之间具有差异性，这说明学习者在写评语的时候或多或少都有一些反思意识，评价量规对学习者起到了一定的引导作用，评语质量是有一定保证的。

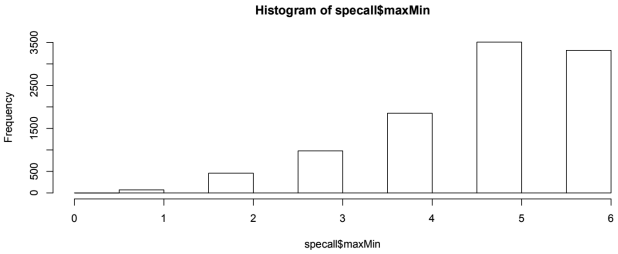


图 10　match 值极差图

2. 一成的学习者得到的作业评语对其改进作业无帮助

从图 9 可以看出，有 20219 条评语与评价量规没有关系，match 值为 0。进一步分析发现：共有 4473 名学习者（近一半）的作业得到的评语中有 match 值为 0 的评语，其中，991 名学习者（占总被评价人数的 11.2%）的作业得到的所有评语 match 值都为 0，这些评语长度的均值是 4.249 个字，第一四分位数是 2 个字，第三四分位数是 4 个字，即这些 match 值为 0 的评语基本上字数都很少。从这 991 人作业所获得的成绩来看，均值 13.17，也就是说，所得到的评语字数虽少，但并不是因为他们的作业做得不够好，或者说，评语字数虽少，且与评价量规无关，但并不影响学习者的作业成绩，只是对学习者改进作业无帮助，这是有改善空间的。

3. 该课程学习者对研究题目这一维度关注较多

图 11 为这 7 万多条评语中涉及每个维度的评语条数直方图，可以看出，学习者在互评的时候比较关注“研究题目”“研究过程”“研究方法”这三个项目申请书中最核心的内容。其中，对选题尤其在意，约 49% 的评语都谈及了研究题目这一维度，位居 6 个作业评价维度之最。这样的结果也符合我们的预期，因为对于一项研究来说，研究选题决定整个研究的定位，如果选题不合适，那么后续研究也很难继续开展。在所有评语中，对于关键词的评价是 6 个维度中所占比例最少的，仅有 18897 条评语涉及（约占 23.8%），这主要是因为课程团队在评价量规中阐明了关键词的选取只需要满足两个基本要求，一是个数，二是合适与否，而我们在浏览学习者作业的过程中也发现多数申请书在关键词的选取上都能符合上述要求，没有太多需要进一步修改的地方，所以评语也较少涉及对关键词的评价。

有 31.1%的评语只对研究题目维度进行了评论，这也许是因为评价者打开一份作业首先看到的是研究题目，课程团队给出的评价量规中第一个也是研究题目，这种先入为主的顺序也许会使得学习者撰写的评语中多涉及研究题目。但这目前只是推测，还需要后续研究加以验证。

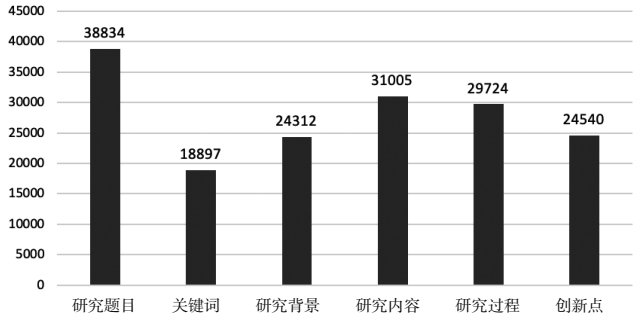
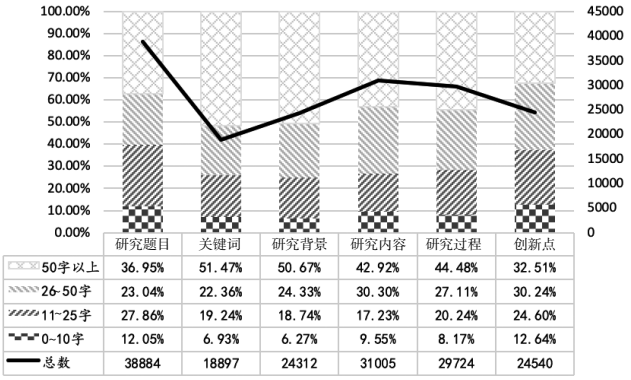


图 11　评价量规维度与评语内容匹配图

4. 评语太短，质量往往不高

为了进一步验证不同评语长度区间的评语质量有何特点，我们将所有评语的评语长度划分为四个数据量大致相同的区间，分别是 0~10 字（28210 条）、11~25 字（22673 条）、26~50 字（12775 条）和 50 字以上（15629 条），并将所有评语及每个评语长度区间内的评语分别与上文建立的关键词词典进行字符串匹配，结果如图 12 所示。



	研究题目	关键词	研究背景	研究内容	研究过程	创新点
50字以上	36.95%	51.47%	50.67%	42.92%	44.48%	32.51%
26-50字	23.04%	22.36%	24.33%	30.30%	27.11%	30.24%
11~25字	27.86%	19.24%	18.74%	17.23%	20.24%	24.60%
0~10字	12.05%	6.93%	6.27%	9.55%	8.17%	12.64%
总数	38884	18897	24312	31005	29724	24540

图 12　各区间评语质量分布图

我们发现 0~10 字区间内的评语与各维度的匹配百分比均不超过 13%，位居四个区间最末，这也再次证明上文对评语长度的研究发现：太短的评语缺乏反思意识。11~25 字以及 26~50 字两个区间评语匹配数量基本趋同，除了在研究题目维度 11~25 字区间内的

评语匹配数量略高于 26~50 字区间的评语,其他 5 个维度都是后者略高于前者。

5. “长评语”中有复制量规文字的现象

按照评语与评价量规匹配度定义评语质量,则 50 字以上区间内的评语在评语质量方面遥遥领先,所占比例居四个区间之最。因此,为了进一步探究这些评语长度较长的评语质量如何,撰写这些评语的学习者反思意识程度如何,我们从评语字数大于 50 字的评语中随机抽取(R 语言的 sample 函数)了 300 条进行文本分析,分析结果发现这类评语呈现出以下几类特点:

少数评语会根据课程团队给出的评价量规进行评价,评语内容非常细致且全面,例如:题目一般就是研究的中心和重点,此文题目很明确,是对学习焦虑的研究,但是,从选题的依据开始,思路就比较混乱,只是对前人研究过的内容的一种罗列,并没有说明为什么要研究它;研究的内容只字未提“焦虑”,只是将一些英语学不好的原因罗列在上面,没有体现怎样去研究;研究思路做成了如何促成英语学习的多维需求的论述,偏离了课题设计的要求;研究的创新之处太多,不符合实际;通篇文字臃长,没有结构和层次。但这类评语在 300 份抽样中所占比例较少。

从 50 字以上评语随机抽样的文本分析结果显示,约 36.35%的学习者撰写的评语出现了粘贴复制评价量规的情况,例如:“提出了研究问题,且绘制了清晰的研究内容之间的关系图,或用文字步步论证了逻辑关系,对研究的重点和难点有清晰的认识和合理的判断。能够为每个研究问题选择合适的研究方法,且说明了之前的研究基础以及具备本研究所需要的条件,研究过程完整,考虑了信效度”,这部分评语反映出学习者在对他人作业进行评价的过程中有应用评价量规的意识,但是没有结合作业中的具体内容给出有针对性的改进建议,反思程度有限,这一结论也与我们对评语长度的分析结果一致:长评语也可能缺乏反思意识。此外,数据分析结果显示:有 762 名学习者(约占比 7%)会选择将同样一份评语复制给另一份作业。

有部分学习者评价的作业和自己感兴趣的研究主题相关,所以会在评语中表达出自己对该申请书的兴趣,或者是对该研究主题的赞许,在对作业本身评价之外附加一些自己对该研究主题的赞许之词。例如:“看了您的开题报告,不禁让我眼前一亮,我们学校也非常注重青年新教师的培养,开展了‘ssss’青蓝工程,每一个新任教师都有自己的师父,师父准备‘xx 杯’课程,让徒弟学习,每个徒弟也要根据师父的指导,自己准备‘yy 杯’课程,其他教师给出意见,对每位新教师都是很大的帮助。期末的时候师父徒弟同时准备同课异构。希望您的研究成果能够及时得到推广,能让更多的新教师受益!”该学习者共评价了 7 份作业,评语长度(单位:字)依次为 177、128、54、46、26、24、2,可见这类学习者表现出对感兴趣话题的联想和反思意识,但并不是对所有作业都能够一视同仁,这也在情理之中。确实存在极少数评语是直接从其他地方拷贝的文字,和同伴互评活动本身毫无关系,属于投机分子,钻系统自动统分的空子。

(四)学习成效和评语质量呈现显著正相关

前面我们发现“成绩好的学生互评评语也会较长”,那么成绩好的学生(因为有较好的学习方法)做互评的时候是不是会更多参考评价量规呢?即我们想研究成绩好的学生是不是评语质量也会比较好。根据学员 ID 将 match 值和课程成绩进行关联,得到了包含 match 值和课程成绩的数据记录。基于此,绘制了 match 值和课程成绩的误差条形图,如图 13 所示。

由于这门课程的作业互评是在最后一周,如果学习者顺利完成了前四周的学习内容和学习活动,那么进入第五周学习时基本学业表现都已达到合格线。从我们的统计结果来看,确实有同伴互评评语记录的学习者中约 83.2%都属于合格学习者,所以即使撰写的评语和评价量规吻合度为 0,即 match 值为 0,这类学习者的平均学业表现也高达 71.75 分,属于合格学习者。

我们发现,随着评语质量逐渐提高,学习者的学业表现确实亦呈现上升趋势,且计算结果表明评语质量和学业表现存在显著正相关(r=0.07,p=0.000<0.01)。

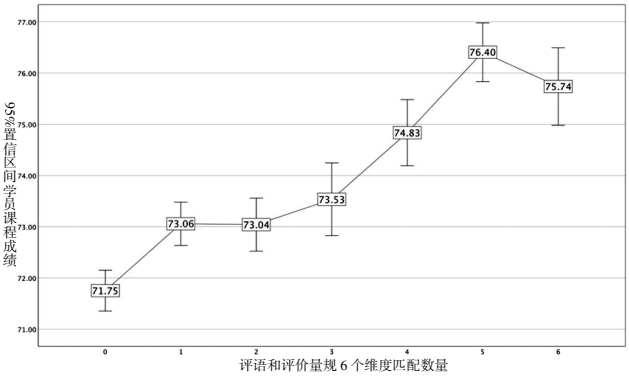


图 13　评语质量和学业表现的误差条形图

四、研究发现的价值

本研究基于一门 MOOC 课程 11 期学习者作业互评数据,有以下发现:

第一,在互评活动中大多数学习者都有一定的反思意识。

这门 MOOC 要求每位学习者要评 6 份作业,但

是人均评语数是 7.78 条，即人均批改了近 8 份作业，“超额”完成了互评作业任务。这说明 MOOC 学习者是很愿意看同伴作业的，同伴互评活动受到学习者欢迎。

本研究将“反思意识”操作化定义为“学习者在进行作业互评的时候，有无应用作业互评量规的意识”。学习者反思意识的强度操作化定义为两部分：评语长度和评语质量，其中“评语质量”转化定义为“评语和评价量规之间的吻合度”。

在分析数据的时候看到有很多评语字数不多，有一些评语存在复制的情况，但是当我们分析“有多少学习者的所有评语字数都不多”的时候，发现这样的学习者很少。大多数学习者还是会因不同的作业而给出不同长度的评判，有的长、有的短，这是合乎情理的。例如：如果看到的作业本身不认真，也就很难有兴趣对它多做点评。这个假设在研究中得到了证实：成绩好的学习者的作业确实得到的长评语较多。从评语长度和评语质量的关系图（图 13）中可以看出，长评语涉及评价量规的维度也会多一些。学习者喜欢并愿意分析、借鉴认真完成的作业。

另外，MOOC 学习者的作业水平有很大的相似性，在刚开始看同伴作业的时候，评语可能会写得长一些，后面见到的作业有太多类似的问题，就可能不太想写评语了，评语字数会越写越少，或者开始复制之前的评语，这也是人之常情，并不能因此就认为是学习者参与互评活动不认真。如果一名学习者对所批改的 6 份作业（这里是项目申请书）有一两份仔细阅读和认真分析了，本课程通过观摩同伴作业进行反思和拓展思路的教学目的就达到了。

那么，会不会出现成绩好的学习者在互评作业中收获更多，而成绩差的学习者其作业得到的评语字数都很少或质量不高的情况呢？表 3 对学生成绩与所获得评语长度的分析揭示：每个分数段的平均评语长度差异不大。从评语质量的分析中可以看出，确实有一成的学习者所得到的作业评语都与评价量规无关，虽然这些学习者的作业成绩并不低，但是所得到的评语无助于他们改进作业也是事实。换句话说，作为评语的施予方，互评活动的教学目标基本达到，但是作为评语的接受方，互评活动的价值和作用还有改进空间。如果平台能够增加对评语的反馈或再评价功能，也许会督促学习者写出更有价值的评语。

综上所述，我们可以得出这样的结论：MOOC 作业互评活动受到学习者欢迎，大多数学习者都是认真参与互评活动的，有一定的反思意识，但是反思质量有待提升，需要强调评分量规对作业评判和评语的指导价值。

第二，学习者评语长度有习惯阈值，评语长度可以作为课程成绩预测指标。

从图 5 可以看出，学生评语长度似乎有个规律，对应每个评分维度大概会写 25~35 个字。这让我们联想到网购平台的客户留言往往也是要求达到 25~35 个字，才能获得某种返点奖励。这可能就是人在网上留言的自然阈值，25~35 个字的留言要求不难达到，且通常可以写出实质性内容。由此，我们建议以后在布置互评作业任务时，不妨要求学习者评语至少写 25 个字，也许可以提高评语的针对性。如果 MOOC 平台有能力拒收 25 字以下的评语或者拒绝复制评语操作，也许可以让作业互评质量有明显提升。

本研究用数据证实了成绩合格的学习者学业表现和评语字数存在显著正相关，合格学习者撰写的评语字数稳定在 25~50 字，且和评价量规具有较高的吻合度，这说明合格学习者在同伴互评过程中撰写的评语具有较为稳定字数区间及内容特点，而这两类特征都是可以由计算机进行大规模验证的，可以作为 MOOC 学习者课程成绩的预测指标。

通过本研究，我们也发现了这门 MOOC 开展同伴互评活动需要改进的地方：

第一，需要更明确阐述互评评语的教学价值。

如果既需要评价者以专家的水平给出评价，还要求评价者给出的评语能被被评价者理解，这对于评价者来说其实是很困难的[7]。虽然目前使用分项量规的评价方式确实能够提高同伴互评的信效度[8]，但并不一定能够提高评语的质量。学习者需要教学团队更明确地提点，才能够意识到写评语的过程也是一个运用知识、展示所学的学习活动，避免随意评论或者生硬地拷贝评价量规进行评价。在另一门 MOOC 中，我们曾尝试用课程论坛组织大家研讨“你希望收到什么样的作业评语”，让学习者设身处地地考虑作业评语对于作业改进的价值、对于自我反思、取长补短的意义，取得了一定的效果。

第二，可增加同伴互评活动的交互性。

已有研究表明，同伴互评活动如果能够更具有交互性，随之产生的如争辩和质疑等交互行为将促进学习者的反思过程，因此，允许被评价者根据得到的反馈评语对评价者做出回应和质疑也是很多同伴互评活动设计的方式之一[9]。如果平台能够支持对评语的再评价或反馈是最理想的，可以有效督促提升评语质量。在平台不支持互评过程申辩的情况下，也可以通过在论坛中开设同伴互评结果讨论区板块，让学习者

分享作业样本和相应评语,建构出一个同伴互评评语分享的在线学习空间。特别是优秀评语的分享,会让学习者看到别人如何认真对待互评活动,也会起到很好的榜样作用。

[参考文献]

[1] TOPPING K.Using peer assessment to inspire reflection and learning[M]. New York:Routledge,39-41.

[2] FALCHIKOV N,GOLDFINCH J. Student peer assessment in higher education:a meta-analysis comparing peer and teacher marks[J]. Review of educational research,2000,70(3):287-322.

[3] HANRAHAN S J ,ISAACS G.Assessing self-and peer-assessment: the students´ views [J].Higher education research & development, 2001,20(1):53-70.

[4] MEEK S E M,BLAKEMORE L,MARKS L. Is peer review an appropriate form of assessment in a MOOC? student participation and performance in formative peer review[J]. Assessment & evaluation in higher education,2017,42(6):1-14.

[5] POL J V D, BERG M V D, ADMIRAAL W F, et al. The nature, reception, and use of online peer feedback in higher education[J]. Computers and education, 2008,51(4):1804-1817.

[6] SCHWEGLER A F,ALTMAN B W. Analysis of peer review comments: QM recommendations and feedback intervention theory[J]. American journal of distance education,2015,29(3):186-197.

[7] CLARK H H, BRENNAN S A. Grounding in communication [M]// RESNICK L B,LEVINE J M,TEASLEY S D.Perspectives on socially shared cognition. Washington,DC:American Psychological Association,1991:127-149.

[8] 范逸洲,冯菲,刘玉,汪琼.评价量规设计对慕课同伴互评有效性的影响研究[J].电化教育研究,2018,39(11):45-51.

[9] MORRIS J.Peer assessment: a missing link between teaching and learning? a review of the literature [J]. Nurse education today, 2001,21(7):507-515.

Study on the Relationship between Reflective Awareness and Learning Outcomes in Peer Assessment of MOOCs

WANG Qiong, OUYANG Jiayu, FAN Yizhou

(Graduate School of Education, Peking University, Beijing 100871)

[Abstract] Peer assessment is an instructional activity commonly used in MOOCs. In order to help learners confidently comment on others´ assignments and promote their self-reflection, the instructors usually provide a rubric as the reflection framework and hope that learners can deepen their understanding of the teaching objectives in the process of writing comments with the rubric. However, not all students understand the real intention of peer assessment. This study is interested in how many learners in MOOCs have reflective awareness during homework mutual assessment, the degree of reflective awareness, and whether there is a correlation between reflective awareness and academic performance. Through in-depth analysis of 79287 pieces of peer mutual assessment data in the course "How Teachers Do Research", it is found that, in the absence of natural intervention, most learners have certain reflective awareness, but the quality of the written comments still needs to be improved. The length and quality of the comments of qualified students are positively correlated with their learning outcomes, which indicates that the length and quality of the comments could be used as a predictor of learning outcomes. This study attempts to use quantitative methods to analyze the teaching significance of learners´ comments in MOOCs and provide reference for subsequent studies.

[Keywords] MOOC; Peer Assessment; Reflective Awareness; Number of Comments; Quality of Comments; Learning Performance