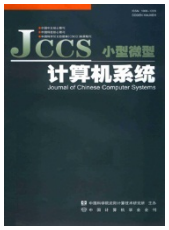

小型微型计算机系统
Journal of Chinese Computer Systems
ISSN 1000-1220,CN 21-1106/TP

《小型微型计算机系统》网络首发论文

题目：一种基于认知诊断的主观题同行互评技术
作者：许嘉，李秋云，刘静，吕品，于戈
收稿日期：2021-01-20
网络首发日期：2021-07-07
引用格式：许嘉，李秋云，刘静，吕品，于戈．一种基于认知诊断的主观题同行互评技术．小型微型计算机系统.
https://kns.cnki.net/kcms/detail/21.1106.tp.20210706.1549.027.html

小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems

一种基于认知诊断的主观题同行互评技术

许 嘉[1,2,3]，李秋云[1]，刘 静[1]，吕 品[1,2,3]，于 戈[4]

[1]（广西大学 计算机与电子信息学院，广西 南宁 530004）

[2]（广西大学 广西多媒体通信网络技术重点实验室，广西 南宁 530004）

[3]（广西大学 广西高校并行与分布式计算重点实验室，广西 南宁 530004）

[4]（东北大学 计算机科学与工程学院，辽宁 沈阳 110819）

E-mail ：lvpin@gxu.edu.cn

摘 要：针对 MOOCs 平台上大规模主观题作业的同行互评问题，研究人员基于概率模型对评价者的可靠性和偏见进行建模，提出了许多估计主观题作业真实分数的有效技术。然而，现有技术均未同时考虑评价者在待评价作业中的答题表现以及评价者的历史答题表现这两方面因素对其可靠性的影响。鉴于此，提出了基于认知诊断的主观题同行互评技术：首先以评价者的历史答题记录为输入，基于流行的认知诊断模型量化评价者对主观题作业的掌握程度；其后同时基于评价者对主观题作业的掌握程度以及评价者在该主观题作业中取得的真实分数对评价者的可靠性建模；最后结合对评价者偏见的建模提出了估计主观题作业真实分数的同行互评概率模型。真实课堂实验表明，在同行互评活动中，本文提出的同行互评技术对主观题作业真实分数的估计更为准确，比相关技术在真实分数估计误差方面平均降低了 42%。

关键词：同行互评；认知诊断；DINA 模型；主观题；真实分数估计

中图分类号：TP391 **文献标识码：**A

A Peer Grading Technology for Subjective Questions Based on Cognitive Diagnosis

XU Jia[1,2,3]，LI Qiu-yun[1]，LIU JING[1]，LÜ Pin[1,2,3]，YU Ge[4]

1 (*School of Computer Electronics and Information, Guangxi University, Nanning 530004, China*)

2 (*Guangxi Key Laboratory of Multimedia Communications and Network Technology, Nanning 530004, China*)

3 (*Guangxi Colleges and University Key Laboratory of Parallel and Distributed Computing, Nanning 530004, China*)

4 (*School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China*)

Abstract: To solve the peer grading problem of large-scale subjective questions at MOOCs platforms, researchers have proposed many effective technologies to estimate true scores of subjective questions based on probability models which model peer graders' reliability and bias. However, the existing technologies do not take into account both of a grader's performance gained in the graded subjective question and the grader's historical question-answering performance. In view of this, a peer grading technology for subjective questions based on cognitive diagnosis is proposed in this paper. First, based on a grader's historical question-answering records, the grader's competence to a subjective question is computed by using a popular cognitive diagnosis model. Second, a grader's grading reliability to a subjective question is modeled by considering both of the grader's competence to the question and the estimated true score of the grader in the question. Finally, by combining the modeling of a grader's bias, two probability models are proposed to estimate the true scores of subjective questions in peer grading activities. Real classroom experiments show that, in peer grading activities, the proposed peer grading technology gets more accurate estimates for the true scores of subjective questions, reducing the error of true score estimation by on average of 42% compared with the state-of-the-art technology.

Key words: peer grading; cognitive diagnosis; DINA model; subjective question; true score estimation

收稿日期：2021-01-20 收修改稿日期：2021-03-02 基金项目：国家自然科学基金项目（62067001，U1811261）资助；“广西八桂学者”专项经费项目资助； 广西自然科学基金项目（2019JJA170045）资助；广西高等教育本科教学改革工程项目（2020JGA116，2017JGZ103）资助. 作者简介：**许嘉**，女，1984 年生，博士，副教授，硕士生导师，CCF 会员，CCF 数据库专委会委员，研究方向为教育数据分析挖掘；**李秋云，**女，1996 年生，硕士，CCF 会员，研究方向为主观题互评技术；**刘静，**女，1995 年生，硕士，CCF 会员，研究方向为主观题互评技术；**吕品(通讯作者)，**男，1983 年生，博士，副研究员， CCF 会员，CCF 协同计算专委会委员，研究方向为物联网技术；**于戈**，男，1962 年生，博士，教授，博士生导师，CCF 会员，研究方向为数据库理论与技术。

1 引 言

随着大数据、云计算和互联网技术的不断发展，以Coursera、edX、中国大学MOOC和学堂在线为代表的在线教育平台的兴起给平台上的任课教师带来了严峻的教学挑战。一个最突出的教学挑战在于教师如何高效批改大规模选课学生在平台上提交的作业。鉴于做作业能够帮助学生巩固和内化知识，是至关重要的教学活动，各大在线教育平台都提供了客观题（例如选择题和判断题）的自动批改功能，减轻了任课教师的教学负担。相对于客观题，主观题（例如简答题和应用题）更能考察学生的语言表达能力、知识运用能力与创新思维能力，所以主观题的考察对于很多在线课程而言是必不可少的[1]。然而，由于没有唯一标准答案，主观题的批改很难由计算机自动完成[2]，需要任课教师花费大量精力逐份手工批改，导致他们无法将精力用于课程内容及活动的改进提高。可见，如何减轻任课教师的主观题批改负担是当前教育研究领域亟待解决的重要问题。

为了有效降低任课教师的主观题作业批改负担，国内？外各大在线平台与科研机构提出了不少主观题评判的技术，这些技术可分为两类：基于自然语言处理的评判技术[3-5]和基于同行互评的评判技术[6-10]。其中，基于自然语言处理的评判技术通过分析学生答案与教师给的参考答案之间的匹配程度来实现主观题的自动判分。然而，基于自然语言处理的评判技术通常依赖于特定领域的知识，只适用于解决面向特定领域的主观题评分问题，因此鲜有在线教育平台提供基于自然语言处理的主观题评判功能。基于同行互评的评判技术是当下不少主流在线教育平台（例如Coursera和中国大学MOOC）提供的主观题评判功能。该类技术将主观题批改任务的子集分派给每个学生，然后基于多名学生对某主观题的评分来估计该题的真实分数。基于同行互评的主观题评判技术对于教师与学生而言都有积极益处：一方面减轻了任课教师的主观题作业批改负担；另一方面要求学生评判他人的主观题作业，不但能够让他们学习到不同的解题思路，还能提高他们的课程参与度[11,12]。因此，基于同行互评的主观题评判技术成为当下解决大规模主观题评判问题的主流技术和目前智能教育领域的研究热点，关注于提出提高同行互评质量的方法[13]。

本文考虑基于基数估计的同行互评场景，即每名同行评价者针对每道主观题给出一个数值型的评价分数。基于同行互评的主观题评判方法的研究难点在于如何利用多个同行给出的评价分数估计被评价者的真实分数。大多数在线教育平台只是简单基于各个评价分数的均值或中位数来估计被评价者的真实分数。然而，由于同行评价者的打分质量受其可靠性、偏见等因素的影响[14]，简单用各个评价分数的均值或中位数估计被评价者的真实分数往往不够准确[15]。近年来，研究人员将同行评价者的评分可靠性及评分偏见作为模型的随机变量，构建了估计被评价主观题作业真实分数的概率模型，能够利用变量间的依赖关系提高估计的准确性[6-9]。然而，现有研究方法均假设同行评价者的可靠性只与其当前作业的答题情况相关，未同时考虑同行评价者对主观题考察的知识点的掌握程度（由其历史答题结果数据诊断得到）对其评分可靠性造成的影响，因而存在局限性。对284名同行评价者针对三道主观题作业给出的2109条互评打分记录进行统计分析。具体而言，首先以这些同行评价者的历史答题结果数据为输入并利用流行的认知诊断DINA模型[16]诊断得到他们对主观题考察的知识点的掌握程度，并进而量化每个同行评价者对每道主观题的掌握程度值。之后，计算由每名同行评价者对每道主观题的掌握程度值组成的序列与每名同行评价者对每道主观题的评分误差值序列之间的皮尔逊相关系数。由于两个序列的皮尔逊相关系数为-0.673，表明评价者的可靠性还受其对该主观题掌握程度的影响：评价者的掌握程度越低，则平均评分误差越大，可靠性越低；评价者的掌握程度越高，则平均评分误差越小，可靠性越大。因此，在对同行评价者的可靠性进行建模时，应该同时考虑评价者对待评价习题的掌握程度信息。

鉴于此，本文提出了一种基于认知诊断的主观题同行互评技术，包括PG_8和PG_9两个概率模型。该技术在现有概率模型的基础上[9]，同时基于同行评价者在本次作业中的答题表现（对应于本次作业取得的真实分数）以及评价者的历史答题表现（对应于基于历史答题记录诊断得到的该评价者对本次作业题的掌握程度）对评价者的可靠性进行建模，以期最终提高概率模型估计主观题作业真实分数的准确性。PG_8和PG_9的区别在于：PG_8假设评价者的评分可靠性服从伽马分布；PG_9则假设评价者的评分可靠性服从高斯分布。综上，本文的主要贡献包括：

（1）提出了改进现有同行评价概率模型的思路，即应同时以认知诊断得到的同行评价者对主观题的掌握程度信息和评价者在该主观题中取得的真实分数信息作为评价者评分可靠性的建模依据，以期进一步提高概率模型对主观题作业真实分数的估计准确性。

（2）基于由284名学生参与的3次主观题作业的互评活动收集真实互评数据集，并基于该数据集评估提出的互评技术和相关互评技术的有效性。实验结果表明本文提出的基于认知诊断的主观题互评技术在提高对主观题作业真实分数的估计准确性方面比其它相关技术更具优势。

本文剩余部分的内容组织如下。第2部分阐释了相关研究工作。第3部分给出了预备知识。第4部分给出了基于认知诊断的同行互评技术，包含PG_8和PG_9两个概率模型。第5部分为实验。最后，第6部分总结了全文。

2 相关工作

2.1 基于自然语言处理的主观题评判技术

基于自然语言处理的主观题评判技术从题目本身的特性出发，利用自然语言处理、机器学习等技术实现主观题的自动评判。例如，文献[5]基于自然语言处理技术对开放式数学问题的每一个解答转变为数字特征，再通过聚类分析发现解答中正确、部分正确以及不正确的解答结构，从而实现了对该类问题的自动判分。文献[3]针对英文论文写作题给出了自动判分的解决方案，该方案利用潜在语义分析和学习向量量化算法来提升自动判分的准确率。文献[17]针对英语简答题设计了自动判分方法，该方法利用同义词词典和衡量语义距离的两种自然语言处理方法来解决标准文本相似度

衡量方法对于同义词的匹配不够准确的问题。文献[4]则基于潜在语义分析的奇异值分解策略设计了日语短文的自动评分系统。基于自然语言处理的主观题评判技术为主观题的自动评分提供解决思路，也取得了不错的评分效果。然而，该类技术通常依赖特定领域的知识来优化自然语言的处理过程，从而保证自动判分的准确性，因而只适用于解决特定领域的主观题自动判分问题，很难在其它领域推广使用。

2.2 基于同行互评的主观题评判技术

基于同行互评的主观题评判问题即让每名评价者对分配给其的一部分主观题作业进行评判，最终基于各个评价者反馈的评判信息估计每份主观题作业的质量。由于评价者的态度和能力存在差异，与众包问题类似[18][19]，基于同行互评的主观题评判问题需要解决的核心问题是对评价者反馈的评价信息进行质量控制。按照评价者反馈的评价信息形式的不同，基于同行互评的主观题评价技术可分为序数（Ordinal）估计技术和基数（Cardinal）估计技术两类。

序数估计技术要求每名评价者对分配给其的主观题作业给出表征作业质量高低的排名反馈，系统则基于所有评价者给出的作业间的偏序排名信息估计每份作业的质量[20]。序数估计技术通常利用基于配对比较的方法[21,22]、贝叶斯生成法[23]和矩阵分解方法[24]来估计主观题作业的质量。序数估计的方法不要求同行评价者给出主观题作业的具体分数，降低了评价者的评判难度。然而，该类技术存在两大问题[25]：首先，评价者由于评判经验有限，很难对质量相差不大的两份主观题作业给出它们的合理排序；其次，仅依赖作业间的偏序排名信息很难量化两份作业之间的质量差异。

与序数估计技术不同，基数估计技术要求每名评价者对分配给其的每份主观题作业都给出一个量化分数，系统继而基于不同评价者针对同一份作业给出的多个评价分数估计作业的真实分数。主流的基数估计方式有两种：加权求和的估计方式[25-28]和基于概率模型的估计方式[6-9]。其中，加权求和的估计方式依据同行评价者的评分准确性和信任度给他们赋以不同的权重，然后以同行评价者针对主观题作业给出的评价分数为输入，通过加权求和的方法来估计该作业的真实分数。系统会根据同行评价者在新的互评活动中的评分表现来迭代更新其权重信息。另一类方式是通过构建概率模型来估计主观题作业的真实分数。本文提出的基于认知诊断的主观题互评技术就属于这类方法。这类方法的主要实现思路是将待估计的主观题作业的真实分数、同行评价者的可靠性及偏见都建模为满足一定概率分布的隐含变量，然后基于能观察到的同行评价者的评分信息来推演以上各个隐含变量的值。具体而言，Piech 等人[6]首先提出了估计主观题作业真实分数的三个概率模型，即 PG_1（考虑了评价者当前的可靠性和偏见），PG_2（在 PG_1 的基础上考虑了评价者的历史偏见），PG_3（在 PG_1 的基础上将评价者当前可靠性设定为评价者当前作业真实分数的线性函数的随机变量）。考虑到 PG_3 模型所设置的评价者的可靠性是关于评价者真实分数的线性函数这一假设过于严格，Mi 等人将评价者的可靠性建模为满足形状参数为其真实分数的伽马分布或均值为其真实分数的高斯分布，分别得到了 PG_4 模型和 PG_5 模型[7]。研究表明一名同行评价者的评分偏见会受到其朋友的评分偏见的影响[29,30]，为了提高对评价者偏见建模的准确性，Chan 等人利用学堂在线平台上收集到的学生间的社交关系信息优化对评价者偏见的建模，扩展了 PG_1、PG_4、PG_5 这三个概率模型[8]。然而上述概率模型均认为评价者针对不同主观题作业给出的评价分数之间是相互独立的，存在局限性。因此，Wang 等人在概率建模时引入了评价者的相对分数信息（即同一个评价者对不同作业评分之间的差值），提出了 PG_6 模型（构建在 PG_4 之上），PG_7 模型（构建在 PG_5 之上）[9]。这两个概率模型由于引入了评价者的相对分数信息，降低了数据稀疏性给参数估计带来的负面影响，从而有效提高了对主观题真实分数估计的准确性。然而，PG_6 模型与 PG_7 模型仅基于同行评价者针对当前主观题作业取得的真实分数对其可靠性进行建模。PG_6 模型与 PG_7 模型是当前最好的同行互评概率模型，实验部分将针对这两种相关模型进行比较分析。

综上，基于概率模型的基数估计方法是目前实现主观题评判的主流方法，近年来研究人员们提出了不少相关工作。然而，现有研究工作在概率建模时均未同时考虑影响同行评价者评分可靠性的两大因素，即其在本次作业中的答题表现（对应于本次作业取得的真实分数）以及其的历史答题表现（对应于基于历史答题记录诊断得到的该评价者对本次作业题的掌握程度），因而限制了它们对于主观题真实分数的估计准确性。

3　预备知识

认知诊断以认知心理学和心理计量学为理论基础，通过构建具有认知诊断功能的心理计量模型，能够基于被试的历史答题结果数据诊断其对不同技能（知识点）的掌握程度，从而为教学提供重要依据，是当下教育评估领域的研究热点[31-33]。作为最流行的认知诊断模型之一，DINA 模型[16]在实现对被试知识点掌握程度的精准建模的同时具有较好的解释性，近年来受到广泛的关注和研究[34,35]。以同行评价者的历史答题结果数据为诊断基础，本文正是基于 DINA 认知诊断模型来量化评价者对主观题作业的掌握程度。

给定被试集合 $C=\{c_1, ..., c_M\}$，习题集合 $E=\{e_1, ..., e_N\}$，则记录被试和其答题结果之间关联关系的响应矩阵 $\boldsymbol{R}$ 可表示为 $\boldsymbol{R}=[r_{mn}]_{M\times N}$，其中 $r_{mn}=1$ 表示被试 c_m 答对了习题 e_n（$r_{mn}=0$ 则表示答错了该题）。设习题集合 E 考察的知识点集合为 $KP=\{kp_1, \ldots, kp_K\}$，则记录习题与其考察的知识点之间关联关系的 $\boldsymbol{Q}$ 矩阵可表示为 $\boldsymbol{Q}=[q_{nk}]_{N\times K}$，其中 $q_{nk}=1$ 表示习题 e_n 考察了知识点 KP_k（$q_{nk}=0$ 则表示未考察该知识点）。DINA 模型将被试 c_m 的知识状态描述为一个向量 $\boldsymbol{\alpha}_m=\{\alpha_{m1}, ..., \alpha_{mK}\}$，称为被试 c_m 的知识点掌握程度向量。其中，α_{mk} 表示被试 c_m 对知识点 kp_k 的掌握程度，且 $\alpha_{mk}\in[0,1]$。$\alpha_{mk}=1$ 说明被试 c_m 完全掌握了第 k 个知识点；$\alpha_{mk}=0$ 则说明被试 c_m 完全没有掌握第 k 个知识点。DINA 认知诊断模型的项目反应函数为：

$$P(r_{mn}=1 \mid \alpha_m) = guess_n^{1-\delta_{mn}} (1-slip_n)^{\delta_{mn}} \quad (1)$$

其中，

$$\delta_{mn} = \prod_{k=1}^{K} \alpha_{mk}^{q_{nk}} \quad (2)$$

公式 2 中，δ_{mn} 表示知识状态为$\boldsymbol{\alpha}_m$的被试 c_m 对习题 e_n 的潜在正确作答概率，即可被定义为被试 c_m 对习题 e_n 的掌握程度值；$slip_n=P(r_{mn}=0 \mid \delta_{mn}=1)$表示被试掌握习题 e_n 考察的所有知识点但是答错该题的概率，被称为失误参数；$guess_n=P(r_{mn}=1 \mid \delta_{mn}=0)$指被试没有掌握习题 e_n 考察的任何一个知识点时但答对该题的概率，被称为猜测参数。DINA 模型利用 EM 算法最大化公式 1 的边缘似然值，从而得到被试 c_m 的知识点掌握程度向量$\boldsymbol{\alpha}_m$。

本文假设参与主观题互评活动的同行评价者在进行主观题作业评判之前完成了该主观题考察的知识点所对应的客观题的习题练习，因而作业互评测试系统能够收集到他们对于这些知识点对应的客观习题的答题结果数据。以某同行评价者的历史答题结果数据和表征习题和主观题作业知识点间考察关系的 $\boldsymbol{Q}$ 矩阵为输入，利用 DINA 认知诊断模型即可求得该同行评价者的知识点掌握程度向量$\boldsymbol{\alpha}$。然后基于$\boldsymbol{\alpha}$和主观题作业所考察的知识点信息即可以利用公式 2 求得该评价者对于该主观题的掌握程度值。

4 同行互评概率模型

本节介绍了基于认知诊断的主观题同行互评技术，具体涉及概率模型 PG_8 与 PG_9。用 U 表示提交主观题作业的被评价者集合，V 表示参与互评的同行评价者集合。考虑到实际教学实践中一般要求提交主观题作业的被评价者都参与该作业的互评活动，因而有$|U|=|V|$。下面给出模型所涉及的重要概念的定义并说明它们在模型中的设定。

真实分数：假设每份被评价者提交的主观题作业对应一个真实分数，且用 s_i 表示被评价者 $u_i \in U$ 所提交作业的真实分数。两个概率模型中均假设变量 s_i 的取值满足高斯分布。

可靠性：可靠性（记为 τ_v）表示同行评价者 $v \in V$ 对主观题作业的评分精度。评价者 v 的可靠性实际反映了 v 给出的主观题作业的评价分数基于其偏见 b_v 修正后的分数与主观题作业真实分数之间的接近程度。给定某主观题作业，本文首先假设评价者 v 对于该作业的评分可靠性 τ_v 满足形状参数为$\theta_1\delta_v+\theta_2 s_v$的伽马分布，得到 PG_8 模型；其次假设 τ_v 满足均值为$\theta_1\delta_v+\theta_2 s_v$的高斯分布，得到 PG_9 模型。其中，δ_v 表示基于 DINA 认知诊断模型得到的评价者 v 对该作业的掌握程度。可见，PG_8 和 PG_9 在对评价者可靠性建模时同时考虑了评价者的对当前作业答题表现（对应$\theta_2 s_v$部分）和评价者的历史答题表现（对应$\theta_1\tau_v$部分）。

偏见：偏见（记为 b_v）是量化同行评价者 $v \in V$ 评分时表现出其评分高于真实分数或其评分低于真实分数的常量。考虑到互评活动中不同的同行评价者的偏见不同（有些给分偏高，有些则给分偏低），因此两个概率模型均认为所有评价者的偏见值的均值为 0，即假设同行评价者 v 的偏见 b_v 服从均值为 0 且方差为 $1/\eta_0$ 的高斯分布。

互评分数：互评分数（记为 z_i^v）表示同行评价者 $v \in V$ 针对被评价者 u_i 提交的主观题作业给出的评价分数。设所有评价者的互评分数集合为 $Z=\{z_i^v \mid u_i \in U, v \in V\}$。两个概率模型均假设变量 z_i^v 服从以高斯分布，且高斯分布的均值等于作业的真实分数 s_i 与评价者 v 的评分偏见 b_v 之和，方差反比于评价者 v 的可靠性 τ_v。在 PG_9 模型中引入了超参数 λ用于调节高斯分布的方差取值。

相对分数：相对分数（记为 d_{ij}^v）表示同行评价者 $v \in V$ 对被评价者 $u_i \in U$ 和 $u_j \in U$ 的主观题作业给出的互评分数间的差值。记面向所有评价者的相对分数集合为 $D=\{d_{ij}^v \mid u_i, u_j \in U, v \in V\}$。相对分数的引入有利于提高对主观题作业真实分数估计的精度。PG_8 模型中，相对分数 d_{ij}^v 被设定为满足均值为两份被 v 评价的主观题作业的真实分数之差（即 s_i-s_j）、且方差为 $2/\tau_v$ 的高斯分布。在 PG_9 模型中同样引入了超参数 λ 用于调节高斯分布的方差取值。

基于以上符号表征，本文的研究问题为：已知所有同行评价者的互评分数集合 Z，面向所有评价者的相对分数集合 D，所有评价者的知识点掌握程度向量$\boldsymbol{\alpha}$构成的矩阵 $\boldsymbol{M}_{|V|\times|KP|}$，通过构建概率模型 PG_8 和 PG_9 推断出每个同行评价者（即 $\forall v \in V$）的可靠性 τ_v、偏见 b_v 以及每个被评价者（即$\forall u_i \in U$）提交的主观题作业的真实分数 s_i，可以形式化表示为 $P(\{b_v|v \in V\}, \{\tau_v|v \in V\}, \{s_i|u_i \in U\} \mid Z, D, \boldsymbol{M})$。表 1 总结了模型涉及的主要符号和相关解释。

表 1 主要符号及其含义

Table 1 Main notations and their descriptions

符号	描述
U	被评价者集合，u_i 表示第 i 个被评价者
V	同行评价者集合，$v \in V$ 表示某评价者
U_v	作业被评价者 v 评判的被评价者集合
V_{u_i}	评判被评价者 u_i 作业的同行评价者集合
$\boldsymbol{\alpha}_v$	评价者 v 对主观题的掌握程度
s_i	被评价者 u_i 的主观题作业的真实分数
τ_v	评价者 v 的可靠性
b_v	评价者 v 的偏见
z_i^v	评价者 v 给被评价者 u_i 的主观题作业的评分
d_{ij}^v	即相对分数，表示评价者 v 给被评价者 u_i 的作业评分与给 u_j 的作业评分之差

图 1 展示了 PG_8 和 PG_9 的概率图模型。可见，同行评价者 v 针对被评价者 u_i 的主观题作业给出的互评分数 z_i^v、v 针对被评价者 u_i 和被评价者 u_j 给出的评价分数之间的相对分数 d_{ij}^v、v 的潜在正确作答概率δ_v是概率图模型中的观测变量。而 u_i 的主观题作业的真实分数 s_i、v 的偏见 b_v、v 的可靠性 τ_v 则是概率模型估计的隐含变量，且这些隐含变量的先验分布由超参数 μ_0、γ_0、θ_1、θ_2、η_0 和 β_0 所确定。由图可知，这些隐含变量彼此间是相联系的。因而，为了估计这些隐含变量的值，基于每个隐含变量的近似后验分布信息，并利用 Gibbs 采样技术[36]对每个隐含变量的取值进行采样。具体而言，Gibbs 采样技术：首先基于每个隐含变量的近似后验分布信息运行若干次 Gibbs 采样以生成该变量的若干个样本，得到该变量的样本集；其后，当隐含变量样本的分布逐渐趋于收敛和稳定时，基于隐含变量的样本集推断变量的真实值。例如，假定基于 Gibbs 采样技术所得到的被评价者 u_i 的主观题作业真实分数 s_i 的样本集为$\{s_i^1, s_i^2, \ldots, s_i^{I_G}\}$且 I_G 为采样的次数，则可基于样本集中样本的平均值来估计 s_i。考虑到 Gibbs 采样过程存在老化阶段（Burn-in 阶段），这时得到的隐含变量的样本不准确，因而基于 Gibbs 采样技术生成隐含变量的样本集时需要丢弃在老化阶段生成的样本（一

般为样本集中的前 n 个样本）。

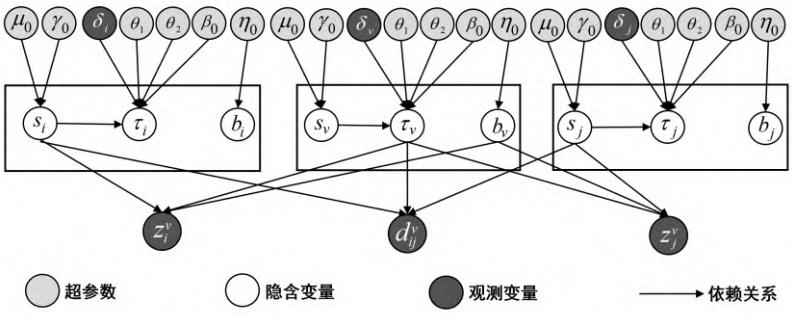

图 1　PG_8 和 PG_9 的概率图模型

Fig. 1 Probabilistic graphical model for PG_8 and PG_9

4.1 PG_8 模型

PG_8 模型扩展了现有的 PG_6 模型[9]，其的生成过程为：

- 对于第 i 个被评价者 u_i 提交的每份主观题作业
  - → 定义隐含变量 s_i（即 u_i 的真实分数）$s_i \sim N(\mu_0, 1/\gamma_0)$
- 对于每个同行评价者 v
  - → 定义隐含变量 τ_v（即 v 的可靠性）$\tau_v \sim \Gamma(\theta_1\delta_v + \theta_2 s_v, 1/\eta_0)$
  - → 定义隐含变量 b_v（即 v 的偏见）$b_v \sim N(0, 1/\eta_0)$
- 对于每个互评分数 z_i^v
  - → 定义可观测变量 $z_i^v \sim N(s_i + b_v, 1/\tau_v)$
- 对于每个相对分数 d_{ij}^v
  - → 定义可观测变量 $d_{ij}^v \sim N(s_i - s_j, 2/\tau_v)$

由于概率模型 PG_8 中的隐含变量 s_i 没有闭式解（close-form solution），因而采用近似离散推断的策略得到该隐含变量的近似后验分布。概率模型 PG_8 中隐含变量的近似后验分布的推断结果如下：

$$s \propto \frac{\beta_0^{\theta_2 s_i}\tau_i^{(\theta_2 s_i - 1)}}{\Gamma(\theta_1\delta_i + \theta_2 s_i)} \times \exp(R(s_i - \frac{Y}{R})^2) \tag{3}$$

$$\text{其中 } R = \gamma_0 + \sum_{v \in V_{u_i}} \tau_v + \sum_{v \in V_{u_i}} \sum_{u_j \in U_v} \frac{\tau_v}{2},$$

$$Y = \mu_0\gamma_0 + \tau_v(\sum_{v \in V_{u_i}} (z_i^v - b_v) + \sum_{v \in V_{u_i}} \sum_{u_j \in U_v} \frac{(d_{ij}^v + s_j)}{2})$$

$$\tau \sim \Gamma(\theta_1\delta_v + \theta_2 s_v + \frac{|U_v|}{2}, \beta_0 + \frac{\sum_{u_i \in U_v} (z_i^v - s_i - b_v)^2 + \sum_{u_i, u_j \in U_v} \frac{1}{2}(d_{ij}^v - s_i + s_j)}{2}) \tag{4}$$

$$b \sim \mathcal{N}(\frac{\sum_{u_i \in U_v} \tau_v (z_i^v - s_i)}{\eta_0 + |U_v|\tau_v}, \frac{1}{\eta_0 + |U_v|\tau_v}) \tag{5}$$

4.2 PG_9 模型

PG_8 模型与 PG_9 模型的区别在于 PG_8 模型假同行设评价者的可靠性满足伽马分布而 PG_9 模型则假设同行设评价者的可靠性满足高斯分布。PG_9 模型扩展了现有的 PG_7 模型[9]，其的生成过程为：

- 对于第 i 个被评价者 u_i 提交的每份主观题作业
  - → 定义隐含变量 s_i（即 u_i 的真实分数）$s_i \sim N(\mu_0, 1/\gamma_0)$
- 对于每个同行评价者 v
  - → 定义隐含变量 τ_v（即 v 的可靠性）$\tau_v \sim N(\theta_1\delta_v + \theta_2 s_v, 1/\eta_0)$
  - → 定义隐含变量 b_v（即 v 的偏见）$b_v \sim N(0, 1/\eta_0)$
- 对于每个互评分数 z_i^v
  - → 定义可观测变量 $z_i^v \sim N(s_i + b_v, \lambda/\tau_v)$
- 对于每个相对分数 d_{ij}^v
  - → 定义可观测变量 $d_{ij}^v \sim N(s_i - s_j, 2\lambda/\tau_v)$

由于 PG_9 模型中的隐含变量 s_i 和 τ_v 没有闭式解，因而采用近似离散推断的策略得到该隐含变量的近似后验分布。概率模型 PG_9 中隐含变量的近似后验分布的推断结果如下：

$$s \propto \frac{\beta_0^{\theta_2 s_i}\tau_i^{(\theta_2 s_i - 1)}}{\Gamma(\theta_1\delta_i + \theta_2 s_i)} \times \exp(R(s_i - \frac{Y}{R})^2) \tag{6}$$

$$\text{其中 } R = \gamma_0 + \sum_{v \in V_{u_i}} \frac{\tau_v}{\lambda} + \sum_{v \in V_{u_i}} \frac{\tau_v * (|U_v| - 1)}{2\lambda},$$

$$Y = \gamma_0\mu_0 + \frac{\tau_v}{\lambda}(\sum_{v \in V_{u_i}} (z_i^v - b_v) + \frac{\sum_{v \in V_{u_i}} \sum_{u_j \in U_v} (d_{ij}^v + s_j)}{2})$$

$$\tau \propto \tau_v^{\frac{|U_v|}{2}} \times \exp(-\frac{\beta_0}{2}[\tau_v - (\theta_1\delta_v + \theta_2 s_v + \sum_{u_i \in U_v} \frac{(z_i^v - s_i - b_v)^2}{\lambda\beta_0} + \sum_{u_i, u_j \in U_v} \frac{(d_{ij}^v - s_i + s_j)^2}{2\lambda\beta_0})]^2) \tag{7}$$

$$b \sim \mathcal{N}(\frac{\sum_{u_i \in U_v} \frac{\tau_v}{\lambda}(z_i^v - s_i)}{\eta_0 + |U_v|\frac{\tau_v}{\lambda}}, \frac{1}{\eta_0 + |U_v|\frac{\tau_v}{\lambda}}) \tag{8}$$

4.3 真实分数估计步骤

利用 PG_8 模型和 PG_9 模型即可估计一份主观题作业的真实分数，具体分为以下四个步骤：

步骤一：认知诊断。以所有同行评价者的历史答题记录为输入，利用 DINA 模型诊断得到记录了他们对所有知识点的掌握程度信息的矩阵 $\boldsymbol{M}$。

步骤二：推理。由于概率模型中的各个变量是相互联系的，因而基于模型中观测变量的观测值（包括同行评价者

v 的潜在正确作答概率 $_v$、互评分数 z_i^v 和相对分数 d_{ij}^v）推断模型中隐含变量（包括同行评价者的偏见 b_v、可靠性 τ_v 和被评价者 u_i 的主观题作业的真实分数 s_i）的后验概率分布是一个循环推理的过程，最终推理得到 PG_8 模型中各个隐含变量的近似后验分布（循环推理得到的近似后验概率分布如公式 3-5 所示）以及 PG_9 模型中各个隐含变量的近似后验分布（循环推理得到的近似后验概率分布如公式 6-8 所示）。

步骤三：采样。以互评分数集合、相对分数集合和步骤一得到的知识点的掌握程度矩阵 $\boldsymbol{M}$ 为输入，以 Gibbs 采样技术为采样框架并利用步骤二得到的各个隐含变量的近似后验分布得到概率模型中每个隐含变量的多个样本值。

步骤四：整合。对步骤三得到的概率模型中的每个隐含变量的多个样本值进行整合，进而得到每个隐含变量（包括主观题作业的真实分数）的估计值。

5 实 验

基于真实采集的主观题同行互评数据集，本节对本文提出的基于认知诊断的主观题同行互评技术 PG_8、PG_9 和相关的主观题同行互评技术进行了实验比较。

5.1 数据集

为了验证本文提出的基于认知诊断的同行互评技术对于主观题评判的有效性，基于自主研发的“会了吗”在线教学服务系统[37]收集计算机专业核心主干课“数据库原理”中“关系数据库规范化理论”这一节的真实教学数据，得到涉及关系数据库规范化理论相关知识点的主观题同行互评数据集以及客观题测试结果数据集。

5.1.1 主观题同行互评数据集

在“会了吗”在线教学服务系统中实现了主观题作业的互评功能。通过给“数据库原理”课程的五个本科平行教学班的 284 名学生布置考察了关系数据库规范化理论的三次主观题作业并组织他们进行同行互评从而得到主观题同行互评数据集。每次主观题作业仅包含一道主观题，且布置的三次主观题作业涉及考察关系数据库规范化理论的 11 个知识点，这些知识点和它们的编号分别为：（1）一范式；（2）二范式；（3）三范式；（4）BC 范式；（5）主属性；（6）传递函数依赖；（7）决定因素；（8）函数依赖；（9）码；（10）部分函数依赖；（11）非主属性。这些知识点是数据库原理这门课的教学难点，而主观题形式的作业比客观题形式的作业能更好地帮助学生巩固对这些知识点的学习。图 2 给出了记录了三次主观题作业所考察知识点信息的 Q 矩阵。

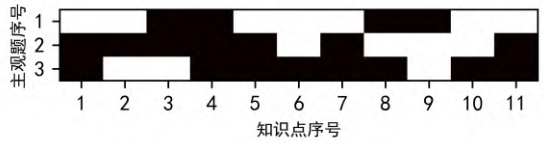

图 2 主观题作业的 Q 矩阵

Fig. 2 The Q matrix of subjective questions

在主观题作业的互评教学活动中，每名学生既是提交主观题作业的提交者（即被评价者）又是评判同行提交的主观题作业的评价者。每个评价者都会收到系统随机给其派发的 3 份主观题作业，并要求其遵循教师制定的评分指导规则完成对这 3 份主观题作业的判分。需要说明的是，为了保证互评的质量，互评活动采用双盲的方式进行。为了评估不同主观题互评技术对于主观题作业真实估计的准确性，邀请拥有 6 年以上“数据库原理”课程教学经验的教师对每份学生提交的主观题作业进行评价打分，并以教师的评分作为该主观题作业的真实分数。表 2 给出了从三次主观题作业的互评教学活动收集到的主观题同行互评数据集的相关统计信息。

表 2 主观题同行互评数据集的统计信息

Table 2 Statistics of our subjective question dataset for peer grading

	作业 1	作业 2	作业 3
作业提交数	268	269	267
教师评价数	268	269	267
同行评价者数	254	260	254
同行互评数	694	725	690
作业分值	20	20	20

5.1.2 历史客观题测试结果数据集

为了能够基于 DINA 模型诊断学生对主观题的掌握程度，要求学生们在“会了吗”在线教学服务系统上完成包含 40 道客观题的在线测试。这些客观题覆盖了三次主观题作业考察的关系数据库规范化理论的 11 个知识点。基于在线测试活动得到的每名学生的客观题测试结果数据和记录了每道客观题考察的知识点信息的 Q 矩阵（如图 3 所示），从而可基于 DINA 模型诊断每名学生对 11 个知识点的掌握程度，进而可计算每名学生对每道客观题作业的掌握程度。

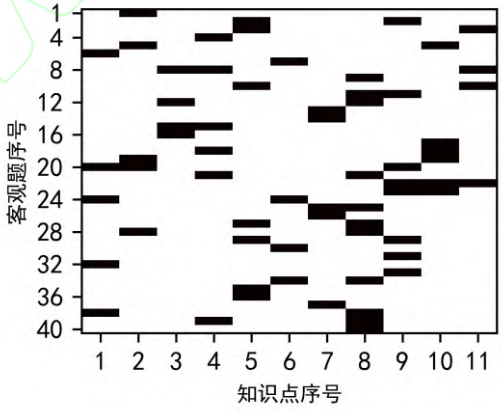

图 3 每道客观题考察的知识点信息的 Q 矩阵

Fig. 3 The Q matrix of objective questions

5.2 参与比较的主观题同行互评技术

为了评估本文提出的 PG_8 模型与 PG_9 模型的有效性，将它们与其它主观题同行互评技术进行比较，具体包括：

- **中位数**：即用一份主观题作业所获得的所有评价分数的中位数估计该作业的真实分数，这也是当今大多数提供主观题互评功能的 MOOC 平台（例如 Coursera 和中国大学 MOOC）采用的估计主观题作业真实分数的方法。
- **均值**：即用一份主观题作业所获得的所有评价分数的均值估计该作业的真实分数。
- **PG_6 和 PG_7[9]**：PG_6 和 PG_7 均是解决主观题同行互评问题的现有最先进概率模型。本文提出的 PG_8 与 PG_9 模

型分别是在 PG_6 和 PG_7 模型的基础上对评价者可靠性进行了建模优化。具体而言，PG_6 和 PG_7 模型在评价者可靠性时仅考虑了其在当前主观题作业中的答题表现，而 PG_8 与 PG_9 模型在对评价者的可靠性进行建模时不但考虑了其在当前作业中的答题表现还考虑了基于其历史答题表现诊断得到的评价者对待评价作业的掌握程度信息，以期提高概率模型对主观题作业真实分数估计的精确性。需要说明的是：（1）PG_8 与 PG_6 相对应，均假设同行评价者互评可靠性取值的先验分布为伽马分布；（2）PG_9 与 PG_7 相对应，均假设同行评价者互评可靠性取值的先验分布为高斯分布。

5.3　实验设置

本文提出的主观题同行互评技术和相关主观题同行互评技术 PG_6 和 PG_7 均是利用概率模型对同行评价者的互评可靠性和互评偏见进行建模，因而都使用了一些超参数。为这些超参数设置合理的值对准确估计主观题作业的真实分数非常重要。对于概率模型中的真实分数变量 s_i 服从的高斯分布的超参数，即均值 μ_0 和方差 $1/\gamma_0$，分别设置为所有主观题作业互评分数的均值和方差。根据文献[7,9]的参数设置，本文的具体调整策略为：对于 PG_8 和 PG_6，在其它参数取值固定的前提下，以 50 为步长尝试超参数 β_0 在[150, 400]范围中的不同取值，然后以该技术所得到的对真实分数最准确的估计值为该技术的最终估计值；对于 PG_9 和 PG_7，在其它参数取值固定的前提下，以 0.2 为步长尝试超参数 λ 在[0.6, 1.6]范围中不同取值，然后以该技术所得到的对真实分数最准确的估计值为该技术的最终估计值。由于基于概率模型的同行互评技术在估计主观题作业真实分数时具有一定的随机性，因此对于超参数集合的每种设定，每种技术都执行 10 次真实分数的推断算法。对于基于概率模型的同行互评技术中每个需要估计的隐含变量，推断算法均迭代运行 600 次 Gibbs 采样获取隐含变量的样本值，并设定前 60 次采样得到的样本为老化阶段的样本，这些老化阶段的样本将不参与对真实分数的估计运算。

所有参与比较的主观题同行互评技术均基于 Python（v3.7）语言实现，并在配备了 i5-8500 3GHZ CPU、8GB 内存、1TB 硬盘，运行了 64 位 Windows 10 操作系统的服务器上进行统一实验测试。

5.4　实验结果

5.4.1　同行互评技术的估计准确性

采用不同技术给出的对主观题真实分数的估计值和主观题作业真实分数之间的均方根误差（即 RMSE）作为不同同行互评技术有效性的评估指标。RMSE 被广泛应用于评估同行互评技术有效性[6,8]。表 3 展示了不同主观题同行互评技术估计主观题作业真实分数的准确性。需要说明的是，表中的 RMSE 表示互评技术 10 次迭代得到的 RMSE 的平均值，而 STD 表示 RMSE 的标准差。由表 3 可知，本文提出的基于认知诊断的同行互评技术 PG_8 和 PG_9 在三份主观题作业中的估计准确率均高于比其他技术。由于同时考虑了同行评价者在本次作业中的答题表现以及评价者的历史答题表现对其评分可靠性的影响，PG_8 和 PG_9 技术对三次作业真实分数的平均估计误差比 PG_6 和 PG_7 技术平均降低了 42%。实验结果证实了结合本次作业中的答题表现以及评价者的历史答题表现建模可靠性对于基数同行互评估计的有效性。

表 3 估计真实分数的准确性

Table 3 The error of true score estimation

	作业 1		作业 2		作业 3	
	RMSE	STD	RMSE	STD	RMSE	STD
均值	4.61	0.00	4.16	0.00	4.53	0.00
中位数	5.09	0.00	4.59	0.00	5.04	0.00
PG_6	3.32	0.01	2.67	0.01	3.32	0.02
PG_8	2.31	0.01	1.69	0.01	**1.30**	**0.01**
PG_7	2.46	0.01	2.56	0.00	2.82	0.01
PG_9	**1.57**	**0.01**	**1.28**	**0.00**	1.63	0.01

5.4.2　同行互评技术的最大估计误差

通过衡量主观题作业真实分数估计值与教师批改分数之间的最大评分偏差来分析同行互评技术的评估表现，如表 4 所示。从表中可看出，均值技术与中位数技术的最大评分偏差是最大的，而基于认知诊断的同行互评技术 PG_8 和 PG_9 在三份主观题作业中的最大评分偏差是最小的，说明同行评价者对主观题作业考察的知识点的掌握程度信息使概率模型能更有效地保障对每个学生的主观题作业真实分数的估计准确性。同时还可观察到，PG_8 和 PG_9 技术对三次作业真实分数估计的最大评分误差均低于 PG_6 和 PG_7 技术，进一步表明了同时考虑影响可靠性的两方面因素（即同行评价者在本次作业中的答题表现以及评价者的历史答题表现）能够提升对主观题作业真实分数估计的精确性。

表 4 真实分数估计值与真实分数间的最大评分偏差

Table 4　Maximum deviation between an estimated grade and ground truth for all students

	作业 1	作业 2	作业 3
均值	18	8	16
中位数	18	8	16
PG_6	10.87	6.31	10.46
PG_7	10.74	6.36	10.81
PG_8	6.03	5.38	**4.26**
PG_9	**5.54**	**4.12**	4.94

5.4.3　同行互评技术的超参数敏感性

为了表明 PG_8 技术中的超参数 β_0 和 PG_9 技术中的超参数 λ 对主观题作业真实分数估计的影响，本文采取固定其他超参数值的策略并对这两个超参数的值进行了实验分析。在实验中为了测试模型的敏感性，将 PG_8 中的超参数 β_0 设置在[150, 400]范围内以 50 为步长变化，实验结果如图 4；将 PG_9 中的超参数 λ 设置在[0.6, 1.6]范围内以 0.2 为步长变化，实验结果如图 5。图 4 和图 5 的结果表明：在合理的取值范围内，这两种技术对超参数值具有鲁棒性，它们对主观题作业真实分数的估计误差都控制在可接受的范围。

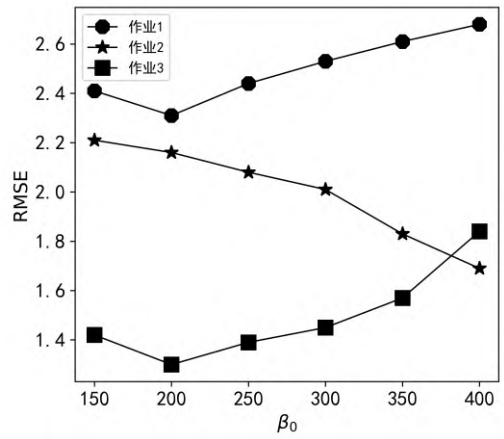

图 4 PG_8 技术的超参数敏感性分析

Fig. 4 Sensitivity analysis of hyper-parameter for PG_8

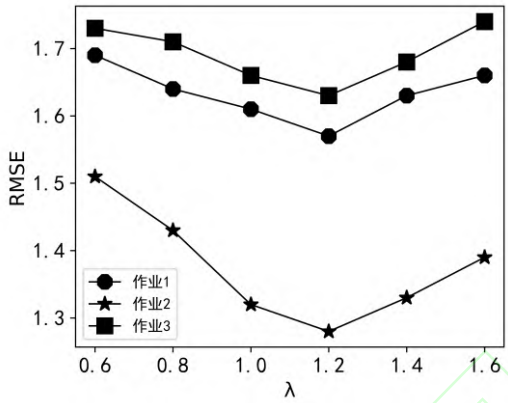

图 5 PG_9 技术的超参数敏感性分析

Fig. 5 Sensitivity analysis of hyper-parameter for PG_9

6 总结

同行互评是当前大型开放式网络课程（MOOCs）平台用以解决大规模主观题作业评价的主流方式。同行评价者的评分偏见和评分可靠性是未知的，因此基于多个同行评价者给出的评价分数估计主观题作业的真实分数是一个具有挑战的问题。现有同行互评技术利用概率模型对同行评价者的评分可靠性和评分偏见进行建模，有效提高了估计主观题作业的真实分数的准确性。然而，这些技术均未同时考虑同行评价者在本次作业中的答题表现以及评价者的历史答题表现对其评分可靠性的影响。鉴于此，本文在现有概率模型的基础上提出了基于认知诊断的主观题同行互评技术，包含 PG_8 和 PG_9 两个概率模型。PG_8 和 PG_9 利用教育评估领域流行的认知诊断 DINA 模型诊断得到同行评价者对主观题的掌握程度信息并结合评价者在待评价作业中的答题表现对评价者评分可靠性进行建模，实验证实 PG_8 和 PG_9 比相关最好的同行技术在提升主观题作业真实分数估计准确性方面更有优势。

References:

[1] Paré D E, Joordens S. Peering into large lectures: examining peer and expert mark agreement using peerScholar, an online peer assessment tool[J]. Journal of Computer Assisted Learning, 2008, 24(6): 526-540.

[2] Caragiannis I, Krimpas G A, Voudouris A A. Aggregating partial rankings with applications to peer grading in massive online open courses[C]//Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2015: 675-683.

[3] Ratna A A P, Raharjo B S, Purnamasari P D, et al. Automatic essay grading system with latent semantic analysis and learning vector quantization[C]// Proceedings of the 3rd International Conference on Communication and Information Processing (ICCIP), 2017: 158-163.

[4] Ratna A A P, Santiar L, Ibrahim I,et al. Latent semantic analysis and winnowing algorithm based automatic japanese short essay answer grading system comparative performance[C]//IEEE 10th International Conference on Awareness Science and Technology (iCAST), 2019: 1-7.

[5] Lan A S, Vats D, Waters A E, et al. Mathematical language processing automatic grading and feedback for open response mathematical questions[C]//Proceedings of the Second ACM Conference on Learning @ Scale (L@S), 2015: 167-176.

[6] Piech C, Huang J, Chen Z, et al. Tuned models of peer assessment in moocs[C]//Proceedings of the 6th International Conference on Educational Data (EDM), 2013: 153-160.

[7] Mi F, Yeung D Y. Probabilistic graphical models for boosting cardinal and ordinal peer grading in moocs[C]//Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI), 2015: 454-460.

[8] Chan H P, King I. Leveraging social connections to improve peer assessment in moocs[C]//Proceedings of the 26th International Conference on World Wide Web (WWW), 2017: 341-349.

[9] Wang T Q, Li Q, Gao J, et al. Improving peer assessment accuracy by incorporating relative peer grades[C]//Proceedings of the 12th International Conference on Educational Data (EDM), 2019: 450-455.

[10] Song Y, Hu Z, Guo Y, et al. An experiment with separate formative and summative rubrics in educational peer assessment[C]//2016 IEEE Frontiers in Education Conference (FIE), 2016: 1-7.

[11] Kulkarni C E, Wei W P, Le H, et al. Peer and self assessment in massive online classes[J].ACM Transactions on Computer-Human Interaction, 2013, 20(6): 1-31.

[12] Gehringer E F. A survey of methods for improving review quality[C]//New Horizons in Web Based Learning-ICWL 2014 International Workshops,2014: 92-97.

[13] Wang W Y, An B，Jiang Y C. Optimal spot-checking for improving the evaluation quality of crowdsourcing: application to peer grading systems[J]. IEEE Transactions on Computational Social Systems, 2020, 7(4): 940-955.

[14] Alfaro L D, Shavlovsky M. Dynamics of peer grading: an empirical study[C]//Proceedings of the 9th

International Conference on Educational Data (EDM), 2016: 62-69.
[15] Capuano N, Caballé S. Towards adaptive peer assessment for MOOCs[C]//10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2015: 64-69.
[16] Torre D L. DINA model and parameter estimation: a didactic[J]. Journal of Educational and Behavioral Statistics, 2008, 34(1): 115-130.
[17] Omran A M B, Aziz M J A. Automatic essay grading system for short answers in english language[J]. Journal of Computer Science, 2013, 9(10): 1369-1382.
[18] Gao Li-ping, Jin Tao. Long-term crowdsourcing quality control strategy for dynamically selecting worker model[J]. Journal of Chinese Computer Systems, 2020, 41(10): 2017-2023.
[19] Zheng Zhi-yun, Jiang Guo-lin, Zhang Xing-jin, et al. Crowdsourcing quality evaluation algorithm based on sliding task window[J]. Journal of Chinese Computer Systems, 2017, 38(9): 2125-2129.
[20] Wauthier F L, Jordan M I, Jojic N. Efficient ranking from pairwise comparisons[C]//Proceedings of the 30th International Conference on Machine Learning (ICML), 2013，(3): 109–117.
[21] Shah N B, Bradley J K, Parekh A, et al. A case for ordinal peer-evaluation in moocs[C]//In NIPS Workshop on Data Driven Education,2013: 1-8.
[22] Raman K, Joachims T. Methods for ordinal peer grading[C]//The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2014: 1037-1046.
[23] Waters A E, Tinapple D, Baraniuk R G. Bayesrank: a bayesian approach to ranked peer grading[C]// Proceedings of the Second ACM Conference on Learning @ Scale (L@S), 2015: 177-183.
[24] Luacesa O, D éza J, Betanzosb A A, et al. A factorization approach to evaluate open-response assignments in moocs using preference learning on peer assessments[J]. Knowledge-Based Systems, 2015, 85(9): 322-328.
[25] Alfaro L D, Shavlovsky M. Crowdgrader: a tool for crowdsourcing the evaluation of homework assignments[C]//The 45th ACM Technical Symposium on Computer Science Education (SIGCSE), 2014: 415–420.
[26] Walsh T. The peerRank method for peer assessment[C]//2014-21st European Conference on Artificial Intelligence (ECAI), 2014: 909–914.
[27] Gutierrez P, Osman N, Sierra C. Collaborative assessment[C]//In Proceedings of the 17th International Conference of the Catalan Association for Artificial Intelligence (CCIA), 2014: 136-145.
[28] Garcia-Martinez C, Cerezo R, Bermudez M, et al. Improving essay peer grading accuracy in massive open online courses using personalized weights from student's engagement and performance[J]. Journal of Computer Assisted Learning, 2019, 35(1): 110-120.
[29] Singla P, Richardson M. Yes, there is a correlation: from social networks to personal behavior on the web[C]// Proceedings of the 17th International Conference on World Wide Web (WWW), 2008: 655-664.
[30] Yang S H, Long B, Smola A J, et al. Like like alike: joint friendship and interest propagation in social networks[C]//Proceedings of the 20th International Conference on World Wide Web (WWW), 2011: 537-546.
[31] Wu R Z, Liu Q, Liu Y P, et al. Cognitive modelling for predicting examinee performance[C]//Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI), 2015: 1017-1024.
[32] Zhu T Y, Liu Q, Huang Z Y, et al. MT-MCD: a multi-task cognitive diagnosis framework for student assessment[C]//Database Systems for Advanced Applications-23rd International Conference (DASFAA), 2018(2): 318-335.
[33] Cheng S, Liu Q, Chen E H, et al. DIRT: deep learning enhanced item response theory for cognitive diagnosis[C]//Proceedings of the 28th International Conference on Information (CIKM), 2019: 2397-2400.
[34] Zhu Tian-yu, Huang Zhen-ya, Chen En-hong, et al. Cognitive diagnosis based personalized question Recommendation[J]. Chinese Journal of Computers, 2017, 40(1): 178-193.
[35] Wang Chao, Liu Qi, Chen En-hong, et al. The rapid calculation method of DINA model for large scale cognitive diagnosis[J]. Acta Electronica Sinica, 2018, 46(5): 1047-1055.
[36] Geman S, Geman D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, 6(6): 721-741.
[37] Xu J, Wang J B, Wang X X, et al. iTest: a novel online testing system based on the weChat platform[J]. Computer Applications in Engineering Education,2019, 27(4): 885-893.

附中文参考文献：

[18] 高丽萍，金涛. 动态选取工作者模型的长时众包质量控制策略[J]. 小型微型计算机系统，2020，41(10): 2017-2023.
[19] 郑志蕴，江国林，张行进,等. 基于滑动任务窗的众包质量评估算法[J]. 小型微型计算机系统，2017，38(9): 2125-2129.
[34] 朱天宇，黄振亚，陈恩红,等. 基于认知诊断的个性化试题推荐方法[J]. 计算机学报, 2017, 40(1):178-193.
[35] 王超，刘淇，陈恩红,等. 面向大规模认知诊断的DINA 模型快速计算方法研究[J]. 电子学报，2018，46(5): 1047-1055.