ПРИМЕНЕНИЕ ПИРИНГОВОЙ ОЦЕНКИ ПИСЬМЕННЫХ РАБОТ СТУДЕНТОВ В ПОТОКОВЫХ ОЧНЫХ КУРСАХ

Д.А. Королев, А.С. Кольбе, А.В. Паволоцкий

Национальный исследовательский университет «Высшая школа экономики»
ул. Мясницкая, 20, Москва, Россия, 101000

При сокращении контактного времени в учебных дисциплинах требуются инструменты сохранения активных форм изучения для студентов, предполагающие самостоятельную исследовательскую работу. Одной из таких форм является пиринговая оценка работ (peer assessment), позволяющая выйти за узкие рамки автоматически проверяемых тестов на уровень небольших исследовательских работ. В статье приводятся результаты двухлетнего эксперимента по использованию пиринга на потоковом курсе в МИЭМ НИУ ВШЭ с учетом специфики аудитории относительно массовых онлайн курсов.

Полученные результаты и выводы послужили основой для моделирования способов оценивания письменных работ при пиринговой проверке без участия преподавателя. В эксперименте были отработаны схемы оценивания и мотивации для управления соотношением количества авторов и рецензентов, а также общие вопросы использования предложенной системы оценки в курсе.

Ключевые слова: кросс-рецензирование, пиринг, автоматизация проверки, очное обучение в вузе, оценивание работ студентов, электронная поддержка очных курсов

Введение

Пиринговая проверка, а также оценка, рецензирование (от англ. peer assessment (1), evaluation, review) — это способ проверки и оценивания письменных работ, когда работа одного автора проверяется несколькими независимыми рецензентами, обычно из одного с автором круга. В данном случае проверку осуществляли студенты того же потока. Как правило, для повышения объективности оценки используется «слепой» или «двойной слепой» метод проверки, когда автор не знает, кто проверяет его работы, и когда проверяющий такжене знает, кто автор данной работы. Такой способ проверки или рецензирования активно используется в научном мире при рецензировании публикаций перед печатью [3], а также в синхронных массовых онлайн-курсах, когда невозможно проверить задание автоматически, а ручная проверка при аудитории в несколько тысяч человек просто невозможна [4].

Использование пиринга позволяет не только сократить время, потраченное преподавателем на проверку работ, но и изменить сам подход к обучению. По наблюдениям [8], в среднем на прочтение и комментирование студенческой работы преподаватель тратит 20—40 минут, а в случае с массовым образованием количество работ и, соответственно, время труда пропорционально возрастают. Автор отмечает, что задания с peer review развивают у студентов навык чтения научной литературы и писательский навык — студентам приходится писать научные работы на доступном языке, а также учат составлять конструктивную критику, в том числе отрицательную, что обычно вызывает сложности у студентов.

Для peer review как формы учебной активности важной проблемой является мотивация студентов [2]. Yanqing Wang, Yaowen Liang, Luning Liu and Ying Liu разработали свою систему "EduPCR4" для проведения peer review программного кода [9]. Система предлагает три типа баллов для мотивации студентов: баллы, начисляемые за предоставление работы в срок, качественные баллы, зависящие от задания, и бонусные баллы, которые могут быть как положительными, так и отрицательными — зависит от единства мнения рецензентов. Данная модель мотивирует студентов к участию как автором, так и рецензентом.

Обработка рецензий и расчет итоговой оценки является ключевой задачей, связанной с peer review. Hoi K. Suen рассмотрел распространенные методы обработки результатов peer review [7]. Calibrated Peer Review (CPR) — подход, зависящий от реальной успеваемости рецензента, что позволяет задать весовой коэффициент его рецензии. Bayesian post hoc stabilization построен на Байесовских моделях. Однако данный метод не учитывает систематические ошибки. Последний подход широко применяется на популярном онлайн-ресурсе Coursera.org.

Подход к построению учебного курса

Для краткого курса длительностью всего восемь недель на практике возможно провести всего две работы с пиринговой проверкой, учитывая, что на написание и на проверку работы дается по одной неделе, после каждой работы нужно оставить время на разбор итогов (коротко на лекции и подробно — на сайте поддержки), чтобы в студенты учли свои ошибки в следующей работе или к зачету. К тому же это весьма ресурсоемкое мероприятие и для студентов, и для преподавателей.

В такой ситуации хотелось избежать обычного для онлайн-курсов ground-truth тестирования, когда студентам дается набор эталонных, заранее проверенных экспертами работ и по результатам проверки этих работ студентами устанавливается, насколько оценка каждого студента близка к экспертной, далее их голоса учитываются с соответствующими поправочными коэффициентами.

Подход к оценке работ студентов в курсе

Студенты имеют разные амбиции, интересы и способности. Обычный учебный план предполагает одинаковую траекторию для всего потока. Система оценок в курсе предполагала накопление баллов, и различные учитываемые в накопленной оценке активности в сумме давали существенно больше баллов, чем максимально возможная оценка в ведомости. Студент мог сам выбирать, каким образом ему зарабатывать себе баллы, и мог посчитать, какие виды активности вероятнее принесут ему желаемые баллы. При этом, компенсируя отсутствие «отрицательной оценки» в традиционном понимании, здесь в шкале оценки отдельных видов работ могли присутствовать действительно отрицательные величины, т.е., начинающиеся с отрицательных чисел, и это наглядно демонстрировало влияние того или иного достижения или провала в каждом из заданий. Это не новый подход, он встречается в зарубежной практике, например, в [9], но в российских вузах он, если и практикуется, то на уровне инициативы преподавателей.

Пиринговая проверка работ также не является распространенным инструментом в учебной практике в России, но отдельные преподаватели используют этот инструмент в своих курсах, например, в НИУ ВШЭ [6].

Оценки за работы и за их рецензирование начинались с отрицательной шкалы. Так, при общей шкале возможных баллов около 200 (100 баллов в курсе соответствовали максимальной оценке) письменная работа оценивалась в диапазоне 30 баллов, но от –9 до +20. Все задания были необязательными, но оговаривалось, что «необязательно» не значит «бесплатно». Так, не написавшие эссе получали «0», а не выполнившие в срок рецензирование получали минимальный балл, т.е., «2». Работы, уличенные в превышении допустимого уровня заимствований, автоматически получали минимальный балл, т.е., «9».

Критерии оценки

К заданию прилагалась таблица критериев оценки (2) с детальным описанием не только самих критериев, но и необходимого содержания по каждому критерию для получения того или иного балла. Критерии относились как к содержанию работы, так и к форме. Таблица критериев оценки — это самый важный элемент в постановке работы над пиринговой проверкой. Если требования сформулированы четко и заставляют обратить внимание автора (а позднее — рецензента) на различные стороны работы, то и сам процесс написания работы, удовлетворяющей требованиям, и процесс проверки таких работ становится полезным для студента, так как учит объективности и непредвзятости.

Рецензенты

Большинство студентов из рассматриваемой выборки (второй курс бакалавриата инженерного факультета) не имели навыков написания академических текстов, равно как и навыков проверки чужих работ — в школе этому не учат, а в институте ко второму курсу заняться научной или педагогической деятельностью они еще не успевают. Отсюда возникло опасение, что не только работы будут в среднем низкого качества, но, что важнее, их проверка окажется в руках столь же неквалифицированных рецензентов. Требовалось обеспечить значительное превосходство числа рецензентов над числом авторов, поскольку одну работу должны проверить несколько человек и только в этом случае можно выявить возможные расхождения во мнениях. В самой системе оценок в курсе закладывались стимулирующие меры для рецензентов и преграждающие — для авторов работ. Формально любой студент мог не писать и не проверять эти работы, если был уверен в успешности своих усилий в практической области (лабораторные работы дают достаточно высокий балл, если их выполнять добросовестно).

Постановка эксперимента и используемые инструменты.

Эксперимент проводился в 2013—2014 и 2014—2015 учебных годах. Основной «площадкой» был потоковый курс «Компьютерная графика» на втором курсе бакалавриата факультета информационных технологий и вычислительной техники МИЭМ НИУ ВШЭ.

На этапе подготовки курса были изучены варианты существовавшего на тот момент программного обеспечения [10] и сервисов (например, iPeer (3)), но более детальное ознакомление с ними показало, что те немногие инструменты, которые удалось найти, были ориентированы на другой формат работы и не подходили, а для проведения исследований требовалась максимальная гибкость и возможность исправлять замеченные недоработки в сжатые сроки. В этой ситуации роль базы для экспериментов легла на сервисы Google Apps (тексты Google Documents, таблицы Google Spreadsheets и формы Google Forms), а также сервис Blogger для публикации материалов курса.

Обратная связь от студентов принималась через формы, что позволяло надежно собирать их ответы в таблицы с точным указанием времени отправки. Таблицы, в свою очередь, позволяют автоматизировать обработку поступающей информации при помощи формул и скриптов. Результаты публиковались на сайте. Так из документооборота были исключены этапы переписки со студентами по электронной почте или в месенджерах. Это отнюдь не исключало переписку для поддержки по содержательным или организационным вопросам, но все учебные транзакции стали проходить через формы и регистрироваться в таблицах, исключая человеческий фактор из обработки.

Далее, средствами Forms требовалось создать тесты, а средствами Spreadsheets — весь цикл подготовки и обработки информации пиринговой проверки. Для создания тестов Forms подходит ограниченно, но в данном случае тесты проводились больше для напоминания студентам о курсе, заставляя их вернуться к темам лекций после этих лекций, фактически заменяя «повторение пройденного материала». Существенно сложнее оказалось реализовать при помощи таблиц весь цикл обработки пиринговой проверки эссе. В первый год применялась полуавтоматическая обработка данных: использовались лишь формулы в таблицах, что не позволяло работать с персональной рассылкой электронной почты, ограничивало обработку массивов данных. На этом этапе были сформулированы задачи для автоматизации обработки и выявлены недостатки алгоритмов расчета оценок. С другой стороны, табличное представление информации давало полную картину хода работы: все вычисления были детально видны на листах таблиц, и любой студент, не согласившийся с оценкой, мог видеть, как она формировалась и что помешало ему получить желаемый балл.

Ход работы

Работа позиционировалась как добровольная, но сама эта активность была для студентов необычным нововведением, количество сданных работ примерно соответствовало ожидаемому (рассчитывалось, что работы напишут 20% от потока, а задания на их проверку выполнят 80%). Работы прислали 23 автора (19%), в то же время специфику начисления баллов за рецензирование поняли не все, и количество присланных рецензий оказалось ниже ожидаемого (69%). Тем не менее заложенный в механизм оценки расчет сработал — количество рецензий на одну работу было достаточным не только чтобы увидеть разные мнения рецен-

зентов, но и позволяло применять статистические методы для определения наиболее адекватных оценок.

Описанный ранее подход к оцениванию показал сильное «размытие» оценок (4). И сильные, и слабые работы получали незначительно различающиеся баллы, большинство оценок, отклоняющихся к верхней или нижней границе шкалы, нивелировалось оценками, попадающими в «безопасный» диапазон в середине шкалы.

Чтобы выявить недобросовестных рецезнентов и показать студентам ориентиры на конкретных примерах, на сайте поддержки подробно разбирались все процессы оценивания и выборочного контроля (5).

Второй год эксперимента проходил на потоке 180 человек, и для проведения курса на базе GoogleSpreadsheets были созданы скрипты, что позволило вынести обработку данных из таблиц и использовать их только для сбора информации из форм и наглядного представления результатов обработки. Для наглядности также выводились некоторые промежуточные значения, что позволяло сохранить для студентов возможность видеть принцип формирования их оценки.

Во избежание конфликтов и непонимания принципа расчета оценок в первый год применялась простая формула: считалось среднее арифметическое по оценкам всех рецензентов данной работы по каждому критерию (их было четыре), после чего преподавателем вручную проверялись десять самых неоднозначных работ. Неоднозначность выявлялась по среднеквадратичному отклонению в оценках рецензентов. Далее оценки преподавателя (экспертные оценки) сравнивались с оценками каждого из рецензентов в установленном доверительном интервале (плюс-минус 1 балл по каждому критерию), и, если расхождения превышали допустимый порог, это отмечалось в таблице. Если рецензия имела более 50% критериев (в данном случае три из четырех оценок выходили за доверительный диапазон) с отклоненной оценкой, то она аннулировалась, что исключало ее из расчета общей оценки за работу и снимало все баллы, начисляемые рецензенту за проверку данной работы.

На второй год эксперимента было решено ввести непрерывную весовую шкалу для рецензий, по которой считать и весовой коэффициент оценки рецензента в групповой оценке, и оценку самой рецензии.

1. Для каждой компоненты оценки каждой работы определяется среднее арифметическое значение набора. Средние арифметические значения компонент становятся «эталонными» оценками.

2. Для каждой компоненты оценки каждой работы находится максимальное отклонение от эталонной соответствующей оценки Это отклонение становится длиной шкалы.

3. Для каждой компоненты оценки каждого рецензента находится весовой коэффициент соответствия эталонной оценке, приведенный по шкале (от 0 до 1).

4. Итоговая оценка работы рассчитывается как сумма всех компонент оценки по всем рецензентам для данной работы, умноженных на соответствующий весовой коэффициент.

На рисунке показан пример расчета оценки для одного критерия из 6 баллов. Закрашенные клеточки — это поступившие оценки рецезентов: «1» не поставил

никто, «2» — один, «3» — двое и т.д. Средний балл в таком случае будет равен 4,5, там проходит ось мнения большинства. Длина шкалы — от оси до самой удаленной оценки (2), т.е., 2,5 балла. Вес оценок рецензентов будет падать пропорционально удалению от оси. Так, каждые 0,5 балла при линейном падении веса будут отнимать 20%. Любопытно, что при таком подсчете никто не получит полный балл, так как оценка имеет целочисленные значения и даже самые близкие к оси оценки получат вес 80%, а соседние с ними — по 40%. Компенсировать этот недостаток несложно, пропорционально «подтянув» после всех расчетов все значения, чтобы максимальные достигли 100%, но в эксперименте этого не делалось.

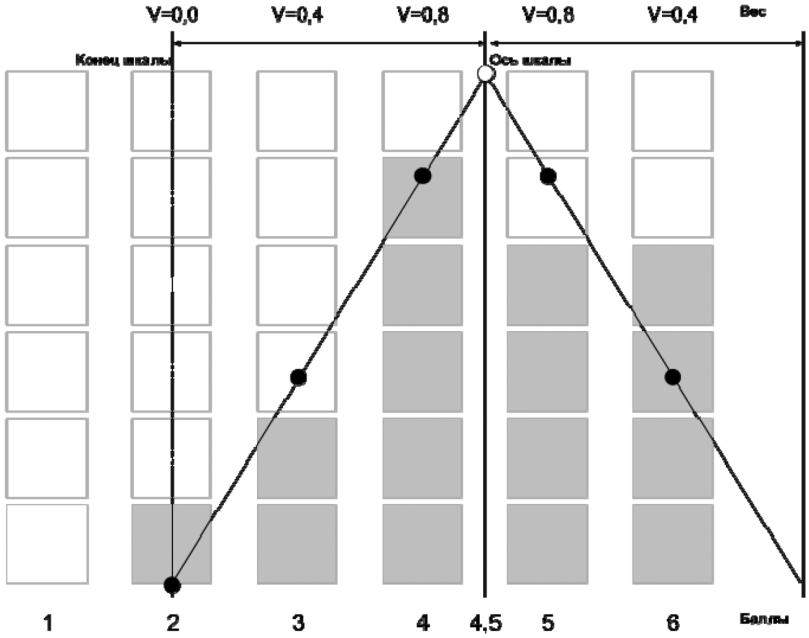

Рис. Расчет весовых коэффициентов рецензий с линейной весовой шкалой (показан расчет по одному критерию с максимумом оценки 6 баллов)

Шкала в таком подходе — это максимальное расстояние от мнения большинства до самого отдаленного от него мнения отдельного рецензента, поэтому на рисунке шкала уходит в область несуществующих значений оценки. Поскольку разные критерии имели разное количество баллов, то и предельная длина шкалы варьировалась. На практике она зависела еще и от единодушия рецензентов — если все сошлись во мнениях и поставили одинаковую оценку, то единственный, поставивший иной балл, попадал на край шкалы и такая оценка обесценивалась. При таком подходе «удаленность» отдельно взятой оценки от «эталонной» (т.е., средней) определяла падение веса этой оценки в итоговой оценке, равно как и оценку за саму рецензию.

На практике проводилась частичная проверка работ преподавателем и далее экспертные оценки принимались за «эталонную», что смещало точку отсчета шкалы. В такой ситуации случалось, что немногие поставившие далекую от мнения большинства оценку, оказывались у основания шкалы, а большинство теряло вес в соответствии с удаленностью от нового центра истины.

Последующая работа

По итогам эксперимента были собраны все полученные от рецензентов оценки, и этот массив данных стал основой для следующего исследования. Зная оценки всех участников эксперимента, включая преподавателя, можно построить формулы, которые дадут высокую степень корреляции с экспертной (преподавательской) оценкой. В данной статье мы не будем останавливаться на этом исследовании, кратко оно отражено в докладе [1].

Результаты и выводы

Численные итоги активности студентов в ходе экспериментов 2013 и 2014 гг. отражены в таблице. Здесь мы видим соотношение числа авторов и рецензентов как следствие мер по стимулированию рецензентов и демонстрации ответственности авторов за некачественную работу, заложены в системе оценки работ в курсе. Каждому студенту предлагалось проверить по три работы, поэтому число рецензий в среднем втрое больше числа рецензентов.

Таблица

Активность студентов в ходе экспериментов

	Эссе-1 2013	Эссе-2 2013	Эссе-1 2014	Эссе-2 2014
Эссе, шт.	23	12	8	12
Рецензии (всего), шт.	249	270	363	453
Рецензенты, человек	83	90	131	149

Даже детализированная по критериям оценки работ схема оценки при пиринговой проверке склонна к смещению к «безопасному диапазону». В такой ситуации рецензент имеет меньше шансов выпасть из доверительного интервала (в первом запуске эксперимента) или на попасть на край весовой шкалы (во втором запуске). Решить это можно введением контрольных вопросов с однозначной оценкой, пересекающихся с оценкой по основным выбранным критериям. Например, если в шкале критериев приводятся ориентиры по количественным характеристикам (допустим, ссылкам на источники), то явное указание численных характеристик (допустим, 3 из 10 минимально необходимых) будет значить, что по соответствующему критерию («Обоснованность») высокий балл уже не может быть выставлен, поскольку в таблице критериев приводятся соответствующие численные ориентиры.

По итогам проведенных экспериментов был получен массив оценок и набор анкет с отзывами студентов. Численные данные были обезличены для возможности публичного использования и послужили основным материалом для последующих исследований. В частности, путем моделирования различных рабочих ситуаций и наборов рецензентов были получены формулы расчета оценки для первой (заменяющего ground-truth этап) и последующих пиринговых проверок, а также вычислены желательные и минимальные соотношения количества рецензентов к количеству авторов.

ПРИМЕЧАНИЯ

(1) Cornell University, Center for Teaching Excellence. Peer assessment https://www.cte.cornell.edu/teaching-ideas/assessing-student-learning/peer-assessment.html
(2) Критерии оценки эссе — Компьютерная графика 2013. Сайт поддержки курса. http://cg-2013.blogspot.ru/2013/09/blog-post_24.html
(3) iPeer. Веб-приложение для проведения пиринговой оценки. https://sourceforge.net/projects/ipeer/
(4) "Эссе-1. Оценки работ". Компьютерная графика 2013. URL: http://cg-2013.blogspot.ru/2013/10/1.html
(5) "Как проверяются рецензии". Компьютерная графика 2013. . URL: http://cg-2013.blogspot.ru/2013/10/blog-post_6.html

ЛИТЕРАТУРА

[1] *Кольбе А.С., Королев Д.А.* Исследование применимости пирингового метода оценки в очных курсах // Управление качеством: Материалы 14-й международной научно-практической конференции. М.: ПРОБЕЛ-2000, МАТИ, 2015. С. 219—229.
[2] *Est vez-Ayres I. et al.* An Algorithm for Peer Review Matching in Massive Courses for Minimising Students' Frustration //J. UCS. 2013. T. 19. №. 15. P. 2173—2197.
[3] Hames, Ethical Guidelines for Peer Reviewers. Commitee on Publication Ethics (COPE), March, 2013. № 1. URL: http://publicationethics.org/files/Peer%20review%20guidelines_0.pdf
[4] *Kloos C.D., Mu oz-Merino P.J., Alario-Hoyos C., Est vez A., Fern ndez-Panadero C.* Mixing and blending MOOC Technologies with face-to-face pedagogies, 2015 IEEE Global Engineering Education Conference (EDUCON), Tallinn, 2015. P. 967—971.
[5] *Piech C., Huang J., Chen Z., Do C., Ng A., Koller D.* Tuned models of peer assessment in MOOCs // arXiv preprint arXiv:1307.2579. 2013.
[6] *Stognieva O.* Implementing peer assessment in a russian university esp classroom / Journal of language and education. 2015. № 4 (4). С. 63—73.
[7] *Suen H.K.* Peer assessment for massive open online courses (MOOCs) // The International Review of Research in Open and Distributed Learning. 2014. 15(3).
[8] *Taylor S.M.* Can Peer Review Help Johnny Write Better? Critiquing classmates' work helps students better diagnose their own writing problems, but many do not know how to comment constructively. Here are successful peer-review strategies // Education Digest. 2014, 80(4). P. 4—10.
[9] *Wang Y., Liang Y., Liu L., Liu Y.* A Motivation Model of Peer Assessment in Programming Language Learning // arXiv preprint arXiv:1401.6113. 2014.
[10] *Williams R.* Automated essay grading: An evaluation of four conceptual models // Expanding Horizons in Teaching and Learning. Proceedings of the 10th Annual Teaching Learning Forum. Perth, Australia: Curtin University of Technology, 2001. P. 7—9.

APPLICATION OF PEER-TO-PEER ASSESSMENT OF WRITTEN WORKS OF STUDENTS IN STREAM INTERNAL COURSES

D.A. Korolev, A.S. Kolbe, A.V. Pavolotsky

National research university «Higher School of Economics»
Myasnitskaya str., 20, Moscow, Russia, 101000

At reduction of contact time in subject matters the instruments of preservation of active forms of studying for students assuming independent research work are required. One of such forms is peer-to-peer evaluation of works (peer assessment) allowing to go beyond a narrow framework of automatically

checked tests for the level of small research works. In article results of a two-year experiment on use of a piring on a stream course are given in MIEM Higher School of Economics National Research University taking into account specifics of audience rather mass online of courses.

The received results and conclusions have formed a basis for modeling of ways of estimation of written works at peer-to-peer check without participation of the teacher. In an experiment schemes of estimation and motivation for management of a ratio of number of authors and reviewers have been fulfilled, and also the general questions of use of the offered system of assessment it is aware.

Key words: peer assessment, automated assessment, classroom studies, assessment of assignments, electronic support for classroom courses

REFERENCES

[1] Kolbe A.S., Korolev D.A. *Issledovanie primenimosti piringovogo metoda ocenki v ochnyh kursah* [Research of applicability of a peer-to-peer method of assessment in internal courses]. *Upravlenie kachestvom: Materialy 14-j mezhdunarodnoj nauchno-prakticheskoj konferencii* [Quality management: Materials of the 14th international scientific and practical conference]. M.: PROBEL-2000,MATI, 2015. Pp. 219—229.

[2] *Est vez-Ayres I. et al.* An Algorithm for Peer Review Matching in Massive Courses for Minimising Students' Frustration //J. UCS. 2013. T. 19. №. 15. P. 2173—2197.

[3] Hames, Ethical Guidelines for Peer Reviewers. Commitee on Publication Ethics (COPE), March, 2013. № 1. URL: http://publicationethics.org/files/Peer%20review%20guidelines_0.pdf

[4] *Kloos C.D., Mu oz-Merino P.J., Alario-Hoyos C., Est vez A., Fern ndez-Panadero C.* Mixing and blending MOOC Technologies with face-to-face pedagogies, 2015 IEEE Global Engineering Education Conference (EDUCON), Tallinn, 2015. P. 967—971.

[5] *Piech C., Huang J., Chen Z., Do C., Ng A., Koller D.* Tuned models of peer assessment in MOOCs // arXiv preprint arXiv:1307.2579. 2013.

[6] *Stognieva O.* Implementing peer assessment in a russian university esp classroom / Journal of language and education. 2015. № 4 (4). С. 63—73.

[7] *Suen H.K.* Peer assessment for massive open online courses (MOOCs) // The International Review of Research in Open and Distributed Learning. 2014. 15(3).

[8] *Taylor S.M.* Can Peer Review Help Johnny Write Better? Critiquing classmates' work helps students better diagnose their own writing problems, but many do not know how to comment constructively. Here are successful peer-review strategies // Education Digest. 2014, 80(4). P. 4—10.

[9] *Wang Y., Liang Y., Liu L., Liu Y.* A Motivation Model of Peer Assessment in Programming Language Learning // arXiv preprint arXiv:1401.6113. 2014.

[10] *Williams R.* Automated essay grading: An evaluation of four conceptual models // Expanding Horizons in Teaching and Learning. Proceedings of the 10th Annual Teaching Learning Forum. Perth, Australia: Curtin University of Technology, 2001. P. 7—9.