Качество образовательного процесса

ПРОБЛЕМЫ ВЗАИМНОГО ОЦЕНИВАНИЯ В УЧЕБНОЙ РАБОТЕ СТУДЕНТОВ

М.П. Карпенко, *доктор технических наук*
В.А. Басов, *кандидат физико-математических наук*
Т.Ю. Семенова, *кандидат социологических наук, доцент*
А.В. Слива, *кандидат технических наук*
В.Н. Фокина, *кандидат социологических наук*

Показана важная роль взаимного оценивания в современном высшем образовании с точки зрения эффективности в учебном процессе и в плане решения проблемы оценивания возрастающего количества выполняемых обучающимися учебных заданий. Сформулированы и решены две из наиболее значимых задач, возникающих при взаимном экспертном оценивании учебной работы студентов – отбрасывание «грубых ошибок» оценивания и оценка добросовестности работы студентов как экспертов. Приведены примеры, иллюстрирующие на практике работоспособность предлагаемых решений для построения пятибалльных оценок работы экспертов.

Ключевые слова: *взаимное оценивание, эксперты, грубые ошибки измерений, математическое ожидание, дисперсия, статистическая гипотеза, недобросовестный эксперт.*

Последние 15–20 лет в практике оценки студенческих работ, особенно в университетах США, Великобритании, Ирландии, Канады и Австралии все более широко применяется взаимное оценивание (Peer Assessment) студентами различных письменных и устных работ коллег по обучению [1; 9–11]. Большинство исследователей указывает, что выступление студента в роли эксперта, оценивающего письменную или устную работу коллег по обучению, является важным и эффективным видом учебной работы в современной коллегиальной образовательной среде.

На наш взгляд преимущества взаимного оценивания в учебном процессе наиболее полно изложены в работе [2], где указано, что оно:

«• помогает студентам развить понимание собственной работы через оценивание работ других обучающихся;

• развивает важные жизненные навыки оценки и анализа;

• поддерживает независимое и самостоятельное обучение;

• дает студентам чувство сопричастности (к учебе других студентов) и, таким образом, повышает мотивацию;

• трактует оценивание как часть учебы, поэтому ошибки рассматриваются как возможности, а не признак неудачи;

• использует оценку работ других студентов для создания модели самооценки учащимся результатов собственного обучения (метапознание), стимулируя глубокое обучение, в противовес поверхностному;

• содействует распространению «учебных сообществ»;

• снижает количество оценок, выставляемых педагогами, но улучшает качество обучения;

• время для обдумывания или обсуждения с критикующим другом может помочь личности «встать в сторону» от своей собственной работы и обратить внимание на комментарии других;

• поднять внимание студентов к мероприятиям, таким, как презентации или групповые выступления коллег, когда студенты проводят их оценивание;

• обеспечивает более точную обратную связь по таким компонентам учебного процесса, как совместная работа, поскольку студенты часто лучше преподавателя могут судить об индивидуальном вкладе в работу;

• помогает уточнить критерии оценивания;

• дает студентам более широкий диапазон обратной связи с учебным процессом».

В настоящее время польза взаимного оценивания и его эффективность в учебной работе не вызывает сомнений. При этом следует отметить еще одну важную проблему, которую позволяет решить использование в учебном процессе взаимного оценивания. Это резкое увеличение в последние годы количества выполняемых студентами письменных и устных работ, ориентированных на практическое применение полученных знаний [3], что во многом связано с переходом к компетентностному обучению.

В условиях резкого роста количества аттестуемых работ студентов оказалось, что даже в традиционных вузах, становится весьма проблематичным оперативно организовать проверку необходимого количества письменных работ традиционными методами, т. е. силами преподавателей.

Ситуация усугубляется тем, что все большее количество вузов реализует электронное обучение, дистанционные образовательные технологии. В таких вузах роль и функции

преподавателей принципиально меняются: занятия в таких вузах ведутся роботами (компьютерными программами), а преподаватель взаимодействует со студентами через телекоммуникации, опосредованно, в асинхронном режиме. Преподаватель в этих условиях разрабатывает новые виды электронных занятий и включаемого в них образовательного контента, готовит ответы для асинхронных консультаций студентов по их вопросам, касающимся изучаемого предмета и т. д.

Производительность преподавателя в условиях электронного обучения, дистанционных образовательных технологий на порядки выше, чем в традиционных вузах [4]. Поэтому, по сравнению с традиционными вузами, количество преподавателей в вузах, реализующих указанные инновационные технологии, существенно ниже. Отметим, что в таких вузах количество подлежащих проверке письменных и устных работ за семестр, как например, в СГА, составляет порядка 20. При численности студентов порядка 100 000 это означает приблизительно 2 миллиона студенческих работ за семестр, что делает обеспечение их проверки традиционным способом абсолютно невозможным.

Ответом на сложившуюся ситуацию явился наблюдаемый во всем мире бум создания и внедрения автоматизированных систем оценки письменных работ студентов. В настоящее время с использованием такого рода систем, среди которых наиболее популярны Criterion (Educational Testing Service, США), Intelligent Essay Assessor (Pearson Education Technologies Inc., США), IntelliMetric (Vantage Learning, СШАhttp://www.redorbit.com/news/education/1524723/400000_essays_later_vantage_learning_spells_success_in_over_half/), Project Essay Grader (Measurement Inc., США) и другие, ежегодно оцениваются миллионы и миллионы письменных работ.

Существующие в настоящее время системы автоматизации проверки письменных работ, хотя уже и получили широкое развитие, но еще недостаточно интеллектуальны, чтобы помимо выставления оценки произвести содержательное оценивание, «разбор полетов», т. е. дать оценку с указанием студенту на допущенные ошибки и недоработки.

Не следует забывать и об устных работах – их содержательное оценивание автоматизации поддается еще сложнее. Для этого необходимо перевести устную работу в текстовый вид (автоматически, иначе теряется смысл устной работы), но при этом неизбежно добавляются ошибки распознавания речи при переводе устной информации в текстовую. Кроме того, в этом случае ситуа-

ция осложняется необходимостью включения в текст пояснительных рисунков и таблиц. В результате – те же проблемы, что и с письменными работами, но еще сложнее. Видеозапись устной работы в обозримом будущем оценивать в автоматическом режиме не представляется возможным.

В этих условиях использование взаимного оценивания работ коллег самими студентами представляется в настоящее время единственным реальным выходом, особенно в условиях электронного обучения, дистанционных образовательных технологий.

В настоящей работе не будут затронуты такие важные вопросы, как формирование критериев взаимного оценивания и его влияние на результаты обучения. Мы хотим сосредоточить внимание на вопросах оценивания качества работы студента как эксперта, оценивающего учебную работу своих коллег по «учебному сообществу». Это два взаимосвязанных вопроса: первый – отбрасывание «грубых ошибок» и второй – собственно оценка работы эксперта.

Отсеивание «грубых ошибок» оценивания

Эксперт, производящий оценку, по существу является прибором, измеряющим в данном случае качество работы оцениваемого студента. Рассмотрим оценивание результатов какой-либо учебной работы одного конкретного студента n студентами, выступающими в роли экспертов (взаимное оценивание). **Пусть $x_1, x_2,\ldots x_n$ – расположенные по не убыванию значения соответствующих оценок** – выборка объемом n результатов измерений. Как и в любых измерениях возникает проблема исключения грубых ошибок.

Наличие грубой ошибки в выборке нарушает характер распределения случайной величины X – в нашем случае оценки результатов работы студента, т.е. нарушает однородность наблюдаемой статистики измерений. Поэтому выявление грубых ошибок можно интерпретировать как проверку однородности наблюдений. Поэтому исключение грубых ошибок будем проводить на основе проверки гипотезы о том, что все элементы выборки принадлежат одной генеральной совокупности.

При этом в случае взаимного оценивания следует учитывать, что зачастую мы будем находиться в условиях малой выборки – количество студентов n, проводящих взаимное оценивание, вполне может оказаться достаточно малым, в частности $n \le 10$. Поэтому для исключения грубых ошибок следует применять специфические статистические ме-

тоды, ориентированные именно на малые выборки, такие как критерий вариационного размаха [5] ($n \geq 5$); критерий Романовского и Ирвина [6] ($n \geq 3$); вариационный критерий Диксона [7] ($n \geq 4$) и пр.

Для решения о выборе конкретного критерия из перечисленных выше был проведен эксперимент. Случайным образом были выбраны оценки 30 различных студенческих работ 10-ю экспертами, оценивавшими работы с точностью до 0,1.

В результате оказалось, что самыми мощными из рассмотренных оказались вариационный критерий Диксона и «безымянный», а самым слабым – критерий вариационного размаха. Исходя из простоты расчета, по нашему мнению, предпочтительней использовать критерий Диксона, в котором нет необходимости рассчитывать математическое ожидание и стандартное отклонение, а только размах $R = x_n - x_1$, $(x_n - x_{n-1})/R$ и $(x_2 - x_1)/R$.

Оценка работы студентов, как экспертов при взаимном оценивании

Приведем описание этого подхода на примере оценки десятью студентами-экспертами 40 студенческих работ. По каждой из этих 40 работ студенты, выполняющие взаимное оценивание (функции эксперта), дают 10 оценок по 5-бальной шкале с точностью до 1/10 балла, из которых с использованием критерия Диксона отбрасываются «грубые ошибки».

Приведем пример использования критерия Диксона. Пусть за некоторую работу эксперты выставили $\boldsymbol{k} = 10$ оценок, которые расположены по неубыванию (табл. 1).

Проверим на «грубые ошибки» – минимальное и максимальное значения оценок. В нашем примере размах выборки составляет $R = 2{,}3 = 4{,}9 - 2{,}6$. Примем для определенности уровень риска 0,1. Тогда при $k = 10$ критическое значение критерия Диксона $K_{\text{Дкр}}(0{,}1;\ k)$ равно (табл. 2) $K_{\text{Дкр}}(0{,}1;\ k) = 0{,}35$.

Проверим на грубую ошибку максимальную оценку – 4,9. Разность между 10-м и 9-м значениями в упорядоченной по возрастанию выборке (строка 2 табл. 1) равна $4{,}9 - 4{,}8 = 0{,}1$, поэтому значение критерия Диксона $К_д$ для оценки 4,9 составит

$$K_д = 0{,}1/2{,}3 = 0{,}043 < 0{,}35 = K_{\text{Дкр}}(0{,}1;\ k).$$

Отсюда следует, что на принятом уровне значимости 0,1 значение оценки 4,9 не отбрасывается.

Проверим теперь на грубую ошибку оценку 2,6. Разность между вторым и первым значениями упорядоченной по возрастанию выборки составит $4{,}0 - 2{,}6 = 1{,}4$. Тогда получим

$$K_д = 1{,}4/2{,}3 = 0{,}61 > 0{,}35 = K_{\text{Дкр}}(0{,}1;\ k).$$

Таблица 1

Пример оценивания одной студенческой работы 10-ю студентами-экспертами

Номер оценки в порядке ее возрастания *m*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_m	2,60	4,00	4,10	4,30	4,40	4,5	4,60	4,70	4,80	4,90
Номер эксперта *n*	1	10	8	4	3	2	5	6	7	9

Таблица 2

Критические значения критерия Диксона при уровне риска 0,1

k	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$K_{Дкр}(0,1; k)$	0,68	0,56	0,48	0,43	0,4	0,37	0,35

Следовательно, при выбранном уровне риска $q = 0,1$ значение оценки, равное 2,6, является грубой ошибкой и должно быть отброшено.

Аналогично проверяются на грубые ошибки оценки всех 10 студентов-экспертов по всем 10 проверяемым ими студенческим работам. Первоначально всем студентам-экспертам присваивается 40 баллов. Если у какого-либо эксперта оценка попадает в число грубых ошибок, то каждое такое попадание приводит к вычитанию из текущей оценки эксперта одного балла. Таким образом, итоговая оценка работы студентов, осуществляющих взаимное оценивание, будет находиться в промежутке от 0 до 40 баллов.

Пусть Q – количество оценок этого эксперта, признанных по критерию Диксона грубыми ошибками. Тогда определим первоначально оценку работы эксперта как

$$Z = 40–Q.$$

Теперь отобразим оценку работы студента-эксперта в 5-балльную шкалу.

В первом приближении отображение оценок Z работы эксперта в 5-балльную шкалу проведем следующим способом:

$40 \geq Z \geq 35 \rightarrow$ 5 баллов,

$34 \geq Z \geq 25 \rightarrow 4$ балла,

$24 \geq Z \geq 10 \rightarrow 3$ балла,

$Z < 10 \rightarrow 2$ балла.

Следует отметить, что наблюдения за взаимным оцениванием и анализ его практических результатов позволил выявить одну существенную особенность работы студентов в качестве экспертов. Бóльшая часть студентов ответственно относилась к оцениванию, выставляя продуманные оценки. Однако, определенная доля студентов ставила оценки, не затрудняя себя анализом работы. Это означает, что если у какого-то эксперта количество оценок, отличающихся «грубо вверх» имеет статистически незначимую разницу от количества оценок, отличающихся «грубо вниз», то этот эксперт ставит оценки «на авось», не продумывая (недобросовестный эксперт). Указанное наблюдение потребовало корректировки первого приближения оценки работы экспертов в случаях, когда количество «грубых ошибок» в их оценках достаточно велико – $Q \geq 16$ ($Z < 25$).

Если такое значительное количество «грубых ошибок» эксперт допустил, делая их «на авось», т. е. не продумывая оценку, то пятибалльная оценка работы эксперта должна быть уточнена в сторону снижения:

$Z < 25 \rightarrow 2$ балла.

Если же окажется, что крен в сторону завышения или занижения оценки у какого-то эксперта является статистически значимым, то это означает, что выдаваемое им оценивание студентов представляет собой некоторую позицию эксперта, пусть отличную от других, но базирующуюся на неких неслучайных основах. В этом случае пятибалльная оценка работы эксперта, по нашему мнению, должна быть уточнена в сторону повышения:

$24 \geq Z \geq 10 \rightarrow 4$ балла,

$Z < 10 \rightarrow 3$ балла.

Для завершения построения метода оценки работы студентов-экспертов при взаимном оценивании остается изложить решение задачи проверки случайности/неслучайности различия количества грубых «завышений» и «занижений» при оценивании студенческих работ.

Пусть число грубых ошибок эксперта $40 \geq n \geq 16$, что отражает случаи, когда необходима корректировка первичной пятибалльной оценки эксперта. Введем случайную величину – V, такую, что $Vi = -1$ – если эксперт грубо занизил оценку и $Vi = 1$ – если грубо завысил.

В предположении случайности проставления экспертом заниженных и завышенных оценок (не задумываясь), среднее значение $V = 0$. Выборочное среднее распределено около него по нормальному закону с неизвестной дисперсией.

Поэтому задача проверки случайности выставления экспертом

заниженных и завышенных оценок математически сводится к проверке статистической гипотезы равенства математического ожидания нормального распределения нулю при неизвестной дисперсии.

В этом случае рекомендуется использовать критерий [8] $T=\overline{V}\sqrt{n}/s$ (s – **несмещенная оценка дисперсии** V), имеющий **распределение Стъюдента с $n-1$ степенью свободы**.

Пусть из n грубых ошибок из которых $n1$ **– в сторону занижения** и $n2$ **– в сторону завышения** ($n1+n2=n$). Выборочное среднее $\overline{V}$ в этом случае получим как $(n2-n1)/n$.

Тогда получим **несмещенную** оценку D выборочной дисперсии V:

$$D=\frac{n1\left(\frac{-(n-n1)-n2}{n}\right)^2+n2\left(\frac{(n-n2)+n1}{n}\right)^2}{n-1}.$$

Таким образом получим:

$$D=4\frac{n1n2(n2+n1)}{(n-1)n^2}.$$

Отсюда

$$s=\sqrt{D}=2\sqrt{\frac{n1n2}{n(n-1)}}.$$

Теперь получаем, что выборочное значение критерия T вычисляется следующим образом

$$T=(n2-n1)/n\times\sqrt{n}\frac{2}{\sqrt{\frac{n1\times n2}{n(n-1)}}}=\frac{n2-n1}{2}\sqrt{\frac{n-1}{n1\times n2}}.$$

Если при заданном уровне риска $|T|<T_{кр}$ (двусторонняя критическая область), то гипотеза равенства 0 математического ожидания распределения V принимается **(т.е. считается, что грубые ошибки вверх и вниз эксперт делал случайным образом),** иначе – отвергается **(т. е. существенные отличия в оценках эксперта – это его позиция).**

Выберем, например, уровень значимости (риска) $q=0{,}1$. По таблице критических значений (точек) распределения Стъюдента для этого уровня значимости сформируем табл. 3 этих критических точек для

Таблица 3

N	**16**	**17**	**18**	**19–21**	**22**	**23**	**24–28**	**29–31**	**32–38**	**39**	**40**
$T_{кр}$	1,75	1,75	1,74	1,73	1,72	1,72	1,71	1,70	1,69	1,68	1,68

$k = n–1$ степени свободы (соответственно, для $n = k + 1$, n от 16 до 39).

ПРИМЕРЫ.

1. Пусть эксперт сделал $\boldsymbol{n = 30}$ грубых ошибок оценивания, из которых $\boldsymbol{n1 = 20}$ – занижения и $\boldsymbol{n2 = 10}$ – завышения. Из табл. 3 получим $T_{кр} = 1,7$

$$T = \frac{10-20}{2}\sqrt{\frac{29}{10\times 20}} = -1,9.$$

То есть $|T| = 1,9 > 1,7$.

В данном случае имеем дело с неслучайной позицией эксперта. Ему следует выставить оценку 4 ($Z = 40–30 = 10$).

2. Пусть эксперт сделал $\boldsymbol{n = 20}$ грубых ошибок оценивания, из которых $\boldsymbol{n1 = 8}$ – занижения и $\boldsymbol{n2 = 12}$ – завышения. Из табл. 3 получим $T_{кр} = 1,73$

$$|T| = \frac{12-8}{2}\sqrt{\frac{19}{8\times 12}} = 0,88 < 1,73.$$

Следовательно, у рассматриваемого эксперта имеем дело со случайным различием количества заниженных и завышенных оценок работы студентов за веб-инар. В этом случае, поскольку $\boldsymbol{Z} = 40–20 = 20 < \boldsymbol{25}$, то эксперт за оценивание получит оценку 2 балла.

В заключение отметим, что пороговые значения в рассмотренном оценивании экспертов-студентов – 10, 25 и 35 баллов – выбирались экспертами-преподавателями и при оценивании различных специфических студенческих работ могут пересматриваться. Эти значения не являются принципиальными, и предлагаемый подход к оценке работы экспертов-студентов при взаимном оценивании носит достаточно общий характер и, по нашему мнению, может применяться для самых различных студенческих работ.

Литература

1. Maeve Foreman. Peer Assessment of Problem Based Learning – Fostering Reflective Practice in Social Work Students. School of Social Work and Social Policy, Trinity College, College Green, Dublin 2, Ireland AISHE Readings: 2007, Number 1. http://www.aishe.org/readings/2007-1/No-20.html.

2. Peer and self assessment. http://www.teachingexpertise.com/articles/peer-and-self-assessment-2867.

3. Карпенко М.П. Качество высшего образования. М.: Изд-во СГУ, 2012.

4. Тихонов А.Н. и др. Управление современным образованием. М.: Вита, 1998.

5. Третьяк Л.Н. Обработка результатов наблюдений: Учеб. пособие. Оренбург: ГОУ ОГУ, 2004.

6. Метод исключения резко выделяющихся значений результатов испытаний. ГОСТ 10518-88 [Электронный ресурс] // Режим доступа: http://www.docload.ru/Basesdoc/8/8480/index.htm

7. Сергеев А.Г., Крохин В.В. Метрология. М.: Логос, 2001.

8. Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высшая школа, 1998.

9. C. Philip Wheater, A. Mark Langan and Peter J. Dunleavy. Department of Environmental and Geographical Sciences, Manchester Metropolitan University. Students assessing student: case studies on peer assessment. Planet. No. 15. December 2005.

10. Gloria Yi-Ming Kao. Enhancing the quality of peer review by reducing student "free riding": Peer assessment with positive interdependence. British Journal of Educational Technology. Volume 44, Issue 1, pages 112–124, January 2013.

11. Glyn Thomas, Dona Martin, Kathleen Pleasants. Using self- and peer-assessment to enhance students' future-learning in higher education. La Trobe University, Volume 8. Issue 1, 2011, http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1112&context=jutlp

THE QUALITY OF THE EDUCATIONAL PROCESS

Karpenko M.P., *Doctor in Technical Sciences, Modern University for the Humanities*
Basov V.A., *PhD in physico-mathematical sciences, Modern University for the Humanities*
Semenova T. Ju., *PhD in Sociology, Modern University for the Humanities*
Sliva A.V., *PhD in Technical Sciences, Modern University for the Humanities*
Fokina V.N., *PhD in Sociology, Modern University for the Humanities*

Problems of Peer Assessment in Academic Work of Students

The article shows the importance of peer assessment in modern higher education both in terms of effectiveness in the learning process, and in terms of solving the problem of estimating the increased number of training tasks performed by trainees. Formulated and solved two most important problems arising in mutual expert evaluation of academic students work: drop "blunders" of evaluation and assessment of the integrity of the students as experts. Some examples are given to display practical performance of proposed solutions for the formation of a five-point evaluations of experts.

Key words: *peer assessment, experts blunders measurements expectation, variance, sampling, statistics, statistical hypothesis, unscrupulous expert.*