# Экспериментальное исследование расхождения отметок, выставляемых студенческим работам преподавателями

**Михаил Петрович Карпенко,**
доктор технических наук, профессор,
президент Современной гуманитарной академии.

В результате экспериментального исследования установлено, что расхождения отметок, выставляемых студенческим работам преподавателями-асессорами, чрезмерно велики и объясняются, во-первых, индивидуальными психосоциальными характеристиками преподавателей (систематическая часть отклонения от условно объективных отметок), и во-вторых, остаточным разбросом (случайная часть отклонения), подчиняющимся закону нормального распределения. Предложен метод введения поправок в отметки, значительно нивелирующий влияние индивидуальных психосоциальных характеристик, и критерий количественной оценки профессионализма преподавателей-асессоров. Сделан вывод о желательности комиссионной оценки студенческих работ или применения алгоритмических методов количественной оценки.

*Ключевые слова*: студенческая работа, отметка, преподаватель-асессор, психосоциальные характеристики, препозиции, отклонения отметок, случайный разброс, стандарт отклонения.

Выставление обучающимся отметок за выполнение учебных заданий и в ходе различного рода аттестаций является основной, а для традиционных дидактик – единственной мерой влияния преподавателя на учебный процесс, обеспечения контроля усвоения учащимися учебного материала и продвижения по учебному плану. Следует заметить, что в российской педагогической практике термины «отметка» и «оценка» обычно

воспринимаются как синонимы, но в практике англоязычных школ термин «оценка» (асессмент) понимается более широко и может включать в себя поликритериальный анализ ученической работы, который может быть расширен до рецензии, тогда как термин «отметка» (грейдинг) употребляется для выражения однозначной, как правило, цифровой интерпретации уровня качества работ или ответов учащихся.

При этом необходимо учитывать типы учебных дисциплин и разделить их на естественнонаучные или технические (в которых решаются числовые задачи и значения искомых величин или параметров имеют однозначную трактовку, а, следовательно, влияние индивидуальных внутренних оценочных шкал преподавателей невелико или вообще отсутствует) и дисциплины гуманитарного типа, в которых преподаватели выступают в роли экспертов (асессоров – в англоязычной терминологии), осуществляют асессмент.

К необходимости определения объективных и субъективных факторов, влияющих на оценки преподавателей, приводят все расширяющиеся масштабы применения в образовательной практике асессмента - взаимной оценки работ студентами. Распространение исследований на преподавателей вызвано двумя существенными причинами.

Первая из них. Масштабными экспериментальными исследованиями, проведенными в 2015-2016гг. в Современной гуманитарной академии (в эксперименте приняло участие 16 тысяч студентов бакалавриата), было обнаружено явление препозиции асессоров – устойчивого настроя внутренних оценочных шкал асессоров на завышение отметок (оверстейтеры), либо на их занижение (лоустейтеры), причем у различных асессоров наблюдалась разная степень препозиции.[1] Нет никаких оснований предполагать, что явление препозиции не распространяется на преподавателей, а оно до последнего времени не исследовалось.

Во-вторых, по литературным данным, имеется ряд предложений проводить калибровку или ранжирование студентов-асессоров [2], но поскольку проблему отсутствия эталонов студенческих работ определенного

качества [3] приходится решать, принимая за эталоны отметки, выставляемые отдельными преподавателями [4], то надо установить релевантность этих эталонов.

С целью оценки типовых психосоциальных характеристик преподавателей-асессоров в Современной гуманитарной академии в 2018 году был проведен эксперимент.

На добровольной основе в эксперименте в качестве асессоров, выставляющих отметки студенческим работам, приняли участие 38 профессиональных преподавателей высшей школы. Все они имели педагогический стаж более 10 лет и ученую степень, 9 преподавателей имели степень доктора наук и звание профессора. Среди асессоров было 22 женщины и 16 мужчин.

Для оценки были представлены 40 актуальных студенческих творческих работ в виде кратких рефератов и эссе различного типа по 20-ти учебным дисциплинам, которые условно можно представить в виде трех блоков: психолого-педагогический, экономико-управленческий и правовой. Студенческие работы для эксперимента отбирались методом случайной выборки.

Каждому асессору, как правило, давалось на оценку 10 работ, из них 6 по его узкой специальности и 4 из близкого ему блока, но по дисциплине, в которой, строго говоря, данный асессор не являлся специалистом. В соответствии с этим отметки, выставляемые работам, имели код «узкий специалист» (с дисциплинарной специализацией) / «широкий специалист». Для равномерного (с учетом действия различных факторов) распределения работ между асессорами была разработана специальная методика.

Работы предоставлялись асессорам анонимно. Для отметок применялась традиционная для российской педагогики балльная шкала с крайними значениями баллов: 5 – высший балл, 2 – низший балл. Крайние баллы соответствуют часто применяемой международной процентной шкале:

100% - 0%. Для пересчета балльной шкалы в процентную применялась формула:

$$a(ij)\% = [a(ij)балл - 2] \times 33{,}33,$$

где $a(ij)$ - отметка, выставленная *i*-той работе *j*-тым асессором.

Работы оценивались и отметки выставлялись по следующим предположительно независимым критериям:

1. последовательность и логика изложения материала;

2. степень раскрытия заданной темы реферата, эссе.

Эти два критерия рекомендованы асессорами, которые в своей педагогической практике считают их наиболее важными при оценке качества студенческой творческой работы.

К этим критериям были добавлены еще три критерия, необходимых для градуировки оценивающих автоматов при использовании цифровых технологий на основе лингвистических подходов:

3. профессионализм;

4. общекультурный уровень;

5. научный уровень.

Кроме того, асессорами было предложено дать общую оценку работы, исходя из общего впечатления после ее прочтения:

6. общая оценка работы.

Примерно такие же критерии применяют для определения качества студенческих работ зарубежные исследователи [5].

Дальнейший анализ показал, что только 26,5% асессоров воспользовалось представленной им возможностью дать общую оценку, отвлекаясь от частных критериев, а 73,5% проставили общую отметку как среднюю арифметическую из пяти ранее выставленных отметок по отдельным критериям.

Таким образом, для всех работ отметки выставлялись по шести критериям. Всего было выставлено 2148 отметок.

Все дальнейшие расчеты проводились в процентной шкале (0 - 100).

Исследовалась попарная корреляция отметок по всем шести критериям. Было установлено, что резкие колебания (флуктации) отметок по критериям присутствуют только у трех из 38 асессоров (8%). В результате корреляционного анализа была получена матрица (табл. 1) средних по всем асессорам значений коэффициентов корреляции отметок, выставленных по выше приведенным критериям.

Таблица 1

| NN | КРИТЕРИИ | Коэффициент корреляции по NN критериев | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | Последовательность и логика изложения материала | 1,000 | 0,779 | 0,708 | 0,720 | 0,749 | 0,865 |
| 2 | Степень раскрытия темы | | 1,000 | 0,744 | 0,742 | 0,744 | 0,897 |
| 3 | Профессионализм | | | 1,000 | 0,778 | 0,817 | 0,863 |
| 4 | Общекультурный уровень | | | | 1,000 | 0,823 | 0,847 |
| 5 | Научный уровень | | | | | 1,000 | 0,865 |
| 6 | Отметка по общему впечатлению | | | | | | 1,000 |

Полученные данные показывают очень высокую степень корреляции отметок по всем парам критериев. Чем объяснить этот факт? Вероятно, человеческой психике вообще свойственно давать явлениям и событиям обобщенные интегральные оценки.

Поскольку в своей массе отметки по различным критериям близки друг к другу, все дальнейшие расчеты велись по интегральному критерию, представляющему собой среднее арифметическое по шести частным критериям, в дальнейшем все расчеты велись по интегральному критерию.

В основе расчета находится исходная матрица отметок: *i* (номера работ) – *j* (номера асессоров). Математическая обработка проведена двумя ветвями: ветвь работ и ветвь асессоров.

Расчет по ветви работ.

$$\bar{a}(i)=\frac{1}{n(i)}\sum a(ij), \qquad (1)$$

где: $\bar{a}(i)$ - среднеарифметическая отметка *i*-той работы; $n(i)$ - количество отметок, полученных *i*-той работой от различных асессоров; $a(ij)$ - отметка, выставленная *i*-той работе *j*-тым асессором.

$$\delta(ij)=a(ij)-\bar{a}(i), \qquad (2)$$

где: $\delta(ij)$ - отклонения от средней отметки исходных отметок, выставленных *i*-той работе *j*-тым асессором.

$$\left.\begin{aligned}\bar{\delta}(i)&=\frac{1}{n(i)}\sum\delta(ij)\\ \sigma(i)&=\sqrt{\frac{1}{n(i)}\sum\left[\delta(ij)-\bar{\delta}(i)\right]^2}\end{aligned}\right\}, \qquad (3)$$

где: $\bar{\delta}(i)$ - среднее арифметическое отклонение отметок i-той работы, выставленных j-тыми асессорами; $\sigma(i)$ - среднее квадратическое отклонение, рассчитанное по величинам $\bar{\delta}(i)$.

Изучение средних квадратических отклонений [$\sigma(i)$] (стандартов отклонения) показало, что их можно считать реализациями случайной величины, распределенной по закону, близкому к нормальному.

По критерию Пирсона была проведена проверка нормального распределения случайной величины. Для 7-ми интервалов (число степеней свободы К=7 – 3= 4) и обычно применяемого уровня значимости (α =0.05) критериальное значение $\chi^2$ = 9,5. Рассчитано эмпирическое значение $\chi^2$ = 8,62,то есть меньше критериального, что подтверждает гипотезу о нормальном распределении случайной величины (стандартов расхождения отметок).

Рассчитано математическое ожидание (средняя величина) стандарта отклонения, оно оказалась равным

$$\bar{\sigma}(i) = 22{,}5\%$$ (0,68 балла в 5-балльной шкале).

В соответствии с правилом «трех сигм» (3-sigma rule) можно утверждать, что расхождение отметок асессоров-преподавателей может достигать

$$3\bar{\sigma}(i) = 67{,}5\%$$ (более 2 баллов в 5-балльной шкале).

Данный результат показывает ненадежность экспертных оценок, даваемых даже опытными преподавателями, имеющими ученые степени и звания. Мы стоим перед необходимостью замены субъективных экспертных решений более объективными - формальными, вырабатываемыми тщательно сконструированными алгоритмами, или же введения комиссионных оценочных процедур, в значительной степени нивелирующих эффект действия «человеческого фактора».

С другой стороны, заметим, что интеллектуальный робот (не искусственный интеллект, бесконтрольно подбирающий сам алгоритмы своей работы и принятия решений, - с ним еще надо разбираться), выставляя оценки, может опираться только на заложенные в него алгоритмы, все действия робота и принимаемые им решения детерминированы и не являются случайными величинами. А, следовательно, в качестве асессора робот совершенно надежен и может работать один (другое дело, что алгоритм и критериальная система, им используемые, должны быть достаточно совершенны).

В продолжение расчета определялось действие гендерного фактора.

$$\frac{\bar{a}(ij)(жен.) - \bar{a}(ij)(муж.)}{\bar{a}(ij)(жен)} * 100\% = 2\%$$ (0,06 балла).

В среднем отметки, выставленные женщинами, выше отметок, выставленных мужчинами, но эта разница настолько мала (в 11 раз меньше

$\bar{\sigma}(i)$), что позволяет сделать вывод об отсутствии влияния гендерного фактора на оценки студенческих работ.

Исследовалось влияние фактора «дисциплинарной специализации». Получен результат

$$\frac{\bar{a}(ij)(шир.спец) - \bar{a}(ij)(дисп.спец)}{\bar{a}(ij)(шир.спец)} * 100\% = 3{,}3\% \text{ (0,1 балла)}$$

В среднем отметки, выставленные «широкими специалистами», выше отметок, выставленных «дисциплинарными специалистами», но эта разница настолько мала [в 6,5 раз меньше $\bar{\sigma}(i)$], что позволяет сделать вывод о незначительном влиянии фактора «дисциплинарной специализации» асессоров на оценки студенческих работ.

Расчет по ветви асессоров.

На первом этапе задействования ветви асессоров, отходящей от исходной матрицы, для каждого асессора по каждой, проверяемой им работе, рассчитываются средние пир- отметки – то есть отметки, выставленные за эту работу другими асессорами-коллегами.

В сфере образования термин «пир» обычно употребляется для обозначения взаимных оценок обучающихся, сверстников, но здесь данный термин использован, чтобы подчеркнуть, что асессоры являются коллегами, находящимися в равном положении друг к другу.

$$\bar{A}_p(i) = \frac{1}{N(i)-1} \sum A_p(ij), \qquad (4)$$

где: $\bar{A}_p(i)$ - средняя пир - отметка *i*-той работы; $N\ (i)$ - количество асессоров, участвующих в оценивании i-ой работы; $A_p(ij)$ - отметка *i*-той работы, выставленная *j*-тым пир-асессором.

Для дальнейших рассуждений введем понятие «диссектная» (англ. – dissect – анализировать, критически разбирать) для отметки, выставленной асессором, деятельность которого в данный момент анализируется. Соответственно: дис-отметка, $A_d(ij)$.

На втором этапе рассчитывается матрица отклонений дис-отметок, выставленных *j*-тым асессором (дис-асессором) *i*-той работе, от средней пир-отметки.

$$\Delta_d(ij) = A_d(ij) - \overline{A}_p(i) \quad , \qquad (5)$$

где: $\Delta_d(ij)$ - отклонение дис-отметки *j*-того асессора от средней пир-отметки *i*-той работы; $A_d(ij)$ - дис-отметка *j*-того асессора по *i*-той работе; $\overline{A}_p(i)$ - по формуле (4).

Общее рассеяние точек можно объяснить действием двух факторов:

- фактор препозиции, по-видимому, связанный с социо-психологическим настроем асессора, влияние которого в первом приближении можно выразить линейным трендом;

- остаточный случайный разброс отметок, отражающий социо-психологическую неустойчивость асессора, а может быть его небрежность, недостаточную сформированность внутренних шкал оценки студенческих работ.

С целью исследования систематического фактора препозиции для каждого асессора были определены дополнительные величины:

$$\left.\begin{aligned} \overline{\Delta_d}(j) &= \frac{1}{N(j)} \sum \Delta_d(ij) \\ \overline{\omega}(j) &= \frac{1}{N(j)} \sum \left[\Delta_d(ij) / A_d(ij)\right] \end{aligned}\right\} \qquad (6)$$

где $\overline{\Delta_d}(j)$ - среднее отклонение отметок, выставленных $j$-тым асессором; $\overline{\omega}(j)$ - среднее отношение отклонения дис-отметки к ее величине (удельное отклонение); $N(j)$, $\Delta_d(ij)$, $A_d(ij)$, - по формулам (4), (5).

По этим данным построен график, представленный на рис. 1.

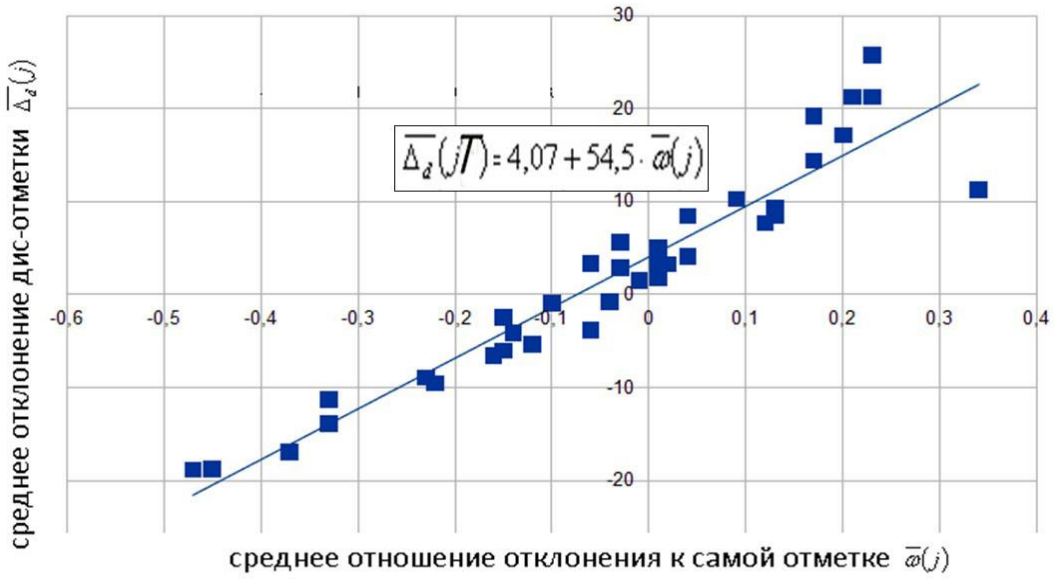



Рис. 1 График зависимости среднего отклонения дис-отметки $\left[\overline{\Delta}_d(j)\right]$ от среднего отношения отклонения к самой отметке $\left[\overline{\omega}(j)\right]$.

Вид поля точек показывает, что они достаточно тесно группируются вблизи линии общего тренда. При этом знак среднего отклонения дис-отметки от пир-отметки убедительно делит всех асессоров на две группы: оверстейтеров (плюсовое отклонение), завышающих отметки, и лоустейтеров (минусовое отклонение), занижающих отметки. Однако степень их препозиции различна. Можно выделить в центре тренда группу асессоров, у которых отклонения невелики – например, не превышают по абсолютной величине 5 % (процентная шкала). Это асессоры с нейтральной препозицией.

Отклонения до 15 % можно считать ординарными препозициями, а свыше 15 % – экстроординарными препозициями. Результаты расчетов по полученному в эксперименте сочетанию препозиций различной степени приведены в таблице 2.

Таблица 2

| ПРЕПОЗИЦИИ АСЕССОРОВ | | | | |
|---|---|---|---|---|
| Лоустейтеры | | Нейтральная | Оверстейтеры | |
| Экстраординарные | Ординарные | | Ординарные | Экстраординарные |

| | | | | |
|---|---|---|---|---|
| | | препозиция | | |
| 3 чел. | 7 чел. | 15 чел. | 8 чел. | 5 чел. |
| 8%∑N(j) | 18%∑N(j) | 40%∑N(j) | 21%∑N(j) | 13%∑N(j) |

В качестве критерия работы асессора логично принять стандарт отклонений [*σ(j)*] выставленных им дис-отметок, который отражает действие двух факторов – препозиции и случайной диссипации (все величины уже представлены в предыдущей формуле).

$$\sigma(j)=\sqrt{\frac{1}{N(j)}\sum\Delta_d^2(ij)} \qquad (7)$$

Компенсировать случайную вариабельность отметок при традиционной дидактике можно только комиссионным способом – увеличивая число асессоров, выставляющих отметки одной и той же работе. Поскольку, как было показано выше, случайная величина (разброс отметок) подчиняется нормальному закону распределения, то можно ожидать, что стандарт отклонения будет снижаться пропорционально корню квадратному из числа асессоров.

Но действие фактора препозиции можно устранить, введя поправки в фактические отметки, выставленные асессорами. Приведем математический аппарат для расчета поправок.

$$K=1-\bar{\omega}(j), \qquad (8)$$

где *K* - поправочный коэффициент; $\bar{\omega}(j)$ – по формуле (6).

Далее вычисляются все исправленные показатели.

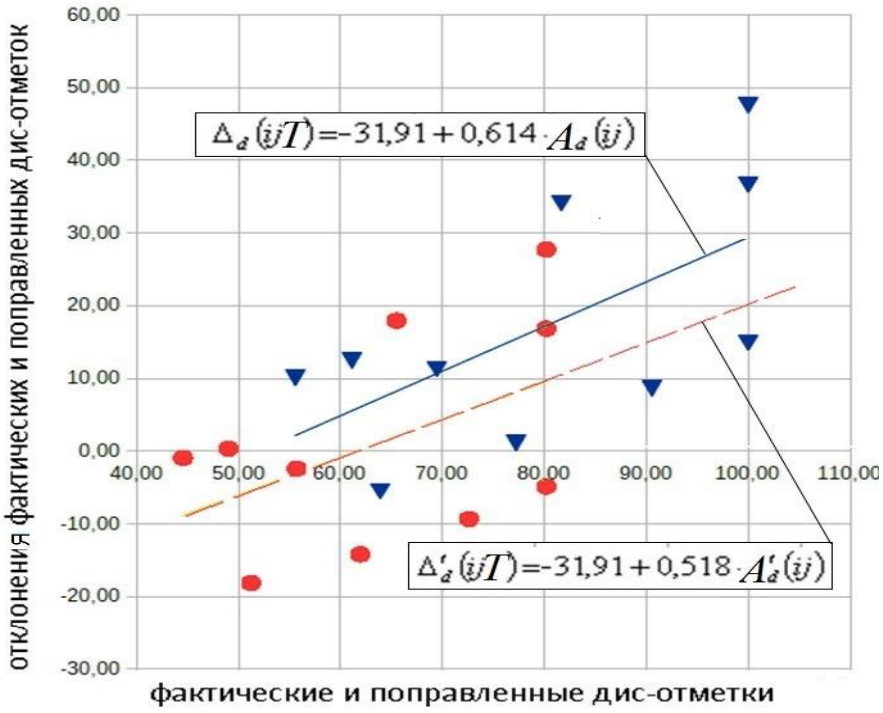
$$\left.\begin{aligned} A_d'(ij)&=K\cdot A_d(ij)\\ \Delta'(ij)&=A_d'(ij)-A_d(ij)+\Delta(ij)\omega\\ \sigma'(j)&=\sqrt{\frac{1}{N(j)}\sum\Delta'(ij)^2}\end{aligned}\right\} \qquad (9)$$

где $A_d^{'}(ij)$ – поправленная дис-отметка $i$-той работе $j$-того асессора; $\Delta'(ij)$ – отклонение поправленной дис-отметки; $\sigma'(j)$ – стандарт отклонения поправленных отметок $j$-того асессора.

Можно считать, что именно эта величина $[\sigma'(j)]$ отражает действие фактора остаточного случайного разброса. Зная общий критерий качества $[\sigma(j)]$ и критерий случайного разброса $[\sigma'(j)]$, можно определить критерий препозиции $[\sigma_{\text{пр}}(j)]$ в виде

$$\sigma_{\text{пр}}(j) = \sqrt{\sigma^2(j) - \sigma'^2(j)} \qquad (10)$$

В качестве примера проанализируем работу асессора № 1, графически представленную на рис. 2.




▼ – фактические значения дис-отметок, $A_d(ij)$.

● – поправленные значения дис-отметок, $A_d^{'}(ij)$.

Рис. 2. Графическое представление фактически выставленных асессором и поправленных дис-отметок.

Как видно из представленных данных, удалось установить, что асессор №1 является оверстейтером (положительная препозиция) и в среднем завышает отметки. Чтобы устранить систематическую составляющую оверстейтера, рассчитан поправочный коэффициент к его отметкам, $K = 0{,}802$ . Переход на поправленные отметки позволил уменьшить среднее отклонение от пир-отметок в 13 раз (1,30 вместо 17,14) и разброс отклонений на 13% (14,93 вместо 16,82).

Аналогичные расчеты были проведены по данным всех 38 асессоров. При этом отметки с экстраординарными отклонениями (более 27 % в процентной шкале) не учитывались при расчете средних пир-отметок.

Для каждого асессора были определены поправочные коэффициенты для устранения систематической (препозиционной) составляющей отклонений дис-отметок.

Сумма отклонений от средних пир-отметок оценивались по абсолютной величине, чтобы положительные и отрицательные взаимно не гасились – знак препозиции определялся по удельному отклонению. Средние отклонения поправленных дис-отметок уменьшились в 2,5 раза (3,66 вместо 8,99).

Уменьшился и разброс поправленных дис-отметок (случайная составляющая) средний стандарт отклонения сократился на 45% (14,97 вместо 21,67).

Стандарты поправленных отклонений, рассчитанные для отдельных асессоров, распределены по нормальному закону, что было проверено по критерию Пирсона. Критериальная величина ($\alpha = 0{,}05$ , к = 5-3=2) составила $\chi^2_{кр} = 6{,}0$ , тогда как эмпирическая меньше ($\chi^2_{эмп} = 4{,}72$).

Таким образом, введение поправок в фактически выставленные асессорами отметки позволили заметно приблизить их к более надежным средним пир-отметкам и практически устранить действие систематического фактора.

В результате проведенного эксперимента можно сделать следующие выводы:

1. Установлено, что индивидуальные препозиции и случайный разброс приводят к тому, что выставляемые учебным работам отметки могут отклоняться от объективно правильных на недопустимо большие величины. Бороться с этим явлением можно введением комиссионного метода оценивания работ и (или) разработкой алгоритмов оценивания по формальным критериям, что позволяет применять роботы для оценивания студенческих работ.

2. Установлено, что как правило, асессоры, даже в лице опытных и квалифицированных преподавателей, не используют поликритериальные системы оценивания, предпочитая применять интегральную оценку учебных работ.

3. Установлено, что сторонние факторы (исследовался гендерный фактор и фактор дисциплинарной специализации асессоров) практически не влияют на оценивание учебных работ.

4. Разработан понятийный и математический аппарат обработки статистических данных оценивания учебных работ с получением показателей, определяющих одновременно качество работ и качество оценивания., что особенно важно при использовании в учебном процессе коллегиальной среды.

5. Установлены примерные (т.к. объем статистики невелик) статистические показатели экспертной оценки учебных работ (деление асессоров на лоустейтеров, оверстейтеров и нейтральных препозиционеров, параметры функций распределения показателей процесса оценивания учебных работ).

6. Разработаны численные показатели профессионализма асессоров с учетом действия факторов препозиции и случайного разброса.

7. Разработан метод определения необходимости и расчета индивидуальных поправок в фактически выставленные асессорами отметки за учебные работы.

## Литература

1. Карпенко М.П. Экспериментальное исследование социально-психологических характеристик студентов по результатам массового ассессмента // Психология обучения. 2017.№ 10 с. 5-21.

2. Piech C. Tuned models of peer assessment in MOOCs. arxiv preprint arxiv: 1307.2579.2013.

3. Michael Wong. Online Peer Assessment in MOOCs: Students Learning from Students. March 28, 2013, URL: http://ctlt. ubc.ca/2013/03/28/online-peer-assessment-in-moocs-students-learning-from-students (дата обращения 19.04.2017).

4. Edx. Open Response Assessments. URL:https://edx.readthedocs.io/projects/open-edx-building- and- running- a-course/en/named- release- birch/exercises- tools/open_ response_assessments/index.html (дата обращения 04.07.2017)

5. Peer and Self Assessment of Student Work. URL:http://www.ryerson.ca/content/dam/It/recources/handouts StudentPeerAssessment.pdf (дата обращения 19.05.2017).