УДК: 159.9.018
doi: 10.11621/vsp.2020.04.03

ТОЧНОСТЬ ЭКСПЕРТНОЙ ТЕХНОЛОГИИ ОБНАРУЖЕНИЯ МОШЕННИЧЕСТВА НА УДАЛЕННЫХ ТЕСТОВЫХ ЭКЗАМЕНАХ (ПРОКТОРИНГ)

А.Г. Шмелев

Факультет психологии МГУ имени М.В. Ломоносова, Москва, Россия.
Для контактов. E-mail: ags12@ht.ru

Актуальность и цель работы. В данном эмпирическом исследовании предпринята попытка измерения точности обнаружения мошенничества («читинга») на основе экспертных оценок с использованием видеозаписей хода выполнения тестовых экзаменов в режиме онлайн.

Выборка. 35 испытуемых выполняли в режиме онлайн тест на эрудицию из 30 заданий с выбором из 4-х вариантов ответа. Половина испытуемых (18 человек) были «подставными читерами» — использовали шпаргалки с правильными ответами.

Ход и методы. Видеозапись процесса тестирования включала «захват экрана», на котором наблюдатель-эксперт («проктор») мог наблюдать все перемещения курсора, видеть запись мимики и зрительный фокус внимания испытуемого в отдельном окне (запись с фронтальной камеры), слышать речевое проговаривание испытуемым условий задания и ответов («устное решение»). В оценивании видеозаписей приняло участие 14 экспертов, из которых 8 показали удовлетворительные результаты по уровню точности в обнаружении читинга (точность-accuracy, измеренная с помощью коэффициента Карра, у них была выше 0,5).

Выводы. Выявлена высокая асимметричная прогностичность экспертных оценок: более точные эксперты допускают пренебрежимо малое (около 5 процентов) число ошибок типа «ложная тревога», но сравнительно много ошибок типа «пропуск». Сделаны рекомендации по практическому использованию методики экспертной оценки в сочетании с автоматическим хронометрическим анализом степени атипичности протоколов и последующим контрольным очным тестированием всех заподозренных испытуемых (экзаменуемых).

Ключевые слова: онлайн-обучение, онлайн-тестирование, онлайн-экзамены, метод независимых судей, точность экспертных оценок, надежность экспертных оценок, конкордация экспертных оценок, достоверность тестирования, фальсификация при тестировании, прокторинг, читинг.

Для цитирования: *Шмелев А.Г.* Точность экспертной технологии обнаружения мошенничества на удаленных тестовых экзаменах (прокторинг) // Вестник Московского университета. Серия 14. Психология. 2020. № 4. С. 44–66. doi: 10.11621/vsp.2020.04.03

Поступила в редакцию: 17.07.2020 / Принята к публикации: 24.08.2020

ACCURACY OF EXPERT FRAUD DETECTION TECHNOLOGY IN REMOTE TEST EXAMS (PROCTORING)

Alexander G. Shmelev

Lomonosov Moscow State University, Moscow, Russia.
Corresponding author. E-mail: ags12@ht.ru

The purpose (objective) of the empirical study is the measurement of the accuracy of expert-proctors in detecting cheating in online testing.

Sample of the study. 35 test takers passed an online test of general knowledge on the basis of 30 multiple choice questions. Half of the subjects (18 persons) were "artificial cheaters" — they used cheat sheets with correct answers.

Methods. The video recording of the testing process included a "screen capture" so that expert-proctors could observe all cursor movements, see a recording of the subject's facial expressions and a visual focus of attention in a separate window (recording from the front camera), and could listen to the subject pronouncing the task conditions and answers ("oral decision"). 14 experts took part in rating of video recordings, of which 8 experts showed satisfactory results in terms of the level of accuracy in detecting cheating (their accuracy that was measured using the Kappa coefficient was higher than 0.5).

Conclusions. A high asymmetric validity of expert assessments is revealed. More accurate experts allowed a negligible (about 5 percent) number of errors of the "false alarm" type, but a relatively large number of errors of the "skip" type. Recommendations are made for the practical use of the expert assessment method in combination with automatic chronometric analysis of the degree of atypical protocols and subsequent control of face-to-face offline testing of all suspected subjects (examinees).

Keywords: online training, online testing, online exams, method of independent judges, the accuracy of expert ratings, the reliability of expert ratings, concordation of expert ratings, reliability of testing, falsification upon testing, proctoring, cheating.

For citation: Shmelev, A.G. (2020) Accuracy of expert fraud detection technology in remote test exams (proctoring). *Vestnik Moskovskogo Universiteta. Seriya 14. Psikhologiya [Moscow University Psychology Bulletin], 4,* P. 44–66. doi: 10.11621/vsp.2020.04.03

Received: July 17, 2020 / Accepted: August 24, 2020

Введение: теоретико-методический и прикладный контекст исследования

Стремительный прогресс в области информационно-коммуникационных технологий приводит к парадоксальным ситуациям. Яркий пример — прокторинг. Эта технология видеонаблюдения за выполнением удаленного экзамена (экзамена в режиме онлайн) появилась в нулевые годы. Первая компьютерная прикладная система на английском языке создана была в 2008 году (ProctorU, 2008), и в настоящее время множество подобных софтверных систем уже получили широкое распространение в практике современного электронного (цифрового) образования (e-learning). Появились аналогичные проекты и в русскоязычном Интернете (ProctorEdu.ru, 2015). Но следы научно-методических исследований по этой проблематике найти крайне трудно. Автор статьи не без труда нашел научные публикации по этой тематике лишь на английской языке и… ни одной на русском (см. на русском языке только веблиографию, отражающую прежде всего рекламу коммерческих софтверных решений по проведению прокторинга, отдельные интервью с разработчиками, заказчиками и т.п.). К моменту написания этой статьи термина «прокторинг» еще даже не было в русскоязычной Википедии. А вместе с тем вопрос о точности экспертных оценок с помощью прокторинга стоит как весьма актуальная, практически-значимая проблема.

Для того чтобы быть уверенными в объективности и достоверности экспертных оценок на основе видеонаблюдения, мы должны измерить соответствие экспертных оценок и параллельной объективной информации о том, пользуется или не пользуется экзаменуемый (тестируемый) запрещенными источниками (шпаргалками и подсказками). Аксиомой для научно-обоснованного применения метода

экспертных оценок служит следующий постулат: при определении конкретного состава экспертов, допущенных до ответственной работы, надлежит руководствоваться не только и не столько формальным авторитетом экспертов, сколько реальной валидностью их заключений. Как свидетельствует опыт, нередки случаи, когда авторитет эксперта не дает адекватной эмпирически измеренной валидности (Орлов, 2002). В то же время в русскоязычных публикациях на тему прокторинга мы не нашли ни одного исследования по его реальной валидности, измеренной статистическими методами. А ведь обнаружение экспертов, которые просто не обучены вести адекватное наблюдение, должно приводить к их дисквалификации. В то же время работа по согласованию экспертных оценок, должна приводить к росту числа обученных экспертов и росту общей точности (валидности) усредненных экспертных оценок, если это сочетается с увеличением числа обученных экспертов, оценивавших «спорные случаи», как это происходит при использовании метода «снежного кома» для формирования команды экспертов (Орлов, 2002).

Как известно, для измерения точности экспертных оценок применяется чаще всего процедура анализа их согласованности (конкордации). Согласование мнений экспертов по технике Дельфи предусматривает многократное повторение заочных (дистанционных) опросов независимых экспертов и соответствующее число шкалирования[1] их оценок (Ядов, 2003). В наших собственных прикладных исследованиях, проведенных автором на базе компьютерной системы онлайн-шкалирования, также основной мерой для измерения валидности-надежности экспертных оценок служил расчет их конкордации, в частности, путем подсчета ранговой или точечно-бисериальной корреляции индивидуальных оценок с усредненными оценками, полученными от всех экспертов (Шмелев, 2002, 2006).

В случае прокторинга ситуация выглядит несколько иначе: мошенничество — это не скрытый (латентный) психический процесс, который происходит в глубине мозговой психической активности экзаменуемого, а это определенное объективное поведение, которое может и должно иметь внешние объективные поведенческие про-

[1] Термин «шкалирование» в данной статье употребляется в двух разных значениях в зависимости от смысловых контекстов: 1) шкалирование (scaling) как метод пересчета сырых баллов в стандартизированные, статистически более обоснованные (Ядов, 2003); 2) шкалирование (rating) как оценочная экспертная процедура с использованием шкал, имеющих несколько оценочных градаций (в контексте проведенного нами эксперимента). — Примечание автора.

явления. Эти проявления отражаются не только в появлении высоких результатов (особенно высоких в случае не экзаменов вообще, а именно тестовых экзаменов, где ответы очень легко подделать, если нет надлежащего контроля за достоверностью процесса тестирования). Эти проявления отражаются в объективном наличии у мошенничающих экзаменуемых (читеров) соответствующих файлов с электронными шпаргалками (с помеченными в них правильными, ключевыми ответами), в особенностях временной динамики (задержки или необъяснимо быстрые решения заданий), в их вербальном и невербальном поведении, в частности, в движениях глаз, поворотах головы, в напряжении мимической мускулатуры и т.п.

Работа с одной и той же «шпаргалкой» приводит к тому, что в массиве собранных протоколов (массиве экзаменационных работ) появляются протоколы с очень высоким уровнем сходства не только на множестве правильных ответов, но и на множестве ошибочных ответов. Недаром до появления доступного массового видеонаблюдения методические усилия по обнаружению «читинга» (мошенничества в ходе компьютерного тестирования) предпринимались, в частности, в направлении выявления подмножества экзаменуемых, которые дают сходные паттерны ответов (Macmanus et al., 2005). Другим эмпирико-статистическим направлением в поисках объективных индикаторов читинга следует считать работы Ван дер Линдена (van der Linden, 2006), в которых была обоснована статистически значимая зависимость между читингом и латентным временем решения отдельных тестовых заданий (вопросов). Чаще, чем скоростные ответы, возникают именно задержки, так как читеры в ходе тестового экзамена тратят время на поиск подсказки в обширном перечне заданий, которые надо просмотреть и найти то, которое в данный момент предъявлено на экране компьютера и требует решения. Следует заметить, что аналогичный хронометрический анализ числа отклонений от нормальной длительности в решении тестовых заданий был внедрен автором статьи в рамках проекта «Олимпиада Телетестинг» еще в конце 90-х годов (см. Шмелев, 2013).

Исследованием, наиболее близким к поставленной нами задаче, оказалась диссертационная работа Чиа Ян Чуанга из Аризонского университета (Chia-Yuan Chuang, 2015). В этом эксперименте 42 студента-психолога должны были ответить на 2 теста из 10 тестовых вопросов по теме «Изучение языка программирования Python». Вначале испытуемые выполняли на компьютере первый тест без использования вспомогательного материала («шпаргалки»), а за-

тем — второй тест (параллельную форму) с использованием такого материала, в который они могли «подглядывать». Видеозапись поведения этих 42 испытуемых во время выполнения теста анализировалась сразу с помощью трех разных методов: 1) анализ временных параметров длительности решения каждого задания (time delay), 2) формализованный анализ поворотов головы и фокуса зрительного внимания (VFOA), 3) экспертная оценка мимики, выражающей на лице испытуемых аффективное состояние «замешательства» (confusion). Как видим, в третьем случае фактически применялись элементы прокторинга — видеонаблюдения экспертов за поведением испытуемых, но в данном случае два проктора-наблюдателя решали более узкую задачу: будучи сертифицированными специалистами-психологами по распознаванию эмоциональных состояний по выражению лица (по системе FACS в соответствии с методикой Экмена–Фризена) они должны были фиксировать все появления признаков «замешательства». С помощью коэффициента Kappa (по-русски — каппа) был измерен уровень надежности-согласованности этих экспертов-кодировщиков, который достиг на данном материале требуемого высокого значения 0,9 (о том, что такое коэффициент Kappa см. Cohen, 1960; формулу и правила подсчета Kappa очень быстро можно найти в англоязычной Википедии, в статье Inter-rater reliability, 2020). На основе статистической обработки полученных результатов с использованием кросс-валидизации по методу LOOCV (leave-one-out cross-validation), автор эксперимента пришел к выводу, что все три показателя значимо коррелируют с обнаружением читинга и их системная комбинация дает точность-accuracy в районе 70% (по всем испытуемым в среднем), при этом точность-precision оказалась равной 68%, а доля ошибок типа «ложная тревога» составила 8%[2]. Следует заметить, что с нашей точки зрения этот уровень достигнутой точности следует считать недостаточно высоким.

[2] На четырехклеточной таблице сопряженности (ЧТС), которую в бизнес-литературе чаще называют «матрица ошибок», точность-accuracy рассчитывается как относительное превышение процента совпадений над ожидаемым процентом совпадений, а точность-precision — как доля верных обнаружений «читинга» по отношению ко всем обнаружениям (включающим также «ложные тревоги»). К сожалению, в русском языке эти два разных показателя точности переводятся одним словом «точность», что создает путаницу особенно в умах у гуманитарных специалистов. В данной статье автор предлагает переводить термин precision как «избирательная точность», или «избирательность». Это подобно тому, как в медицинской диагностике уже давно прижились разные термина для обозначения «точности-чувствительности» и «точности-специфичности». — Примечание автора.

Цель эмпирического исследования

В нашем эксперименте мы поставили прежде всего задачу выяснить, как растет уровень точности, если к видеонаблюдению (прокторингу) привлечь не двух, а более независимых экспертов, которые будут использовать сразу несколько критериев для оценки поведения (многокритериальное оценивание), и если при этом произвести отсев экспертов, которые плохо справляются с задачей, и оставить для расчета точности только оценки реально успешных (обученных) экспертов.

Методика эмпирического исследования

В мае-июне 2020 года автором данной статьи был спланирован и проведен компьютеризированный эксперимент по проверке экспертной точности прокторинга. В ходе этого эксперимента испытуемые выполняли в режиме онлайн тест на эрудицию из 30 заданий (по 10 заданий на знание литературы, истории и географии) и сопровождали свои действия максимально полными устными комментариями: формулировали устно условие задания (вопрос), варианты ответа, которые можно было рассматривать как правильные, и выбранный ответ. Все задания содержали только один правильный ответ из четырех вариантов. При этом испытуемые включали в фоновом режиме программу типа Screen Recorder, которая осуществляла запись рабочего экрана (и всех перемещений курсора на нем), речи испытуемого и его лица с фронтальном камеры компьютера (или смартфона).

Общее число испытуемых, которые полностью справились с программой эксперимента (некоторые столкнулись с техническими трудностями в ходе пробной записи демоверсии теста), составило 35 человек. Все они — зарегистрированные участники онлайн-проекта «Клуб Испытателей Тестовых Технологий» (сайт в Интернете — www.testbytest.ru). 15 участников были мужского пола, 20 — женского. Возраст участников — от 17 до 57 лет. Средний возраст — 32 года.

Из числа испытуемых 18 человек вошли в группу условных «читеров» («мошенников»): они получили от организатора эксперимента за день до выполнения основного теста файл-шпаргалку, в котором правильный ответ для каждого из 30 заданий был помечен символом «!!!» (три восклицательных знака). Задача этой группы состояла в том, чтобы максимально незаметно воспользоваться шпаргалкой и получить как можно больше очков — дать как можно больше пра-

вильных ответов, так что они имели внешнюю денежную мотивацию для качественного выполнения теста. Участники из подгруппы «честных» испытуемых вообще не знали о существовании группы «читеров». Обыкновенные испытуемые («честные») получили вместе с приглашением на эксперимент информацию о том, что за факт участия (предоставление качественной видеозаписи выполнения теста) получат 500 рублей, а также получат по 100 рублей премии за каждые дополнительные 10 процентов правильных ответов (10 пунктов на шкале ППО), начиная с 50 процентов. Участники из подгруппы «читеров» вместе с приглашением получили информацию о том, что за факт участия они получат 1000 рублей, а также по 100 премиальных рублей за каждые дополнительные 10 процентов правильных ответов. Насколько эта мотивация участников оказалась эффективной, показали результаты эксперимента.

Своеобразной второй группой обследованных в данном случае явились те, кто выполнял задачу прокторинга — оценивал достоверность поведения испытуемых по видеозаписи устного решения тестовых заданий. Этот второй контингент участников эксперимента составили 14 штатных сотрудников Центра «Гуманитарные технологии» (включая автора статьи), среди которых преобладали методисты — разработчики тестовых методик и психологи-консультанты по профессиональной ориентации[3]. Из 14 прокторов 8 человек — мужчины, 6 — женщины. Возраст — от 22 до 65 лет, средний — 35 лет. Задача эксперта заключалась в том, чтобы сразу после просмотра видеозаписи (длительность каждой — от 10 до 15 минут), оценить поведение испытуемого по 5 пятибалльным оценочным шкалам. Ниже названия шкал и пояснения к ним даются в точном текстуальном соотвествии с тем, как это было сформулировано в инструкции эксперту на его личной веб-странице:

1. РЕЧЬ — насколько естественные речевые интонации и плавный темп речи (не вызывающие подозрений). Слишком быстрый темп или большие задержки при переходе к решению (в отличие от обычного для заданий разной длины) должны приводить к сни-

[3] Автор статьи выносит персональную благодарность за участие в эксперименте в роли экспертов следующим коллегам (по алфавиту в именительном падеже): Балаян Степан, Белорусец Арсений, Васильцов Михаил, Дерябо Ольга, Дулькина Анастасия, Завоеванная Наталья, Кузнецов Кирилл, Лисица Ирина, Лобов Вадим, Окатова Мария, Серебряков Алексей, Тимофеев Антон, Яшина Светлана. В табл. 2 с результатами эксперимента инициалы экспертов-прокторов изменены с учетом использованных экспертами псевдонимов. — Примечание автора.

жению оценок — до 1 или 2. Высший балл 5 следует ставить, когда речь не вызывает никаких подозрений.

2. МИМИКА и ГЛАЗА — насколько естественная мимика (без напряжения особого напряжения) и движения глаз (сфокусированность взгляда именно на экране с заданием).

3. МОТОРИКА — насколько естественный темп работы с экранным интерфейсом (управление курсором, выбор ответа и т.п.)

4. СОПРЯЖЕННОСТЬ[4] — отдельная общая оценка согласованности (скоординированности) всех трех компонентов поведения — речи, мимики и моторики. Обратите внимание на следующее: если испытуемый знаком с заданием (читает его не первый раз и уже знает ответ), то он, как правило, несколько поспешно указывает на ответ вначале РУКАМИ (курсором), а речь как бы запаздывает — комментирует уже произведенное действия. В то же время, когда честный испытуемый принимает решение ПЕРВЫЙ РАЗ, то он, как правило, его вначале проговаривает вслух, а лишь позднее (либо в самом конце зачитывания ответа) производит моторное действие.

5. ОБЩАЯ ОЦЕНКА — оценка достоверности поведения испытуемого. Обратите внимание, что эта оценка НЕ ДОЛЖНА быть получена формальным путем, то есть НЕ должна быть арифметическим средним предыдущих оценок. Например, если лишь по одному признаку (движения глаз) Вы заподозрили неладное и поставили низкий балл 1 или 2, то общая оценка может быть РАВНА именно этому низкому баллу. Это действие по принципу «закона конвоя»: скорость движения конвоя равна скорости самого медленного судна.

В качестве софтверной платформы для выполнения онлайн-тестирования и сбора экспертных оценок использовалась система HT-LINE, разработанная в Центре «Гуманитарные технологии» (см. сайт в Интернете www.ht-line.ru). Каждый участник эксперимента, включая и испытуемых, и экспертов, получал в этой системе свою

[4] Понятие «сопряженность» в данном случае навеяно опытом создания самых первых «лай-детекторов» в истории экспериментальной психологии — методиками, в которых эмоциональная напряженность испытуемого пытались регистрировать через рассогласование речевой и моторной реакции в ассоциативном эксперименте: в случае аффективно значимого слова-стимула моторная реакция опережала речевую или, напротив, сильно отставала, что происходило реже (Лурия, 1928). — Примечание автора.

«личную страницу» (или «личный кабинет»), из которого мог вызывать разные тесты, оценочные процедуры и прикреплять ссылки на файл с видеозаписью. На рис. 1 и рис. 2 в приложении показаны снимки экрана (скриншоты), иллюстрирующие, как выглядела видеозапись, которую оценивал проктор в ходе экспертного оценивания, а также страница с пятью оценочными шкалами.

Результаты и их обсуждение

На самом первом шаге в ходе обработки полученных результатов мы попытались проверить, насколько высокой оказалась мотивация испытуемых из группы «читеров». Оказалось, что практически все «читеры» показали по основному тесту более высокие результаты, чем испытуемые из группы «честных». Все они показали от 70 до 100 баллов на шкале ППО — процентов правильных ответов на вопросы теста (21 правильный ответ из 30 и выше). А в группе «честных» испытуемых только один показал балл ППО = 96% и вклинился по этому показателю в группу в высокую группу, а остальные — ниже 70% и большинство даже меньше половины — ниже 50% на шкале ППО. Таким образом, данный тест для большинства испытуемых оказался сложным и читерам, очевидно, помогло наличие «шпаргалки с правильными ответами», то есть они были серьезно замотивированы на то, чтобы показать высокий результат, подглядывая в шпаргалку, и при этом не быть раскрытыми в ходе данной экспериментальной «деловой игры».

На втором шаге мы проделали объективный хронометрический анализ ответов на каждое задание в каждом из 35 протоколов тестирования, полученных от участников-испытуемых. При этом для каждого из 30 тестовых заданий (items) рассчитывались показатели среднего времени решения (в секундах), а также интервалы «плюс-минус сигма», то есть границы нормальной продолжительности решения ниже и выше среднего на одно стандартное отклонение. Далее для каждого из 35 испытуемых подсчитывался сырой балл «хронометрической атипичности протокола», равный числу выходов за пределы нормального интервала продолжительности (как в сторону задержек, так и в сторону слишком быстрых ответов). На рис. 3 показана гистограмма частотного распределения испытуемых по количеству допущенных ими «выходов» за пределы нормальных хронометрических интервалов. Как видим, большинство испытуемых допускали не более 10 таких выходов из 30 возможных (по числу тестовых заданий в тесте). Это две трети выборки (24 из 35). Но среди

этих «нормальных испытуемых», работавших в нормальном темпе оказалось, увы, немало «читеров». «Увы» — это потому, что хотелось бы увидеть, что протоколы всех читеров являются атипичными. То есть, на нижнем полюсе построенной шкалы «хронометрической атипичности» наблюдается низкая дифференцирующая способность этой шкалы в отношении различения целевой (критериальной) бинарной переменной «читер — честный». Но… зато на верхнем полюсе наблюдается другая картина: в группе «читеров» оказались все те 7 человек, которые показали высокий балл «хронометрической атипичности» (более 15 выбросов из 30 возможных — в более чем половине тестовых заданий).

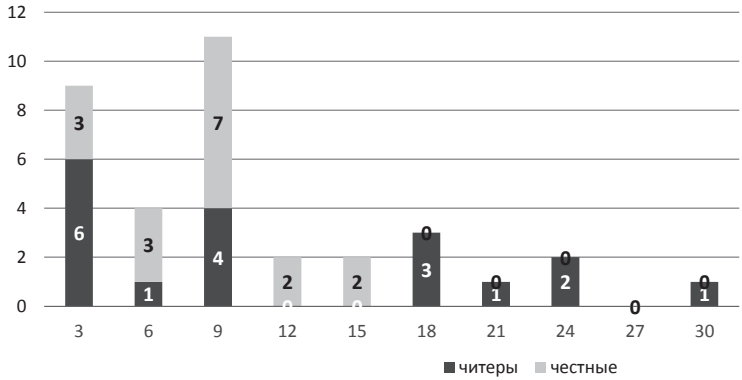

Рис. 3. Гистограмма частотного распределения испытуемых на шкале «число атипичных длительностей решений» (с точностью до интервала в 3 атипичных решения из 30 возможных)

Таким образом, четырехклеточная таблица сопряженности (она же — матрица ошибок, или confusion matrix) для данного показателя оказалось асимметричной (табл. 1).

Таблица 1

Четырехклеточная таблица сопряженности для предиктора «число хронометрически атипичных решений»

	Реальные читеры	Реальные честные
Подозреваемые испытуемые (более 15 атипичных решений)	A = 7	B = 0
Испытуемые вне подозрений (менее 16 атипичных решений)	C = 11	B = 17

Общая точность-accuracy, подсчитанная по табл. 1 в виде коэффициента Каппа, оказывается совсем невысокой, только 0,38. Из-за того, что столбцы таблицы оказываются симметричными по числу случаев (сумме частот по столбцам), то известный коэффициент дискриминативности (КД1) равен для этой таблички тоже 0,38, как и фи-коэффициент (Аптон, 1972). Но зато один из значимых показателей, а именно Precision (прецизионная точность, или, как мы предлагаем его называть, «избирательная точность») оказывается в данном случае равным 1.0, хотя выявлен он и на совсем небольшой выборке в 7 человек (сумма частот по первой строке). Иными словами, мы выявили в нашем эксперименте такую точку отсечения на шкале «хронометрически атипичных решений», которая дает нам нулевое число «ложных тревог»: все испытуемые, которые оказываются выше этой точки отсечения, оказались в группе «читеров»[5]. В целом этот уровень точности для хронометрических данных соответствует известным нам, процитированным в этой статье выше зарубежным исследованиям Ван дер Линдена и Чиа Ян Чуанга.

На третьем шаге мы построили таблицы сопряженности, аналогичные той, которая изображена в табл. 1, для каждого отдельного эксперта и посчитали показатели общей точности с помощью коэффициентов Карра и традиционного коэффициента линейной корреляции Пирсона, который легко каждый пользователь может найти среди стандартных статистических функций в популярной программе MS Excel. В табл. 2 отражены полученные значения для 13 экспертов (один эксперт не выполнил норматив по числу видеозаписей — не менее 30 процентов из всего массива — и был исключен из дальнейшего анализа)[6]. Следует специально подчеркнуть, что

[5] За пределами эксперимента, явившегося предметом данной статьи, остаются данные автора, полученные им на первом массовом конкурсном проекте «Лидеры России» в 2017 году: на выборке в 84 тысячи участников конкурса, протестированных по тесту общих знаний, тогда тоже была обоснована точка отсечения «атипичных протоколов», равная значению ППО = 50% (половина правильных ответов на задания теста). — Примечание автора.

[6] Коэффициент корреляции Пирсона мы привели для того, чтобы читателю легче было воспринимать материал этой статьи, а для этого опереться на нечто знакомое и привычное. Для весьма симметричных столбцов, какими оказались ЧТС в данном эксперименте, корреляция Пирсона очень неплохо отражает общую точность каждого эксперта. Линейная корреляция значений коэффициентов Линкор и Каппа в данном случае оказалась равной 0,92, что говорит о практической взаимозаменимости этих коэффициентов в нашем случае. Но… в практической работе, когда читеров значительно меньше в выборке, чем «честных испытуемых», надо обязательно считать коэффициент Каппа. — Примечание автора.

данные в табл. 2 отражают 5-ю шкалу экспертной оценки — субъективную «общую оценку достоверности». Дихотомическое огрубления 5-й шкалы (перевод в бинарную шкалу) достигалось таким образом: балл 1 в пользу достоверности присваивался шкальным субъективным оценкам 4–5, а балл 0 — в пользу «мошенничества» субъективным оценкам 1–3 (как всегда, наши эксперты при использовании пятибалльной шкалы допускали асимметрию в пользу более высоких оценок, так что балл 3 фактически получил смысл «ниже среднего»).

Таблица 2

Индивидуальные показатели общей точности (accuracy) для 13 экспертов-прокторов (Линкор и Каппа), а также согласованности со «средним экспертом» (Косинус)

Эксперт	А.С.	М.В.	И.В.	О.Д.	А.Д.	Н.С.	М.О.	А.А.	А.Г.	С.Б.	К.К.	С.Я.	В.М.
Линкор	0,82	0,61	0,42	0,84	0,61	0,54	0,69	0,52	0,81	0,48	0,69	0,81	0,18
Каппа	0,7	0,54	0,25	0,88	0,65	0,42	0,77	0,42	0,76	0,37	0,69	0,8	0,24
Косинус	0,88	0,86	0,74	0,82	0,76	0,8	0,86	0,82	0,86	0,82	0,4	0,54	0,36

Средний показатель точности экспертов Каппа равен 0,58 и значительно превосходит точность объективной шкалы «хронометрической атипичности». Только у трех экспертов из 13 (под инициалами И.В., С.Б. и В.М.) обнаружены низкие показатели общей точности — ниже 0,38, что уступает хронометрической шкале. Если же отобрать для «практической работы» только тех 8 экспертов, у которых Каппа выше 0,5, то для этих более подготовленных экспертов среднее значение Каппа будет равно уже 0,72, что оказывается на удивление близким к тому значению, которое получено в процитированном выше эксперименте Чиа Ян Чуанга. При этом сводная ЧТС для этих восьми лучших экспертов принимает следующий вид, как это показано в табл. 3.

Что еще примечательного мы видим в табл. 3? — Опять, как и в случае с хронометрической шкалой, мы видим, что в таблице значительно меньше ошибок типа «ложная тревога» (правая верхняя клетка), чем ошибок типа «пропуск» (6 в клеточке В против 24 в клеточке С). При этом специфичность (контраст частот во втором столбце) выше, чем чувствительность (контраст частот в первом столбце). Показатель «избирательной точности» (presicion) равен 82 / 82 + 6) = 0,93, что приближает каждого из подготовленных экс-

Таблица 3

Суммарная четырехклеточная таблица сопряженности для 8 лучших экспертов-прокторов

	Реальные читеры	Реальные честные
Подозреваемые испытуемые (более 15 атипичных решений)	A = 82	B = 6
Испытуемые вне подозрений (менее 16 атипичных решений)	C = 24	D = 106

Пояснение: сумма частот в таблице больше 35, так как эта частотная таблица явилась результатом суперпозиции — поклеточного суммирования — восьми индивидуальных частотны таблиц).

пертов к высокоэффективной индивидуальной работе. Ведь дело в том, что в реальных условиях (отличных от искусственных условий нашего эксперимента, где группа читеров была искусственно составлена из «подставных добровольцев»), частота ошибок типа «пропуск» должна упасть естественным образом, и главное требование — не допускать ошибок типа «ложная тревога».

На четвертом шаге мы произвели особый анализ — такой, который не мог себе позволить наш зарубежный коллега Чиа Ян Чуанг. Дело в том, что в его эксперименте участвовали лишь 2 эксперта, а в нашем 14 экспертов, из которых мы смогли выявить 8 подготовленных прокторов, способных выявлять читеров. Кроме того, в нашем эксперименте практически все эти 8 прокторов оценивали всех испытуемых — все 35 видеозаписей, а не разделялись парами по разным видеозаписям. Это позволило нам суммировать данные от 8 экспертов и построить особую шкалу «суммарной экспертной оценки» (по всем 5 шкалам от всех 8 экспертов по всем 35 испытуемым). Все 35 испытуемых были отранжированы по убыванию балла «суммарной экспертной достоверности», так что возникли только 2 ошибки в этом отранжированном списке: только 1 «честный» испытуемый оказался ниже некоторых «читеров» и только 1 «читер» оказался выше некоторых «честных» испытуемых. ЧТС для этой «суммарной экспертной достоверности» приобрела следующий вид (см. табл. 4).

Коэффициент общей точности Каппа для этой шкалы приблизился уже к 0,9 (точное значение 0,89), что позволяет смело утверждать, что назначение группы квалифицированных экспертов для работы по обнаружению читеров даст нам необходимый уровень точности в их работе.

Таблица 4

Четырехклеточная таблица сопряженности для шкалы «суммарной экспертной достоверности» (по 8 лучших экспертам)

	Реальные читеры	Реальные честные
Подозреваемые испытуемые (более 15 атипичных решений)	17	1
Испытуемые вне подозрений (менее 16 атипичных решений)	1	16

На пятом шаге обработки мы поставили своей целью измерить, как связаны между собой внутренняя согласованность экспертных оценок (конкордация) и объективная точность различения группы «читеров» и группы «честных». В табл. 2 в третьей строке для 13 экспертов приводятся значения коэффициента «косинус», который измеряет степень согласованности оценок данного эксперта с оценками остальных экспертов (с профилем усредненных экспертных оценок). Ранговая корреляция между показателями Каппа и Косинус оказалась невысокой (0,38), что говорит о том, что в данном случае такая шкала как «согласованность с остальными экспертами» в незначительной степени отражает квалификацию экспертов по обнаружению «реальных читеров». Требуется ввести в систему обучения и измерения квалификации экспертов именно реальных «читеров», а привычного расчета внутренней согласованности между экспертами в данном случае недостаточно.

На шестом шаге в ходе обработки результатов мы проанализировали, насколько удачно работали эксперты с различными 5-ю оценочными шкалами критериями. На рис. 4 отражается сравнительная эффективность всех пяти шкал, которую отражает в данном случае линейная корреляция (строго говоря, точечно-бисериальная корреляция, так как одна из двух переменных, а именно целевой критерий, является в данном случае бинарной переменной «1 или 0»).

Как видим на рис. 4, мы получили несколько неожиданный результат: суммарный показатель по всем пяти шкалам оказался даже менее прогностичен, чем оценки по отдельной 5-й шкале, которую можно считать «субъективной интегральной оценкой». Это можно, очевидно, объяснить тем, что среди отдельных шкал не было ни од-

Рис. 4. Сравнительная диаграмма, показывающая различия в корреляции отдельных экспертных шкал с целевым показателем «наличие — отсутствие читинга» (*число пар наблюдений в данном случае равно 35 с использованием данных от 8 лучших экспертов)*

ной, которая бы отражала важный фактор, различающий «читеров» и «честных» в нашем эксперименте: эксперты легко могли видеть, что читеров отличает более высокая «эрудиция», ведь, вспомним, что все они (практически все) попали в группу «высокоэрудированных». Так что если в ходе видеозаписи устного решения проктор видит, что испытуемый раз за разом дает правильные ответы почти на все задания, некоторые из которых весьма трудные (практически 100 процентов правильных ответов), то в своей итоговой оценке по 5-й шкале проктор, конечно, учитывает этот факт. В целом мы видим, что работали почти все шкалы, но относительно более слабый результат получен по шкале номер 2 «глаза и мимика». Этот результат явился следствием технических недостатков методики, которые вскрыл наш эксперимент уже после его проведения: окно, в котором эксперты-прокторы могли наблюдать за движениями глаз и мимикой, оказалось слишком маленьким и это создавало чисто перцептивные трудности для экспертов.

На последнем, седьмом шаге выполненной нами обработки результатов мы решили проверить с помощью кросс-валидизации

устойчивость экспертных оценок трех лучших экспертов к расщеплению выборки оцениваемых испытуемых пополам[7]. К сожалению, после расщепления пополам объем подвыборки оцениваемых испытуемых для эксперта С.Я. оказался недостаточным. Таким образом, мы публикуем коэффициенты Каппа для случайной половины выборки по трем следующим лучшим экспертам: А.С. (0,62), О.Д. (0,78), А.Г. (0,65). Таким образом, по половине выборки тройка экспертов сохранила свои лидирующие позиции среди группы экспертов и это хороший результат. Среднее значение оказалось равным Каппа = 0,68, что неплохо, так как близко к зарубежным показателям точности в районе 0,7. Но это говорит одновременно о следующем: для достижения требуемых высоких показателей точности-валидности выше 0,9 необходимо организовать обучение экспертов на выборках объемом примерно в 50 человек. Тем самым нашей выборки объемом в 35 тестируемых субъектов (и тем более ее половины объемом в 18 человек) оказывается, строго говоря, недостаточно, так что требуется выборка примерно в 1,5 больше по объему (50 против 35). Хотя при этом опять-таки для трех лучших экспертов вполне удовлетворительным оказывается при расщеплении выборки показатель «точности-избирательности» (precision). Желающие легко могут самостоятельно подсчитать этот показатель для следующей четырехклеточной таблицы сопряженности (табл. 5).

Таблица 5

Четырехклеточная таблица сопряженности на половине от выборки испытуемых (по 3 лучшим экспертам)

	Реальные читеры	Реальные честные
Подозреваемые испытуемые (более 15 атипичных решений)	17	1
Испытуемые вне подозрений (менее 16 атипичных решений)	7	25

[7] В данном случае нами был применен самый простой и понятный для читателей, не искушенных в анализе данных (data analysis), метод кросс-валидизации — более простой по сравнению с более изощренными современными компьютерными методами K-fold cross validation, которые требуют просчета устойчивости выработанного правила на K равных по объему случайных подвыборках; а в данном случае K = 2 (см. статью в англоязычной Википедии — cross-validation). — Примечание автора.

Легко видеть, что доля «реальных читеров» среди всех подозреваемых в табл. 5 составляет 17 из 18, то есть 94%. В тоже время доля «отрицательного правильного прогноза» опять-таки невелика (25 из 32, то есть 0,78), что говорит о значительном количестве «пропусков». Вывод в этом случае напрашивается такой: даже самые успешные эксперты, выявленные в нашем исследовании, нуждаются в дополнительном анализе происхождения ошибок типа «пропуск» и в тренинге на более обширном материале, чтобы сократить эти ошибки. Тем не менее уже в настоящем виде методика экспертной оценки может быть использована для выявления читеров с пренебрежимой долей ошибок типа «ложная тревога». Таким образом, если мы направляем таких подозреваемых на контрольное (повторное тестирование в очном режиме), то в очень малой степени сомневаемся в том, что все эти испытуемые такого контрольного тестирования заслуживают. Сама же перспектива контрольного очного перетестирования является, как известно, серьезным сдерживающим фактором, резко сокращающим число участников тестирования, готовых рисковать своей репутацией и прибегнуть к фальсификациям — к различным стратегиям «читинга» (мошенничества).

Обсуждая результаты проведенного эксперимента, мы не можем не отметить ряд недостатков и условностей самой методики и организационной схемы:

1. Выборка размером в 35 испытуемых является недостаточно репрезентативной для категоричных выводов. В дальнейшем планируется в ходе практико-ориентированных массовых прикладных работ с использованием данной методики значительно расширить выборку. Но сам факт фактического совпадения показателей точности с данными зарубежных исследований позволяет с определенным доверием относиться и к другим результатам проведенного исследования, несмотря на ограниченный объем выборки.

2. Сама методика формирования группы «подставных испытуемых» нуждается в совершенствовании. В данном случае эта организационно-методическая схема во многом базировалась на доверии к испытуемым. Ведь каждый из них, получив доступ к ключевым ответам на задания задолго до выполнения теста, мог просто выучить эти ключевые ответы вместе со всеми 30 заданиями. Но… тот факт, что результаты испытуемых из группы «подставных» не были 100-процентными, говорит о том, что эти испытуемые проявили добросовестность (в каких-то случаях подкрепленную занятостью

или ленью) и этого не сделали — не выучивали задания[8]. Тем самым полученные нами результаты позволяют считать этот недостаток в целом незначительным.

Выводы и перспективы исследования

В результате проведенного нами эмпирического исследования можно сделать следующие выводы:

1. Нами получены результаты, которые по уровню точности обнаружения «читинга» в целом соответствуют аналогичным зарубежным исследованиям, что повышает уровень доверия к этим результатам.
2. Подготовленные эксперты могут эффективно обнаруживать «читинг» с помощью интегральной субъективной оценки достоверности поведения испытуемых на основе видеозаписи устного решения тестовых заданий в режиме онлайн (при выполнении удаленного тестового экзамена).
3. Разработана валидная методика измерения индивидуальной эффективности отдельных экспертов, которая позволяет отсеивать неэффективных экспертов и не допускать их к работе.
4. Объективная шкала «хронометрическая атипичность протоколов» уступает по точности любому коллективу из двух-трех подготовленных экспертов и тем более многочисленному коллективу экспертов. Но, будучи гораздо более дешевой по затратам, эта шкала может быть эффективно использована на ранних массовых этапах тестирования с целью эффективного выявления «читеров», так как обладает высокой «избирательной точностью» (presicion, или точность положительного прогноза). В то же время квалифицированных экспертов следует привлекать для более скрупулезного анализа спорных случаев или видеозаписей «контрольных сеансов дистанционного тестирования».
5. На данном материале обнаружена слабая скоррелированность объективной меры точности экспертных оценок и тра-

[8] Фактически выучил задания только один испытуемый из группы «подставных», который откровенно признался в этом в ходе телефонного опроса, проведенного уже по итогам эксперимента с испытуемыми, показавшими «подозрительно высокие результаты». — Примечание автора.

дикционного показателя согласованности экспертных оценок, что отменяет использование показателя согласованности для обучения и отбора экспертов в данном случае.

6. Созданная в ходе исследования эталонная выборка объемом в 35 видеозаписей, из которых примерно половина продемонстрировала «реальный читинг», может служить эффективным материалом для обучения экспертов, по крайней мере для надежного тренинга экспертов до такого уровня «избирательной точности», который позволяет избегать ошибок типа «ложная тревога» Но эта же выборка пока является недостаточной для снижения доли ошибок типа «пропуск». То есть, пока тренировочный материал позволяет формировать коллектив экспертов, работающих с асимметричной точностью — в пользу испытуемых (спорные случаи кодируются как «честные»).
7. Разработанную методику в силу асимметричной прогностичности следует применять на практике не изолированно, но в системном сочетании с такой организационной методической схемой тестовых экзаменов, которая предполагает проведение вслед за дистанционным онлайн-экзаменом очного контрольного выборочного перетестирования — на выборке тех экзаменуемых, которые заподозрены коллективом обученных экспертов в читинге.
8. Перспективы дальнейшего исследования таковы. Надо практически измерить, насколько справедлива следующая правдоподобная, но, строго говоря, еще не доказанная гипотеза: повторные обучающие серии экспериментального обучения экспертов позволяют повысить точность работы экспертов.

СПИСОК ЛИТЕРАТУРЫ

Аптон Г. Анализ таблиц сопряженности. М: Финансы и статистика, 1972. 143 с.

Лурия А.Р. Диагностика следов аффекта. Сопряженная моторная методика и её применение в исследовании аффективных реакций // Проблемы современной психологии. М., 1928. Т. 3. С. 46.

Орлов А.И. Экспертные оценки. Учебное пособие. М.: МЗ-Пресс, 2002. 31 с.

Шмелев А.Г. Многокритериальная оценка пользовательских интерфейсов портальных проектов // Интернет-порталы: содержание и технологии». М.: Просвещение, 2002. Т. 2. С. 346–361.

Шмелев А.Г. Согласование экспертных оценок с помощью Интернет-технологии шкалирования // Экспертиза в современном мире: от знания к деятельности. М.: Смысл, 2006. С. 175–185.

Шмелев А.Г. Практическая тестология. Тестирование в образовании, прикладной психологии и управлении персоналом. М.: Маска, 2013. 688 с.

Ядов В.А. Стратегия социологического исследования. М.: Академкнига, Добросвет, 2003. 596 с.

Chia-Yuan Chuang. Improving Proctoring by Using Non-Verbal Cues During Remotely Administrated Exams. A Dissertation Presented for the Degree Doctor of Philosophy. Arizona state university, 2015. 79 p.

Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 1960, 20* (1), 37–46.

Macmanus I.C., Lissauer T.J., Williams S. Detecting cheating in written medical examinations by statistical analysis of similarity of answers: Pilot study. British Medical Journal (online), 2005, 1064-6.

van der Linden W.J. A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 2006, *3 1* (2), 181–204.

Cross-validaion (2020). Ссылка: https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation

Inter-rater reliability (2020). Ссылка: https://en.wikipedia.org/wiki/Inter-rater_reliability

ProctorU. (2008). Pupilcity ProctorU. Ссылка: http://www.proctoru.com/

ProctorEDu (2015). Тесты в бизнесе и прокторинг. Ссылка: https://proctoredu.ru/interview

VProctor. (2014). VProctor. Ссылка: http://vproctor.com/

Что такое прокторинг и чем он полезен на экзаменах, тестировании и в обучении (2020). Ссылка: https://finacademy.net/materials/article/proktoring

REFERENCES

Apton G. Analiz tablic soprjazhennosti. Moscow: Finansy i statistika, 1972. 143 s.

Chia-Yuan Chuang. Improving Proctoring by Using Non-Verbal Cues During Remotely Administrated Exams. A Dissertation Presented for the Degree Doctor of Philosophy. Arizona state university, 2015. 79 p.

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20 (1), 37–46.

Jadov V.A. Strategija sociologicheskogo issledovanija. Moscow: Akademkniga, Dobrosvet, 2003. 596 s.

Lurija A. R. Diagnostika sledov affekta. Soprjazhennaja motornaja metodika i ejo primenenie v issledovanii affektivnyh reakcij. — V kn: Problemy sovremennoj psihologii, t. 3. Moscow., 1928, s. 46.

Macmanus I.C., Lissauer T.J., Williams S. Detecting cheating in written medical examinations by statistical analysis of similarity of answers: Pilot study. British Medical Journal (online), 2005, 1064-6.

Orlov A.I. Jekspertnye ocenki. Uchebnoe posobie. Moscow: MZ-Press, 2002. 31 s.

Shmelev A.G. Mnogokriterial'naja ocenka pol'zovatel'skih interfejsov portal'nyh proektov. — V sbornike Internet-portaly: soderzhanie i tehnologii, Moscow: Prosveshhenie, 2002, tom 2, s. 346–361.

Shmelev A.G. Prakticheskaja testologija. Testirovanie v obrazovanii, prikladnoj psihologii i upravlenii personalo. Moscow: Maska, 2013. 688 s.

Shmelev A.G. Soglasovanie jekspertnyh ocenok s pomoshh'ju Internet-tehnologii shkalirovanija. — V sbornike "Jekspertiza v sovremennom mire: ot znanija k dejatel'nosti". Moscow: Smysl, 2006, s. 175–185.

van der Linden W.J. A lognormal model for response times on test items. Journal of Educational and Behavioral Statistics, 2006, 31 (2), 181–204.

Cross-validaion (2020). https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation

Inter-rater reliability (2020). https://en.wikipedia.org/wiki/Inter-rater_reliability

ProctorU. (2008). Pupilcity ProctorU. http://www.proctoru.com/

ProctorEDu (2015). Тесты в бизнесе и прокторинг. https://proctoredu.ru/interview

VProctor. (2014). VProctor. http://vproctor.com/

Chto takoe proktoring i chem on polezen na jekzamenah, testirovanii i v obuchenii (2020). https://finacademy.net/materials/article/proktoring

ИНФОРМАЦИЯ ОБ АВТОРЕ

Шмелев Александр Георгиевич — доктор психологических наук, профессор, профессор факультета психологии МГУ имени М.В. Ломоносова, Москва, Россия. E-mail: ags12@ht.ru

ABOUT THE AUTHOR

Alexander G. Shmelev — Doctor of Psychology, Professor, Faculty of Psychology, Lomonosov Moscow State University, Moscow, Russia. E-mail: ags12@ht.ru

Приложение 1.

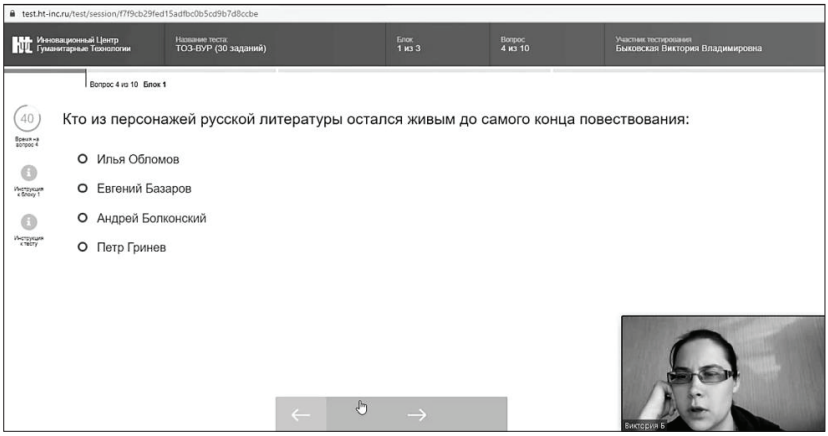

Рис. 1. Экран видеозаписи устного решения в том виде, как он был представлен эксперту-проктору *(в правом нижнем углу по записи воспроизводится мимическая динамика испытуемого, синхронизированная с его действиями по выбору правильного ответа)*

Приложение 2.

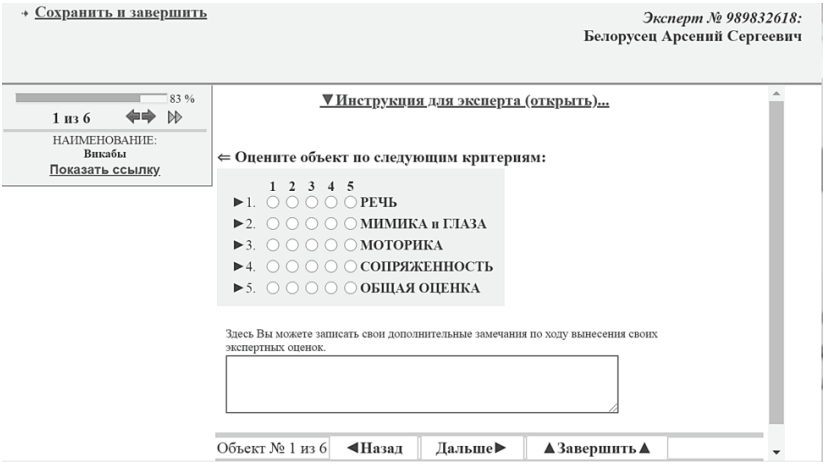

Рис. 2. Веб-интерфейс эксперта-проктора: *пять оценочных шкал для оценки поведения испытуемого после наблюдения за видеозаписью устного решения*