Парадокс децентрализации в образовательном оценивании: валидность Всероссийской олимпиады школьников

Гулов Артем Петрович – д-р пед. наук, доцент кафедры английского языка, МГИМО МИД РФ (Москва, Россия); gulov@tea4er.org

Аннотация. *Всероссийская олимпиада школьников (ВсОШ) охватывает более 6 млн. участников ежегодно, определяя их образовательные траектории, но функционирует без систематической психометрической валидации. Это противоречит международным стандартам для высокоставочных систем оценивания. Целью исследования была комплексная оценка валидности ВсОШ путем анализа литературы и качественного исследования практик региональных жюри (37 экспертов из 12 регионов). Использованы полуструктурированные интервью с двойным кодированием (к=0,76). Выявлены пять устойчивых проблем: отсутствие пилотирования заданий (89,2% экспертов), непрозрачность критериев (20% регионов), локальные интерпретации стандартов (70,3%), конфликты интересов в жюри (54,1%) и нерегулируемые апелляции (39,2%). ВсОШ не достигает ни одного из шести компонентов единой концепции валидности Мессика. Предложена гибридная модель реформы: 70% заданий разрабатываются федерально, 30% региональных разработок проходят внешний контроль с анализом смещений.*

Ключевые слова. *Валидность тестов, справедливость оценивания, олимпиады, децентрализованные системы оценивания.*

Всероссийская олимпиада школьников (ВсОШ) занимает уникальное место в системе российского образования. Это крупнейшая образовательная конкурсная система в стране, охватывающая ежегодно более 6 млн. школьников по 24 предметам. Значимость этой системы определяется не только ее масштабностью, но и тем, что она напрямую влияет на образовательные траектории ее участников и их поступление в университеты.

Достижение определенного статуса в олимпиаде – от призера регионального этапа до победителя федерального уровня – открывает для школьников реальные льготы при поступлении в высшие учебные заведения. Эти льготы варьируются от снижения баллов единого государственного экзамена (ЕГЭ) до полного зачисления в университет без необходимости проходить вступительные испытания. Таким образом, результаты ВсОШ приобретают судьбоносное значение для миллионов молодых людей, определяя доступ к качественному высшему образованию и, в конечном итоге, к карьерным перспективам. Однако парадокс заключается в том, что эта система с высокими ставками на кону функционирует без систематической психометрической валидации. ВсОШ принципиально отличается от централизованных систем оценивания, таких как ЕГЭ в России, SAT в США или Gaokao в Китае. Последние подвергаются строгой проверке надежности и справедливости на федеральном уровне. ВсОШ же разрабатывается децентрализованно, во всех регионах страны, без обязательного пилотирования заданий, без анализа смещений в измерениях при сравнении разных групп участников (так называемого дифференциального функционирования пункта, в англоязычной литературе – DIF) и без применения других

стандартных процедур проверки качества и справедливости, которые требуют международные стандарты для высокоставочных (high stakes) систем оценивания. Центральная предметно-методическая комиссия как федеральный институт ежегодно публикует методические рекомендации для разработчиков школьного и муниципального этапов, оставляя право создания комплектов заданий региональным органам управления образованием. Единые комплекты по всей стране начинают функционировать только с регионального этапа, по сути 3 уровня отбора.

Это противоречие между высокими ставками, которые определяют жизненные пути десятков миллионов людей, и отсутствием систематической проверки качества и справедливости оценивания составляет центральную проблему настоящего исследования. Такое положение дел вызывает серьезные вопросы о справедливости системы и о том, может ли она в действительности обеспечить честное и объективное оценивание для всех участников, независимо от их региона проживания, социального происхождения или других характеристик.

Вопросы качества, справедливости и надежности оценивания давно находятся в центре внимания мировой педагогической науки. Согласно авторитетным стандартам, разработанным Американской ассоциацией исследователей в образовании, Американской психологической ассоциацией и Национальным советом по образовательным измерениям, любая система оценивания, которая оказывает существенное влияние на жизненные пути участников, должна соответствовать строгим и четко определенным требованиям.

С.Мессик [1] предложил единую концепцию валидности, раскрывающую сложность и многомерность этого понятия. Валидность, согласно этому подходу, включает шесть взаимосвязанных компонентов, каждый из которых критически важен для справедливого оценивания:

• *валидность содержания (content validity)* – задания в достаточной мере и репрезентативно охватывают содержание и дисциплину, по которой проводится оценивание;

• *валидность конструкта (construct validity)* – задания действительно измеряют те компетенции и знания, которые задумано измерять, а не что-либо другое;

• *структурная валидность (structural validity)* – структура полученных результатов соответствует теоретической модели развития компетенции, которая лежит в основе теста;

• *валидность обобщаемости (generalizability validity)* – результаты, полученные в одном контексте, могут быть обоснованно распространены на другие контексты и группы участников;

• *внешняя валидность (external validity)* – результаты тестирования коррелируют и согласуются с внешними критериями, такими как успешность обучения в университете или дальнейшие достижения в выбранной области;

• *валидность последствий (consequential validity)* – социальные и педагогические последствия использования результатов оценивания справедливы и не наносят вреда отдельным группам участников.

Более поздние разработки М.Т.Кейн [2] расширили понимание данного концепта, подчеркнув, что валидность – это не просто набор внутренних технических характеристик теста, а прежде всего вопрос о том, насколько корректно интерпретируются и используются результаты оценивания в конкретном контексте и для конкретных целей. Надежность, т.е. внутренняя согласованность и стабильность оценок, является необходимым условием для любого справедливого оценивания. Т.Бентон [3] в своем исследовании показывает, что для достижения минимально приемлемого уровня надежности

(коэффициент альфа Кронбаха $\alpha \geq 0{,}85$, что рекомендуется для высокоставочных систем) тест должен содержать от 40 до 60 заданий в зависимости от их типа и особенностей. Это означает, что на каждом этапе олимпиады должно быть не менее 40 независимых пунктов оценивания, чтобы можно было достаточно надежно различать уровни компетенции участников. Одна из наиболее актуальных и практически важных проблем в психометрике – это выявление и анализ скрытых смещений в тестовых заданиях. Дифференциальное функционирование пункта (DIF) – это явление, при котором определенное задание систематически выгодно одной группе участников (например, мальчикам) и невыгодно другой (например, девочкам), даже если оба участника обладают одинаковым общим уровнем компетенции. Такие смещения могут быть результатом непреднамеренной культурной предвзятости в формулировке задания, половых стереотипов или других скрытых факторов. Авторы исследования [4] убедительно показывают, что анализ дифференциального функционирования пункта должен быть не факультативным, а обязательной, рутинной частью разработки и проверки качества любого высокоставочного теста. Это означает, что перед тем, как использовать тест с реальными участниками, разработчики должны проверить каждое задание на наличие таких смещений. Другое исследование [5] предоставляет наглядный пример значимости такого анализа. После выявления и удаления пунктов с выраженным дифференциальным функционированием оценка гендерных различий в исследуемом конструкте изменилась на 30–40%. Это не просто статистическое изменение – это кардинальное изменение интерпретации результатов, которое может привести к совершенно разным выводам о том, существуют ли реальные различия между группами или они являются артефактом смещения заданий.

В российском образовательном контексте систематическое изучение валидности высокоставочных систем оценивания только начинает развиваться. В.А.Болотов [6] в своем анализе исторического развития российской системы оценки качества образования отмечает, что ЕГЭ, несмотря на более чем 20-летнюю историю своего существования, только в последние 5–10 лет начал подвергаться более или менее систематической психометрической проверке. Это означает, что даже в централизованной, подконтрольной федеральному уровню системе вопросы валидации долгое время не получали должного внимания. Фундаментальное исследование О.Д.Федорова и К.Д.Веринчук [7], посвященное валидности ЕГЭ по истории, выявило три критических угрозы справедливости оценивания: во-первых, критерии оценивания открытых заданий часто бывают нечеткими и допускают двусмысленность; во-вторых, при применении этих критериев различными экспертами проявляется высокая субъективность; в-третьих, уровень подготовки экспертов по критериям оценивания существенно различается между регионами. Особенно важен вывод авторов: если даже централизованная система оценивания, разрабатываемая и контролируемая на федеральном уровне, обнаруживает такие проблемы, то ВсОШ, являющаяся децентрализованной системой с минимальным федеральным контролем, нуждается в валидации еще больше. С.В.Иванова и А.А.Малыгин [8] поднимают глубокий философский и педагогический вопрос: доверие к системе оценивания – это не просто техническое требование, а фундаментальное условие для справедливого и продуктивного образовательного процесса. Авторы показывают: когда участники системы (школьники, учителя, родители) теряют доверие к механизмам оценивания, это влечет далеко идущие негативные последствия – падение учебной мотивации, де-

градацию этических норм в образовательном сообществе и углубление социального расслоения.

В отличие от большинства международных олимпиад, где главная цель заключается в выявлении и ранжировании узкой группы лучших участников, декларируемая цель ВсОШ состоит в развитии креативности и творческого мышления у широкого круга школьников, а не только у топ 1% сильнейших [9]. Это предполагает, что олимпиадные задания должны не просто эффективно выделять наиболее подготовленных учащихся, но и вовлекать в содержательное интеллектуальное усилие участников со средним уровнем подготовки, а также оценивать не только объем и точность предметных знаний, но и более сложные когнитивные и творческие компетенции. Вместе с тем, операционализация такого ключевого целевого ориентира, как «креативность», в реальных олимпиадных заданиях остается неявной и неформализованной, что создает прямую угрозу конструктной валидности: остается неясным, какие именно аспекты творческого мышления предполагается измерять и с помощью каких типов задач.

Ситуацию усугубляет то, что в международной исследовательской повестке содержательная валидность олимпиад также изучена фрагментарно. Так, исследование [10], проведенное на материале Южноафриканской математической олимпиады (140 заданий за 2006–12 гг.), выявило серьезный дисбаланс в покрытии содержательных областей: доля заданий по числам и операциям составила 36,3% при планируемых 20%, тогда как геометрия была представлена лишь в 7,3% заданий при требуемых 27,5%, а темы графиков функций и десятичных дробей вовсе отсутствовали. Аналогично, в отношении Международной биологической олимпиады исследование Л.Флориан и его коллег [11], охватившее 703 задания из различных лет проведения, выявило, что теоретические задания не проходят формального пилотирования с психометрической проверкой в той же степени, что и практические туры олимпиады, что отражает асимметрию в требованиях к валидации разных компонентов экзамена. Для международных олимпиад по физике и химии (IPhO, IChO) публикации о содержательной валидности существенно ограничены, что не позволяет судить о полноте и сбалансированности охвата предметных областей. В совокупности это дает основание утверждать, что содержательная валидность олимпиадных систем – как международных, так и национальных – остается либо практически не изученной, либо исследованной лишь частично и фрагментарно, что является серьезным пробелом в свете высоких ставок, с которыми связаны результаты таких соревнований.

На основании анализа международных стандартов и российских исследований, а также выявленного дефицита работ о валидации олимпиадных систем сформулирована цель и задачи нашего исследования. Его цель – комплексно оценить психометрическую валидность заданий Всероссийской олимпиады школьников, объединив систематический анализ международной литературы с качественным исследованием практик региональных экспертных комиссий.

Для достижения этой цели решаются следующие задачи:

• выявить ключевые расхождения между требованиями международных стандартов [12] и существующим состоянием валидации в системе ВсОШ;

• провести качественное эмпирическое исследование на основе полуструктурированных интервью с 37 экспертами из 12 регионов России, участвующими в разработке и оценивании олимпиадных заданий;

• предложить реалистичные пути реформирования ВсОШ, позволяющие приблизить ее к международным стандартам ва-

лидности и справедливости оценивания с учетом российских условий.

Необходимо восполнить значительный пробел в отечественной педагогической литературе и дать научно обоснованные рекомендации по совершенствованию одной из ключевых образовательных систем России.

Методология. Нами использован смешанный дизайн, интегрирующий два взаимодополняющих компонента: обзор литературы и качественное исследование в формате полуструктурированных интервью. Первый компонент представляет собой анализ англоязычных и русскоязычных источников, отобранных по критериям релевантности к проблематике психометрической валидности, дифференциального функционирования заданий [13–15], олимпиадного движения и справедливости процедур оценивания. Второй компонент – качественное исследование, основанное на полуструктурированных интервью (N = 37) с экспертами региональных команд жюри из 12 субъектов Российской Федерации. В состав выборки вошли 28 человек (75.7%) – члены региональных жюри, которые проверяют муниципальный этап ВсОШ, 9 человек (24.3%) – методисты предметных комиссий регионального этапа, входящие в организационный комитет, отобранные по критериям: не менее трех лет участия в работе жюри и опыт оценивания не менее 100 работ. Интервью проводились в октябре–ноябре 2025 г. в очном формате или онлайн (Zoom), их продолжительность составляла 45–60 мин. В качестве инструмента использовался полуструктурированный протокол с 10 основными вопросами, направленными на выяснение практик разработки и утверждения заданий, наличия и форм пилотирования, интерпретации и вариативности федеральных критериев оценивания между регионами, механизмов обеспечения объективности, типичных сложностей в оценивании, случаев несправедливого оценивания, процедур работы с апелляциями, возможных направлений улучшения и наличия конфликтов интересов в жюри. Полученные расшифровки интервью были подвергнуты тематическому анализу по 15 кодам; для обеспечения надежности проведено двойное кодирование 40% материала, при этом коэффициент согласия Cohen's Kappa составил 0.76, что свидетельствует о хорошем уровне межкодировочного согласия [14]. Процедуры качества включали получение информированного согласия всех участников, гарантирование конфиденциальности и предварительное одобрение исследовательского протокола этическим комитетом.

Результаты и обсуждение. Качественное исследование позволило выявить пять устойчивых проблем в практике оценивания, которые носят системный характер и проявляются в большинстве опрошенных регионов. Первая проблема связана с практически полным отсутствием пилотирования олимпиадных заданий: 89.2% экспертов сообщили, что задания перед олимпиадой не пилотируются, еще 8.1% описали лишь эпизодическую неформальную «апробацию на коллегах и друзьях», и лишь 2.7% указали на формальное пилотирование с анализом параметров. В качестве причин назывались дефицит времени (78.4%), ресурсов (70.3%), программного обеспечения для анализа (59.5%) и методических знаний о том, как организовать пилотирование (48.6%). В результате организаторы и жюри не обладают информацией о трудности заданий, их дискриминативности и возможных гендерных или иных смещениях, что создает серьезные угрозы психометрической валидности и справедливости оценивания. Вторая проблема – непрозрачность критериев оценивания. По данным интервью, примерно в 20% случаев критерии вообще не раскрываются участникам и их родителям, в большинстве слу-

чаев они предоставляются лишь частично (ориентировочные формулировки, укрупненные уровни оценивания без детальных расшифровок), и лишь немногие эксперты описывают ситуацию, когда критерии можно считать по-настоящему полностью открытыми. При этом, даже когда официальные критерии формально опубликованы, скрытыми остаются дополнительные схемы проверки, которыми жюри фактически руководствуется при выставлении баллов, а также внутренние инструкции и разъяснения, позволяющие понять процедуру оценивания во всей полноте. В совокупности это создает ситуацию, когда участники и их наставники видят лишь «верхний слой» нормативных документов, тогда как реальные практики оценивания опираются на непубличные регламенты и неформализованные договоренности внутри жюри, что усиливает асимметрию информации, снижает доверие к системе и затрудняет внешний контроль за справедливостью процедуры. Обоснования закрытости при этом во многом повторяются и не являются взаимоисключающими: один и тот же эксперт мог называть сразу несколько причин, поэтому приведенные доли не суммируются до 100%. Чаще всего закрытость рассматривается как способ снижения риска «натаскивания» (43.2% респондентов упоминают этот мотив), как устоявшаяся традиция (40.5%), а также как инструмент снижения конфликтности, поскольку открытые критерии, по мнению экспертов, провоцируют споры при апелляциях (37.8%). Такая непрозрачность усиливает асимметрию информации между жюри и участниками, снижает доверие к системе и дополнительно затрудняет контроль за справедливостью оценивания. Третья проблема – локальные интерпретации критериев оценивания, характерные примерно для 70.3% опрошенных. В интервью регулярно описывается ситуация, когда в одном регионе сформировались исторически «жесткие» стандарты по отдельным предметам, тогда как в другом требования существенно мягче. Один из экспертов отмечает, что традиция строгости восходит к авторитету выдающегося ученого советского периода; хотя его уже нет, устойчивая норма оценивания сохраняется, в то время как в соседнем регионе за ту же работу участник получает другие баллы. В результате один и тот же уровень решения задания может приводить к разным оценкам в зависимости от региона, что нарушает принцип эквивалентности условий и подрывает идею общенационального статуса олимпиады. Четвертая проблема связана с конфликтами интересов внутри жюри. Более половины экспертов (54.1%) сообщают, что в их жюри работают учителя из тех же школ, где обучаются участники, 37.8% отмечают случаи, когда члены жюри ранее готовили участников как репетиторы, а 29.7% прямо указывают на известные случаи предвзятости в пользу «своих» учеников. Формально предпринимаются меры смягчения: в 73.0% случаев декларируется запрет на оценивание собственных учеников, а в 89.2% используется шифрование работ. Однако сами эксперты признают, что эти механизмы не всегда работают безупречно: по стилю решения, почерку или содержанию можно догадаться, чью работу они оценивают, и это сохраняет риск систематической предвзятости. Пятая проблема касается нерегулируемых и непоследовательных практик рассмотрения апелляций. По оценкам 39.2% экспертов, им известны случаи нечестного рассмотрения апелляций, когда решения принимались скорее в интересах удобства жюри, чем в интересах справедливости. Около 32.4% респондентов отмечают, что апелляции нередко отклоняются «пакетно» или без детального разбора ради экономии времени, а 27.0% указывают на обратную крайность: жюри иногда «жалеет» участника и необосно-

ванно поднимает баллы, руководствуясь эмоциональными, а не профессиональными соображениями. В совокупности это означает, что апелляционный механизм не выполняет функции институциональной защиты прав участника и сам становится источником дополнительных искажений.

Теоретическая интерпретация этих пяти проблем выводит к парадоксу децентрализации ВсОШ. В децентрализованных системах, где ответственность за разработку и оценивание заданий распределена между множеством региональных команд, по мере роста степени децентрализации (D) требуемый уровень централизованного контроля качества и валидации должен возрастать, тогда как фактический контроль, напротив, имеет тенденцию снижаться. В предельном случае высокой децентрализации, когда регионы разрабатывают около 60% заданий ($D \approx 0.95$), теоретически необходимый уровень контроля становится максимальным, тогда как реальные механизмы контроля остаются минимальными, что создает дефицит примерно на уровне 80% от требуемого. Использование концепции М.Липски [16] о чиновниках, работающих на местах (street-level bureaucrats), позволяет объяснить, почему именно такой набор проблем возникает в децентрализованной олимпиадной системе: региональные эксперты, действующие в условиях хронического дефицита времени, ресурсов, данных и четких психометрических требований сверху, фактически вырабатывают собственные локальные нормы и практики; отсутствие ресурсов для пилотирования делает параметры заданий неизвестными; непрозрачность критериев превращается в способ управления при высокой дискреции; конфликты интересов и нерегулируемые апелляции отражают отсутствие эффективных формальных механизмов внешнего контроля; локальные интерпретации критериев становятся естественным продуктом автономии, не сдерживаемой едиными стандартами валидации. Все это позволяет рассматривать выявленные проблемы как предсказуемое следствие структурного парадокса децентрализации при отсутствии сопоставимого по масштабу контроля качества, а не как набор случайных индивидуальных ошибок.

Таблица 1/ Table 1

Цитаты из интервью / Quotes from the interview

№	Проблема	Цитата
1	Отсутствие пилотирования	«У нас в комиссии 4 основных человека. Мы собираемся один раз за 2 недели до дедлайна, обсуждаем задания, которые придумали дома, голосуем. Если 3 из 4 согласны – включаем в тест. Никакого пилотирования не проводим. Как проверим, что задание хорошее? Ну, по опыту чувствуем»
2	Отсутствие пилотирования	«Мы не знаем, будут ли задания трудными или легкими, пока не проведем олимпиаду. Часто бывает, что задание 5, которое мы считали сложным, решают все, а задание 2, которое казалось простым, не решает никто. Это означает, что мы неправильно оценили трудность»
3	Региональная вариативность	«Я проверяла задания из соседнего региона. У них продуктивное задание оценивается по 5 критериям, у нас – по 4. У них максимум 25 баллов, у нас – 20. Как сравнивать результаты? Это невозможно»

№	Проблема	Цитата
4	Закрытость критериев	«Мы публикуем общие критерии, но детали остаются у жюри. Например, за эссе мы даем 15 баллов, но как они распределяются между аргументацией, стилем и фактами – это наше внутреннее знание. Если участник спросит, мы объясним, но заранее не раскрываем»
5	Локальные интерпретации критериев	«Мы очень строго оцениваем грамматику в письменных работах: одна ошибка – минус 1 балл. В других регионах, я знаю, более либерально: условно минус 0,5 балла или даже игнорируют мелкие ошибки. Это создает неравенство»
6	Локальные интерпретации творчества	«Мы пытаемся оценивать креативность в эссе, но у нас нет четкого определения, что это. Один эксперт считает креативным необычный поворот мысли, другой – использование редких цитат. Это создает разногласия»
7	Конфликты интересов	«В нашей комиссии 15 человек, из них 3 – учителя из лицеев, которые готовят победителей. Они проверяют работы, в том числе своих учеников. Это создает конфликт интересов. Я несколько раз замечал, что работы их учеников получали более высокие баллы, чем аналогичные работы из обычных школ»
8	Конфликты интересов	«Нет процедуры, когда эксперты заполняют декларацию об отсутствии конфликта интересов. Мы просто предполагаем, что все честны. Но это не гарантирует справедливость»
9	Злоупотребления при апелляции	«У нас было несколько случаев, когда участник пришел на апелляцию с просьбой повысить баллы, а жюри нашло дополнительные ошибки и снизило баллы. Теперь родители и учителя боятся подавать апелляции. Они говорят: “Лучше не рисковать, вдруг снизят еще больше”»
10	Злоупотребления при апелляции	«Был случай, когда участник пришtл на апелляцию, а жюри сказало: “Если вы будете настаивать на повышении баллов, мы пересмотрим всю вашу работу и можем снизить баллы за другие задания”. Это запугивание. Участник отказался от апелляции»

Обсуждение результатов показывает, что Всероссийская олимпиада школьников в ее текущей конфигурации системно не достигает тех требований к качеству измерения, которые закреплены в современной теории валидности и реализуются в сопоставимых по масштабу системах оценивания. Систематический обзор и эмпирические данные позволяют утверждать, что дефицит валидности носит не точечный, а многоуровневый характер, затрагивая содержательные, конструктные, структурные аспекты измерения, обобщаемость и внешние связи результатов, а также практически игнорируя анализ последствий олимпиадного отбора для различных групп участников. При этом выявленные в качественном исследовании практики региональных жюри демонстрируют, что существующая модель децентрализации не сопровождается эквивалентным по силе и глубине механизмом централизованного контроля качества, что приводит к фрагментации стандартов и подрывает идею единого всероссийского соревнования.

Сопоставление с другими системами оценки с высокими ставками подчеркивает, что уникальность ВсОШ заключается не только в ее масштабе, но и в отсутствии стабильной психометрической инфраструктуры. Тогда как единый государственный экзамен, стандартизированные вступи-

тельные тесты США и китайский экзамен Gaokao выстраивают различные, но институционально закрепленные комбинации централизованной разработки материалов, многоступенчатой проверки заданий и регулярного мониторинга качества, олимпиадная система фактически передаёт ключевые решения на региональный уровень без сопоставимого сопровождения методами образовательных измерений. В результате комбинация очень высокой децентрализации с минимальным уровнем формализованной валидации делает ВсОШ исключением среди крупных систем, определяющих образовательные траектории миллионов учащихся, и создает значительные риски как для точности измерения, так и для восприятия справедливости процедуры участниками и их наставниками.

Предложенные практические варианты реформы демонстрируют, что на уровне образовательной политики существует, по сути, выбор между тремя типами компромиссов. Радикальная централизация разработки заданий могла бы обеспечить максимальный контроль качества, но чревата утратой региональной автономии, сужением разнообразия содержательных подходов и значительным ростом затрат, что делает этот сценарий малореалистичным в ближайшей перспективе. Модель жесткого федерального контроля при сохранении полной региональной разработки, напротив, опирается на идею широкой сети специалистов по образовательным измерениям в регионах и дорогостоящей инфраструктуры пилотирования и анализа, что также выглядит чрезмерно амбициозным для одномоментной реализации. Наиболее взвешенным решением в этих условиях представляется гибридная модель, в которой основная масса заданий проходит полный цикл проверки на федеральном уровне, а региональные разработки интегрируются через обязательную внешнюю экспертизу и анализ смещений, что позволяет одновременно повысить валидность и сохранить ограниченную, но содержательно значимую автономию субъектов.

Интерпретируя результаты, важно учитывать ограничения проведенного исследования. Выборка экспертов, несмотря на охват разных типов регионов и ролей в системе, не гарантирует полноты представления о практике малых и отдаленных территорий, где ресурсы и институциональные условия могут существенно отличаться. Одномоментный характер сбора данных не позволяет судить о динамике изменений: возможно, в отдельных регионах уже запущены инициативы по усилению прозрачности и валидации, которые пока не успели проявиться в виде устойчивых практик. Кроме того, опора на самоотчеты экспертов и анализ доступных документов не заменяет необходимого количественного подтверждения гипотез о наличии смещений и различий между группами участников; прямой анализ дифференциального функционирования заданий и продольное отслеживание образовательных траекторий победителей и призеров остаются задачами для последующих исследований. Тем не менее, совокупность представленных данных позволяет говорить о структурном характере проблем валидности в ВсОШ и задает рамку для дальнейшей научной и управленческой работы по преобразованию олимпиады из преимущественно традиционной конкурсной практики в систему отбора, опирающуюся на современные стандарты образовательных измерений и справедливости.

Заключение. Всероссийская олимпиада школьников сегодня определяет жизненные траектории более 6 млн. школьников ежегодно, но при этом фактически функционирует без полноценной психометрической валидации, что находится в прямом противоречии с современными международными стандартами образовательных измерений и требованиями к систе-

мам с высокими ставками. Это означает, что инструмент, от которого зависят доступ к лучшим университетам, стипендии и академическая репутация участников, не проходит тех процедур проверки качества, которые считаются обязательными даже для существенно менее масштабных экзаменационных систем.

Проведенный анализ показал, что ВсОШ не достигает ни одного из шести ключевых компонентов валидности в рамках единой концепции С.Мессика: содержательная валидность обеспечивается лишь частично, конструктная и структурная валидность практически не подтверждены эмпирически, валидность обобщения и внешняя валидность ограничены, а валидность последствий фактически отсутствует как предмет систематического мониторинга. Пять критических проблем – отсутствие пилотирования заданий, непрозрачность критериев оценивания, локальные интерпретации стандартов, конфликты интересов в жюри и нерегулируемые практики рассмотрения апелляций – последовательно проявляются как в литературе, так и в интервью с экспертами и выглядят не набором случайных отклонений, а скорее устойчивым паттерном работы системы.

Теоретическая интерпретация этих проблем через призму децентрализованной архитектуры ВсОШ показывает, что в основе кризиса валидности лежит структурное противоречие: по мере того, как ответственность за разработку и оценивание заданий передается на региональный уровень, требования к централизованному контролю качества должны возрастать, но фактический контроль при этом ослабевает. В результате формируется «парадокс децентрализации», при котором максимальная автономия регионов сочетается с минимальной степенью психометрической и организационной проверки, создавая критический дефицит доверия, прозрачности и воспроизводимости результатов. В таких условиях даже добросовестные практики отдельных жюри не способны компенсировать системные риски неэквивалентности условий и возможной дискриминации различных групп участников.

Предложенная гибридная модель реформы, в рамках которой около 70% олимпиадных заданий разрабатываются и полностью валидизируются на федеральном уровне, а оставшиеся 30% создаются регионами, но проходят обязательный внешний контроль (включая анализ смещений и структурных характеристик), предлагает реалистичный путь к разрешению этого противоречия. Такой подход позволяет совместить необходимый уровень стандартизации и контроля качества с сохранением содержательной автономии и инициативы регионов, а также выстроить поэтапную трехлетнюю траекторию перехода к устойчивой системе валидации, включающей подготовку специалистов по образовательным измерениям, создание региональных центров экспертизы и регулярную публикацию технических отчетов.

Путь вперед для российской олимпиады видится в немедленном развертывании систематической психометрической валидации ВсОШ на основе гибридной модели, с четким законодательным и организационным закреплением процедур апробации, анализа качества заданий и мониторинга последствий олимпиадного отбора. Это потребует политической воли, целевых инвестиций и переобучения ключевых кадров, но без такого шага система будет и дальше опираться на традицию и репутацию при явном несоответствии современным научным и этическим требованиям к справедливому отбору одаренных школьников. В долгосрочной перспективе внедрение валидации не только повысит качество и справедливость олимпиад, но и сможет восстановить доверие участников, их семей и академического сообщества к ВсОШ как к действительно надежному и честному механизму выявления и поддержки талантов.

ЛИТЕРАТУРА

1. *Messick S.* Validity // Educational Measurement / Ed. by R.L.Linn. N.Y.: American Council on Education, 1989. P. 13–103.

2. *Kane M.T.* Validating the interpretations and uses of test scores // Journal of Educational Measurement. 2013. № 50 (1). P. 1–73.

3. *Benton T.* (2024). How long should a high-stakes test be? // Research Matters 38. Cambridge, 2024. P. 28–47.

4. *Martinková P., Drabinová A., Zvára K., Potužníková E.* Why differential item functioning analysis should be a routine part of developing conceptual assessments // CBE–Life Sciences Education. 2017. № 16 (2). [Электронный ресурс]. URL: https://doi.org/10.1187/cbe.16-10-0307.

5. *Fleishman J.A., Spector W.D., Altman B.M.* Impact of differential item functioning on age and gender comparisons of functional disability. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences. 2002. No. 57 (5). P. 275–284. [Электронный ресурс]. URL: https://doi.org/10.1093/geronb/57.5.S275.

6. *Болотов В.А.* Прошлое, настоящее и возможное будущее российской системы оценки качества образования // Вопросы образования. 2018. № 3. С. 287–297.

7. *Федоров О.Д., Веринчук К.Д.* ЕГЭ по истории и валидность его результатов: эксперты региональных предметных комиссий об открытых заданиях // Вопросы образования. 2021. № 3. С. 189–211. [Электронный ресурс]. URL: https://doi.org/10.17323/1814-9545-2021-3-189-211.

8. *Иванова С.В., Малыгин А.А.* Доверие как социально-дидактическая проблема при измерении образовательных результатов // Ценности и смыслы. 2024. № 3 (91). С. 53–71.

9. Приказ Министерства просвещения Российской Федерации от 27 ноября 2020 г. № 678 «Об утверждении Порядка проведения всероссийской олимпиады школьников». [Электронный ресурс]. URL: https://docs.cntd.ru/document/573339049 (дата обращения 14.01.2026).

10. *Engelbrecht J., Mwambakana J.* Validity and diagnostic attributes of a Mathematics Olympiad for Junior High School Contestants // Journal of Educational Measurement. 2016. № 53 (1). P. 18–37. URL: https://doi.org/10.1111/jedm.12098.

11. *Florian L., Keßler C., Müller K., Schaper N.* Assessing High Performers in the Life Sciences: Characteristics of Exams Used at the International Biology Olympiad (IBO) and Their Implications for Life Science Education // CBE–Life Sciences Education. 2019. URL: https://doi.org/10.1187/cbe.19-10-0215.

12. American Educational Research Association, American Psychological Association, National Council on Measurement in Education // Standards for Educational and Psychological Testing. AERA, 2014.

13. *Bond T.G., Fox C.M.* Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd ed.). Routledge, 2015.

14. *Cohen J.* Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates, 1988.

15. *Holland P.W., Wainer H.* (Eds.). Differential Item Functioning. Lawrence Erlbaum Associates, 1993.

16. *Lipsky M.* Street-level bureaucracy: Dilemmas of the individual in public services. Russell Sage Foundation, 1980.

Дата поступления – 13.01.2026

Decentralization paradox in educational assessment: validity of the All-Russian School Olympiad

Artem P. Gulov – Dr. Sci. (Pedagogics), associate professor of the Department of the English language, MGIMO University (Moscow, Russia); gulov@tea4er.org

Abstract. *The All-Russian School Olympiad encompasses over 6 million participants annually, determining their educational trajectories, but operates without systematic psychometric validation. This fact*

contradicts international standards for high-stakes assessment systems. The study aimed to comprehensively assess the validity of the All-Russian Olympiad through literature review and qualitative research of regional jury practices (37 experts from 12 regions). Semi-structured interviews with dual coding (κ=0.76) were employed. Five persistent problems were identified: absence of item piloting (89.2% of experts), lack of transparency in evaluation criteria (20% of regions), local interpretations of standards (70.3%), conflicts of interest within juries (54.1%), and unregulated appeals procedures (39.2%). The All-Russian Olympiad fails to meet any of the six components of Messick's unified validity framework. A hybrid reform model is proposed: 70% of items are developed at the federal level, while 30% of regional developments undergo external review with differential item functioning analysis.

Key words. *Test validity, assessment fairness, academic olympiads, decentralized assessment systems.*

REFERENCES

1. *Messick S.* Validity. *Educational Measurement*. Ed. R.L.Linn. N.Y.: American Council on Education, 1989. P. 13–103.

2. *Kane M.T.* Validating the interpretations and uses of test scores. *Journal of Educational Measurement*. 2013. No. 50 (1). P. 1–73.

3. *Benton T.* (2024). How long should a high-stakes test be? *Research Matters 38*. Cambridge, 2024. P. 28–47.

4. *Martinková P., Drabinová A., Zvára K., Potužníková E.* Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE–Life Sciences Education*. 2017. No. 16 (2). URL: https://doi.org/10.1187/cbe.16-10-0307.

5. *Fleishman J.A., Spector W.D., Altman B.M.* Impact of differential item functioning on age and gender comparisons of functional disability. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences.* 2002. No. 57 (5). P. 275–284. URL: https://doi.org/10.1093/geronb/57.5.S275.

6. *Bolotov V.A.* Proshloye, nastoyashcheye i vozmozhnoye budushcheye rossiyskoy sistemy ocenki kachestva obrazovaniya [Past, present and possible future of the Russian system for assessing the quality of education]. *Voprosy obrazovaniya* [Questions of Education]. 2018. No. 3. P. 287–297.

7. *Fedorov O.D., Verinchuk K.D.* EGE po istorii i validnost' yego rezul'tatov: ekspertu regional'nykh predmetnykh kommissiy ob otkrytykh zadaniyakh [Unified State Examination in History and validity of its results: experts of regional subject commissions on open-ended tasks]. *Voprosy obrazovaniya* [Questions of Education]. 2021. No. 3. P. 189–211. URL: https://doi.org/10.17323/1814-9545-2021-3-189-211.

8. *Ivanova S.V., Malygin A.A.* Doveriye kak social'no-didakticheskaya problema pri izmerenii obrazovatel'nykh rezul'tatov [Trust as a socio-didactic problem in measuring educational outcomes]. *Tsennosti i smysly* [Values and Meanings]. 2024. No. 3 (91). P. 53–71.

9. Prikaz Ministerstva prosveshcheniya Rossijskoj Federacii ot 27 noyabrya 2020 g. № 678 «Ob utverzhdenii Poryadka provedeniya vserossijskoj olimpiady shkol'nikoV» [Ministry of Education of the Russian Federation. Order No. 678 of November 27, 2020 "On the approval of the Procedure for conducting the All-Russian olympiad for schoolchildren"]. URL: https://docs.cntd.ru/document/573339049 (accessed 14.01.2026).

10. *Engelbrecht J., Mwambakana J.* Validity and diagnostic attributes of a Mathematics Olympiad for Junior High School Contestants. *Journal of Educational Measurement.* 2016. No. 53 (1). P. 18–37. URL: https://doi.org/10.1111/jedm.12098.

11. *Florian L., Keßler C., Müller K., Schaper N.* Assessing High Performers in the Life Sciences: Characteristics of Exams Used at the International Biology Olympiad (IBO) and Their Implications for Life Science Education. *CBE–Life Sciences Education.* 2019. URL: https://doi.org/10.1187/cbe.19-10-0215.

12. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for Educational and Psychological Testing. AERA, 2014.

13. *Bond T.G., Fox C.M.* Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd ed.). Routledge, 2015.

14. *Cohen J.* Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates, 1988.

15. *Holland P.W., Wainer H.* (Eds.). Differential Item Functioning. Lawrence Erlbaum Associates, 1993.

16. *Lipsky M.* Street-level bureaucracy: Dilemmas of the individual in public services. Russell Sage Foundation, 1980.

Submitted – 13.01.2026