Психология. Журнал Высшей школы экономики. 2024. Т. 21. № 1. С. 101–122.
Psychology. Journal of the Higher School of Economics. 2024. Vol. 21. N 1. P. 101–122.
DOI: 10.17323/1813-8918-2024-1-101-122

КАК СОВРЕМЕННАЯ ПСИХОМЕТРИКА МОЖЕТ ИЗМЕНИТЬ ТЕСТИРОВАНИЕ ИНТЕЛЛЕКТА ДЕТЕЙ И ПОДРОСТКОВ В РОССИИ

Т.Н. КАНОНИР[a], Д.А. ФЕДЕРЯКИН[b], Т.И. ЛОГВИНЕНКО[c], Е.А. ОРЕЛ[a], А.А. КУЛИКОВА[a]

[a] *Национальный исследовательский университет «Высшая школа экономики», 101000, Россия, Москва, ул. Мясницкая, д. 20*

[b] *Университет им. Иоганна Гутенберга в г. Майнц, 55128, Германия, ул. Якоба Вельдера, 9, офис 14/219*

[c] *Научно-технологический университет «Сириус», 354340, Россия, Краснодарский край, Федеральная территория «Сириус», Олимпийский просп, д. 1*

How Modern Psychometrics Can Change Intelligence Testing of Children and Adolescents in Russia

T.N. Kanonire[a], D.A. Federiakin[b], T.I. Logvinenko[c], E.A. Orel[a], A.A. Kulikova[a]

[a] *HSE University, 20 Myasnitskaya Str., Moscow, 101000, Russian Federation*

[b] *University of Mainz, 9 Jakob-Welder-Weg, Mainz, 55128, Germany, 14/219*

[c] *Sirius University of Science and Technology, 1 Olympic Ave, Sirius Federal Territory, 354340, Krasnodar region, Russian Federation*

Резюме

Оценка уровня интеллектуального развития у детей и подростков является важным элементом психодиагностики и позволяет понять природу трудностей, связанных с обучением. Тесты на интеллект имеют более чем столетнюю историю, но даже современные версии наиболее популярных тестов интеллекта (например, тест Векслера или Стенфорда–Бине) ограничены наследием своих бумажных версий. В то же время с появлением вычислительных мощностей и компьютеризации тестов психометрика претерпела значительное развитие и открыла

Abstract

The assessment of intellectual development in children and adolescents is an essential component of psychodiagnostics, facilitating an understanding of the nature of learning difficulties. Intelligence tests have a history that spans more than a century, and even contemporary versions of widely used intelligence tests (such as the Wechsler or Stanford-Binet tests) are limited by the influence of their traditional paper-based formats. Nevertheless, the advent of computational capabilities and the digitalization of tests have led psychometrics to

Исследование выполнено при поддержке Российского научного фонда, проект № 23-28-01399.
The study was funded by Russian Science Foundation (RSF), project # 23-28-01399.

новые возможности в оценивании. Применение современной психометрики в полной мере может привести к революционным изменениям в измерении интеллекта. В статье предложено описание текущей ситуации в сфере измерения интеллекта детей и подростков в России и мире и рассмотрено применение новых методологических решений для улучшения качества оценивания. Представлены такие возможности современной психометрики, как использование инновационных типов заданий, использование универсального дизайна при разработке теста, компьютерное адаптивное тестирование и многоступенчатое тестирование, применение многомерных моделей IRT для обоснования конструктной и критериальной валидности, переход от дискретных норм к непрерывному нормированию, лонгитюдные измерения. Совокупность этих свойств компьютерного формата тестирования интеллекта позволит 1) повысить мотивацию респондентов и сократить время прохождения теста, 2) оценивать прогресс ребенка и делать это так часто, как необходимо, предъявляя каждый раз разные версии теста, 3) существенно снизить ошибки при администрировании теста и подсчете результатов, а также обеспечить его защиту от использования непрофессионалами.

Ключевые слова: тесты интеллекта, психометрика, современная теория тестирования (IRT), компьютерное адаптивное тестирование, многоступенчатое тестирование, дети школьного возраста.

Канонир Татьяна Николаевна — доцент, Институт образования, Национальный исследовательский университет «Высшая школа экономики», доктор психологических наук, PhD.
Сфера научных интересов: измерения в социальных науках, интеллект, субъективное благополучие.
Контакты: tkanonir@hse.ru

Федерякин Денис Александрович — научный сотрудник, Департамент экономического образования, Университет им. Иоганна Гутенберга в г. Майнц (Германия).

experience substantial development, opening new possibilities for assessment. The complete application of modern psychometrics holds the promise of revolutionizing the measurement of intelligence. This paper presents the current situation of measuring intelligence in children and adolescents in the world and in Russia. It describes the potentialities presented by modern psychometrics, including the utilization of innovative task formats such as technology-enhanced items, the integration of universal design principles in test development, the adoption of computerized adaptive testing (CAT) and multistage testing (MST) methodologies, the application of multidimensional IRT models to establish both construct and criterion validity, the shift from discrete norms to continuous scaling, and the incorporation of longitudinal measurements. The combination of these features within a computerized intelligence testing framework will enhance respondents' motivation and simplify the test administration process. It will facilitate frequent assessments of a child's progress by presenting varying test versions on each occasion, thereby significantly reducing errors in administration and scoring, while also safeguarding against misuse by non-professionals.

Keywords: intelligence tests, psychometrics, item response theory (IRT), computerized adaptive testing, multistage testing, school-children.

Tatjana N. Kanonire — Associate Professor, Institute of Education, HSE University, PhD in Psychology.
Research Area: measurement in social sciences, intelligence, subjective well-being.
E-mail: tkanonir@hse.ru

Denis A. Federiakin — Research Fellow, The Chair of Economic Education, University of Mainz.

Сфера научных интересов: психометрика, вычислительные социальные науки.
Контакты: denis.federiakin@uni-mainz.de

Логвиненко Татьяна Игоревна — клинический психолог, старший специалист, Научный центр когнитивных исследований, Научно-технологический университет «Сириус».
Сфера научных интересов: процессы чтения, язык и речь, расстройства обучения, психометрика, нейронауки.
Контакты: logvinenkota.spb@gmail.com

Орел Екатерина Алексеевна — старший научный сотрудник, Центр психометрики и измерений в образовании; Национальный исследовательский университет «Высшая школа экономики», кандидат психологических наук, PhD.
Сфера научных интересов: психодиагностика, разработка инструментов измерения.
Контакты: eorel@hse.ru

Куликова Алёна Александровна — научный сотрудник, Центр психометрики и измерений в образовании; Национальный исследовательский университет «Высшая школа экономики», кандидат наук об образовании.
Сфера научных интересов: измерения в психологии и образовании, разработка инструментов оценивания, социально-эмоциональное развитие.
Контакты: aponomareva@hse.ru

Research Area: psychometrics, computational social science.
E-mail: denis.federiakin@uni-mainz.de

Tatiana I. Logvinenko — clinical psychologist, Research Fellow, Center for Cognitive Sciences, Sirius University of Science and Technology.
Research Area: reading, language, learning disabilities, psychometrics, neuroscience.
E-mail: logvinenkota.spb@gmail.com

Ekaterina A. Orel — Senior Research Fellow, Center of Psychometrics and Measurement in Education, HSE University, PhD in Psychology.
Research Area: psychodiagnostics, test development.
E-mail: eorel@hse.ru

Alena A. Kulikova — Research Fellow, Center of Psychometrics and Measurement in Education, HSE University, PhD in Education.
Research Area: psychological and educational assessment, test development, social and emotional development.
E-mail: aponomareva@hse.ru

Область оценивания интеллектуального развития — одна из старейших прикладных сфер психологии. С появления тестов на интеллект можно начать отсчет существования психометрики. Именно они стали первыми комплексными инструментами психодиагностики, а некоторые из них, появившись в первой половине XX в., и сейчас являются наиболее популярными для исследований и практики — достаточно упомянуть тест Стенфорда–Бине, тест Векслера или прогрессивные матрицы Равена. Эти тесты пережили множественные изменения с момента своего создания, с учетом межпоколенческой динамики коэффициента интеллекта (IQ), концептуальных представлений об интеллекте и технологических возможностей. Особую роль тестирование интеллекта играет в психодиагностике детей и подростков, так как на основании ее результатов принимаются важные решения, имеющие большое влияние на дальнейшую жизнь ребенка. В этой статье сделан фокус на особенностях оценивания интеллекта у данных возрастных групп, описываются современное состояние, вызовы и перспективы тестирования интеллекта детей и подростков в России.

Разработчики наиболее используемых в мире тестов на интеллект регулярно выпускают их обновленные версии с новыми нормами, измененными заданиями и дополненными теоретическими обоснованиями; многие тесты имеют компьютерную форму. Однако насколько существенны эти изменения и позволяют ли они решать все задачи, стоящие перед тестами интеллекта, в том числе появившиеся в недавнее время? В рамках данной статьи мы проанализируем методологические аспекты оценивания интеллектуального развития у детей школьного возраста, рассмотрим мировой и российский опыт и предложим возможные пути развития тестов на интеллект с учетом современных подходов к разработке тестов и к психометрике.

Индивидуальные различия в интеллектуальном развитии являются важной составляющей психодиагностики у детей школьного возраста. Оценка уровня интеллектуального развития является одним из основных критериев для постановки различных диагнозов и принятия важных решений, например, о рекомендуемой учебной программе. Так, результаты оценки интеллекта являются критерием «отсечения» в случае задержки умственного развития, предположения о наличии специфических расстройств развития учебных навыков, а также будут источником важной сопутствующей информации в случае расстройства аутистического спектра (World Health Organization, 2019). В других случаях, например при диагностике интеллектуальных способностей в процессе профориентации, особенности интеллектуального развития сами оказываются фокусом оценивания. Но в каждом из этих случаев критически важным является точное и валидное измерение интеллекта каждого ребенка, что, несомненно, предъявляет повышенные требования к качеству используемых тестов.

Несмотря на то что сегодня существует много теоретических подходов к интеллекту (Kaufman et al., 2013), в сфере оценивания лидируют психометрические теории, объясняющие индивидуальные различия в интеллектуальных способностях. В частности, к наиболее популярным теориям этой группы относится модель когнитивных способностей Кеттелла–Хорна–Керролла (Cattell–Horn–Carroll theory of cognitive abilities, CHC-модель), объединившая в себе модель флюидного и кристаллизованного интеллекта Р. Кеттелла и Дж. Хорна и модель трех страт Дж. Керролла (Carroll, 1993; Horn, 1976; McGrew, 2009). Тесты, имеющие в своей основе психометрические модели интеллекта, такие как тест Векслера (Wechsler Intelligence Scale for Children-V, WISC-V), тест Вудкока–Джонсона (Woodcock–Johnson-III), батарея тестов Кауфмана (Kaufman Adolescent and Adult Intelligence Test, KAIT; Kaufman Assessment Battery for Children, KABC), «Дифференциальный тест способностей» (Differential Ability Scales, DAS), корректируют свои теоретические обоснования в соответствии с CHC-моделью, так как она имеет сильные эмпирические основания, легко операционализируется для применения в психодиагностических целях, а накопленный эмпирический и практический опыт позволяет делать с ее помощью точные и развернутые интерпретации полученных индивидуальных результатов.

В России ситуация с инструментами оценивания интеллектуального развития детей и подростков, их качеством и пригодностью для индивидуального оценивания более острая. Во многом это связано с фактической остановкой разработки и внедрения инструментов диагностики индивидуальных различий на период с 1936 г. по конец 1960-х гг. И хотя сегодня ситуация в области разработки тестов и психометрики изменилась в лучшую сторону, все равно можно констатировать существенные дефициты, на которые еще в 2008 г. указывал Н.А. Батурин, а именно: малое количество качественных отечественных психодиагностических методик и использование (часто неправомерное) устаревших или неадаптированных зарубежных методик; нехватка специально подготовленных разработчиков методик и психометриков; низкая психодиагностическая и психометрическая культура пользователей методик; неконтролируемое распространение психодиагностических методик (Батурин, 2008).

В 2004 г. на страницах журнала «Психология. Журнал Высшей школы экономики» развернулась значимая для отечественной науки об интеллекте дискуссия. О возможностях и ограничениях тестирования интеллекта высказались А.Г. Шмелев, Д.Б. Богоявленская, М.А. Холодная, Д.В. Ушаков. Были поставлены вопросы методической и методологической зрелости пользователей и разработчиков тестов (Шмелев, 2004), обсуждалась принципиальная возможность получения валидных индивидуальных результатов тестирования интеллекта (Холодная, 2004), приводились доказательства прогностической валидности тестов интеллекта (Ушаков, 2004). В заключительном слове Д.В. Ушаков предложил четкую программу действий по формированию культуры эффективной разработки и применения тестов интеллекта в России: 1) создание праксиологии тестирования, 2) накопление массива исследований с целью получения нормативных данных для разных групп респондентов, 3) разработка тестов интеллекта нового поколения, которые «фиксировали бы не только текущий срез когнитивной системы, но и историю ее развития» (Ушаков, 2004, с. 90). Некоторые шаги в реализации этой программы были сделаны. К ним можно отнести и создание А.Г. Шмелевым профессионального экспертного сообщества в области тестирования и психодиагностики, и выпуск «Стандарта тестирования персонала организации» (Батурин и др., 2015), и попытку Н.А. Батурина наладить выпуск «Ежегодника тестовых рецензий и обзоров» (к сожалению, состоялись только два выпуска — в 2010 и в 2013 гг. (Батурин, Эйдман, 2010, 2013)), и обсуждение существующей культуры разработчиков тестов, оказывающих влияние на качество инструментов (Науменко, Орел, 2010; Поддьяков, 2007), и разработку собственной структурно-динамической теории Д.В. Ушаковым (Ушаков, 2011). Более того, дискуссия вокруг темы интеллекта в России продолжает развиваться: например, Д.В. Ушаков и А.А. Григорьев выводят ее на новый концептуально-философский уровень (Ushakov, Grigoriev, 2016); М.А. Холодная продолжает развивать и дополнять онтологическую теорию интеллекта (Kholodnaya, Volkova, 2016) и предлагает альтернативы для доминирующей психометрической модели, которые также могут выступить основаниями для новых подходов к оценке (Холодная, 2015).

В данной работе мы хотели бы подробнее затронуть вопрос оценивания, измерения интеллекта. В следующем параграфе мы подробнее раскроем современное состояние российской сферы тестирования интеллекта и приведем обзор методик, предназначенных для оценки интеллекта детей и подростков.

В российском контексте можно выделить несколько сфер психологической практики, где тесты интеллекта используются наиболее активно: клиническая практика, сфера образования, а также профориентация и профессиональный отбор. В клинической практике и в сфере образования основу диагностики составляет сочетание анализа клинической картины и результатов стандартизированных методик. В качестве стандартизированных методик оценивания интеллектуального развития детей школьного возраста в России, как правило, используются детский вариант теста Векслера (WISC-R) и прогрессивные матрицы Равена (Raven's Progressive Matrices, RPM). Эти тесты фигурируют в том числе и в утвержденных Министерством здравоохранения в 2021 г. клинических рекомендациях в отношении интеллектуальных нарушений. Применяются также и другие адаптированные методики, например, культурно-свободный тест Кеттелла (Culture Fair Intelligence Test, CFIT), тест структуры интеллекта Амтхауэра (Intelligence Structure Test, IST). Среди отечественных разработок можно выделить «Подростковый интеллектуальный тест» и «Универсальный интеллектуальный тест», Санкт-Петербург — Челябинск (соответственно, ПИТ СПЧ и УИТ СПЧ), разработанные Н.А. Батуриным, Н.А. Курганским, И.М. Дашковым и Л.К. Федоровой, а также «Тест интеллектуального потенциала сокращенный» (ТИПС, авторы — А.Г. Шмелев и коллектив лаборатории «Гуманитарные технологии»). Однако все перечисленные инструменты обладают рядом существенных недостатков. Подробнее остановимся на особенностях перечисленных тестов.

Прогрессивные матрицы Равена в России часто используются как в исследованиях, так и в практике, однако вопрос определения нормативных показателей для них до конца не решен. Разработка норм критична для любого теста, который планируется для индивидуальной психодиагностики, так как только наличие норм позволяет сравнить результат индивида с референтной популяцией или, что реже применяется в психологии, соотнести полученный результат с определенным критерием. Именно это позволяет сделать вывод о более низком или высоком уровне выраженности измеряемого признака у конкретного человека. Попытки разработать российские нормы для прогрессивных матриц Равена предпринимались как минимум трижды: для стандартной версии И.Э. Щеткиной на выборке 432 девятиклассников 14–15 лет (Равен и др., 2002), для двадцатиминутной версии на выборке 7894 человек (Давыдов, Чмыхова, 2016), из которых 907 человек в возрасте 10–16 лет, а также для «Стандартных прогрессивных матриц плюс» на выборке из 1890 учащихся в возрасте 11–16 лет (Сорокова, Юркевич, 2014). Таким образом, имеющиеся нормы не охватывают младший школьный возраст. Также в настоящий момент ведется работа по стандартизации прогрессивных матриц Равена в лаборатории возрастной психогенетики (ПИ РАО) под руководством С.Б. Малых, но

результаты пока не опубликованы. Однако даже появление надежных нормативных значений не закроет потребности в комплексной оценке интеллекта, так как этот тест оценивает только способность к абстрактным рассуждениям (abstract reasoning).

Инструментом комплексной оценки мог бы стать детский вариант теста Векслера. Однако, в то время как в мире доступна уже его пятая версия, в России можно приобрести только версию WISC-R 1974 г. (Wechsler, 1974; Филимоненко, Тимофеев, 2001, 2016, 2020). Эта устаревшая версия имеет множество ограничений. Во-первых, устарела теоретическая модель, на базе которой она была разработана. Во-вторых, часть заданий уже не является релевантной современному контексту. В-третьих, изданные в России версии предлагают при подсчете результатов для перевода полученных сырых баллов в стандартизированные использовать нормативные таблицы оригинальной версии 1949 г. В-четвертых, в руководстве к методике нет релевантной информации о проведенных исследованиях по адаптации, валидизации или стандартизации методики в России. Таким образом, циркулирующий в России детский вариант теста Векслера обладает недостаточными и противоречивыми данными о надежности, устаревшими или неполными нормативными данными, что делает решения, принятые на основе результатов этой версии теста, заведомо не валидными.

В российской практике также используется культурно-свободный тест флюидного интеллекта Р. Кеттелла, изначально разработанный для людей от 8 до 60 лет. Исходя из информации о стандартизации в руководстве российской адаптации теста (Денисов, Дорофеев, 2003), сложно понять, какие нормы представлены авторами (оригинальные авторские, полученные в 1950-х гг., или собранные на русскоязычной выборке перед публикацией адаптированной версии). Так или иначе, их можно считать устаревшими, и, значит, согласно данным исследований, они могут давать завышенные показатели IQ ввиду большого временно́го разрыва.

Для российской выборки также был адаптирован тест структуры интеллекта Амтхауэра (Ясюкова, 2002), предназначенный для оценки интеллекта у респондентов старше 13 лет. Однако данная российская адаптация не содержит нормированных оценок по субтестам для оценки разных компонентов интеллектуальных способностей, что сужает спектр применения теста.

Первыми из отечественных методик, предназначенных для оценки интеллекта подростков и взрослых, можно назвать уже упомянутые разработки группы исследователей из ЮУрГУ и СПбГУ — «Подростковый интеллектуальный тест» для детей 10–15 лет (Батурин, Курганский, 2005) и «Универсальный интеллектуальный тест», стандартизированный для выборки подростков 13–17 лет. Кроме того, по словам авторов, была разработана адаптивная версия теста УИТ СПЧ (Батурин и др., 2011). Эта линейка использует в качестве прототипа тесты Векслера и Амтхауэра. Важное ограничение данных тестов заключается в том, что они не были валидизированы на клинических подгруппах, стандартизация проходила на выборке

обычных школьников и позволяет оценивать интеллектуальные способности, только если общий IQ > 70 (Орел, 2010).

«Тест интеллектуального потенциала сокращенный», разработанный компанией «Гуманитарные технологии», может быть использован для оценки интеллекта у подростков (8–11-й класс) и взрослых. Задания методики схожи с заданиями тестов Векслера и Амтхауэра. Данный тест был апробирован на выборке 6000 старших школьников. Существует также «Краткий тест отбора» (КТО) тех же авторов, который предназначен для быстрой оценки интеллектуальных способностей старших подростков и взрослых. Однако данные тесты заточены под нужды бизнеса и профориентации, они рассчитаны на старший возраст и не проходили апробацию на клинических подгруппах. Иными словами, данные тесты не предназначены для применения на всех ступенях школьного образования и для принятия образовательных решений.

Кроме того, для оценки интеллектуального развития применяются, например, критериально ориентированные тесты, разработанные в Психологическом институте Российской академии образования: «Школьный тест умственного развития» (ШТУР), «Тест умственного развития для абитуриентов и старшеклассников» (АСТУР), «Тест умственного развития младших школьников» (ТУРМШ) (Акимова и др., 1988; Гуревич, Акимова, 2008), а также патопсихологические методики (например, руководства С.Я. Рубинштейн или В.М. Блейхера, И.М. Крук). Однако и линейка тестов ПИ РАО, и патопсихологическое методики не базируются на нормативных данных, что противоречит фундаментальному принципу вычисления коэффициента интеллекта и потому выходит за пределы внимания данного обзора.

Также тесты интеллекта или отдельных интеллектуальных способностей разрабатывают или адаптируют некоторые коммерческие и консалтинговые компании, психологические службы профильных образовательных или профессиональных организаций для целей психологического отбора старших подростков и взрослых, однако, как правило, в публичном поле для этих инструментов нельзя найти даже минимальные данные об их психометрических свойствах.

Названными выше методиками не ограничивается список инструментов, используемых для диагностики разных аспектов интеллекта в России, однако они являются одними из наиболее распространенных и отражают актуальные проблемы в сфере психодиагностики интеллекта детей и подростков.

Итак, учитывая обозначенные проблемы с качеством используемых российскими специалистами инструментов оценивания интеллектуального развития (устаревшие нормы, не всегда релевантные стимульные материалы, ограниченные возможности разносторонней оценки и, как следствие, не поддающиеся корректной интерпретации результаты тестирования), можно было бы предложить провести адаптацию какого-либо существующего теста на интеллект с хорошей репутацией и обоснованной валидностью, с его последующей стандартизацией для российской популяции. Однако в случае адаптации необходимо учитывать как минимум два обстоятельства. Во-первых, адаптация таких инструментов сопряжена с постоянными выплатами право-

обладателям за каждое пройденное тестирование, что сильно удорожает их использование. Во-вторых, такие тесты, несмотря на свою многолетнюю популярность и последовательные шаги по поддержанию их валидности, тоже имеют существенные ограничения, являющиеся, по нашему мнению, следствием их длительной истории существования. Несмотря на то что правообладатели и разработчики этих тестов постоянно обновляют и пересматривают теоретические обоснования, содержание субтестов и нормы, необходимость поддерживать преемственность каждой последующей версии существенно ограничивает возможности внесения серьезных изменений, направленных на решение специфических проблем, таких как, например, оценка прогресса интеллектуального развития с помощью современных психометрических методов (Wilson et al., 2012), применение технологий компьютерного адаптивного тестирования (Magis, Barrada, 2017) или непрерывное нормирование тестов (Lenhard et al., 2018). В этом контексте нехватка качественного инструментария для оценки интеллектуального развития в России представляет собой одновременно и вызов, и возможность, а именно — возможность разработать оригинальный тест на новых методологических основаниях.

Методологические возможности современной психометрики

В данном разделе статьи мы сфокусируемся на описании требований к новому поколению тестов на интеллект, которые будут использовать в полной мере психометрические возможности, предлагаемые современной методологией измерений в социальных науках, а также на перечислении ограничений традиционных подходов к тестированию интеллекта.

Компьютерная форма

Зачастую компьютеризация тестирования понимается разработчиками и пользователями тестов как простой перенос бланковой формы в компьютерную. Однако сегодня цифровая среда предоставляет большое разнообразие форм заданий. К таким типам заданий могут быть отнесены задания с мультимедиа-элементами, сценарные задания, игровые и симуляционные форматы заданий и пр., и это только часть форм, которые не могут быть реализованы ни в каком другом виде, кроме компьютерного. На сегодняшний день исследователями предложено множество классификаций инновационных типов заданий (technology-enhanced items) (Parshall et al., 2009; Scalise, Gifford, 2006). Использование их в полной мере может сделать процедуру тестирования неотличимой от компьютерной игры. Это позволяет сохранить мотивацию респондентов к прохождению теста, что особенно важно для учеников школы.

Тем не менее, несмотря на привлекательность таких заданий для пользователей, есть несколько ограничений, которые нужно учитывать в тестировании интеллекта. Во-первых, разработка подобных заданий требует более длинного цикла разработки инструмента и дополнительных шагов валидизации получаемых выводов

для целей психодиагностики. С использованием такого рода заданий возрастает сложность разработки и валидизации алгоритмов начисления баллов (von Davier, Halpin, 2013). Во-вторых, непривычный вид заданий теста может вызвать недоверие у пользователей и сомнения в валидности получаемых выводов. Поэтому разработка нового поколения тестов интеллекта должна быть ориентирована на сохранение баланса между интерактивными заданиями, которые позволяют разнообразить процесс прохождения теста и обеспечить мотивацию респондентов, и более привычными формами заданий, что сохранит уверенность пользователей в валидности выводов.

Хочется отметить и еще одно потенциальное преимущество использования компьютерной формы для комплексных тестов, к которым относятся и тесты на интеллект. Как правило, такие тесты имеют достаточно сложную процедуру администрирования, включающую в себя предъявление заданий, фиксацию ответов, подсчет баллов на уровне субшкал и в целом по тесту, а также трансформацию сырых баллов в стандартизированные. Предъявление теста в компьютерной форме дает возможность максимально автоматизировать эту процедуру, а значит, и уменьшить риск потенциальных ошибок. К тому же компьютерная форма позволяет защитить тест от несанкционированного использования людьми, не имеющими соответствующей квалификации для его использования, что является важным элементом защиты инструмента и поддержания его валидности.

Универсальный дизайн теста

Тесты на интеллект, как правило, используются в психодиагностических целях, что определяет их достаточно широкую целевую аудиторию, в том числе включающую респондентов с ограниченными возможностями здоровья (ОВЗ). В то же время различные ОВЗ могут диктовать и различные условия оптимальной формы предъявления материала, которые можно учесть на этапе разработки нового теста. Подход к разработке тестов, который заранее учитывает особенности всех групп, входящих в целевую популяцию теста, получил название универсального дизайна (Ketterlin-Geller, 2005). Главная задача этого подхода — обеспечить справедливое оценивание конструкта, придерживаясь следующих принципов: 1) включить в целевую популяцию теста все группы респондентов, для которых этот тест будет применяться; 2) дать детальное определение конструкту, чтобы все возможные барьеры, не связанные с конструктом (например, сенсорные, физические и т.д.), могли быть предугаданы и устранены из заданий; 3) разрабатывать задания с учетом предыдущих пунктов, чтобы сделать их доступными для всех респондентов популяции и исключить возможные искажения (bias); 4) изначально планировать дизайн теста таким образом, чтобы можно было реализовать аккомодации (например, увеличение размера изображения); 5) инструкции теста представлять в простой и ясной форме; 6) сделать так, чтобы задания теста были максимально «читабельны» и понятны, если специфика конструкта не требует иного; 7) все тексты, таблицы, изображения и т.д. представлять в форме,

максимально удобной для восприятия, если только другого не требуется для измерения целевого конструкта (Thompson et al., 2004). Разработка теста с самого начала, в отличие от адаптации, позволяет придерживаться этих принципов, а использование компьютерной формы делает применение аккомодаций более доступным.

Адаптивность тестирования и правила остановки

Развитие компьютерных технологий позволило внедрить компьютерное адаптивное тестирование (КАТ, computerized adaptive testing) и его вариацию — многоступенчатое тестирование (МСТ, multistage testing) (Magis, Barrada, 2017). Эти подходы к тестированию основаны на применении моделей современной теории тестирования (IRT) (van der Linden, 2016), способных предсказывать вероятность выполнения всех откалиброванных заданий в банке заданий с использованием профиля ответов респондента, наблюдаемого в данный момент (Chang, Ying, 1996). В КАТ на основе этого предсказания каждое следующее задание выбирается респонденту таким образом, чтобы оно давало наибольшее количество информации об уровне его способности. В итоге два разных респондента могут получить два непересекающихся варианта теста, при этом сопоставимость их результатов будет сохранена за счет того, что параметры заданий известны до тестирования. Однако при таком подходе необходим большой банк откалиброванных заданий, как правило, около 200 на каждую субшкалу (Sahin, Anil, 2017), что требует огромных трудовых затрат от разработчиков теста.

В качестве компромиссного варианта, лишенного многих ограничений КАТ, психометриками рассматривается МСТ (Magis et al., 2017). В таком дизайне теста задания предъявляются группами (обычно по 3–7 штук), и настройка теста на респондента происходит в момент выбора следующей группы. Это позволяет существенно сократить необходимый объем банка, а также обеспечить бóльшую содержательную сопоставимость между индивидуализированными вариантами теста (Crotts et al., 2012). Существенным преимуществом МСТ, по сравнению с классическим КАТ, является стоимость разработки, поскольку требует меньшего количества заданий, сохраняя все достоинства КАТ.

Большинство инструментов диагностики интеллекта, разработанных и валидизированных на данный момент, основаны на классической линейной структуре теста, при которой все респонденты получают все задания (Oakland et al., 2016), а их результаты обрабатываются в рамках классической теории тестирования. При этом в некоторых из них (как, например, в тесте Векслера) используется правило остановки — инструкции для администраторов прекратить тестирование по субшкале, когда достигнуто определенное количество ошибок. В классической теории тестирования такое правило приводит к нарушению фундаментальных ее допущений, т.е. в таких тестах сырые баллы респондентов, получивших разные наборы заданий, нельзя сравнивать друг с другом. Подобные сравнения возможны только в IRT — основе КАТ и МСТ.

Более того, сами по себе правила остановки в тестировании являются большой темой исследований и подразумевают немалое количество психометрических допущений, которые должны быть изучены и валидизированы (Wang et al., 2013). Поэтому выбору корректной методологии стоит уделить особое внимание при разработке нового поколения тестов.

Также традиционная практика разработки компьютерных адаптивных тестов почти полностью игнорирует методологические исследования алгоритмов выбора первых заданий КАТ на основе контекстных переменных (например, социально-демографических), характеризующих респондента. Реализация таких алгоритмов позволяет существенно сократить длину адаптивных тестов и повысить их надежность за счет того, что на пилотном этапе устанавливаются регрессионные отношения между измеряемой способностью и различными контекстными характеристиками, после чего на их основе предсказывается ее ожидаемый уровень, и с этого момента начинается тестирование (Vie et al., 2018). Такие алгоритмы позволяют преодолеть проблему «холодного старта» (cold start problem) в КАТ, когда в силу отсутствия какой-либо информации о респонденте на старте тестирования алгоритм вынужден допускать, что у всех респондентов один и тот же средний уровень способности. Более того, использование контекстных характеристик в качестве коллатеральной информации в моделях латентной регрессии (подробнее об этом скажем далее) позволяет повысить надежность докладываемых баллов, не изменяя их интерпретацию (Федерякин и др., 2021).

Подходы к оценке интеллекта как композитного конструкта

С самого начала исследования интеллекта Ч. Спирменом (Spearman, 1904) он понимался как композитный конструкт, состоящий из некоторого общего фактора (интегративной способности успешно решать разнообразные интеллектуальные задачи) и специфических факторов (способностей решать частные интеллектуальные задачи конкретных типов). Так, Спирмен обнаружил, что различные частные способности коррелируют друг с другом положительно (positive manifold), что привело его к гипотезе о существовании некоторого общего фактора интеллекта. Но даже если вслед, например, за Дж. Керроллом не придерживаться идеи наличия общего фактора, интеллект все равно остается сложным и многомерным конструктом. Такого рода конструкты и способы их математического описания являются одной из популярных тем исследований в современной психометрике (Wilson, Gochyyev, 2020).

В силу дизайна инструментов измерения интеллекта и диагностических потребностей практиков, новое поколение тестов на интеллект должно иметь возможность представлять результаты респонденту как по каждому из компонентов интеллекта, так и по уровню интеллекта в целом. Исторически все алгоритмы КАТ разрабатывались применительно к одномерному тестированию, которое выдает результат только по одной шкале. Многомерный КАТ является относительно новой разработкой в психометрике (Piton-Gonçalves, Aluísio, 2012). Несмотря на повышение сложности этих алгоритмов, они обладают

огромным преимуществом: позволяют учитывать корреляции между результатами по разным шкалам. Это дает возможность использовать их баллы как коллатеральную информацию в многомерных моделях, что повышает надежность (Федерякин и др., 2021). Однако многие существующие многомерные тесты все еще строятся не на многомерных алгоритмах КАТ и МСТ, а на нескольких одномерных, что приводит к относительно низким значениям надежности измерения (Wang et al., 2004).

В традиционных психометрических моделях для композитных конструктов исследователи вынуждены делать допущение о том, что различные способности не коррелируют друг с другом, что приводит к дальнейшим неверным или даже невозможным интерпретациям. Наиболее ярко это проявляется в бифакторных моделях. Так, например, разработчики вынуждены допускать, что при измерении некоторого общего фактора математической грамотности с помощью заданий по алгебре и геометрии балл общей способности (математическая грамотность) не связан ни с первой (алгебра), ни со второй (геометрия) оцениваемыми областями знаний, а сами баллы по геометрии и алгебре не коррелируют друг с другом. Это приводит к тому, что исследователи используют только общий фактор и обращаются со специфическими факторами как с «шумом». Для практических целей эти факторы обладают крайне ценной диагностической информацией, но она становится недоступной для интерпретации в простых бифакторных моделях в силу математических допущений. Более того, надежности таких специфических факторов в ортогональных бифакторных моделях чрезвычайно низки, что дополнительно запрещает их использование в практике. В целом, то же касается и моделей с факторами второго порядка, которые в высокой степени математически родственны бифакторным (Gignac, 2016; Rijmen, 2010; Schmid, Leiman, 1957).

В ответ на это ограничение методологи разработали серию частично косоугольных бифакторных моделей, которые пытаются преодолеть описанную проблему интерпретации. Однако интерпретация таких моделей все еще остается сложной для практических целей. Для решения этой проблемы Д.А. Федерякин и М. Вилсон (Federiakin, Wilson, 2023) разработали новую бифакторную модель, которая полностью снимает эту проблему интерпретации, существующую в психометрике с 1937 г. (Holzinger, Swineford, 1937). Это полностью косоугольная бифакторная модель, которая оценивает все корреляции всех факторов (специфических — между собой и с общим) и при этом возвращает высокие показатели надежности всех баллов. Ее применение для тестов интеллекта нового поколения позволит с большей доказательной силой одновременно сообщать баллы и по общему, и по специфическим факторам.

Новые способы анализа критериальной валидности

Применение современных психометрических методов позволит улучшить предсказательную валидность тестов интеллекта. За последние 25 лет разработаны модели IRT, способные напрямую оценивать взаимосвязь измеряемой

латентной характеристики с какими-либо внешними наблюдаемыми переменными, — модели латентной регрессии (Adams et al., 1997; Christensen et al., 2004). Они характеризуются тем, что более точно, нежели классические регрессионные модели, выявляют связи между латентной характеристикой и факторами индивидуального и институционального уровня, ассоциированными с ее развитием (De Boeck, Wilson, 2004). Однако до сих пор эти модели не использовались для анализа отсроченных во времени достижений (distal life outcomes), — например, в исследованиях предсказательной валидности тестов интеллекта с применением IRT.

Результаты, полученные с помощью более простых моделей, могут содержать искажения в их интерпретации, что ставит под сомнение результаты исследования критериальной валидности, так как интерпретации связей результатов с внешними переменными будут также искажены. Несмотря на то что уже предложены отдельные работы для изучения критериальной валидности более современными методами (Nylund-Gibson et al., 2019), пока что разработчики тестов уделяют мало внимания вопросам преодоления неоднородности дисперсии распределения способности респондентов (гетероскедастичности). Гомоскедастичность является одним из фундаментальных требований к линейным моделям, которое, если нарушается, может существенно исказить оценки параметров модели (в данном случае — оценки критериальной валидности теста). Традиционные IRT-модели и факторно-аналитические модели допускают, что дисперсия способности является одинаковой для всех групп респондентов, что периодически оспаривается современными исследованиями интеллекта, например, в контексте половых различий (Johnson et al., 2008). Однако современная психометрика предлагает методы преодоления этого допущения, моделирующие дисперсию различных групп как функцию от контекстных переменных (Fischer, Molenaar, 2012). Таким образом, новые методологические решения при разработке тестов на интеллект могут дать ценный вклад в предметную дискуссию касательно вопроса критериальной валидности и однородности распределения результатов тестов интеллекта.

Новые способы нормирования тестов

Разработанный не так давно (Lenhard et al., 2018; Zachary, Gorsuch, 1985) и стремительно набирающий популярность метод непрерывного нормирования тестов может произвести революцию в сообщении тестовых баллов респондентам. Традиционные дискретные нормы тестов выглядят как большие таблицы, выделяющие по определенным критериям отдельные категории популяции, с которыми должен сравниваться респондент, а также нормативные значения тестовых баллов для этих категорий. Это приводит к парадоксальным ситуациям: например, когда нормы представлены по возрастным категориям, то разница тестирования всего в один день может переместить респондента в следующую возрастную подгруппу с более высокими нормами.

Непрерывное нормирование позволяет сгладить этот парадокс, сжимая огромные нормирующие таблицы до одного уравнения линейной регрессии. В этом уравнении среднее (ожидаемое) значение IQ предсказывается на основе нескольких социально-демографических переменных, с помощью умножения их на соответствующие коэффициенты и суммирования, что объясняет ту часть дисперсии способности, которая связана с различиями по этим контекстным переменным. Тогда дисперсия остатков начинает играть роль истинных индивидуальных различий в конструкте. Стандартизация этих различий и перевод на традиционную IQ-шкалу (нормальное распределение со средним 100 и стандартным отклонением 15) позволяет в этом случае сравнивать между собой по уровню интеллектуального различия респондентов даже из разных возрастных и демографических групп. Компьютерная форма тестирования позволяет автоматизировать процесс. В данном случае подобные сравнения носят характер «X имеет более высокий IQ при сравнении с Y точно такого же возраста и пола». Такое непрерывное нормирование позволяет индивидуально таргетировать нормативную группу под конкретного респондента, за счет допущения о функциональной связи между социально-демографическими переменными и интеллектом. Как результат, это существенно снижает требования к объему выборки при нормировании теста, но повышает требования к качеству регрессионной модели, служащей для выработки норм (Lenhard et al., 2018).

Лонгитюдные измерения и возможность отслеживания изменений

Еще один важный аспект, пока не реализованный в современных тестах интеллекта, — это использование принципов измерения прогресса с помощью IRT, а не классической теории тестирования. IRT способна измерить положение респондентов в разные моменты времени на единой шкале логитов, отражающих вероятность решения заданий (Embretson, 1991). Для оценки прогресса требуется изменение содержания теста от предыдущего момента измерения к следующему: обычно в него включаются более трудные задания, отражающие повышение уровня развития оцениваемого конструкта у респондентов. В итоге процедуры обработки результатов из классической теории тестирования становятся неприменимы в данном случае, потому что они работают на уровне сырых баллов, которые несравнимы между фактически разными тестами из разных моментов времени. Эта проблема усугубляется, если волн измерения конструкта больше, чем две.

Однако измерение прогресса возможно с применением лонгитюдных моделей IRT (Wilson et al., 2012). Сопоставление разных шкал может достигаться внедрением якорных заданий в два «смежных» варианта теста (пре-тест и пост-тест) (Loyd, Hoover, 1980). В этом случае сопоставимость шкал достигается с помощью допущения о том, что поскольку сами по себе якорные задания не изменились, не изменились и их психометрические параметры. Однако это требует повторного предъявления якорных заданий, что может вызвать претензии пользователей из-за знакомства респондентов с тестовыми заданиями.

Другой альтернативой для измерения изменений является использование КАТ или МСТ, где становится возможным предъявление респондентам абсолютно не пересекающихся вариантов тестов, которые различаются по трудности. Тем не менее в этом случае также сохраняется сопоставимость оценок параметров респондентов за счет того, что параметры заданий были оценены на одной шкале и зафиксированы при разработке банка заданий. Ценность инструментов измерения интеллекта нового поколения будет заключаться, в частности, в том, что они будут способны оценивать именно изменения в интеллекте на длительной дистанции, что откроет новые возможности для лонгитюдных исследований интеллекта. Однако для этого потенциал лонгитюдного применения таких инструментов должен быть заложен в их дизайн с самого начала разработки.

Общие выводы

Использование всех возможностей, которые предоставляет современная психометрика, без преувеличения позволит произвести революцию в измерении интеллекта, поскольку она стоит гораздо ближе к вычислительным наукам о поведении, чем к набору догм о том, как разрабатывать психологические тесты. Современные психометрические средства повышения точности, аутентичности и справедливости инструментов измерения, сбора продвинутых свидетельств валидности, а также интерпретации результатов сильно превосходят те средства, которые были доступны на момент разработки используемых сегодня тестов интеллекта детей и подростков. Эти средства были созданы до начала 1990-х гг., когда основное количество тестов проводилось в бумажном виде. С тех пор психометрика испытала настоящую революцию в исследованиях, связанную как с ростом доступных вычислительных мощностей, так и с компьютеризацией оценивания. Однако практика психодиагностики интеллекта в основной своей массе осталась на уровне развития тридцатилетней давности. Введение новых технологий в оценивание интеллекта важно не только для того, чтобы синхронизировать практику психодиагностики и передовые современные разработки, но и для того, чтобы повысить качество оценивания, например, снижая риск ошибок и некорректных использований тестовых баллов и их интерпретаций и подавляя возможности нежелательных социальных последствий тестирования. Более того, это позволит ставить новые и более доказательно отвечать на уже существующие исследовательские вопросы по отношению к конструкту интеллекта.

И хотя дефицит современных тестов на интеллект — это общая проблема, для России она стоит особенно остро, так как на сегодняшний день мы ощущаем нехватку валидных стандартизированных инструментов. Учитывая ограничения существующих инструментов и те возможности, которые открывает перед разработчиками и пользователями тестов современная психометрика, разработка принципиально нового теста для оценки интеллектуального развития представляется самым перспективным направлением.

В российском контексте более релевантно разрабатывать свои собственные оригинальные комплексные тесты, отвечающие требованиям современной психодиагностики, чем адаптировать или дорабатывать зачастую устаревшие, созданные в парадигме бумажных линейных тестов инструменты. Такие современные тесты должны иметь компьютерную форму, позволяющую реализовать преимущества многоступенчатого тестирования, непрерывное нормирование, возможность оценки прогресса, быть доступными для реализации универсального дизайна. Их разработка будет способствовать развитию сразу нескольких направлений: психодиагностики, исследований интеллекта и психометрики.

Литература

Акимова, М., Борисова, Е., Гуревич, К., Козлова, В., Логинова, Г. (1988). Школьный тест умственного развития. В кн. А. А. Бодалев, В. В. Столин (ред.), *Практикум по психодиагностике. Психодиагностические материалы* (с. 213–214). М.: Изд-во Московского университета.

Батурин, Н. А. (2008). Современная психодиагностика России. *Психология. Психофизиология*, *32*(132), 4–9.

Батурин, Н. А., Вучетич, Е. В., Костромина, С. Н., Кукаркин, Б. А., Куприянов, Е. А., Лурье, Е. В., Митина, О. В., Науменко, А. С., Орел, Е. А., Полетаева, Ю. С., Попов, А. Ю., Потапкин, А. А., Симоненко, С. И., Синицына, Ю. Д., Шмелев, А. Г. (2015). Российский стандарт тестирования персонала (временная версия, созданная для широкого обсуждения в 2015 году). *Организационная психология*, *5*(2), 67–138.

Батурин, Н. А., Дашков, И. М., Курганский, Н. А. (2011). Создание адаптивного теста интеллекта на основе апробированного тестового материала. *Вестник Южно-Уральского государственного университета. Серия: Психология*, *14*(29), 15–19.

Батурин, Н., Курганский, Н. (2005). Разработка и стандартизация теста интеллекта для среднего школьного возраста. *Психологическая наука и образование*, *3*, 74–85.

Батурин, Н., Эйдман, Е. Н. (ред.). (2010). *Ежегодник профессиональных рецензий и обзоров. Методики психологической диагностики и измерения* (т. 1). Челябинск: Изд-во ЮУрГУ.

Батурин, Н., Эйдман, Е. Н. (ред.). (2013). *Ежегодник профессиональных рецензий и обзоров. Методики психологической диагностики и измерения* (т. 2). Челябинск: Изд-во ЮУрГУ.

Гуревич, К. М., Акимова, М. К. (2008). *Психологическая диагностика* (3-е изд., перераб.). СПб.: Питер.

Давыдов, Д. Г., Чмыхова, Е. В. (2016). Применение теста Стандартные прогрессивные матрицы Равена в режиме ограничения времени. *Вопросы психологии*, *4*, 129–139.

Денисов, А., Дорофеев, Е. (2003). *Интеллектуальный тест Р. Кеттелла. Диагностика культурно-независимого интеллекта (методическое руководство)*. СПб.: ИМАТОН.

Науменко, А., Орел, Е. (2010). А судьи кто? Индивидуальные особенности разработчиков и характеристики тестовых заданий. *Психологические исследования*, *3*(12). https://doi.org/10.54359/ps.v3i12.911

Орел, Е. А. (2010). Универсальный интеллектуальный тест Санкт-Петербург – Челябинск Модернизированный, УИТ СПЧ-М. В кн. *Ежегодник профессиональных рецензий и обзоров. Методики психологической диагностики и измерения* (с. 209–224). Челябинск: Издательский центр ЮУрГУ.

Поддьяков, А. (2007). Тестирование интеллекта, конкуренция и рефлексия. *Рефлексивные процессы и управление*, *2*(7), 46–56.

Равен, Дж., Курт, Дж., Равен, Дж. (2002). *Руководство к Прогрессивным Матрицам Равена и Словарным Шкалам*. М.: Когито-Центр.

Сорокова, М. Г., Юркевич, В. С. (2014). Стандартизация теста «СПМ Плюс Равена» на московской выборке. *Дефектология*, *6*, 28–37.

Ушаков, Д. В. (2004). Тесты интеллекта, или горечь самопознания. *Психология. Журнал Высшей школы экономики*, *1*(2), 76–93.

Ушаков, Д. (2011). *Психология интеллекта и одаренности*. М.: Институт психологии РАН.

Федерякин, Д. А., Угланова, И. Л., Скрябин, М. А. (2021). Новые источники информации в компьютерном тестировании. *Вестник Томского государственного университета*, *465*, 179–187. https://doi.org/10.17223/15617793/465/24

Филимоненко, Ю. И., Тимофеев, В. И. (2001). *Тест Д. Векслера (диагностика структуры интеллекта (детский вариант)). Методическое руководство*. М.: ИМАТОН.

Филимоненко, Ю. И., Тимофеев, В. И. (2016). *Тест Д. Векслера. Диагностика структуры интеллекта (детский вариант)*. СПб.: ИМАТОН.

Филимоненко, Ю. И., Тимофеев, В. И. (2020). *Тест Д. Векслера (диагностика структуры интеллекта (детский вариант)). Методическое руководство*. СПб.: ИМАТОН.

Холодная, М. А. (2004). Психологическое тестирование и право личности на собственный вариант развития. *Психология. Журнал Высшей школы экономики*, *1*(2), 66–75.

Холодная, М. А. (2015). Интеллект, креативность, обучаемость: Ресурсный подход (о развитии идей В.Н. Дружинина). *Психологический журнал*, *36*(5), 5–14.

Шмелев, А. Г. (2004). Тест как оружие. *Психология. Журнал Высшей школы экономики*, *1*(2), 40–53.

Ясюкова, Л. (2002). *Тест структуры интеллекта Р. Амтхауэра (IST): Методическое руководство*. СПб.: ИМАТОН.

Ссылки на зарубежные источники см. в разделе References.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47–76.

Akimova, M., Borisova, E., Gurevich, K., Kozlova, V., & Loginova, G. (1988). Shkol'nyi test umstvennogo razvitiya [School test for intellectual development]. In A. A. Bodalev & V. V. Stolin (Eds.), *Praktikum po psikhodiagnostike. Psikhodiagnosticheskie materialy* [Tutorial in psychodiagnostics. Psychodiagnostic materials] (pp. 213–214). Moscow: Moscow University Press.

Baturin, N. A. (2008). Sovremennaya psikhodiagnostika Rossii [Contemporary Psychodiagnostics in Russia]. *Psikhologiya. Psikhofiziologiya, 32*(132), 4–9.

Baturin, N. A., Dashkov, I. M., & Kurganskii, N. A. (2011). Sozdanie adaptivnogo testa intellekta na osnove aprobirovannogo testovogo materiala [The development of an intelligence test on the basis of the approbated testing material]. *Vestnik Yuzhno-Ural'skogo Gosudarstvennogo Universiteta. Seriya: Psikhologiya, 14*(29), 15–19.

Baturin, N., & Eidman, E. N. (Eds.). (2010). *Ezhegodnik professional'nykh retsenzii i obzorov. Metodiki psikhologicheskoi diagnostiki i izmereniya* [A yearbook of professional reviews. Methods of psychological diagnostics and measurement] (Vol. 1). Chelyabinsk: YuUrGU.

Baturin, N., & Eidman, E. N. (Eds.). (2013). *Ezhegodnik professional'nykh retsenzii i obzorov. Metodiki psikhologicheskoi diagnostiki i izmereniya* [A yearbook of professional reviews. Methods of psychological diagnostics and measurement] (Vol. 2). Chelyabinsk: YuUrGU.

Baturin, N., & Kurganskii, N. (2005). *Razrabotka i standartizatsiya testa intellekta dlya srednego shkol'nogo vozrasta* [Development and standardization of the intelligence test for the preteen age]. *Psikhologicheskaya Nauka i Obrazovanie, 3*, 74–85.

Baturin, N., Vuchetich, E., Kostromina, S., Kukarkin, B., Kupriyanov, E., Lurie, E., Mitina, O., Naumenko, A., Orel, E., Poletaeva, Y., Popov, A., Potapkin, A., Simonenko, S., Sinitsina, Y., & Shmelyov, A. (2015). Russian Standard for Personnel Testing (Interim version, designed for a discussion). *Organizatsionnaya Psikhologiya [Organizational Psychology], 5*(2), 67–138. (in Russian)

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*(3), 213–229.

Christensen, K. B., Bjorner, J. B., Kreiner, S., & Petersen, J. H. (2004). Latent regression in loglinear Rasch models. *Communications in Statistics-Theory and Methods, 33*(6), 1295–1313.

Crotts, K., Sireci, S. G., & Zenisky, A. (2012). Evaluating the content validity of multistage-adaptive tests. *Journal of Applied Testing Technology*, *13*(1). https://jattjournal.net/index.php/atp/article/view/48368

Davydov, D. G., & Chmykhova, E. V. (2016). Primenenie testa Standartnye progressivnye matritsy Ravena v rezhime ogranicheniya vremeni [Application of the Raven Standard Progressive Matrices with the time limit mode]. *Voprosy Psikhologii, 4*, 129–139.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach* (Vol. 10, pp. 978–1001). New York, NY: Springer.

Denisov, A., & Dorofeev, E. (2003). *Intellektual'nyi test R. Kettella. Diagnostika kul'turno-nezavisimogo intellekta (metodicheskoe rukovodstvo)* [R. Cattell's Intelligence Test. Diagnostics of culturally independent intelligence (methodical instructions)]. Saint Petersburg: IMATON.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495–515.

Federiakin, D. A., Uglanova, I. L., & Skryabin, M. A. (2021). New sources of information in computerized testing. *Vestnik Tomskogo Gosudarstvennogo Universiteta [Tomsk State University Journal], 465*. 179–187**.** https://doi.org/10.17223/15617793/465/24

Federiakin, D., & Wilson, M. R. (2023). *Identification and interpretation of the completely oblique Rasch Bifactor Model* [Manuscript under review]. HSE University, Russian Federation; Johannes Gutenberg University of Mainz, Germany; University of California in Berkeley, USA.

Filimonenko, Yu. I., & Timofeev, V. I. (2001). *Test D. Vekslera (diagnostika struktury intellekta (detskii variant)). Metodicheskoe rukovodstvo* [D. Wechsler's Test (diagnostics of intellect structure (for childen)). Methodical instructions]. Saint Petersburg: IMATON.

Filimonenko, Yu. I., & Timofeev, V. I. (2016). *Test D. Vekslera. Diagnostika struktury intellekta (detskii variant)* [D. Wechsler's Test. Diagnostics of intellect structure (for children)]. Moscow: IMATON.

Filimonenko, Yu. I., & Timofeev, V. I. (2020). *Test D. Vekslera (diagnostika struktury intellekta (detskii variant)). Metodicheskoe rukovodstvo* [D. Wechsler's Test (diagnostics of intellect structure (for children). Methodical instructions]. Saint Petersburg: IMATON.

Fischer, G. H., & Molenaar, I. W. (2012). *Rasch models: Foundations, recent developments, and applications*. Springer Science & Business Media.

Gignac, G. E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence*, *55*, 57–68.

Gurevich, K. M., & Akimova, M. K. (2008). *Psikhologicheskaya diagnostika* [Psychological diagnostics] (3rd ed.). Saint Petersburg: Piter.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54.

Horn, J. L. (1976). Human abilities: A review of research and theory in the early 1970s. *Annual Review of Psychology*, *27*(1), 437–485.

Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science*, *3*(6), 518–531.

Kaufman, J. C., Kaufman, S. B., & Plucker, J. A. (2013). Contemporary theories of intelligence. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 811–822). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.013.0051

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *The Journal of Technology, Learning and Assessment*, *4*(2). https://ejournals.bc.edu/index.php/jtla/article/view/1649

Kholodnaya, M. A. (2004). Psychological testing and the right of a person to choose her/his own path of development. *Psychology. Journal of Higher School of Economics, 1*(2), 66–75. (in Russian)

Kholodnaya, M. A. (2015). Intelligence, creativity, learning capability: resource approach (on development of V.N. Druzhinin's ideas). *Psikhologicheskii Zhurnal, 36*(5), 5–14. (in Russian)

Kholodnaya, M. A., & Volkova, E. V. (2016). Conceptual structures, conceptual abilities and productivity of cognitive functioning: The ontological approach. *Procedia – Social and Behavioral Sciences*, *217*, 914–922. https://doi.org/10.1016/j.sbspro.2016.02.040

Lenhard, A., Lenhard, W., Suggate, S., & Segerer, R. (2018). A continuous solution to the norming problem. *Assessment*, *25*(1), 112–125.

Loyd, B. H., & Hoover, H. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*(3), 179–193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software*, *76*(1), 1–19.

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Springer.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Naumenko A., & Orel E. (2010). Who are the judges? Individual traits of test developers and test items characteristics. *Psikhologicheskie Issledovaniya [Psychological Studies], 3*(12). https://doi.org/10.54359/ps.v3i12.911 (in Russian)

Nylund-Gibson, K., Grimm, R. P., & Masyn, K. E. (2019). Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 967–985. https://doi.org/10.1080/10705511.2019.1590146

Oakland, T., Douglas, S., & Kane, H. (2016). Top ten standardized tests used internationally with children and youth by school psychologists in 64 countries: A 24-year follow-up study. *Journal of Psychoeducational Assessment*, *34*(2), 166–176. https://doi.org/10.1177/0734282915595303

Orel, E. A. (2010). Universal'nyi intellektual'nyi test Sankt-Peterburg – Chelyabinsk Modernizirovannyi, UIT SPCh-M [Universal intelligence test Saint Petersburg — Chelyabinsk, Modernized, UIT SPCh-M]. In *Ezhegodnik professional'nykh retsenzii i obzorov. Metodiki*

psikhologicheskoi diagnostiki i izmereniya [A yearbook of professional reviews. Methods of psychological diagnostics and measurement] (pp. 209–224). Chelyabinsk: Izdatel'skii tsentr YuUrGU. https://publications.hse.ru/chapters/67367817

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2009). Innovative items for computerized testing. In *Elements of adaptive testing* (pp. 215–230). New York, NY: Springer New York.

Piton-Gonçalves, J., & Aluísio, S. M. (2012). An architecture for multidimensional computer adaptive test with educational purposes. In *Proceedings of the 18th Brazilian Symposium on Multimedia and the Web* (pp. 17–24). New York, NY: Association for Computing Mashinery. https://doi.org/10.1145/2382636.2382644

Poddiakov, A. (2007). Testirovanie intellekta, konkurentsiya i refleksiya [Testing for intelligence, competition and reflexion]. *Refleksivnye Protsessy i Upravlenie, 2*(7), 46–56.

Raven, J., Court, J., & Raven, J. (2002). *Rukovodstvo k Progressivnym Matritsam Ravena i Slovarnym Shkalam* [Instructions for the Raven Progressive Matrices and verbal scales]. Moscow: Kogito-Tsentr. (Original work published 1998)

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361–372.

Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*, 321–335.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning and Assessment*, *4*(6). https://ejournals.bc.edu/index.php/jtla/article/view/1653

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53–61.

Shmelev, A. G. (2004). A test as a weapon. *Psychology. Journal of Higher School of Economics, 1*(2), 40–53. (in Russian)

Sorokova, M. G., & Yurkevich, V. S. (2014). Standartizatsiya testa "SPM Plyus Ravena" na moskovskoi vyborke [Standartization of the test "SPM Plus Raven" on a Moscow sample]. *Defektologiya, 6*, 28–37.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology, 15*(2), 201–293.

Thompson, S., Thurlow, M., & Malouf, D. B. (2004). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology*, *6*(1), 1–15.

Ushakov, D. V. (2004). Intelligence tests, or the bitter taste of self-knowledge. *Psychology. Journal of Higher School of Economics*, *1*(2), 76–93. (in Russian)

Ushakov, D. (2011). *Psikhologiya intellekta i odarennosti* [The psychology of intelligence and giftedness]. Moscow: Institute of Psychology of the RAS.

Ushakov, D., & Grigoriev, A. (2016). Macropsychology of intelligence: Through emotions to theoretical depth. *Psychology. Journal of Higher School of Economics, 13*(4), 629–635.

Van der Linden, W. J. (2016). *Handbook of item response theory: Volume 1. Models*. CRC Press.

Vie, J. J., Popineau, F., Bruillard, É., & Bourda, Y. (2018). Automated test assembly for handling learner cold-start in large-scale assessments. *International Journal of Artificial Intelligence in Education*, *28*, 616–631. https://doi.org/10.1007/s40593-017-0163-y

Von Davier, A. A., & Halpin, P. F. (2013). Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations. *ETS Research Report Series*, *2013*(2), i–36.

Wang, C., Chang, H. H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *37*(2), 99–122.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*(1), 116–136.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children-Revised (WISC-R)*. San Antonio, TX: The Psychological Corporation.

Wilson, M., & Gochyyev, P. (2020). Having your cake and eating it too: Multiple dimensions and a composite. *Measurement*, *151*, 107–247. https://doi.org/10.1016/j.measurement.2019.107247

Wilson, M., Zheng, X., & McGuire, L. (2012). Formulating latent growth using an explanatory item response model approach. *Journal of Applied Measurement*, *13*(1), 1–22.

World Health Organization. (2019). *International Classification of Diseases 11th Revision*. https://icd.who.int/en

Yasyukova, L. (2002). *Test struktury intellekta R. Amtkhauera (IST): Metodicheskoe rukovodstvo* [The Amthauer Intelligence Structure Test (IST): Methodical instructions]. Saint Petersburg: IMATON.

Zachary, R. A., & Gorsuch, R. L. (1985). Continuous norming: Implications for the WAIS-R. *Journal of Clinical Psychology*, *41*(1), 86–94.