Федеральное государственное бюджетное
образовательное учреждение высшего образования
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ
ПСИХОЛОГО-ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ»

М.Г. Сорокова, Е.Ю. Карданова,
Н.П. Радчикова, В.В. Федоров

РУКОВОДСТВО ПО СТАНДАРТИЗАЦИИ ПСИХОДИАГНОСТИЧЕСКОГО ИНСТРУМЕНТАРИЯ: ТРЕБОВАНИЯ И ОЦЕНКА КАЧЕСТВА

Учебное пособие

Москва
2024

**Федеральное государственное бюджетное
образовательное учреждение высшего образования
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ
ПСИХОЛОГО-ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ»**

**М.Г. Сорокова, Е.Ю. Карданова,
Н.П. Радчикова, В.В. Федоров**

Руководство по стандартизации психодиагностического инструментария: требования и оценка качества

Учебное пособие

**Москва
2024**

УДК 159
ББК 88.9
Р84

Рецензенты

Орел Е.А., к.психол.н., заведующая проектно-учебной лабораторией моделирования и оценивания компетенций в высшем образовании Центра психометрики и измерений в образовании, Институт образования, Национальный исследовательский университет «Высшая школа экономики» (ФГБОУ ВО НИУ ВШЭ).

Одинцова М.А., к.психол.н., доцент, заведующая кафедрой Психологии и педагогики дистанционного обучения, факультет Дистанционного обучения, Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ).

Авторы и разработчики

Сорокова М.Г., д.пед.н., к.физ.-мат.н., доцент, руководитель Научно-практического центра по комплексному сопровождению психологических исследований PsyDATA, заведующая кафедрой «Цифровое образование», ФГБОУ ВО МГППУ.

Карданова Е.Ю., к.физ.-мат.н., доцент, научный руководитель Центра психометрики и измерений в образовании, Институт образования НИУ ВШЭ, ординарный профессор НИУ ВШЭ, научный руководитель магистерской программы «Обучение и оценивание как наука», Институт образования, ФГБОУ ВО НИУ ВШЭ

Радчикова Н.П., к.психол.н., доцент, ведущий научный сотрудник Научно-практического центра по комплексному сопровождению психологических исследований PsyDATA, ФГБОУ ВО МГППУ.

Федоров В.В., старший преподаватель кафедры «Социальная психология развития» факультета «Социальная психология», старший научный сотрудник Научно-практического центра по комплексному сопровождению психологических исследований PsyDATA, ФГБОУ ВО МГППУ.

Сорокова М.Г., Карданова Е.Ю., Радчикова Н.П., Федоров В.В.

Р84 Руководство по стандартизации психодиагностического инструментария: требования и оценка качества : учебное пособие / М.Г. Сорокова, Е.Ю. Карданова, Н.П. Радчикова, В.В. Федоров ; под ред. Сороковой М.Г. — М. : ФГБОУ ВО МГППУ, 2024. — 48 с.

ISBN 978-5-94051-331-5

Рекомендовано Ученым советом Московского государственного психолого-педагогического университета для студентов, обучающихся по направлениям и специальностям подготовки УГСН 37.00.00 «Психологические науки», УГСН 44.00.00 «Образование и педагогические науки»

УДК 159
ББК 88

ISBN 978-5-94051-331-5

Содержание

1. Введение

1.1. Актуальность и цели разработки Руководства

В настоящее время в связи с укреплением позиций доказательного подхода в психологии и образовании все большую роль приобретает построение практической деятельности специалистов в этих областях на основе данных научных исследований. Пользование валидным и надежным диагностическим инструментарием является необходимой составляющей доказательного подхода и актуально для специалистов-исследователей и практиков в любой сфере психологии и образования.

Вместе с тем, в психологии ощущается явный дефицит психодиагностического инструментария, удовлетворяющего международным требованиям. Исследователи и практики зачастую пользуются методиками, которые вообще не проходили процедуры стандартизации или же проходили их достаточно давно и используются скорее «по традиции», чем на основе принципов доказательности. В ряде случаев используемые методики поступают из источников, надежность которых не подтверждена. В дальнейшем это обстоятельство приводит к снижению качества профессиональной деятельности педагогов-психологов образовательных организаций.

Первая попытка систематизировать сферу психодиагностических методик и дать пользователям инструмент, с помощью которого они могли бы подбирать качественные методики под свои задачи, была предпринята в начале 2010-х годов. Тогда, по инициативе Н.А. Батурина из Южно-Уральского Федерального Университета, запустился Ежегодник Тестовых Рецензий и Обзоров, в котором публиковались рецензии на психодиагностические методики. К сожалению, свет увидели всего два выпуска: в 2010 и в 2013 году. Однако это была важная попытка самоорганизации профессионального психологического сообщества: разработчикам методик предлагалось добровольно присылать свои инструменты и технические отчеты к ним на рецензию, и каждый инструмент рецензировался двумя экспертами. Это не было двойное слепое рецензирование, как в научных журналах (после публикации рецензий, естественно, раскрывались и имена авторов, и имена рецензентов), однако это давало возможность узнавать, кто какими темами занимается, и выстраивать профессиональные связи. Ежегодник основывался на опыте Британской и Американской Психологических Ассоциаций, где подобная практика существует уже десятки лет. К сожалению, после того, как Н.А. Батурин отошел от активного руководства проектом, выпуск Ежегодника прекратился, однако это был важный опыт, предваряющий создание «Открытого реестра психодиагностических методик, вызывающих доверие профессионального сообщества».

Вторая важная инициатива, которая предваряет создание данного учебного пособия — это Российский стандарт тестирования персонала,

выпущенный в 2015 году коллективом авторов под руководством А.Г. Шмелева [Российский стандарт тестирования персонала / 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. // Организационная психология. 2015. Т. 5. № 2. С. 67–138. URL: https://orgpsyjournal.hse.ru/2015–5–2/152057297.html (дата обращения 13.02.2022)]. Стандарт описывает принципы разработки и администрирования тестирования в сфере оценки персонала и опирается на лучшие мировые образцы в этой области. В авторском коллективе удалось собрать всех ключевых персоналий, заинтересованных в продвижении передового опыта в сфере психологического тестирования. Стандарт описывает, какие процедуры должны быть проведены и описаны в руководстве к инструментам, применяемым в оценке персонала. Таким образом, стандарт предваряет данное учебное пособие в сфере психологического тестирования взрослых.

В 2020 в рамках Госзадания Министерства просвещения РФ коллектив авторов МГППУ подготовил документ «Система функционирования психологических служб в общеобразовательных организациях. Методические рекомендации» [Система функционирования психологических служб, 2020]. Документ рекомендован Министерством просвещения РФ и размещен на сайте Общероссийской общественной организации «Федерация психологов образования России» https://rospsy.ru/node/759. В структуру этого документа входит «Открытый реестр психодиагностических методик, вызывающих доверие профессионального сообщества». В этот реестр включены стандартизованные методики (основные), но также широко применяемые методики, о валидности и надежности которых нет достаточных сведений (условно рекомендуемые). Необходимо продолжать совершенствовать реестр за счет стандартизации условно рекомендуемых методик, а также включения в него нового стандартизованного психодиагностического инструментария. Методология и методика проведения стандартизации должны соответствовать международным и российским стандартам.

Цели Руководства:

- Предоставить разработчикам психодиагностических инструментов (методик, тестов, шкал, опросников) описание основных процедур для создания надежного и валидного психодиагностического инструментария в соответствии с классической теорией тестирования (КТТ). В качестве дополнения могут быть использованы методы современной теории тестирования (IRT).
- Определить параметры для оценки качества психодиагностических инструментов (методик, тестов, шкал, опросников) экспертным сообществом на основе статьи в рецензируемом журнале с результатами стандартизации. Такая оценка может стать основанием для размещения психодиагностической методики в составе диагностического модуля

цифровой платформы психологической службы системы образования или профильных университетов в рамках создания цифровой образовательной среды (ЦОС).

- Предоставить пользователям психодиагностического инструментария — исследователям и психологам-практикам организаций и учреждений системы образования — ориентиры доказательности валидности и надежности применяемых инструментов. Информация о психометрических характеристиках диагностических методик может способствовать принятию обоснованных решений об их применении, повышению квалификации пользователей и росту культуры их применения в целом.
- Способствовать дальнейшему развитию и расширению «Открытого реестра психодиагностических методик, вызывающих доверие профессионального сообщества».

1.2. Стандартизация психодиагностического инструментария как требование доказательного подхода в психологии и образовании

Доказательный подход первоначально происходит из области медицинских исследований, а в психологии и образовании он получил широкое распространение и международное признание с 2010-х годов. Его актуальность обусловлена двумя основными идеями. Решения о вмешательствах и реформах в социальной сфере должны приниматься на основе результатов научных исследований, а не методом проб и ошибок, т.к. она затрагивает интересы миллионов людей. Научные исследования должны быть основой практической деятельности специалистов в области психологии и образования.

Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ) и общероссийская общественная организация «Федерация психологов образования России» (ФПОР) содействуют реализации доказательного подхода в психологии и образовании с целью повышения эффективности мер воздействия на образование и социальную сферу и совершенствования социальной политики. **Основные положения доказательного подхода** включают:

- Теоретическое обоснование программы, технологии, методики. Авторы и разработчики должны сформулировать, какие представления отечественной психологической науки (культурно-историческая и деятельностная научная школа — Л.С. Выготский, А.Н. Леонтьев, П.Я. Гальперин, А.В. Запорожец и др.) или теорий и подходов других научных школ, получивших международное признание, в том числе зарубежных, лежат в основе их программы, технологии, методики. Важно также сослаться на результаты отечественных и международных исследований, современные публикации в рейтинговых научных журналах России и других стран по данной тематике.

- Корректный дизайн исследования и использование стандартизованных инструментов психолого-педагогических измерений. Если измерения и оценка психических функций и процессов, личностных или характерологических особенностей, метапредметных компетенций и др. производились без таких инструментов, то уже на этапе сбора эмпирических данных к ним возникают вопросы, а анализ таких данных имеет гораздо меньше смысла.
- Заслуживающие доверия процедуры сбора эмпирических данных. В частности, в процессе сбора эмпирических данных не должно быть тенденциозности, искажающей результаты исследования. Исследования должны проводиться на больших выборках, желательно объемом до нескольких тысяч.
- Анализ данных должен проводиться с использованием релевантных количественных и качественных методов. Необходимо применение стандартных статистических пакетов для анализа данных научного исследования (SPSS, Statistica, R, KNIME Analytics Platform и др.).
- Результаты количественного и качественного анализа данных должны быть правильно интерпретированы. Необходимо обеспечить воспроизводимость результатов исследований. Базы данных исследований должны быть общедоступны в репозиториях научных данных.
- Результаты исследований должны публиковаться в рецензируемых научных журналах, обсуждаться на научных вебинарах, конференциях. Так они становятся предметом дискуссий широкого профессионального сообщества и подвергаются дополнительной проверке со стороны профильных специалистов.

Более подробно о доказательном подходе в психологии и образовании, экспертизе и верификации программ, технологий, практик в парадигме доказательного подхода см. в [Доказательный подход: Руководство по верификации программ, технологий, практик в образовании и социальной сфере : учебное пособие / М.Г. Сорокова, О.А. Ульянина, Г.В. Семья, О.И. Леонова, Д.В. Лубовский, Е.И. Исаев, Т.Г. Подушкина, Н.П. Бусыгина, Н.П. Радчикова, А.А. Шведовская ; под ред. Марголиса А.А., Сороковой М.Г., Семья Г.В. — М. : Издательство ФГБОУ ВО МГППУ, 2024.]

1.3. Открытый реестр психодиагностических методик

В «Отрытом реестре психодиагностических методик, вызывающих доверие профессионального сообщества» представлен психодиагностический инструментарий, рекомендованный для использования психологами образовательных организаций. Это составная часть документа **«Система функционирования психологических служб в общеобразовательных организациях. Методические рекомендации»** [Система функционирования психологических служб, 2020]**.** Документ подготовлен рабочей группой в составе: Рубцов В.В. (руководитель), Сергоманов П.А.

(соруководитель), Леонова О.И. (ответственный секретарь), Абушкин Б.М., Алехина С.В., Банников С.Г., Вихристюк О.В., Гаязова Л.А., Делибалт В.В., Драганова О.А., Дубровина И.В., Егоренко Т.А., Егорова М.А., Забродин Ю.М., Зарецкий В.К., Исаев Е.И., Клюева Т.Н., Лавриненко О.А., Лобанова А.В., Марголис А.А., Пахальян В.Э., Романова Е.С., Ослон В.Н., Сафронова М.А., Семья Г.В., Сорокова М.Г., Чиркина Р.В., Шарабарина О.Д., Шведовская А.А., Шумакова Н.Б.

Структура Реестра включает следующие позиции: Порядковый номер методики в Реестре; Автор методики, название методики; Измеряемый конструкт; Возрастная группа; Параметры стандартизации, доказательность: перечислены результаты стандартизации и проверенные параметры; Источник: даны ссылки на статьи в рецензируемых научных журналах, где подробно описаны все процедуры стандартизации и проверенные параметры; Год, наличие компьютерной версии методики; Статус: основная (т.е. методика, прошедшая процедуры стандартизации и валидизации) или условно рекомендуемая (т.е. диагностические методики, в настоящее время широко используемые и/или вызывающие доверие профессионального сообщества, но пока не обладающие всеми необходимыми сведениями их доказательности); Целевая группа в соответствии с классификацией, представленной в документе «Система функционирования психологических служб в общеобразовательных организациях. Методические рекомендации».

В этом документе на основе анализа нормативных правовых актов, а также научно-методических документов в системе образования и социальной защиты населения представлена следующая **классификация целевых групп детей**, в отношении которых в общеобразовательных организациях реализуются программы адресной психологической помощи: Норма (нормотипичные дети и подростки с нормативным кризисом взросления); Дети, испытывающие трудности в обучении; Уязвимые категории детей. В Уязвимых категориях детей выделены 2 подкатегории: Дети, находящиеся в трудной жизненной ситуации и Одаренные дети. К числу Детей, находящиеся в трудной жизненной ситуации, относятся 3 группы: Дети-сироты и дети, оставшиеся без попечения родителей; Обучающиеся с ОВЗ, дети-инвалиды; Дети с отклоняющимся поведением (девиантное поведение детей и подростков, суицидальное поведение детей и подростков). Взрослые представляют отдельную целевую группу.

Диагностические методики, представленные в Реестре, сгруппированы в соответствии со следующими **разделами**: Развитие основных психических функций; Метапредметные компетенции и универсальные учебные действия; Социальное развитие и морально-ценностная сфера; Эмоционально-волевая сфера; Коммуникативная сфера; Поведенческая сфера; Профессиональная направленность, мотивация, характерологические особенности; Диагностика профессиональных и личностных проблем взрослых участников образовательного процесса. В последнем разделе представлен психодиагностический инструментарий для работы с персоналом образовательных

организаций, учреждений социальной защиты населения, организаций для детей-сирот и детей, оставшихся без попечения родителей, и родителями.

1.4. Этапы разработки и стандартизации психодиагностической методики (инструмента)

Разработка и стандартизация психодиагностической методики (инструмента, теста) включает следующие основные этапы [Анастази, Урбина, 2007; Бурлачук, 2006; Клайн, 1994; Митина, 2011; Шмелев, 2013].

1) Анализ научной литературы, теоретическое обоснование и содержательное описание исследуемого конструкта в рамках выбранной теории, концептуального подхода или с ориентацией на гипотезы, сформулированные на основе теоретических предположений об измеряемом свойстве.
2) Разработка заданий методики (вопросов теста, пунктов опросника и др.).
3) Подбор стандартизованных инструментов для проверки разных типов валидности разрабатываемой методики.
4) Формирование выборки стандартизации, репрезентативной относительно обследуемой категории испытуемых, для которых предназначен инструмент, и сбор эмпирических данных и создание базы данных.
5) Количественный анализ эмпирических данных в стандартных статистических пакетах (таких как Excel, SPSS, KNIME Analytics Platform, RStudio, Winsteps и др.) с целью проверки психометрических характеристик инструмента (описательная статистика для всех измеряемых параметров, проверка нормальности распределений, проверка надежности и валидности методики, расчет нормативных значений показателей, оценка ошибки измерения и расчет доверительных интервалов и др.). Интерпретация результатов.
6) В случае необходимости задания (вопросы, пункты опросника) могут быть скорректированы (заменены на новые, удалены и т.д.), что весьма типично для процесса разработки и валидизации диагностической методики. Тогда предшествующий этап рассматривается как ее апробация, формируется новая стандартизационная выборка, как правило, большего объема, и процесс расчета психометрических характеристик выполняется снова.
7) Если расчеты показывают хорошее соответствие эмпирических данных измеряемым теоретическим конструктам (достаточную надежность, валидность, ясную факторную структуру, интерпретируемость и высокий процент объясняемой общей дисперсии в модели эксплораторного факторного анализа (ЭФА), хорошие и удовлетворительные показатели соответствия в модели конфирматорного факторного анализа (КФА), индексы соответствия и параметры модели IRT и др.), авторы разрабатывают описание инструмента и подробную инструкцию для пользователей.

8) По результатам стандартизации пишется статья в рецензируемый научный журнал. Методика считается стандартизованной, если результаты ее стандартизации опубликованы в научном журнале, где представлены перечисленные выше параметры стандартизации, а также описание инструмента с инструкцией для пользователей.
9) Стандартизованная методика может быть выложена на институциональную цифровую платформу для использования исследователями научных и образовательных организаций всех видов, в том числе подведомственных Министерству просвещения РФ, Министерству образования и науки РФ, а также психологами-практиками в системе образования и социальной защиты населения и т.д. Базы данных могут быть представлены в репозиториях научных данных открытого доступа.

Примеры представления результатов разработки и стандартизации психодиагностического инструментария в научной периодике см. в списке литературы в разделе «**Статьи по стандартизации и валидизации психодиагностических инструментов**».

Адаптация зарубежного инструментария. Русскоязычные версии зарубежных стандартизованных методик также должны повторно проходить процедуры валидизации в новых социо-культурных условиях на русскоязычных выборках: только в этом случае они могут применяться отечественными специалистами — исследователями и практиками. Описанным выше процедурам валидизации предшествует двойной перевод методики с языка оригинала на русский язык и обратно с целью максимально точной передачи содержания пунктов оригинального опросника отечественной аудитории.

1.5. Формат представления психодиагностической методики

1.5.1. Назначение и краткое описание инструмента

Описание содержит общую характеристику инструмента, чтобы дать пользователю четкое представление о его назначении, сфере применения и характерных особенностях. В описании обычно приводятся:

- авторы и название инструмента и дополнительная информация о его разработке (является ли он авторским или адаптацией зарубежного инструмента на российской выборке; возможно, это авторская разработка на основе уже известных российских или зарубежных опросников или их модификация; новая версия инструмента и др.);
- измеряемые конструкты (параметры, признаки, психолого-педагогические характеристики), для которых дается краткая характеристика;
- возрастные и другие специфические особенности группы, для которой предполагается использование инструмента (в описании инструментов для обследования обучающихся учреждений среднего, среднего профессионального и среднего специального образования бывает полезно привести также рекомендуемый диапазон классов школы или колледжа);

- источник: ссылки на статьи в рецензируемых научных журналах, где подробно описаны все процедуры стандартизации и приведены результаты вычисления соответствующих психометрических характеристик, а также официальные печатные издания и сборники, где опубликована методика;
- наличие компьютерной версии инструмента (есть или нет);
- целевая группа в соответствии с классификацией, представленной в документе «Система функционирования психологических служб в общеобразовательных организациях. Методические рекомендации», см. п. 1.3. настоящего Руководства;
- сфера применения или отнесение к группе диагностических методик (например, в соответствии с «Открытым реестром психодиагностических методик», т.е. «развитие основных психических функций», «метапредметные компетенции и универсальные учебные действия» и др., см. п. 1.3. настоящего Руководства);
- примерное время обследования;
- дополнительная информация, важная с точки зрения авторов и разработчиков инструмента (исторический контекст, сведения об особенностях применения и др.).

1.5.2. Процедура проведения тестирования: инструкции, основные требования

В этом разделе дается подробное описание материала методики, например, количество и размер изображений, их расположение на бланке или экране компьютера, особенности их компоновки и др.

Приводится подробная инструкция для пользователей — исследователей и психологов образовательных организаций всех видов, учреждений системы социальной защиты населения и др. — по проведению обследования. Указывается, является ли процедура обследования индивидуальной или групповой (указать примерные размеры группы); нужны ли «тренировочные задания»; в каких выражениях рекомендуется давать инструкцию испытуемым; на что следует обратить внимание и что следует проверить в процессе и по окончании тестирования; ограничения по времени (если есть).

Указывается способ оценки результатов (scoring), т.е. порядок начисления тестовых баллов как для отдельных субшкал, так и для итоговой шкалы методики (если это предусмотрено). Для этого приводится ключ к инструменту для расчета «сырых баллов». Указывается также способ перевода «сырых баллов» в стандартные баллы в соответствии со шкалами станайнов, стенов, Т-значений, z-значений и др., широко используемыми в современных международных исследованиях для сопоставления показателей испытуемого по разным психодиагностическим методикам (тестам, опросникам, инструментам), а также по субшкалам одного и того же инструмента. Приводится интерпретация результатов, которая может быть полезна и служить ориентиром психологу при обследовании конкретного испытуемого.

1.5.3. Описание выборки стандартизации

Сначала производится сбор эмпирических данных стандартизационной выборки испытуемых соответствующих категорий. Если методика разрабатывалась для подростков 13–15 лет, то испытуемые должны относиться именно к этой категории, а методика в дальнейшем должна применяться также для этой категории подростков, а не, например, для взрослых или младших школьников.

Выборка стандартизации должна быть достаточно большого **объема**. По рекомендации Европейской федерации психологических ассоциаций (EFPA) https://www.efpa.eu/, для решений с низкими и высокими ставками, при принятии которых хотя бы частично учитываются результаты тестирования, даны разные ориентировочные показатели [EFPA, 2013, P. 35]. Для высоких ставок при принятии нетривиальных решений с привлечением результатов тестирования объемы выборок стандартизации больше. Для низких ставок хорошим считается объем выборок от 300 испытуемых, для высоких — от 400, объем стандартизационной выборки 1000 и более испытуемых считается отличным в обоих случаях. Ориентиром могут быть также исследования по стандартизации методик, опубликованные в рейтинговых зарубежных научных журналах: нередки объемы выборок порядка 1000 испытуемых и выше.

Процедуры отбора выборки стандартизации. Согласно терминологии социологических и маркетинговых исследований и опросов, выборки делятся на два типа: вероятностные и невероятностные.

Вероятностные выборки могут формироваться следующим образом:

- Случайная выборка (простой случайный отбор). Такая выборка предполагает однородность генеральной совокупности, одинаковую вероятность доступности всех элементов, наличие полного списка всех элементов. При отборе элементов может быть использован генератор случайных чисел.
- Механическая (систематическая) выборка. Разновидность случайной выборки, упорядоченная по какому-либо признаку (алфавитный порядок, номер телефона, дата рождения и т.д.). Первый элемент отбирается случайно, затем, с шагом n отбирается каждый k-ый элемент. Размер выборки при этом — N=n*k.
- Стратифицированная (районированная). Применяется в случае неоднородности генеральной совокупности. Генеральная совокупность разбивается на группы (страты), например регион проживания, возрастная группа, пол и т.д. В каждой страте отбор осуществляется случайным или механическим образом.
- Серийная (гнездовая или кластерная) выборка. При серийной выборке единицами отбора выступают не сами объекты, а группы (кластеры или гнёзда). Группы отбираются случайным образом. Внутри групп обследуются все объекты.

Невероятностные выборки. Отбор в такой выборке осуществляется не по принципам случайности, а по субъективным критериям — доступности, типичности, равного представительства и т.д.

- Квотная выборка. Изначально выделяется некоторое количество групп объектов (например, школьники-подростки 13–15 лет, проживающие в городах с населением более 1 миллиона, от 100 тысяч до 999 тысяч, менее 100 тысяч; дети-сироты и дети, оставшиеся без попечения родителей; обучающиеся с ОВЗ и дети-инвалиды различных нозологических групп; учителя в возрасте 20–30 лет, 31–45 лет и 46–60 лет и др.). Для каждой группы задается количество объектов, которые должны быть обследованы. Количество объектов, которые должны попасть в каждую из групп, задается, чаще всего, либо пропорционально заранее известной доле группы в генеральной совокупности, либо одинаковым для каждой группы. Внутри групп объекты отбираются произвольно.
- Метод снежного кома. Выборка строится следующим образом. У каждого респондента, начиная с первого, просят контакты его друзей, коллег, знакомых, которые подходили бы под условия отбора и могли бы принять участие в исследовании. Таким образом, за исключением первого шага, выборка формируется с участием самих объектов исследования. Метод часто применяется, когда необходимо найти и опросить труднодоступные группы респондентов (например, родителей детей с ОВЗ или инвалидностью конкретной нозологической группы; респондентов, принадлежащих к одной профессиональной группе; респондентов, имеющих какие-либо схожие хобби/увлечения и т.д.).
- Стихийная выборка. Опрашиваются наиболее доступные респонденты. Типичные примеры стихийных выборок — опросы в газетах, журналах, соцсетях; анкеты, отданные респондентам на самозаполнение; большинство интернет-опросов. Размер и состав стихийных выборок заранее не известен и определяется только одним параметром — активностью респондентов.
- Выборка типичных случаев. Отбираются единицы генеральной совокупности, обладающие средним (типичным) значением признака. При этом возникает проблема выбора признака и определения его типичного значения.

Стандартизационная выборка должна быть **репрезентативной** для целевой и возрастной группы, однако вопрос о репрезентативности конкретной выборки не всегда имеет однозначный ответ и может быть полемически заостренным. В некоторых источниках выборку рекомендуется считать репрезентативной, либо если эта выборка вероятностная, либо если выборка квотная, то есть если состав выборки по ряду переменных (например, возраст, пол, образование, регион проживания, тип поселения и др.) аналогичен составу целевой и возрастной группы. При этом в случае валидизации опросников, применимых для более узких целевых групп, например,

для детей с ОВЗ и инвалидностью или для выявления подростков с риском суицидального поведения, в выборку стандартизации должны быть включены нормотипичные дети и подростки из соответствующей популяции.

1.5.4. Психометрические характеристики теста

Психометрические характеристики теста должны быть приведены в соответствие с классической теорией тестов тестирования (см. п. 2) и включать показатели заданий теста (трудность и дискриминативность), а также показатели качества всего теста — его надежность и валидность [Анастази, Урбина, 2007; Бурлачук, 2006; Клайн, 1994; Митина, 2011; Шмелев, 2013]. В некоторых случаях можно заменить и/или дополнить эти расчеты психометрическими характеристиками современной теории тестов тестирования (IRT): характеристические кривые пунктов методики (item characteristic curve, ICC), информационные кривые отдельных пунктов и всего теста (item и test information curve) (см. п. 3), а также карта переменных всего теста и т.д..

1.5.5. Описательная статистика для всех измеряемых показателей, проверка нормальности распределений

Для всех субшкал и итоговой шкалы опросника приводятся следующие показатели описательной статистики: среднее (M), медиана (Me), стандартная ошибка среднего, среднеквадратическое (стандартное) отклонение (SD), асимметрия, стандартная ошибка асимметрии, эксцесс, стандартная ошибка эксцесса. Проверка нормальности распределения проводится с помощью критерия Колмогорова — Смирнова, критерия Шапиро — Уилка или других аналогичных критериев. Полезно также построить гистограммы полученных распределений, чтобы оценить форму распределения визуально.

Стоит отметить, что значение критерия Колмогорова-Смирнова (как и показатели других критериев проверки формы распределения) зависит от размера выборки и дает статистически значимые различия при минимальных отклонениях от нормальности, если выборка велика, поэтому обычно рекомендуется рассматривать и другие признаки нормальности. В пользу симметричности распределения данных говорит совпадение средних значений и медиан, а также то, что значения асимметрии лежат в пределах от –0,5 до 0,5. Более радикальные авторы допускают даже асимметрию в пределах от –2 до 2 [Structural equation modeling, 1995, pp. 56–75]. Некоторые авторы предлагают рассматривать отношения асимметрии и эксцесса к их стандартным ошибкам как показатель согласия с нормальным распределением или его отсутствия. Данное отношение интерпретируется как свидетельство отклонения от нормальности, если оно более 1 [Наследов, 2004, С. 60] или более 2 [Weinberg & Abramowitz, 2002, P. 79], что обычно верно для небольших выборок, используемых в психологических исследованиях. В случае выборок среднего размера (от 50 до 300 человек) предельное соответствующее значение отношения

равно 3,29. В специальной литературе встречаются также и другие рекомендации. В любом случае, соответствие формы распределения нормальному должно быть обосновано.

1.5.6. Стандарты и нормативные шкалы

В психодиагностике широкое применение получили стандартные показатели, которые рассчитываются на основе линейного или нелинейного преобразования первичных показателей («сырых баллов») стандартизационной выборки, если ее распределение согласуется с нормальным [Бурлачук, Морозов, 2002]. К числу таких показателей относятся z-значения, Т-значения, стандартный IQ и некоторые другие. Например, если распределения показателей по двум разным тестам оба согласуются с нормальным, и это подтверждается соответствующим статистическим критерием, то с помощью z-преобразования они легко могут быть сопоставлены друг с другом. В случае нормального распределения выборки стандартизации интервал $[M - \sigma; M + \sigma]$ определяет границы нормы исследуемого признака.

Шкалы стенов и станайнов являются примерами нелинейных преобразований «сырых баллов» в стандартную шкалу. Перевод первичных тестовых баллов в эти шкалы производится с помощью процентилей. Стандартные шкалы чрезвычайно важны для обеспечения сопоставимости результатов, полученных по разным тестам или субшкалам одного и того же теста, имеющим разный диапазон значений. Это дает возможность их совместной интерпретации, сведения оценок к единой системе. Они позволяют также определить место индивидуального результата испытуемого в распределении групповых данных выборки стандартизации и таким образом облегчить их интерпретацию.

2. Классическая теория тестирования (КТТ) как теоретико-методологическая основа стандартизации психодиагностического инструментария

2.1. Историческая справка

История развития тестологии насчитывает более 100 лет. Появление первых тестов индивидуальных различий в сфере сенсомоторных, интеллектуальных функций и перцептивных процессов относится к концу 19 века и связано с работами Ф. Гальтона, Дж. Кеттела, Дж. Ястрова. В этот же период Э. Крепелин создал большую серию тестов для измерения памяти, утомляемости, отвлекаемости, а его ученик А.Эрн применил тесты восприятия, памяти, ассоциаций и моторики для изучения взаимосвязи психических функций. В начале 20-го века широкую известность приобрели серии интеллектуальных тестов А. Бине, многократно переработанных и модифицированных. Одним из самых удачных вариантов таких тестов стала шкала Станфорд — Бине. Тестирование приобрело массовый характер в США в Первую мировую войну в связи с потребностями армии быстро определить интеллектуальный уровень более миллиона новобранцев, а вскоре после ее окончания их начали систематически использовать в школах и при поступлении в колледж. Со 2-й половины 1940-х широкое распространение получили комплексные батареи способностей, предназначенные для получения индивидуальных профилей, а с начала 1950-х — тесты для измерения личностных характеристик [Анастази, Урбина, 2007].

Первые стандартизованные тесты для оценки результатов школьного обучения появились на рубеже 20-го века. Так, Г. Эббингауз предъявлял учащимся тесты арифметического счета, объема памяти и завершения предложений, а в 1920-е годы появились батареи достижений, позволяющие сравнить выполнение заданий по разным школьным предметам относительно единой системы нормативов. Развитию тестирования в США и Европе способствовало также создание региональных и национальных программ оценки образовательных достижений. В нашей стране в современный период начало широкому использованию тестов академических достижений в школах положило введение ЕГЭ в середине 2000-х.

Рука об руку с развитием тестологии шло создание соответствующего математического обеспечения для стандартизации тестов. Снова упомянем Ф. Гальтона и его ученика К. Пирсона, внесших значительный вклад в разработку методов математической статистики для анализа данных по индивидуальным различиям. Исследованием взаимосвязей различных психологических характеристик занимался Ч. Спирмен, а дальнейшее развитие

его идей привело к созданию такого мощного статистического метода многомерного анализа данных, как факторный анализ, в частности, в работах Т.Л. Келли и Л.Л. Терстоуна. Начиная с 1950-х годов одновременно с разработкой математического обоснования факторного анализа этот метод становится общенаучным. В нашей стране с начала 1990-х с распространением персональных компьютеров и появлением общедоступных профессиональных статистических пакетов факторный анализ входит в основной инструментарий психологии и других социальных наук [Анастази, Урбина, 2007; Наследов, 2004].

Если становление психодиагностики как науки, разработка и применение тестов в исследовательских и практических целях в США и западной Европе хотя и переживало периоды кризисов, острых дискуссий и критики этого направления, но происходило непрерывно, то история развития этой науки в России гораздо более драматична [Бурлачук, 2006]. В дореволюционной России начала XX в. тесты были хорошо известны и популярны, а сторонники естественно-научного направления в развитии психологии — А.П. Нечаев, Н.Е. Румянцев, Г.И. Россолимо, Ф.Е. Рыбаков — способствовали распространению идей тестирования и разрабатывали методики измерения общих способностей, пространственного воображения и др. Вместе с тем высказывались мнения о границах применения тестов. Так, Г.И. Челпанов подчеркивал, что психологические тесты могут применяться исключительно только для научных исследований, но не для практических целей. В советский период 1920–1930-х гг. практика тестирования получает широкое распространение в педологии и психотехнике. Тесты активно используются прежде всего в учебно-воспитательных учреждениях. Плодотворно работают в области психологического тестирования М.Я. Басов, М.С. Бернштейн, П.П. Блонский, Г.И. Россолимо и др. Ведутся дискуссии об инструментах измерения индивидуальных различий, недостаточном теоретическом обосновании тестов, перспективах их использования в широкой практике и др. Дискуссии, как и тестирование, были прерваны известным постановлением ЦК ВКП(б) «О педологических извращениях в системе наркомпросов» от 4 июля 1936 г.

Началом эпохи возрождения психодиагностических исследований в советской психологии принято считать 1969 год, когда Центральный совет Общества психологов СССР официально признал необходимость развития психодиагностики как науки. В предшествующий этому событию и в последующий периоды значительный вклад в становление советской психодиагностики внесла Ленинградская психологическая школа во главе с Б.Г. Ананьевым. В комплексном исследовании человека, реализованном под его руководством, использовались тесты интеллекта, личностные опросники и другие диагностические инструменты. Психологами и психиатрами Ленинградского психоневрологического института, основанного

В.М. Бехтеревым, проводились клинико-психологические исследования с использованием зарубежных психодиагностических методик.

Дискуссии о тестах периодически вспыхивали вплоть до середины 1970-х гг., но к началу 1980-х значительно сократились. Иногда тесты, объявляемые «количественным подходом» к диагностике психических явлений, противопоставлялись «качественному», который критиками признавался единственно верным. Однако с конца 1960-х — середины 1970-х появляется все больше публикаций, посвященных возможностям применения тестов в образовании, медицине, спорте и других сферах жизнедеятельности человека. Тесты все более активно используются в медико-психологических исследованиях, изучении разных возрастных групп, профориентации и профотборе, судебно-психологической экспертизе. Появляются и теоретические публикации, анализирующие состояние дел в зарубежных исследованиях (Л.Ф. Бурлачук и др.). В 1980-е гг. продолжается обсуждение общих и частных проблем психодиагностики, при этом широко привлекается собственный исследовательский опыт (Е.Т. Соколова, В.С. Аванесов, М.М. Кабанова, Б.В. Кулагина, Л.Ф. Бурлачук и др.). Делаются первые шаги на пути разработки оригинальных методик, многие из которых сегодня достаточно активно используются в психологических исследованиях (А.Е. Личко и др., 1983; В.М. Мельников и Л.Т. Ямпольский, 1985; К. Акимова и др., 1988; В.В. Столин и др., 1988; А.А. Кроник, 1991 и др.). Проводится работа по повышению квалификации психологов в области психометрики и конструирования тестов. Обращается внимание и на разработку этических норм, которыми должны руководствоваться создатели и пользователи психологических тестов.

В настоящее время в нашей стране все большее распространение тесты получают в таких сферах применения, как психологическое, образовательное и кадровое тестирование [Шмелев, 2013]. Во всех трех случаях объект тестирования один и тот же — это человек как эмпирический носитель какого-либо интересующего нас свойства или функции. Однако предметы тестирования — разные. В случае психодиагностики предметом являются психические свойства человека: константность восприятия, объем внимания, прочность памяти, черты темперамента и т.д. В случае образовательного тестирования — это знания, умения, навыки и познавательные способности — «образовательные компетенции». В кадровой тестологии предмет — это профессиональные и производственные компетенции, т.е. свойства, которые влияют на эффективность профессиональной деятельности. Эти предметы пересекаются. Действительно, хотя в учебной и профессиональной деятельности общие способности человека нередко напрямую не влияют на эффективность деятельности в конкретных ситуациях, а проявляются лишь косвенно — через конкретные умения и навыки, тем не менее более глубокая образовательная и профессиональная диагностика восходит к уровню общих способностей и личностных качеств человека.

Тестирование в каждой из трех упомянутых областей имеет свою специфику. Большинство образовательных тестов — это так называемые «предметно-ориентированные» тесты знаний. Это типичный пример ТРМ — «тестовых рейтинговых методик» — в отличие от ТДМ — «тестовых диагностических методик». Задания в этих тестах относятся к определенным традиционным академическим предметам — математике, родному или иностранному языку, истории, физике. Это тестирование в условиях образовательных учреждений организуют и проводят не школьные психологи, а сами педагоги-предметники или соответствующие структурные подразделения — учебно-производственные лаборатории и т.д. В то же время школьные психологи могут проводить психологические тесты — на умственное развитие, профориентационные тесты, учитывающие широкий комплекс факторов: интересы, мотивы, акцентуации характера и др. В профессиональном тестировании большую роль играют тесты профессиональных знаний, в которых в последнее время на место теоретических вопросов все чаще приходят кейс-задания — задачи, описывающие практические производственные ситуации и предлагающие выбор способа действия в этой ситуации. В программах тестирования персонала также часто используются психологические и психофизиологические тесты.

2.2. Понятие теста в широком и узком смыслах

В различных источниках можно встретить различные трактовки термина «тест». Отчасти это связано с областью применения диагностических методик — психология, образование или кадровый отбор персонала организаций, — задачами диагностики, целевой группой испытуемых, спецификой их возраста, особенностей развития и др. Например, в Российском стандарте тестирования персонала [Российский стандарт тестирования персонала, 2015, С. 73–74] используется понятие теста в узком смысле. Методика оценки может называться тестом, если обладает одновременно следующими признаками: стандартизированный набор вопросов или заданий иного типа; одна или несколько измерительных шкал, позволяющих выразить результаты количественно; связь каждого ответа на каждое задание с одной или несколькими измерительными шкалами (наличие «ключей к тесту»); стандартизированная процедура проведения, включающая однозначную (стандартную) инструкцию для тестируемого, правила использования вспомогательной информации, правила завершения или приостановки тестирования и т.п.; возможность автоматической (без участия человека) обработки результатов, то есть формализованная процедура подсчета баллов по шкалам с помощью весовых коэффициентов (ключей); тестовые нормы — фиксированные границы перевода тестовых баллов в оценочные категории; формализованная модель интерпретации результатов и/или рекомендации по принятию тех или иных решений, связанные

с определенными интервалами значений на шкале (шкалах) и сочетаниями значений шкал (при наличии двух шкал и более); направленность на индивидуальную количественную оценку какой-либо характеристики одного человека (а не группы, коллектива и т.п.).

В международных стандартах и рекомендациях по тестированию можно встретить более широкое понимание теста. Согласно Руководству по использованию тестов Международной комиссии по тестированию (International Test Commission) 2001 года [International Guidelines for Test Use, 2001, P. 96] тестирование может включать в себя широкий спектр процедур для психологической, профессиональной и образовательной оценки. Тестирование может содержать процедуры измерения как нормального, так и абнормального или дисфункционального поведения. Процедуры тестирования обычно разрабатываются для проведения в тщательно контролируемых или стандартизированных условиях, которые включают протоколы систематической оценки. Эти процедуры обеспечивают меры производительности и включают в себя выводы из образцов поведения. Они также включают процедуры, которые могут привести к качественной классификации или упорядочению людей (например, с точки зрения типа) [International Guidelines for Test Use, 2001, P. 96].

В стандартах Европейской федерации психологических ассоциаций (European Federation of Psychological Associations. EFPA), предназначенных для описания и оценки психологических и образовательных тестов, предложена еще более широкая трактовка этого термина, как любого «… оценочного устройства или процедуры, в которой образец поведения испытуемого в определенной области получается, а затем количественно оценивается с использованием стандартизированного процесса». Авторы документа ссылаются на «Стандарты образовательного и психологического тестирования» (1999) Американской ассоциации исследователей образования (American Educational Research Association), Американской психологической ассоциации (American Psychological Association) и Национального совета по измерению в образовании (National Council on Measurement in Education). [EFPA, 2013, P. 3]

Настоящее Руководство по стандартизации психодиагностического инструментария предназначено, прежде всего, для оценки качества тестов, опросников, диагностических методик, разработанных для их использования исследователями и психологами-практиками образовательных организаций. Специфика этой сферы деятельности состоит в работе с широким спектром возрастов детей — от дошкольного до юношеского возраста, а также со взрослыми участниками образовательного процесса — педагогами школы, социальными работниками, родителями. Помимо возрастной дифференциации, требующей соответствующей дифференциации психодиагностических методик, имеются также методики, предназначенные для разных целевых групп — нормативно развивающихся детей и подростков,

детей с отклонениями в поведении, обучающихся с ОВЗ, детей-инвалидов и др. При таком многообразии возрастных и целевых групп существует широкий спектр диагностических инструментов, отличных по структуре от опросников со шкалами Ликерта, и для их стандартизации приходится применять разные наборы процедур и методов математической статистики. Вот почему для наших целей более пригодно широкое понимание термина «тест». В данном Руководстве мы также используем термин «инструмент» как обобщающее понятие для таких терминов, как «психодиагностическая методика», «тест», «опросник». Этот термин также широко используется в зарубежной научной периодике, стандартах и практических руководствах по тестированию для пользователей.

2.3. Общая характеристика Классической теории тестирования (КТТ)

В основе КТТ лежит трактовка оценок по тесту как случайных величин, принимающих в процессе тестирования заранее не известные значения. С точки зрения теории вероятностей, однократное тестирование понимается как испытание, в результате которого получается наблюдаемое значение случайной величины, которое в ту или иную сторону отклоняется от истинной оценки испытуемого по данному тесту. Наблюдаемое значение отклоняется от истинного из-за действия многочисленных случайных факторов, таких как возможное угадывание ответа, волнение и общее самочувствие испытуемого, его способность концентрироваться, понимание инструкции, формулировка вариантов ответов и др. В результате при гипотетическом многократном повторении испытания наблюдаемые оценки будут отклоняться от истинного значения из-за ошибок измерения, которые тоже являются случайными величинами. Основное уравнение КТТ имеет вид:

$$X_i = T_i + E_i$$

где X_i — наблюдаемый результат *i-го* испытуемого выборки по данному тесту; T_i — его истинный балл; E_i — суммарная ошибка измерения при оценке T_i с помощью теста.

Ошибки измерения бывают систематические и случайные. Систематические ошибки порождаются как недостаточным качеством теста, так и неправильными условиями его проведения, а случайные ошибки — индивидуальными особенностями испытуемых и нарушением условий тестирования. Ошибки измерения снижают надежность теста и достоверность его результатов.

С теоретической точки зрения, коэффициент корреляции между истинной и наблюдаемой оценками по тесту называется **показателем надежности теста**, однако на практике это определение не используется, т.к. нам недоступны ни истинные оценки по тесту, ни вся совокупность наблюдаемых оценок.

Принципы (аксиомы, постулаты) КТТ [Крокер, Алгина, 2010]:

1) Среднее значение ошибок измерения для генеральной совокупности испытуемых равно 0.
2) Корреляция между истинной оценкой и ее ошибочным компонентом равна 0.
3) Корреляция между ошибочными компонентами оценок по двум параллельным тестам равна 0.

Отсюда следует фундаментальное соотношение КТТ: $\sigma_x^2 = \sigma_T^2 + \sigma_E^2$, т.е. дисперсия наблюдаемых баллов σ_x^2 равна сумме дисперсии истинных баллов σ_T^2 и дисперсии ошибок измерения σ_E^2. Это равенство удобно переписать в виде: $\rho = \frac{\sigma_T^2}{\sigma_x^2} = 1 - \frac{\sigma_E^2}{\sigma_x^2}$, где через ρ обозначен **коэффициент надежности теста**. Коэффициент надежности определяется как отношение дисперсии истинной оценки к дисперсии наблюдаемой оценки и интерпретируется как доля дисперсии наблюдаемых баллов, которая может быть объяснена истинной вариацией истинных оценок испытуемых. **Показатель надежности** равен квадратному корню из **коэффициента надежности** и равен коэффициенту корреляции между истинными и наблюдаемыми баллами. Пусть, например, коэффициент надежности теста получился равным 0,81. Это означает, что 81% дисперсии наблюдаемых баллов может быть объяснено дисперсией истинных оценок испытуемых. При этом показатель надежности в данном случае будет равен 0,90 (квадратный корень из 0,81), что означает, что корреляция между наблюдаемыми и истинными оценками испытуемых для данного теста равна 0,90.

В психометрике используются две группы методов оценки надежности теста по выборке: процедуры, требующие двукратного предъявления теста, и процедуры, требующие однократного предъявления теста, в том числе методы, основанные на ковариациях заданий. Коэффициент надежности зависит от свойств выборки стандартизации, поэтому потенциальным пользователям тестов рекомендуется убедиться, что сообщаемый в документах к тесту коэффициент надежности получен на выборке, подобной по составу той группе, для которой будет использоваться тест. В противном случае надежность надо переоценить заново, применительно к имеющейся группе испытуемых. Более подробно о методах оценки надежности теста будет сказано ниже.

Интерпретация индивидуальных оценок испытуемых по некоторому тесту становится более корректной при использовании доверительных интервалов для истинных значений тестовых баллов. Доверительный интервал, с вероятностью 95% покрывающий истинное значение тестового балла, строится с помощью стандартной ошибки измерения $\sigma_{\hat{A}}$ и статистики одновыборочного t-критерия Стьюдента с уровнем значимости p = 0,05: $(x_i - 1{,}96 \cdot \sigma_E,\ x_i + 1{,}96 \cdot \sigma_E)$, где x_i –тестовый балл i-го испытуемого. Стандартная ошибка измерения оценивается с помощью коэффициента надежности теста: $\sigma_E = \sigma_x \cdot \sqrt{1 - \rho}$.

Классическая теория позволяет оценить и качество заданий (пунктов, вопросов) психодиагностической методики. Для этого используются две характеристики заданий — трудность и дискриминативность. Трудность характеризует, насколько трудно испытуемым согласиться с утверждением. Если задание имеет один правильный ответ и предполагается дихотомическая оценка 0/1, то трудность задания рассчитывается как доля испытуемых, выполнивших задание верно (получивших 1 балл за выполнение задания). Если задание имеет несколько градаций ответа и предполагает политомическую оценку (как, например, в шкалах Ликерта), то трудность определяется как отношение среднего балла по заданию в данной группе участников к разности между наибольшим и наименьшим баллами за задание. Средний балл за задание вычисляется как сумма баллов по этому заданию всех испытуемых, деленная на число испытуемых. Трудность может принимать значения от 0 до 1, при этом, чем больше значение, тем легче задание. Принято задания с трудностью более 0,8 считать легкими, менее 0,2 — трудными, с трудностями в диапазоне между этими значениями — средней трудности. Рекомендуется большинство заданий в тесте делать средней трудности, однако для более точного оценивания в тесте нужны и легкие задания, и трудные.

Дискриминативность задания (другое название дифференцирующая способность задания) характеризует способность задания различать испытуемых с различным уровнем выраженности измеряемой черты. Дискриминативность может вычисляться разными способами, самым популярным является вычисление корреляции между баллами по заданию и общим баллом по тесту. Если для данного задания указанный коэффициент корреляции отрицателен или близок к нулю, это означает, что испытуемые с высоким общим баллом по тесту (т.е. с высоким уровнем выраженности измеряемой черты) имеют по данному заданию баллы, меньшие, чем испытуемые с низким баллом по тесту. Причина этого может быть, как в неправильном оценивании задания (например, ошибка в ключах), так и в самом задании — например, некорректной формулировке. Допустимыми считаются значения дискриминативности большие 0,2–0,3. Задания с низкой дискриминативностью должны быть обязательно проанализированы, изменены или удалены из теста.

Основными психометрическими свойствами тестов в классической теории являются надежность, валидность, репрезентативность и достоверность. Заметим, что трактовка этих понятий может несколько отличаться в разных источниках.

Надежность — это характеристика психодиагностической методики, отражающая точность психодиагностических измерений, а также устойчивость или стабильность результатов, полученных при помощи данной методики, к действию посторонних случайных факторов. **Валидность** теста — это устойчивость результатов к воздействию со стороны других психических свойств и компетенций, не являющихся предметом

измерения в данном сеансе тестирования [Шмелев, 2013]. Валидность отражает степень сфокусированности теста именно на заявленном свойстве, степень целевой направленности измерения. Так, например, слабое владение клавиатурой может быть причиной низкой валидности компьютерных тестов — особенно тех, которые предполагают не выбор заданных ответов, а ввод слов. **Репрезентативность** тестовой шкалы — это степень соответствия реальных тестовых норм, полученных на выборке стандартизации, тем идеальным тестовым нормам, которые были бы получены, если бы была протестирована вся популяция испытуемых, на которых планируется тест использовать. Наличие репрезентативности главным образом зависит от сходства демографических характеристик, которыми обладает стандартизационная выборка, с той социально-демографической группой, на которой применяется тест. Наконец, **достоверность** теста связана с защитой от такого универсального фактора, присутствующего во всех ситуациях тестирования, как желание показать высокий или социально одобряемый результат, что приводит к списыванию, подсказкам, подтасовкам и т.д. Итак, надежность — это точность и стабильность процедуры измерения; валидность — соответствие теста измеряемому свойству; репрезентативность — точность определения тестовых норм; достоверность — устойчивость теста к фальсификации.

2.4. Проверка надежности инструмента

Надежность теста характеризует точность измерений, свободу от ошибки измерения, устойчивость результатов теста к воздействию со стороны различных случайных факторов-помех, таких как посторонние отвлекающие звуки речи и шумы, различия в освещенности, мелькание посторонних предметов на фоновом зрительном поле, вибрация, сбои в электросети и др., приводящих к ошибке измерения и разбросу значений вокруг истинного показателя [Шмелев, 2013, С. 67–71]. Между величиной ошибки измерения и надежностью существует обратная зависимость: чем меньше интервал, в котором находится истинное значение тестового балла, тем выше точность измерения, тем выше стабильность процедуры: при каждом последующем тестировании результаты оказываются ближе к предыдущим.

Выделяют разные виды надежности:

1) **Надежность самого измерительного инструмента (теста, психодиагностической методики)** проверяется 4-мя способами: как надежность теста-ретеста (ретестовая надежность), как надежность параллельных форм теста, как надежность частей теста, и как внутренняя согласованность (согласованность пунктов).

Надежность теста-ретеста (ретестовая надежность) определяется как коэффициент корреляции между результатами тестирования одной и той же выборки респондентов два раза через некоторый интервал времени при одних и тех же условиях. Надёжность тем выше, чем ее показатель

сильнее приближается к 1. Хорошими показателями считаются: 0,8–0,9 [Шмелев, 2013; Митина, 2011], в среднем 0,8 [Крокер, Алгина, 2010].

Надежность параллельных (альтернативных, эквивалентных) форм теста оценивается с помощью коэффициента корреляции между результатами тестирования одной и той же выборки по двум параллельным формам теста при одних и тех же условиях. Методика признаётся надёжной, если полученный коэффициент надёжности не ниже 0,75–0,85. Для тестов достижений приводятся ориентировочные значения 0,8–0,9 [Крокер, Алгина, 2010]. Согласно КТТ [Крокер, Алгина, 2010] две формы теста (два теста) являются параллельными, если каждый испытуемый имеет одну и ту же истинную оценку по обеим формам; дисперсии ошибок для двух форм равны. Такие тесты, как следствие, на одной и той же выборке будут иметь равные средние значения баллов и равные дисперсии. Обычно в литературе используется упрощенное определение параллельных форм. Различные формы (варианты) теста считаются параллельными, если они разработаны на основе одной спецификации, имеют одинаковое количество заданий попарно равной трудности с совпадающими характеристиками, и порождают на одной и той же выборке идентичные распределения наблюдаемых баллов (распределения с одинаковыми средними, дисперсиями и другими выборочными характеристиками).

Надежность расщепления теста пополам определяется путем анализа согласованности результатов отдельных совокупностей тестовых вопросов или единичных пунктов теста, например, методом расщепления, при котором вычисляется коэффициент корреляции между значениями двух эквивалентных по характеру групп пунктов теста (например, четные и нечетные вопросы). При этом, в результате расщепления длина теста уменьшается в два раза, поэтому коэффициент надежности, вычисленный этим способом, недооценивает (занижает) коэффициент надежности всего теста, т.к. более длинные тесты более надежны. Чтобы преодолеть эту проблему, для коррекции коэффициента надежности в данном случае используют формулу Спирмена-Брауна. Отметим, что в современной психометрической практике метод используется довольно редко, особенно для психодиагностических методик, так как в его основе лежит маловероятное предположение о параллельности двух половин теста при расщеплении.

Внутренняя согласованность используется, когда надо проанализировать всю совокупность пунктов, относящихся к данной шкале в целом. Например, для дихотомических шкал можно использовать формулу Кьюдера-Ричардсона, а для других шкал — коэффициент надежности альфа Кронбаха. Анализ внутренней согласованности предполагает также вычисление альфа Кронбаха при удалении пункта опросника: если при этом коэффициент альфа уменьшается, то данный пункт хорошо «работает» на данную субшкалу. Принято считать, что коэффициент альфа задает нижнюю границу коэффициента надежности. Например, если коэффициент альфа получился равным

0,8, то мы можем утверждать, что, по крайней мере, 80% дисперсии наблюдаемых оценок является следствием дисперсии истинных баллов испытуемых.

2) Второй вид надежности — **стабильность (постоянство во времени) измеряемого свойства (характеристики).**

Стабильность измеряемого свойства определяется в зависимости от теоретической концепции, лежащей в основе разрабатываемой методики и вычисляется обычно как коэффициент корреляции между результатами теста и ретеста через определенный промежуток времени, когда имеются подтверждения устойчивости самой методики.

3) Третий вид надежности — **относительная независимость результатов от личности экспериментатора (константность).**

Надежность тем выше, чем ее показатель сильнее приближается к 1. Методика признается надежной, если полученный коэффициент надежности не ниже 0,75 [Бурлачук, 2006; Бурлачук, Морозов, 2002]. При этом ориентировочные значения для тестов достижений могут быть 0,80–0,90 [Крокер, Алгина, 2010].

Диагностическая методика всегда снабжается подробными инструкциями по ее применению, правилами и примерами, указывающими, как проводить опрос, однако полностью регламентировать манеру поведения экспериментатора невозможно. Показатель константности вычисляется как коэффициент корреляции результатов двух измерений, проведенных в одинаковых условиях на одной и той же выборке респондентов, но разными экспериментаторами. Если в присутствии нового экспериментатора сдвиг произошел у всех испытуемых в равной степени, то сам по себе этот факт на надёжность методики влияния не окажет. Надёжность изменится, если действие экспериментатора окажет различное влияние на испытуемых: у одних результаты увеличатся, у других уменьшатся, у третьих останутся прежними. Коэффициент корреляции не должен быть меньше 0,8 [Митина, 2011]

2.5. Проверка валидности инструмента

2.5.1. Характеристика различных видов валидности

Валидность теста — характеристика пригодности оценочной информации для принятия правильных решений на ее основе. Валидность теста говорит нам о том, что тест измеряет и насколько хорошо он это делает, какие выводы можно сделать из полученных по нему показателей [Анастази, Урбина, 2007, С. 134] Выделяют различные виды валидности теста. Заметим, что терминология в этой области может быть очень разнообразной, а трактовка понятий может различаться в разных источниках. Проверка валидности методики называется валидизацией и для большинства видов валидности определяется с помощью внешнего критерия — независимого от методики показателя проявления изучаемого свойства. Термин «стандартизация» более широкий и подразумевает проверку всех психометрических характеристик инструмента, включая вычисление нормативных значений.

Конструктная валидность теста показывает, насколько его результаты могут рассматриваться в качестве меры некоторого теоретического конструкта или свойства. [Анастази, Урбина, 2007, С. 149] Конструктная валидность характеризуется способностью теста к измерению такой черты, которая была обоснована теоретически (как теоретический конструкт). Когда сложно найти адекватный прагматический критерий, может быть выбрана ориентация на гипотезы, сформулированные на основе теоретических предположений об измеряемом свойстве. Подтверждение этих гипотез свидетельствует о теоретической обоснованности методики.

Сначала необходимо, насколько возможно полно, содержательно описать конструкт, для измерения которого предназначен тест. Это достигается за счет формулирования гипотез о нем, предписывающих, с чем данный конструкт должен коррелировать, а с чем не должен. После чего данные гипотезы проверяются. Это наиболее эффективный способ валидизации для личностных опросников, для которых установление единственного критерия их обоснованности является затруднительным. Приведем примеры конструктов: базовые навыки чтения, математики, базовая фонематическая грамотность первоклассников; функциональное владение нормами языка; компетенции 4К (soft skills) — критическое и креативное мышление, коммуникация, кооперация; информационно-коммуникационная компетентность; универсальные учебные действия (планирование, рефлексия, построение рассуждений и способов решения проблем поискового характера, способность классифицировать, моделировать приводить доказательство и др.), личностные свойства (гибкость, самостоятельность, ответственность и др.), невербальный и вербальный интеллект, школьная тревожность и др.

Конструктная валидность — это базисное понятие, включающее ее остальные виды. Конструкты — это широкие категории, выводимые логическим путем из общих признаков, свойств или черт, обнаруживающих себя в непосредственно наблюдаемых поведенческих переменных. Это «латентные переменные», т.е. теоретические категории, недоступные непосредственному наблюдению. [Анастази, Урбина, 2007, С. 134]. Проверка конструктной валидности подразумевает применение методов эксплораторного факторного анализа (ЭФА) с целью выделения субшкал опросника и конфирматорного факторного анализа (КФА) для проверки соответствия эмпирических данных ранее выделенным субшкалам [Митина, 2011; Наследов, 2013; Сорокова, 2014; Structural equation modeling, 1995]. Основанием для включения пункта опросника в конкретную субшкалу служит его большая факторная нагрузка (по модулю не менее 0,5) на соответствующий фактор и одновременно малые факторные нагрузки на остальные факторы.

Критериальная валидность показывает связь результатов оценки измеряемого конструкта с внешним критерием, релевантность которого обосновывается данными предыдущих исследований. Например, для валидизации

тестов интеллекта часто используются показатели учебных достижений: школьные оценки, показатели тестов академических достижений по учебным предметам, данные о переводе в следующий класс, а также рейтинги учащихся, составляемые педагогами [Анастази, Урбина, 2007, С. 141–142]. Критериальная валидность может проверяться с использованием других уже стандартизованных методик, измеряющих теоретически близкие конструкты. Для этого рассчитывают корреляции данного теста с другими валидизированными тестами и анализируют, являются ли эмпирически полученные корреляции ожидаемыми и теоретическими объяснимыми.

В связи с понятием «критериальная валидность» также используются термины «внешняя валидность», «конвергентная валидность», «дивергентная валидность», «дифференциальная валидность», которые могут по-разному пониматься в классической психометрике. Например, конвергентная валидность может трактоваться как прямые корреляции субшкалы теста с теоретически близкими субшкалами других валидизированных тестов или обратные корреляции с конструктами, выражающими противоположные тенденции. Дифференциальная валидность некоторыми авторами понимается как теоретически ожидаемое отсутствие корреляций субшкалы теста с субшкалами других валидизированных тестов (мы в данном случае будем использовать понятие «дивергентная валидность»), а другими — как способность теста выявлять различия по субшкалам для разных категорий испытуемых, например, по полу, возрасту, клиническому диагнозу и др.

Конвергентная валидность — частный случай конструктной валидности (и критериальной валидности), который выражается в требовании статистической зависимости (коррелированности) двух тестовых показателей, если они должны быть направлены на измерение концептуально родственных психических свойств индивида. Другими словами, конвергентная валидность — это корреляции теста с теми переменными, с которыми он должен коррелировать (прямо или обратно) из теоретических предположений. **Дивергентная валидность** — частный случай конструктной валидности (и критериальной валидности), который выражается в требовании статистической независимости (некоррелированности) двух тестовых показателей, если они не должны быть направлены на измерение концептуально родственных психических свойств индивида. Другими словами, **дивергентная валидность** — это низкие корреляции с нерелевантными тесту переменными, т.е. с теми, от которых он должен отличаться (заметим, что А. Анастази и С. Урбина в этом случае используют термин «дифференциальная валидность» [Анастази, Урбина, 2007, С. 152]). Пример конвергентной валидности: тест базовых математических навыков первоклассников, предположительно, должен прямо коррелировать с тестом на способность к логическим рассуждениям. Пример дивергентной валидности: тест базовой фонематической грамотности, предположительно, не должен показывать значимых корреляций с тестами креативности или тревожности.

Дифференциальная и **дискриминантная валидность** часто понимаются как способность методики выявлять различия по социально-демографическим и другим факторам. Чаще всего исследуется влияние факторов «пол», «возраст», «образование», «этнокультурная принадлежность», а также «клинический диагноз» или «нозология» и других релевантных факторов на показатели по субшкалам и итоговым шкалам опросника.

Основной процедурой определения валидности является корреляционный анализ связи результатов теста с критериальными характеристиками исследуемого свойства. Еще одним распространенным способом характеристики диагностической эффективности (дифференциальной валидности) методики является **сравнение контрастных групп**. При этом, как правило, используют сложный критерий, в котором отражено комплексное влияние различных факторов. Например, при определении валидности теста интеллекта показатели детей с диагностированной задержкой психического развития могут быть сопоставлены с показателями нормотипичных школьников того же возраста. Аналогично валидность личностного опросника, предназначенного для выявления «уровня невротизации», определяется на основе сравнения его результатов у пациентов, страдающих неврозами, и практически здоровых людей. Сравнение выраженности тестовых баллов в контрастных группах обычно проводят с помощью t-критерия Стьюдента для двух независимых выборок или его непараметрического аналога — критерия Манна-Уитни. Если внешний критерий имеет более двух уровней реализации, то может применяться однофакторный дисперсионный анализ или его непараметрический аналог критерий Краскела-Уоллиса.

Следует отметить, что **не существует однозначных правил** относительно того, какими должны быть показатели валидности и надёжности, чтобы методика была принята к использованию. На деле **допускаемые величины показателей валидности и надёжности** складываются исходя из общепринятой практики работы исследователей над различными тестами. Как правило, низким признаётся показатель валидности порядка 0,2–0,3, средним 0,4–0,5 и высоким — свыше 0,6. Рекомендуемая ретестовая надёжность не должна быть ниже 0,7. Надёжность расщепления или параллельных форм также не ниже 0,7. Хорошим показателем считается более 0,8, а очень хорошим — более 0,9 [Митина, 2011]

Другие виды валидности. Помимо перечисленных видов валидности в современной литературе встречаются также другие виды, например, прагматическая валидность, частными случаями которой служат прогностическая, текущая и ретроспективная валидность; инкрементная валидность; экспертная валидность и др. [Митина, 2011] При этом терминология может отличаться в разных источниках.

Прагматическая валидность — определение практической ценности методики с точки зрения ее эффективности, значимости, полезности. Критериями прагматической валидности выступают показатели, обладающие

непосредственной ценностью для определенных областей практики, например, «успеваемость», «количество принимаемых обезболивающих», и т.д.

Прогностическая валидность отражает информацию о том, с какой степенью точности и обоснованности методика (инструмент, тест) позволяет судить о диагностируемом психологическом качестве спустя определенное время после измерения. Обычно определяется по достаточно надежному внешнему критерию, но информация о нем собирается спустя определенное время после измерения по методике. Основной процедурой определения прогностической валидности является корреляционный анализ связи результатов теста с критериальными характеристиками исследуемого свойства.

Текущая валидность (диагностическая, конкурентная) — это соответствие результатов валидизируемого теста независимому критерию, отражающему состояние исследуемого тестом качества в момент проведения исследования. Основной процедурой определения текущей валидности является корреляционный анализ связи результатов теста с критериальными характеристиками исследуемого свойства. Отличие текущей валидности от прогностической валидизации заключается в том, что оба источника информации об испытуемом — и тест, и критерий — «работают» фактически на одном и том отрезке времени, то есть совпадают в реальном масштабе времени. **Конкурентная валидность** — наличие высокой корреляции между разрабатываемой и существующей методикой, измеряющей то же свойство, валидность которой уже подтверждена. При наличии такой связи можно говорить о том, что разрабатываемая методика измеряет то же свойство, что и существующая.

Ретроспективная валидность определяется на основе критерия, отражающего события или состояние качества в прошлом. Может быть использована для быстрого получения сведений о предсказательных возможностях методики. Так, для проверки того, в какой мере хорошие результаты теста способностей соответствуют быстрому обучению, можно сопоставить прошлые оценки успеваемости, прошлые экспертные заключения и т.д. у лиц с высокими и низкими на данный момент диагностическими показателями.

Инкрементная валидность — психометрическая характеристика теста, состоящая в относительном приращении точности отнесения испытуемого к определенной категории над возможной точностью отнесения, которая имелась уже до проведения данного теста. Проверяет, объясняет ли характеристика, измеряемая с помощью данного теста, дополнительную часть дисперсии прогнозируемого результата по сравнению с дисперсией, объясняемой конкурирующими конструктами. Часто вычисляется с помощью сравнения нескольких регрессионных моделей. На первом шаге в регрессионную модель вводятся только конкурирующие конструкты (модель I). На следующем шаге в модель в качестве предиктора добавляется характеристика, измеряемая с помощью данного теста (модель II), и оценивается разница в процентах объясненной дисперсии между моделями I и II.

Экспертная (оценочная) валидность — это степень соответствия между показателями, полученными испытуемыми по данной методике, и оценками измеряемого свойства сторонними экспертами (близкими, педагогами, коллегами по работе и т.д. — то есть людьми, хорошо знающими обследуемого).

Основной процедурой определения валидности является корреляционный анализ связи результатов теста с критериальными характеристиками исследуемого свойства. Еще одним распространенным способом характеристики диагностической эффективности (дифференциальной валидности) методики является сравнение контрастных групп. При этом, как правило, используют сложный критерий, в котором отражено комплексное влияние различных факторов. Например, при определении валидности теста интеллекта показатели детей с диагностированной задержкой психического развития могут быть сопоставлены с показателями нормотипичных школьников того же возраста. Аналогично валидность личностного опросника, предназначенного для выявления «уровня невротизации», определяется на основе сравнения его результатов у пациентов, страдающих неврозами, и практически здоровых людей.

Содержательная валидность — показывает степень репрезентативности содержания пунктов методики измеряемой области психического свойства. Содержательная валидность закладывается на этапе разработки методики — при подборе пунктов будущей методики. Разработке конкретного содержания пунктов методики предшествует системный анализ соответствующей литературы и консультации со специалистами по диагностируемой области.

2.5.2. Использование факторного анализа для проверки валидности инструмента

Если пункты (вопросы) теста представляют собой шкалы Ликерта (в другой транскрипции, Лайкерта) диапазоном не менее 3-х градаций, при проверке валидности может быть использован факторный анализ. Факторный анализ применяется для проверки того, что определенное подмножество вопросов теста измеряют заранее определенный конструкт (характеристику). Валидность проверяется по отношению к внутренней структуре теста: исследуются отношения между пунктами теста, субшкалами теста и их соответствие предполагаемой, теоретически обоснованной структуре. Говоря в широком смысле, факторный анализ может быть разделен на эксплораторный факторный анализ (ЭФА) и конфирматорный факторный анализ (КФА) [Митина, 2011; Наследов, 2013; Сорокова, 2014; Structural equation modeling, 1995].

ЭФА может быть использован для выявления структуры полученных данных. В этом случае он может показать, как вопросы и конструкты связаны между собой и помочь в разработке новых теорий. Именно поэтому лучше применять ЭФА на ранних стадиях разработки диагностического

инструментария. ЭФА также может использоваться для проверки одномерности диагностического инструмента. Перед применением ЭФА рекомендуется проверить, значимо ли отличаются корреляции между переменными, которые планируется использовать в ЭФА, от нуля с помощью Bartlett`s Test of Sphericity (в тесте сферичности Бартлетта значения $p < 0,05$ указывают на то, что данные вполне приемлемы для проведения факторного анализа) и на степень применимости факторного анализа к данной выборке с помощью Kaiser-Meyer-Olkin Measure of Sampling Adequacy (критерий адекватности выборки Кайзера-Мейера-Олкина показывает долю дисперсии переменных, которая может быть обусловлена лежащими в их основе факторами. KMO тест: если значение теста менее 0,05 — ФА неприменим к выборке; более 0,5 низкая адекватность; более 0,6 удовлетворительная адекватность; более 0,7 приемлемая адекватность; более 0,8 высокая адекватность; более 0,9 безусловная адекватность) [Наследов, 2013]. Основанием для принятия решения о количестве выделяемых факторов с помощью ЭФА могут служить несколько подходов [Lee, Ashton, 2007]: 1) метод Кеттелла — диаграмма «каменистой осыпи» / scree plot (оставляют то количество факторов, на котором диаграмма «ломается» сильно последний раз — факторы, образующие сам склон, считаются важными для рассмотрения, а факторы, представляющие «каменистую осыпь», — нет); 2) критерий Кайзера (оставляют то количество факторов, для которых собственные значения корреляционной матрицы выше или равны единице) [Крамер, 2007]; 3) процент общей объясненной дисперсии (оставляют то количество факторов, которые объясняют 60% и более общей дисперсии; 4) параллельный анализ (Parallel analysis) — оставляют то количество факторов, для которых собственные значения статистически достоверно превышают собственные значения факторов из сгенерированной случайным образом матрицы, аналогичной исходной матрице данных, где отсутствует факторная структура [Horn, 1965]; 4. MAP (Minimum average partial). Метод MAP заключается в вычислении минимального числа компонентов, для которых среднеквадратичная частная корреляция является минимальной [Velicer, 1976].

КФА обычно используется для подтверждения предварительно обоснованной теоретической модели и показывает, соответствуют ли полученные эмпирические данные модели. КФА следует использовать, когда теоретические конструкты хорошо поняты и четко сформулированы, а валидность внутренней структуры диагностического инструмента (отношения между пунктами) была проверена в сходных случаях. Соответствие модели КФА эмпирическим данным проверяются с помощью критерия Хи-квадрат Пирсона и ряда индексов соответствия, причем рекомендуется приводить как можно больше различных индексов. Наиболее часто используемыми являются критерии согласия, такие как GFI (Goodness of fit), AGFI (Adjusted goodness of fit), PGFI (Parsimonious goodness-of-fit index), IFI (Bollen's Incremental Fit Index), TLI

(Tucker Lewis index), CFI (Comparative fit index) (>0,95 показывают хорошее согласие; >0,90 приемлемое согласие); квадратный корень среднеквадратической ошибки аппроксимации RMSEA (Root mean square error of approximation) и границы его 90 % доверительного интервала (<0,05 показывает хорошее соответствие, <0,06–0,08 приемлемое соответствие, <0,1 слабое соответствие); стандартизированный корень среднего остатка SRMR (Standardized root mean square residual) (<0,06 показывает хорошее соответствие, <0,06–0,08 приемлемое соответствие,); относительный (или нормированный) хи-квадрат χ^2/df (<3 показывает хорошее соответствие) [Клайн, 1994; Byrne, 2010; Brown, 2014]. В психометрической практике для оценки соответствия модели эмпирическим данным чаще используются индексы TLI, CFI, RMSEA и SRMR. Связи между факторами в КФА должны хорошо воспроизводить матрицу корреляций шкал между собой и с общим (итоговым) баллом инструмента.

Оценка моделей в КФА проводится с помощью разных методов: ML (Maximum Likelihood) – метод максимального правдоподобия (применяется для непрерывных нормально распределенных переменных), MLR (Robust Maximum Likelihood) – робастный метод максимального правдоподобия (для непрерывных переменных, допускает ненормальное распределение переменных), WLSMV (Robust Weighted Least Squares) – метод взвешенных наименьших квадратов и DWLS (Diagonally Weighted Least Squares) – метод диагонально взвешенных наименьших квадратов (для порядковых переменных, типа шкал Лайкерта) [Brown, 2014; Xia, 2019].

В случае, если модель оказалась слабо согласованной с данными, то возможны несколько направлений коррекции модели: 1) фиксирование нулевых значений для статистически недостоверных параметров (удаление «лишних» стрелок); 2) удаление «лишних» переменных, статистически недостоверно связанных с остальными переменными или особенно сильно нарушающих согласие модели; 3) освобождение фиксированных нулевых параметров, повышающих согласие модели (добавление новых стрелок, опираясь на индексы модификации/ Modification Indices); 4) добавление в модель новых переменных, например вторичных факторов [Наследов, 2013].

Прежде, чем диагностический инструмент может использоваться в научных и практических целях, результаты ЭФА должны быть подтверждены с помощью КФА, причем это подтверждение не рекомендуется проводить на той же самой выборке испытуемых, что и первоначальный ЭФА, что может быть достигнуто с помощью деления одной большой выборки на две случайные части.

2.6. Корреляции между шкалами и с общим (итоговым) баллом инструментария

Корреляции между шкалами и с общим (итоговым) баллом инструментария должна воспроизводить обоснованные теоретически предполагаемые отношения между конструктами.

2.7. Расчет нормативных значений: тестовые нормы

Стандартизация психологического теста состоит в определении норм выполнения теста, а также преобразовании нормальной (или искусственно нормализованной) шкалы оценок в новую шкалу, основанную уже не на количественных эмпирических закономерностях, а на его относительном месте в распределении результатов в выборке испытуемых. Нормы разрабатываются для различных возрастов, профессий, пола и пр. Наличие нормативных данных (норм) в стандартизованных методах психодиагностики является их существенной характеристикой. Нормы необходимы при интерпретации тестовых результатов (первичных показателей) в качестве эталона, с которым сравниваются результаты тестирования.

Расчет тестовых норм. Расчет репрезентативных норм для субшкал и итоговых шкал опросника – необходимая часть процедур стандартизации. Дополнительно могут быть рассчитаны шкалы стенов или станайнов для обеспечения сопоставимости результатов по шкалам данного опросника и других методик, а также для удобства интерпретации результатов индивидуального обследования.

Для каждой стандартной шкалы существуют свое среднее арифметическое и стандартное отклонения, которые заранее известны. Традиционно чаще всего используются следующие шкалы.

Таблица 1

Типы стандартных шкал

Тип шкалы	Среднее (M)	Стандартное отклонение (s)
Шкала станайнов	5	2
Шкала стэнов	5,5	2
Т-шкала	50	10
Шкала IQ	100	(иногда 12,14,16,18)
Пятибалльная шкала	3	1
Семибалльная шкала	4	1

Чтобы пользователю не приходилось делать каждый раз вычисления и переводить результаты, полученные по тесту, в одну из стандартных шкал, в руководстве к тесту всегда должна даваться специальная таблица перевода первичных баллов по тесту, которые в принципе можно получить (эти баллы часто называют «сырыми»), в одну из стандартных шкал. В такой таблице каждому «сырому» баллу теста соответствует какой-то балл по стандартной шкале.

В то же время, любой человек, умеющий вычислять среднее арифметическое и стандартное отклонение для группы данных, может перевести полученный в «сырых» баллах результат в одну из стандартных шкал. Для этого сначала тестовые баллы испытуемых переводят в z-оценки, которые

вычисляются как отношение разности данного тестового балла и среднего по выборке к стандартному отклонению по выборке. Затем полученные z-оценки подставляют в формулу:

$$T = M + sZ,$$

где T — балл по новой шкале;
M —среднее значение признака по новой шкале;
s —стандартное отклонение в выборке по новой шкале;
Z — балл по Z-шкале (z-оценка).

Для тестов, надежность которых выше 0,7 (а это свойственно большинству профессиональных методов), доверительный интервал для 5-процентного уровня не превышает 2 станайна, а для 1-процентного уровня — 3 станайна. Поэтому, когда в руководстве не приводятся доверительные интервалы и пользователь даже не знает точных коэффициентов надежности, чтобы вычислить их самостоятельно, можно исходить из приблизительной оценки: 2 станайна — минимальный интервал, на который следует обращать внимание при интерпретации различий в индивидуальных результатах обследуемых; различия в 3 станайна высоко достоверны.

Кроме того, шкала станайнов обладает еще рядом удобных для пользователя свойств. Например, 1 станайн, охватывающий нижние 4% обследуемых, интерпретируется как «крайне низкий». А при анализе данных тестов интеллекта нижние 2% традиционно признаются легкой степенью умственной отсталости. От «ниже среднего» (3) такой результат отделяют как раз 2 станайна, а от «среднего» (4) — 3. Результат в 2 станайна отделяют от «среднего» 2 станайна. С вероятностью 95% можно утверждать, что истинный показатель обследуемого находится, как минимум, ниже среднего в целом по группе, на которой были получены нормы. Интервал от 4 до 6 станайнов охватывает «средние» результаты (по 27% от центра распределения), что традиционно интерпретируется как «норма». Различия между результатом в 3 станайна («ниже среднего») и 7 станайнов («выше среднего») достоверны на 1% уровне. Все сказанное характерно и для верхней части шкалы.

Таблица 2

Интерпретация шкалы станайнов и Т-шкалы

Станайн/Т-шкала	Характеристика результата	Охватываемый процент обследуемых
1 / до32	Очень низкий	4
2 / 33–37	Низкий	7
3 / 38–42	Ниже среднего	12
4 / 43–47	Средний	17
5 / 48–52	Средний	20
6 / 53–57	Средний	17
7 / 58–62	Выше среднего	12
8 / 63–67	Высокий	7
9 / более 68	Очень высокий	4

3. Введение в Современную теорию тестирования (IRT)

3.1. Основные положения современной теории тестирования (IRT)

Современная теория тестирования — Item Response Theory (IRT), которую также называют «Теория латентных черт», «Теория моделирования и параметризации тестов» [Нейман, Хлебников, 2000; Карданова, 2008], и т.д., возникла в середине 20-го века. **Цель IRT**: создать основу для оценки того, насколько хорошо работает весь тест для измерения определенной черты (характеристики) и насколько хорошо работает каждый отдельный пункт теста [Baker, 1985; Embretson, Reise, 2000].

Современная теория тестирования включает набор математических моделей, которые описывают в терминах теории вероятности связь между ответом респондента на вопрос теста с его уровнем выраженности «латентной характеристики», которая измеряется данным тестом и с характеристиками самого вопроса. Эта латентная характеристика представляет собой гипотетический конструкт — свойство, черту, способность, которая предполагается существующей, но не измеряемой прямо (например, интеллект, тревожность, экстраверсия-интроверсия и т.д.). Предполагается, что уровень выраженности данной характеристики — единственный фактор, который определяет ответ на каждый вопрос (пункт) шкалы. Модели IRT основаны на вероятности того, что респондент даст определенный ответ на данный вопрос в соответствии со своим уровнем измеряемой латентной характеристики.

Параметрами моделей IRT обычно служат параметры респондентов и параметры заданий. Эти параметры оцениваются методами математической статистики. Полученные оценки находятся на метрической шкале, причем эта шкала единая и для параметров респондентов, и для параметров заданий. Таким образом, используя модели IRT и зная оценки параметров, можно вычислить вероятность ответа в любой категории для любого респондента в любом задании теста, что позволяет строить прогнозы и сокращать время тестирования. Следует отметить, что для применения современной теории тестирования требуется специализированное программное обеспечение, например, программы Winsteps, ConQuest, XCalibre и др., или статистический пакет R.

В основе моделей IRT лежит характеристическая кривая задания (item characteristic curve (ICC)), которая в простейшем дихотомическом случае определяет зависимость вероятности правильного ответа на данный вопрос теста от уровня способности респондентов. Характеристические кривые различных заданий могут отличаться положением кривой вдоль оси

переменной и формой (крутизной) кривой. Эти два свойства характеристической кривой связаны с двумя свойствами заданий — трудностью и дискриминативностью.

IRT — статистическая теория, позволяющая численно определить уровень выраженности способности/латентной черты у респондентов, измеряемой данным тестом. При этом IRT позволяет преодолеть одно из важнейших ограничений классической теории тестирования — зависимость характеристик пунктов теста от контингента испытуемых и зависимость характеристик испытуемых от самих пунктов теста. Если используемая модель IRT соответствует эмпирическим данным, то одни и те же пункты теста сохраняют свои статистические свойства (например, трудность и дискриминативность) при использовании на разных выборках респондентов, а уровни выраженности способности респондентов не зависят от конкретного набора тестовых пунктов, на которых были получены.

Предположения IRT:

1) Одномерность. Только один конструкт (одна черта) может измеряться набором пунктов теста. Следовательно, предполагается, что некий один фактор объясняет большую часть вариативности (дисперсии) ответов на все пункты теста. Данное предположение в простейшем случае может быть проверено по аналогии с КТТ — с помощью эксплораторного факторного анализа (20 % или более общей дисперсии должно приходиться на первый фактор) или конфирматорного факторного анализа (для проверки гипотезы о единственном факторе). При этом, стоит отметить, что в IRT разработаны и другие методы.

Разумеется, современная оценочная практика выходит за рамки одномерности. Многие конструкты имеют сложную природу и состоят из субконструктов-составляющих. Например, в компетенции «работа в команде» выделяют несколько составляющих, связанных как с коммуникативными навыками (прием и передача информации), так и с саморегуляцией и целеполаганием (постановка групповых целей, расстановка приоритетов, распределение ролей). В таких случаях важно оценить и общий уровень способности, и способности по отдельным составляющим (шкалам). Специально для таких конструктов разработаны более сложные модели IRT — многомерные, бифакторные и др. Предположение одномерности в этом случае понимается следующим образом: тест должен измерять столько конструктов, сколько в него заложено разработчиками теста в соответствии с теоретической моделью конструкта.

2) Локальная независимость. Связи между ответами респондента на задания теста могут быть объяснены только его уровнем способности. Свойство локальной независимости нарушается, например, тогда, когда одно задание теста содержит информацию, полезную при ответе на другое задание теста. Свойства одномерности и локальной независимости взаимосвязаны. Если тест одномерный, что свойство локальной

независимости выполняется. Однако свойство локальной независимости может выполняться даже, если тест не является одномерным. Можно сказать, что размерность теста равна числу латентных характеристик, необходимых для достижения локальной независимости.

3) Характеристическая кривая задания отражает истинную связь между ненаблюдаемыми переменными (уровнем способности) и наблюдаемыми (ответами на задание).

3.2. Модели IRT и информационные функции

Модели IRT. В IRT используются различные модели, число которых постоянно растет. Наиболее часто используемыми являются следующие.

1) Модели для дихотомических (бинарных) данных: 1PL, 2PL, 3PL.. Данные модели различаются по числу параметров заданий, включенных в модель.

Модель 1PL включает только один параметр задания — его трудность. Обычно трудность изменяется от –3 до +3, где 0 соответствует среднему уровню, положительные значения — более трудным заданиям и отрицательные — более легким.

Модель 2PL включает помимо трудности еще один параметр задания — его дискриминативность. Чем больше значение параметра дискриминативности, тем круче наклон ICC и тем эффективнее этот вопрос может различать респондентов по уровню их латентной характеристики.. Обычно параметр дискриминативности изменяется от 0 до 2 (чем больше, тем лучше).

Модель 3PL включает помимо трудности и дискриминативности еще один параметр задания — параметр угадывания. Этот параметр обеспечивает ненулевую нижнюю асимптоту для характеристической кривой задания и представляет вероятность испытуемых с низким уровнем подготовки выполнить задание верно

2) Модели для политомических (небинарных) шкал. В психодиагностике часто применяются шкалы с политомической оценкой, например, шкалы типа Ликерта. Специально для таких шкал разработаны модели Rating Scale Model (RSM) и Modified Graded Response Model (M-GRM), а также их более общие случаи — модели Partial Credit Model (PCM), Graded Response Model (GRM), Generalized Partial Credit Model (GPCM) и др. В этих моделях в качестве параметров заданий выступают трудности выбора различных категорий в пунктах теста, а в некоторых моделях еще и другие характеристики заданий.

Test information curve. В IRT особое внимание обращают на информационную функцию (test information curve), которая показывает, насколько хорошо каждый уровень выраженности черты/характеристики оценивается тестом. Информационная функция вычисляется на основании суммирования информации, полученной от каждого пункта (утверждения) опросника. График информационной функции теста показывает, как методика

оценивает измеряемый конструкт в полном диапазоне значений этого конструкта: чем больше информации соответствует данному значению измеряемого конструкта, тем меньше ошибка его измерения. Наименьшая ошибка измерения будет у респондентов, чей уровень измеряемого конструкта находится вблизи пика информационной кривой, и именно для этих респондентов методика будет наиболее подходящей.

Таким образом, при применении IRT для оценки диагностического инструмента должно быть сделано следующее:

1) Проведена проверка размерности теста. Если инструмент включает в себя несколько шкал, IRT применяется к каждой из шкал отдельно или используются более сложные модели IRT.
2) Определена и описана модель, которая будет использована. Проверено соответствие имеющихся данных и используемой модели. В случае конкурирующих моделей следует привести сравнение моделей с целью выбора наиболее подходящей. Сделать это можно с помощью различных статистических критериев, например, индексов соответствия AIC, BIC или др.)
3) Рассчитаны параметры модели: получены оценки параметров заданий в соответствии с используемой моделью IRT и параметров испытуемых (уровни способности)
4) Оценено согласие эмпирических данных с используемой моделью. Сделать это можно с помощью специальных статистических критериев, а также через построение и анализ характеристических кривых заданий — теоретических и эмпирических.
5) Проверено качество теста как измерительного инструмента. В частности, оценивается надежность измерений и некоторые другие индексы, характеризующие точность измерений, строятся и анализируются информационные кривые для каждого пункта теста и график информационной функции для всего теста (test information curve). Дополнительно обычно строится карта переменных, представляющая собой график, на котором показаны оценки трудности заданий и параметров испытуемых на одной шкале. Карта позволяет оценить, насколько тест подходит по трудности данной выборке, как распределены трудности заданий и оценки испытуемых.
6) Проведен перевод параметров респондентов на шкалу, выбранную для сообщения результатов тестирования.

3.3. Оценивание сложных конструктов

Как отмечалось выше, многие конструкты имеют сложную природу и состоят из субкомпонент, каждая из которых может также делиться на отдельные составляющие. Например, критическое мышление включает две составляющие: анализ информации и построение вывода и аргументации. Составляющая «Анализ информации» включает в себя навыки работы с информацией в соответствии с целями и условиями поставленной задачи.

Составляющая «Вывод и аргументация» включает в себя построение собственного вывода и аргументов к нему в отношении решаемой проблемы с помощью результатов, полученных на этапе анализа. Инструменты оценивания таких конструктов должны оценивать все эти переменные с учетом их взаимосвязей и предоставлять доказательства того, что оценивается именно то, что планировалось оценивать, в силу чего стандартные инструменты оценивания мало подходят для оценивания таких сложных конструктов.

Чтобы оценить сложные конструкты, необходимо провести наблюдение за тем, как учащиеся принимают решения и действуют в различных ситуациях в реальной жизни. Поэтому инструмент измерения сложных конструктов должен, во-первых, позволять респонденту поведенчески проявлять сложную структуру навыка и, во-вторых, приближать ситуацию тестирования к реальной жизни. Например, это могут быть инструменты на основе заданий сценарного типа, игр, симуляций. Такие задания часто относят к аутентичным заданиям, что значит, что они соответствуют реальной коммуникативной среде и встречаются в реальной жизни. Например, поиск информации из надежного источника для решения задачи, уместная реакция на сообщение, написание письма, покупка билета на сайте и т.д. Такой подход лучше других подходит для оценки сложных конструктов.

Такие инструменты состоят из субшкал, которые определенным образом вкладываются в итоговый тестовый балл. На языке психометрики такие инструменты являются многомерными, и задача состоит в том, чтобы оценить, по возможности, как общую сформированность всего конструкта у респондентов, так и его отдельные составляющие. Информация как об уровне развития целостной характеристики, так и её составных частей, является важной для исследователей, а также для лиц, принимающих решения на основе результатов измерений. Это позволяет учитывать развитые и развивающиеся черты или способности респондентов, улучшать работу, например, корректируя практики преподавания.

Проводить обоснование психометрического качества таких инструментов можно с применением IRT. Для сообщения пользователям как общего балла по тесту, так и баллов по компонентам теста, требуется проведение дополнительных психометрических исследований, доказывающих валидность всех сообщаемых баллов. В статье [Федерякин Д.А., Ларина Г.С., Карданова Е.Ю., 2021] на примере инструмента PROGRESS-ML, направленного на измерение базовой математической грамотности учащихся в 3-ем классе, демонстрируется спектр таких исследований.

Заключение

Данное Руководство имело целью помочь разработчикам психодиагностических инструментов (методик, тестов, шкал, опросников) в создании надежного и валидного психодиагностического инструментария. Публикация информации о психометрических характеристиках диагностических методик позволит исследователям и психологам-практикам принимать обоснованные решения об их применении, будет способствовать росту культуры их применения в целом.

В разделе «Статьи по стандартизации и валидизации психодиагностических инструментов» приведены кейсы по разработке, стандартизации и валидизации психодиагностических инструментов для широкого спектра психологических характеристик, представленные в публикациях в рецензируемых научных журналах.

Литература

Международные и российские стандарты, руководства-прототипы и литература по общим вопросам

1. Доказательный подход: Руководство по верификации программ, технологий, практик в образовании и социальной сфере : учебное пособие / М.Г. Сорокова, О.А. Ульянина, Г.В. Семья, О.И. Леонова, Д.В. Лубовский, Е.И. Исаев, Т.Г. Подушкина, Н.П. Бусыгина, Н.П. Радчикова, А.А. Шведовская ; под ред. Марголиса А.А., Сороковой М.Г., Семья Г.В. — М. : Издательство ФГБОУ ВО МГППУ, 2024.
2. Российский стандарт тестирования персонала / 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. // Организационная психология. 2015. Т. 5. № 2. С. 67–138. URL: https://orgpsyjournal.hse.ru/2015–5–2/152057297.html (дата обращения 13.02.2022).
3. Система функционирования психологических служб в общеобразовательных организациях: методические рекомендации / Авт. коллектив. М.: Издательство ФГБОУ ВО МГППУ, 2020. — 120 с.
4. EFPA review model for the description and evaluation of psychological and educational tests. Test review form and notes for reviewers. Version 4.2.6. EFPA, 2013.
5. International Guidelines for Test Use. January 2001. International Journal of Testing 1:93–114.

Литература по классической, современной теории тестов и количественному анализу данных

6. Анастази А., Урбина С. Психологическое тестирование. — 7-е изд. — СПб: Питер, 2007. — 688 с.
7. Бурлачук Л.Ф. Психодиагностика: Учебник для вузов. — СПб.: Питер, 2006. — 351 с.
8. Бурлачук Л.Ф., Морозов С.М. Словарь-справочник по психодиагностике. — 2-е изд., перераб. и доп. —СПб.: Питер, 2002. — 528 с.
9. Карданова Е.Ю. Моделирование и параметризация тестов: основы теории и приложения. — М.: Федеральный центр тестирования, 2008. — 292 с.
10. Клайн П. Справочное руководство по конструированию тестов: Введение в психометрическое проектирование / Пер. с англ.; Под ред. Л.Ф. Бурлачука. — Киев, 1994. — 283 с.

11. Крамер Д. Математическая обработка данных в социальных науках: современные методы: учеб. пособие для студентов высших учебных заведений / Дункан Крамер; пер. с англ. И.В. Тимофеева, Я.И.Киселевой; науч. ред. О.В. Митина. — М.: Издательский центр «Академия», 2007. — 288 с.
12. Крокер Л., Алгина Дж. Введение в классическую и современную теорию тестов: учебник / Пер. с англ. Н.Н. Найденовой, В.Н. Симкина, М.Б. Челышковой; под общ. ред. В.И. Звонникова, М.Б. Челышковой — М.: Логос, 2010–668 с.
13. Митина О.В. Разработка и адаптация психологических опросников. — М.: Смысл, 2011. — 235 с.
14. Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных — СПб.: Речь, 2004. — 392 с.
15. Наследов А.Д. IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных. — СПб.: Питер, 2013. — 416 с.
16. Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. — М.: Прометей, 2000. — 169 с.
17. Сорокова М.Г. Методы математической статистики в психологии [Электронный ресурс]: учебное пособие. — Саарбрюкен: Palmarium Academic Publishing, 2014. — 405 с. — URL: https://psychlib.ru/mgppu/SMm-2014/MMa-405.htm#$p1 (дата обращения: 17.07.2023).
18. Шмелев А.Г. Практическая тестология. Тестирование в образовании, прикладной психологии и управлении персоналом. — М.: ООО «ИПЦ «Маска»», 2013. — 688 с.
19. Baker F.B. The basics of item response theory. — ERIC Clearinghouse on Assessment and Evaluation, College Park, MD, 1985. — 186 p.
20. Brown, T. A. (2014). Confirmatory factor analysis for applied research. New York, NY: Guilford Press.
21. Byrne, B.M. Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming. 2nd ed. Multivariate applications series. New York: Taylor & Francis Group, 2010. 396 p.
22. Embretson S.E., Reise S.P. Item response theory for psychologists. — Mahwah, NJ: Lawrence Erlbaum, 2000. — 384 p.
23. Horn, J.L. A rationale and test for the number of factors in factor analysis. Psychometrika 30, 179–185 (1965). https://doi.org/10.1007/BF02289447
24. Lee K., Ashton M.C. Factor analysis in personality research // Robins R.W., Fraley R.C., Krueger R.F. (eds). Handbook of research methods in personality psychology. N.Y.: The Guilford Press, 2007. P. 424–443.
25. Structural equation modeling: Concepts, issues and applications / Hoyle R.H. (Ed.). –Newbery Park, CA: Sage; 1995. — 312 p.
26. Velicer W.F. Determining the number of components from the matrix of partial correlations // Psychometrika. 1976. V. 41. P. 321–327.
27. Weinberg Sh.L., Abramowitz S.K. Data Analysis for the Behavioral Sciences Using SPSS. — Cambridge University Press, 2002. — 626 p.

28. Xia, Y., Yang, Y. RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. Behav Res 51, 409–428 (2019). https://doi.org/10.3758/s13428-018-1055-2

Статьи по стандартизации и валидизации психодиагностических инструментов

29. Антипкина И.В. Анализ опросника дошкольной родительской вовлеченности с использованием рейтинговой модели Раша // Современная зарубежная психология. 2018. Т. 7. № 3. С. 75–86. http://doi.org/10.17759/jmfp.2018070307
30. Александрова О.В., Дерманова И.Б. Семантический дифференциал жизненной ситуации // Консультативная психология и психотерапия. 2018. Том 26. № 3. С. 127–145. doi:10.17759/cpp.2018260307
31. Гордеева Т.О., Сычев О.А., Гижицкий В.В., Гавриченкова Т.К. Шкалы внутренней и внешней академической мотивации школьников // Психологическая наука и образование. 2017. Том 22. № 2. С. 65–74. doi:10.17759/pse.2017220206
32. Денисенкова Н.С., Тарунтаев П.И., Федоров В.В. Адаптация опросника родительского посредничества детской медиа-активности Г. Нимрод, Д. Лемиш, Н. Элиас на российской выборке родителей старших дошкольников // Психолого-педагогические исследования. 2023. Том 15. № 3. С. 96–114.
33. Ениколопов С.Н., Цибульский Н.П. Психометрический анализ русскоязычной версии Опросника диагностики агрессии А. Басса и М. Пери // Психологический журнал. 2007. № 1. С. 115–124.
34. Иванова А.Е. Валидизация опросника поведенческих характеристик младших школьников // Современная зарубежная психология. 2018. Т. 7. № 3. С. 86–95. http://doi.org/10.17759/jmfp.2018070308
35. Канонир Т.Н., Угланова И.Л., Куликова А.А. Мониторинг субъективного благополучия в школе: оценка в рамках современной теории тестирования // Мониторинг общественного мнения: Экономические и социальные перемены. 2022. № 4. С. 247–272. http://doi.org/10.14515/monitoring.2022.4.2010
36. Марголис А.А., Сорокова М.Г., Шведовская А.А., Радчикова Н.П. Разработка и стандартизация опросника «Шкала отношения к вакцинации от COVID-19» // Психология. Журнал Высшей школы экономики. 2022. Т. 19. № 3. С. 454–474. DOI: 10.17323/1813-8918-2022-3-454-474
37. Мешкова Н.В., Ениколопов С.Н., Митина О.В., Мешков И.А. Адаптация опросника «Поведенческие особенности антисоциальной креативности» // Психологическая наука и образование. 2018. Том 23. № 6. С. 25–40. doi:10.17759/ pse.2018230603

38. Моросанова В.И., Бондаренко И.Н., Фомина Т.Г. Создание русскоязычной версии опросника проявлений психологического благополучия (ППБП) для обучающихся подросткового возраста // Вопросы психологии. 2018. № 4. С. 103–109.
39. Никитская М.Г., Угланова И.Л. Русскоязычная версия опросника целей учебных достижений: разработка, валидизация и исследование функциональных возможностей // Психологическая наука и образование. 2021. Т. 26. № 5. С. 67–84. http://doi.org/10.17759/pse.2021260506
40. Орел Е.А., Куликова А.А. Анализ психометрических характеристик инструмента оценки социально-эмоциональных навыков в начальной школе [Электронный ресурс] // Современная зарубежная психология. 2018. Том 7. № 3. С. 8–17. doi:10.17759/jmfp.2018070301
41. Петренко В.Ф., Митина О.В. Методика «Сказочный семантический дифференциал»: диагностические возможности // Психологическая наука и образование. 2018. Том 23. № 6. С. 41–54. doi:10.17759/pse.2018230604
42. Полякова Ю.М., Сорокова М.Г., Гаранян Н.Г. Факторная структура и надежность шкалы взаимной адаптации в паре (DAS) в российской выборке // Консультативная психология и психотерапия. 2018. Том 26. № 3. С. 105–126. doi:10.17759/cpp.2018260306
43. Сирота Н.А., Московченко Д.В., Ялтонский В.М., Ялтонская А.В. Разработка русскоязычной версии опросника проблемного использования социальных сетей // Консультативная психология и психотерапия. 2018. Том 26. № 3. С. 33–55. doi:10.17759/cpp.2018260303
44. Сорокова М.Г., Одинцова М.А., Радчикова Н.П. Шкала оценки цифровой образовательной среды (ЦОС) университета // Психологическая наука и образование. 2021. Том 26. № 2. С. 52–65. doi:10.17759/pse.2021260205
45. Татарко А.Н., Лебедева Н.М. Разработка и апробация сокращенной версии методики «Социальные аксиомы» М. Бонда и К. Леунга // Культурно-историческая психология. 2020. Том 16. № 1. С. 96–110. doi:10.17759/chp.2020160110
46. Угланова И.Л., Брун И.В., Васин Г.М. Методология Evidence-Centered Design для измерения комплексных психологических конструктов [Электронный ресурс] // Современная зарубежная психология. 2018. Том 7. № 3. С. 18–27. doi: 10.17759/jmfp.2018070302
47. Федерякин Д.А., Ларина Г.С., Карданова Е.Ю. Измерение базовой математической грамотности в начальной школе // Вопросы образования. 2021. № 2. С. 199–226. http://doi.org/10.17323/1814-9545-2021-2-199-226
48. Холмогорова А.Б., Воликова С.В., Сорокова М.Г. Стандартизация опросника «Семейные эмоциональные коммуникации» // Консультативная психология и психотерапия. 2016. Том 24. № 4. С. 97–125. doi:10.17759/cpp.2016240405

49. Шинина Т.В., Митина О.В. Разработка и апробация опросника «Готовность подростков к самостоятельной жизни»: оценка и развитие жизненных навыков // Психологическая наука и образование. 2019. Том 24. № 1. С. 50–68. doi:10.17759/pse.2019240104
50. Шумакова Н.Б., Щебланова Е.И., Сорокова М.Г. «Климат в классе» — стандартизация русскоязычной версии модифицированного опросника «Школьный климат» // Психология. Журнал Высшей школы экономики. 2023. Т. 20. № 2. С. 231–256.http://doi. 10.17323/1813-8918-2023-2-231-256
51. Gu L., Liu O.L., Xu J., Kardanova E., Chirikov I., Li G., Hu S., Yu N., Ma L., Guo F., Su Q., Shi J., Shi H., Loyalka P. Validating the Use of Translated and Adapted HEIghten® Quantitative Literacy Test in Russia, in: Assessment of Learning Outcomes in Higher Education. Cross-National Comparisons and Perspectives. Springer, 2018. doi Ch. 13. P. 267–284. http://doi.org/10.1007/978-3-319-74338-7_13
52. Mielenz T., Edwards M., Callahan L. Item Response Theory Analysis of Two Questionnaire Measures of Arthritis-Related Self-Efficacy Beliefs from Community-Based US Samples // Arthritis. 2010. 416796. 10.1155/2010/416796.
53. Schivinski B., Brzozowska-Woś M., Buchanan E.M., Griffiths M.D., Pontes H.M. Psychometric assessment of the Internet Gaming Disorder diagnostic criteria: An Item Response Theory study // Addictive Behaviors Reports, 2018 Jun 30; 8: 176–184. doi: 10.1016/j.abrep.2018.06.004
54. Shaw A., Liu O.L., Gu L., Kardanova E., Chirikov I., Li G., Hu S., Yu N., Ma L., Guo F., Su Q., Shi J., Shi H., Loyalka P. Thinking critically about critical thinking: validating the Russian HEIghten® critical thinking assessment // Studies in Higher Education. 2020. Vol. 45. No. 9. P. 1933–1948. http://doi.org/10.1080/03075079.2019.1672640
55. Tyumeneva Y.A., Kardanova E., Kuzmina Y. Grit: Two Related but Independent Constructs Instead of One. Evidence from Item Response Theory // European Journal of Psychological Assessment. 2019. Vol. 35. No. 4. P. 469–478. http://doi.org/10.1027/1015-5759/a000424

Руководство по стандартизации психодиагностического инструментария: требования и оценка качества

Учебное пособие

Дизайнер – *М.Ю. Степаненкова*

Компьютерная верстка – *М.В. Мазоха*

Подписано в печать: 27.05.24. Формат: 60x90/16.
Гарнитура Time. Усл. печ. п. 3,0. Усл.-изд. л. 2,8.
Электронное издание. Печать по требованию.

Московский государственный психолого-педагогический университет
127051, г. Москва, ул. Сретенка, д. 29