



УДК 316.77+808.5+001.891.57+371.279

Научная статья / **Research Full Article**DOI: [10.15293/2658-6762.2504.11](https://doi.org/10.15293/2658-6762.2504.11)Язык статьи: русский / **Article language: Russian**

Проблема обеспечения аутентичности и надежности LSP-тестов в рамках интегративной модели валидации: тематический обзор

Е. А. Балыгина¹, Ю. В. Яровикова¹, А. В. Гузова¹¹ Московский государственный психолого-педагогический университет, Москва, Россия

Проблема и цель. В статье представлен обзор научной литературы по проблеме обеспечения аутентичности и надежности LSP-тестов (Language for Specific Purposes) в рамках интегративной модели валидации. Исследуются актуальные вопросы валидации LSP-тестов, выявляются основные проблемы, связанные с их низкой валидностью, подчеркивается важность интеграции требований профессионального контекста в процесс тестирования. Цель статьи заключается в систематизации теоретико-методологических оснований интегративной модели валидации и выявлении условий преодоления ключевого противоречия между аутентичностью и надежностью LSP-тестов.

Методология. Исследование носит обзорный характер и основано на систематическом анализе отечественных и зарубежных работ, посвященных валидации LSP-тестов. Используются методы сравнительно-сопоставительного анализа, синтеза и обобщения данных.

Результаты. Основные результаты заключаются в следующем: установлены основные принципы интегративного подхода к валидации LSP-тестов (контекстуальная детерминированность валидности, социально обусловленный характер тестового конструкта и непрерывность процесса валидации); выявлено ключевое преимущество интегративного подхода, заключающееся в гибкости и адаптивности, позволяющее учитывать динамику профессиональной среды; определены ограничения, такие как неопределенность конструкта, субъективность оценивания и высокая ресурсоемкость. Для преодоления противоречия между аутентичностью и надежностью предложены методологические условия: четкое определение границ конструкта, разработка эмпирически обоснованных критериев оценки и стандартизация процедуры оценивания с учетом требования адаптивности к профессиональному контексту. Авторы отмечают, что унификация эмпирически подтвержденной системы дескрипторов в рамках контекстно-ориентированного подхода создает условия для преодоления противоречия между требованиями аутентичности и стабильности измерений.

Заключение. В заключении делается вывод о том, что интегративная модель валидации предоставляет теоретико-методологическую основу для создания LSP-тестов, которые сочетают высокую надежность с аутентичностью. Это открывает новые перспективы

Библиографическая ссылка: Балыгина Е. А., Яровикова Ю. В., Гузова А. В. Проблема обеспечения аутентичности и надежности LSP-тестов в рамках интегративной модели валидации: тематический обзор // Science for Education Today. – 2025. – Т. 15, № 4. – С. 246–272. DOI: <http://dx.doi.org/10.15293/2658-6762.2504.11>

✉ Автор для корреспонденции: А. В. Гузова, sashenka_0879@mail.ru

© Е. А. Балыгина, Ю. В. Яровикова, А. В. Гузова, 2025

для разработки инструментов оценки, соответствующих современным требованиям профессиональной коммуникации.

Ключевые слова: LSP-тесты (*Language for Specific Purposes*); интегративная модель валидации; аутентичность тестирования; надежность языковых тестов; стандартизация критериев оценки; профессионально ориентированная коммуникация; измерительные инструменты; методология языкового тестирования.

Постановка проблемы

В эпоху глобализации знание английского языка для специальных целей (ESP) становится одним из основных условий способности специалиста к профессиональному развитию и активному взаимодействию с зарубежными коллегами. Снятие языковых барьеров расширяет круг его возможностей для трудоустройства и получения дополнительного образования за рубежом.

Актуальность проблемы совершенствования языковой подготовки, направленной на развитие навыков профессионально-ориентированной коммуникации, обуславливает необходимость поиска новых форм оценки уровня английского языка для специальных целей, измеряющих подготовленность специалиста к выполнению профессиональных задач с использованием английского языка. Переход от широкой лингвистической ориентированности тестов, измеряющих уровень общего английского языка (General English), к узкой профессиональной направленности ESP-тестов потребовал от разработчиков измерительных инструментов решения целого ряда проблем, связанных с необходимостью совершенствования методики разработки и применения языковых тестов в образовательном и профессиональном контексте. В частности, перед разработчиками ESP-тестов стоит за-

дача добиться соответствия структуры и содержания теста целевым потребностям тестируемых (target needs). Как отмечают Т. Hutchinson и А. Waters¹, ESP – это не просто «английский язык плюс терминология», а системный подход к обучению, требующий глубокого анализа потребностей обучающихся (needs analysis), которые должны сформировать способность эффективного решения профессиональных задач на английском языке.

Очевидно, что валидизация ESP-тестов, направленная на подтверждение связи содержания теста с целевыми потребностями тестируемых, должна фокусироваться не только на объективных характеристиках теста (таких как статистическая надежность и содержательная релевантность), но и на динамике его взаимодействия с соответствующим образовательным или профессиональным контекстом. Традиционные методы валидации языковых тестов рассматривают тестирование как изолированный акт, где взаимодействие «тестируемый – тест» не выходит за рамки ситуации проведения теста. Однако в реальности применение различных измерительных инструментов, служащих средством дифференциации испытуемых по определенным признакам, может оказывать значительное влияние на образовательные и профессиональные траектории развития этих людей. Более того, участники

¹ Hutchinson T., Waters A. English for specific purposes: a learning-centred approach. – Cambridge: Cambridge

University Press, 1987. – 183 p. URL: http://assets.cambridge.org/9780521318372/excerpt/9780521318372_excerpt.pdf



тестирования и другие пользователи теста, различающиеся по уровню образования и осведомленности о содержании и целях теста, формируют собственное представление о справедливости и обоснованности существующих способов интерпретации его результатов. Такая обратная связь и социальные последствия внедрения тестирования в образовательную и профессиональную среду должны приниматься во внимание при подтверждении валидности ESP-теста как надежного инструмента оценки уровня сформированности англоязычной коммуникативной компетенции.

Современные исследования² [1–5] подчеркивают, что тест для специальных целей – это не изолированный инструмент измерения, а элемент экосистемы профессиональной подготовки, где валидность формируется через взаимодействие теста, тестируемого и общества. Валидизация LSP-теста должна включать получение эмпирических данных, доказывающих, что тестовые результаты предсказывают профессиональные успехи участника в данной профессиональной области. При этом валидизация рассматривается как непрерывный процесс, требующий постоянного мониторинга и подтверждения обоснованности результатов тестирования в различных условиях его проведения.

Современный подход к валидизации языковых тестов, учитывающий их взаимодействие с контекстом, носит *интегративный характер*, акцентируя взаимообусловленность различных аспектов валидности: содержательную валидность (соответствие струк-

туры и содержания теста его целям), конструктивную валидность (связь теста с теоретической моделью измеряемой способности), критериальную валидность (корреляцию тестовых результатов с внешними показателями) и социальную валидность (социальную значимость теста с точки зрения участников тестирования и этичность практики его применения).

Новая модель «унифицированной» валидности базируется на идеях S. Messick, который трактует это понятие как «интегрированное оценочное суждение о степени, в которой теоретические и эмпирические данные подтверждают адекватность и уместность интерпретаций и действий, основанных на результатах тестов»³.

Детальное описание валидизации как процедуры сбора эмпирических доказательств справедливости и обоснованности принимаемых на основе тестирования практических решений представлено в работах M. Kane, в которых он придерживается концепции «основанной на аргументах» валидизации (argument-based validation)^{4, 5}, развивающей идеи S. Messick. В качестве «аргументов», требующих эмпирического подтверждения, выступают такие «суждения», как «результаты теста предсказывают профессиональные успехи участника», «задания теста отражают содержание учебной программы», «на результаты тестирования не влияют условия его проведения» и пр.

В работах C. Chapelle и U. Knoch [6–7] подчеркивается, что условием обеспечения валидности LSP-теста является его *аутентич-*

² Knoch U., Macqueen S. Assessing English for Professional Purposes. – New York: Routledge, 2020. – 211 p.

³ Messick S. Validity. In: Linn R. L. (Ed.) Educational measurement. – New York: American Council on Education and Macmillan, 1989. – P. 13.

⁴ Kane M. An argument-based approach to validity // Psychological Bulletin. – 1992. – Vol. 112. – P. 527–35.

⁵ Kane M. Validation. In: Brennan R. L. (Ed.). Educational measurement. – Washington, DC: National Council on Measurement in Education and the American Council on Education, 2006. – P. 17–64.



ность, понимаемая как связь содержания теста с целевой областью применения языка (TLU domain – Target Language Use domain). Конструирование тестовых заданий должно основываться на детальном анализе особенностей коммуникации в соответствующей профессиональной сфере. При этом обоснование валидности LSP-теста должно включать демонстрацию репрезентативности заданий теста по отношению к целевой области, которая проявляется в их способности отражать реальные ситуации профессионального общения.

Таким образом, на современном этапе развития измерительных инструментов в фокусе внимания их разработчиков оказываются не только объективные характеристики теста, но и его динамические свойства, актуализирующиеся в конкретных условиях тестирования в процессе проведения оценки. В частности, U. Knoch, опираясь на интегративный подход к валидации S. Messick, обосновывает «социально-ориентированную теорию» конструкта, создавая теоретическую базу для систематического анализа методов оценки английского языка для специальных целей с учетом контекста, в котором проводится тестирование.

В рамках данной модели конструкт не является статичной сущностью, изолированной от контекста. Напротив, он формируется и актуализируется во взаимодействии с конкретными условиями тестирования. Критически важным аспектом этой динамики выступает заявленный конструкт (stated construct), который служит «средством коммуникации между различными социальными мирами», участвующими в тестировании⁶.

Заявленный конструкт в исследовании U. Knoch определяется как «публичное лицо» теоретического и операционального конструктов измерительных инструментов, предлагаемых для коммерческого использования. Заявленный конструкт предназначен для того, чтобы в доступном и понятном виде представить потенциальным покупателям объект тестирования и обрисовать направления его практического применения. Например, на главных страницах сайтов компаний-разработчиков тестов размещены следующие заявленные конструкты: “the language proficiency of people who want to study or work where English is used as a language of communication”⁷; “English proficiency for visas, study, registration and work in healthcare”⁸. Эти утверждения строятся по формуле: «широкая способность для определенной цели» (a broad ability for purpose). Представляя информацию о тесте в доступной и краткой форме, такие формулы информируют пользователей о связи соответствующего измерительного инструмента с определенной профессиональной областью.

Связывая абстрактную измеряемую способность с конкретными социальными целями (например, с работой в сфере здравоохранения), заявленный конструкт выступает как неотъемлемый компонент процесса валидации, одной из целей которого является достижение согласия между разработчиками и пользователями теста о возможности его применения для решения практических задач. Понимание валидации как процесса непрерывной коммуникации и социального согласования представлений о тестовом конструкте –

⁶ Macquene S., Pill J. & Knoch U. Language test as boundary object: perspectives from test users in the healthcare domain // Language Testing. – 2016. – Vol. 33 (2). – P. 271–288.

⁷ International English Language Testing System (IELTS, 2018). URL: <https://www.ielts.net/what-is-the-ielts-exam-structure/>

⁸ Occupational English Test (OET, 2019). URL: <http://www.occupationalenglishtest.org/>.

это ключевое отличие модели U. Knoch от традиционных контекстно-независимых психометрических моделей валидности.

U. Knoch подчеркивает, что практика применения теста в значительной степени определяется тем, каким образом этот инструмент воспринимается всеми участниками процесса оценки – экзаменаторами и испытуемыми. В связи с этим исследователь выделяет в значении конструкта такой аспект, как «*воспринимаемый конструкт*» (*perceived construct*)⁹, обозначающий то, каким образом тестовый конструкт представлен в сознании пользователей. Как правило, пользователи воспринимают конструкт через собственную «языковую призму», которая, например, может включать их представление о том, какие именно знания и навыки обуславливают успешность профессионально ориентированной коммуникации. Исследования, изучающие процесс принятия решений в языковом тестировании [8–12], показывают, что выставляемые экзаменаторами оценки зависят от их субъективного восприятия критериев оценивания.

Несмотря на то, что U. Knoch подчеркивает необходимость унифицированного подхода к интерпретации тестовых результатов, в ее модели субъективный опыт пользователей признается конститутивным элементом тестового конструкта. Это означает, что вместо игнорирования субъективных факторов модель интегрирует их в процесс обоснования валидности. Валидность измерительного инструмента неотделима от анализа того, как его конструкт актуализируется в сознании всех участников тестовой ситуации. Если представления пользователей о конструкте (воспринимаемый

конструкт) расходятся с другими его измерениями, это побуждает разработчиков к совершенствованию измерительного инструмента. Например, если экзаменаторы систематически придают большее значение грамматической точности, чем предусмотрено предписанными критериями оценки, это требует либо пересмотра этих инструкций и дополнительного тренинга экзаменаторов, либо изменения операционального конструкта.

Несмотря на то, что контекстно-ориентированная валидизация позволяет добиться соответствия валидируемого инструмента потребностям пользователей, повышая его аутентичность и социальную валидность, зависимость тестового конструкта от контекста может сделать его более «размытым». Так, включение в тест комплексных заданий, моделирующих ситуации общения в профессиональном контексте, позволяет учесть специфику среды, но при этом приводит к возникновению неопределенности относительно того, что именно измеряет тест: иноязычную коммуникативную компетенцию или профессиональные навыки (например, умение интерпретировать результаты медицинских анализов)¹⁰.

«Размытость» тестового конструкта, обусловленная высокой интегрированностью профессиональной компетентности и языковой способности, в свою очередь, может стать причиной снижения надежности тестового инструмента [13]. Например, достаточно трудно выработать единые критерии оценки устной речи тестируемых в условиях, когда их ошибки в использовании терминологии при выполнении теста могут объясняться как незнанием устоявшихся эквивалентов этих тер-

⁹ Knoch U., Macqueen S. Assessing English for Professional Purposes. – New York: Routledge, 2020. – 211 p.

¹⁰ Тихонова Т. В. О проблемах языкового тестирования // Вестник КАСУ. – 2008 – № 2. – С. 164–167.
URL: <http://www.vestnik-kafu.info/journal/14/540>

минов во втором языке, так и неверной интерпретацией их значения в силу недостаточно высокой профессиональной компетентности.

Кроме того, зависимость тестового конструкта от специфики профессиональной среды создает методологические сложности в операционализации измерительных инструментов, под которой понимается перевод абстрактной теоретической модели (например, «способность к профессиональной коммуникации») в конкретные измерительные процедуры. Ключевая проблема заключается в противоречии между необходимостью стандартизации – фундаментальным требованием психометрической надежности – и вариативностью коммуникативных стратегий. Так, если экзаменатор обязан снижать балл за отклонение от шаблона «идеального ответа», теряется ценность адаптивной коммуникации – ключевого навыка в любой профессии¹¹.

Трудности в операционализации конструкта также связаны с динамичностью практики профессионального общения: чем точнее тест репрезентирует существующий профессиональный контекст, тем быстрее он утрачивает актуальность. Эта дилемма ярко проявляется в таких областях, как медицина, где стремительная цифровизация документооборота требует постоянной корректировки критериев оценки письменной речи, или IT-секторе, где лексикон ежегодно пополняется неологизмами [6].

Следствием динамичности профессиональной среды становится невозможность разработки универсальных оценочных шкал: стандартизированные дескрипторы, соответствующие требованиям определенного временного периода, утрачивают валидность при попытке их длительного применения. Таким

образом, сама природа интегративной валидации как механизма адаптации к динамическим потребностям участников тестирования обуславливает возникновение системного противоречия между аутентичностью теста и психометрической строгостью, ставя под вопрос возможность создания стабильных измерительных инструментов, обеспечивающих воспроизводимость результатов.

У. Кнош описывает конфликт между аутентичностью и надежностью следующим образом: тесты либо становятся надежными «измерителями абстракций», утрачивая связь с практикой профессионально ориентированного общения, либо превращаются в аутентичные «симуляторы реальности», которые в силу низкой стандартизованности не способны обеспечить воспроизводимость результатов [6].

Проведенный анализ теоретических и практических аспектов языкового тестирования в профессиональной сфере демонстрирует, что, несмотря на значительный накопленный опыт в области валидации языковых тестов, ряд ключевых вопросов остается открытым. К их числу относится противоречие между *аутентичностью*, понимаемой как соответствие содержания языкового теста особенностям коммуникации в целевой предметной области, и его *надежностью*, которая достигается посредством стандартизации тестовых заданий, условий тестирования и критериев оценки. Это противоречие проявляется, в частности, в трудностях, возникающих при стандартизации критериев оценки устных речевых навыков в сфере профессионального общения и при моделировании реальных профессиональных ситуаций в условиях тестирования.

¹¹ Тихонова Т. В. О проблемах языкового тестирования // Вестник КАСУ. – 2008. – № 2. – С. 164–167.
URL: <http://www.vestnik-kafu.info/journal/14/540>

В связи с этим нами была сформулирована *научная проблема исследования*, заключающаяся в низкой валидности LSP-тестов, обусловленной трудностями с обеспечением их надежности и аутентичности в профессиональном контексте.

Актуальность настоящей работы обусловлена недостаточной разработанностью вопросов, связанных с источниками низкой валидности тестовых оценок в сфере профессионально ориентированного общения, и потребностью в улучшении качества тестов как инструментов, обеспечивающих достоверную оценку профессиональной языковой компетенции.

Цель статьи заключается в систематизации теоретико-методологических оснований интегративной модели валидации языковых тестов в профессиональном контексте и выявлении условий преодоления ключевого противоречия между требованиями обеспечения аутентичности и надежности LSP-тестов.

Методология исследования

Исследование носит обзорный характер и направлено на систематизацию и критический анализ существующих в отечественной и зарубежной литературе данных о теоретических основаниях и опыте применения интегративного подхода к разработке и валидации LSP-тестов.

Объектом исследования является процесс валидации LSP-теста, предназначенного для оценки уровня сформированности иноязычной коммуникативной компетенции в сфере профессионального общения.

Предметом исследования выступает методика валидации LSP-теста, используемая для подтверждения его аутентичности и соответствия потребностям и социальным запросам тестируемых.

Методологическую основу исследования составил систематический аналитический обзор литературы, посвященной изучению теоретико-методологических оснований интегративной модели валидации LSP-тестов и особенностей ее реализации в профессиональном контексте. Для достижения поставленной цели были применены методы сравнительно-сопоставительного анализа, систематизации, обобщения и синтеза.

В результате систематического поиска и отбора источников в базах данных Scopus и Web of Science были выявлены релевантные научные публикации (статьи, монографии) за период 2001–2024 гг. Выборка исследования включала теоретические работы по проблемам валидации LSP-тестов и публикации, посвященные изучению кейсов разработки тестов для различных профессиональных областей (ОЕТ для медиков, LSTE/Т для правоохранительных органов). Кроме того, были изучены научные работы по теме стандартизации и обеспечения надежности тестирования.

В *задачи* настоящего исследования входило следующее.

1. *Сравнительная оценка методологических рамок традиционных психометрических моделей и современного интегративного подхода к валидации с точки зрения их релевантности для LSP-тестирования.*

Следует отметить, что традиционно валидность трактовалась как стабильное свойство измерительного инструмента, определяемое степенью соответствия содержания теста его теоретическому конструкту (содержательной валидностью) и способностью теста изме-



рять заявленный конструкт (конструктивной валидностью)¹². Кроме того, валидность теста понималась как наличие корреляции между результатами теста и внешним критерием (например, последующим измерением прогнозируемых качеств), что получило название критериальной валидности¹³ [14–19].

В современной тестологии утвердилось унифицированное понимание валидности как интегрального конструкта, объединяющего различные ее аспекты. Согласно этой парадигме, валидность интерпретируется как динамическое свойство измерительного инструмента, зависящее, в частности, от социальных последствий тестирования. Основоположник унифицированного понимания валидности S. Messick указывает, что тест имеет высокий уровень валидности, когда интерпретация его результатов позволяет принимать обоснованные и этически приемлемые решения^{14, 15, 16, 17}. Например, LSP-тест считается валидным, если высокие баллы коррелируют с экспертной оценкой профессиональной коммуникативной компетенции тестируемых и подтверждаются наблюдением за их поведением в профессиональной среде (например, во время презентаций или совещаний). При этом практическая ценность теста находится в прямой зависимости от содержательной и конструктивной валид-

ности: чем точнее определен заявленный конструкт и чем более полно он отражен в заданиях теста, тем выше вероятность того, что решения, принимаемые на основе результатов тестирования, окажутся верными и справедливыми.

Ключевой вклад S. Messick заключается в смещении акцента с объективных характеристик теста и переориентации процесса валидации на анализ последствий тестирования. Он впервые предложил включать в руководства к тестам эмпирически подтвержденные суждения о связи результатов с конкретными практическими решениями. Именно такие суждения формируют «публичное измерение» валидности. Например, подтверждение того, что LSP-тест репрезентативен и прогнозирует профессиональный успех, повышает его воспринимаемую валидность как инструмента для принятия справедливых решений о допуске к профессиональной сертификации.

Аналогично Американская ассоциация исследований в области образования (American Educational Research – AERA) определяет валидность как степень, в которой «эмпирические и теоретические данные подтверждают адекватность интерпретации тестовых результатов с точки зрения их практических последствий»¹⁸. Согласно этому определению,

¹² Cronbach L. J. & Meehl P. E. Construct validity in psychological tests // Psychological Bulletin. – 1955. – Vol. 52 (4). – P. 281–302.

¹³ Муратова И.А. О конструктивной валидности лингводидактических тестов // Социология. – 2016. – № 2. – С. 186–189. URL: <https://www.elibrary.ru/item.asp?edn=xwrcmd&ysclid=m979pnzt4s82466209> EDN: XWRCMD

¹⁴ Chapelle C. A. Validity in language assessment // Annual review of applied linguistics. – 1999. – Vol. 19. – P. 254–272.

¹⁵ Fulcher G. Practical Language Testing. – London: Hodder Education, 2010. – 352 p.

¹⁶ Green P. Green's Medical Symptom Validity Test: User's manual. – Canada: Green's Publishing, 2004. – 100 p.

¹⁷ Popham J. W. The ABCs of Educational Testing: Demystifying the Tools That Shape Our Schools. – California: Corwin, 2017. – 176 p.

¹⁸ American Educational Research Association, American Psychological Association & National Council on Measurement in Education. Standards for educational and psychological testing. – Washington, DC: Author, 2014. – P. 11.

валидность инструмента зависит от его способности выступать в качестве основы для принятия решений об образовательной или профессиональной траектории тестируемых.

Развивая идеи S. Messick, C. Chapelle¹⁹ и U. Knoch [6; 7] выделяют четыре ключевых аргумента (суждения), лежащих в основе валидации:

а) *аутентичность*, проявляющаяся в способности заданий отражать особенности коммуникации в конкретной профессиональной сфере;

б) *обоснованность интерпретации результатов*. Тест должен обладать прогностической валидностью, демонстрируя корреляцию между баллами и результатами деятельности, успешность которой зависит от измеряемой способности;

в) *обоснованность и справедливость решений*. Инструмент должен надежно дифференцировать уровни компетенции, необходимые для успешного выполнения профессиональных задач. При этом пороговые баллы должны быть установлены таким образом, чтобы минимизировать ошибки классификации (такие как ложноположительные или ложноотрицательные результаты);

г) *положительный эффект применения*. Использование теста должно способствовать улучшению академической или профессиональной среды. Например, включение ESP-тестирования в систему сертификации врачей должно оказывать положительное воздействие на качество предоставляемых медицинских услуг. При этом тестирование не должно создавать необоснованных барьеров, связанных с культурной или социальной принадлежностью тестируемых (например, вследствие

включения в тестовые задания культурно-специфичной информации). Практика применения теста должна оказывать положительное влияние на доступность образования или возможность трудоустройства.

2. В числе первостепенных задач исследования – *анализ опыта применения интегративного подхода к валидации LSP-тестов в зарубежной и российской практике* (на примере тестов для физиков, преподавателей и медиков) и *выявление основных методологических подходов к обеспечению аутентичности LSP-тестов* (таких как наблюдение, дискурсивный анализ, анализ литературы или оценка экспертов в целевой профессиональной области).

Исследователи, обсуждая проблему низкой аутентичности существующих измерительных инструментов, указывают на необходимость при разработке критериев коммуникативной компетенции учитывать результаты ее «аутентичной оценки» (*indigenous assessment*), которая проводится не преподавателями-лингвистами, а экспертами в целевой профессиональной области²⁰. Такой подход позволит определить именно те аспекты коммуникативной компетенции, которые существенны для успешного решения профессиональных задач. При этом результаты экспертной оценки и мнение преподавателей-лингвистов о наиболее важных составляющих успешной коммуникации могут не совпадать.

Привлечение экспертов в целевой области может помочь разработчикам LSP-тестов выявить особенности взаимодействия лингвистической и предметной компетенций в процессе профессиональной коммуникации. В частности, «аутентичная оценка» может

¹⁹ Chapelle C., Enright M. K., Jamieson J. M. (Eds.). Building a Validity Argument for the Test of English as a Foreign Language. – New York: Routledge, 2008. – 384 p.

²⁰ Jacoby S. Science as performance: socializing scientific discourse through conference talk rehearsals. – Los Angeles: UCLA PhD thesis, 1998. – 218 p.



предоставить необходимые эмпирические данные для выделения порогового уровня владения языком для специальных целей.

Этот вывод подтверждается рядом исследований, поставивших себе задачу определить те аспекты языковой способности, которые, по мнению профессионалов в соответствующей сфере, обеспечивают успешность их коммуникации друг с другом. Предметом изучения в таких исследованиях становятся особенности использования второго языка специалистами в ситуациях профессионального общения.

В работах Дж. Пилла и Т. Макнамары [20] обсуждается опыт привлечения экспертов из целевой профессиональной области к обсуждению соответствия шкалы оценки специфическим требованиям к коммуникативной компетенции (J. Pill, T. McNamara [20]). В частности, эксперты в области медицины или авиации оценивали минимальный уровень коммуникативной компетенции, необходимый для выполнения таких профессиональных задач, как общение с пациентами или взаимодействие с авиадиспетчерами.

Успешная попытка найти способы повышения валидности теста ОЕТ была предпринята в диссертации Т. Дж. Н. Пилла, называющейся «То, что ценят врачи во время консультаций и каковы последствия специализированного языкового тестирования» («What doctors value in consultations and the implications for specific-purpose language testing»)²¹. Для того чтобы ответить на вопрос, какие именно аспекты коммуникативной компетентности ценят врачи, исследователь проанализировал их отзывы на видеозаписи сцен общения врачей-ординаторов с пациентами.

Т. Дж. Н. Пилл выявил два аспекта коммуникативной компетенции, которые врачи считали особенно важными: *вовлеченность клинициста* (clinician engagement) и *управление взаимодействием* (management of interaction). *Вовлеченность клинициста* в процесс общения подразумевает не только способность сообщить пациенту необходимую информацию, но и умение адаптировать коммуникацию к его потребностям и ожиданиям. *Умение управлять взаимодействием* означает способность продемонстрировать свое понимание состояния пациента и той информации, которую он сообщает. Врач должен уметь задавать вопросы, слушать пациента и реагировать на его ответы.

Результаты исследования Т. Дж. Н. Пилл послужили основой для пересмотра конструкта теста ОЕТ: в его состав были включены прагматические аспекты коммуникации, описанные исследователем как *вовлеченность клинициста* (clinician engagement) и *управление взаимодействием* (management of interaction).

В исследовании К. Элдер, Т. Макнамары и др. [21], которое также привлекло врачей к экспертной оценке наиболее важных аспектов профессиональной коммуникативной компетенции, приводится пример, иллюстрирующий отличия общего языка и языка для специальных целей: врачу в некоторых случаях, не рекомендуется использовать прямые вопросы при общении с пациентом (например, Do you want to harm yourself?). В такой ситуации предпочтительным будет вопрос: “Sometimes when people feel down, they feel like escaping/hurting themselves. Do you ever feel like that?” (C. Elder, T. McNamara, H. Kim, J. Pill, T. Sato [21]).

²¹Pill T. J. H. What doctors value in consultations and the implications for specific-purpose language testing (Unpublished doctoral thesis). – Australia: University of Melbourne, 2013. – 180 p.

Авторы исследования [21] приходят к выводу, что повышение валидности измерительного инструмента стало возможным в результате преодоления его «лингвистической» ориентированности, обусловленной достаточно узким взглядом экспертов-лингвистов на понятие профессиональной коммуникативной компетенции. Привлечение к процессу валидации теста экспертов в области медицины, указавших на наиболее важные аспекты взаимодействия врачей и пациентов, позволило уточнить представление об этом конструкте [21].

В работе В. Ву и Ч. Стэнсфилда [22] была предпринята попытка создать модель коммуникации лиц, подозреваемых в криминальной деятельности, на основе анализа и интерпретации их разговоров, записанных в ходе оперативных мероприятий (W. M. Wu, C. W. Stansfield [22]). Были выявлены такие особенности их речи, как использование коротких предложений, кодированной лексики, частый переход от основной темы к обсуждению «пустых» тем (small talk), смешение языков и др. Анализ и систематизация выявленных особенностей речи позволили разработать задания для теста, оценивающего навыки аудирования кандидатов на должность лингвиста в правоохранительных органах в рамках проекта, известного как Listening Summary Translation Exam in Taiwanese (LSTE/T).

Для того чтобы получить представление об особенностях коммуникации в TLU разработчики теста могут самостоятельно проводить анализ рабочих операций (a job analyses), опираясь на мнение экспертов в соответствующей области или участников тестирования. Например, в работе С. Макквин, Ш. Яхалом и

др.²² проводится изучение требований к навыкам письменной речи медицинских работников в Австралии. Проведя опрос среди специалистов сферы здравоохранения, исследователи смогли выяснить, какие именно формы письменной коммуникации распространены в этой области и особенно важны с точки зрения профессионалов. Было установлено, что несмотря на широкое использование компьютеров, умение создавать рукописный текст остается высоко ценимым качеством в медицинской сфере. Результаты этого исследования подтвердили обоснованность включения в тест ОЕТ (Occupational English Test) заданий, проверяющих умение врача выписывать направление на обследование, рецепты и другие медицинские документы без использования компьютера.

3. Одной из задач исследования также являлось *определение ключевых проблем в применении основных методологических подходов к обеспечению аутентичности LSP-тестов*.

Несмотря на то, что участие экспертов может повысить содержательную валидность измерительного инструмента, у этого подхода есть ряд недостатков. По мнению E. Manias и T. McNamara²³, главная сложность заключается в том, что эксперты, не обладая специальной подготовкой в области лингвистики, часто смещают внимание с коммуникативных навыков испытуемых на их знание соответствующей предметной области. Это приводит к несоответствию между тестовыми результатами и заявленным конструктом, снижая валидность измерительного инструмента.

²² Macqeen S., Yahalom S., Kim H., Knoch U. Exploring writing demands in healthcare settings. – Australia: University of Melbourne, 2012. – 31 p.

²³ Manias E., McNamara T. Standard setting in specific-purpose language testing: What can a qualitative study add? // Language Testing. – 2016. – Vol. 33 (2). – P. 235–249.

В работе S. Jacoby и T. McNamara²⁴ описан практический опыт наблюдения за физиками, взаимодействующими друг с другом в лаборатории и во время ситуаций, моделирующих общение на конференции в ходе представления научных работ с помощью презентаций и их последующего обсуждения. Целью исследования было охарактеризовать устную академическую речь, в которой участвовали как носители родного языка (L1), так и пользователи второго языка (L2). Исследование показало, что внимание участников было сосредоточено преимущественно на том, насколько полно и точно выступающим коллегам удалось донести до слушателей содержание своего доклада.

Оценивая эти выступления, физики, принимавшие участие в опросе, обращали внимание, прежде всего, на такие аспекты устной речи, как ее релевантность, согласованность с визуальной информацией, ясность, краткость, аргументированность и точность. Намного реже участники опроса комментировали лингвистические аспекты выступления. Орфографические и грамматические ошибки упоминались в их ответах только в тех случаях, когда они встречались в письменных материалах презентаций. Если критические замечания в отношении ясности и краткости устной речи коллег все же высказывались, участники опроса отмечали, что эти проблемы не привели к помехам в коммуникации.

Таким образом, исследование S. Jacoby и T. McNamara показало, что в реальных ситуациях профессионального общения в фокусе внимания профессионалов оказывается не знание языка как таковое, а комплекс коммуника-

тивных умений, позволяющих успешно осуществлять совместную профессиональную деятельность.

В качестве ограничения данного исследования следует отметить, что суждения участников, специализирующихся в области физики, об особенностях профессионально ориентированной речи их коллег не всегда могут служить основой для уточнения показателей сформированности коммуникативной компетентности в профессиональной сфере. Очевидно, что их восприятие лингвистической составляющей профессионального общения как незначительной проблемы, которая может быть легко преодолена при условии наличия у выступающего знания предметной области, достаточно субъективно и не отражает действительной роли языковой способности в эффективной коммуникации.

4. В рамках данного исследования ставилась и решалась задача, заключающаяся в *установлении основных факторов, снижающих уровень надежности в профессиональном языковом тестировании.*

В работе Б. Дейгера, К. Ван Горпа и Т. Демистера [23] указывается, что даже при использовании единой системы уровней CEFR результаты тестирования могут значительно различаться в зависимости от методики конструирования теста, способов операционализации конструкта и особенностей администрирования процедуры оценивания (B. Deygers, K. Van Gorp, T. Demeester [23]). В этом исследовании проблема надежности языкового тестирования рассматривается на примере сравнения двух тестов нидерландского языка – ITNA (Интегрированного теста на уровень владения нидерландским языком) и STRT

²⁴ Jacoby S., McNamara T. Locating competence // English for Specific Purposes. – 1999. – Vol. 18 (3). – P. 213–241.

(Стандартного теста на уровень владения нидерландским языком). Оба инструмента предназначены для оценки коммуникативной компетенции в академической и профессиональной сферах. Так как оба теста (ITNA и STRT) используют шкалу дескрипторов CEFR для определения уровня владения языком, предполагалось, что результаты оценивания навыков устной речи, проведенного с их помощью, будут сопоставимы. Однако сравнение тестовых результатов 82 студентов, проходивших тестирование, показало, что вероятность получения количества баллов, соответствующего уровню B2 при использовании ITNA, значительно ниже, чем в случае применения STRT ($p < 0.05$).

Несопоставимость результатов тестирования, вероятно, объясняется различными подходами к операционализации измеряемого конструкта или описанию уровней владения языком при конструировании этих тестов [23]. Кроме того, на результаты тестирования могли повлиять различия в уровне профессиональной подготовки преподавателей, оценивающих речевые навыки участников: экзаменаторы ITNA, как правило, являются опытными преподавателями нидерландского языка, в то время как экспертами STRT становятся студенты, имеющие начальную подготовку в области лингвистики.

Таким образом, несмотря на наличие общей рамки для описания языковых уровней, предоставляемой CEFR, их конкретное содержание может существенно различаться в каждом из этих измерительных инструментов. Кроме того, к снижению конвергентной валидности анализируемых тестов может приводить использование различных методов оценки. В заключение исследователи делают вывод, что не следует ожидать, что применение шкалы CEFR позволит автоматически до-

биться полного совпадения результатов применения различных оценочных инструментов [23]. Для повышения конвергентной валидности этих тестов требуется более глубокий анализ факторов, влияющих на их валидность и надежность.

5. Наконец, исследование ставило перед собой задачу, связанную с *определением методологических условий преодоления ключевого противоречия между требованиями обеспечения аутентичности и надежности LSP-тестов*.

Возникновение противоречия между надежностью и аутентичностью теста обусловлено сложностью стандартизации заданий теста и процедуры оценки в условиях контекстно-ориентированного тестирования. В частности, попытки достижения аутентичности через моделирование сценариев профессионально ориентированного общения увеличивают вариативность в интерпретации критериев оценки [23].

Повышению валидности процедуры оценивания способствует разработка четких критериев оценки и их эмпирическое подтверждение. Например, для оценки коммуникативной компетенции могут использоваться такие критерии, как приемлемость содержания и уместность оцениваемого высказывания (или дискурса) в данной ситуации общения. При этом важно учитывать, что без эмпирического подтверждения связи критериев оценки с профессиональным успехом тестируемых шкала оценок остается абстрактным конструктом, открытым для субъективной интерпретации, что приводит к снижению прогностической валидности теста [13].

Я. Х. Халстийн [24; 25] также указывает на необходимость дополнительного эмпирического подтверждения шкалы дескрипторов, используемых для оценки уровня владения иностранным языком в «Общеввропейской



шкале языковой компетенции» (The Common European Framework of Reference for Languages: Learning, Teaching, Assessment – CEFR). По его мнению, сопоставление результатов тестирования, проведенного на основе CEFR, и эмпирических данных о пороговом уровне владения языком позволило бы уточнить показатели минимального уровня языковой компетенции, необходимого для успешной коммуникации (J. H. Hulstijn [24; 25]).

Попытка эмпирической верификации связи шкалы оценки с уровнем профессиональной коммуникативной компетентности была предпринята в исследовании С. Си [26], в котором было показано, что баллы по тесту TOEFL iBT Speaking могут использоваться для распределения между сотрудниками обязанностей ассистента преподавателя: участники, получившие высокий балл, успешно справляются с такими задачами (teaching assistant assignments), как проведение лекций и семинаров; средний балл указывает на способность тестируемых консультировать студентов по различным вопросам, оказывать помощь при выполнении домашних заданий и т. п.; участники с низким баллом могут быть допущены только к проверке домашних заданий и контрольных работ (X. Xi [26]). Исследователь пришел к выводу о дифференцирующем потенциале теста TOEFL iBT Speaking, установив корреляцию между результатами этого теста и локального теста, разработанного в трех американских университетах для отбора иностранных ассистентов преподавателей (ITA – International Teaching Assistants).

В исследовании С. Си [26] для минимизации ложноположительных результатов использовались ROC-кривые, которые помо-

гают найти точку, где количество ложноположительных результатов минимально, а уровень истинных положительных результатов достаточно высок. Корреляция между результатами TOEFL iBT Speaking и локальными тестами для ИТА подтвердила, что пороговые баллы, определенные методом ROC-анализа, адекватно отражают требования учебных заведений к языковой подготовке преподавателей. Данное исследование демонстрирует механизм решения конфликта между надежностью и аутентичностью теста через синтез статистических методов (таких как корреляционный анализ, метод ROC-кривых) и контекстно-ориентированной интерпретации шкалы оценки. Такой подход повышает прогностическую валидность теста.

В монографии U. Knoch и S. Macqueen²⁵ указываются следующие способы обеспечения надежности результатов тестирования.

1. Включение в тест достаточного количества заданий для оценивания различных аспектов измеряемой способности (для того, чтобы определить, сколько именно заданий должен содержать тест для получения надежного результата, может использоваться теория обобщаемости (G-теория)).

2. Проведение процедуры уравнивания (equating), позволяющей добиться эквивалентности различных версий теста. Результатом уравнивания должно являться то, что баллы по всем версиям теста имеют одно и то же значение. Иначе говоря, должна обеспечиваться «надежность параллельных форм теста» (parallel-forms reliability).

3. Мониторинг согласованности оценок различных экзаменаторов. Для преодоления больших расхождений между выставленными

²⁵ Knoch U., Macqueen S. Assessing English for Professional Purposes. – New York: Routledge, 2020. – 211 p.

оценками должны разрабатываться специальные правила и процедуры. Это позволит добиться согласованности оценок, выставляемых экзаменаторами (inter-rater reliability).

U. Knoch указывает, что валидизация теста должна включать доказательство того, что влияние факторов, не включенных в тестовый конструкт, сведено к минимуму. Результаты тестирования должны быть стабильными при повторных измерениях несмотря на возможное изменение набора заданий, репрезентирующих измеряемый конструкт, состава экзаменаторов и условий тестирования. Иначе говоря, результаты теста не могут быть случайными и зависеть от факторов, не связанных с уровнем знаний и способностей тестируемого²⁶.

Результаты исследования

Основные принципы интегративного подхода к валидизации языковых тестов для специальных целей (LSP)

Проведенный систематический анализ научной литературы позволил выявить основные принципы интегративного подхода к валидизации языковых тестов для специальных целей (LSP).

1. *Контекстуальная детерминированность валидности.* Валидность LSP-теста не существует вне профессионального контекста (TLU domain). Ее доказательство требует эмпирического подтверждения *аутентичности* теста, проявляющейся в соответствии тестовых заданий практике профессионально ориентированного общения, *прогностической валидности*, определяемой как способность те-

стовых результатов прогнозировать профессиональную успешность участников, и *социальной валидности*, заключающейся в обоснованности и справедливости решений, принимаемых на основании тестовых результатов.

2. *Социально обусловленный характер тестового конструкта.* Тестовый конструкт (например, «профессиональная коммуникативная компетенция») рассматривается как результат согласования представлений разработчиков и пользователей о том, что измеряет валидируемый тест и каковы последствия его применения в профессиональном контексте. Подчеркивается роль диалога между разработчиками тестов и пользователями в обеспечении эффективного и этичного использования психометрических инструментов.

3. *Валидизация* понимается как *непрерывный процесс* мониторинга социальных последствий тестирования, получения обратной связи и адаптации к меняющейся профессиональной среде. Динамическая природа валидности исключает возможность ее окончательной верификации, требуя постоянного поиска новых эмпирических данных, подтверждающих способность теста достоверно оценивать измеряемую способность.

Контекстуальная детерминированность интегративной модели валидизации объясняет, почему она оказалась особенно востребованной именно в области LSP-тестирования. В отличие от общего языка, где коммуникативный контекст относительно универсален, оценка уровня иностранного языка для специальных целей требует учета особенностей целевой области применения

²⁶ Мороз Л. С. Методы определения надежности и валидности тестов для контроля знаний // Труды БГТУ. Серия 3: Физико-математические науки и информатика. – 2010. – № 6. – С. 176–179. URL:

<https://www.elibrary.ru/item.asp?id=22884215>
EDN: TGUDXD

языка, не учитываемых в традиционных психометрических подходах.

Преимущества и ограничения интегративного подхода к валидации LSP-тестов

Ключевым преимуществом интегративной модели валидации, выявленным в ходе систематического анализа литературы, является ее *гибкость* и *адаптивность*. Эта адаптивность достигается благодаря систематическому привлечению разнообразных источников данных, таких как экспертные оценки, отзывы работодателей и тестируемых, результаты мониторинга профессиональной деятельности. Ориентация на потребности пользователей теста позволяет сформировать объективное и многомерное представление об измеряемом конструкте, что, в свою очередь, создает методологические предпосылки для достижения *аутентичности* теста.

В рамках интегративной парадигмы происходит переосмысление понятия *аутентичности* теста: аутентичность перестает быть всего лишь статичным мерилем соответствия содержания теста целевой области применения языка, превращаясь в динамический показатель репрезентативности теста в отношении постоянно меняющейся практики профессиональной коммуникации. Например, в тестах для медиков (ОЕТ) такой подход позволяет оперативно интегрировать новые цифровые форматы коммуникации (телеконсультации, электронные истории болезни), обеспечивая тем самым соответствие тестового конструкта актуальным требованиям современной клинической среды.

Обобщение опыта валидации LSP-тестов показало, что применение интегративного подхода позволяет добиться принципиально иного уровня аутентичности. Так, в языковых тестах для медиков (Occupational English Test) задания теперь воспроизводят реальные профессиональные сценарии, вплоть

до стрессовых ситуаций, требующих быстрого принятия решений. В отличие от традиционных тестов, ориентированных на проверку лингвистических знаний, в современных LSP-тестах оценивается, насколько эффективно специалист способен решить профессиональную задачу. Такой подход требует тесного сотрудничества лингвистов с профессионалами конкретной отрасли – пилотами, врачами, инженерами. Взаимодействие с экспертами позволяет выявить коммуникативные потребности, необходимые для успешной профессиональной деятельности.

Однако, как показало исследование, у интегративной модели есть и серьезные *ограничения*.

1. *Неопределенность конструкта*, возникающая вследствие ориентированности на достижение динамической аутентичности. Моделирование комплексных профессиональных сценариев (например, кризисной коммуникации в медицине) в процессе тестирования, хотя и повышает репрезентативность заданий теста, неизбежно приводит к трудностям в определении объекта оценки, обусловленным высокой интегрированностью языковой способности и профессиональной компетентности. Подобная неопределенность в отношении того, что измеряет тест, снижает его конструктивную валидность и усложняет операционализацию, что, в свою очередь, негативно влияет на воспроизводимость результатов при повторных измерениях.

2. *Субъективность в оценивании*, связанная с тем, что даже стандартизированные системы могут допускать вариативность в понимании критериев оценки. Это, в свою очередь, обусловлено комплексным характером заданий, моделирующих ситуации профессионального общения. В тестах, содержащих такие задания, дескрипторы уровней могут включать требования к языковой способности



и к профессиональной компетентности одновременно, что может приводить к расплывчатости критериев оценки. Например, дескриптор продвинутого уровня для оценки эссе языкового теста может указывать на логичность структуры и точное использование профессиональной терминологии в качестве основных критериев оценки.

Размытость критериев может объясняться трудностью разграничения языковой компетенции и профессиональных умений при разработке тестов, ориентированных на профессиональный контекст. Эта неопределенность приводит к субъективной интерпретации критериев оценивания: разные экзаменаторы, руководствуясь одной и той же рубрикой, неосознанно придают различный вес лингвистическим и профессиональным аспектам выполнения задания. Субъективность оценок, в свою очередь, напрямую ведет к снижению надежности: если критерии оценки допускают широкую интерпретацию, то результаты тестирования становятся несопоставимыми.

3. *Ресурсоемкость.* Реализация интегративной валидации требует огромных ресурсов и затрат времени. Необходимо проводить детальный анализ рабочих ситуаций, привлекать экспертов, разрабатывать критерии оценки, которые будут признаны справедливыми как лингвистами, так и профессионалами. Процедуры глубокого анализа TLU, привлечение экспертов и эмпирическое подтверждение обоснованности и справедливости решений, принимаемых на основе тестирования, делают разработку в 3–5 раз дороже традиционных тестов²⁷. Кроме того, обеспечение

аутентичности требует постоянного обновления тестов в ответ на изменения практики профессионального общения.

Методологические условия преодоления противоречия между требованиями аутентичности и надежности LSP-тестов в рамках интегративного подхода

Проведенный анализ успешных кейсов реализации интегративной модели валидации (ОЕТ, LSTE/Т и др.) позволил выявить условия, при которых интегративный подход минимизирует противоречие между аутентичностью и надежностью.

1. *Четкое определение границ тестового конструкта.* Преодолеть неопределенность тестового конструкта, связанную с высоким уровнем интегрированности языковой способности и профессиональной компетентности, можно посредством включения в тестовую спецификацию точного определения измеряемых аспектов коммуникативной компетентности (например, способность уместно использовать терминологические единицы в профессиональном дискурсе, создавать связный текст и т. п.). В спецификацию также может быть включен список профессиональных навыков, которые не должны становиться объектом оценки. Привлечение экспертов в соответствующих профессиональных областях к разработке и валидации языкового теста позволит четко разграничить общую языковую способность и языковые навыки, необходимые для успешной профессиональной деятельности. Кроме того, для статистического подтверждения того, что тест измеряет заявленный конструкт, могут использоваться методы факторного анализа.

²⁷ Knoch U., Macqueen S. Assessing English for Professional Purposes. – New York: Routledge, 2020. – 211 p.

2. *Разработка эмпирически обоснованных критериев оценки.* Эта стратегия предполагает получение эмпирического подтверждения связи критериев оценки с конкретными требованиями к коммуникативной компетентности тестируемых. Например, исследование С. Си [26] продемонстрировало эффективность синтеза статистических методов и контекстного анализа для установления пороговых баллов в тесте TOEFL iBT Speaking. Такой подход позволил снизить субъективность оценивания, трансформируя абстрактные дескрипторы уровней владения языком в описание минимальных требований к языковой компетенции, необходимой для решения определенного круга профессиональных задач.

При этом следует учитывать, что критерии не являются статичными. Их валидность должна постоянно подтверждаться посредством сбора данных о последствиях тестирования (например, о профессиональной успешности тестируемых). Кроме того, необходима постоянная адаптация дескрипторов уровней к актуальным коммуникативным потребностям тестируемых и изменениям в целевой области применения языка.

3. *Стандартизация процедуры оценивания.* Разработка эмпирически подтвержденных критериев оценки должна сопровождаться стандартизацией процедуры оценивания, что предполагает проведение постоянного мониторинга согласованности оценок экзаменаторов (inter-rater reliability). Унифицировать процедуру оценивания также можно посредством детализации дескрипторов и формализации правил оценивания. Например, может быть установлен список параметров, которые нужно учитывать при оценке логичности структуры эссе (таких как, наличие связующих элементов, последовательность изложения и т. п.). При этом могут быть установ-

лены отдельные списки параметров для лингвистической и профессиональной составляющих коммуникативной компетенции. Такой подход позволит снизить вариативность процесса оценивания, вызванную неоднозначностью комплексных заданий, моделирующих ситуации профессионального общения.

Таким образом, преодоление конфликта между требованиями аутентичности и надежности LSP-теста основано на взаимодействии принципов эмпиризма, стандартизации и контекстуализации: критерии оценки должны быть унифицированы, эмпирически обоснованы, учитывать особенности профессионально ориентированного общения и постоянно верифицироваться в динамической профессиональной среде.

В рамках такого подхода стандартизированные критерии отражают не абстрактные представления об измеряемой способности, а эмпирически подтвержденные требования к профессиональной коммуникативной компетенции. Это позволяет преодолеть конфликт между требованиями аутентичности и надежности: единая система дескрипторов, фиксируя существенные параметры профессиональной коммуникации, в условиях контекстно-ориентированной валидации сохраняет адаптивность к вариативным контекстам применения языка. Стандартизация приобретает динамический характер, превращаясь в инструмент сохранения баланса между стабильностью измерений и адаптивностью тестового конструкта.

Заключение

Проведенный анализ научной литературы позволяет сделать следующие ключевые выводы.

1. Интегративный подход к валидации LSP-тестов принципиально меняет традици-

онное понимание качества языкового тестирования, смещая фокус с психометрических показателей на соответствие теста особенностям *целевой области применения языка*. К основным принципам этого подхода можно отнести контекстуальную детерминированность валидности, социально обусловленный характер тестового конструкта и непрерывность процесса валидации.

2. Главное преимущество интегративного подхода заключается в способности обеспечивать *аутентичность и адаптивность* тестовых заданий в условиях динамической профессиональной среды.

3. Основные *ограничения интегративного подхода* обусловлены неоднозначностью тестового конструкта, вызванной взаимосвязанностью языковых и профессиональных аспектов коммуникативной компетенции; сложностями стандартизации оценочных крите-

риев для комплексных заданий, воспроизводящих ситуации реального профессионального взаимодействия.

4. Преодоление противоречия между аутентичностью и надежностью требует реализации *принципов эмпиризма и контекстуализации* при определении границ тестового конструкта и обосновании критериев оценки. Унификация эмпирически подтвержденной системы дескрипторов в рамках контекстно-ориентированного подхода создает условия для преодоления противоречия между требованиями аутентичности и стабильности измерений.

Таким образом, интегративная модель валидации предоставляет теоретико-методологическую основу для создания LSP-тестов, которые, сохраняя психометрическую строгость, являются аутентичными инструментами оценки профессиональной коммуникативной компетенции.

СПИСОК ЛИТЕРАТУРЫ

1. Whyte S. Revisiting Communicative Competence in the Teaching and Assessment of Language for Specific Purposes // Language Education & Assessment. – 2019. – Vol. 2 (1). – P. 1–19. DOI: <https://doi.org/10.29140/lea.v2n1.33.hal-02120612>
2. Бондарева Е. В., Польшина Ю. А., Сорокина В. В. Основные свойства теста в терминосистемах языкового тестирования и оценивания в английском, русском и испанском языках // Вестник Самарского технического университета. Серия: Психолого-педагогические науки. – 2020. – № 1. – С. 20–36. URL: <https://www.elibrary.ru/item.asp?id=42668685> DOI: <http://doi.org/10.17673/vsgtu-pps.2020.1.2>
3. Inozemtseva K. M., Morozova E. V., Kolesnikov I. M. Assessment of ESP students' learning outcomes in a digital learning environment // RUDN journal of informatization in education. – 2022. – № 19 (4). – С. 300–311. URL: <https://www.elibrary.ru/item.asp?id=50144997> DOI: <http://doi.org/10.22363/2312-8631-2022-19-4-300-311>
4. Khosravani M., Rostamian M., Ashraf H. A structural equation modeling of English tests' social and educational consequences: exploring target, leverage, risk, and critical variables // Language Testing in Asia. – 2022. – Vol. 12 (1). – P. URL: <https://languagetestingasia.springeropen.com/articles/10.1186/s40468-022-00177-2> DOI: <https://doi.org/10.1186/s40468-022-00177-2>
5. Ван Ц. Валидность тестовых заданий «Гаокао» на понимание русскоязычного текста // Ярославский педагогический вестник. – 2024. – № 4. – С. 51–61. URL: <https://elibrary.ru/ILLDUX> DOI: <http://dx.doi.org/10.20323/1813-145X-20244-139-51>



6. Knoch U., Chapelle C. Validation of rating processes within an argument-based framework // *Language Testing*. – 2017. – Vol. 35 (4). – P. 477–499. DOI: <https://doi.org/10.1177/0265532217710049>
7. Chapelle C. A. An introduction to language testing's first virtual special issue: Investigating consequences of language test use // *Language Testing*. – 2020. – Vol. 37 (4). – P. 638–645. DOI: <https://doi.org/10.1177/0265532220928533>
8. Eckes T. Operational rater types in writing assessment: Linking rater cognition to rater behavior // *Language Assessment Quarterly An International Journal*. – 2012. – Vol. 9 (3). – P. 270–292. DOI: <http://dx.doi.org/10.1080/15434303.2011.649381>
9. Gauthier G., St-Onge C., Tavares W. Rater cognition: Review and integration of research findings // *Medical Education*. – 2016. – Vol. 50 (5). – P. 511–522. DOI: <https://doi.org/10.1111/medu.12973>
10. Luo L. Pursuing Authenticity in ESP Testing – the need for interdisciplinary collaboration // *The Journal of Teaching English for Specific and Academic Purposes*. – 2019. – Vol. 7 (2). – P. 159–169. DOI: <https://doi.org/10.22190/JTESAP1902159L>
11. Giraldo F. Language Assessment Practices and Beliefs: Implications for Language Assessment Literacy. – 2019. – Vol. 26 (1). – P. 35–61. DOI: <https://doi.org/10.19183/how.26.1.481>
12. Dong M. Structural relationship between learners' perceptions of a test, learning practices, and learning outcomes: A study on the washback mechanism of a high-stakes test // *Studies in Educational Evaluation*. – 2020. – Vol. 64. – P. 100824. URL: <https://www.sci-hub.ru/10.1016/j.stueduc.2019.100824> DOI: <https://doi.org/10.1016/j.stueduc.2019.100824>
13. Davies A. The logic of testing languages for specific purposes // *Language Testing*. – 2001. – Vol. 18 (2). – P. 133–147. DOI: <http://dx.doi.org/10.1177/026553220101800202>
14. Kane M. T. Explicating validity // *Assessment in Education: Principles, Policy and Practice*. – 2015. – Vol. 23 (2). – P. 198–211. DOI: <http://doi.org/10.1080/0969594X.2015.1060192>
15. Shepard L. A. Evaluating test validity: reprise and progress // *Assessment in Education: Principles, Policy & Practice*. – 2016. – Vol. 23 (2). – P. 268–280. DOI: <http://dx.doi.org/10.1080/0969594X.2016.1141168>
16. Green P., Flaro L. Results from three Performance Validity Tests in children with intellectual disability // *Applied Neuropsychology. Child*. – 2016. – Vol. 5 (1). – P. 25–34. DOI: <http://doi.org/10.1080/21622965.2014.935378>
17. Lu W., Zeng Y., Chen J. Proposing a Framework of Validity Evidence for a Score Report // *Creative Education*. – 2021. – Vol. 12 (8). – P. 1912–1925. DOI: <http://doi.org/10.4236/ce.2021.128146>
18. Варакута А. А., Шеломенцев П. Ю., Андриенко Е. В. Определение различных типов системы оценки и оценочных процедур в контексте исторического развития // *Science for Education Today*. – 2021. – № 6. – С. 121–141. URL: <https://www.elibrary.ru/item.asp?id=47447641> DOI: <http://dx.doi.org/10.15293/2658-6762.2106.07>
19. Palmour L. Assessing speaking through multimodal oral presentations: The case of construct underrepresentation in EAP contexts // *Language Testing*. – 2023. – Vol. 41(1). – P. 9–34. DOI: <https://doi.org/10.1177/02655322231183077>
20. Pill J., McNamara T. How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals // *Language testing*. – 2016. – Vol. 33 (2). – P. 217–234. DOI: <https://doi.org/10.1177/0265532215607402>
21. Elder C., McNamara T., Kim H., Pill J., Sato T. Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialist can tell us // *Language and Communication*. – 2017. – Vol. 57. – P. 14–21. DOI: <https://doi.org/10.1016/j.langcom.2016.12.005>



22. Wu W. M., Stansfield C. W. Towards authenticity of task in test development // Language Testing. – 2001. – Vol. 18 (2). – P. 187–206. DOI: <https://doi.org/10.1177/026553220101800205>
23. Deygers B., Van Gorp K., Demeester T. The B2 Level and the Dream of a Common Standard // Language Assessment Quarterly. – 2018. – Vol. 15 (1). – P. 44–58. DOI: <https://doi.org/10.1080/15434303.2017.1421955>
24. Hulstijn J. H. The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency // The Modern Language Journal. – 2007. – Vol. 91 (4). – P. 663–667. DOI: https://doi.org/10.1111/j.1540-4781.2007.00627_5.x
25. Hulstijn J. H. The Common European Framework of Reference for Languages: A challenge for applied linguistics // ITL-International Journal of Applied Linguistics. – 2014. – Vol. 165 (1). – P. 3–18. DOI: <https://doi.org/10.1075/itl.165.1.01hul>
26. Xi X. Validating TOEFL® iBT Speaking and Setting Score Requirements for ITA Screening // Language Assessment Quarterly An International Journal. – 2007. – Vol. 4 (4). – P. 318–351. DOI: <https://doi.org/10.1080/15434300701462796>

Поступила: 04 мая 2025

Принята: 05 июля 2025

Опубликована: 31 августа 2025

Заявленный вклад авторов:

- Е. А. Балыгина (руководитель исследования и основной исполнитель): концепция и дизайн исследования, интерпретация результатов исследования, литературный обзор.
- Ю. В. Яровикова (исполнитель) сбор материалов, литературный обзор, оформление текста статьи.
- А. В. Гузова (исполнитель): автор-корреспондент, сбор материалов, оформление текста статьи.

Все авторы ознакомились с результатами работы и одобрили окончательный вариант рукописи.

Информация о конфликте интересов:

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов в связи с публикацией данной статьи

Информация об авторах

Балыгина Елена Анатольевна

кандидат филологических наук, доцент,
кафедра зарубежной и русской филологии,
Московский государственный психолого-педагогический университет,
123290, г. Москва, ул. Шелепихинская набережная, д. 2А, Россия.
ORCID ID: <https://orcid.org/0000-0002-5558-1389>
SPIN-код: 8671-5242
E-mail: baliginaea@mgppu.ru



Яровикова Юлия Владимировна

кандидат филологических наук, доцент,
кафедра зарубежной и русской филологии,
Московский государственный психолого-педагогический университет,
123290, г. Москва, ул. Шелепихинская набережная, д. 2А, Россия.
ORCID ID: <https://orcid.org/0000-0002-2097-8597>
SPIN-код: 4627-7910
E-mail: yarovikovayuv@mgppu.ru

Гузова Александра Викторовна

кандидат педагогических наук, доцент,
кафедра зарубежной и русской филологии,
Московский государственный психолого-педагогический университет,
123290, г. Москва, ул. Шелепихинская набережная, д. 2А, Россия.
ORCID ID: <https://orcid.org/0000-0001-8335-6528>
SPIN-код: 8315-0290
E-mail: sashenka_0879@mail.ru



The problem of ensuring authenticity and reliability of LSP tests within the framework of an integrative validation model: A thematic review

Elena A. Balygina¹, Yulia V. Yarovikova¹, Alexandra V. Guzova  ¹

¹ Moscow State University of Psychology and Education, Moscow, Russian Federation

Abstract

Introduction. The article presents a review of Scholarly literature on the problem of ensuring the authenticity and reliability of LSP tests (Language for Specific Purposes) within the framework of the integrative validation model. The study examines current issues of LSP test validation, identifies the main problems associated with their low validity, and emphasizes the importance of integrating the requirements of the professional context into the testing process. The purpose of the article is to systematize the theoretical and methodological foundations of the integrative validation model and identify the conditions for overcoming the key contradiction between the authenticity and reliability of LSP tests.

Materials and Methods. The study is of a review nature and is based on a systematic analysis of Russian and international works devoted to the validation of LSP tests. The methods of comparative analysis, synthesis and generalization of data are used.

Results. The outcomes of the study are as follows: the basic principles of the integrative approach to the validation of LSP tests (contextual embeddedness of validity, socially determined nature of the test construct and continuity of the validation process) have been established; the key advantage of the integrative approach has been revealed, which consists in flexibility and adaptability, allowing to take into account the dynamics of the professional environment; the limitations have been identified, such as the vagueness of the construct, subjectivity of assessment and high resource intensity. In order to overcome the contradiction between authenticity and reliability, the following methodological conditions have been proposed: a clear definition of the boundaries of the construct, the development of empirically substantiated assessment criteria and standardization of the assessment procedure taking into account the requirement of adaptability to the professional context. The authors note that the unification of an empirically confirmed system of descriptors within the context-oriented approach creates conditions for overcoming the contradiction between the requirements of authenticity and stability of measurements.

For citation

Balygina E. A., Yarovikova Yu. V., Guzova A. V. The problem of ensuring authenticity and reliability of LSP tests within the framework of an integrative validation model: A thematic review. *Science for Education Today*, 2025, vol. 15 (4), pp. 246–272. DOI: <http://dx.doi.org/10.15293/2658-6762.2504.11>

  Corresponding Author: A. V. Guzova, sashenka_0879@mail.ru

© E. A. Balygina, Yu. V. Yarovikova., A. V. Guzova, 2025

Conclusions. The study concludes that the integrative validation model provides a theoretical and methodological basis for creating LSP tests that combine high reliability with authenticity. This opens up new prospects for developing assessment tools that meet the modern requirements of professional communication.

Keywords

Language for Specific Purposes tests; Integrative validation model; Testing authenticity; Language test reliability; Standardization of assessment criteria; Professionally oriented communication; Measurement tools; Language testing methodology.

REFERENCES

1. Whyte S. Revisiting Communicative competence in the teaching and assessment of language for specific purposes. *Language Education & Assessment*, 2019, vol. 2 (1). pp. 1-19. DOI: <http://dx.doi.org/10.29140/lea.v2n1.33>
2. Bondareva E. V., Polshina Yu. A., Sorokina V. V. Main qualities of test in language testing and assessment terminology systems in English, Russian and Spanish languages. *Bulletin of the Samara State Technical University. Series: Psychological and Pedagogical Sciences*, 2020, vol. 17 (1), pp. 20-36. (In Russian) URL: <https://www.elibrary.ru/item.asp?id=42668685> DOI: <https://doi.org/10.17673/vsgtu-pps.2020.1.2>
3. Inozemtseva K. M., Morozova E. V., Kolesnikov I. M. Assessment of ESP students' learning outcomes in a digital learning environment. *RUDN Journal of Informatization in Education*, 2022, vol. 19 (4), pp. 300-311. (In Russian) URL: <https://www.elibrary.ru/item.asp?id=50144997> DOI: <https://doi.org/10.22363/2312-8631-2022-19-4-300-311>
4. Khosravani M., Rostamian M., Ashraf H. A structural equation modeling of English tests' social and educational consequences: Exploring target, leverage, risk, and critical variables. *Language Testing in Asia*, 2022, vol. 12 (1). URL: <https://languagetestingasia.springeropen.com/articles/10.1186/s40468-022-00177-2> DOI: <https://doi.org/10.1186/s40468-022-00177-2>
5. Wang Ju. Validity of Gaokao tests for understanding the Russian-language text. *Yaroslavl Pedagogical Bulletin*, 2024, no. 4, pp. 51-61. (In Russian) URL: <https://elibrary.ru/ILLDUX> DOI: <http://dx.doi.org/10.20323/1813-145X-2024-4-139-51>
6. Knoch U., Chapelle C. A. Validation of rating processes within an argument-based framework. *Language Testing*, 2017, vol. 35 (4), pp. 477-499. DOI: <https://doi.org/10.1177/0265532217710049>
7. Chapelle C. A. An introduction to language testing's first virtual special issue: Investigating consequences of language test use. *Language Testing*, 2020, vol. 37 (4), pp. 638-645. DOI: <https://doi.org/10.1177/0265532220928533>
8. Eckes T. Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly: An International Journal*, 2012, vol. 9 (3), pp. 270-292. DOI: <http://dx.doi.org/10.1080/15434303.2011.649381>
9. Gauthier G., St-Onge C., Tavares W. Rater cognition: Review and integration of research findings. *Medical Education*, 2016, vol. 50 (5), pp. 511-522. DOI: <https://doi.org/10.1111/medu.12973>
10. Luo L. Pursuing Authenticity in ESP Testing – the need for interdisciplinary collaboration. *The Journal of Teaching English for Specific and Academic Purposes*, 2019, vol. 7 (2), pp. 159-169. DOI: <https://doi.org/10.22190/JTESAP1902159L>



11. Giraldo F. Language Assessment Practices and Beliefs: Implications for Language Assessment Literacy, 2019, vol. 26 (1), pp. 35-61. DOI: <https://doi.org/10.19183/how.26.1.481>
12. Dong M. Structural relationship between learners' perceptions of a test, learning practices, and learning outcomes: A study on the washback mechanism of a high-stakes test. *Studies in Educational Evaluation*, 2020, vol. 64, pp. 100824. URL: <https://www.sci-hub.ru/10.1016/j.stueduc.2019.100824> DOI: <https://doi.org/10.1016/j.stueduc.2019.100824>
13. Davies A. The logic of testing languages for specific purposes. *Language Testing*, 2001, vol. 18 (2), pp. 133-147. DOI: <http://dx.doi.org/10.1177/026553220101800202>
14. Kane M. T. Explicating validity. *Assessment in Education: Principles, Policy and Practice*, 2015, vol. 23 (2), pp. 198-211. DOI: <http://doi.org/10.1080/0969594X.2015.1060192>
15. Shepard L. A. Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy & Practice*, 2016, vol. 23 (2), pp. 268-280. DOI: <http://dx.doi.org/10.1080/0969594X.2016.1141168>
16. Green P., Flaro L. Results from three performance validity tests in children with intellectual disability. *Applied Neuropsychology. Child*, 2016, vol. 5 (1), pp. 25-34. DOI: <http://doi.org/10.1080/21622965.2014.935378>
17. Lu W., Zeng Y., Chen J. Proposing a framework of validity evidence for a score report. *Creative Education*, 2021, vol. 12 (8), pp. 1912-1925. DOI: <http://doi.org/10.4236/ce.2021.128146>
18. Varakuta A. A., Shelomentsev P. Yu., Andrienko E. V. Defining different types of assessment systems and assessment procedures in the context of historical development. *Science for Education Today*, 2021, vol. 11 (6), pp. 121-141. (In Russian) URL: <https://www.elibrary.ru/item.asp?id=47447641> DOI: <http://dx.doi.org/10.15293/2658-6762.2106.07>
19. Palmour L. Assessing speaking through multimodal oral presentations: The case of construct underrepresentation in EAP contexts. *Language Testing*, 2023, vol. 41 (1), pp. 9-34. DOI: <https://doi.org/10.1177/02655322231183077>
20. Pill J., McNamara T. How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals. *Language Testing*, 2016, vol. 33 (2), pp. 217-234. DOI: <https://doi.org/10.1177/0265532215607402>
21. Elder C., McNamara T., Kim H., Pill J., Sato T. Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialist can tell us. *Language and Communication*, 2017, vol. 57, pp. 14-21. DOI: <https://doi.org/10.1016/j.langcom.2016.12.005>
22. Wu W. M., Stansfield C. W. Towards authenticity of task in test development. *Language Testing*, 2001, vol. 18 (2), pp. 187-206. DOI: <https://doi.org/10.1177/026553220101800205>
23. Deygers B., Van Gorp K., Demeester T. The B2 level and the dream of a common standard. *Language Assessment Quarterly*, 2018, vol. 15 (1), pp. 44-58. DOI: <https://doi.org/10.1080/15434303.2017.1421955>
24. Hulstijn J. H. The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, 2007, vol. 91 (4), pp. 663-667. DOI: https://doi.org/10.1111/j.1540-4781.2007.00627_5.x
25. Hulstijn J. H. The common european framework of reference for languages: A challenge for applied linguistics. *ITL-International Journal of Applied Linguistics*, 2014, vol. 165 (1), pp. 3-18. DOI: <https://doi.org/10.1075/itl.165.1.01hul>



26. Xi X. Validating TOEFL® iBT speaking and setting score requirements for ITA screening. *Language Assessment Quarterly An International Journal*, 2007, vol. 4 (4), pp. 318-351. DOI: <https://doi.org/10.1080/15434300701462796>

Submitted: 04 May 2025

Accepted: 05 July 2025

Published: 31 August 2025



This is an open access article distributed under the [Creative Commons Attribution License](#) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. (CC BY 4.0).

The authors' stated contribution:

E. A. Balygina

Contribution of the co-author: (head of the study and main author of the study): concept and design of the study, interpretation of the results of the study, literary review.

Yu. V. Yarovikova

Contribution of the co-author: (author of the study): collection of materials, literary review, formatting the text of the article.

A. V. Guzova

Contribution of the co-author: (author of the study): corresponding author, collection of materials, formatting the text of the article.

All authors reviewed the results of the work and approved the final version of the manuscript.

Information about competitive interests:

The authors declare no apparent or potential conflicts of interest in connection with the publication of this article

Information about the Authors

Elena Anatolyevna Balygina

Cand. Sci. (Philology), Docent,

Associate Professor,

Department of Foreign and Russian Philology,

Moscow State University of Psychology and Education,

2A, Shelepihinskaya Naberezhnaya, 123290, Moscow, Russian Federation.

ORCID ID: <https://orcid.org/0000-0002-5558-1389>

E-mail: baliginaea@mgppu.ru





Yulia Vladimirovna Yarovikova

Cand. Sci. (Philology), Docent,

Associate Professor,

Department of Foreign and Russian Philology,

Moscow State University of Psychology and Education,

2A, Shelepikhinskaya Naberezhnaya, 123290, Moscow, Russian Federation.

ORCID ID: <https://orcid.org/0000-0002-2097-8597>

E-mail: yarovikovayuv@mgppu.ru

Alexandra Victorovna Guzova

Candidate of Pedagogical Sciences,

Docent, Associate professor,

Department of Foreign and Russian Philology,

Moscow State University of Psychology and Education,

2A, Shelepikhinskaya Naberezhnaya, 123290, Moscow, Russian Federation.

ORCID ID: <https://orcid.org/0000-0001-8335-6528>

E-mail: sashenka_0879@mail.ru