УДК 372.881.1
DOI: 10.17223/19996195/48/16

ОТ ASTP ДО ACTFL: РАЗВИТИЕ РЕЙТИНГОВЫХ ШКАЛ ДЛЯ ОЦЕНИВАНИЯ УСТНОЙ РЕЧИ В США

С.Р. Балуян

Аннотация. Современный подход к тестированию устной речи сформировался на основе многолетних исследований, направленных на разработку инструментов для измерения того, насколько человек хорошо владеет языком. Среди этих инструментов наиболее важными являются рейтинговые (оценочные) шкалы. В научном сообществе сложилось мнение, что методика оценивания навыков устной речи с помощью оценочных шкал впервые была разработана в Институте дипломатической службы Госдепартамента США (шкала FSI) в 1950-х гг. При этом возникает вопрос: есть ли свидетельства применения подобной методики до появления этих шкал? И как результаты теоретических и практических исследований способствовали дальнейшему ее развитию? В данном исследовании, основанном на изучении архивных материалов и вторичных источников, делается попытка ответить на поставленные вопросы. Вторая мировая война, а также последующие войны и международные конфликты, в которых участвовали США, потребовали смещения фокуса в обучении иностранным языкам на устную речь, что повлекло за собой смещение акцентов и в языковом тестировании. Оценивание навыков говорения, основанное на субъективном суждении экзаменатора, который мог лишь интуитивно определить, какую оценку поставить тестируемому, не отвечало требованиям времени. Решением проблемы оказалось создание нового поколения измерителей – прямых тестов, снабженных рейтинговыми шкалами, в которых критерии для оценивания были представлены в виде уровневых дескрипторов. Выявлено, что первые рейтинговые шкалы для определения уровня владения устной речью были предложены В. Каулферсом, Ф. Агардом и Г. Данкелом еще в 1940-х гг., за десятилетие до появления шкалы FSI. Краткий исторический обзор развития рейтинговых шкал позволил также проследить их эволюцию, определить современное состояние и предположить, что дальнейшее развитие методики оценивания устной речи будет связано не только с созданием компьютерных и интернет-версий тестов, но и их оцениванием с помощью компьютерных технологий. Сегодня решение таких глобальных задач невозможно без широкого междисциплинарного подхода и совместных усилий высококлассных специалистов из разных стран. Знание и профессиональная оценка событий в истории лингводидактики и языкового тестирования, а также опора на предыдущий опыт могут определить их путь к успеху.

Ключевые слова: история языкового тестирования; устное тестирование; критериальное оценивание; рейтинговые шкалы; оценочные шкалы; шкала FSI; шкала ILR; шкала ACTFL.

Введение

Основой эффективной коммуникации на изучаемом иностранном языке является умение говорить на нем. Развитие навыков устной речи

является важнейшей целью обучения, это то, ради чего большинство людей начинают изучать иностранный язык и на что направляют основные усилия их учителя. Так было не всегда. Трудности на пути к правильной и беглой речи через развитие и комплексное использование множества разнообразных автоматизированных навыков и умений, отсутствие большого прикладного значения часто вынуждали методистов выдвигать на первый план другие цели, такие как чтение и письмо. Тем не менее, роль устной речи в структуре обучения периодически менялась в зависимости от господствующей модели обучения иностранным языкам. Менялись и методы тестирования. В данной работе термин «тестирование» используется в своем широком для него значении в английском языке – «контроль с помощью проверочных заданий».

Понятие «устная речь» включает в себя интегрированные умения: продуктивные умения производить речь в звуковой форме – говорение, и рецептивные умения понимать звучащую речь – аудирование. Ввиду того, что оценивание аудирования не представляет особой трудности и возможно с помощью традиционных тестовых методик, не требующих применения оценочных шкал, которые являются объектом данного исследования, термин «устная речь» часто используется в значении, синонимичном термину «говорение».

В ранней истории языкового тестирования, в ее донаучный период, навыки говорения если и оценивались от случая к случаю, то суждения об их сформированности строились на интуиции, и говорить о какой-либо точности тестовых измерений не приходилось. В современный период устный экзамен занимает значительное место в практике обучения иностранным языкам, а усилия тестологов направлены на повышение объективности при их оценивании. Особенно сложной задачей при тестировании продуктивных видов речевой деятельности в целом и говорения в частности является максимальное освобождение оценки от субъективного мнения экзаменатора, что обеспечило бы повышение надежности тестов, и, следовательно, их общей эффективности. Решением проблемы оказалось создание нового поколения измерителей – уровневых рейтинговых (оценочных, критериальных) шкал, в которых критерии оценки представлены в виде дескрипторов. Работа над их совершенствованием идет уже не одно десятилетие. Они и сегодня находятся в центре внимания тестологов в странах с развитой тестовой культурой.

Отдельные вопросы истории создания и развития рейтинговых шкал рассматривались в работах Д. Барнуэлла, П. Лоу, Г. Солленбергера, Б. Спольски, Г. Фулчера и др., однако пока еще не удалось собрать целостную картину их эволюции.

В научном сообществе сложилось мнение, что первая критериальная шкала, известная как шкала FSI, была создана в 1950-х гг. в Ин-

ституте дипломатической службы Госдепартамента США (Foreign Services Institute, FSI) и предназначалась для оценивания уровня владения языком с помощью тестов в формате устного собеседования. При этом возникает вопрос: есть ли свидетельства применения подобной методики до появления этих шкал? И как результаты теоретических и практических исследований способствовали ее дальнейшему развитию? В данной работе, основанной на изучении архивных материалов и вторичных источников, делается попытка ответить на поставленные вопросы.

Методология

Для решения задач исследования в рамках системного подхода к изучению педагогических явлений (П.К. Анохин, И.В. Блауберг, В.Н. Садовский, Э.Г. Юдин), системно-исторического подхода (М.С. Бургин, Ф.Ф. Королев, А.И. Ракитов) был применен комплекс методов, как общенаучных, так и специальных, свойственных историко-научным исследованиям. Выбор их определялся логикой исследования, что позволило обеспечить объективность и научную достоверность его результатов.

Критический анализ литературы по вопросам зарождения и эволюции методов языкового тестирования позволил собрать научные данные по теме исследования, определить существующие точки зрения на проблему и достижения в данной области. Прежде всего, это работы Л. Бахмана, Д. Барнуэлла, Дж. Кларка, А. Дейвиса, Ф. Хинофотис, Дж. Оллера-мл., Б. Спольски. Российские тестологи также уделяли внимание отдельным аспектам изучаемой проблемы, среди них А.А. Алексеева, Н.И. Башмакова, Н.Н. Кукуева, С.П. Макушева, А. Павленко, И.Ю. Павловская, И.А. Рапопорт, Р. Сельг, И. Соттер, С. П. Суворов. В 1969 г. в Таганроге были опубликованы результаты исследования И.А. Цатуровой «Из истории развития тестов в СССР и за рубежом», большая часть которого посвящена развитию тестовых методов контроля за рубежом, что вполне объяснимо: в XX в. именно в странах Запада, и особенно в США, тесты приобрели большое значение в психологии, педагогике, медицине, промышленном производстве и других областях жизнедеятельности общества. Вопросы истории развития языкового тестирования и, в частности, тестирования устной речи рассматриваются и в других ее работах [1, 2].

Данные были также получены в результате анализа первичных источников – самих рейтинговых шкал для оценивания устных тестов, разработанных на различных этапах развития языкового тестирования.

Нарративный метод был использован для сведения собранных фактов в единое повествование и их интерпретации, исходя из опреде-

ленных причинно-следственных связей, для последующего первичного анализа.

Исторический (историко-генетический) метод позволил рассмотреть предмет исследования в его развитии, помог найти истоки рейтинговых шкал для оценивания при языковом тестировании, выявить преемственность в их создании и использовании, лучше понять современное состояние проблемы и прогнозировать их развитие в будущем.

Сравнительный метод был применен для асинхронного сопоставления рейтинговых шкал в разных временных отрезках.

Для комплексного исследования процессов и фактов в истории развития рейтинговых шкал для оценивания устной речи также были проведены семантико-терминологический анализ, логический анализ, синтез и интерпретация полученной информации.

Исследование и результаты

Экзамен с участием специально подготовленных рейтеров (экспертов), использующих методику оценивания навыков говорения с помощью рейтинговых шкал, как холистических, так и аналитических, сегодня все чаще становится частью коммуникативного обучения иностранным языкам и проверки общего уровня владения им. Рейтинговая шкала представляет собой упорядоченный набор уровневых дескрипторов, т.е. содержит описание умений, характерных для каждого уровня компетенции, определяемой данной шкалой. Методика оценивания с помощью рейтинговых шкал дает возможность не ограничиваться лишь выставлением того или иного балла по итогам устного тестирования, но также позволяет более полно судить об общем уровне владения говорением, дифференцируя умения и навыки по степени сформированности (например, хорошее произношение, достаточно беглая речь, обширная лексика, однако много грамматических ошибок и нарушения в использовании регистра общения). В настоящее время уже сложно представить себе, что языковое тестирование когда-то обходилось без этой методики.

История образовательных институтов США не знает времени, когда не существовало каких-либо форматов оценивания. Языковое тестирование появилось на рубеже XIX и XX вв. с началом использования первых «объективных» тестов в практике обучения современным иностранным языкам. В каждый из периодов истории языкового тестирования доминировала определенная «школа» со своей методикой педагогического измерения. Смена этих методик была детерминирована лингвистическими, историческими и культурными факторами.

Ранняя история языкового тестирования ассоциируется с так называемым грамматико-переводным методом обучения языкам

(Grammar Translation Method), имевшем целью обучить студентов чтению и переводу и создать иллюзию их «эрудированности». Вот почему в тот период устному экзамену не было уделено должного внимания. Напротив, широко распространенный «прямой метод» обучения (Direct method), делавший упор именно на говорении, подразумевал обязательное оценивание устной речи. Однако в то время задача объективного оценивания устной речи считалась неразрешимой, и такие тесты обладали чрезвычайно низкой надежностью. Оценивание говорения основывалось на субъективном суждении опытного экзаменатора, который мог лишь интуитивно определить, какую оценку поставить тестируемому.

Основоположником тестовых технологий и методики использования оценочных шкал для научного изучения индивидуальных особенностей считается английский исследователь Френсис Гальтон, хотя, судя по результатам некоторых исследований, он не был первым, кто разработал систематический метод оценивания личности [3, 4]. Прошло несколько десятилетий, прежде чем эта методика была адаптирована для применения в языковом тестировании.

Первой рейтинговой шкалой для оценивания уровня владения иностранным языком, как было отмечено выше, признается шкала FSI. Однако исследование позволило выявить, что еще в 1940-х гг. У. Каулферс, и позднее Ф. Агард и Г. Данкел предложили подобную методику и даже разработали простейшие шкалы для оценивания навыков говорения.

Как известно, в области военной подготовки обучение иностранным языкам и языковое тестирование всегда являлись важным звеном, поскольку во всех конфликтах и миротворческих операциях умение свободно общаться является жизненно необходимым. В годы Первой мировой войны быстрыми темпами развивалось психологическое тестирование, основные принципы которого были экстраполированы на педагогическое, и, в частности, языковое тестирование. В результате широкое применение получили так называемые объективные тесты с заданиями множественного выбора, установления соответствия, альтернативного выбора и т.д.

Вторая мировая война также оказала значительное влияние на развитие языкового тестирования. Необходимость подготовки кадров, владеющих иностранными языками, стала очевидной, как только было принято решение о вступлении США в войну. По Армейской специализированной программе подготовки кадров (Army Specialized Training Program, ASTP) в 1943–1944 гг. прошли обучение 140 тыс. человек. В рамках этой программы на базе 55 университетов США были подготовлены медицинские работники, инженеры, ученые, дантисты, психологи. Программа также включала обучение иностранным языкам с упором на разговорную речь – на 500 интенсивных языковых курсах велась подготовка по 30 (по другим данным 34) иностранным языкам.

Г. Майрон отмечает, что до 1940-х гг. главной целью обучения иностранным языкам в США оставалось чтение. Американские солдаты не были подготовлены к выполнению своих обязанностей, требующих владения иностранными языками [5]. Армейская программа подготовки кадров, а также последующие войны и международные конфликты, в которых участвовали США, потребовали смещения фокуса на устную речь, т.е. говорение и понимание речи на слух, что повлекло за собой смещение акцентов и в языковом тестировании. Ф. Гиго пишет, что армии нужны были тесты, которые оценивались бы с помощью критериев и позволяли бы определить два положительных уровня владения языком – высший «expert» и достаточный «competent»:

1. Trainees who have satisfied the institutional authorities that they can both comprehend and speak the language as well as a person with the same amount of formal schooling should speak his mother tongue, will be graduated from Term 6 and will be designated on availability reports as expert [6. P. 359] (Слушателям, которые убедили руководство учебного заведения в том, что они понимают и говорят на языке так же, как носители языка с таким же как у них уровнем образования, будет зачтен 6-й семестр и присвоен уровень «высший» – здесь и далее перевод наш. – *С.Б.*).

2. Trainees who have satisfied the institutional authorities that they can readily comprehend the language as spoken by one adult native to another and can speak the language well enough to be intelligible to natives on non-technical subjects of military importance, will be graduated from Term 6 and will be designated on availability results as competent [Ibid. P. 360] (Слушателям, которые убедили руководство учебного заведения в том, что они понимают разговор двух взрослых носителей языка и говорят на языке на военные нетехнические темы так, что их понимает носитель языка, будет зачтен 6-й семестр и присвоен уровень «достаточный»).

Несмотря на простоту и лаконичность, именно эта двухуровневая шкала может претендовать на роль первой рейтинговой шкалы для устных языковых тестов, снабженной критериями оценивания. Тем не менее в практике тестирования она широко не использовалась. В оценивании преобладали традиционные трехбалльная (отлично, хорошо, удовлетворительно) и стобалльная (процентный балл) системы.

В начале 1940-х гг. В. Каулферс предложил методику оценивания навыков чтения, аудирования и говорения с помощью ранжированных критериев, тем самым приблизившись к армейским требованиям к языковому тестированию. Он отмечал, что характер тестовых заданий должен быть таким, чтобы предоставить конкретные и очевидные доказательства готовности испытуемого к общению в реальной жизненной ситуации, когда отсутствие способности понимать и говорить может

стать серьезным препятствием для безопасности или эффективного исполнения своих военных обязанностей [7. P. 137].

Для теста, состоящего из трех частей (получение услуг, запрашивание информации, предоставление информации), Каулферс предложил две категории для оценивания: объем устного высказывания и качество устного высказывания.

Шкала для оценивания объема высказывания выглядела так:

A. Can make known only a few essential wants in set phrases or sentences. – Может выразить только несколько простых желаний с помощью стандартных фраз или предложений.

B. Can give and secure the routine information required in independent travel abroad. – Может предоставлять и получать информацию, самостоятельно путешествуя за границей.

C. Can discuss common topics and interests of daily life extemporaneously. – Может обсуждать ежедневные общие темы и интересы без предварительной подготовки.

D. Can converse extemporaneously on any topic within the range of his knowledge or experience – Может говорить на любые темы в пределах своих знаний и опыта без предварительной подготовки [Ibid. P. 144].

Шкала для оценивания качества устной речи выглядела так:

0. Unintelligible or no response – Ответ непонятный или нет ответа.

A literate native would not understand what the speaker is saying, or would be confused or mislead. – Грамотный носитель языка не понял бы, что говорят, или понял бы неправильно.

1. Partially intelligible – Ответ частично понятный.

A literate native might be able to guess what the speaker is trying to say. The response is either incomplete, or exceedingly hard to understand because of poor pronunciation or usage. – Грамотный носитель языка мог бы предположить, что пытается сказать говорящий. Ответ либо неполный, либо чрезвычайно трудно понять из-за плохого произношения или использования языка.

2. Intelligible but labored – Ответ понятный, но затрудненный.

A literate native would understand what the speaker is saying, but would be conscious of his efforts in speaking the language. The delivery is hesitating, or regressive, but does not contain amusing or misleading errors in pronunciation or usage. – Грамотный носитель языка понял бы говорящего, но заметил бы усилия, которые прилагаются для высказывания. Речь с паузами хезитации, но не содержит забавных или препятствующих пониманию ошибок в произношении или использовании языка.

3. Readily intelligible – Ответ совершенно понятный.

A literate native would readily understand what the speaker is saying, and would not be able to identify the speaker's particular foreign nationality. – Грамотный носитель языка легко понял бы говорящего и

не был бы в состоянии определить, что собеседник иностранец [7. P. 144].

Следующая шкала была предложена для оценивания навыков аудирования:

0 Cannot understand the spoken language. – Не может понять речь говорящего.

1–5 Can catch a word here and there and occasionally guess the general meaning through inference. – Улавливает отдельные слова и от случая к случаю догадывается об общем значении сказанного из контекста.

6–10 Can understand the ordinary questions and answers relating to the routine transactions involved in independent travel abroad. – Может понять простые вопросы и ответы, относящиеся к обычным ситуациям во время путешествия за границу.

11–15 Can understand ordinary conversation on common non-technical topics, with the aid of occasional repetition or paraphrastic restatements. – Может понять простой разговор на общие нетехнические темы, изредка нуждаясь в повторении и перефразировании.

16–20 Can understand popular radio talks, talking-pictures, ordinary telephone conversations, and minor dialectical variations without difficulty. – Без труда может понять популярные радиопередачи, звуковые фильмы, обычный телефонный разговор, а также распознать небольшие диалектные вариации в речи [Ibid. P. 139].

Тест, предложенный Каулферсом, и рекомендуемая им методика оценивания были сходны с тестами, принятыми позже, в следующее десятилетие, для оценивания устной речи. Однако практического применения в то время они так и не получили.

В 1944 г. Ф. Агард и Г. Данкел начали глубокое изучение проблем в области обучения иностранным языкам, включая языковое тестирование, которое стало известно под названием «Чикагское исследование» (Chicago study). Результаты его были опубликованы в 1948 г. Относительно возможности контроля навыков говорения они сделали следующий вывод: «Что касается продуктивных тестов устной речи, мы не обнаружили таковых для общего пользования» [8. С. 55]. Они же сами и попытались восполнить этот пробел, разработав инновационный для того времени тест, измеряющий, по их терминологии, «communicative ability-intelligibility», то есть то, что мы сегодня называем коммуникативными умениями.

Тест имел следующую структуру. Часть первая предлагала серию картинок для описания. Задание второй части под названием «развернутая речь» было сложнее – тестируемый должен был продолжительно говорить на заданную тему без предварительной подготовки.

Для низкого уровня владения языком, например, предлагалось такое задание:

You are talking with a Spanish-speaking person who has never been to the United States. Describe to him the town or city in which you live. – Вы говорите с носителем испанского языка, который никогда не был в Соединенных Штатах. Опишите ему город, в котором живете.

Для стимулирования продолжения речи также задавались дополнительные вопросы:

This person is also interested in what a North American home looks like. Describe to him the home in which you live. – Этот человек также интересуется тем, как обычно выглядит дом в Северной Америке.

Для высокого уровня задание было более сложным:

You have met a young German in Europe who seems to you to have the makings of an outstanding American citizen. You resolve to try to convince him that he should emigrate to the United States. Talk to him about the United States so that you may help him decide whether he would like to come. – Вы встретили молодого немца в Европе и подумали, что он мог бы стать замечательным гражданином Америки. Вы решаете убедить его переехать жить в Соединенные Штаты. Поговорите с ним о США так, чтобы помочь ему решить, хочет ли он этого или нет.

Для стимулирования продолжения речи, при необходимости, также задавались дополнительные вопросы:

Your young German friend is interested in American schools. Describe to him life at the school you attend. – Вашего молодого немецкого друга интересует вопрос об американских школах. Опишите жизнь в школе, которую вы посещаете.

Стимулы для третьей части, проверяющей диалогическую речью, были записаны на фонограф. Сначала тестируемый слышал голос, который задавал вопрос на иностранном языке, затем второй голос, который указывал на родном языке, как на него отвечать:

1-й голос: ¿Cómo está usted? – Как дела?

2-й голос: Tell him you're fine and ask him how he is. – Скажи, что у тебя все хорошо, и спроси, как у него дела.

Для оценивания каждой части теста была разработана шкала с баллами от 0 до 2. Первая и третья части оценивались холистически. Дескрипторы каждого из уровней шкалы соответствовали степени успешности передачи сообщения:

Part 1. Picture series – Часть первая. Серия картинок

2 – Conveys a simple description completely and correctly. – Речь при простых описаниях правильная и полностью понятная.

Conveys the simple description completely and correctly, but elaborates and in so doing makes some error or errors of vocabulary, grammar, or pronunciation – errors which interfere little with the understandability of the utterance. – Речь при простых описаниях полная и правильная, но при развертывании речи допускает ошибку или ошибки в использовании

лексики, в грамматике или произношении, которые не влияют на понимание высказывания.

1 – Conveys the simple description with one or more errors of vocabulary, grammar, or pronunciation, these errors being such as not to interfere with the understandability of simple description. – При простых описаниях допускает одну или больше ошибок в использовании лексики, в грамматике или произношении, которые не влияют на понимание высказывания.

0 – Conveys very little meaning. – В высказывании мало смысла.

Conveys the wrong meaning. – Значение высказывания искажается.

Makes errors which obscure the meaning. – Допускает ошибки, которые делают речь непонятной.

Says nothing – Ничего не говорит [8. С. 57].

Part 3. Conversation – Часть третья. Разговор

2 – Expresses ideas accurately. – Точно передает смысл.

1 – Partially incorrect. – Частично неправильно.

Conveys the correct idea but has one or more errors of grammar. – Смысл передает правильно, но с одной или двумя грамматическими ошибками.

Conveys almost the correct idea, having one or two errors of vocabulary. – Смысл передает почти правильно, но с одной или двумя ошибками в использовании лексики.

0 – Only small part of idea conveyed. – Передает смысл частично.

Wrong idea conveyed. – Смысл передает неправильно.

Not understandable. – Не понятно.

No utterance made – Не высказывается [Там же. С. 59].

Для оценивания второй части экзаменатор использовал четырехкомпонентную аналитическую шкалу, проверяющую беглость речи, владение лексикой, произношение и грамматику:

Fluency – Беглость речи

2 – Speaks smoothly, phrasing naturally according to his thoughts. – Говорит гладко, естественно деля высказывание на смысловые группы в соответствии с речевым намерением.

1 – Occasionally hesitates in order to search for the right word or to correct an error. – От случая к случаю делает паузы в поисках правильного слова или для исправления ошибки.

0 – Speaks so haltingly that it is difficult to understand the thought he is conveying. – Говорит так сбивчиво, что трудно понять смысл сказанного.

Vocabulary – Лексика.

2 – Vocabulary adequate for expressing the ideas he wishes to convey. – Словарный запас достаточен для передачи смысла высказывания.

1 – Manages to convey his ideas in part, but in several instances uses an incorrect word or fails to find any word to use. – Может частично пере-

дать смысл высказывания, однако в некоторых случаях использует неправильное слово или не находит нужного слова.

0 – Cannot communicate his thought because he does not have an adequate vocabulary. – Не может передать смысл высказывания из-за недостатка словарного запаса.

Pronunciation and Enunciation – Произношение и дикция.

2 – Sufficiently approaches native speech to be completely understandable. – Достаточно близко к речи носителя языка и полностью понятно.

1 – Can be understood, though with difficulty, because there are sounds which he does not utter correctly. – Можно понять, но с определенными трудностями, так как некоторые звуки произносит неправильно.

0 – Would not be understood by natives because his pronunciation is so different from theirs. – Носителю языка не было бы понятно, так как его произношение сильно отличается от их произношения.

Grammatical Correctness – Грамматическая правильность.

2 – Speaks correctly with no serious errors in correct grammatical usage. – Говорит правильно без серьезных грамматических ошибок.

1 – Speech is understandable, but there are serious grammatical errors. – Речь понятна, но присутствуют серьезные грамматические ошибки.

0 – Speech is not readily understandable because it is so full of grammatical errors– Речь почти непонятна из-за большого количества грамматических ошибок [8. С. 58].

Тесты для определения уровня практического владения языком Агарда и Данкела, снабженные критериями для их оценивания, были инновационными для того времени. Система критериально-ориентированного тестирования, которую они рекомендовали, была внедрена государственными учреждениями Соединенных Штатов лишь десять лет спустя, когда проявились последствия недостаточного внимания властей к вопросу обучения иностранным языкам. Нехватка кадров, владеющих языками, во время войны с Японией, а затем и с Кореей, а также усиление холодной войны продемонстрировали, насколько важно незамедлительно приняться за решение проблемы. В 1952 г. правительством США был издано постановление о национальной мобилизации и кадрах, на основании которого комиссии по делам гражданской службы было поручено составить список служащих дипломатического корпуса, владеющих иностранными языками [9]. Декану Языковой школы Института дипломатической службы поступило распоряжение из офиса Госсекретаря Дина Ачесона (Dean Acheson) с требованием разработать критерии, которые помогли бы определить уровень языковой подготовки служащих. В служебной записке указывалось, что кри-

терии должны дифференцировать измеряемые уровни от «отсутствие знания» иностранного языка до «полное владение» [10].

Специально созданная для решения этой задачи межведомственная комиссия под руководством декана, доктора Генри Ли Смита (Dr. Henry Lee Smith), приступила к разработке дескрипторов уровней владения языком для последующего тестирования и составления требуемого списка служащих различных ведомств с указанием уровня владения ими иностранным языком. В состав комитета были включены представители различных государственных учреждений, заинтересованных в решении поставленной задачи.

Члены комиссии понимали, что особой проблемой устного тестирования является получение пригодного для анализа образца речи, подлежащего надежному и валидному оцениванию. Такие образцы они получили от сотрудников дипломатического корпуса, владеющих иностранными языками в различной степени. Затем приступили к созданию шкал. Задача была сложная, но шаг за шагом они разработали стандартизованную шестиуровневую шкалу – от 0 («не владеет языком») до 5 («эквивалентный уровню образованного носителя языка»). Эти баллы были соотнесены с кратким описанием умений и навыков, соответствующих каждому уровню.

В 1952 г. критерии прошли апробацию, а в 1956 г. подверглись серьезной переработке. Если в версии 1952 г. описание каждого уровня ограничивалось одним предложением, то в модернизированной версии оно расширилось до 100 слов. И с 1956 г. (или 1957 г., по другим данным) с помощью теста в формате устного собеседования (Oral Interview, OI), снабженного этими оценочными шкалами, начали определять уровень практического владения языком. Несмотря на то, что время от времени в шкалы вносились какие-то поправки, в целом, дескрипторы уровней владения устной речью остались теми же. Как считает Б. Спольски, в истории языкового тестирования тест в формате устного собеседования, созданный в Институте дипломатической службы Государственного департамента США в период с 1952 по 1956 г., занимает почетное место первого инструмента для измерения уровня практического владения языком [11. С. 99]. В 1958 г. устное тестирование для определения уровня практического владения языком стало обязательным для всех чиновников дипломатической службы.

Вскоре предложенная методика стала активно применяться и за пределами Госдепартамента и его подразделений – в Министерстве обороны, Центральном разведывательном управлении, Федеральном бюро расследований, а также в Корпусе мира [12–14].

Стандартизация методики оценивания с помощью четко структурированного теста в формате собеседования, снабженного рейтинговыми шкалами, позволила повысить его объективность. Эта инновация

привела к существенному повышению межрейтерской надежности теста (inter-rater reliability) и, следовательно, общей надежности теста, изначально обладающего высокой валидностью. Тест использовали не только в США. Он получил широкое признание в мире и был известен как *Интервью FSI* (FSI interview), или просто FSI.

Интерес властей к вопросам обучения иностранным языкам и языковым тестам возобновился в 1960-х гг. Одной из причин опять явилась война – на этот раз война во Вьетнаме. В 1968 г. была разработана новая модернизированная шкала усилиями Межведомственного языкового круглого стола (Interagency Language Roundtable, ILR). Под таким названием несколько различных учреждений правительства объединились с целью координации работы и обмена информацией, связанной с владением госслужащими иностранными языками. Шкала содержала дескрипторы базовых уровней во всех четырех видах речевой деятельности – говорении, аудировании, чтении и письме – и была размещена в справочнике для персонала правительства США. Задача определения уровня владения языком госслужащих была, наконец, решена. В 1976 г.у шкалой ILR 1968 г. начали пользоваться в НАТО.

В последующие годы работа над совершенствованием шкалы ILR была продолжена. К 1985 г.у в нее были внесены изменения – включено полное описание уровней «плюс» в систему оценивания. Были добавлены уровни 0+, 1+, 2+, 3+ или 4+ для более точного оценивания в тех случаях, когда языковые и речевые умения значительно превышают требования одного уровня, но еще не полностью соответствуют критериям следующего уровня. В итоге получилось 11 уровней оценки [14]. Шкалами можно было пользоваться также для оценивания сформированности отдельных навыков, таких как чтение, говорение, аудирование, письмо, перевод, аудиоперевод, устный перевод и межкультурное общение.

С тех пор официальные дескрипторы уровней владения языком известны как «шкала ILR» (ILR Scale) или «Определения ILR» (ILR Definitions). Несмотря на то, что сами тесты в различных правительственных учреждениях отличаются в зависимости от их потребностей, все они пользуются шкалой, которая стала стандартным инструментом измерения уровня практического владения языком.

В 1978 г. Служба тестирования в образовании (Educational Testing Service, ETS) получила грант от Министерства образования США на разработку дескрипторов уровней владения иностранным языком для использования в сферах образования, бизнеса и других. В основу новой шкалы легла та же шкала FSI, и работа по ее модернизации продолжалась вплоть до окончания гранта. Затем исследования были продолжены в Американском совете по обучению иностранным языкам (American Council on the Teaching of Foreign Languages, ACTFL) и в 1981 г.,

наконец, была предложена новая десятиуровневая шкала с усовершенствованными дескрипторами уровней, ставшая известной как шкала ACTFL. Для подготовки экзаменаторов, использующих эту шкалу, были организованы специальные тренинги. С того времени ACTFL и языковые службы правительства США координируют работу двух крупнейших тестовых систем.

В разработку широко известного стандартизованного теста ACTFL *Интервью для определения уровня владения устной речью* (Oral Proficiency Interview, OPI) также были заложены принципы теста *Интервью FSI*. В основном, тест OPI использовался для носителей английского языка, однако с созданием компьютерной версии теста (OPIc) в сотрудничестве с LanguageTesting International (LTI) – подразделением компании Samsung, он стал доступнее и набирает популярность за пределами страны. Тест оценивается сертифицированными экспертами в США по шкалам ACTFL или ILR.

Заключение

Таким образом, в США рейтинговые шкалы для измерения уровня владения устной речью приобрели современный вид в результате многолетних исследований тестологов. Пионерами в этой области традиционно считаются сотрудники Института дипломатической службы при Государственном департаменте и других госучреждений, заинтересованных в определении уровня владения языком своих сотрудников. Разработанные ими в 1950-х гг. и в 1960-х гг. шкалы FSI и ILR получили широкую известность и стали основой для создания других подобных шкал.

Однако изучение архивных материалов и вторичных источников позволило сделать выводы о более раннем использовании методики оценивания уровня владения устной речью с помощью рейтинговых шкал. Идея критериально-ориентированного измерения в лингводидактике зародилась еще в 1940-е гг. и нашла отражение в работах В. Каулферса, а также Ф. Агарда и Г. Данкела. Краткий исторический обзор развития рейтинговых шкал позволил также проследить их эволюцию, определить современное состояние и предположить, что дальнейшее развитие методики оценивания устной речи будет связано не только с созданием компьютерных и интернет-версий тестов, но и их оцениванием с помощью компьютера. Сегодня решение таких глобальных задач невозможно без широкого междисциплинарного подхода и совместных усилий высококлассных специалистов из разных стран. Знание и профессиональная оценка событий в истории лингводидактики и языкового тестирования, а также опора на предыдущий опыт могут определить их путь к успеху.

Литература

1. ***Цатурова И.А.*** Из истории развития тестов в СССР и за рубежом. Таганрог, 1969.
2. ***Цатурова И.А., Балуян С.Р.*** Тестирование устной коммуникации. М. : Высш. шк., 2004.
3. ***Ellson D.G., Ellson E.C.*** Historical note on the rating scale // Psychological Bulletin. 1953. № 50 (5). P. 383–384.
4. ***McReynolds P., Ludwig K.*** On the history of rating scales // Personality and Individual Differences. 1987. № 8 (2). P. 281–283.
5. ***Myron H.*** Teaching French to the Army // The French Review. 1944. № 17 (6). P. 345–52.
6. ***Ghigo F.*** Standardized Tests in the ASTP at the University of North Carolina // The French Review. 1944. № 17 (6). P. 358–360.
7. ***Kaulfers W.V.*** War-time developments in modern language achievement testing // Modern Language Journal. 1944. № 28. P. 136–150.
8. ***Agard F., Dunkel H.*** An investigation of second-language teaching. Chicago : Ginn & Company, 1948.
9. ***Barnwell D.P.*** A History of Foreign Language Testing in the United States: from its beginnings to the present. Tempe, Arizona : Bilingual Press, 1996.
10. ***Stansfield C.W.*** ACTFL Speaking proficiency guidelines. ERIC Digest. Washington, DC : ERIC Clearinghouse on Languages and Linguistics, 1992. Retrieved from ERIC database (ED 347 852).
11. ***Spolsky B.*** Measured words. Oxford : Oxford University Press, 1995.
12. ***Sollenberger H.*** Development and Current Use of the FSI Oral Interview Test // Direct testing of speaking proficiency. Theory and application / Clark L.D. (ed.). Princeton, NJ : Educational Testing Service, 1978. P. 1–12.
13. ***Wilds C.P.*** The Oral Interview Test // Testing Language Proficiency / R.L. Jones, B. Spolsky (eds). Washington, DC : Center for Applied Linguistics,1975. P. 29–38.
14. ***Lowe P.*** The ILR Oral Interview: Origins, Applications, Pitfalls, and Implications // Die Unterrichtspraxis. 1983. № 16 (2). P. 230–244.

Сведения об авторе:
Балуян Светлана Размиковна – доктор педагогических наук, доцент, профессор кафедры лингвистического образования, Южный федеральный университет (Ростов-на-Дону, Россия). E-mail: baluyans@sfedu.ru

Поступила в редакцию 7 ноября 2019 г.

ASTP TO ACTFL: THE DEVELOPMENT OF RATING SCALES FOR ORAL LANGUAGE TESTING IN THE USA

Baluyan S.R., D.Sc. (Education), Professor at the Department of Linguistics, Southern Federal University (Rostov-on-Don). E-mail: baluyans@sfedu.ru

DOI: 10.17223/19996195/48/16

Abstract. The modern approach to oral language testing is the outgrowth of quite a long research and cooperative development of measurement instruments intended to evaluate how well a person speaks a language. Among these instruments the most important are rating (assessment) scales. Most studies seem to agree that the Foreign Services Institute rating scale devised in 1950-s in the USA originates rating scale methodology in oral language testing. The question then arises: is there evidence of the applications of performance rating scales prior to them? And how have they evolved over time? The present study based on data obtained from an archival database and secondary sources attempts to fill this gap and get an in-depth historical understanding of the subject. World War II and posterior wars and interna-

tional conflicts, in which the United States participated, demanded to shift the language teaching curricula toward more aural and oral methods and to focus on speaking ability training which resulted in a shift in testing practices. The routine of testing speaking skills based on intuitive judgement of the experienced examiner did not score the tests accurately and so didn't meet the demands of the time. The problem was solved by introducing a new generation of measuring instruments – direct tests where the test taker's performance was judged against a fixed set of criteria described in rating scales on several ascending levels or bands. This work's originality is rooted in providing evidence of several applications of rating scales earlier in 1940-s by W. Kaulfers, F. Agard and H. Dunkel. A brief chronological account helped to trace back their evolution and detect the current state. It also helped "looking back to look forward" and get some perspectives for their future. We suggest that further development of oral language assessment techniques would be associated not only with the construction of computer and Internet versions of tests, but also their computer-aided assessment. Today, the solution of such global tasks is impossible without a broad interdisciplinary approach and joint efforts of highly qualified specialists from different countries. Knowledge and professional comprehension of events in the history of applied linguistics and language testing, as well as reliance on previous experience, can determine their success.
Keywords: Language testing history; oral testing; criterion referenced assessment; rating scales; assessment scales; FSI scale; ILR scale; ACTFL scale.

Referenses

1. Tsaturova, I.A. (1969) From the history of the development of tests in the USSR and abroad. Taganrogю
2. Tsaturova, I.A., Baluyan, S.R. (2004) Testing oral communication. M.: Higher School.
3. Ellson, D.G. & Ellson, E.C. (1953). Historical note on the rating scale. *Psychological Bulletin, 50*(5), 383-384. http://dx.doi.org/10.1037/h0054149
4. McReynolds, P.& Ludwig, K. (1987). On the history of rating scales. Personality and Individual Differences. 8 (2), 281-283. https://doi.org/10.1016/0191-8869(87)90188-7
5. Myron, H. (1944). Teaching French to the army. *The French Review, 17* (6), 345–352.
6. Ghigo, F. (1944). Standardized Tests in the ASTP at the University of North Carolina. *The French Review, 17*(6), 358-360. Retrieved from http://www.jstor.org/stable/381627
7. Kaulfers, W.V. (1944). War-time developments in modern language achievement testing. *Modern Language Journal, 28*, 136-150.
8. Agard, F and Dunkel, H (1948). *An investigation of second-language teaching.* Chicago: Ginn & Company.
9. Barnwell, D. P. (1996). A *History of Foreign Language Testing in the United States: from its beginnings to the present.* Tempe, Arizona: Bilingual Press.
10. Stansfield, C.W. (1992, September), ACTFL Speaking proficiency guidelines. *ERIC Digest, ERIC Clearinghouse on Languages and Linguistics*, *Washington, DC*. Retrieved from ERIC database (ED 347 852).
11. Spolsky, B. (1995). *Measured words.* Oxford: Oxford University Press.
12. Sollenberger, H, (1978). Development and Current Use of the FSI Oral Interview Test. In L.D. Clark (Ed), *Direct testing of speaking proficiency. Theory and application* (pp. 1-12). Princeton, NJ: Educational Testing Service.
13. Wilds, C.P. (1975). The Oral Interview Test. In R.L. Jones & B. Spolsky (Eds), *Testing Language Proficiency* (pp. 29-38). Washington, DC: Center for Applied Linguistics.
14. Lowe, P. (1983). The ILR Oral Interview: Origins, Applications, Pitfalls, and Implications. *Die Unterrichtspraxis / Teaching German, 16*(2), 230-244. doi:10.2307/3530138

Received 7 November 2019