

УДК 371.13:808.5

Балуян С.Р.

ПОДГОТОВКА ЭКСПЕРТОВ ПО ОЦЕНКЕ УРОВНЯ ВЛАДЕНИЯ УСТНОЙ РЕЧЬЮ

Ключевые слова: языковое тестирование, устный тест, эксперт, рейтинговая шкала.

Надежность и валидность тестовых результатов являются важными количественными характеристиками эффективности тестов. Стремление повысить надежность и валидность тестов являлось движущей силой развития тестовых методов контроля и в области обучения иностранным языкам на протяжении всей его истории.

С появлением коммуникативной теории языка и его применения практика языкового тестирования поднялась на качественно новую ступень. Коммуникативные языковые тесты являются прежде всего практическими тестами, требующими оценивания в процессе выполняемого тестируемым коммуникативного действия – рецептивного, продуктивного либо того и другого. Можно утверждать, что «прямые» коммуникативные тесты являются валидным инструментом и для оценивания владения устной речью. При этом повышение надежности таких тестов является серьезной задачей, решение которой позволит существенно повысить общее качество устных тестов. Одним из решений проблемы обеспечения эффективности тестов устной коммуникации является качественная подготовка экспертов (рейтеров) для их оценивания.

Применение экспертной оценки ранее не поощрялось из-за ее относительной субъективности, когда самым важным в тестировании считалось обеспечение надежности оценки. Сегодня в связи с развитием методов «прямого» тестирования владения иностранным языком введение эксперта в процесс оценивания является необходимым, хотя и проблематичным, так как оценка, получаемая испытуемым, отражает не только качество его ответа, но и профессиональные качества эксперта. Согласно принятой в международном сообществе тестологов терминологии

в области обучения иностранным языкам под «прямым» языковым тестом (direct language test) понимается тест для определения уровня сформированности коммуникативной компетенции в продуктивных видах речевой деятельности – говорении и письме, при выполнении которого тестируемый «продуцирует» речь – письменную или устную, которую оценивают эксперты. К «прямым» тестам относятся тесты в формате интервью, ролевых игр, написания эссе и т.д. При «непрямом» тестировании судить о сформированности компетенций в рецептивных видах речевой деятельности экзаменаторам приходится опосредованно, так как напрямую наблюдать процесс, например, понимания прочитанного или восприятия речи на слух не представляется возможным.

Профессионализм эксперта в большой мере зависит от уровня организации специальной подготовки, во время которой он учится однозначно интерпретировать схемы оценивания, тщательно разработанные специалистами. Эксперт обеспечивается специальными ранжированными шкалами, которые значительно облегчают оценивание и позволяют максимально объективизировать его. В шкалах представлен ряд качественных характеристик, описывающих ту или иную степень проявления коммуникативной компетенции тестируемого, и соответствующий им рейтинг. Такая схема оценивания используется в тестах, задания которых требуют распространенного ответа.

Существует два подхода к измерению при «прямом» коммуникативном тестировании – целостный (холистический) и аналитический. Им соответствуют два типа шкал. Целостная шкала – это набор общих характеристик (дескрипторов) компетенций, соответствующих определенному уровню

сформированности коммуникативной компетенции тестируемого, с помощью которого эксперт оценивает уровень владения языком в целом. Аналитическая шкала – многомерный инструмент эксперта, состоящий из набора шкал, каждая из которых содержит подробные, специфические дескрипторы компетенций, соответствующих определенному уровню сформированности отдельных компонентов коммуникативной компетенции, с помощью которого эксперт оценивает уровень владения данной компетенцией и уровень владения языком в целом. Каждый из указанных подходов к оцениванию находит своих сторонников.

Учитывая многокомпонентность коммуникативной компетенции, уровень ее сформированности можно определить, пользуясь целостной шкалой. При таком подходе тестируемые решают коммуникативные задачи, а эксперты устанавливают степень успешности решения этих задач. Однако при использовании целостных шкал возникает ряд проблем, основными из которых являются проблемы интерпретации результатов, трудности в установлении уровней владения различными компонентами коммуникативной компетенции, проблема «размытости границ» между уровнями, проблема определения относительной значимости компонентов оценивания и др. С помощью целостных шкал невозможно точно определить, что именно отражает тестовый балл. Например, оценка «Коммуникация в основном эффективна» отражает лишь общее впечатление от ответа тестируемого, а не его владение определенными компонентами коммуникативной компетенции. Трудность установления уровня сформированности коммуникативной компетенции заключается и в том, что тестируемый может владеть

ее разными компонентами на разных уровнях (например, при достаточно беглой речи допускать большое количество грамматических ошибок), и тогда при оценивании возникает необходимость отдавать предпочтение одним компонентам в ущерб другим. При этом разные эксперты могут по-разному оценивать «вес» компонентов, вследствие чего результат тестирования окажется искаженным.

Использование аналитических шкал имеет, на наш взгляд, заметное преимущество. Во-первых, они дают возможность представить результат тестирования не просто в виде определенного балла, но также позволяют более полно судить о коммуникативной компетенции испытуемого, дифференцируя его коммуникативные умения. Например, результат тестирования может быть сформулирован следующим образом: «Имеет широкий словарный запас, но постоянно делает грамматические ошибки». Применение аналитических шкал позволяет проводить диагностику, поэтому их иногда называют диагностическими. Вторым преимуществом аналитических шкал является то, что ими могут успешно пользоваться эксперты, не имеющие большого опыта в оценивании «прямых» тестов. Опрос экспертов показал, что даже самые опытные из них при целостном оценивании интуитивно принимают во внимание отдельные специфические компоненты умений и навыков, подвергаемых тестированию. Но, к сожалению, с помощью только аналитических шкал очень сложно выявить общий уровень владения языком, т.е. коммуникативной компетенции тестируемого, состоящей из большого количества разнообразных компонентов и субкомпонентов. Феномен многокомпонентности коммуникативной компетенции создает огромные

трудности для тестологов, которые вынуждены строго ограничивать объекты тестирования и концентрировать свое внимание лишь на некоторых из них ввиду того, что эксперты в состоянии оценить не более трех-четырех критериев сразу. Однако современный взгляд на коммуникативную компетенцию как сложную систему требует учета и других ее важнейших компонентов.

Исследования последних лет показали, что наилучшим решением этой проблемы является сочетание аналитических шкал с целостной шкалой, отражающей общее впечатление экспертов о коммуникативной компетенции тестируемого. Подобная комбинированная шкала была разработана для широко известного в мире американского теста TSE (Test of Spoken English) и применяется для оценивания других устных тестов, составленных на его основе, при этом баллу за общее впечатление придается значительно больший вес, чем баллам по аналитическим шкалам. Этот принцип был реализован и при разработке шкалы для «Теста коммуникативной компетенции в устной речи», который успешно использовался в процессе вступительных испытаний абитуриентов в ТТИ ЮФУ в течение десяти лет.

Участие в тестировании более одного эксперта приводит к повышению надежности теста. Обычно тест оценивают двое экспертов. В некоторых случаях эксперт может и не присутствовать во время тестирования, а оценивать ответы позже на основе видеопленки или аудиозаписи. В этом случае его роль схожа с ролью маркера, который оценивает другие виды тестов, например, бланковые, непосредственно не встречаясь с лицами, выполнившими их. При «прямом» тестировании говорящего роли эксперта и собеседника (интервьюера) часто совмещаются.

Самой сложной и дорогой частью при организации «прямого» тестирования в настоящее время является подготовка экспертов, от которых в значительной мере зависит надежность оценивания. Дело в том, что в подобных тестах главным источником нестабильности результатов выступают оценки экспертов. Надежность «прямых» тестов обычно выражается через: 1) коэффициент согласованности показателей разных экспертов (inter-rater reliability), который определяется как коэффициент корреляции между результатами оценивания двумя или несколькими экспертами одного и того же контингента тестируемых. По данным литературы (Schulz, Strecker, Francis, Shohamy), этот коэффициент считается приемлемым в пределах 0,80–0,90; 2) коэффициент согласованности показателей одного эксперта (intra-rater reliability), который определяется как коэффициент корреляции между результатами оценивания одним и тем же экспертом двух последовательных тестирований одного и того же контингента тестируемых параллельной формой того же теста; 3) коэффициент ретестовой надежности теста (test-retest reliability), который вычисляется как коэффициент корреляции между результатами двух последовательных тестирований одного и того же контингента тестируемых одним и тем же тестом.

Вопрос о тщательной подготовке экспертов для повышения надежности тестов продуктивных умений исследовался в США еще на заре становления языкового тестирования в связи с необходимостью повысить качество оценивания письменных сочинений. Однако наиболее глубокие исследования о влиянии степени подготовленности экспертов на результаты оценивания были связаны с появлением устно-

го теста Института дипломатической службы (FSI). Например, Г. Хеннинг провел эксперимент с участием экспертов, которые не прошли специальную подготовку, а ознакомились со шкалами теста FSI и прослушали записанные образцы собеседования самостоятельно. Коэффициент согласованности показателей разных экспертов составил 0,93 [7]. И. Шохами получила коэффициент согласованности показателей разных экспертов 0,98, проведя подобное собеседование на иврите [8].

В 1979 г. Дж. Фритт опубликовал результаты небольшого, но интересного исследования. Группа участников в течение двух дней обучалась по программе подготовки экспертов и затем оценила большое количество образцов собеседования. Совпадение их оценок с оценками опытных экспертов при оценивании первых восьми работ составило 84%, далее эта цифра выросла до 96%. Еще более впечатляющими были данные, полученные у параллельной группы. Эта группа не участвовала в тренинге экспертов, а изучила схему оценивания теста самостоятельно и прослушала записи с образцами собеседования. Их результаты были почти такими же – 84% при оценивании первых восьми работ и 94% при оценивании остальных. Результаты этого исследования подтвердили возможность адаптации теста FSI к проведению в других учреждениях [4].

В 1970-х гг. носители языка все чаще стали привлекаться для оценивания речи иностранцев. Исследования Г. Гюнтерман привели к выводу, что носители испанского языка достаточно «негативно» реагировали на некоторые ошибки тестируемых, хотя понимали, что они говорят [6]. К. Частайн выяснил, что носители языка меньше всего обращали внимание на

ошибки, связанные с использованием имперфекта/претерита, тогда как эти ошибки были в центре внимания преподавателей при обучении испанскому языку. При этом наиболее строго они оценивали неумение использовать инфинитив после предлога, что не являлось объектом пристального внимания преподавателей. То, что было важно при обучении языку для преподавателей, для носителей языка оказалось не столь существенным [2]. Эти выводы дали материал для дальнейшего исследования проблем, связанных с характером ошибок и оцениванием.

Исследования К. Энц привели к выводу, что грамматические ошибки тестируемых с хорошим произношением оцениваются носителями языка строже, чем такие же ошибки тестируемых с плохим произношением [3].

В большей части экспериментов с так называемыми «наивными» экспертами – носителями языка (naïve native raters) в качестве экспертов принимали участие студенты. Исследования Р. Полицера, а также Р. Ван, Д. Мейер и Ф. Лиренза показали, что такие факторы, как возраст и образование, существенно образом влияют на отношение экспертов к ошибкам. Другими важными факторами являются знание экспертом родного языка тестируемых и степень его заинтересованности в теме разговора [9].

Разные исследования дали различные результаты о надежности оценок опытных экспертов и экспертов-студентов – от большого расхождения в оценках двух групп экспертов [5] до мнения, что «оценки обеих групп экспертов, к удивлению, в целом стабильны» [1].

В настоящее время в странах с развитой тестовой культурой установилась общепринятая процедура подготовки экспертов. Она обычно проходит в фор-

ме учебных занятий в течение нескольких дней, чаще всего по выходным дням. Будущие эксперты отбираются из числа желающих и обычно имеют определенный опыт в области тестовых измерений. Занятия начинаются с ознакомления с тестом и критериями оценивания. Затем обучаемым дают послушать запись ответов, соответствующих всем уровням, представленным на шкале. Эта запись известна как *запись эталонных ответов* (benchmark tape). При ее создании для каждого уровня отбираются те ответы тестируемых, по которым у опытных экспертов не возникает разногласий. Каждый из уровней в записи чаще всего бывает представлен не одним ответом, а несколькими, что дает возможность показать и некоторые вариации в ответах внутри одного уровня.

Далее будущие эксперты практикуются в оценивании образцов разноразрядных ответов тестируемых. Они называют свою оценку вслух, обосновывают принятое решение, сравнивают результаты своего оценивания с результатами других обучаемых и оценками опытных рецензентов. Несоответствия детально фиксируются и обсуждаются, при этом особое внимание уделяется тому, как разрядные дескрипторы и термины интерпретируются разными экспертами. Это помогает им научиться правильно соотносить ответы и критерии. Такая же работа обычно предлагается в качестве домашнего задания, и на следующем занятии обсуждение принятых решений продолжается до тех пор, пока не будут согласованы результаты оценивания всех образцов.

Тренинг заканчивается квалификационным тестом, содержащим такие же задания, и при успешном их выполнении обучаемые получают аккредитацию в качестве экспертов. Вместе

с тем для обеспечения максимальной объективности оценивания при «прямом» тестировании устанавливается постоянный контроль над работой экспертов. Изучение бланков с оценками при расчете коэффициента согласованности показателей разных экспертов помогает выявить возможные ошибки в их работе. Слишком большое расхождение оценок какого-либо эксперта с оценками других анализируется опытными экспертами, и при необходимости данный эксперт проходит дополнительный тренинг.

С развитием техники значительно расширились возможности качественной подготовки экспертов. Среди учреждений, которые активно занимаются исследованиями в области использования инновационных технологий не только для разработки устных тестов и их проведения, но и для профессиональной подготовки экспертов, можно назвать Центр прикладной лингвистики (Center for Applied Linguistics – CAL), основанный в 1986 г. в США. Особый интерес вызывает дистанционный курс тренинга для экспертов, который проводится в режиме онлайн с целью научить их надежно оценивать устную речь (Assessment Training Online – ATOL).

Тренинг начинается с базового курса, который предусматривает знакомство со шкалами ACTFL и включает практическую работу по оцениванию образцов устной речи на их основе. Обучение проводится с помощью материала для чтения, дискуссионных форумов и чатов, домашних заданий, пересылаемых по электронной почте.

Вторая часть курса посвящена подготовке экспертов для оценивания уровня владения устной речью применительно к арабскому или испанскому языкам. Обучаемые также оценивают множество образцов речи и выполняют контрольные задания с целью

определить уровень своей подготовки. Обе части курса были опробованы специалистами, которые пришли к выводу, что дистанционное обучение экспертов – вполне эффективная, удобная и дешевая альтернатива традиционным курсам.

Будущее языкового тестирования связано с надежным и валидным тестированием устной и письменной коммуникации, в том числе и в формате видеоконференции в режиме реального времени. На помощь экспертам в скором времени должны прийти (и работа в этом направлении уже ведется) инновационные технологии и оценивающие компьютерные программы, которые снимут нагрузку с экспертов и позволят максимально объективизировать оценивание «прямых» коммуникативных тестов.

Таким образом, разработка качественных ранжированных шкал для оценивания и качественная профессиональная подготовка экспертов, умеющих однозначно интерпретировать их, являются главными условиями для решения проблемы обеспечения эффективности «прямых» тестов. Тренинг экспертов является самой сложной и дорогой частью организации тестирования. Он чаще всего проходит в форме учебных занятий и заканчивается квалификационным тестом, успешное выполнение которого позволяет получить аккредитацию в качестве эксперта. При этом практическая работа экспертов также должна осуществляться под постоянным контролем. Только при участии высококвалифицированных экспертов возможно максимально объективизировать результаты тестирования устной речи при обучении иностранным языкам.

Литература

1. Callaway D. Accent and the evaluation of ESL oral proficiency // Occasional papers in linguistics,

-
- Southern Illinois University. ERIC EDRS ED 144406. 1977.
2. *Chastain K.* Native-speaker reaction to instructor-identified student second language errors // *Modern Language Journal*. 1980. № 64.
 3. *Ensz K.* French attitudes towards speech errors // *Modern Language Journal*. 1982. № 66.
 4. *Frith J.R.* Testing the FSI testing kit // *ADFL Bulletin*. 1979. № 11.
 5. *Galloway V.* Perceptions of the communicative efforts of American students of Spanish // *Modern Language Journal*. 1980. № 64.
 6. *Guntermann G.* A study of the frequency and communicative effect of errors in Spanish // *Modern Language Journal*. 1978. № 62.
 7. *Henning G.* Oral proficiency testing: Comparative validities of interview, imitation, and completion models // *Language Learning*. 1983. № 33.
 8. *Shohamy E.* Rater reliability of the oral interview speaking test // *Foreign Language Annals*. 1983. № 16.
 9. *Vann R.J., Meyer D.E., Lirenz F.* Error gravity: A study of faculty opinion of ESL errors // *TESOL Quarterly*. 1984. № 18 (3).