Установление стандарта проходного балла:
валидность, надежность, практичность

С.Э.Шевченко,
Е.Ф.Прохорова

Проблема, заявленная в теме данной презентации, рассматривается на примере установлении стандарта проходного балла для рецептивных видов речевой деятельности, уровень владения которыми проверяется в трех разделах Теста второго сертификационного уровня СПбГУ, а именно, «Аудирование», «Чтение» и «Практическое использование языкового материала». Процедура установления стандарта проходного балла проводилась в СПбГУ в рамках международной экспертизы по установлению соответствия Теста второго сертификационного уровня (английский язык) требованиям уровня B2 Общеевропейской шкалы иноязычной компетенции.

Цель презентации – попытаться ответить на вопрос: как проблема валидности, надежности и практичности (важная проблема для любой экзаменационной системы) решается в рамках установления стандарта проходного балла (далее - установление стандарта) для Теста второго сертификационного уровня.

В языковом тестировании существуют различные подходы к рассмотрению данной проблемы и различные методы установления стандарта. В утвержденной Советом Европы процедуре по соотнесению экзамена с Общеевропейской шкалой описаны десять таких методов, при этом в исследовании Фелянки Кафтанджиевой[1] представлено около 60 методов установления стандарта проходного балла, которые использовались только в последние два десятилетия. Сужая рамки этой проблемы, мы остановимся на одном из таких методов, который нашел применение в крупнейших экзаменационных системах, и входит в число методов, рекомендованных Советом Европы[2], а именно, на Голландской версии метода закладок[3].

Но прежде всего, необходимо уточнить значение самого понятия «установление стандарта». Два автора, Cizek and Bunch, предлагают такое определение: установление стандарта порогового балла это «процедура, которая позволяет её участникам, использующим конкретный метод, применить свои оценки таким образом, чтобы перевести политические позиции уполномоченных организаций в определенные позиции на балльной шкале»[4].Таким образом, процедура установления стандарта предполагает принятие решения относительно целевого уровня владения английским языком. В контексте СПбГУ это решение о необходимости для выпускников нашего университета иметь уровень иноязычной коммуникативной компетенции не ниже B2. Опуская политическую часть определения, в данной презентации мы в большей мере сосредоточимся на методе.

1Kaftanjieva F. Methods for Setting Cut Scores in Criterion referenced Achievement Tests, ©EALTA, Cito, Arnhem, 2010.

[2] Council of Europe (2009): Relating Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A Manual.

[3] Оригинальное название - *The Cito variation on the Bookmark method*. *Cito* – аббревиатура названия Голландского института образовательных измерений, поэтому авторы данной статьи сочли возможным дать такой перевод названия этого метода.

[4]Cizek, G.J. and Bunch, M.B. (2007): *Standard Setting: a guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage, P.18.

Голландскую версию метода закладок можно отнести к наиболее передовым методам, которые основываются на применении современной теории тестов IRT. Как известно, основным положением IRT является установление вероятностной связи/зависимости между наблюдаемыми результатами тестирования и латентными параметрами тестируемых и заданий теста.5

В основе Голландской версии метода закладок лежит однопараметрическая логистическая модель (OPLM). Эта модель объединяет математические достоинства однопараметрической модели Раша с характеристиками двухпараметрической логистической модели, которая оценивает не только параметры трудности заданий, но и учитывает коэффициенты дискриминации заданий как заданные константы [6].

Базовым понятием этой модели является понятие способности тестируемого, выраженной в виде латентной переменной, которая отражает соответствие между наблюдаемыми данными ответов тестируемых и теоретическим предположением вероятности правильного ответа.

Как показал опыт экспертизы теста в СПбГУ, использовать Голландскую версию метода закладок могут только те экзаменационные системы, которые готовы к требуемому уровню проведения процедуры установления стандарта. Центр лингводидактического тестирования СПбГУ работал над созданием независимой экзаменационной системы шесть лет, и выбор данного метода для установления стандарта в объективных видах речевой деятельности в рамках международной экспертизы Теста второго сертификационного уровня неслучаен.

В основе всех методов установления стандарта, предлагаемых в процедуре Совета Европы, лежит понятие «пороговый тестируемый» (Borderline Person). В контексте соотнесения Теста второго сертификационного уровня с Общеевропейской шкалой – это студент, который владеет минимально допустимым набором языковых навыков и речевых умений уровня B2 Общеевропейской шкалы. И соответственно тот проходной балл, к которому пришли эксперты в результате установления стандарта, есть количественный показатель минимально допустимого уровня иноязычной коммуникативной компетенции испытуемого для уровня B2.

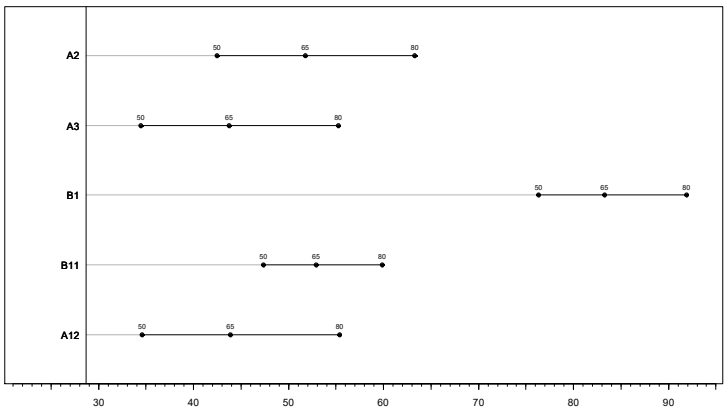
Большая часть методов при анализе тестовых заданий использует дихотомическую модель ответов экспертов на вопрос: может или не может пороговый тестируемый правильно выполнить тестовое задание. По мнению Reckase такое упрощение задачи экспертов может вести к ошибке/погрешности определения проходного балла[7].

Отличительной особенностью Голландской версии метода закладок является использование понятия различных уровней вероятности правильного ответа на тестовое задание. Эксперты использовали шкалу вероятности правильного ответа (от 0,50 до 0,80), которая соотносится с 3 уровнями компетенции порогового тестируемого: (высокий, средний, и низкий (недостаточный). Эти уровни, а также статистические характеристики тестовых вопросов отражены на оценочной карте, «рабочем инструменте» экспертов (Рис.1).

---

[5]Bachman L.F. (1990) Fundamental Considerations in Language testing, Oxford University Press, P.204.
[6]Verhelst, N.D., Glas, C.A.W.&Verstralen, H.H.F.M.(1995) One-Parameter Logistic Model OPLM, Cito, P.1.
[7]Reckase, M. D. (2006a): A Conceptual Framework for a Psychometric Theory for Standard Setting with Examples of Its Use for Evaluating the Functioning of Two Standard Setting Methods. Educational Measurement: Issues and Practice, 2006, 25(2), 4–18.

Рис 1.*Фрагмент оценочной карты 5 тестовых заданий раздела «Аудирование»*



Каждый тестовый вопрос (обозначения вопросов даны по вертикали) в Оценочной карте эксперта представлен в виде отрезка. Его левый конец показывает уровень сложности задания (чем больше отрезок смещен вправо, тем сложнее тестовый вопрос), длина отрезка отражает коэффициент дискриминативности (чем длиннее отрезок, тем меньше дискриминативная способность вопроса).

Эксперты принимали решение, исходя не только из своего видения тестового вопроса, то есть, субъективной оценки, но и учитывая статистические характеристики каждого тестового вопроса, основанные на реальных результатах выполнения этого вопроса тестируемыми.

Таким образом, наличие нескольких уровней вероятности, применение объективных квалиметрических данных, помогающих эксперту принять решение, позволило минимизировать ошибку оценок экспертов, что, несомненно, позволяет говорить о большой валидности, надежности и достоверности стандарта проходного балла.

Другим преимуществом Голландской версии метода закладок является возможность переноса стандарта проходного балла, установленного в рамках экспертизы, на другие версии Теста Второго сертификационного уровня, что говорит о его практичности. Данный метод дает возможность проанализировать ответы тестируемых в различных вариантах Теста на единой шкале латентной переменной способности тестируемых. При этом стандарт остается неизменным, а проходной балл для разных вариантов Теста второго сертификационного уровня будет несколько изменяться. Как мы помним, вероятность успешного выполнения теста зависит как от способности тестируемого, так и от трудности заданий. Не бывает заданий и тестов абсолютно одинаковой сложности, поэтому и проходной балл будет изменяться. Однако изменяться он будет в определенном, достаточно узком диапазоне, поскольку все тестовые вопросы калибруются, то есть анализируются их количественные показатели, что не только позволяет убедиться в том, что тест проверяет то, что запланировано проверить, но и даёт возможность минимизировать ошибку в оценке результатов тестирования.

В соответствии с требованиями процедуры Совета Европы(Council of Europe, 2009) по каждому из трех разделов экзамена проводились раунды установления стандарта (УС) и раунды перекрестной валидизации (ПС). (Для каждого раздела были использованы три версии Теста Второго сертификационного уровня.)

В рамках каждого раунда технология оценивания тестовых заданий по каждому разделу экзамена включала в себя 3 основные операции/процедуры:

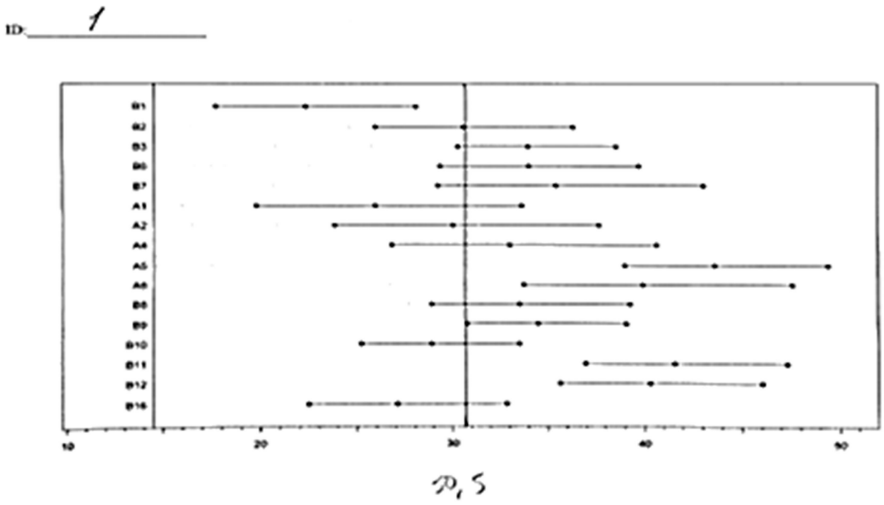
- работая с Буклетом тестовых заданий раздела теста, каждый эксперт анализировал каждый тестовый вопрос, и принимал решение, какой уровень компетенции должен иметь пороговый тестируемый, чтобы дать правильный ответ на данный вопрос. Свое решение эксперты заносили в Оценочную таблицу тестовых вопросов по уровню компетенции тестируемого.

Таблица 1. *Оценочная таблица тестовых вопросов по уровню компетенции тестируемого. Раздел «Чтение»*

	No mastery	Moderate mastery (weak)	Moderate mastery (rather good)	Full mastery
item	< 50%	50% - 65%	65% - 80%	> 80%
A2		+		
A3			+	
B1			+	
B11		+		
A12				+

- После анализа всех вопросов эксперт переносил результаты своего оценивания в «Оценочную карту тестовых заданий», где он мог соотнести свою оценку каждого тестового вопроса со статистическими характеристиками (коэффициент дискриминативности и коэффициент трудности) тестового вопроса.

Рис.2 *Оценочная карта тестовых заданий, заполненная экспертом. Раздел «Чтение».*

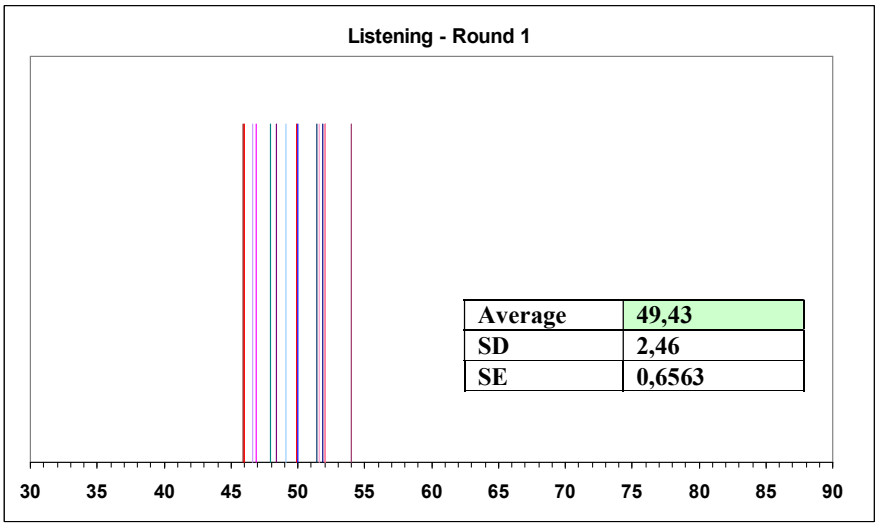


Важно иметь в виду, что результатом работы экспертов был количественный показатель латентной способности тестируемого (по горизонтальной оси на Рис. 2), то есть его компетенции (минимального набора речевых умений и навыков, позволяющих отнести его к уровню В2 ОШИИК) в определенном виде речевой деятельности.

- Среднеарифметическое значение персональных оценок экспертов считалось результатом работы всей группы экспертов в том или ином раунде.

После первого раунда, экспертам представлялась дополнительная информация по среднеарифметическому значению оценки (Average), стандартному отклонению (SD) и стандартной ошибке (SE) группы экспертов в целом. Каждый эксперт, на основе графического изображения, мог проанализировать, насколько его оценка соотносится с результатами работы группы в целом.

График 1. *Распределение оценок независимых экспертов. Раунд 1. Раздел «Аудирование»*

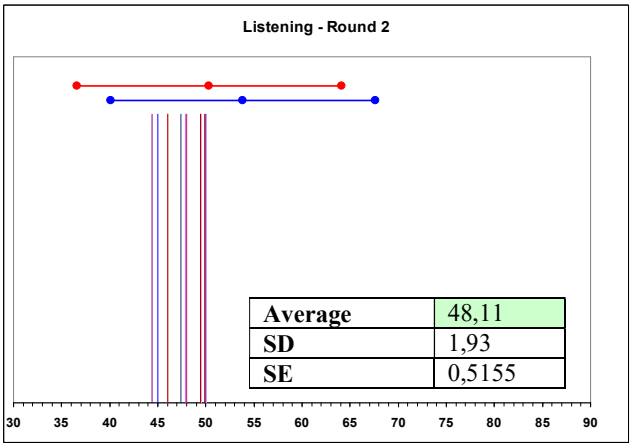


Прежде чем перейти ко второму раунду эксперты имели возможность обсудить возникшие вопросы и результаты оценивания в мини-группах.

После второго раунда экспертам предлагалось проанализировать информациюо соотнесении значений латентной переменной способности порогового тестируемого установленных экспертами и ее средним показателем в двух версиях теста.

Ниже дано графическое изображение этой информации по разделу «Аудирование», которое было предоставлено экспертам. На нем можно увидеть, что среднеарифметическое индивидуальных оценок экспертов (48,11) значительно ниже среднего показателя латентной переменной способности (середина горизонтальных отрезков) тестируемых в двух вариантах теста 2013 г.

График 2.



Основной задачей этапа перекрестной валидизации является определение уровня устойчивости/надежности результатов оценивания в условиях изменения процедуры установления стандарта проходного балла. Этот этап позволил показать, что при смене варианта теста, эксперты продемонстрировали устойчивость и стабильность оценок. А другая группа экспертов (представлявших 4 страны) показала, в целом, схожие результаты, используя другой метод.

Важным показателем внутренней валидности процедуры установления стандарта является характеристика согласованности работы (суждений) экспертов, измеряемая через относительную стандартную ошибку.

Значения этого показателя, представленные в Таблице 2 (раздел «Аудирование») , наиболее ярко иллюстрируют, как согласованность экспертов росла от раунда к раунду, об этом говорит уменьшение относительной стандартной ошибки от 0.048 в раунде 1 до 0.028 в раунде перекрестной валидизации.

Таблица 2.

Exam section	Round 1	Round 2	Round 3	Cross validation Round 1 (CVBM)	Cross validation Round 2 (CVBM)	Cross validation Round (AM)
Listening	**0.048**	**0.038**	**-**	**0.028**	**-**	**0.152**

Как можно заметить, значения относительной ошибки измерений для экспертов, которые работали с методом Ангоффа, несколько выше. Мы предполагаем, что это обусловлено тем, что данный метод не предусматривает использования эмпирических данных (процент выполнения тестовых вопросов), и тем, что у экспертов не было возможности обсуждать результаты со своими коллегами, так как они работали дистанционно. Тем не менее, в целом, можно сделать вывод о довольно высокой точности, согласованности и устойчивости оценок экспертов.

Еще одним показателем внутренней валидности процедуры установления стандарта является точность выбранного метода, измеренная через отношение стандартной ошибки полученного стандарта к стандартной ошибке измерений[8].Считается, что стандартная ошибка установленного стандарта не должна значительно превышать стандартную ошибку измерений. По самым строгим меркам, как отмечает Jaeger,[9] их отношение не должно превышать 0.25. Результаты произведенных расчетов представлены в Таблице 3. Как можно заметить,полученные показатели по всем разделам значительно меньше этой цифры, что позволяет сделать вывод о высокой точности и устойчивости использованного метода.

[8]Council of Europe (2009) Relating Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A Manual.Chapter 7, P.104. http://www.coe.int/t/dg4/linguistic/Publications_EN.asp#P182_5822

[9]Jaeger, R. Selection of Judges for Standard-Setting. // Educational Measurement: Issues andPractice, 1991, 10 (2), 3-6.

Таблица 3.

	Listening		Reading		Use of English	
	round 2	cross val.	round 3	cross val.	round 2	cross val. 2
Standard error of the standard (SE_S)	0.255	0.239	0.230	0.276	0.119	0.129
Standard error of measurement (SE_M)	2.333	2.334	2.189	2.171	2.106	2.189
SE_S/SE_M	**0.109**	**0.102**	**0.105**	**0.127**	**0.057**	**0.059**

Отталкиваясь от контекста соотнесения Теста второго сертификационного уровня с Общеевропейской шкалой иноязычной коммуникативной компетенции, можно подвести такой итог: использование и освоение данного метода не только позволило получить валидные и надежные результаты установления стандарта, а значит и экспертизы, но и дало возможность продолжать использовать самые передовые методы определения проходного балла, которые гарантируют максимальную объективность оценивания.

Список литературы

1. Bachman L.F. (1990) Fundamental Considerations in Language testing, Oxford University Press, P.204.
2. Council of Europe (2001): Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.
3. Council of Europe (2009): Relating Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A Manual.
4. Cizek, G.J. and Bunch, M.B. (2007): *Standard Setting: a guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage, P.18.
5. Kaftanjieva F. (2010) Methods for Setting Cut Scores in Criterion referenced Achievement Tests, ©EALTA, Cito, Arnhem,.
6. Reckase, M. D. (2006): A Conceptual Framework for a Psychometric Theory for Standard Setting with Examples of Its Use for Evaluating the Functioning of Two Standard Setting Methods. Educational Measurement: Issues and Practice, 2006, 25(2), 4□18.
7. Verhelst, N.D., Glas, C.A.W. &Verstralen, H.H.F.M.(1995) One-Parameter Logistic Model OPLM, Cito, P.1.