Тагаева Гульмира Сарыгуловна,
кандидат педагогических наук,
Лаборатория теории и практики оценивания достижений учащихся,
Кыргызская академия образования,
Кыргызская Республика, город Бишкек

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ТЕСТОВЫХ ЗАДАНИЙ В ПРОЦЕССЕ СТАНДАРТИЗАЦИИ ИНСТРУМЕНТАРИЕВ ОЦЕНИВАНИЯ СФОРМИРОВАННОСТИ КОМПЕТЕНТНОСТЕЙ УЧАЩИХСЯ

Тагаева Гульмира Сарыгуловна,
педагогика илимдеринин кандидаты,
Кыргыз билим берүү академиясы,
Окуучулардын жетишкендиктерин баалоонун теориясы жана практикасы лабораториясы,
Кыргыз Республикасы, Бишкек шаары

ОКУУЧУЛАРДЫН КОМПЕТЕНТТҮҮЛҮГҮНҮН КАЛЫПТАНУУСУН БААЛООНУН ИНСТРУМЕНТАРИЙЛЕРИН СТАНДАРТТАШТЫРУУ ПРОЦЕССИНДЕ ТЕСТТИК ТАПШЫРМАЛАРДЫ СТАТИСТИКАЛЫК ИШТЕП ЧЫГУУ

Tagaeva Gulmira Sarygulovna,
Head of the Laboratory,
Theory and Practice of Assessing the Achievements of Students,
Kyrgyz Academy of Education,
Kyrgyz Republic, Bishkek city

STATISTICAL PROCESSING OF TEST TASKS IN THE PROCESS OF STANDARDIZATION OF TOOLS FOR ASSESSING THE FORMATION OF STUDENTS' COMPETENCIES

Аннотация: *В статье рассматривается проблема статистической обработки тестовых заданий в процессе стандартизации инструментариев оценивания сформированности компетентностей учащихся. Описываются признаки педагогического теста. Представлена сравнительная таблица современной и классической теорий тестов. Даны результаты апробационного исследования проверки качества тестовых заданий.*

Аннотация: *Макалада окуучулардын компетенттүүлүгүнүн калыптануусун баалоонун инструментарийлерин стандартташтыруунун процессинде тесттик тапшырмаларды статистикалык иштетүүнүн көйгөйлөрү каралат. Педагогикалык тесттин белгилери сүрөттөлөт. Тест теориясынын азыркы (заманбап) жана классикалык теориялардын салыштырмалуу таблицасы берилет. Тесттик тапшырмалардын*

сапатын текшерүүнүн апробациялык изилдөөнүн натыйжалары берилди.

Annotation: *The article discusses the problem of statistical processing of test items in the process of standardization of tools for assessing the formation of students' competencies. The features of the pedagogical test are described. A comparative table of modern and classical test theories is presented. The results of a trial study of the quality control of test tasks are given.*

Ключевые слова*: психометрия, педагогический тест, шкала оценивания, номинальная шкала, дихотомическая шкала, интервальная шкала, шкала отношений, теория тестирования, качественный тест, инструментарии оценивания, надежность, валидность.*

Түйүндүү сөздөр*: психометрия, педагогикалык тест, баалоо шкаласы, номиналдык шкала, интервалдык шкала, катыштар шкаласы, дихотомиялык шкала, тестирлөө теориясы, сапаттуу тест, баалоо инструментарийлери, ишенимдүүлүк, валиддүүлүк.*

Key words: *psychometrics, pedagogical test, grading scale, nominal scale, dichotomous scale, interval scale, relationship scale, testing theory, quality test, assessment tools, reliability, validity.*

Вопросы статистической обработки результатов тестирования являются важными в процессе стандартизации тестов. Они освещены в трудах таких ученых как Аванесов В.С., Анастази А., Майоров А.Н., Стивенс С.С., Урбина С., Челышкова М.Б., Лорд Ф.М. и т. д.

С внедрением современных технологий возникает необходимость совершенствования техники обработки результатов тестового контроля. Вопросы надежности шкалирования результатов тестирования в педагогических измерениях так же важны, как и проблема подготовки качественного теста.

Статистическую обработку результатов тестирования рассматривает наука **психометрия**. Это специальная наука, изучающая теорию и методику измерений в социальных науках (психологии, педагогике), а также математике и статистике.

Основной целью психометрии является создание измерительных инструментариев (таких как анкеты, опросники, тесты и методики описания (оценки) личности), валидизация инструментариев, разработка процедур измерения.

Каждый измерительный инструментарий должен иметь психометрическое сопровождение.

Уже на стадии разработки составитель предполагает, какая шкала оценивания должна использоваться при проведении педагогического теста. Что такое педагогический тест? Педагогический тест – это система заданий специфической формы определенного содержания и возрастающей трудности, позволяющей объективно оценить структуру и эффективно измерить уровень знаний, умений, навыков и представлений [1].

Шкала – это средство фиксации результатов измерения определенных свойств объектов путем упорядочивания их в определенную числовую систему, в которой отношение между отдельными результатами выражено в соответствующих числах. В педагогических измерениях шкалы различаются в зависимости от характера свойств, лежащих в основе их построения.

В практике педагогических тестов и психологии измеряют на разных уровнях шкал и на разных видах шкал.

Уровни шкал:

- номинальная шкала,
- ординальная шкала (ранговая),
- интервальная шкала,
- шкала отношений.

Виды шкал:

- z AM = 0 SD = 1
- IQ AM = 100 SD = 15
- T AM = 50 SD = 10
- STININE AM = 5 SD = 2
- PISA AM = 500 SD = 100

С.С. Стивенс [2] различает четыре типа (уровня) шкал: номинальная шкала, шкала порядка (ранговая), интервальная шкала и шкала отношений.

Измерения на первых двух шкалах считается качественными, на следующих двух – количественными. Шкалы качественных измерений называют «дискретными», а количественных – «непрерывными». В каждой из этих шкал определены свойства чисел, приписываемых объектам.

1. *Номинальная, шкала* наименований, самая простая: например, зачет-незачет. Шкала наименований устанавливает критерии, позволяющие распределить измеряемые объекты по состоянию измеряемого свойства на несколько классов (или категорий). При этом каждый объект должен попасть в определенный класс, в котором объектам приписывается одно и то же число. Объекты одного класса считаются одинаковым и по состоянию измеряемого свойства. Таким образом, номинальная шкала используется, когда устанавливается принадлежность к какой-либо группе по одному признаку (вероисповедание, пол, место жительство) [3, с. 27]. С помощью номинальной шкалы измеряются только качественные признаки, поэтому обработка производится не с самими числами, а с удельными весами. Примером является дихотомическая шкала: за выполненное задание дается 1 балл, а за невыполненное или неправильно выполненное задание – 0 баллов. Методы обработки таких результатов оценивания знаний называют «статистикой качественных признаков». Данные, соответствующие номинальным шкалам, составляют наблюдаемые значения частот появления каждой из разновидностей изучаемой переменной. Эти результаты используются при построении матриц результатов.

2. *Ординальная (порядковая или ранговая) шкала* (например, пятибалльная и др.) – это шкала, результаты измерений по которой невозможно сравнить между собой. В пределах этой шкалы можно только упорядочить тестовые задания, в порядке возрастания или убывания оценок измеряемых параметров. На такой шкале оцениваются только качественные признаки, например, оценка A>B [4, с. 49-54]. Используя ординальную шкалу, нельзя найти среднюю величину и сравнить результаты в числах. Можно фиксировать только место в шкале.

3. *Интервальная* (шкала равных единиц) шкала является шкалой более высокого уровня – количественная. Здесь можно задать разность оценок (X1-X2), абсолютное значение которой трактуется как расстояние между двумя элементами множества, выраженное в определенных единицах. Для такой шкалы характерно отсутствие начала отсчета, равного нулю, но допустимы различные арифметические действия над числами. Эта шкала задает взаимное положение измеряемых объектов относительно друг друга, но не показывает расположение объектов относительно начала координат. Так, например, разности баллов 38-35 и 5-2 одинаковы, а смысл их разности может быть равным. К результатам измерения на такой шкале применимы почти все статистические операции.

Интервальная или нормальная – это такая шкала, у которой задано начало отсчета. На такой шкале можно определить не только единицы измерения, но и понятие нормы (местоположения от начала координат). На такой шкале оцениваются количественные признаки.

4. *Шкала отношений* – самый высокий

уровень измерений: не только приписывание числа измеряемому объекту, но также не допускает все арифметические действия над этими числами (статистические операции), а также устанавливать равенство отношений чисел, приписываемых объектам, что вытекает из фиксированного положения нуля [4, с. 50]. Любая интервальная шкала может использоваться в качестве отношений, если в рамках проводимого измерения задать начало отсчета. На шкале отношений к полученным результатам применимы все известные понятия и методы математической статистики. Содержательная статистическая обработка и интерпретация результатов измерений по этим шкалам могут быть только в том случае, когда методы обработки адекватны тем шкалам, к которым отнесена исходная информация.

Основной целью педагогического тестирования является надежное измерение уровня учебных достижений испытуемых в определенной области. Традиционные тестирования (классические) используют порядковые шкалы, отличающиеся друг от друга длиной, масштабом и значением центрального индекса. Балл тестируемого определяется количеством правильно выполненных заданий «А» из общего числа заданий «В». Тогда отношение А/В можно выразить в процентах и получить 100 балльную шкалу, называемую процентной.

Окончательный балл участников тестирования зависит от относительных успехов каждого по сравнению с успехами других тестируемых [5, 6]. Такие шкалы называют процентильными. Как и процентные, они имеют ранговый смысл. Недостатком этих шкал является невозможность сравнения полученных результатов, между распределением результатов разных выборок тестируемых.

При разработке тестов необходимо предусмотреть форму представления полученных результатов, определяемую как целями тестирования, так и индивидуальными особенностями лиц, которые будут эти результаты использовать. Как правило, форма представления результатов содержит в себе текст, графики, диаграммы, профили и т. д.

Табличная форма представления результатов может представлять собой:

1) значения данных;

2) таблица перевода табличных значений в шкальные;

3) таблицы прогноза, на основе которых осуществляется предсказание успехов в обучении, академической успеваемости или продуктивности деятельности.

Графическая форма представления результатов может представлять собой:

1) график, иллюстрирующий изменение измеряемой величины;

2) диаграмму, иллюстрирующие в наглядной форме представление оценок по тесту.

Для стандартизации необходимо провести несколько апробационных тестирований (претестов). В лаборатории теории и практики оценивания КАО по результатам нескольких апробационных претестов были получены надежные тесты по чтению и пониманию.

Целью апробационного тестирования является проверка функционирования заданий (анализ тестовых заданий) и всего теста в целом, исследование системообразующих свойств теста, оценивание его надежности и валидности. Для начала выбирается теория тестирования, модель, методы оценки качества и валидизации инструментариев, методы представления результатов тестирования.

Теория тестирования – теория, обеспечивающая общие подходы к связыванию наблюдаемых переменных (тестовые баллы участников тестирования) с ненаблюдаемы-

ми переменными (истинные баллы участников, уровни подготовленности (способности) участников).

В настоящее время существуют две теории тестирования – классическая теория тестирования (КТТ) и современная (IRT).

Классическая теория тестирования (КТТ):

- первая половина 20-го века;
- достоинства – простота обработки и интерпретации результатов;
- обладает целым рядом существенных недостатков.

Характеристики заданий в классической теории тестирования:

• трудность задания (коэффициент решаемости): доля испытуемых, выполнивших задание верно (получивших 1 балл за выполнение задания для дихотомических заданий);

• дискриминативность (дифференцирующая способность задания): способность задания различать испытуемых с различным уровнем подготовки;

• надежность – характеристика точности и устойчивости результатов оценки;

• валидность – характеристика пригодности оценочной информации для принятия правильных решений на ее основе.

Item Response Theory (IRT) (Modern Test Theory, современная теория тестирования):

- вторая половина 20-го века;
- позволяет преодолеть недостатки КТТ;
- открывает возможности для использования новых технологий тестирования и дополнительного анализа данных.

Таблица 1. Сравнение современной и классической теорий тестов по Кардановой Е. Ю. [7]

	Классическая теория тестирования (КТТ)	**IRT (модели Раша)**
1	Оценки сложности тестовых заданий зависят от уровня подготовленности конкретной выборки испытуемых	Оценки сложности тестовых заданий инвариантны относительно контингента испытуемых, по результатам тестирования которых они получены
2	Оценки уровня подготовленности испытуемых (первичные баллы) зависят от уровня трудности конкретного теста	Оценки уровня подготовленности испытуемых инварианты относительно тестовых заданий, по результатам выполнения которых они получены
3	Ошибка измерения является величиной постоянной для всех испытуемых. Ошибка измерения заданий не оценивается	Ошибка измерения оценивается индивидуально для каждого испытуемого и каждого задания. Причём ошибка подсчитывается непосредственно, а не косвенно
4	Методы оценивания надёжности требуют существенных ограничений и дают искажённые результаты	Возможно оценить отдельно надёжность измерения испытуемых и надёжность оценивания заданий теста
5	Шкала первичных баллов является порядковой. Никакое преобразование первичных баллов в КТТ не повышает уровня шкалы	Шкала логитов является интервальной, что даёт возможность перейти от ранжирования испытуемых и заданий к измерению соответственно уровня подготовленности и уровня

		трудности
6	Нормальное распределение баллов испытуемых и сложности заданий теста играет существенную роль	Нормальность распределения параметров не требуется
7	Способы установления соответствия между баллами испытуемых, выполнявших различные варианты, требуют трудновыполнимых предположений	Возможно выполнить процедуру выравнивания показателей различных вариантов и осуществить шкалирование на единой метрической шкале. Возможно создание банков заданий
8	Не подходит для компьютерного адаптивного тестирования	Вся теория компьютерного адаптивного тестирования базируется на IRT
9	Анализ концентрируется только на оценивании трудности заданий и мер испытуемых	Возможен анализ влияния дополнительных факторов на оценки параметров заданий и мер испытуемых
10	Искусственное назначение весов заданиям может привести к искажению информации об уровне подготовленности испытуемых	Вес (информационный вклад) тестового задания может быть вычислен отдельно вне зависимости от характеристик других заданий

Самая простая модель IRT – это модель Раша.

Дихотомическая модель Раша – это модель для двухступенчатых видов ответов.

Например, **да/нет** или **правильно/неправильно**.

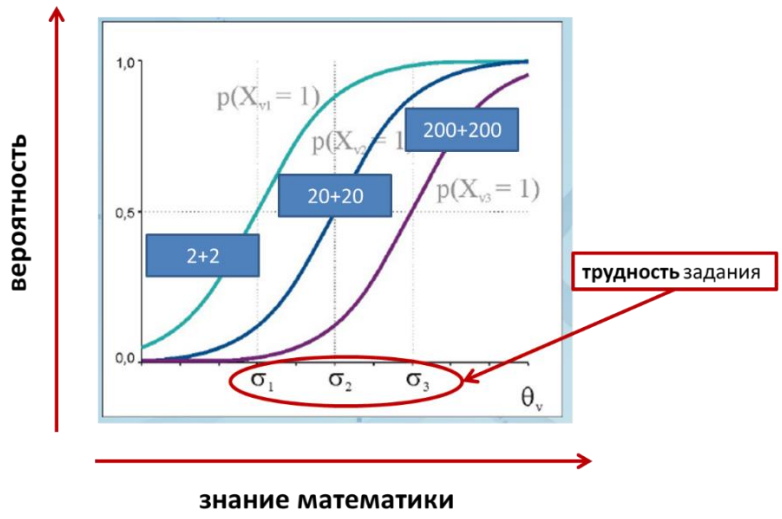
Важные параметры:

- **вероятность решения** задания **p**
- **выражение свойства** (знание математики) **θ (Theta)**
- **трудность** item/задания **σi**

Параметры описываются на одной шкале.

Приведем пример из математики.

Функция item характеристики описывает **взаимосвязь между знанием математики и определённым ответом в тесте** (правильно/не правильно).



При модели Раша параметры (выражение свойства/трудность) определяются по **максимум-лайклихуд-принципу** (Maximum-Likelihood – Prinzip): **принцип максимальной вероятности.**

Определяется, с какой вероятностью модель действительна на основе данных.

Параметры заданий и личные параметры определяются вместе, с целью достичь максимальную вероятность (Maximum Likelihood).

Этот процесс итеративный (постепенный), **результат не однозначный**.

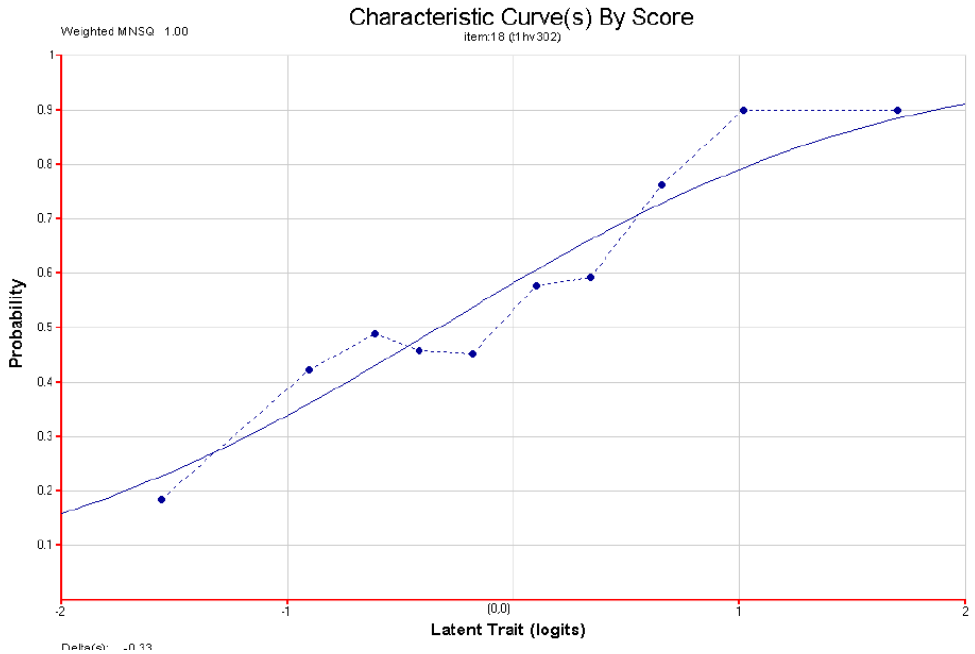
При шкалировании **каждое задание необходимо проверять подходит ли оно к модели**.

Эта проверка называется **Item-Fit**. Задания, которые не подходят к модели, исключаются.

Важные параметры:

- **вероятность решения** задачи p;
- **выражение свойства** (знание) θ;
- **трудность** item/задания σ_i;
- **Weighted Mean Square (MNSQ);**
- **T**
- **Дискриминация**
- У **совершенного** Item-Fit величина **MNSQ** = 1.
- **Плохой** Item-Fit когда **MNSQ>1**.
- **MNSQ<1** указывает **слишком хороший** Item-Fit. IC-функция слишком крутая, это значит слишком чётко разделяет сильных/слабых.
- **T>2** значительное отклонение от ожидаемого результата
- **Дискриминация**

Item-Fit определяется через **то, что мы ожидаем по модели, и то, что мы по-настоящему измерили** (данные)!

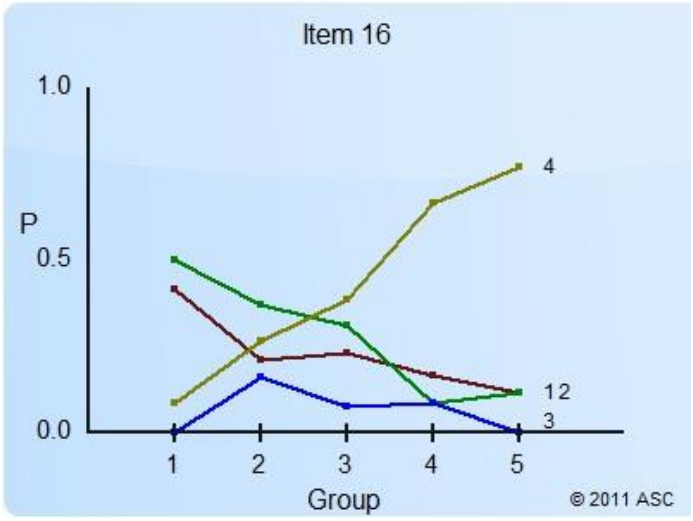


Совершенный Item-Fit

VARIABLES		WEIGHTED FIT	
	item	MNSQ	CI	T
18	t1hv302	1.00	(0.92, 1.08)	-0.1

item:18 (t1hv302)
Cases for this item 298 Discrimination 0.43
Item Threshold(s): -0.33 Weighted MNSQ 1.00
Item Delta(s): -0.33

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
0	0.00	146	48.99	-0.43	-8.20(.000)	-0.59	0.73
1	1.00	152	51.01	0.43	8.20(.000)	0.02	0.80



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
16	16	4	Yes	4	1	

Item statistics

N	P	Total Rpbis	Total Rbis	Alpha w/o
83	0,47	0,45	0,56	0,64

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	17	0,20	-0,14	-0,20	14,65	4,57	Maroon	
2	21	0,25	-0,29	-0,40	13,62	4,04	Green	
3	5	0,06	-0,05	-0,10	15,00	2,74	Blue	

4	39	0,47	0,45	0,56	19,00	3,95	Olive	**KEY**
Omit	1	0,01	-0,10	-0,33	5,00	0,00		
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	17	0,42	0,21	0,23	0,17	0,12	Maroon	
2	21	0,50	0,37	0,31	0,08	0,12	Green	
3	5	0,00	0,16	0,08	0,08	0,00	Blue	
4	39	0,08	0,26	0,38	0,67	0,77	Olive	**KEY**

Перед нами образец статистической характеристики одного конкретного задания.

N – количество учащихся, принявших участие в претесте (здесь: 83).

P – трудность задания (01 – 09) (здесь: 05) (50% не могли ответить правильно).

Total Rpbis – дискриминативность, т.е. способность делить учащихся на слабых и сильных (02<) здесь показатель дискриминативности (хороший: 04).

Alpha w/o – надёжность (08<) (здесь: 0,64), т.е. приближена к 08, так как параметр надежности ещё зависит от качества каждого тестового задания в тесте. Если говорить о дистракторах, то только третий выбрали мало учащихся (5).

Сводная таблица коэффициентов для анализа заданий

MNSQ	0.8 – 1.25
T	< 1.96
rit	> 0.3
% of tot	< 30 тяжёлое = 50 средние > 80 лёгкое

34

Таким образом, чтобы проверить качество заданий, необходимо:

- проверить Ключ!
- проверить дистракторы!
- проверить формулировку тестового задания! [9].

Также необходимо определить длину теста и время тестирования. Здесь мы основывались на проведенных в лаборатории оценки образовательных достижений учащихся Кыргызской академии образования в с участием

учащихся начальных классов [8]. Исходя из этого, мы посчитали важным следующее:

- время тестирования должно определяться по расположению максимума дисперсии тестовых результатов и не превышать 40 минут;

- длина теста не должна превышать 20-30 заданий, в предположении, что на выполнение одного задания потребуется не более одной минуты;

- тестирование необходимо проводить в первой половине дня;

- тестирование желательно проводить в середине недели.

Литература:

1. *Аванесов В.С. Теория и практика педагогических измерений. – ЦТ и МКО УГТУ-УПИ, 2005. г.*
2. *Стивенс С.С. Экспериментальная психология. – М., 1960. – 316с.*
3. *Ингенкамп К. Педагогическая диагностика. / – М.: Педагогика, 1991. – 240 с.*
4. *Нейман Ю.М. Вопросы точностных расчетов в теории моделирования и параметризации педагогических тестов // Труды центра тестирования. – М., 1989 – Выпуск 2.*
5. *Клайн П. Введение в психометрическое программирование: Справочное руководство по конструированию тестов. – Киев: ПАН Лтд, 1994. – 276 с.*
6. *Челышкова М.Б. Теоретико-методологические и технологические основы адаптивного тестирования в образовании. – дисс…док. пед.наук. М., 2001. – 324 с.*
7. *Карданова Е.Ю. Преимущества современной теории тестирования по сравнению с классической теорией тестирования. Вопросы тестирования в образовании. – 2004, № 10.*
8. *Мамытов А., Мыкыева М. Определение продолжительности тестирования образовательных достижений учащихся начальных классов. // Известия КАО, № 2 (26), – Бишкек, 2013. – С. 3-7.*
9. *В статье использованы лекции обучающих семинаров Левин Юлии, симспециалиста GIZ.*

Рецензент:
Калдыбаев С.К.
доктор педагогических наук, профессор