ПСИХОМЕТРИЧЕСКИЙ АНАЛИЗ РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ В РАМКАХ ВНЕШНЕЙ ОЦЕНКИ УЧЕБНЫХ ДОСТИЖЕНИЙ В 2014 ГОДУ

ПРИМБЕТОВА ГУЛЬЖАН СЕРИКБАЕВНА
руководитель управления по организации формирования тестовых заданий для среднего образования Национального центра тестирования, канд. пед. наук

E-mail: primbetova1@yandex.ru

Астана, Казахстан

ИСКАКОВА АЛЬМИРА МУХТАРОВНА
ведущий эксперт управления по организации формирования тестовых заданий для среднего образования Национального центра тестирования

E-mail: iskakova_am@mail.ru

Астана, Казахстан

АННОТАЦИЯ. В статье представлены результаты психометрического анализа тестовых заданий мониторингового исследования ВОУД-2014. В качестве примера показан анализ заданий теста по географии. Результаты психометрического анализа тестовых заданий будут описаны по следующим параметрам: соответствие порогового балла тестов ВОУД порогу, принятому в мониторинговых исследованиях; дифференцирующая способность заданий; анализ согласия тестовых заданий с моделью. В статье также представлены характеристические кривые заданий, находящихся в согласии с моделью, и дан подробный анализ заданий, не находящихся в согласии с моделью.

КЛЮЧЕВЫЕ СЛОВА: психометрический анализ заданий, дифференцирующая способность, коэффициент корреляции, дистрактор.

PSYCHOMETRIC ANALYSIS OF THE EXTERNAL ASSESSMENT OF EDUCATIONAL ACHIEVEMENTS RESULTS IN 2014

GULJAN PRIMBETOVA
Head of the department for development of test items in secondary education, National testing center, PhD in Education
E-mail: primbetova1@yandex.ru
Astana, Kazakhstan

ALMIRA ISKAKOVA
Lead expert of the department for development of test items in secondary education, National testing center
E-mail: iskakova_am@mail.ru
Astana, Kazakhstan

ABSTRACT. The article discusses the results of psychometric analysis of the test items operated by the External assessment of educational achievements results in 2014. The analysis of items on Geography test is provided as an example. The results of psychometric analysis are described according to the parameters as follows: compliance of the assessment benchmarks with the ones commonly used in monitorings; differentiation ability of the test items; analysis of the accordance between the test items and assessment model. The authors also provide characteristic curves of the test items in accordance with assessment model, and detailed analysis of test items disaccording to the model.

KEYWORDS: psychometric analysis of test items, differentiation ability, coefficient of correlation, distractor.

Важным составляющим всех экзаменов и мониторингов, проводимых на национальном уровне, является инструментарий измерения. Инструментарий измерения высокого качества является залогом получения реальной и объективной информации о достижениях обучающихся. В РК в качестве инструментария измерения результатов обучения используются тестовые задания, поэтому оценка их качества осуществляется путем проведения психометрического анализа.

Психометрия является разделом прикладной статистики, которая касается создания и валидизации тестовых заданий. По результатам психометрического анализа осуществляется нормирование и шкалирование результатов тестирования, установление пороговых баллов, оценка качества тестовых заданий.

При проведении психометрического анализа используются различные программы обработки результатов тестирования. В нашей статье будут показаны результаты психометрического анализа тестовых заданий по географии, использованных в рамках ВОУД-2014. Статистическая обработка проводилась с помощью программ Winsteps, IATA, RUMM.

ПРОВЕРКА СООТВЕТСТВИЯ ПОРОГОВОГО БАЛЛА ТЕСТОВ ВОУД ПОРОГУ, ПРИНЯТОМУ В МОНИТОРИНГОВЫХ ИССЛЕДОВАНИЯХ

ВОУД является мониторинговым исследованием и направлено на проверку усвоения обучающимися учебной программы, то минимально подготовленный учащийся должен выполнить 50% из всех заданий.

Для проверки соответствия порогового балла тестов ВОУД-2014 порогу, установленному в мониторинговых исследованиях, нами использованы методы Nedelsky, Ebel.

Метод Nedelsky разработан для установления пороговых оценок. Процедура метода Nedelsky предполагает привлечение экспертов, которые выбирают вероятности правильного выполнения тестовых заданий с множественным выбором. При выборе вероятностей суждения эксперты основываются на оценках правдоподобия того, что определенная воображаемая группа пограничных экзаменующихся будет в состоянии исключить неправильные варианты при выборе ответов к заданиям. Первая часть работы экспертов состоит из обсуждения экспертами описания

пограничного экзаменующего. Эксперты, использующие метод Nedelsky, рассматривают индивидуально каждое задание, уделяя особое внимание вариантам выбора ответов к заданиям. Затем они снова анализируют каждое задание в тесте и для него идентифицируют варианты ответов, которые, как они полагают, гипотетический минимально компетентный экзаменующийся, типичный для выборки, где будет применяться тест, исключит из набора ответов как неправильные. Величина, обратная числу оставшихся вариантов ответа, становится для каждого задания «величиной Недельски». Эта величина интерпретируется как вероятность того, что пограничный ученик выполнит задание верно [4, с. 246–247].

Другой популярный метод – это метод, предложенный Робертом Эбелем (Метод Ebel). «Существенная особенность данного метода состоит в том, что в нем от каждого эксперта требуется не одна, а две оценки: первая – приближенная оценка трудности каждого задания, вторая – суждение относительно релевантности задания поставленной цели создания теста. Обе оценки носят не вероятностный количественный характер в виде долей или процентов, а получаются путем размещения суждений экспертов по заданиям в соответствии с отдельными оценочными категориями. Так, в первом случае при оценивании трудности заданий эксперты относят их к одной из трех категорий: легкое, среднее и трудное, а во втором случае – к одной из четырех категорий релевантности целям тестирования: существенное, важное, приемлемое и сомнительное. После вынесения двух суждений относительно отдельных заданий эксперты предлагают процент заданий, на которые гипотетически минимально компетентные экзаменующиеся должны ответить правильно. Преимущество метода Эбеля состоит в том, что его можно использовать вне зависимости от форм заданий в тесте во всех тех случаях, когда выбирается дихотомическая оценка результатов выполнения заданий теста [4, с. 251–255].

На примере тестов географии покажем результаты проведенной работы. Согласно экспертной оценке по географии, пороговые значения приведены в таблице 1:

Таблица 1. Пороговые оценки по географии

Методы установления пороговых баллов	Тест для школ с казахским языком обучения	Тест для школ с русским языком обучения
Метод Nedelsky	9,5	10
Метод Ebel	10	9,5

По данным, полученным экспертным путем, можно понять, что между порогом, определенным по тестам, и порогом, принятым в мониторинговых исследованиях, разницы нет. Следовательно, уровни трудностей заданий в тестах, использованных в ВОУД-2014, определены правильно.

ОЦЕНКА ДИФФЕРЕНЦИРУЮЩЕЙ СПОСОБНОСТИ ТЕСТОВЫХ ЗАДАНИЙ

Рассмотрим первый параметр качества тестовых заданий. Это дифференцирующая способность заданий. Дифференцирующая способность показывает, насколько эффективно тестовое задание различает учащихся, овладевших и не овладевших учебным материалом. Дифференцирующая способность тестового задания позволяет выявлять сильных и слабых учащихся, дифференцировать испытуемых по уровню подготовленности [3, с. 296]. Для вычисления коэффициента дифференцирующей способности мы использовали метод крайних групп. Доля членов крайних групп может изменяться в широких пределах в зависимости от величины выборки. Чем больше выборка, тем меньшей долей испытуемых можно ограничиться при выделении групп с высоким и низким результатами. В нашем докладе мы использовали 27-процентную группу, так как при таком процентном соотношении достигается максимальная точность определения дифференцирующей способности [3, с. 164].

Индекс дифференцирующей способности D определяется как разность между долей лиц, правильно решивших задачу, из «высокопродуктивной» и «низкопродуктивной» групп и вычисляется по формуле:

$$D = \frac{N_{n_{max}}}{N_{max}} - \frac{N_{n_{min}}}{N_{min}},$$

где:

$N_{n_{max}}$ – количество испытуемых в группе лучших, верно выполнивших задание;

$N_{n_{min}}$ – количество испытуемых в группе худших, верно выполнивших задание;

N_{max} – общее количество испытуемых в группе лучших;

N_{min} – общее количество испытуемых в группе худших.

Коэффициент дифференцирующей способности может принимать значения от -1 до +1. Результат D ≥ 0,3 считается удовлетворительным. Если значение коэффициента близко к 0, то задачи должны рассматриваться как некорректно сформулированные.

Согласно проведенному анализу тестовых заданий по географии для школ с русским языком обучения выявлено, что 95% заданий признаны пригодными, 5% тестовых заданий имеют низкую дифференцирующую способность. Этот означает, что 5% тестовых заданий не различают учащихся, овладевших и не овладевших учебным материалом. Это допустимое количество заданий с низким показателем дифференцирующей способности.

АНАЛИЗ СОГЛАСИЯ ЗАДАНИЙ ТЕСТА С МОДЕЛЬЮ

Рассмотрим статистические показатели заданий теста для того, чтобы оценить, находятся ли все задания теста в согласии с моделью.

В таблице 2 даны статистические показатели заданий теста. Статистики согласия приведены в последних двух столбцах таблицы. Простая статистика согласия более чувствительна к экстремально неожиданным ответам, когда сильный испытуемый неожиданно неправильно отвечает на легкое задание или, наоборот, слабый испытуемый неожиданно правильно отвечает на трудное задание. Взвешенная статистика позволяет уменьшить влияние экстремально неожиданных ответов. Поскольку мы анализируем результаты мониторингового исследования, которое не является широкомасштабным исследованием и не является формой контроля «с высокими ставками», каковыми являются национальные экзамены, то допустимым интервалом для статистик согласия является (0,5; 1,5).

Из таблицы 2 видно, что у двух заданий статистики согласия не входят в указанный интервал (№ 6, № 16) и коэффициенты корреляции этих заданий низкие (0,14 и 0,21 соответственно). Отклонение статистик согласия у данных заданий от нормы незначительное. Из полученных после статистической обработки результатов тестирования видно, что 80% тестовых заданий находятся в хорошем согласии с моделью.

Отдельно рассмотрим задания, находящиеся в согласии с моделью, и задания, не находящиеся в согласии с моделью.

Таблица 2. Статистические показатели заданий теста

№	Оценка трудности	Ошибка измерения	Коэффициент корреляции	Статистики согласия	
				Взвешенная (INFIT)	Простая (OUTFIT)
1	-2,16	0,24	0,48	0,80	0,62
2	-2,10	0,91	0,37	0,91	0,98
3	-0,59	1,04	0,37	1,04	1,11
4	-0,75	0,18	0,41	0,98	0,98
5	0,73	0,18	0,43	0,98	1,01
6	1,12	0,18	0,14	1,27	1,77
7	1,26	0,19	0,44	0,93	1,19
8	-1,37	0,20	0,41	0,95	0,92
9	0,67	0,17	0,54	0,86	0,82
10	0,64	0,17	0,51	0,90	0,86
11	0,58	0,17	0,50	0,91	0,88
12	0,46	0,17	0,54	0,87	0,84
13	0,02	0,17	0,36	1,08	1,11
14	0,95	0,18	0,53	0,86	0,9
15	-0,29	0,17	0,50	0,92	0,82
16	-0,94	0,18	0,21	1,18	0,48
17	0,25	0,17	0,36	1,08	1,12
18	0,49	0,17	0,37	1,06	1,14
19	0,31	0,17	0,44	1,00	0,96
20	0,73	0,18	0,28	1,16	1,34

АНАЛИЗ СОГЛАСИЯ ЗАДАНИЙ С МОДЕЛЬЮ

При исследовании заданий теста анализируется согласие с моделью измерения ответов всех испытуемых на каждое задание теста. В этом случае для каждого задания теста рассматриваются общая и взвешенная статистики и их стандартизированные версии. Суммарные сведения по тестовым заданиям и тестируемым представлены в таблицах 3 и 4.

Таблица 3. Общие характеристики тестируемых (испытуемых)

	Total score	Count	Measure	Model error	Взвешенные статистики согласия (INFIT)		Общие статистики согласия (OUTFIT)	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	10,1	20	0,08	0,54	1,00	0,0	1,04	0,0
S.D	3,8	0	1,10	0,09	0,21	0,9	0,43	0,9
MAX	19,0	20	3,27	1,04	1,72	2,5	3,07	3,0
MIN	2,0	20	-2,61	0,49	0,62	-2,2	0,49	-2,0

Средние значения общей (OUTFIT MNSQ) и взвешенной общей статистики согласия (INFIT MNSQ) находятся в пределах 0,5–1,5. Средние значения стандартизированной общей (OUTFIT ZSTD) и стандартизированной взвешенной общей статистики (INFIT ZSTD) находятся в пределах -2–2. Следовательно, задания теста хорошо согласуются с моделью. Однако предельные значения далеки от допустимых границ.

АНАЛИЗ УРОВНЯ ПОДГОТОВЛЕННОСТИ ИСПЫТУЕМЫХ

При исследовании уровня подготовленности испытуемых анализируется согласие с моделью измерения ответов каждого испытуемого на все задания теста. В этом случае для каждого испытуемого рассматриваются общая OUTFIT MNSQ и взвешенная INFIT MNSQ статистики и их стандартизированные версии (OUTFIT ZSTD и INFIT ZSTD).

Ошибка измерения очень низкая, она равна 0,08. Коэффициент Альфа-Кронбаха уровня подготовленности испытуемых равен 0,74. Стандартное отклонение равно 1,10. Значения общей (OUTFIT MNSQ) и взвешенной общей статистики согласия (INFIT MNSQ) находятся в пределах 0,5–1,5. Значения стандартизированной общей (OUTFIT ZSTD) и стандартизированной взвешенной общей статистики (INFIT ZSTD) находятся в пределах -2–2. Следовательно, профиль ответа испытуемых, выполнявших данный вариант теста, находится в хорошем согласии с моделью.

Таблица 4. Общие характеристики тестовых заданий по географии

	Total score	Count	Measure	Model error	Взвешенные статистики согласия (INFIT)		Общие статистики согласия (OUTFIT)	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	86,5	171	0,00	0,18	0,99	-0,1	1,04	0,2
S.D	30,4	0	0,98	0,02	0,12	1,4	0,26	1,6
MAX	148,0	17	1,26	0,24	1,27	2,9	1,77	4,3
MIN	47,0	171	-2,16	0,17	0,80	-2,1	0,62	-1,6

На рисунке 1 представлена карта переменных, на которой показано распределение испытуемых и заданий относительно друг друга на общей метрической шкале. Слева – шкала логитов (уровень подготовленности испытуемых), справа – задания. Более трудные задания и более сильные

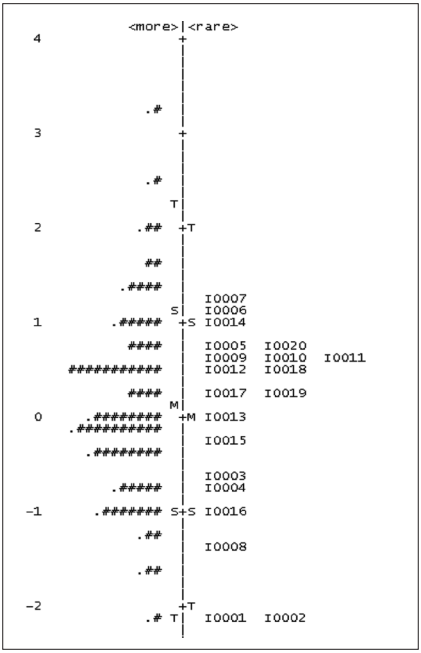

Рисунок 1. Карта переменных

испытуемые расположены в верхней части карты, легкие задания и менее подготовленные испытуемые – в нижней части карты.

Распределение на карте близко к нормальному. Следовательно, тест ориентирован на данную выборку и соответствует уровню подготовленности испытуемых. Об этом говорит центрированность множества заданий относительно выборки тестируемых. Данный тест оптимален по трудности. Но не хватает нескольких более трудных заданий для более точного оценивания сильных испытуемых.

АНАЛИЗ ТЕСТОВЫХ ЗАДАНИЙ, НАХОДЯЩИХСЯ В СОГЛАСИИ С МОДЕЛЬЮ

Основное положение IRT подразумевает, что наблюдаемые результаты выполнения теста порождаются взаимодействием двух множеств: множества значений латентного параметра, характеризующего уровень знаний испытуемого, и множества значений латентного параметра, характеризующего уровень трудности [5, с. 102]. Для сравнения теоретической вероятности с эмпирической поступают следующим образом, всю выборку испытуемых разбивают на три группы в соответствии с оценками их уровней подготовленности. В первую группу отбирают испытуемых

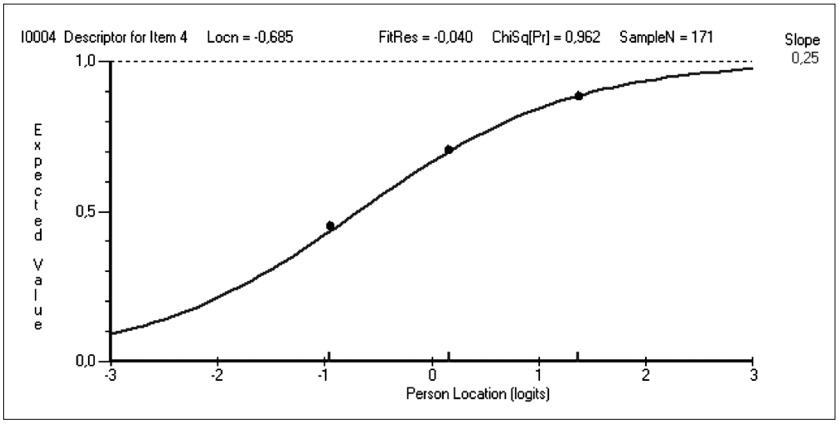

Рисунок 2. Характеристическая кривая задания № 4

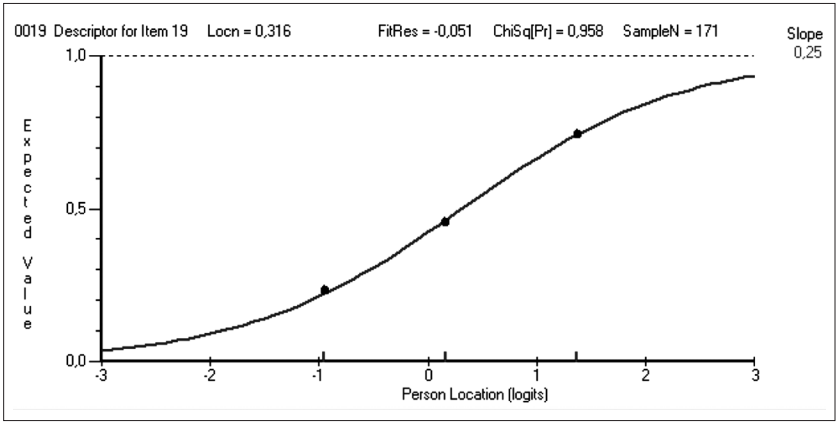

Рисунок 3. Характеристическая кривая задания № 19

с высоким уровнем подготовленности, во вторую – со средним уровнем подготовленности и в третью группу – с низким уровнем подготовленности. Для каждой группы вычисляют среднее значение оценок подготовленности испытуемых данной группы и эмпирическую вероятность правильного ответа на рассматриваемое задание. Если задание находится в хорошем согласии с моделью, то эти три точки будут находиться в достаточной близости от модельной характеристической кривой задания. Для примера рассмотрим характеристические кривые заданий, которые находятся в хорошем согласии с моделью (рис. 2, 3).

На рисунках 2 и 3 видно, что все три точки заданий № 4, 19 находятся на модельной кривой. Это означает, что задание функционирует в полном согласии с моделью измерения.

АНАЛИЗ ТЕСТОВЫХ ЗАДАНИЙ, НЕ НАХОДЯЩИХСЯ В СОГЛАСИИ С МОДЕЛЬЮ

Тест, как видно из полученных данных таблицы 4, одномерный. Но задания № 6 и № 16 действуют обособленно. По этой причине проанализируем задания на предмет их корректности, правильности формулировки дистракторов, соответствия учебной программе и материалам учебника.

Формулировки в обоих тестовых заданиях соответствуют требованиям, предъявляемым к заданиям подобной формы. Дистракторы сформулированы правильно, грамматически согласованы с условием задания. Термины, использованные в заданиях, соответствуют терминам, принятым в географии. Условие задания выглядит следующим образом: «Вид специализации, предусматривающий выполнение трех основных стадий изготовления машин: заготовка деталей – механическая обработка – сборка».

Тестовое задание № 6 является заданием репродуктивного характера, выполнение которого требует от учащегося применения знаний в знакомой ситуации. От учащегося требуется продемонстрировать знание той или иной специализации производства. В учебниках по географии (9-й класс) описания всех существующих в производстве специализаций

Таблица 5. Выбираемость вариантов ответов тестового задания учащимися

Варианты ответов на задание	A	B	C	D	E
Количество учащихся, выбравших вариант ответа (в %)	24,5	9,9	8,7	26,9	30

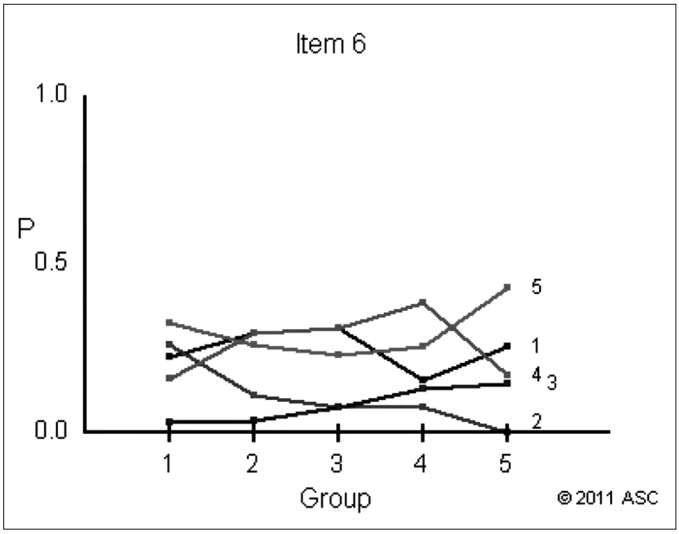

Рисунок 4. Дистракторный анализ тестового задания № 6

даны подробно, поэтому выполнение данного задания не должно вызывать никаких проблем.

Проведем дистракторный анализ данного тестового задания. При проведении дистракторного анализа определяется процент выполняемости учащимися предложенных вариантов ответов. Процент выбираемости представлен в таблице 5.

Дистракторный анализ тестового задания № 6 графически изображен на рисунке 4.

Согласно требованиям тестологии дистрактор считается корректным и работающим, если его выбирают более 5% учащихся. Данные таблицы 5 и рисунка 4 показывают, что все дистракторы корректные и хорошо функционируют. Таким же образом был проведен дистракторный анализ задания № 16. Результат анализа аналогичный.

Таким образом, проведенный анализ содержания заданий (№ 6, 16) показал, что тестовые задания корректны, дистракторы сформулированы корректно и работают правильно. Следовательно, проблема несоответствия тестовых заданий модели не в самих заданиях, а в чем-то другом. Для выявления этой проблемы подробнее рассмотрим тестовое задание № 16.

Условие задания сформулировано следующим образом: «На 1 января 2010 года в городе X проживало 600 тысяч человек. За 2010 год родилось 20 тысяч, умерло 5 тысяч, прибыло 40 тысяч, выбыло 15 тысяч человек. Общий прирост города X за 2010 год…». Тестовое задание № 16 является заданием продуктивного характера, при выполнении которого от учащегося требуется продемонстрировать умение применять знания в незнакомой ситуации, умение отбирать необходимую информацию. В условии задания есть лишняя информация. Это информация о населении города X на начало 2010 года. Данная информация не нужна при вычислении общего прироста населения. От учащегося требуется продемонстрировать умение отбирать необходимую информацию.

Умение применять знания в незнакомой ситуации учащийся должен продемонстрировать при решении задачи. Так как в учебниках географии (9-й класс) даны только формулы расчета естественного и искусственного прироста населения, а формула расчета общего прироста населения не дана, то он, используя имеющиеся готовые формулы, должен определить формулу расчета общего прироста населения и выполнить вычисление.

Проведенные нами беседы с учащимися показали, что учащиеся на уроках редко решают географические задачи, хотя все учебники содер-

жат достаточное количество задач для формирования навыков решения задач по данному предмету.

Возможные причины: недостаточно хороший уровень подготовленности учителей географии, то есть неумение решать географические задачи учителями. Пути решения проблемы – организация специальных курсов повышения квалификации для учителей географии по решению задач.

Таким образом, психометрический анализ результатов тестирования в рамках ВОУД-2014 позволил выявить следующее:

1) варианты тестов сформированы правильно и позволяют точно определить минимальный уровень подготовленности учащихся 9-х классов;

2) тестовые задания корректны, находятся в согласии с моделью;

3) 95% тестовых заданий позволяют дифференцировать учащихся в соответствии с их уровнями подготовленности.

СПИСОК ЛИТЕРАТУРЫ

1) Инструкция по проведению Внешней оценки учебных достижений : утв. приказом и.о. министра образования и науки Республики Казахстан от 06.04.2012 № 151 // Сайт Комитета по контролю в сфере образования и науки Министерства образования и науки Республики Казахстан [Электронный ресурс]. – Режим доступа : http://skocontrol.gov.kz/files/instrukciya%20po%20VOUD%20rus.htm.

2) Ефремова, Н.Ф., Михалева, Т. Г. Мониторинг учебных достижений как объект стандартизации / Н. Ф. Ефремова, Т. Г. Михалева // Стандарты и мониторинг в образовании. – 2009. – № 3. – С. 12–13.

3) Майоров, А. Н. Теория и практика создания тестов для системы образования / А. Н. Майоров. – М. : Интеллект-Центр, 2002. – 296 с.

4) Звонников, А. И. Контроль качества обучения при аттестации: компетентностный подход : учеб. пособие / А. И. Звонников. – М, 2012.

5) Карданова, Е. Ю. Моделирование и параметризация тестов: основы теории и приложения / Е. Ю. Карданова. – М., 2008.