



Сравнительный анализ характеристик тестовых заданий, уровня подготовленности абитуриентов и успеваемости студентов

В.В. Малиновский, А.А. Чиркина

Учреждение образования «Витебский государственный университет имени П.М. Машиерова»

Проведен сравнительный анализ характеристик тестов, уровня подготовленности абитуриентов и успеваемости студентов первого курса за период 2006–2012 годов. Установлено, что уровень трудности тестовых заданий и их дифференцирующая способность не изменяются по годам. Среди тестовых заданий преобладают задания среднего уровня сложности. Обнаружена умеренная корреляционная зависимость между результатами централизованного тестирования по всем предметам и показателями успеваемости студентов. Результаты централизованного тестирования не дают возможности статистически достоверно предсказать «успешность–неуспешность» обучения на первом курсе. Величина среднего балла аттестата коррелирует с успешностью обучения для всех групп студентов. Получение по предметам централизованного тестирования баллов в диапазоне 0–9 не означает, что студент не сможет успешно учиться.

Ключевые слова: централизованное тестирование, характеристика тестов, успеваемость, абитуриенты, студенты.

Comparative analysis of test characteristics, the level of university applicants' readiness and students' academic performance

V.V. Malinovsky, A.A. Chirkina

Educational establishment «Vitebsk State University named after P.M. Masharov»

The comparative analysis of test characteristics, the level of university applicants' readiness and first year students' academic performance for the period of 2006–2012 has been carried out. It has been proved that the level of test difficulty and their differentiating ability haven't changed during the above mentioned period. Among test tasks those of medium level prevail. Moderate correlation dependence between the results of centralized testing in all subjects and the indicators of students' academic performance has been found out. Results of centralized testing don't make it possible to have a statistically valid prediction about 'success–failure' of first year students' academic performance. The average score of the certificate of secondary education correlates with successful training for all student groups. The points of 0 to 9 in the subjects of the centralized testing don't mean a student's inability to make good progress in his studies.

Key words: centralized testing, test characteristics, academic performance, university applicants, students.

Тестирование является одной из форм массового контроля знаний, позволяющей эффективно оценить и измерить уровень подготовленности обучаемых. В настоящее время существует два основных подхода к разработке тестовых материалов и интерпретации результатов. Критериально-ориентированный подход позволяет оценить, в какой степени испытуемый усвоил необходимый объем знаний, умений или навыков, а нормативно-ориентированный – предназначен для дифференциации испытуемых по результатам выполнения теста. Задача конкурсного отбора абитуриентов – ранжировать испытуемых по уровню подготовки, определить рейтинг каждого из них. Нормативно-ориентированный тест позволяет сравнивать учебные достижения (уровень знаний и умений) отдельных испытуемых друг с другом на основе распределения тестовых баллов.

Традиционный контроль с оценкой, основанной на опыте и интуиции преподавателя, никогда не был лишен субъективизма, а в ряде случаев порождал конфликты и разочарования со стороны абитуриентов. Использование тестовых технологий исключает субъективный, личностный подход к ученику и гарантирует воспроизводимость результатов при проверке знаний. Н.С. Феськов отмечает, что тестовые технологии как форма стандартизированных педагогических измерений повышают у школьников мотивацию обучения и достижения более высоких результатов и создают равные условия для испытуемых по сложности, объему, времени выполнения. Если же тестирование приобретает статус централизованного, то минимизирует случайные ошибки и погрешности как в процессе проведения, так и на этапе обработки результатов и интерпретации данных [1].

Введение централизованного тестирования (ЦТ) в качестве основной формы вступительных испытаний в вузы Республики Беларусь породило ряд вопросов: соответствуют ли тесты школьной программе, адекватны ли они уровню знаний и умений учеников, насколько точно тесты дифференцируют абитуриентов по уровню их подготовленности по предмету. Республиканский институт контроля знаний неоднократно заявлял, что тесты удовлетворяют требованиям, предъявляемым к контролирующим материалам такого рода. Анализ результатов централизованного тестирования дает возможность характеризовать уровни подготовленности абитуриентов и ранжировать их по этому признаку [2].

Однако ежегодно после окончания приема в вузы в печати возникают вопросы, связанные с условием тестовых заданий, снижением уровня подготовки абитуриентов, целесообразностью учета среднего балла аттестата, введением нижней планки результатов тестирования, позволяющей поступать в вуз. Зачастую обсуждение этих вопросов проводится на уровне субъективных взглядов участников дискуссии и, следовательно, выводы не могут носить достоверный характер. Ответы можно дать на основе статистического анализа результатов централизованного тестирования и итоговой успеваемости студентов первого курса.

Цель исследования: анализ результатов ЦТ, уровня подготовленности абитуриентов и успеваемости студентов. Для реализации цели были поставлены следующие задачи: оценить эмпирические характеристики тестовых заданий (трудности, дискриминативности) и уровня подготовленности абитуриентов; определить связи между результатами централизованного тестирования, средним баллом аттестата и успеваемостью студентов на первом курсе по дисциплинам математического цикла.

Материал и методы. В исследовании были использованы результаты участников ЦТ по математике по пункту тестирования № 703 Витебского государственного университета имени П.М. Машерова с 2006 по 2012 год (в общей сложности 10905 человек, получивших положительные результаты теста), а также итоги первой сессии, средний балл аттестата, результаты ЦТ по русско-белорусскому языку, физике и математике студентов математического факультета специальностей «Прикладная математика», «Математика. Информатика» 2006–2012 годов набора (566 студентов).

При анализе качества тестовых заданий использовались две теории: классическая теория тестирования (Classical Test Theory, CTT) и математическая теория измерений (Item Response Theory, IRT).

Были проверены следующие предположения: характеристики тестовых заданий и уровень подготовленности тестируемых не изменяются по годам; средний балл аттестата не связан с успешностью обучения, его можно не учитывать при приеме в вуз; существует нижняя граница баллов по ЦТ, которая предопределяет «успешность–неуспешность» обучения.

Для проверки этих предположений использовалась статистическая обработка результатов ЦТ и экзаменационных сессий с помощью пакета программ STATISTICA (StatSoft, USA). Применялись следующие методы статистического анализа: проверка нормальности распределения количественных признаков с помощью критериев Лиллиефорса и Шапиро–Уилка, ранговый анализ вариаций по Краскелу–Уоллису, сравнение групп с использованием критерия Манна–Уитни, ранговый корреляционный анализ по Спирмену.

Результаты и их обсуждение. Трудность в классической теории тестов определяется как доля тестируемых, которые справились с заданием: $p_j = R_j/N$, где N – общее количество испытуемых; R_j – число правильных ответов, полученных по заданию с номером j . Недостаток данного показателя: характеристики трудности тестовых заданий как доли правильных ответов зависят от конкретной группы испытуемых, при помощи которых они были получены.

Стандартной мерой дифференцирующей способности задания в классической теории тестирования является значение коэффициента корреляции задания с тестом – корреляция оценок, полученных испытуемыми по заданию, с суммой баллов тех же испытуемых. Мера связи определяется посредством расчета классического коэффициента корреляции Пирсона. Чем выше значение коэффициента корреляции, тем лучше. В качестве нижней границы включения заданий в тест обычно рассматриваются значения $r = 0,3$, и самой нижней, в исключительных случаях, $-r = 0,2$.

Математическая теория измерений IRT является психолого-педагогическим вариантом более общей методологии латентно-структурного анализа (Latent Structure Analyses, LSA), нацеленного на выявление латентных качеств личности посредством вероятностно-статистического моделирования. Однопараметрический вариант IRT предложен Георгом Рашем (G. Rasch). Оценка трудности тестовых заданий на основе модели Раша не зависит от выборки испытуемых, на которых она была получена, а оценка уровня знаний испытуемых аналогично не зависит от используемого набора тестовых заданий. Ключевая идея модели Раша может быть сформулирована следующим образом: вероятность правильного ответа на тестовое задание зависит от уровня подготовленности тестируемого и трудности тестового задания. Основная логистическая однопараметрическая модель Раша:

$$p_{ij} = \frac{e^{d_j(\theta_i - \beta_j)}}{1 + e^{d_j(\theta_i - \beta_j)}},$$

где p_{ij} – вероятность правильного ответа i -го испытуемого на j -е задание; θ_i – уровень подготовленности i -го испытуемого; β_j – мера трудности j -го задания; d – константа шкалирования, которая вводится в связи с использованием логистической функции для описания латентных параметров, распределение которых предполагается нормальным.

Несмотря на то, что в основной формуле используется два параметра β и θ , модель называется однопараметрической, так как вероятность успешного выполнения задания реально зависит не от самих параметров, а от их разности. Существенное преимущество модели Раша по сравнению с классической теорией состоит в том, что в качестве меры трудности заданий и меры подготовленности испытуемых применяется единая метрическая шкала логитов, которая позволяет сопоставить трудность задания β_j и подготовленность испытуемого θ_i . По Рашу определены два логита: логит уровня знаний – натуральный логарифм отношения доли правильных ответов испытуемого на все задания теста к доле неправильных ответов; логит уровня трудности задания – натуральный логарифм отношения доли неправильных ответов на задание теста к доле правильных ответов на это задание по множеству испытуемых. По результатам вычисления θ_i , β_j строятся характеристические кривые тестовых заданий. При геометрической интерпретации параметр β можно рассматривать как характеристику положения кривой задания относительно оси θ . Если уровень подготовленности соответствует трудности задания, то вероятность правильного ответа составляет 50%. Представляет интерес разность между уровнем подготовленности и трудностью задания. Если разность между уровнем подготовленности и уровнем трудности тестового задания равна 1 логиту, то вероятность верного выполнения таким испытуемым такого задания равна 73%, если разность равна -1 логит, то вероятность равна 27%. Большие положительные или отрицательные значения разности не представляют интереса для тестирования, так как вероятность справиться с заданием близка либо к 100% либо к 0%.

По результатам ЦТ по математике по пункту тестирования № 703 были получены основные характеристики тестовых заданий и уровня подготовленности тестируемых в 2006–2012 гг. В табл. 1 представлены пять уровней трудности тестовых заданий по F. Baker и распределение тестовых заданий по трудности с использованием параметров IRT.

Таблица 1

Распределение тестовых заданий по трудности

Градации трудности задания	Мера β_j	Год тестирования						
		2006	2007	2008	2009	2010	2011	2012*
Очень трудные	$\beta_j > 2,6$	2	5	4	7	4	7	4
Трудные	$1,5 < \beta_j < 2,59$	7	2	3	4	5	4	3
Среднего уровня трудности	$-1,49 < \beta_j < 1,49$	16	16	20	16	19	17	20

Легкие	$-2,59 < \beta_j < -1,5$	0	2	3	1	1	2	2
Очень легкие	$\beta_j < -2,6$	0	0	0	2	1	0	0
Всего заданий		25	25	30	30	30	30	29

* Из 30 заданий одно задание было исключено из рассмотрения, так как ни один из тестируемых его не выполнил.

Считается, что на тестовые задания с $\beta_j < -6$ практически каждый испытуемый дает правильный ответ. При $\beta_j > 6$ с заданием не сможет справиться ни один испытуемый. Рекомендуется рассматривать задания с уровнем трудности от -3 до +3. Наибольшую эффективность имеют задания, соответствующие средним значениям уровней подготовленности участников тестирования [3]. В тесте преобладают задания среднего уровня сложности.

Проверка показала, что в большинстве групп распределение тестовых заданий по трудности отличается от нормального. Ранговый анализ вариаций по Краскелу–Уоллису для характеристик «трудность тестового задания» ($p = 0,9490$) и «доля правильных ответов» ($p = 0,7146$) показал, что значимые статистические различия между значениями выбранных показателей по годам отсутствуют (в обоих случаях уровень значимости $p > 0,05$).

Для анализа дифференцирующей способности тестовых заданий была использована двухпараметрическая модель Бирнбаума (A. Birnbaum), которая является развитием однопараметрической модели Раша:

$$p_{ij} = \frac{e^{da_j(\theta_i - \beta_j)}}{1 + e^{da_j(\theta_i - \beta_j)}},$$

где помимо прежних обозначений добавляется еще один параметр a_j , характеризующий дискриминативность задания теста. Считается, что трудность и дискриминативность – взаимосвязанные характеристики тестового задания. Высокая дискриминативность характерна для заданий со средним показателем трудности. Встречающиеся в практике задания с отрицательными значениями a_j не относятся к тестовым заданиям, а потому подлежат удалению. В этом случае более подготовленные испытуемые отвечают правильно с меньшей вероятностью, а менее подготовленные – с большей вероятностью. На практике рекомендуется использовать задания, для которых значение a_j содержится в интервале от 0,5 до 3.

При геометрической интерпретации параметр a_j связан с крутизной кривой задания в точке ее перегиба. Таким образом, значения a_j , близкие к нулю, соответствуют случаю, когда испытуемые с разными уровнями подготовленности правильно отвечают на j -е задание теста с приблизительно равной вероятностью, эти задания являются бесполезными для дифференциации испытуемых.

По уровню дифференцирующей способности F. Baker делит тестовые задания на шесть уровней. В табл. 2 представлено распределение тестовых заданий ЦТ по математике по дискриминативности с использованием параметров IRT.

Таблица 2

Распределение тестовых заданий по дифференцирующей способности

Дифференцирующая способность	Значения параметра крутизны функции, a_j	Год тестирования						
		2006	2007	2008	2009	2010	2011	2012
Практически отсутствует	0	0	0	0	0	0	1	0
Очень низкая	0,01–0,34	6	10	10	9	11	8	5
Низкая	0,35–0,64	17	13	16	16	18	17	18
Средняя	0,65–1,34	2	2	4	5	1	4	6
Высокая	1,35–1,69	0	0	0	0	0	0	0
Очень высокая	>1,70	0	0	0	0	0	0	0

В целом полученные результаты показывают сбалансированность тестов по уровню трудности и дифференцирующей способности. Однако каждый тест по математике содержал ряд очень трудных заданий с низкой различающей способностью. В то же время большая часть заданий со средним уровнем трудности ($\beta_j < -1$) показала хорошую дискриминативность. Сравнение групп по Краскелу–Уоллису для характеристик «дифференцирующая способность задания» ($p = 0,7858$) и

«коэффициент корреляции задания с тестом» ($p = 0,7858$) показало, что значимых статистических различий по годам нет.

Представляет интерес анализ динамики уровня подготовленности тестируемых в 2006–2012 гг. Несмотря на большой объем выборок, во все годы распределение уровня подготовленности тестируемых отличалось от нормального. Так как данные распределены асимметрично, в качестве показателя центральной тенденции ряда использовалась медиана, которая не так сильно подвержена воздействию крайних значений распределения, как среднее значение. Типичный пример распределения показателя «уровень подготовленности тестируемого» (результаты ЦТ 2012 года) представлен на рис.:

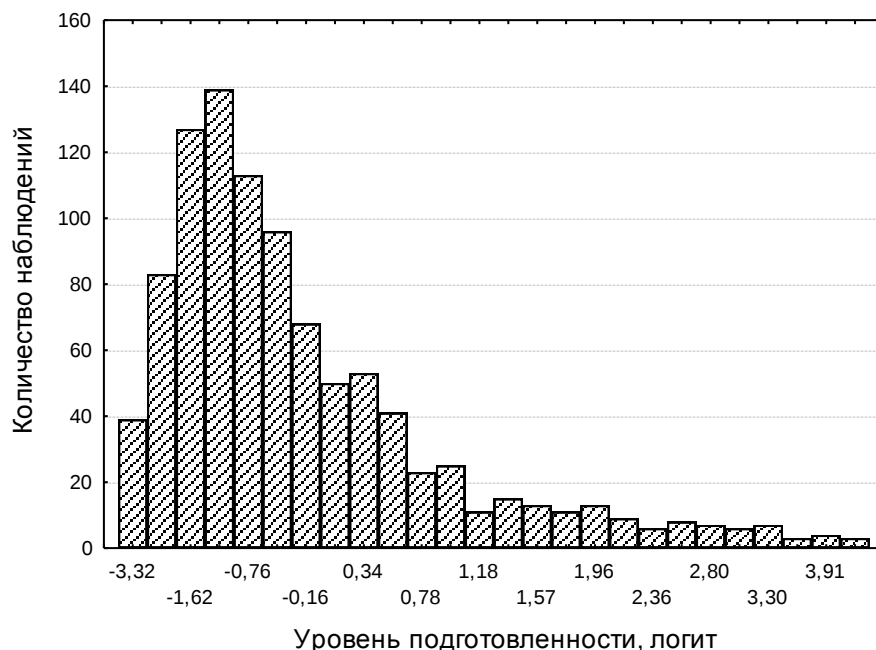


Рис. Гистограмма уровня подготовленности тестируемых в 2012 году.

Характеристики трудности тестовых заданий, уровня подготовленности тестируемых по годам и их разность представлены в табл. 3.

Таблица 3

Характеристики трудности тестовых заданий и уровня подготовленности тестируемых

Год	Количество тестируемых	Трудность тестовых заданий (интервал изменения), логит	Уровень подготовленности (интервал изменения), логит	Разность между уровнем подготовленности и трудностью теста, логит
2006	1994	0,15 (-1,48; 3,85)	-0,47 (-2,50; 5,98)	-0,62
2007	1979	0,24 (-1,90; 4,50)	-0,23 (-2,82; 6,38)	-0,47
2008	1508	0,39 (-1,97; 4,08)	-0,45 (-3,18; 5,10)	-0,84
2009	1844	0,21 (-4,40; 5,40)	-0,62 (-4,44; 7,37)	-0,83
2010	1475	0,23 (-2,99; 3,89)	-0,28 (-3,04; 4,72)	-0,51
2011	1132	0,44 (-2,51; 7,77)	-0,79 (-3,84; 7,83)	-1,23
2012	973	0,05 (-1,94; 6,07)	-0,77 (-3,32; 4,29)	-0,82

В 2006–2012 годах разность между уровнем подготовленности и трудностью теста была отрицательной и колебалась в пределах от -1,23 (2011 г.) до -0,47 (2007 г.) логита. Это означает, что в среднем тестируемые решали задания с вероятностью правильного ответа от 11% до 31%.

Для ответа на вопрос, имеются ли статистически значимые различия между уровнем подготовленности тестируемых в разные годы, было проведено сравнение групп. Попарное сравнение выборок по годам с использованием критерия Манна–Уитни дало следующие результаты (табл. 4).

Таблица 4

Попарное сравнение выборок по годам

Год тестирования	Уровень значимости p					
	2007	2008	2009	2010	2011	2012
2006	0,3966	0,3967	0,1333	0,0084	0,0001*	0,0000*
2007		0,2583	0,0001*	0,0087	0,0000*	0,0000*
2008			0,0495	0,0230	0,0000*	0,0000*
2009				0,0001*	0,0004*	0,0068
2010					0,0000*	0,0000*
2011						0,0693

* Уровень значимости $p < 0,002$ (с учетом поправки Бонферрони для множественных сравнений).

Наиболее значимые отличия в уровне подготовленности тестируемых существуют между абитуриентами 2007 и 2009, 2009 и 2010 годов. Также следует отметить, что в 2011 и 2012 годах уровень подготовленности тестируемых при неизменной трудности тестовых заданий уменьшается.

Далее были проанализированы связи результатов зимней сессии студентов математического факультета с результатами ЦТ. Всего в 2006–2011 годах на первом курсе математического факультета обучались 566 студентов, из них успешно завершили первый семестр 524 студента. По результатам первой сессии были выделены три группы студентов: в первую группу вошли студенты, успешно и в срок сдавшие зимнюю сессию, во вторую – имевшие одну или две задолженности, которые были ликвидированы в установленные сроки, в третью – отчисленные за академическую неуспеваемость.

Проверка результатов ЦТ и показателей успеваемости студентов на нормальность распределения дала отрицательный результат, поэтому в исследовании использовались непараметрические методы. Однако традиционно мы приводим средние значения рассматриваемых показателей (табл. 5).

Таблица 5

Результаты ЦТ, средний балл аттестата и успеваемость студентов 1 курса

Показатель	Среднее	Медиана	Минимум	Максимум
Все студенты первого курса (566 человек)				
Рус./бел. язык (ЦТ)	48,7	47,0	4,0	100,0
Физика (ЦТ)	30,8	28,0	3,0	96,0
Математика (ЦТ)	42,6	42,0	4,0	94,0
Аттестат	8,3	8,4	4,1	9,9
Средний балл по предметам математического цикла	5,6	5,3	4,0	10,0
Итоговый балл за сессию	5,7	5,5	4,0	10,0
Студенты, успешно и вовремя сдавшие первую сессию (398 человек)				
Рус./бел. язык (ЦТ)	52,0	51,0	10,0	100,0
Физика (ЦТ)	32,1	29,0	3,0	96,0
Математика (ЦТ)	45,8	45,0	4,0	94,0
Аттестат	8,5	8,6	5,6	9,9
Средний балл по предметам математического цикла	6,0	6,0	4,0	10,0
Итоговый балл за сессию	6,1	6,0	4,0	10,0
Студенты, имевшие задолженности в первую сессию (126 человек)				
Рус./бел. язык (ЦТ)	38,3	36,0	4,0	77,0
Физика (ЦТ)	26,9	25,0	8,0	69,0
Математика (ЦТ)	32,5	30,5	5,0	80,0
Аттестат	7,7	7,8	4,1	9,4
Средний балл по предметам	4,4	4,0	4,0	7,5

математического цикла				
Итоговый балл за сессию	4,5	4,3	4,0	6,7
Студенты, отчисленные в первую сессию (42 человека)				
Рус./бел. язык (ЦТ)	33,4	31,0	11,0	87,0
Физика (ЦТ)	19,6	16,5	2,0	64,0
Математика (ЦТ)	27,0	23,0	4,0	78,0
Аттестат	7,4	7,5	4,7	9,6

Для анализа взаимосвязи между результатами ЦТ и успеваемостью студентов применялся метод непараметрического корреляционного анализа. В табл. 6 приведены значения коэффициентов ранговой корреляции по Спирмену для рассматриваемых показателей успеваемости.

Таблица 6

Корреляционные зависимости между результатами ЦТ и успеваемостью студентов

Показатель успеваемости	Все студенты первого курса (566 человек)		Студенты, успешно сдавшие сессию (398 человек)		Студенты, имевшие задолженности (126 человек)	
	Ср. балл по предметам математич. цикла	Итоговый балл за зимнюю сессию	Ср. балл по предметам математич. цикла	Итоговый балл за зимнюю сессию	Ср. балл по предметам математич. цикла	Итоговый балл за зимнюю сессию
Рус./бел. язык (ЦТ)	0,408*	0,456*	0,337*	0,395*	0,154	0,179*
Физика (ЦТ)	0,353*	0,378*	0,390*	0,429*	0,190*	0,227*
Математика (ЦТ)	0,322*	0,406*	0,249*	0,334*	0,118	0,255*
Аттестат	0,456*	0,451*	0,382*	0,368*	0,186*	0,165

* Коэффициент корреляции статистически значим ($p < 0,05$).

Закключение. В результате проведения сравнительного анализа характеристик тестовых заданий в 2006–2012 годах существенных изменений в распределении заданий по трудности и дискриминативности не выявлено. Среди тестовых заданий преобладают задания среднего уровня сложности, что позволяет точнее определить средний уровень подготовки контингента участников тестирования. Большинство заданий обладают низкой и очень низкой дифференцирующей способностью. В то же время в тестах присутствуют очень трудные задания с низкой дискриминативностью, целесообразность включения этих заданий в тест вызывает сомнение.

Между результатами ЦТ по всем предметам и показателями успеваемости студентов математического факультета в зимнюю сессию наблюдается умеренная корреляция. Ожидания, что коэффициент корреляции между результатами ЦТ по математике и средним баллом по дисциплинам математического цикла будет близок к единице, не оправдались, самый высокий коэффициент корреляции с показателями успеваемости наблюдался для результатов ЦТ по русско-белорусскому языку и среднего балла аттестата. В группе студентов, имевших задолженности, корреляция между результатами ЦТ и успеваемостью практически отсутствует, это говорит о том, что результаты ЦТ не дают возможности статистически достоверно предсказать «успешность–неуспешность» обучения на первом курсе. Получение по предметам ЦТ низких баллов из диапазона 0–9 не означает, что студент не сможет успешно учиться.

ЛИТЕРАТУРА

1. Феськов, Н.С. Централизованное тестирование: направления, содержание, технологии / Н.С. Феськов // Адукацыя і выхаванне. – 2006. – № 3. – С. 49–52.
2. Феськов, Н.С. Математические методы интерпретации результатов нормативно-ориентированного тестирования / Н.С. Феськов, А.П. Якобчук // Адукацыя і выхаванне. – № 3. – 2007. – С. 76–94.
3. Феськов, Н.С. Педагогический тест как инструмент объективного измерения уровня подготовки / Н.С. Феськов, А.П. Якобчук // Адукацыя і выхаванне. – № 2. – 2007. – С. 47–64.

Поступила в редакцию 08.01.2013. Принята в печать 20.02.2013
Адрес для корреспонденции: e-mail: channa@tut.by – Чиркина А.А.