УДК 371.26

СТАТИСТИЧНИЙ АНАЛІЗ ОБ'ЄКТИВНОСТІ ТЕСТОВИХ ОЦІНОК З МАТЕМАТИЧНИХ ДИСЦИПЛІН

Лісова Т.В.

У статті розглядаються деякі аспекти зовнішньої валідності тестів, які використовуються викладачами вузу у поточному та підсумковому контролі з математичних дисциплін. Проводиться статистичний аналіз об'єктивності тестових оцінок шляхом їх порівняння з поточними оцінками за інші види роботи та з екзаменаційними оцінками, отриманими впродовж кількох років навчання, з основних математичних дисциплін. Аналізуються причини невисокої валідності та шляхи її підвищення.
Ключові слова: тестовий контроль, об'єктивність оцінювання, система оцінювання, валідність тесту.

В статье рассматриваются некоторые аспекты внешней валидности тестов, которые используются преподавателями вузов во время текущего и итогового контроля по математическим дисциплинам. Проводится статистический анализ объективности тестовых оценок путем их сравнения с текущими оценками за другие виды работы и с экзаменационными оценками, полученными на протяжении нескольких лет обучения, по основным математическим дисциплинам. Анализируются причины невысокой валидности и пути ее повышения.
Ключевые слова: тестовый контроль, объективность оценивания, система оценивания, валидность теста.

The author of the article investigates certain aspects of external validity of tests used at high school for progress monitoring and final control in mathematical disciplines. Statistical analysis of the objectivity of test scores is carried out through comparison with progress monitoring results and examination marks in major mathematical disciplines throughout several academic years. Causes of low validity and ways to increase it are analyzed.
Key words: testing, assessment objectivity, assessment system, test validity.

Функції викладача сучасного вузу не повинні обмежуватися лише накопиченням та поширенням наукової інформації. Викладач повинен також керувати пізнавальною діяльністю студентів та контролювати її результати. У процесі навчання завжди тією чи іншою мірою присутній контроль, результати якого виражаються певними оцінками або балами. Відношення до оцінювання результатів навчання не завжди було однозначним: від надмірної гуманізації навчального процесу до жорсткого контролю з примушенням до навчання. Правоту прибічників різних підходів визначає саме життя, але ціною масштабних новацій в освіті може бути ціле втрачене покоління. Наприклад, відміна оцінок, усіх екзаменів та різних перевірок у післяреволюційний час (1918) призвела до зниження якості освіти, погіршення дисципліни та втрати мотивації до навчання. Зворотній же процес посилення педагогічного контролю, який розпочався у 30-х рр. XX ст., призвів до кін. 50-х рр. до утвердження ідей авторитарності, антигуманізму та формалізму в освіті, а якість навчання при цьому не підвищилася [4, с. 20].

Сучасний процес освітньої діяльності також немислимий без контролю. Можуть змінюватися лише форми та методи його проведення, пріоритети в оцінках та способи їх виставлення, інтенсивність проведення контрольних заходів та ступінь впливу на учасників процесу навчання, акценти в інтерпретації результатів контролю. Контроль стимулює навчання і впливає на поведінку студентів, а з упровадженням нових інтенсивних методів навчання незмінно з'являються і нові методи та форми контролю, у яких тісно переплітаються його діагностична, навчальна та виховна функції. Традиційну систему контролю у вузі утворюють екзамени, заліки, усні опитування, письмові контрольні роботи, реферати, колоквіуми, лабораторні роботи, семінари, курсові, дипломні роботи тощо. Кожна із форм має свої особливості,

але вдале поєднання різних форм контролю свідчить про рівень організації навчального процесу у вузі і є одним із важливих показників педагогічної кваліфікації викладача. Завдяки технічному оснащенню вузів поряд із традиційними методиками оцінювання усе ширше застосовується тестовий контроль знань, змінюється психологічна звичка викладачів та студентів до самої процедури тестування. Педагогічне тестування як один із методів контролю засвоєння знань, умінь і навичок має ряд переваг перед традиційними методами контролю: 1) більш висока об'єктивність, оскільки до мінімуму зводиться вплив суб'єктивних факторів; 2) більш висока диференційність тестової оцінки, оскільки є можливість використовувати шкали, які містять більше градацій; 3) більш висока ефективність, оскільки тестування можна проводити одночасно з великими групами студентів за незначних затратах на обробку результатів. Але слід також розуміти, що тестування, як будь-який метод контролю, має свої обмеження та недоліки, особливо, якщо використовуються не стандартизовані тести. Тому у багатьох випадках найкращий результат отримується у поєднанні тестового та традиційного методів контролю. В умовах оцінювання успішності студентів за накопичувальною (наприклад, 100-бальною) шкалою залишається лише вдало вибрати види контролю та розподілити між ними кількість балів.

Не перший рік на фізико-математичному факультеті Ніжинського університету використовується тестовий поточний та підсумковий контроль у комп'ютерній або паперовій формі з багатьох математичних дисциплін, тому можливе деяке узагальнення набутого досвіду щодо побудови системи оцінювання, яка б поєднувала різні методи та форми контролю у достатній кількості для одержання об'єктивної картини про стан та рівень успішності з дисципліни. Причому кількість балів, яка відводиться на ту чи іншу форму контролю, повинна адекватно відображати важливість контролю та затрачені студентом зусилля для його проходження. Бажаною також є висока кореляція оцінок, отриманих за різними формами контролю, оскільки вона є свідченням певного рівня об'єктивності.

У даній роботі досліджується місце, вага та роль тестування у системі оцінювання знань студентів із дисциплін "Теорія ймовірностей та математична статистика" та "Математична логіка і теорія алгоритмів". Так, у системі оцінювання з дисципліни "Теорія ймовірностей та математична статистика" на підсумкове комп'ютерне тестування відводилося 25 балів, а на поточний письмовий контроль – 50 балів (5 письмових робіт) із 100 можливих. Решта балів – інші види роботи (розрахунково-графічна робота, колоквіум, усні відповіді). Така значна частка балів на тестування зобов'язує, щоб отримана студентом оцінка була надійною, об'єктивною та не випадковою. Було проведено аналіз умов комп'ютерного тестування та порівняльний аналіз його результатів з результатами поточного письмового оцінювання з даної дисципліни та з екзаменаційними оцінками з інших математичних дисциплін для 38 студентів 4 курсу спеціальності "Математика*" у трьох групах: МЕ (спеціалізація "основи економіки", 17 чол.), МІ ("основи інформатики та освітні вимірювання", 13 чол.) та МФ ("фізика", 8 чол.). Успішність з теорії ймовірностей та математичної статистики порівнювалася з успішністю з дев'яти основних математичних дисциплін, з яких студенти усіх груп мали підсумковою формою контролю екзамен: математичний аналіз; вища геометрія; аналітична геометрія; лінійна алгебра; алгебра і теорія чисел; інформатика; диференціальна геометрія і топологія; методика навчання математики та комплексний аналіз. Таке порівняння може свідчити про рівень зовнішньої валідності тесту та про об'єктивність підсумкової оцінки з даної дисципліни [4, с. 164].

Першою несподіванкою виявилося те, що тестові оцінки були у середньому на 25% нижчі, ніж бали поточного оцінювання у всіх групах (рис. 1), причому інколи ця різниця була відчутнішою для студентів, які мали високі бали за поточні письмові роботи. Останній факт уже вказує на проблеми у тесті, оскільки сильніші студенти, витрачаючи час на розв'язування завдань, припускалися помилок, тоді як слабкі, не впевнені у своїх знаннях студенти, отримали вищі бали за рахунок вдалого угадування.

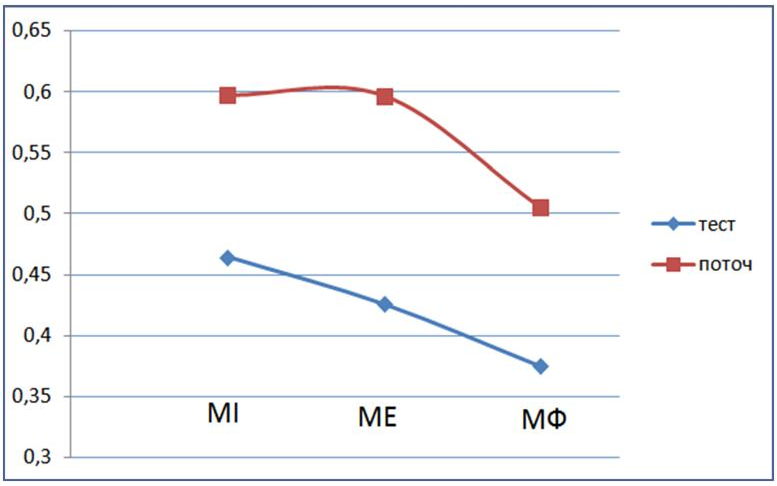

Рис.1. ***Співвідношення між тестовими та поточними балами***

Пояснити ж низькі результати тестового оцінювання порівняно з письмовим можна частково психологічними особливостями процедури тестування. Дійсно, тестування проводилося без попередньої підготовки, час тестування був обмежений, база завдань не була оприлюднена і студенти не мали змоги тренуватися із подібними завданнями у тестовому форматі. Можливо почуття тривоги та напруженості у непідготовлених до тестування в умовах обмеженого часу учасників призвело до занижених результатів. Відомі дослідження [1, с. 48], коли результати тестування учасників, які мали досвід роботи з тестами, були на 11% вищі, ніж результати у групі непідготовлених до роботи з тестовими завданнями учасників. Тому важливо оцінити статистичну значущість відхилення результатів тестування та письмового оцінювання з теорії ймовірностей та математичної статистики, а також порівняти результати тестового, поточного, і як наслідок, підсумкового оцінювання з даної дисципліни із зовнішнім критерієм, наприклад, середнім балом екзаменаційних оцінок з дев'яти основних математичних дисциплін.

Було встановлено, що коефіцієнт кореляції тестових та поточних оцінок із середнім балом був значно вищим у групі МЕ (таблиця 1), яка з теорії ймовірностей та математичної статистики мала екзамен, а не залік, як у інших групах. Це означає, що на даному етапі не зовсім коректно порівнювати екзаменаційні та залікові оцінки, хоча вони і виставляються за єдиною 100-бальною шкалою. У випадку, коли навчання відбувається заради оцінки, а не здобуття знань, така ситуація легко пояснюється небажанням студентів підвищувати свою залікову оцінку, оскільки вона не впливає на призначення стипендії. Очевидно, що картина могла б бути зовсім іншою, якби всі оцінки мали однакову вагу.

Таблиця 1

Група	***Середній бал***	***Коефіцієнт кореляції для тестових оцінок***	***Коефіцієнт кореляції для поточних оцінок***
МІ	74,4	0,3380	0,4266
МЕ	73,4	0,6262	0,8558
МФ	69,9	0,5410	0,6942

Встановлено також, що у всіх групах коефіцієнт кореляції поточних оцінок із середнім балом з дев'яти основних математичних дисциплін був вищим, ніж відповідний коефіцієнт для тестових оцінок. Отже, результати тестування значно гірші за результати поточного оцінювання, причому мають нижчий коефіцієнт кореляції із середнім балом екзаменаційних оцінок. Щоб встановити, чи є ця різниця випадковою, чи зумовлена саме процедурою тестування, проведемо статистичну перевірку гіпотези H_0 про відсутність відхилення у способах оцінювання, попередньо дослідивши можливість об'єднання результатів трьох груп в одну.

Оскільки в усіх трьох групах навчання та оцінювання відбувається в однакових умовах, тому немає підстав вважати, що різні значення середнього бала у різних групах зумовлені виключно складом групи, а не випадковими чинниками. Це можна підтвердити дослідженням однорідності трьох груп за ранговим критерієм Краскела-Уолліса, який застосовується до даних без вимоги їх нормальності [8, с. 193]. Якщо у кожній групі розподіли оцінок є нормальними, можна також провести дисперсійний аналіз для виявлення впливу фактора групи в основній гіпотезі про рівність середніх балів у всіх групах. Застосування спеціальних статистичних пакетів (Statistica, SPSS тощо), або навіть використання деяких додаткових можливостей MS Excel дозволяє проводити зручно та швидко такий аналіз. На рис. 2 наведено результат застосування критерію Краскела-Уолліса у пакеті Statistica до середніх екзаменаційних оцінок у трьох групах. На рівні значущості $\alpha = 0{,}05$ можна стверджувати, що групи є однорідними (спостережене значення критерію не потрапляє у критичну область, значення ймовірності *p* є достатньо високим), тому їх об'єднання в одну групу для подальшого дослідження є обґрунтованим.

Kruskal-Wallis ANOVA by Ranks; Var1 (Spreadsheet3 in Workbook11)
Independent (grouping) variable: Var2
Kruskal-Wallis test: H (2, N= 38) =1,193758 p =,5505

Depend.: Var1	Code	Valid N	Sum of Ranks
1	1	13	282,0000
2	2	17	329,0000
3	3	8	130,0000

Рис. 2. ***Дослідження однорідності груп у Statistica***

Аналогічний висновок отримуємо і за результатами дисперсійного аналізу, проведеного в MS Excel за допомогою процедури "*Однофакторный дисперсионный анализ*" з пакета "*Анализ данных*", попередньо встановивши нормальність розподілів середніх балів у кожній групі. Спостережене значення критерію F значно менше за критичне (не потрапляє у критичну область), значення *p* достатньо високе, тому приймаємо гіпотезу про відсутність відхилень для середніх балів у трьох

групах. Тобто вищий середній бал у групі MI не зумовлений її винятковістю, а знаходиться на одному рівні із балами в інших групах у межах статистичної похибки $\alpha = 0{,}05$. Тому нема підстав розглядати вплив способу оцінювання у кожній групі окремо. Об'єднання їх в одну групу дозволить збільшити об'єм вибірки (38 чол.) та отримати більш достовірні висновки.

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
MI	13	967,7778	74,44444	51,33128
ME	17	1248	73,41176	125,2913
MF	8	559,4444	69,93056	41,58355

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	105,3384	2	52,66921	0,633104	0,5369121	3,267423525
Внутри групп	2911,721	35	83,19203			
Итого	3017,059	37				

Рис. 3. ***Дослідження однорідності груп у MS Excel***

Для виявлення впливу методики оцінювання будемо використовувати *t*-критерій Стьюдента для середніх двох нормально розподілених сукупностей тестових та поточних оцінок. Перевірка нормальності розподілів поточних (рис. 4) та тестових (рис. 5) оцінок проводилася у пакеті Statistica за критерієм хі-квадрат. На рівні значущості обидва розподіли можна вважати нормальними, хоча для тестових оцінок спостережене значення критерію хі-квадрат дуже близьке до критичної області. Це вже є деяким свідченням проблем у тесті. Переважна частка студентів мали схожі результати у діапазоні від 30 до 40% правильних відповідей, що менше від середнього значення. Тобто у цілому для даної групи студентів тест виявився занадто складним.

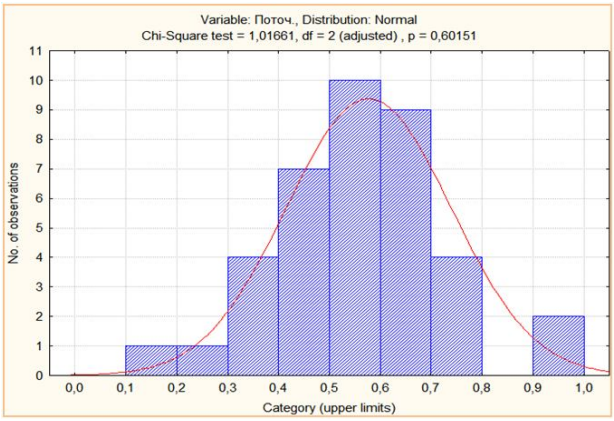

Рис. 4. ***Перевірка нормальності розподілу поточних оцінок***

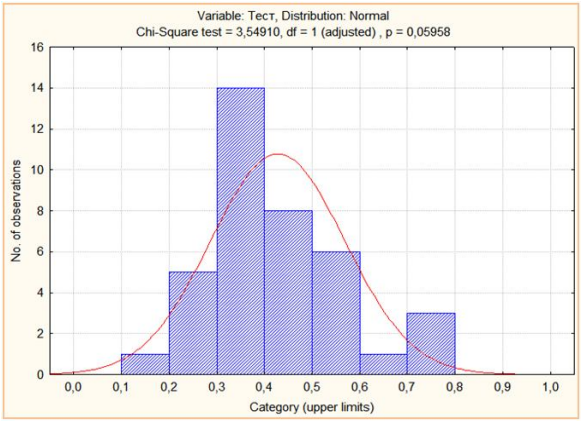

Рис. 5. ***Перевірка нормальності розподілу тестових оцінок***

Результати перевірки гіпотези про відсутність впливу способу оцінювання за допомогою процедури "*Парный двухвыборочный t-тест для средних*" з пакета "*Анализ данных*" MS Excel представлено на рис. 6. Спостережене значення критерію Стьюдента (7,89) значно перевищує критичне для обох типів областей за рівнем значущості 0,05, отже приймаємо альтернативну гіпотезу про те, що середні значення тестових та поточних оцінок суттєво відрізняються, причому ця різниця статистично значуща і зумовлена не випадковими чинниками, а саме процедурою оцінювання. Тому можна зробити висновок, що за результатами одноразового тестування отримана, швидше за все, не зовсім об'єктивна картина успішності з даної дисципліни. Для дослідження впливу тестових оцінок на підсумкову оцінку з теорії ймовірностей та математичної статистики (ТЙіМС) розглядались коефіцієнти кореляції підсумкової оцінки з екзаменаційними оцінками з інших дев'яти дисциплін, серед яких лише з інформатики використовувався тестовий контроль. Коефіцієнт кореляції оцінок з ТЙіМС знаходився у межах від 0,5 (лінійна алгебра) до 0,72 (комплексний аналіз). Така кореляція на загальному фоні не виглядає катастрофічною. Оцінки з методики навчання математики мають коефіцієнт кореляції 0,4 з оцінками з геометрії та з математичного аналізу. Найвищий коефіцієнт кореляції (0,82) мають оцінки з лінійної алгебри з оцінками з аналітичної та

Парный двухвыборочный t-тест для средних		
	Поточне оцінюв.	*Тест*
Среднее	0,577368421	0,428
Дисперсия	0,026187482	0,02
Наблюдения	38	38
Корреляция Пирсона	0,712718496	
Гипотетическая разность	0	
df	37	
t-статистика	7,894022087	
P(T<=t) одностороннее	9,45379E-10	
t критическое односторо	1,687093597	
P(T<=t) двухстороннее	1,89076E-09	
t критическое двухсторон	2,026192447	

Рис. 6. ***Дослідження впливу методики оцінювання***

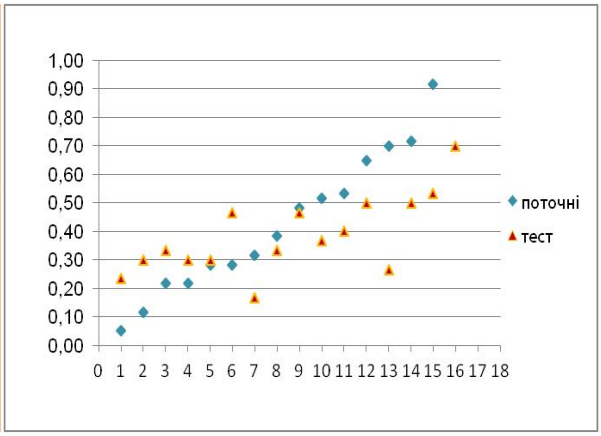

Рис. 7. ***Результати тестування у 1 семестрі***

диференціальної геометрії. Оскільки поточні оцінки мають вищу кореляцію із середнім балом ніж тестові (таблиця 1), то можна зробити висновок, що саме тестові оцінки мали вплив на корельованість оцінок з ТЙіМС з іншими оцінками. Але варто також урахувати, що вибір зовнішнього критерію був не зовсім вдалим, бо оцінки у двох групах з ТЙіМС, які порівнювались з екзаменаційними, були заліковими.

У результаті даного дослідження може сформуватись упереджена думка щодо об'єктивності тестового оцінювання та можливостей тестових завдань нести інформацію про рівень підготовленості учасників тестування, але це зовсім не означає, що потрібно відмовитися від використання тестових методик у практиці оцінювання. Адже професійно складені тестові завдання дозволяють за короткий проміжок часу перевірити знання великої кількості студентів з деяких розділів, а то і з усієї дисципліни, що є дуже важливим в умовах постійно зростаючого навантаження. На основі отриманих результатів можна зробити певні рекомендації щодо визначення місця тестування у системі оцінювання: потрібно зменшити вагу одного тестування, проводити кілька серій тестувань, включаючи і пробні, щоб зменшити психологічний вплив несподіванки та тривожності, практикувати тестові завдання для самоконтролю. Краще, коли банк завдань з однієї дисципліни формує кілька викладачів, тоді кожен може бути експертом для іншого. За можливості необхідно проводити аналіз тестових завдань на складність (наприклад, засобами IRT [6]), відбираючи до тесту завдання різних рівнів складності з найбільшою диференційною здатністю. Важливо також розрахувати оптимальний час тестування.

Вплив "тренованості" студентів можна прослідкувати у разі дослідження результатів тестування з дисципліни "Математична логіка та теорія алгоритмів". У першому семестрі комп'ютерне тестування з вагою 30 балів проводилося без попередньої підготовки (рис. 7). Коефіцієнт кореляції тестового та поточного письмового контролю становив 0,73. Виявлена значна кількість студентів із низькими балами поточного контролю, які мали вищі бали тестового контролю, і навпаки, що свідчить про можливу низьку валідність такого тесту. У наступному семестрі у цій же групі систему оцінювання було вдосконалено: зменшена кількість балів на тестування до 25, оприлюднена у письмовій формі частина бази тестових завдань для самоконтролю та підготовки до комп'ютерного тестування. У результаті отримано високий коефіцієнт кореляції 0,9 тестових та поточних оцінок з даної дисципліни, відсутні несподівані результати (рис. 8).

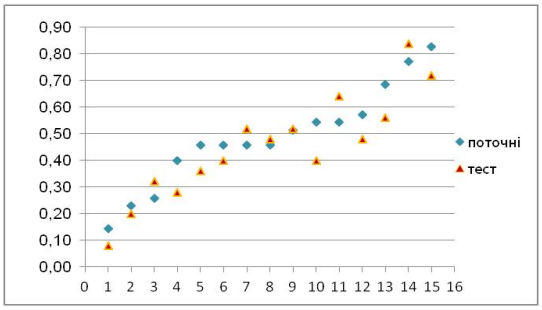

Рис. 8. ***Результати тестування у 2 семестрі***

Запропонована методика дослідження об'єктивності тестових оцінок та визначення зовнішньої валідності тесту дозволяє виявити проблеми у створенні тестів та у процедурі проведення тестування. Але все ж таки для проведення порівняльного аналізу результатів оцінювання з різних дисциплін у межах університету бажано розробити єдині уніфіковані критерії оцінювання хоча б для дисциплін схожих циклів, єдині вимоги до складання тестових завдань та процедури проведення тестування, розробити методику визначення показників валідності та надійності тесту, інтерпретації результатів тестування.

Література

1. Анастази А. Психологическое тестирование / А. Анастизи, С. Урбина. – 7-е изд. – СПб. : Питер, 2007. – 688 с.

2. Вища освіта України і Болонський процес : навчальний посібник. – Тернопіль : Навчальна книга – Богдан, 2004. – 384 с.

3. Гласс Дж. Статистические методы в педагогике и психологии / Дж. Гласс, Дж. Стэнли. – М. : Прогресс, 1976. – 494 с.

4. Звонников В. И. Современные средства оценивания результатов обучения : учеб. пособие для студ. высш. учеб. заведений / В. И. Звонников, М. Б. Челышкова. – М. : Издательский центр “Академия”, 2007. – 224 с.

5. Ким В. С. Тестирование учебных достижений : монография / В. С. Ким. – Уссурийск : Издательство УГПИ, 2007. – 214 с.

6. Лісова Т. В. Аналіз тесту з дисципліни “Теорія керування” засобами IRT / Т. В. Лісова // Науковий часопис НПУ ім. М.П. Драгоманова. Серія № 5 “Педагогічні науки: реалії та перспективи” : зб. наукових праць. – К. : Вид-во НПУ, 2010. – Вип. 23. – С. 174–179.

7. Сидоренко Е. В. Методы математической обработки в психологии / Е. В. Сидоренко. – СПб. : ООО “Речь”, 2003. – 350 с.

8. Тюрин Ю. Н. Анализ данных на компьютере / Ю. Н. Тюрин, А. А. Макаров ; под ред. В. Э. Фигурнова. – М. : ИНФРА-М, 2003. – 544 с.

9. Ravid R. Practical statistics for Educators / Ruth Ravid. – Lanham : University Press of America, 2005. – 230 p.