УДК 303.714, УДК 681.5.01

В.В. Шведова

Національний технічний університет України «Київський політехнічний інститут», Київ

ОЦІНЮВАННЯ НАДІЙНОСТІ ТЕСТОВОГО ПРОСТОРУ КОМПЮТЕРИЗОВАНОЇ СИСТЕМИ ТЕСТУВАННЯ: МЕТОДИ ТА СПОСОБИ ОЦІНЮВАННЯ ПОКАЗНИКА ТА ЙОГО НЕВИЗНАЧЕНОСТІ

Автором систематизовано та проаналізовано методи та способи оцінювання одного з головних показників якості комп'ютеризованої системи тестування (КСТ) – показника надійності; наведено традиційні метрики для оцінювання різних видів надійності тестового простору та проведено їх удосконалення з огляду на особливості шкали, в якій оцінюється показник надійності тестового простору КСТ. Запропоновано аналітичні вирази для оцінювання невизначеності показника надійності тестового простору КТС. Наведено динаміку зміни невизначеності показника надійності при здійсненні моніторингу системи, що дозволяє прогнозувати якісні характеристики КСТ на основі заданого об'єму тестового простору.

Ключові слова: *комп'ютеризована система тестування, тестовий простір, надійність тестового простору, невизначеність.*

Вступ

Основною задачею будь-якої освітньої програми навчання, розробленої з використанням комп'ютерного засобу навчання з комп'ютеризованою системою тестування (КСТ) є забезпечення високої якості навчальної та контрольно-оціночної функції системи [1 – 3], що не поступається аналогічній в традиційній формі навчання. При цьому, КСТ часто виступають як допоміжні засоби, які використовуються при традиційних формах навчання, що значно полегшує діяльність викладача та підвищує ефективність їх праці. КСТ здійснює оцінювання рівня підготовленості студентів на різних етапах навчання, тому більша частина контрольних заходів при залученні комп'ютеризованих засобів навчання проводиться за допомогою КСТ [4 – 10].

Всі переваги автоматизованого контролю з використанням тестових завдань, які зазначені в [11, 12], можна отримати лише при використанні якісної КСТ, яка б забезпечувала отримання дійсно об'єктивних, обґрунтованих оцінок рівня підготовленості [13, 14], виражених в певних одиницях, що супроводжуються значеннями невизначеності цих оцінок; оцінки рівнів підготовленості мають забезпечувати чітке та однозначне їх розуміння різними суб'єктами навчального процесу та бути порівнюваними між собою.

Для реалізації поставлених вимог КСТ має задовольняти певним критеріям щодо її якості [15, 16]. Серед характеристик системи, що безпосередньо визначають якість її функціонування, виділяють показники якості (ПЯ), які підлягають оцінці та перевірці на відповідність щодо норм як перед впровадженням системи в експлуатацію, так і безпосередньо в процесі її експлуатації [17, 18]. Оцінювання ПЯ КСТ впродовж експлуатації повинне мати періодичний характер, що найкращим чином може бути реалізовано шляхом розробки системи моніторингу, яка б в автоматичному або напівавтоматичному режимі реалізувала поставлені задачі.

Тестування з метою оцінювання рівня підготовленості учнів за своєю суттю є емпіричною процедурою отримання первинних даних, які підлягають статистичній обробці за певним алгоритмом (з використанням комп'ютерної системи для проведення обчислень). Результати тестування використовуватимуться в двох аспектах (в залежності від стадії розробки КСТ): на стадії стандартизації – для оцінювання ПЯ КСТ та порівняння отриманих оцінок із встановленим нормами; на стадії експлуатації: для контролю оцінок ПЯ КСТ, отриманих на стадії стандартизації; для уточнення попередніх оцінок ПЯ шляхом зменшення їх невизначеності за рахунок збільшення емпіричних даних при використанні спеціальних алгоритмів обробки даних; оцінювання рівня підготовленості студентів та уточнення і корегування попередніх оцінок рівня підготовленості студентів шляхом генерування індивідуальних рішень, що ґрунтуються на використанні інформації про ПЯ тестових просторів КСТ.

Серед ПЯ КСТ, що характеризують якість процедуру тестування та окремих завдань тесту, найчастіше згадують такі [11 – 13, 19 – 23]:

- **надійність** (характеризує точність процедури тестування та її інваріантність до випадкових чинників, що впливають або можуть впливати на результат дослідження);
- **валідність** (характеризує адекватність процедури тестування поставленій меті);
- **трудність** (характеризує адекватність процедури тестування вибірці опитуваних за складністю змістовного матеріалу);

▪ **розрізняльна здатність** (характеризує чутливість процедури тестування до індивідуальних особливостей (відмінностей) учнів).

На сьогоднішній день використовується два підходи оцінювання ПЯ КСТ, що визначаються шляхом експерименту та пов'язані з контрольно-оціночною функцією системи: класична теорія обробки результатів тестування та теорія моделювання та параметризації тестів (ТМПТ). Автором проведено огляд, класифікацію та критичний аналіз існуючих методів та способів оцінювання ПЯ КСТ, а також запропоновано нові способи оцінювання невизначеності зазначених ПЯ. При цьому автором зазначено, що питання визначення характеристики точності ПЯ майже не висвітлюється в літературі; виключення становить наукова праця [13]. В той же час, при проведенні оцінювання якості КСТ, призначеної для технічних спеціальностей, шляхом встановлення ПЯ, точність оцінювання цих показників є одним з визначальних факторів, що обумовлює необхідність додаткових досліджень КСТ в процесі її експлуатації.

Рис. 1 ілюструє можливості класичного підходу та ТМПТ оцінити окремі ПЯ КСТ.

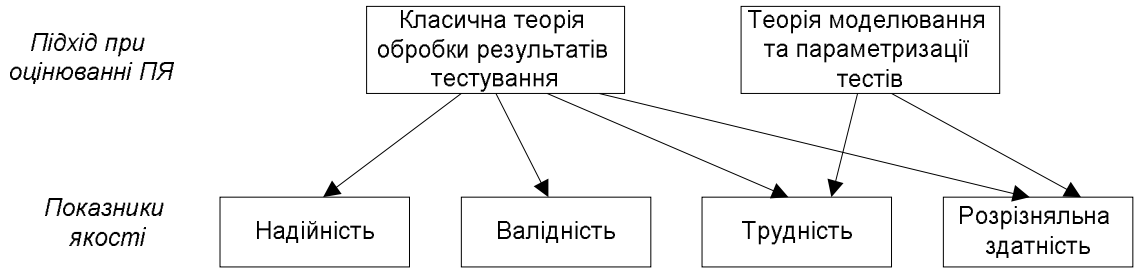

Рис. 1. Систематизація ПЯ КСТ, що оцінюють в рамках класичного підходу та ТМПТ

Оцінювання показника надійності тестового простору КСТ

В статті автором розглянуто та систематизовано методи та способи оцінювання надійності процедури тестування та окремих завдань тесту. Надійність є одним з основних ПЯ, що характеризує придатність тестового простору КСТ до використання. За визначеннями, наведеними в [11, 12, 21, 24] надійність розглядають в декількох аспектах:

а) характеристика точності результатів тестування та їх стійкості до випадкових чинників [12];

б) стійкість результатів, отриманих при повторному контролі [11]; в) характеристика узгодженості (однорідності) завдань в тесті.

Основними методами визначення надійності процедури тестування є такі (табл. 1):

- **Метод «тест-ретест»** (або метод повторного тестування). Метод полягає в багаторазовому повторенні процедури тестування за одним і тим самим тестом на одній і тій самій вибірці опитуваних. Така процедура наближає метод визначення надійності тестового простору до теоретичної моделі надійності, згідно з якою істинний бал опитуваного можна отримати за результатами відповідей опитуваного на даний тест нескінченну кількість разів [12]. Практичне ж застосування даного методу обмежене, зокрема, специфічними вимогами до змісту завдань тесту [11, 19].

- **Метод з використанням паралельних форм тесту**. Метод полягає в багаторазовому повторенні процедури тестування за різними формами тесту. При цьому форми тесту повинні задовольняти вимогу взаємної еквівалентності, тобто бути повністю взаємозамінними при проведенні тестування в процесі експлуатації тестового простору (звідси альтернативні назви – метод еквівалентних або взаємозамінних форм).

- **Метод еквівалентних частин тесту**. Метод полягає в умовному розщепленні тесту на еквівалентні частини та проведенні одного сеансу тестування з метою визначення показника надійності тесту. Тобто опитуваним пропонується єдина форма тесту, а проведене умовне розщеплення тесту на частини визначає лише спосіб отримання потрібного показника. При цьому спосіб розщеплення тесту на частини впливає на показник надійності тесту в контексті того, що недостатня еквівалентність частин призводить до додаткової похибки в оцінці показника і, як наслідок, до зниження його числового значення.

- **Метод, оснований на внутрішній узгодженості завдань тесту**. Даний метод являє собою граничний випадок попереднього, коли кількість частин, на які розщеплюється тест, дорівнює кількості завдань в ньому. Таким чином, метод передбачає одноразову процедуру тестування єдиною формою тесту. Поряд з цим специфічність цього методу породила розмаїття способів, за якими знаходять показник надійності тесту [11].

Для оцінювання надійності окремих завдань тесту використовують метод, оснований на визначенні взаємної узгодженості завдань тесту між собою. При цьому використовується одноразова процедура тестування єдиною формою тесту [25].

Кожен з методів дозволяє оцінити процедуру тестування з різних сторін і, як слушно зазначено в [26], при дослідженні тестового простору надійність, якщо це можливо, доцільно оцінювати різними методами.

Таблиця 1

Методи та способи оцінювання надійності процедури тестування та окремих завдань тестового простору КСТ

Кількість сеансів тестування, необхідних для реалізації методу	Показник, що визначається	Метод визначення показника	Спосіб оцінювання показника за обраним методом
Багаторазова процедура тестування	Надійність процедури тестування	Метод «тест-ретест»	За коефіцієнтом кореляції моментів Пірсона
			За коефіцієнтом рангової кореляції Спірмена
		Метод еквівалентних форм тесту	За коефіцієнтом рангової кореляції Кендала
Одноразова процедура тестування		Метод еквівалентних частин тесту	За рівнянням Спірмена-Брауна
			За формулою Крістофа
			За формулою Фланагана
			За формулою Рюлона
		Метод, оснований на визначенні внутрішньої узгодженості завдань в тесті	За формулою Кьюдера-Річардсона (для дихотомічної шкали оцінювання результатів тестування з використанням коефіцієнту дискримінативності завдань тесту)
			За формулою Гуліксена
			За формулою Кьюдера-Річардсона (для дихотомічної шкали оцінювання результатів тестування без використання коефіцієнту дискримінативності завдань тесту)
			За формулою коефіцієнта альфа (для випадку політомічної оцінки результату тестування)
	Надійність окремих завдань тесту	Метод, оснований на визначенні взаємної узгодженості завдань	За середнім значенням коефіцієнта кореляції даного завдання з іншими завданнями

Кожен з методів дозволяє оцінити процедуру тестування з різних сторін і, як слушно зазначено в [26], при дослідженні тестового простору надійність, якщо це можливо, доцільно оцінювати різними методами. Наведені в табл. 1 методи оцінювання показника надійності процедури тестування запропоновані в рамках класичного підходу. ТМПТ на даний час не має специфічних методів оцінювання цього показника [13]. В той же час в контексті ТМПТ висока надійність процедури тестування має забезпечуватись отриманням стійких та незміщених оцінок рівнів трудності завдань тесту та рівнів підготовленості опитуваних.

Методи, основані на повторному тестуванні, передбачають використання довгих тестів (більше 30 завдань [12]), при цьому завдання повинні бути не складними (дається 1,5 – 3 хвилини з розрахунку на розв'язання одного завдання при сумарному часі тестування 30 – 90 хв. [27]). Такі методи оцінювання показника застосовують для тестового простору, що перевіряють знання фактичного матеріалу або містять завдання на прості розрахунки. В той же час багато тестових просторів з технічних спеціальностей містить задачі, що перевіряють не лише знання, але й вміння та досвід застосовування певних алгоритмів, підходів, методів при розв'язанні задач. Однаковий розв'язок (або навіть схожий алгоритм розв'язання) задачі при повторному тестуванні або при тестуванні еквівалентною формою тесту дасть невиправдано завищений показник надійності тесту, тим самим спотворивши картину щодо істинної надійності тесту. Частою перешкодою для застосування методу еквівалентних частин тесту є складність або неможливість адекватного розбиття тесту на дійсно еквівалентні частини через специфічність змісту тестових завдань.

При проведенні моніторингу ПЯ КСТ в процесі експлуатації до зазначених недоліків методів визначення надійності процедури тестування шляхом багаторазового тестування додаються такі: неможливість проведення повторного тестування на одних і тих самих вибірках опитуваних (метод тест-ретест, метод еквівалентних форм тесту), випадковий вибір завдань тесту або адаптивна процедура тестування унеможливлюють створення еквівалентних частин тесту (метод еквівалентних частин тесту).

Тому при створенні інформаційно-вимірювального та алгоритмічного забезпечення моніторингу КСТ доцільно реалізувати оцінювання надійності тестового простору за одноразовою процедурою тестування єдиною формою тесту.

Аналітичні вирази для способів оцінювання надійності процедури тестування за одноразовою процедурою тестування єдиною формою тесту дозволив обрати спосіб оцінювання показника підчас пілотного експерименту та дослідної експлуатації КСТ. Визначити надійність тесту, як узгодженість завдань з яких він складається можна таким чином [11]:

- За формулою Кьюдера-Річардсона (для дихотомічної шкали оцінювання результатів тестування з використання коефіцієнта дискримінативності завдань тесту):

$$R_{KR\_diff} = 0,5 \cdot \sigma_x^{-2} \left(\sigma_x^2 - \sum_{j=1}^{k} p_j (1 - p_j) \right) + \sqrt{\sigma_x^{-2} \sum_{j=1}^{k} r_{pb\,j}^2 p_j (1 - p_j) + 0,5 \cdot \sigma_x^{-2} \left(\sigma_x^2 - \sum_{i=1}^{k} p_j q_j \right)}, \quad (1)$$

де σ_x – середнє квадратичне відхилення результатів тестування, k – кількість завдань в тесті, p_j – відносна кількість опитаних, які виконали j-е завдання, r_{pbj} – коефіцієнт дискримінативності (розрізняльної здатності) j-го завдання.

- За формулою Гуліксена (більш спрощений варіант порівняно з попереднім для дихотомічної шкали оцінювання результатів тестування):

$$R_G = k\cdot(k-1)\times \times\left[1-\left(\sum_{j=1}^{k} r_{pbj}\sqrt{p_j(1-p_j)}\right)^{-2}\sum_{j=1}^{k} p_j(1-p_j)\right]. \quad (2)$$

- За формулою Кьюдера-Річардсона (для дихотомічної шкали оцінювання результатів тестування без використання коефіцієнта дискримінативності завдань тесту):

$$R_{KR} = k\cdot(k-1)\cdot\sigma_x^{-2}\left(\sigma_x^2-\sum_{j=1}^{k} p_j(1-p_j)\right). \quad (3)$$

- Формула коефіцієнта альфа (для політомічної оцінки результатів тестування):

$$R_\alpha = k\cdot(k-1)\cdot\sigma_x^{-2}\left(\sigma_x^2-\sum_{j=1}^{k}\sigma_{xj}^2\right), \quad (4)$$

де $\sum_{j=1}^{k}\sigma_{xj}^2$ – сума дисперсій результатів за окремими завданнями.

Коефіцієнт дискримінативності, що використовується в формулі Кьюдера-Річардсона (1) та в формулі Гуліксена (2), є одним з ПЯ тестового простору. Тому з метою забезпечення незалежності оцінок ПЯ тестового простору (при умові дихотомічного оцінювання результатів тестування) обрано спосіб розрахунку надійності процедури тестування за формулою Кьюдера-Річардсона (3).

Аналіз статистичних оцінок, що отримують за первинними результатами тестування та традиційно використовують при розрахунку показника надійності процедури тестування за формулою (3), дозволив виявити наступний недолік.

Досліджуваними властивостями в роботі є ПЯ КСТ. Згідно [28] всі властивості (матеріальні і нематеріальні), що характеризують об'єкти матеріального світу, розподіляються на номінальні властивості (nominal properties) і величини (quantities). Якщо властивість об'єкта може проявлятись в більшому або меншому ступеню, тобто підлягає кількісній оцінці, її називають величиною. Величина [29] – властивість спільна в якісному відношенні у багатьох об'єктів та індивідуальна у кількісному відношенні у кожного з них. Величина є емпіричною системою з відношенням $Б = \langle A; R\rangle$, де $A = \{a_1, a_2, \ldots, a_n\}$ – сукупність проявів властивості, R – сукупність відношень між цими проявами. У загальному випадку до R входять відношення еквівалентності, порядку і комбінування $R = \langle \sim; \succ; o\rangle$ [30 – 31]. Відображенням емпіричної системи при вимірюванні є числова система з відношеннями $\mathbf{N} = \langle N, P\rangle$, де N – ряд чисел, а P – ряд відношень, визначених на N. У загальному випадку $P = \langle =; >; +\rangle$. Репрезентаційна умова [32] вимагає, щоб вимірювання було встановлено у відповідності між проявами величини і числами таким чином, щоб відношення між розглянутими проявами величини припускали відношення між їх зображеннями на ряді чисел [33].

При обробці даних тестування розраховують первинні бали завдань тесту c_j, j – номер завдання тесту – як суму правильних відповідей на завдання тесту, та первинні бали опитуваних b_i, i – порядковий номер опитуваного – як суму правильних відповідей кожного опитуваного на тест, що є достатніми статистиками множини вихідних даних тестування [13]. Множина первинних балів завдань тесту c_j та первинних балів опитаних b_i, які використовуються для оцінювання показника надійності процедури тестування, є емпіричною системою, прояви властивості якої знаходяться у відношенні еквівалентності і порядку: $Q = \langle Q; \sim; \succ\rangle$, де Q – прояви властивості; $\sim$, $\succ$ – емпіричні відношення еквівалентності та порядку відповідно. Для зазначених статистик відсутнє відношення пропорційності розмірів властивості та відношення інтервалів. Тому вихідні дані тестування відображаються ординальною шкалою $N = \langle N; =; >\rangle$ з допустимими операціями над елементами шкали в переліку: більше, менше або дорівнює. Таким чином, операції отримання середнього арифметичного та середнього квадратичного відхилення за відомими формулами [34]:

$$\bar{x} = \sum_{i=1}^{n} x_i / n \quad \text{та} \quad \sigma_x = \sqrt{(n-1)^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

(n – кількість опитуваних, x_i – первинний бал i-го опитуваного) присутні в способах визначення надійності за формулами (1) – (3) є допустимими тільки для метричних шкал, на які розповсюджені відношення пропорційності і різниць. Проблема відсутності метричних шкал в області педагогічних вимірювань в рамках класичної теорії обробки результатів тестування помічена в роботі [13]. Однак за традицією, що склалась на практиці та висвітлена в літературі [11 – 12, 21], при оцінювання ПЯ тестового простору за результатами тестування використовують як статистики, розраховані на роботу зі шкалами порядку, так і ті, що розраховані на роботу з метричними шкалами.

З метою усунення зазначеної невідповідності проведено дослідження, що дозволило запропонувати модифікацію формули (3). Оскільки первинні бали завдань тесту c_j та первинні бали опитуваних b_i відображаються ординальними шкалами, то в якості достатніх статистик цих величин запропоновано обирати непараметричні оцінки, які є повністю адекватними шкалі порядку [35 – 36].

Оскільки досліджувані вибірки опитуваних за умовою проблематики дисертації є невеликими (в межах 1 – 4 груп студентів – близько 20…100 осіб), то розподіл результатів тестування (первинних балів) в окремих сеансах тестування рідко є нормальним, на відміну від результатів масштабного тестування [11 – 13]. В такій ситуації найбільш обґрунтованими методами опрацювання експериментальних даних є стійкі непараметричні методи, що основані на медіані. Тобто в якості міри центральної тенденції ряду дослідних даних (в контексті, що розглядається – первинних балів) пропонується обрати медіану ряду, а в якості міри розсіювання одну з запропонованих в [37] непараметричних оцінок:

а) стійка медіанна оцінка середнього квадратичного відхилення:

$$S_1 = \mathrm{med}\left\{\left|x_i' - \mathrm{med}\left\{x_i'\right\}\right|\right\} / 0.675, \qquad (5)$$

де $\mathrm{med}\left\{x_i'\right\}$ – медіана проранжованого в порядку зростання ряду первинних балів опитуваних x_i', i – номер опитуваного.

б) стійка оцінка середнього квадратичного відхилення за вибірковими квартілями:

$$S_2 = \left(x'_{n-k} - x'_k\right) / 1.35, \qquad (6)$$

де n – об'єм вибірки опитуваних, $k = 0.25 \cdot n$,

x'_{n-k}, x'_k – відповідні члени проранжованого ряду первинних балів.

Приклад розрахунку названих статистик на малій вибірці дає можливість встановити придатність використання оцінок S_1 та S_2 для оцінювання надійності тесту за формулою Кьюдера-Річардсона (3). В табл. 2 наведено результати тестування 21 студента, яким було запропоновано по 7 тестових завдань.

Таблиця 2

Результати тестування

Опитуваний	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Проранжовані первинні бали студентів, x_i'	0	1	1	2	2	2	2	3	3	4	4	4	5	5	5	5	6	6	7	7	7
Медіана ряду, $\mathrm{med}\{x_i'\}$	4																				
Різниці $\left\|x_i - \mathrm{med}\{x_i'\}\right\|$	4	3	3	2	2	2	2	1	1	0	0	0	1	1	1	1	2	2	3	3	3
Проранжована вибірка $\left\|x_i - \mathrm{med}\{x_i'\}\right\|$	0	0	0	1	1	1	1	1	1	2	**2**	2	2	2	2	3	3	3	3	3	4
Медіана попереднього ряду	2																				

Для наведеного прикладу маємо:

1. Вибіркове середнє квадратичне відхилення (СКВ) – $S_x = 2.13$.

2. Стійка медіанна оцінка середнього квадратичного відхилення - $S_1 = 2.96$.

3. Стійка оцінка середнього квадратичного відхилення за вибірковими квартілями - $S_2 = 2.22$.

Як найбільш близьку до вибіркового СКВ оцінку міри розсіювання первинних балів обрано СКВ, визначене за вибірковими квартілями. Ця оцінка є стійкою, не залежить від виду розподілу досліджуваної величини та є коректною при роботі з ординальними величинами. Тобто, виходячи з аналізу властивостей досліджуваних величин – ПЯ КСТ – було обрано відповідні статистичні формули для оцінювання центральної тенденції та міри розсіювання досліджуваних емпіричних рядів величин, що за своєю природою є ординальними. На основі цього в роботі модифіковано або обрано нові метрики для оцінювання ПЯ КСТ.

Модифікована формула (3) для оцінювання показника надійності процедури тестування з використанням непараметричних оцінок матиме вигляд:

$$R_{KR} = k \cdot (k-1) \cdot S_{кв}^{-2} \left(S_{кв}^2 - \sum_{j=1}^{k} p_j (1 - p_j)\right), \qquad (7)$$

де $S_{кв}$ – СКВ результатів тестування (первинних балів опитуваних), отримане за вибірковими квартілями (формула (6)).

Коефіцієнт надійності, отриманий за модифікованою формулою (7), може набувати значень від 0 до 1. Значення, близькі до нуля, свідчать при низьку надійність – тобто про непридатність тесту до використання. Збільшення показника надійності свідчить про підвищення надійності процедури тестування.

Визначення надійності процедури тестування за способом Кьюдера-Річардсона вимагає незмінності умов тестування (тобто сталість контингенту опитуваних та набору завдань тесту). Тому цей спосіб можна застосовувати тільки при проведенні пілотного експерименту (при стандартизації КСТ та калібруванні тестового простору), але не можна використовувати при дослідній експлуатації КСТ.

В той же час як на етапі стандартизації, так і на етапі експлуатації системи можна оцінити надійність окремих завдань тесту за взаємною їх узгодженістю в тесті [11, 13].

Для оцінювання надійності завдань тесту спочатку знаходять попарно кореляції між двома завданнями тесту за перетвореним коефіцієнтом Пірсона – коефіцієнтом фі [12]:

$$r_{jl} = \left(p_{jl} - p_j \cdot p_l\right) \cdot \left(p_j \cdot (1-p_j) \cdot p_l \cdot (1-p_l)\right)^{-0.5}, \quad (8)$$

де p_{jl} – доля опитаних, які виконали правильно обидва завдання тесту; p_j, p_l - доля опитаних, які правильно виконали j-е та l-е завдання тесту.

Надійність завдань тесту [11] визначається як корінь квадратний з середнього арифметичного значення кореляцій кожного завдання з усіма іншими. З огляду на зазначені вище застереження щодо використання середнього арифметичного значення при роботі з ординальними величинами, замість нього пропонується оцінювати медіану ранжованого ряду оцінок коефіцієнтів кореляції, що отримані за формулою (8). Тоді надійність окремого завдання тесту оцінюватиметься за формулою:

$$r_j = \sqrt{\underset{l=1,2\ldots k}{\text{med}}\left\{r_{jl}\right\}} \quad \text{для кожного } j = 1,2,\ldots k. \quad (9)$$

Надійність процедури тестування можна оцінити за надійністю окремих завдань тесту з використанням модифікованої автором формули Спірмена-Брауна:

$$R_{SB} = k \cdot \underset{j=1,2\ldots k}{\text{med}}\left\{r_j\right\} \cdot \left[1 + (k-1) \cdot \underset{j=1,2\ldots k}{\text{med}}\left\{r_j\right\}\right]^{-1}, \quad (10)$$

де $\underset{j=1,2\ldots k}{\text{med}}\left\{r_j\right\}$ – медіана ранжованого ряду коефіцієнтів кореляції окремих завдань.

Границі значень показника надійності процедури тестування, визначеного з використанням модифікованої формули Спірмена-Брауна та його інтерпретація аналогічні показнику, отриманому за модифікованою формулою Кьюдера-Річардсона. Модифіковану формулу Спірмена-Брауна доцільно використовувати при синхронному тестуванні під час пілотного експерименту та за умови великого перекриття завдань при асинхронному тестуванні під час дослідної експлуатації КСТ.

Оцінювання невизначеності показника надійності

Отримані шляхом експерименту оцінки ПЯ, зокрема надійності процедури тестування та окремих завдань тесту мають супроводжуватись оцінками їх точності. Оскільки питання оцінювання якості педагогічних вимірювань не розв'язано, автором вперше запропоновані способи оцінювання невизначеності ПЯ відповідно до міжнародного підходу, викладеного в Guide to the expression of uncertainty in measurement [38].

Стандартну невизначеність надійності завдань тесту $u[r_j]$ запропоновано оцінювати як СКВ за вибірковими квартілями ряду значень r_{lj} (формула (8)) для кожного завдання $j = 1,2,\ldots k$. Винайдена формула для визначення розширеної невизначеності надійності завдань тесту з рівнем довіри $P = 0,95$ (за умови нормального розподілу показника):

$$U[r_j] = 2 \cdot u[r_j] = 2 \cdot \frac{\partial r_j}{\partial\left[\underset{l=1,2\ldots k}{\text{med}}\left\{r_{jl}\right\}\right]} \cdot S_{r\_j} = \quad (11)$$

$$= 2 \cdot \left(1 \Big/ \left(2 r_j\right)\right) \cdot S_{\phi i\_j} = r_j^{-1} \cdot S_{r\_j}, \qquad j = \overline{1,k},$$

де S_{r_j} – СКВ за вибірковими квартілями для рядів r_{jl}, $l = \overline{1,k}$.

Стандартну невизначеність показника надійності процедури тестування, отриманого за модифікованою формулою Спірмена-Брауна, автор пропонує оцінювати за виразом:

$$u[R_{SB}] = S_r \cdot \frac{\partial r_j}{\partial\left[\underset{j=1,2\ldots k}{\text{med}}\left\{r_j\right\}\right]} = \quad (12)$$

$$= S_r \cdot k \cdot \left(2 - \underset{j=1,2\ldots k}{\text{med}}\left\{r_j\right\}\right) \cdot \left[1 + (k-1) \cdot \underset{j=1,2\ldots k}{\text{med}}\left\{r_j\right\}\right]^{-2},$$

де S_r – СКВ за вибірковими квартілями для ряду r_j, $j = \overline{1,k}$.

Вираз для оцінювання розширеної невизначеність з рівнем довіри $P = 0,95$ (за умови нормального розподілу) показника надійності, отриманого за модифікованою формулою Спірмена-Брауна, має вигляд:

$$U[R_{SB}] = 2 \cdot u[R_{SB}] =$$

$$= 2 \cdot S_r \cdot k \cdot \left(2 - \underset{j=1,2\ldots k}{\text{med}}\left\{r_j\right\}\right) \cdot \left[1 + (k-1) \cdot \underset{j=1,2\ldots k}{\text{med}}\left\{r_j\right\}\right]^{-2}. \quad (13)$$

Для знаходження довірчого інтервалу показника надійності процедури тестування, визначеної за методом Кьюдера-Річардсона, вибіркове значення показника піддається перетворенню Фішера [39] :

$$Z_{R_{KR}} = 0.5 \ln\left[\left(1 + R_{KR}\right) \cdot \left(1 - R_{KR}\right)^{-1}\right]. \quad (14)$$

Тоді довірчій інтервал для показника $Z_{R_{KR}}$ становитиме:

$$Z = Z_{R_{KR}} \pm Z_{крит} \cdot \left(n-3\right)^{-0.5}, \quad (15)$$

де n – кількість опитаних; $Z_{крит}$ – критичне значення при заданому рівні довіри.

На практиці для показника надійності важливою є лише нижня границя довірчого інтервалу [11], яка дорівнює:

$$Z_L = Z_{R_{KR}} - Z_{крит} \cdot \left(n-3\right)^{-0.5}. \quad (16)$$

Нижня границя довірчого інтервалу показника надійності, отриманого за модифікованою формулою Кьюдера-Річардсона, матиме вигляд:

$$R_L = \left(e^{2Z_L} - 1\right) \cdot \left(e^{2Z_L} + 1\right)^{-1}. \quad (17)$$

Довірчий інтервал (розширена невизначеність) показника надійності може бути представлена як

$$U = R_{KR} - \left(e^{2Z_L} - 1\right) \cdot \left(e^{2Z_L} + 1\right)^{-1}. \quad (18)$$

Для спрощення формули (18) зроблено наступні перетворення. Введено заміну:

$$\Delta_Z = Z_{крит} \cdot (n-3)^{-0.5}, \text{ тоді}$$

$$U = R_{KR} - \left(e^{2Z_{KR}} \cdot e^{-2\Delta_Z} - 1\right) \cdot \left(e^{2Z_{KR}} \cdot e^{-2\Delta_Z} + 1\right)^{-1}.$$

За умови $\lim\limits_{n\to\infty} \Delta_Z \to \overset{0}{\theta}$, де $\overset{0}{\theta}$ – мала величина, при збільшенні вибірки опитуваних n, можна провести заміну виду $e^{-2\Delta_Z} \approx 1 - 2\Delta_Z$, тоді

$$U = R_{KR} - \frac{\left(e^{2Z_{KR}} - 1\right) \cdot \left(1 - 2\Delta_z \frac{e^{2Z_{KR}}}{e^{2Z_{KR}} - 1}\right)}{\left(e^{2Z_{KR}} + 1\right) \cdot \left(1 - 2\Delta_z \frac{e^{2Z_{KR}}}{e^{2Z_{KR}} + 1}\right)}.$$

З урахуванням того, що $2\Delta_z \cdot e^{2Z_{KR}} \cdot \left(e^{2Z_{KR}} - 1\right)^{-1} = \overset{0}{\theta}_1$ мала величина, для якої справедливо $\left(1 - \overset{0}{\theta}_1\right)^{-1} \approx 1 + \overset{0}{\theta}_1$, маємо

$$U \approx R_{KR} - R_{KR}\left(1 + 2\Delta_z \frac{e^{2Z_{KR}}}{e^{2Z_{KR}} + 1}\right) \cdot \left(1 - 2\Delta_z \frac{e^{2Z_{KR}}}{e^{2Z_{KR}} - 1}\right).$$

Після спрощення виразу отримуємо формулу для розрахунку розширеної невизначеності оцінювання показника надійності процедури тестування за модифікованим способом Кьюдера-Річардсона:

$$U[R_{KR}] = R_{KR} \cdot 4 \cdot \Delta_z \cdot e^{2Z_{KR}} \cdot \left(\Delta_z \cdot e^{2Z_{KR}} + 1\right) \cdot \left(e^{4Z_{KR}} - 1\right)^{-1}. \quad (19)$$

Відповідно відносне подання невизначеності показника надійності процедури тестування за модифікованим способом Кьюдера-Річардсона можна подати так:

$$U_{відн}[R_{KR}] = 4 \cdot \Delta_z \cdot e^{2Z_{KR}} \cdot \left(\Delta_z \cdot e^{2Z_{KR}} + 1\right) \cdot \left(e^{4Z_{KR}} - 1\right). \quad (20)$$

В табл. 3 наведено розрахункові значення розширеної невизначеності $U_{відн}[R_{KR}]$ (у відносних одиницях) з рівнем довіри $P = 0,95$ показника надійності в залежності від значень самого показника R_{KR} та кількості опитуваних n (в розрахунках враховано, що при $P = 0,95$ $Z_{крит} = 1,96$).

Таблиця 3

Значення розширеної невизначеності $U_{відн}[R_{KR}]$ (у відносних одиницях) з рівнем довіри $P = 0,95$

R_{KR} \ n	20	100	200	300	400	500
0,5	1,73	0,48	0,30	0,23	0,19	0,17
0,6	1,47	0,38	0,23	0,18	0,15	0,13
0,7	1,28	0,31	0,18	0,14	0,11	0,10
0,8	1,13	0,25	0,14	0,10	0,08	0,07
0,9	1,01	0,20	0,11	0,08	0,06	0,05

З наведених в таблиці розрахунків можна зробити наступні висновки, що при проведенні пілотного експерименту КСТ на малих вибірках опитуваних неможна оцінити надійність процедури тестування і, як наслідок, провести стандартизацію КСТ, оскільки оцінка показника супроводжується великою невизначеністю. Тому збільшення контингенту опитуваних, шляхом накопичення даних – результатів тестування, отриманих під час дослідної експлуатації КСТ дозволяє значно зменшити невизначеність оцінювання надійності процедури тестування. Крім того за заданим значення допустимої похибки оцінювання надійності процедури тестування, можна розрахувати мінімальний об'єм вибірки опитуваних, який би забезпечував задану похибку вимірювання показника. Тобто, проведення моніторингу КСТ впродовж експлуатації є необхідним для зменшення невизначеності оцінок показника надійності тестування.

Допустимі норми на показник надійності, як правило, визначаються емпірично на основі досвіду. З огляду на літературні джерела надійним можна вважати тест, який має коефіцієнт надійності: більше за 0.7 [30]; 0.6 – 0.8 [11]; 0.75 – 0.85 («кращі за надійністю тести» – 0.9) [21]; більше за 0.8 – [22]. Якщо надійність процедури тестування визначається за внутрішньою узгодженістю завдань між собою, тобто за формулами Кьюдера-Річардсона та Спірмена-Брауна, то норми на показних надійності менш жорсткі – надійність не повинна бути меншою за 0.3 [12].

Список літератури

1. *Федорук П.І. Адаптивна система дистанційного начання та контролю знань на базі інтелектуальних Інтернет-технологій: монографія / П.І. Федорук. – Івано-Франківськ: ПрНУ, 2008. – 326 с.*
2. *Колин К.К. Инновационное развитие в информационном обществе и качество образования / К.К. Колин // Открытое образование. – 2009. – №3. – С. 63-71.*
3. *Кондратьев И.Н. Основные составляющие образовательной среды дистанционного образовательного университетского комплекса / И.Н. Кондратьев // Открытое образование. – 2008. – №1. – С. 11-18.*
4. *Беспалько В.П. Программированное обучение (дидактические основы) / В.П. Беспалько. – М.: Высшая школа, 1979. – 300 с.*
5. *Вісмут Дж. Оцінювання для навчання: навч.посіб. / За ред. І.Є. Булах, М.Р. Мруга. – К.: Майстер-клас, 2007. – 170 с.*
6. *Ясулатіс В.А. Дистанційне навчання: метод. Рекомендації / В.А. Ясулатіс. – К.: МАУП, 2005. – 72 с.*
7. *Алексеев Н.А. К вопросу о повышении достоверности оценки при тестовом контроле знаний / Н.А. Алексеев, Н.И. Волков, Т.А. Майорова // Открытое образование. – 2004. – №3. – С. 27-32.*
8. *Волков Н.И. Тестовый контроль знаний: учебное пособие / Н.И. Волков, А.Н. Алексеев, Н.А. Алексеев. – Сумы: ИТД «Университетская книга», 2004. – 109 с.*
9. *Алексеев А.Н. Дистанционное обучение инженерным специальностям: монография / А.Н. Алексеев. – Сумы: ИТД «Университетская книга», 2005. – 333 с.*
10. *Паращенко Л.І. Тестові технології в навчальному закладі: методичний посібник / Л.І. Паращенко, Л.І. Леонський, В.Д. Леонська. – К.: ТОВ «Майстерня книги», 2006. – 217 с.*

11. Бурлачук Л.Ф. Словарь-справочник по психодиагностике: 2-е изд., перераб. и доп. / Л.Ф. Бурлачук, С.М. Морозов. – СПб.: Питер, 2004. – 250 с.

12. Челышкова М.Б. Теория и практика конструирования педагогических тестов: учебное пособие / М.Б. Челышкова. – М.: Логос, 2002. – 432 с.

13. Нейман Ю.М. Введение в теорию моделирования и параметризации педагогических тестов / Ю.М. Нейман, В.А. Хлебников. – М., 2000. – 168 с.

14. Разработка инновационных технологий в области контроля, оценки и коррекции знаний студентов / В.Ю. Преснецов, О.В. Пилипенко, Б.Г. Бобыльов, Л.Ю. Фроленкова // Открытое образование. – 2008. – №2. – С. 48-52.

15. Булах І.Є. Створюємо якісний тест: навч.посіб. / І.Є. Булах, М.Р. Мруга. – К.: Майстер-клас, 2006. – 160 с.

16. Педагогічне оцінювання та тестування. Правила, стандарти, відповідальність / Я.Я. Болюбаш, І.Є. Бухах, М.Р. Мруга, І.В. Філончук. – К.: Майстер-клас, 2007. – 272 с.

17. Шведова В.В. Моніторинг систем дистанційного навчання / В.Д. Ціделко, Н.А. Яремчук, В.В. Шведова // Вища освіта України. – 2006. – №2. – С. 54-61.

18. Шведова В.В. Оцінювання якості контролюючих тестів / В.Д. Ціделко, Н.А. Яремчук, В.В. Шведова // Техніка АПК. – 2006. - №4. – С. 34-36.

19. Аванесов В.С. Тесты в социологическом исследовании / В.С. Аванесов. – М.: Наука, 1982. – 199 с.

20. Аванесов В.С. Композиция тестовых заданий. Учебная книга. 3 изд., доп / В.С. Аванесов. – М.: Центр тестирования, 2002. – 240 с.

21. Гуревич К.М. Психологическая диагностика: учебное пособие / К.М. Гуревич, Е.М. Борисова. – Москва – Воронеж:«МОДЭК», 2001. – 435 с.

22. Бодолаев А.А. Общая психодиагностика / А.А. Бодолаев, В.В. Столин. – СПб.: Реч, 2003 – 440 с.

23. Алексеев А.Н. К вопросу о количественном оценивании результатов тестового контроля знаний / А.Н. Алексеев // Открытое образование. – 2006. – №4. – С. 45-51.

24. Басев Н.В. Педагогика и практическая психология / Н.В. Басев. – Ростов-на-Дону. Феникс, 2000. – 412 с.

25. Шведова В.В. Оцінювання надійності контролюючих тестів під час експлуатації систем дистанційного навчання / В.В. Шведова // Передовий науково-практичний досвід – 2009: Всеукраїнська НПК. Миколаїв, 17 вересня, 2009 р. – С. 154-155.

26. Анастази А. Психологическое тестирование / А. Анастази, С. Урбина. – СПб.: Питер, 2001. – 688 с.

27. Статистический подход к принятию решения по результатам тестирования для тестов закрытой формы / В.Б. Моисеев, В.И. Усманов, К.Р. Таранцева, Л.Г. Пяцублевый // Открытое образование. – 2001. – №4. – С. 37-42.

28. International vocabulary of Basic and general terms in metrology: ISO VIM (DGUIDE 99999): [Revision of the 1993 edition]. – Inte. Org. for Standardization, 2004 – 55 p.

29. Метрологія. Терміни та визначення: ДСТУ 2681-94. – К.: Держстандарт України, 1994. – 68 с. (Національний стандарт України).

30. Орнатський П.П. Вступ до методології науки про вимірювання / П.П. Орнатський. – К.: Віпол, 1994 – 180 с.

31. Орнатский П.П. Теоретические основы информационно – измерительной техники / П.П. Орнатський. – К.: Вища школа. – Головное изд., 1983. – 455 с.

32. Finkelstein L. Theory and Philosophy of Measurement / L. Finkelstein // HandBook of measurement Science. – Sydenham, 1982. – Vol. 1. – Edited By P. H.

33. Шведова В.В. Систематизація вимірюваних величин і складових невизначеності / В.Д. Ціделко, Н.А. Яремчук, В.В. Шведова // Метрологія та вимірювальна техніка. Метрологія – 2006: V НТК. – Х., 2006. – С. 48-50.

34. Гмурман В.Е. Теория вероятностей и математическая статистика: учебн. пособие для вузов – 10-е изд. стер. / В.Е. Гмурман. – М.: Высш. шк., 2004. – 479 с.

35. Шведова В.В. Застосування непараметричних оцінок при визначенні показників якості контролюючих тестів в системах дистанційного навчання / Н.А. Яремчук, В.В. Шведова // Гидротехнологиш, навигация, управление и конструирование авиационно-космической техники : VII Міжнародна НТК. – К., 23-24 квітня, 2009. – С.261-265.

36. Shvedova V.V. The Problems of Evaluation the quality coefficients of distant educational systems (DES) / V.V. Shvedova, N.A. Yaremchuk // IX int. young scientists' conf. on applied physics. Kyiv, 17-20 June, 2009. – P. 106.

37. Грановский В.А. Методы обработки экспериментальных данных при измерениях / В.А. Грановский, Т.Н. Сирая. – Л.: Энергоатомиздат, 1990. – 288 с.

38. JCGM 100: 2008. Evaluation of measurement data – Guide to the expression of uncertainty in measurement, 2008. [Електронний ресурс]. – Режим доступу : http://www.bipm.org/utils/common/documents/jcgm/ JCGM_100_2008_E.pdf.

39. Паніотто В.І. Статистичний аналіз соціологічних даних / В.І. Паніотто, В.С. Максименко, Н.М. Харченко. – К.: Вид. дім «КМ Академія», 2004. – 270 с.

Надійшла до редколегії 25.12.2014

Рецензент: д-р техн. наук, проф. І.В. Руженцев, Харківський національний університет радіоелектроніки, Харків.

ОЦЕНИВАНИЕ НАДЕЖНОСТИ ТЕСТОВОГО ПРОСТРАНСТВА КОМПЬЮТЕРИЗИРОВАННОЙ СИСТЕМЫ ТЕСТИРОВАНИЯ: МЕТОДЫ И СПОСОБЫ ОЦЕНИВАНИЯ ПОКАЗАТЕЛЯ И ЕГО НЕОПРЕДЕЛЕННОСТИ

В.В. Шведова

Автором систематизировано и проанализировано методы и способы оценивания одного из главных показателей качества компьютеризированной системы тестирования (КМТ) – показателя надежности; приведены традиционные метрики для оценивания разных видов надежности тестового пространства и проведено их усовершенствование с точки зрения особенностей шкалы, в которой оценивается показатель надежности тестового пространства КСТ. Приведена динамика изменения неопределенности показателя надежности при осуществлении мониторинга системы, что позволяет прогнозировать качественные характеристики КСТ на основании заданного объема тестового пространства.

Ключевые слова: *компьютеризированная система тестирования, тестовое пространство, надежность тестового пространства, неопределенность.*

ASSESSMENT OF RELIABILITY OF THE TEST AREA FOR COMPUTERIZED TESTING SYSTEM: METHODS AND TECHNIQUES OF EVALUATION OF ITS INDEX AND UNCERTAINTY

V.V. Shvedova

Author systematized and analyzed methods and techniques of evaluation of one of the main indicators of quality of computerized testing system (CTS) - reliability index; given the traditional metrics for evaluation of different types of reliability test space are studied and improvements in terms of features of the scale, which evaluates the performance reliability of the test space CTS. The dynamics of changes of the uncertainty of reliability during monitoring of the system allows to predict the qualitative characteristics of the CTS on the basis of a predetermined volume of the test space.

Keywords: *computerized testing system, the test space, reliability test space, uncertainty.*