УДК 37:004.891.3

О. Г. Колгатін

НАДІЙНІСТЬ РЕЗУЛЬТАТІВ АВТОМАТИЗОВАНОГО ПЕДАГОГІЧНОГО ТЕСТУВАННЯ

Запропоновано метод порівняння якості процедур інтерпретації тестових результатів у системі управління навчальним процесом. Проведені обчислювальні експерименти дозволили визначити умови ефективності інтерпретації результатів тестування на основі класичної процедури, моделі IRT, процедури обчислення тестового бала з урахуванням вагових коефіцієнтів складності завдань і процедури корекції вгадування.

***Ключові слова**: педагогічне тестування, педагогічні вимірювання, надійність результатів, процедура тестування, інтерпретація тестових результатів, модель IRT.*

Постановка проблеми. Невід'ємним складником системи управління навчальним процесом є механізм зворотного зв'язку, який спирається на педагогічні вимірювання. Тому надійність результатів вимірювання визначає якість управління і, врешті-решт, якість навчального процесу. Застосування тестових технологій для педагогічного вимірювання надало педагогам можливість кількісного аналізу точності результатів, що становить основу вдосконалення процедури тестування й інтерпретації результатів вимірювання. Неперервний розвиток тестових технологій, розроблення нових моделей тестування зумовлює потребу в розвитку методів визначення надійності тестових результатів.

Аналіз останніх досліджень. У класичній теорії тестування тест складається з фіксованої послідовності завдань, тестовий бал розраховується як кількість правильно виконаних завдань. Похибка тестового бала визначається через поняття надійності тестових результатів на основі визначення кореляції між результатами за паралельними варіантами тесту, послідовними тестуваннями, частинами тесту або на основі обчислення внутрішньої узгодженості (α-Кронбаха). Залежність похибки від індивідуальних параметрів тестованого не розглядається. Можливість застосування класичної теорії надійності тестових результатів для модифікованих процедур тестування та інтерпретації результатів вимірювання (наприклад, застосування вагових коефіцієнтів, спеціальних алгоритмів подання тестових завдань, врахування вгадування тощо) потребує доведення у кожному конкретному випадку. **В**идатним кроком у розвитку тестових технологій стала модель Г. Раша, яка ґрунтується на вдалій апроксимації залежності ймовірності правильної відповіді на завдання від підготовленості тестованого та параметрів завдання. Це надало можливість динамічно формувати педагогічний тест і забезпечити адекватне визначення підготовленості тестованого в умовах варіації трудності та кількості завдань у тесті. Певні зусилля дослідників були спрямовані на зниження негативного впливу вгадування тестованим правильних відповідей. Так, **В. В**. Кромер [1] запропонував заповняти матрицю результатів тестування значеннями 1, 0 та $\frac{-1}{k-1}$, де k – кількість варіантів відповіді на завдання. У [2] нами запропоновано окремо розглядати компоненти похибки вимірювання під час застосування автоматизованого тестування: вгадування, неуважність, пропуски у структурі навчальних досяг-

нень, недостатня еквівалентність паралельних варіантів тесту, які автоматично створює комп'ютер. Оскільки всі компоненти похибки є незалежними випадковими величинами, то її дисперсія дорівнює сумі дисперсії компонентів: $s_E^2 = s_{вгадування}^2 + s_{неуважності}^2 + s_{структури}^2 + s_{варіантів}^2$. Такий підхід надав можливість запропонувати формулу для дослідження залежності похибки вимірювання, яка пов'язана з вгадуванням, від підготовленості тестованого [3]. П. А. Ротаєнко виконав оцінку похибки вгадування на основі біноміального розподілу ймовірностей [4] що також надає можливість досліджувати залежність похибки вимірювання від підготовленості тестованого. Запропонована нами комбінаторна модель [5] є вільною від припущень про розподіл тестових балів, вона надала можливість проводити дослідження впливу вгадування в тестах, що складаються із завдань з різною ймовірністю надання випадково правильної відповіді. Дослідження у сфері вдосконалення процедур адаптивного й частково адаптивного тестування продовжуються, що потребує розвитку відповідних методів аналізу якості тестових результатів.

Виділення невирішених питань. Усі розглянуті вище підходи передбачають певну процедуру тестування та інтерпретації тестових результатів і не дають можливості порівнювати такі процедури за точністю ранжування тестованих у різних умовах. Видається актуальним завдання побудови метода, який дозволив би моделювати різні підходи до тестування й обчислення тестового бала з можливістю аналізу точності тестових результатів.

Мета даної роботи полягає в розробленні методики порівняння різних процедур тестування та інтерпретації тестових результатів у широкому діапазоні умов застосування тестів.

Виклад основного матеріалу. Будь-яке порівняння має спиратися на певний критерій якості. Але кожна процедура інтерпретації тестових результатів передбачає оригінальний критерій, і різноманітність критеріїв позбавляє дослідника можливості застосувати ці критерії для порівняння різних процедур. Більш того, шкали, за якими визначаються тестові бали, є різними в різних процедурах інтерпретації тестових результатів. Так, за класичною моделлю маємо лінійну шкалу відносно кількості правильно виконаних завдань; моделі з ваговими коефіцієнтами, що враховують трудність або складність завдань, передбачають певні нелінійні шкали; модель IRT, яка започаткована Г. Рашем, передбачає визначення підготовленості тестованого в логітах. Одним із напрямів вирішення проблеми може бути перетворення тестового бала за процентільною шкалою, яка відображає ранжування тестованих за результатами тестування. Але, на наш погляд, такий підхід пов'язаний з певними проблемами застосування статистичних методів для обчислення надійних інтервалів, оскільки зв'язок між різними шкалами є нелінійним. У такій ситуації пропонуємо здійснювати порівняння на підставі методу статистичних випробувань. Нехай для двох тестованих заздалегідь відомо, що один з них підготовлений краще (звісно, потрібна певна міра підготовленості, ця міра може бути визначеною за будь-якою однією шкалою). За результатами тестування після застосування певної процедури інтерпретації даних можливі три ситуації:

- процедура забезпечила правильне ранжування тестованих;
- процедура не виявила різниці в підготовці тестованих;
- процедура призвела до помилки в ранжуванні тестованих.

Оскільки тестові результати містять випадкові похибки, розглянуті вище ситуації виникатимуть випадково. На підставі великої кількості випробувань можливо побудувати розподіл імовірностей розглянутих ситуацій. Звісно, чим більше різниця у підготовленості випробуваних, тим вище ймовірність правильного їх ранжування за певною проце-

дурою інтерпретації тестових балів. Критерієм якості процедури інтерпретації тестових результатів (Q) оберемо різницю між імовірністю правильного та неправильного висновку щодо ранжування тестованих. Саме такий критерій, на наш погляд, є найбільш наочним і зручним. Він змінюється від нуля, коли процедура не забезпечує диференціацію тестованих, тобто кількість правильних і неправильних висновків однакова, до одиниці, коли всі висновки правильні.

Організувати лабораторні випробування з великою кількістю тестованих неможливо, тому потрібно розробити модель тестованого й процедур тестування та інтерпретації тестових результатів. Для цього потрібні певні припущення. Нехай імовірність правильного виконання завдання тесту добре апроксимується трьохпараметричною моделлю Г. Раша. Таке припущення природне, оскільки відомі емпіричні дані свідчать про адекватність моделі Г. Раша. Припустимо також, що параметри кожного завдання не залежать від особистості тестованого. Таке припущення можливо, якщо тест гомогенний, тобто всі завдання перевіряють один елемент навчального матеріалу, що виключає вплив прогалин у структурі навчальних досягнень тестованого. Щоб розширити сферу дослідження, додамо до моделі четвертий параметр, який характеризує неуважність тестованого, тобто ймовірність неправильної відповіді у випадку, коли тестований напевне здатний виконати завдання правильно. Вважатимемо, що параметр неуважності однаковий для всіх тестованих (це штучне обмеження нашої моделі). Таким чином, в основу моделі тестових результатів пропонуємо покласти ймовірність надання правильної відповіді як функцію від параметрів завдання та характеристики уваги випробуваних у вигляді:

$$p = c + \frac{(1-c)\cdot d}{1+\exp(-a\cdot(\theta-b))}, \qquad (1)$$

де θ – підготовленість тестованого, виражена в логітах; a – параметр завдання, що характеризує його роздільну здатність; b – показник трудності завдання; c – ймовірність випадкового надання правильної відповіді (вгадування); d – показник уваги тестованого – ймовірність правильної відповіді за умови, що тестований повністю здатний виконати завдання.

Розглянемо процедуру статистичних випробувань. Вхідними даними є підготовленість кожного з випробуваних θ_1, θ_2 ($\theta_2 > \theta_1$) та параметри кожного завдання тесту. Формуємо вектори відповідей першого й другого тестованого. Випадково, з імовірністю, що обчислюється за формулою (1) призначаємо відповіді правильними або неправильними. Обчислюємо тестовий бал за кожною з досліджуваних процедур інтерпретації тестових результатів. Порівнюємо тестові бали кожної з процедур. Якщо за тестовим балом випробуваний з підготовленістю θ_2 виявляється кращим, то висновок інтерпретації правильний, інакше різниця не виявлена або визначається помилка в ранжуванні. Багаторазово повторюємо випадкове формування вектора відповідей і застосування процедур інтерпретації тестових результатів. В обчислювальних експериментах кількість статистичних випробувань становила 100 000, що за наближеними оцінками з імовірністю не менше 95% забезпечувало дві правильні цифри у шуканому значенні критерію Q.

На рис. 1 і 2 подано результати дослідження якості інтерпретації тестових результатів за класичною процедурою, коли тестовий бал визначається як кількість правильно виконаних завдань. За віссю абсцис відкладено різницю між підготовленістю двох тестованих ($\theta_2 - \theta_1$) в логітах. Ряди даних відображають значення критерію Q як функції від ($\theta_2 - \theta_1$) для різної середньої підготовленості цих тестованих $\theta = (\theta_2 + \theta_1) / 2$.

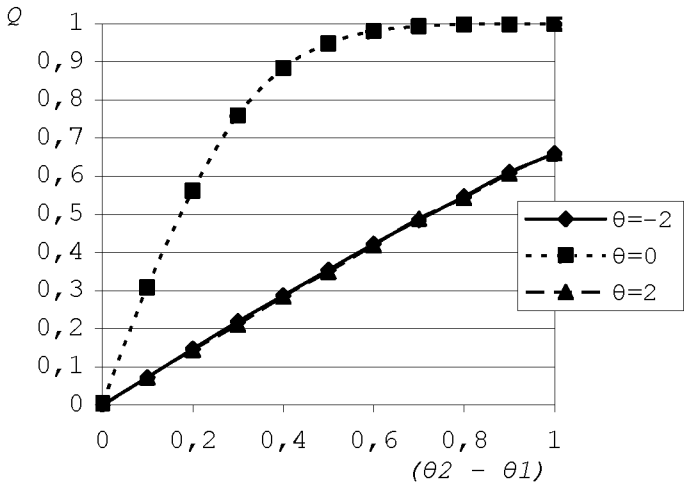

Рис. 1. Критерій Q якості ранжування тестованих за класичною процедурою обчислення тестового бала для тесту, який складається з 31 завдання однакової трудності з параметрами $b = 0$, $a = 2$, $c = 0$, $d = 1$

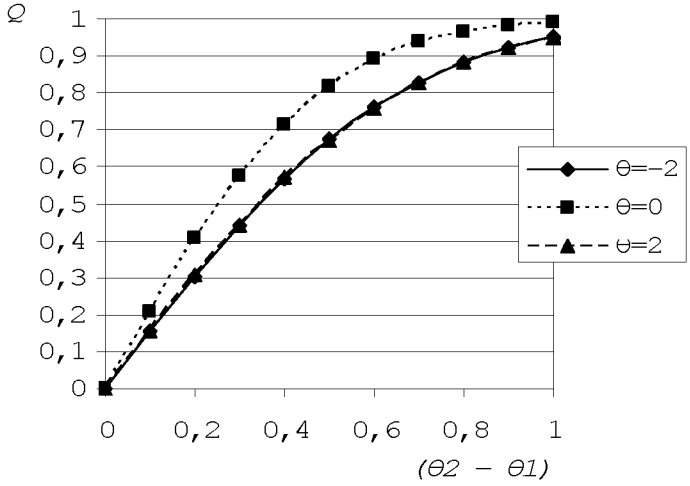

Рис. 2. Критерій Q якості ранжування тестованих за класичною процедурою обчислення тестового бала для тесту, який складається з 31 завдання зростаючої трудності (від $b = –2$ до $b = 2$), з параметрами $a = 2$, $c = 0$, $d = 1$

Результати обчислювальних експериментів збігаються з відомими висновками, що класична процедура інтерпретації тестових результатів забезпечує найкраще розділення тестованих, коли їх підготовленість близька до трудності завдань тесту (рис. 1, ряд даних $\theta = 0$). Але такий тест має вузький робочий діапазон вимірювання і для тестованих з низькою або високою підготовленістю не забезпечує задовільної якості вимірювання. Сучасні педагогічні тести будуються як система завдань зростаючої трудності, що дозволяє суттєво розширити робочий діапазон вимірювання (рис. 2), але чутливість тесту, тобто його здатність розділяти тестованих з невеликою різницею підготовленості, зменшується.

Інтерпретація тестових результатів за моделлю IRT не змінює ранжування тестованих порівняно з класичною процедурою інтерпретації тестових результатів. Це пі-

дтверджується теоретичним аналізом процедури визначення підготовленості тестованого за моделлю IRT і проведеними обчислювальними експериментами. В реальному тестуванні, коли параметри завдань невідомі й обчислюються за результатами тестування, звісно, спостерігатимуться розбіжності в ранжуванні, які викликатимуться похибками визначення параметрів тестових завдань за моделлю Г. Раша.

Проведемо зіставлення певних модернізованих процедур обчислення тестового бала з класичною процедурою. На рис. 3 ряд 1 подає значення критерію якості Q, які відповідають класичній процедурі обчислення тестового бала:

$$y_i = \sum_{j=1}^{m} X_{i,j}, \qquad \text{де } X_{i,j} = \begin{cases} 1 & \text{, правильна відповідь} \\ 0 & \text{, відмова від відповіді} \\ 0 & \text{, неправильна відповідь,} \end{cases}$$

де j – номер завдання; i – номер випробуваного; m – кількість завдань тесту.

Ряд 2 на рис. 3 подає значення критерію якості Q для процедури обчислення тестового бала з корекцією вгадування. Як вже було зазначено вище, вгадування тестованим правильних відповідей призводить до систематичного завищення тестового бала. В класичних гомогенних тестах з фіксованим набором завдань однакової форми й трудності це не призводить до систематичного впливу на ранжування тестованих, хоча випадкова похибка, звісно, збільшується. Але, якщо застосовується критеріально-орієнтована інтерпретація тестових результатів, систематичні похибки тестового бала мають бути скореговані. Для корекції систематичної похибки для випадку тесту з різними за формою завданнями нами на підставі підходу В. В. Кромера [1] була запропонована процедура обчислення тестового бала за формулою [5]:

$$y_i = \sum_{j=1}^{m} X_{i,j} / m, \quad \text{де } X_{i,j} = \begin{cases} 1 & \text{, правильна відповідь} \\ 0 & \text{, відмова від відповіді} \\ \dfrac{-c_j}{1-c_j} & \text{, неправильна відповідь} \end{cases} \qquad (2)$$

де j – номер завдання; i – номер випробуваного; m – кількість завдань тесту; c_j – ймовірність випадкового надання правильної відповіді для j-го завдання.

Ряд 3 на рис. 3 подає значення критерію якості Q для процедури обчислення тестового бала із застосуванням вагових коефіцієнтів, відповідних до трудності завдань – приклади такого підходу досить часто трапляються в літературі й автоматизованих системах тестування. Наприклад, вагові коефіцієнти застосовуються в тестах підсумкової державної атестації для завдань середнього і достатнього рівнів. Оберемо деяку "модельну" процедуру обчислення тестового бала із застосуванням вагових коефіцієнтів для більш трудних завдань:

$$y_i = \sum_{j=1}^{m} X_{i,j} w_j, \qquad \text{де } X_{i,j} = \begin{cases} 1 & \text{, правильна відповідь} \\ 0 & \text{, відмова від відповіді} \\ 0 & \text{, неправильна відповідь,} \end{cases}$$

де j – номер завдання; i – номер випробуваного; m – кількість завдань тесту; w_j – вага завдання, w_j =1 для простих завдань (b = –2 ... –0,8), w_j =2 для завдань середньої трудності (b = –0,7 ... 0,7), w_j =4 для трудних завдань (b = 0,8 ... 2).

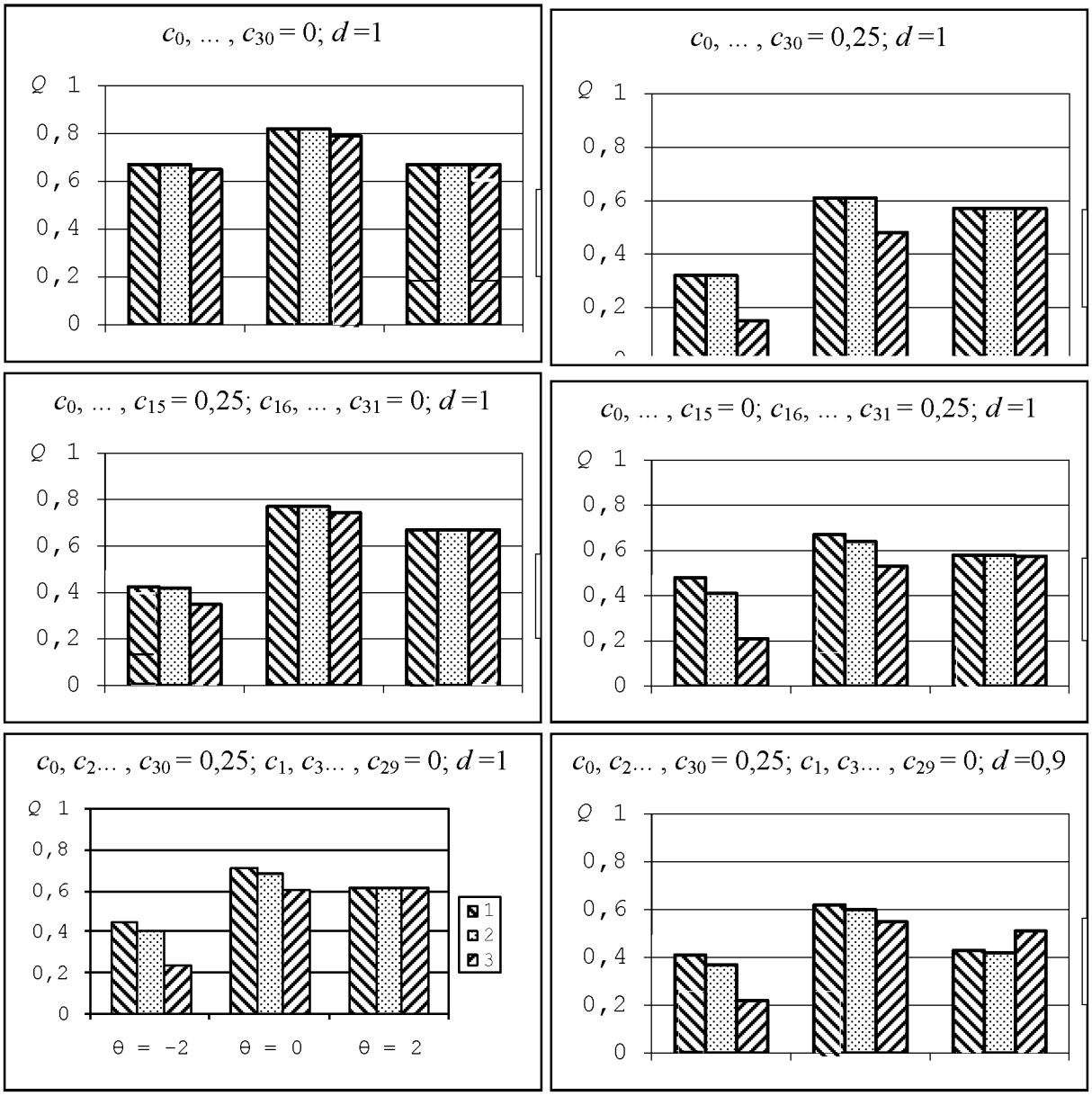

Рис. 3. Вплив вгадування та неуважності на якість інтерпретації тестових результатів: критерій Q обчислено для випадку ранжування тестованих з різницею підготовленості $(\theta_2 - \theta_1) = 0{,}5$ і середньою підготовленістю $\theta = (\theta_2 + \theta_1) / 2$ для тесту, який складається з 31 завдання зростаючої трудності (від $b_0 = -2$ до $b_{30} = 2$), з роздільною здатністю $a = 2$ за різними процедурами обчислення тестового бала (1 – класична; 2 – з поправкою на вгадування; 3 – з ваговими коефіцієнтами)

За результатами обчислювальних експериментів щодо впливу вгадування на якість інтерпретації тестових результатів (рис. 3) бачимо, що можливість надавати випадково правильні відповіді суттєво знижує якість інтерпретації за обраним нами критерієм Q. При цьому можливість вгадування правильних відповідей більш трудних завдань особливо негативно впливає на якість ранжування тестованих. Для тестованих з середньою та доброю підготовленістю ($\theta = 0$ і $\theta = 2$), незалежно від розглянутої процедури обчислення тестового бала, якість ранжування менше у випадку можливості вгадування правильних відповідей трудних завдань.

Порівняння якості ранжування тестованих за різними процедурами обчислення тестового бала дає можливість дійти висновку, що в більшості ситуацій, які розглянуто в

обчислювальних експериментах, якість модернізованих процедур виявилась не кращою, ніж якість класичної процедури. Тому в усіх випадках нормо-орієнтованої інтерпретації тестових результатів має застосовуватися класична процедура.

Процедура обчислення тестового бала з корекцією вгадування в усіх обчислювальних експериментах виявила якість, близьку до якості класичної процедури, крім винятків для практики випадків, коли підготовленість тестованого слабка ($\theta = -2$) та трудні завдання тесту припускають вгадування, а прості – ні. Тому можна рекомендувати корекцію вгадування за процедурою (2) в критеріально-орієнтованих тестах, які не призначено для оцінювання й визначення рейтингу, якщо такі тести містять різні за формою завдання. Прикладом доцільного застосування корекції вгадування може бути тестування в системах педагогічної діагностики, але потрібні додаткові дослідження конкретної системи тестових завдань щодо порівняння випадкової похибки, яка пов'язана з корекцією вгадування, і систематичної похибки, до якої призводить вгадування без застосування корекції.

Процедура із застосуванням вагових коефіцієнтів виявляється доцільною тільки в умовах можливої неуважності тестованих (невисока значущість результатів, занадто ускладнені формулювання завдань, обмеження часу, стомлення тощо) при тестуванні добре підготовлених учнів ($\theta = 2$) за допомогою тесту, що містить багато простих завдань.

Висновки:

1. Запропоновано метод порівняння на основі статистичних випробувань різних процедур інтерпретації тестових результатів.
2. Обчислювальний експеримент підтверджує відомий висновок, що найбільша якість ранжування тестованих забезпечується, якщо тест містить завдання однакової трудності, яка близька до підготовленості тестованих. Але такий тест має вузький діапазон вимірювання.
3. Можливість вгадування правильних відповідей, особливо для трудних завдань, негативно впливає на якість ранжування тестованих.
4. Для тестів з нормо-орієнтованою інтерпретацією результатів треба застосовувати класичну процедуру обчислення тестового бала (без корекції вгадування та вагових коефіцієнтів).
5. Інтерпретація тестових результатів за моделлю IRT не змінює ранжування тестованих порівняно з класичною процедурою інтерпретації тестових результатів.
6. Процедура обчислення тестового бала з корекцією вгадування може бути застосована в тестах з критеріально-орієнтованою інтерпретацією. Корекція вгадування дає можливість зменшити систематичну похибку тестового бала, але вносить додаткову випадкову похибку.
7. Застосування вагових коефіцієнтів для врахування рівня трудності завдань доцільно тільки за умови можливої неуважності тестованих, коли тест містить багато занадто легких для більшості тестованих завдань. В інших випадках застосування вагових коефіцієнтів призводить до суттєвого зниження якості ранжування тестованих.

Напрями подальших розвідок з **проблеми дослідження**: доцільно застосувати запропонований підхід до аналізу якості різних адаптивних і частково адаптивних процедур тестування та інтерпретації тестових результатів.

Список літератури: 1. *Кромер В. В.* О некоторых вопросах тестовых технологий / В. В. Кромер // Тезисы докл. Второй всеросс. конфер. ["Развитие системы тестирования в России"], (Москва, 23-24 ноября 2000 г.). Ч. 4. – М. : Прометей, 2000. – С. 59–61. 2. *Колгатін О. Г.* Автоматизована педагогічна діагностика і точність вимірювання / О. Г. Колгатін // Вісник. Тестування і моніторинг в освіті. – 2006. – № 10–11. – С. 29–33. 3. *Колгатін О. Г.* Статистичний аналіз тесту з різними за формою завданнями /

О. Г. Колгатін // Засоби навчальної та науково-дослідної роботи / За заг. ред. В. І. Євдокимова і О. М. Микитюка; ХДПУ ім. Г. С. Сковороди. – Харків : ХДПУ, 2003. – Вип. 20. – С. 50–54. 4. *Ротаєнко П. А.* Про вірогідність результатів тестування із закритою формою завдань / П. А. Ротаєнко // Комп'ютер у школі та сім'ї. – 2004. – № 6. – С. 12–15. 5. *Колгатін О.* Вплив вгадування на надійність тестових результатів у комп'ютерних системах педагогічної діагностики / Олександр Колгатін // Математика в школі. – 2008. - № 2 (78). – С. 36–41.

Bibliography (transliterated): 1. Kromer V. V. O nekotoryh voprosah testovyh tehnologij / V. V. Kromer // Tezisy dokl. Vtoroj vseross. konfer. ["Razvitie sistemy testirovanija v Rossii"], (Moskva, 23-24 nojabrja 2000 g.). Ch. 4. – M. : Prometej, 2000. – S. 59–61. 2. Kolgatin O. G. Avtomatizovana pedagogichna diagnostika i tochnist' vimirjuvannja / O. G. Kolgatin // Visnik. Testuvannja i monitoring v osviti. – 2006. – № 10–11. – S. 29–33. 3. Kolgatin O. G. Statistichnij anatiz testu z riznimi za formoju zavdannjami / O. G. Kolgatin // Zasobi navchal'noї ta naukovo-doslidnoї roboti / Za zag. red. V. I. Єvdokimova i O. M. Mikitjuka; HDPU im. G. S. Skovorodi. – Harkiv : HDPU, 2003. – Vip. 20. – S. 50–54. 4. Rotaєnko P. A. Pro virogidnist' rezul'tativ testuvannja iz zakritoju formoju zavdan' / P. A. Rotaєnko // Komp'juter u shkoli ta sim'ї. – 2004. – № 6. – S. 12–15. 5. Kolgatin O. Vpliv vgaduvannja na nadijnist' testovih rezul'tativ u komp'juternih sistemah pedagogichnoї diagnostiki / Oleksandr Kolgatin // Matematika v shkoli. – 2008. - № 2 (78). – S. 36–41.

УДК 37:004.891.3

А. Г. Колгатин

НАДЕЖНОСТЬ РЕЗУЛЬТАТОВ АВТОМАТИЗИРОВАННОГО ПЕДАГОГИЧЕСКОГО ТЕСТИРОВАНИЯ

Предложен метод сравнения качества процедур интерпретации тестовых результатов в системе управления учебным процессом. Проведенные вычислительные эксперименты позволили определить условия эффективности интерпретации результатов тестирования на основе классической процедуры, модели IRT, процедуры вычисления тестового бала с учетом весовых коэффициентов трудности заданий и процедуры коррекции угадывания.

***Ключевые слова**: педагогическое тестирование, педагогические измерения, надежность результатов, процедура тестирования, интерпретация тестовых результатов, модель IRT.*

UDC 37:004.891.3

O. Kolgatin

RELIABILITY OF THE RESULTS OF AUTOMATED PEDAGOGICAL TESTING

A method for comparison of quality of procedures of the test results interpretation in systems of learning process control is suggested. Computational experiments gave possibility to determine the conditions of efficiency such procedures of the test results interpretation as the classic procedure, IRT, the use of weight coefficients for the test tasks difficulty, the correction of guessing.

***Keywords:** pedagogical testing and measurement, reliability of results, testing procedures, interpretation of test results, the model of IRT.*

Стаття надійшла до редакційної колегії 3.10.2012