



Ala ȘIȘIANU

lector superior, Universitatea Tehnică a Moldovei

Sporirea calității testului docimologic prin analiza rezultatelor testării

Rezumat: Articolul conține o scurtă descriere a dezvoltării itemilor testului docimologic și o analiză a răspunsurilor la itemi, pentru îmbunătățirea acestora. Sperăm că lucrarea va contribui la o mai bună înțelegere a procesului de concepere a itemilor și a beneficiilor acestora, a importanței validării răspunsurilor la itemi. Dezvoltarea itemului și validarea răspunsului la item sunt procese noi, dinamice, care au nevoie

de o teorie unificată pentru scrierea/formularea itemilor și validarea cercetării. Fără îndoială, elaborarea itemilor continuă să fie un act extrem de creativ. Din moment ce utilizatorii de test conștientizează modul în care rezultatele testului ar trebui sau nu folosite, putem crește calitatea testelor prin dezvoltarea rațională și eficientă a itemului și a procedurilor de validare a răspunsului la item, descrise în articol.

Cuvinte-cheie: test, item, indiciu, dificultate, discriminare, ghicire, omitere.

Abstract: The present article provides a brief treatment of the development of docimological test items and the study of item responses for the improvement or continued use of these items. We hope that our work will help to understand better the process and benefits of well-designed test items and the importance of validating item responses. Item development and item response validation are still new, dynamic processes greatly in need of a unified theory for writing items and validating research. Without doubt, item writing continues to be a highly creative enterprise. Once test users are clear about how test results should and should not be used, we can increase the quality of tests by sensible and effective item development and item response validation procedures described in the article.

Keywords: test, item, difficulty, discrimination, guessing, omitted responses.

INTRODUCERE

Astăzi, ca niciodată, eficiența și eficacitatea formării specialiștilor de înaltă calificare depinde de nivelul cercetării pedagogice și de implementarea în procesul educațional a realizărilor științei pedagogice, care determină, în mare măsură, creșterea lor profesională.

Sistemul învățământului superior din republică mai e în căutarea propriului mecanism de evaluare. Cresc cerințele față de calitatea evaluării, față de noile metode și forme de evaluare și, nu în ultimul rând, față de optimizarea acestora. Odată cu trecerea la sistemul de control al calității, o actu-

alitate deosebită capătă elaborarea unor instrumente care să asigure eficiența în desfășurarea procesului de control continuu și cuprinzător al cunoștințelor. Unul dintre principalele mijloace de control devine sistemul științific bazat pe testarea pedagogică, sistematică, pe parcursul întregului an academic, care face posibilă stimularea studenților pentru activitatea cognitivă și creativă.

Testul docimologic se deosebește de formele obișnuite de testare, în primul rând prin includerea unui număr mai mare de studenți; prin rapiditatea, eficiența și eficacitatea muncii, atât a profesorului, cât și a formabilului. Prin

urmare, acesta trebuie să fie bine elaborat, corect din punct de vedere pedagogic și să corespundă realizărilor moderne ale teoriei și practicii docimologice. Având un potențial pozitiv evident, utilizarea lui, cu siguranță, va spori calitatea învățământului superior.

Prin *evaluare pedagogică* înțelegem sistemul de evaluare a rezultatelor învățării și educației studenților [1]. Scopul unei astfel de evaluări este identificarea nivelului de cunoaștere a programului de studiu și a nivelului de dezvoltare a calităților profesionale și morale ale viitorului specialist. Existența și dezvoltarea diferitelor tipuri de evaluare pedagogică se explică prin rolul stimulat al verificării rezultatelor învățării și a comportamentului studenților. Corectitudinea acestei idei este confirmată atât de practică, cât și de numeroase cercetări [5, p. 5]. Tentativele de a exclude total sau parțial evaluarea din procesul de învățământ au condus, de regulă, la rezultate negative și, mai cu seamă, la reducerea calității educației.

Sistemul evaluării pedagogice în școala superioară are trei funcții interdependente de bază: diagnostică, de învățare și educativă. Sistemul de verificare a rezultatelor învățării și educației include trei tipuri principale de control: curentă, intermediară și finală.

Odată cu trecerea la un nivel calitativ nou în monitorizarea și evaluarea cunoștințelor studenților, o mare importanță revine măsurării rezultatelor asimilării materialului de program.

Scopul cercetării consacrate testelor docimologice este de a repera căile ce ar permite sporirea validității și credibilității acestora. Or, nu putem vorbi despre calitatea unui test fără o amplă și minuțioasă analiză a itemilor acestuia sau, mai bine zis, a răspunsurilor la itemi. În literatura de specialitate sunt prezentate următoarele caracteristici importante ale răspunsurilor la itemi, care determină calitatea testului: indicele de dificultate, indicele de discriminare, indicele de "ghicire" sau pseudo-șansă și cel de omitere a răspunsurilor.

INDICELE DE DIFICULTATE

Una dintre caracteristicile răspunsului la item este *dificultatea*. Scala naturală pentru dificultatea itemului este procentajul subiecților care răspund corect. Plafo-nul oricărui test cu itemi de alegere multiplă (ITAM) este 100%, iar probabilitatea unui răspuns corect determină treapta când subiecții cu cunoștințe minime sau foarte modeste ghicesc răspunsul, adică îl selectează la întâmplare [4, p. 388]. În cazul ITAM (patru variante de răspuns), această treaptă e de 25%, iar în cazul ITAM (trei opțiuni de răspuns) – de 33%. Termenul tehnic obișnuit utilizat pentru marcarea gradului de dificultate a unui item este p , care reprezintă proporția sau procentajul subiecților ce răspund corect la item. Orice item are o anumită dificultate de bază, naturală, una întemeiată pe gradul de performanță al subiecților supuși testării/

evaluării. Valoarea p , așadar, depinde de omogenitatea eșantionului. Dacă eșantionul conține persoane bine pregătite sau cu un potențial înalt, itemii și, respectiv, testele vor apărea ca ușoare, cu valoarea p , de obicei, de peste 90. Dacă eșantionul întrunește persoane nepregătite sau având capacități reduse, itemii și, respectiv, testul va avea o valoare de dificultate sporită: de obicei, sub 40. Care ar fi metodele de control și reglare a valorii p , a indicelui de dificultate? Care sunt cauzele dificultății itemului? Cercetări în domeniul dat sunt destul de puține. Câteva experimente interesante au fost realizate prin anii 1986-87 de Smith, Green și Kramer [Apud 2, p. 208]. Tot de problema dificultății itemilor s-a ocupat și Bejar, care prin analize și probe a demonstrat că dificultatea poate fi redusă utilizând seturi mici, scurte de itemi. Cu cât itemul e mai mare ca volum, cu atât sarcina pusă în fața subiecților va fi mai greu de înțeles. O altă posibilă cauză a valorii p este gradul sau nivelul de pregătire și dezvoltare a eșantionului. Să admitem că valoarea p a unui item este 66. Ce a determinat-o – componența grupului testat sau nivelul de pregătire și instruire? Un indiciu ar fi examinarea performanțelor de testare ale subiecților supuși testării, ale celor instruiți și mai puțin instruiți, ale celor antrenați și mai puțin antrenați, ale celor pregătiți și ale celor nepregătiți. Acest lucru se referă deja la sensibilitatea instruirii și necesită un studiu aparte.

O altă posibilă cauză a valorii p este faptul că itemul nu este cu adevărat relevant pentru domeniul de cunoștințe testat. În această situație, ne-am aștepta ca modelul de performanță să fie neinteligibil sau valoarea p să fie foarte scăzută, deoarece itemul nu se referă la instruire sau pregătire.

INDICELE DE DISCRIMINARE

Indicele de dificultate nu este suficient pentru a putea judeca despre calitatea unui item. Este foarte important ca numărul răspunsurilor la un item să fie în relație cu scorul total al abilității pe care o măsurăm. Acest aspect este evaluat cu ajutorul indicelui de discriminare, care arată cât de bine discriminează itemul între subiecții cu scoruri mari și cei cu scoruri mici, adică puterea de discriminare a itemului. Astfel, un test cu ITAM de evaluare a performanței este un item bun, dacă: (1) majoritatea persoanelor cu scoruri mari la test au răspuns corect la itemul respectiv; (2) cei care au obținut scoruri mici au răspuns greșit la acel item. Dacă majoritatea celor care au obținut scoruri mari au răspuns greșit la un item, atunci este posibil ca subiecții să realizeze interpretări alternative ale răspunsurilor destinate a fi un distractor.

Dacă la un item subiecții care au scoruri ridicate răspund în sens invers scalei, iar subiecții care au scoruri joase oferă răspunsuri în sensul scalei, atunci itemul

respectiv fie este formulat greșit, fie măsoară altceva. Un item care posedă o capacitate de discriminare importantă are un procent mai mare de răspunsuri corecte pentru subiecții care au scoruri ridicate și un procent mai mic de răspunsuri corecte pentru cei care dețin un scor global mic.

Calculul indicelui de discriminare presupune următoarele etape:

- Se calculează scorul total.
- Se formează grupuri distincte de subiecți pe baza scorului total. Se recomandă selectarea unui procent cuprins între 25% și 33%.
- Se calculează indicele de discriminare, utilizând formula $D = (U/NU) - (L/NL)$, unde: U – numărul subiecților din grupul superior care au răspuns corect la item; NU – numărul subiecților din grupul superior; L – numărul subiecților din grupul inferior care au răspuns corect la item; NL – numărul subiecților din grupul inferior.

Un indice pozitiv demonstrează că itemul discrimi-

minează corect. Cu cât indicele de discriminare este mai mare, cu atât face o diferență mai clară între cei cu un potențial mare și cei cu un potențial mic. O valoare negativă a acestui indice este dovada existenței unei probleme în construcția itemului.

Discriminarea oricărui ITAM poate fi studiată din câteva perspective, printre care și cea tabelară și grafică. Aceste metode sunt intuitive și oferă perspective utile privind răspunsurile la itemi. Metoda fundamentală tabelară include media celor care aleg răspunsul corect și media celor care aleg un răspuns incorect. O modalitate bună de a înțelege discriminarea itemului este să reținem că cei care au ales răspunsul corect ar trebui să aibă un scor ridicat la test, iar cei care au ales un răspuns greșit – un scor scăzut. În *Tabelul 1*, primul item are o discriminare bună, al doilea – o discriminare mai redusă, al treilea item nu reușește să discrimineze, iar al patrulea – discriminează negativ. Un astfel de item poate fi introdus în categoria „greșit”.

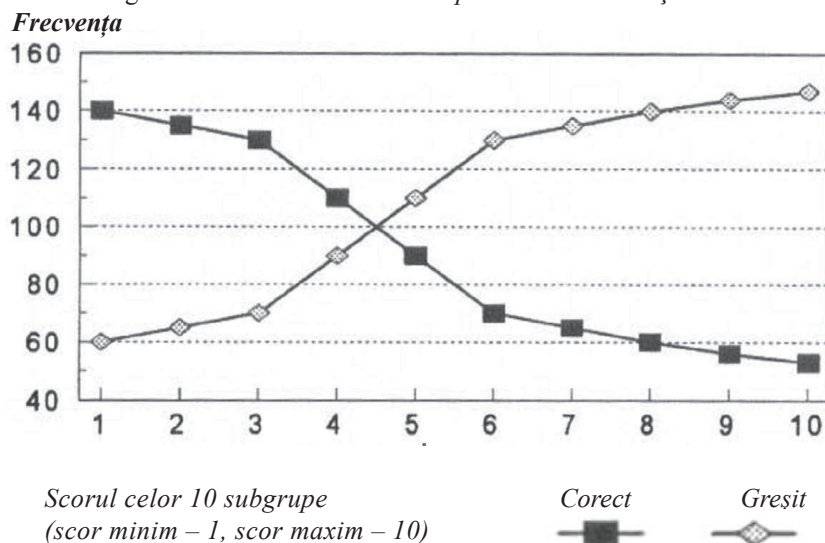
Tabelul 1. Exemple de scor mediu pentru patru tipuri de discriminare

Răspuns	Scor mediu Itemul 1	Scor mediu Itemul 2	Scor mediu Itemul 3	Scor mediu Itemul 4
Corect	90%	70%	65%	65%
Incorect	30%	60%	65%	75%
Discriminare	Foarte înaltă	Moderată	Zero	Negativă

Thomas Haladyna propune o a doua metodă tabelară [2, p. 210], când eșantionul ar fi divizat în subgrupe mai mici, unde grupul 1 îl reprezintă studenții cu cel mai mic scor de răspunsuri corecte, iar 10 – cei cu un scor maxim de răspunsuri corecte. *Tabelul 2* redă această scală de răspunsuri, reprezentată grafic în *Figura 1*:

Grafic este reprezentată linia de urmărire, cunoscută drept *curbă opțională*, pentru alegerea corectă și răspunsurile incorecte. Linia de urmărire poate fi construită în mai multe moduri. Una dintre metodele cele mai simple și cele mai directe o are la dispoziție orice program grafic pe calculator. În *Figura 1*, liniile de urmărire au fost construite în baza rezultatelor din *Tabelul 2*. Linia de urmărire pentru răspunsul corect este în creștere monotonă, în timp ce linia de urmărire pentru răspunsurile colective incorecte este în scădere monotonă. Astfel, linia de urmărire pentru răspunsuri corecte este o imagine-oglină a liniei de urmărire pentru răspunsuri incorecte. Liniile de urmărire oferă posibilitatea ilustrării tendinței discriminatorii a itemilor.

Figura 1. Linia de urmărire a răspunsurilor corecte și incorecte



Tabelul 2. *Scala de răspunsuri*

Studentul	1	2	3	4	5	6	7	8	9	10
Răspuns greșit	140	138	130	109	90	70	63	60	58	56
Răspuns corect	60	62	70	91	110	130	137	140	142	144

Liniile plane (drepte) și cele non-monotone sunt nedorite: liniile plane indică o lipsă de discriminare, iar cele non-monotone sunt greu de explicat din punctul de vedere al comportamentului studenților.

FENOMENUL „GHICIRII”

Testul cu ITAM conține un element de „ghicire”. Orice subiect supus testării se confruntă cu necesitatea de a da un răspuns, fie că știe răspunsul corect, fie că deține cunoștințe parțiale ce-i permit să elimine distractorii indezirabili și să selecteze aleator din răspunsurile rămase sau, pur și simplu, ghicește, în lipsa cunoștințelor elementare.

În teoria testării clasice, se poate ignora influența „ghicitului”, denumit în literatura de specialitate ca „fenomenul hazardului”. Pentru aceasta, ar trebui să se ia în considerare legile probabilității, care determină măsura în care „ghicirea” ar putea influența rezultatele. Probabilitatea de a obține un scor mai mare decât cel meritat prin „ghicire” devine mai mică în cazul când testul este mai mare (conține un număr mai mare de itemi). De exemplu, chiar și într-un test cu patru opțiuni, pentru cei 10 itemi propuși, probabilitatea de a da 10 răspunsuri corecte „ghicite” este de 0.000009. Un alt mod de a privi acest lucru este de a realiza că probabilitatea de a acumula un punctaj de 70% sau mai mult pentru un test cu 10 ITAM prin ghicire aleatorie este mai mică de 0,004. Dacă lungim testul până la 25 de itemi, probabilitatea de a obține un scor mai mare de 70% scade la mai puțin de 0,001.

Deși este adesea numit „indice de ghicire”, modelul de răspuns al itemului cu trei indicii este, de fapt, un nivel de pseudo-șansă. Acest indice nu este destinat să modeleze procesul psihologic de „ghicire”, ci doar să stabilească faptul că există un prag rezonabil pentru indicele de dificultate. Cel de-al treilea parametru este utilizat împreună cu indicele de dificultate și cel de discriminare pentru a calcula scorul subiectului participant la testare. Influența acestui parametru este mică, în raport cu influența parametrului de discriminare.

RĂSPUNSURI OMISE SAU NESOLUȚIONATE

Thomas Haladyna admite că „în funcție de instrucțiunile testului și de alte condiții, subiecții examinați pot omite răspunsurile”, subliniind că: „...o formă gravă a răspunsurilor omise sunt itemii la care nici nu s-a încercat de a da vreun răspuns” [2, p. 217]. De obicei, un șir de răspunsuri la sfârșitul unui set mare de itemi semnaleză existența altor itemi, care nu au fost soluționați. Studiind răspunsurile, este

important de catalogat omiterile ca nerezolvare a itemilor, deoarece amploarea unui astfel de fenomen ca lipsa răspunsului reprezintă o amenințare la adresa interpretărilor sau utilizării valorii testului valid. În unele programe de testare, subiecții sunt descurajați să ghicească și, astfel, au tendința de a omite mai mulți itemi. În plus, având baremul de notare la dispoziție, unii calculează care este scorul minim pentru o notă de trecere și selectează intenționat itemii mai „ușori” sau „favoriți” și îi omit pe ceilalți. Aici ne confruntăm cu fenomenul eschivării de la itemii nedorți.

EVALUAREA DISTRACTORULUI

Utilizatorii de test și analiștii ar trebui să considere distractorul o parte semnificativă a itemului. Într-adevăr, cei aproape 50 de ani de cercetare continuă au scos la iveală o relație-model între alegerea distractorului și scorul total obținut la test. Iată câteva motive pentru studierea performanței distractorului în ITAM:

- Reducerea itemilor „voluminoși”. O soluție ar fi diminuarea „volumului” unui item în favoarea măririi numărului de itemi. Mai bine mai mulți itemi scurți decât câțiva voluminoși.
- Îmbunătățirea itemilor. Obiectivul principal al studierii și evaluării răspunsurilor este îmbunătățirea itemilor. Aceasta înseamnă diminuarea dificultății și îmbunătățirea discriminării.
- Detectarea motivelor problemelor de performanță. Studiul distractorilor poate conduce la concluzii despre motivul pentru care itemul nu funcționează așa cum se presupunea [3]. Întrucât distractorii sunt adesea inspirați din cele mai frecvente erori ale studenților, performanța acestor distractori poate ajuta să se decidă care dintre ei rămân și care trebuie revizuiți. O astfel de analiză are adesea implicații pentru viitoarea instruire sau formare.

În concluzie: Analiza testelor cu ITAM și a indicilor care îi caracterizează ne oferă posibilitatea de a depista problemele cu care se confruntă cadrele didactice și studenții în timpul evaluării. Un studiu aprofundat în domeniu ne va permite să găsim răspunsul la problemele date, soluții pertinente, iar prin aceasta să ne aducem contribuția la dezvoltarea învățământului ingineresc prin sporirea calității și a obiectivității evaluării, dar, mai ales, a calității testelor docimologice.

Luarea unor decizii radicale cu privire la caracteristicile itemilor testului docimologic va influența pozitiv

calitatea acestuia, a procesului de evaluare și a învățământului superior, în general. Anume acesta este scopul nostru comun în pregătirea specialiștilor de înaltă calificare, care își vor pune umărul la dezvoltarea economiei țării, la edificarea tinerei generații.

REFERINȚE BIBLIOGRAFICE:

1. Bhattacharya S. Literature Review on a Comparative Study of the Evaluative Procedures, Examination System and Assessment of the Three Boards – ICSE, CBSE and WBBSE. In: Sundaram: Journal for All Subjects. Vol. 1 (1), pp. 1-22. Pe: www.e-sundaramjournal.in
2. Haladyna T. M. Developing and Validating Multiple-Choice Test Items. 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates, 2004. 305 p.
3. Mukerjee D. R. Testing Reading Comprehension: A Comparative Analysis of a Cloze Test and a Multiple-Choice Test. In: Indian Educational Review, 1991.
4. Wang W.-C. Assessment of Differential Item Functioning. In: Journal for Applied Measurement. Nr. 9 (4), 2008. Hong Kong Institute of Education. pp. 387-408.
5. Асямов С. В. Педагогическое тестирование в системе оценки слушателей. Ташкент: Академия Министерства Внутренних Дел Республики Узбекистан, 2006. 43 p.