Alkalmazott Matematikai Lapok **38** (2021), 39–58.
DOI: 10.37070/AML.2021.38.1.04

ADAPTÍV TESZTEK MINIMÁLIS HOSSZÁNAK, HIBÁJÁNAK, ÉRTÉKELÉSI SZINTJÉNEK ÉS A MEGOLDÓK SZÁMÁNAK ÖSSZEFÜGGÉSEI – ÁLTALÁNOS MEGOLDÁSI ARÁNNYAL

T. KÁRÁSZ JUDIT, TAKÁCS SZABOLCS

A cikkünk alapvetően egy általános formalizmus megalkotását tűzte ki célul, melynek segítségével a papír-ceruza alapú tesztek digitalizációjának egyik problémáját kezelhetjük. A kidolgozott formalizmus általános módon lett megalkotva, azonban egy gyakorlati példán is ismertetjük az eredményeket. Dolgozatunkban bemutatjuk, hogy egy alapvetően papír-ceruza teszten alapuló felmérés esetében, mint amilyen az OECD PISA [8] és az Országos kompetenciamérés [9], annak leendő adaptív változata milyen korlátokkal és előnyökkel szolgálna. A nemzetközi mérések esetében már elkezdődött az a fejlesztési munka, mely ezt az irányt készíti elő, illetve az elmúlt időszak online oktatási tapasztalatai megágyaztak annak, hogy az Országos kompetenciamérés esetében is lehetőség legyen digitális, majd adaptív mérésre való áttérésre. Az ehhez szükséges itemszintű logisztika (feladatbank, modellezési környezet) alapvetően rendelkezésre áll. A szakirodalom azonban általánosságban ezekre a technikai részletekre tér ki – valamint arra, hogy a megbízhatósághoz hány főre van szükség. Cikkünkben mi nem erre a kérdésre voltunk kíváncsiak, ugyanis az itembank kellő méretű, illetve jellemzően egy országos mérésnél a kellő kitöltöttség eredendően adott lesz. Ezzel szemben számunkra az a fontosabb kérdés, hogy milyen mértékben rövidíthető le a teszt úgy, hogy a mostanival megegyező – vagy annál jobb, megbízhatóbb – teszthez jussunk. Minket tehát most nem a kitöltők száma, hanem a teszt hossza érdekelt, erre vonatkozóan végeztünk el úgy szimulációs számításokat, hogy a mostani [8, 9] mérések általánosan kiadott mutatóit vettük alapul.

1. Bevezető

Az adaptív tesztelés nagyobb számítógépes kapacitása napjainkra már nem jelenthet érdemi kifogást a papír-ceruza tesztekkel szemben. Lényegében a legegyszerűbb (Google, Microsoft), leginkább elterjedt platformok is képesek arra, hogy online teszteket vegyünk fel a segítségükkel. Ennél fejlettebb az, ha a tesztünk nemcsak számítógépes, hanem adaptív is. Az adaptív tesztelés [6] során az egymás utáni itemeket a tesztelő rendszer úgy adja a tesztalanynak, hogy az mindig

az aktuálisan számított tudásszintnek megfelelő legyen. Tehát egy rosszul teljesítő alany nem kap megoldhatatlanul nehéz feladatokat, míg az igen jól teljesítő sem kapja a számára unalmasabbnak látszó, könnyű feladatokat. De természetesen az adaptív tesztelés nemcsak teljesítményre, hanem egyéb tesztelésekre is általánosítható. Gondolhatunk itt arra is, hogy ha egy betegség tüneteit nézzük, akkor egy lázas betegnél az alábbi kérdések nem praktikusak: Több-e a testhőmérséklete, mint 36,5 fok?, 37 fok? 37,5 fok? Az egész skálán nem fogunk végigmenni – egészen pontosan nem fogjuk a lázas betegnél az összes alacsony testhőmérsékletre vonatkozó kérdést feltenni. Azaz: ha teszt szinten gondolkodnánk, akkor mondjuk 1/10 fokonként újra és újra kérdeznénk. Ezzel szemben a nincsen lázam, hőemelkedés, láz kérdések, vagy a 38 foknál való kezdés (alatta vagy felette) alkalmas kérdések arra vonatkozóan, hogy viszonylag gyorsan eljussunk a megfelelően kalibrált testhőmérsékletig (vagy természetesen használhatunk hőmérőt is). A valódi kérdés azonban az, hogy ha például egy teljesítmény tesztet akarunk összeállítani, akkor e kérdezési metódussal vajon mennyivel hamarabb tudunk célba jutni [3]? Természetesen ez sok paramétertől függhet (mi a pontosság, milyen típusú kérdésekkel dolgozunk, stb). Ezért rögzítünk néhány alap feltételt.

1) Kizárólag két értékű itemekkel dolgozunk (megoldja/nem oldja meg, illetve például van-e tünet/nincsen tünet) – a súlyosságot nem mérjük.

2) A teszt pontosságát rögzítettnek tekintjük. Ezen azt értjük, hogy ha van egy papír-ceruza alapú tesztünk (vagy egy nem adaptív tesztelési metódusunk/forgatókönyvünk), akkor azt a becslési pontosságot szeretnénk elérni, amit az adott teszt tud.

Megjegyezzük, hogy itt felmerülhet az alábbi kérdés: miért igaz az, hogy kevesebb kérdéssel, gyorsabb méréssel hasonló minőségű és mennyiségű információhoz juthatunk? A válasz kétoldali. Egyik részről nem kérdezünk sok felesleget. A megkérdezett jellemzően a tudásszintjének, teljesítményének megfelelő kérdéseket kap. Másik oldalról a rövidebb idő azt is eredményezheti nagyobb volumenek esetében, hogy azon válaszadók, akik elunják a tesztet és félbehagyják, a rövidebb teszteket mégis kitöltik.

Azaz a kérdésünk: ugyanolyan típusú kérdésekkel, mint az eredeti papír-ceruza teszt, ugyanolyan pontosságot mennyivel gyorsabban tudunk reprodukálni? Azzal a feltételezéssel fogunk még élni, hogy a teszt itemjeinek paramétereit ismerjük, tehát olyan itemekkel dolgozhatunk, melyek nehézsége (feladatmegoldási paraméterei) ismertek a tesztet összeállítók előtt. Ez nem túlzó feltételezés, mert gondolhatunk itt egy betegség esetében arra, hogy egy-egy tünet mennyire súlyos (általában ezzel az orvosok tisztában vannak), illetve egy-egy teljesítményt mérő item esetében is a szakemberek rendelkeznek ezzel az információval a tesztek összeállításakor [2]. Hogy ez mennyire nem túlzó feltételezés, Harrisonék munkájában még az is tisztázott, hogy egy-egy feladat gyermekek és felnőttek számára mennyire nehéz feladat - tehát a szakértők akár korosztályonként is meglehetősen jó becsléseket tudnak mondani a feladatok nehézségét illetően. Ezek után az a

kérdés, hogy ugyanolyan típusú kérdésekkel, ugyanazon jelenséget vizsgálva, hány item segítségével tudunk hasonló becslési pontosságot felmutatni – egy-egy egyénre vonatkoztatva? Megjegyezzük, hogy az eddig hivatkozott művek mindegyike esetében kellően nagy mintákon szimulált vizsgálatokat végeztek annak érdekében, hogy az adott becslési hatékonyságot vizsgálni tudják. Azaz nem a teljes mintafelvétel hibáját szeretnénk csökkenteni, hanem az adott teszten elérhető pontszám hibáját szeretnénk elérni kevesebb item felhasználásával. A kutatókat általánosságban érdeklő és foglalkoztató kérdés az adaptív tesztelés során inkább az, hogy mekkora itembankra van szükség ahhoz, hogy egy adaptív rendszert működtetni lehessen. Gondoljunk ugyanis arra, hogy az IQ-t szeretnénk 100 kérdés helyett csak 10 kérdésből mérni. Ha a kérdések nyilvánosságra kerülnek, használhatatlanná válnak (lásd például [1]), és akkor nagyon gyorsan maguk a tesztek is használhatatlanná válnak. Ezt azzal lehet kivédeni, hogy nagy méretű itembankot hozunk létre, amiből a megkérdezett véletlenszerűen kap kérdéseket – az aktuális szintjének megfelelően.

2. Kuder–Richardson formula

Bár maga a becslési formula meglehetősen régóta ismert [4], a megközelítés általában inkább szimulációk sorát jelentette. Azaz a felmérés során egyfajta mi lenne, ha „lehetőségeket" vizsgálnak, hogy adott mintanagyság esetében egy-egy teszthossz vagy egyéb feltételrendszer esetében hány kérdésre lenne szükség az adott becslési szint teljesítéséhez. A tesztek során az alap definíciónak a teszt megbízhatóságát fogjuk alapul venni. A megbízhatóság a Kuder–Richardson formula szerint az alábbi alakban határozható meg [4, (20) képlet]:

$$r = \frac{L}{L-1}\left(1 - \frac{\sum_{i=1}^{L} p_i q_i}{s^2}\right),$$

ahol L az itemek száma, $\sum_i p_i q_i$ az itemek össz-varianciája, p_i az összes jó válasz aránya (jó válasz / összes esetszám), míg q_i a rossz válaszok aránya. Továbbá s^2 a teljesítmény összesített varianciája. Ez az s^2 alapesetben $N(0,1)$ változók összegzését jelenti (ha a mintaalanyok egymástól függetlenül írják a tesztet/mennek el orvoshoz tünetekkel), tehát s^2 legalábbis nagyságrendileg a minta nagyságával összemérhető.

A teszt a minta növelésével tehát egyre megbízhatóbbá válik, hiszen a hányados határértékben 0-hoz konvergál (feltéve, hogy valóban standard normális határeloszlást tudunk az IRT modellek – a nemzetközi és hazai mérések egyaránt ezzel dolgoznak, lásd bővebben [8, 9] – segítségével meghatározni minden résztvevő pontszámaként). A formula definíciói alapján a teszt standard hibáját (SEM) az

alábbi formula szerint definiálják.

$$SEM = \sqrt{1 - r}.$$

Jól látható tehát az alábbi: Ha a teszt itemjei konzisztensek (azaz megoldhatóságukat tekintve kiegyensúlyozottak), azaz a teszt teljes variabilitásához képest az itemek variabilitása összességében alacsony, akkor a $(\sum pq)/s^2$ kifejezés értéke alacsony lesz. Amennyiben az itemek összvarianciája a teljes teszt/teljesítmények varianciájához képest alacsony marad, úgy az $1-(\sum pq)/s^2$ kifejezés egyre jobban közelít 1-hez. Ebből következően relatíve hosszú tesztek esetében a teljes teszt megbízhatósága 1-hez közelít.

Minél közelebb van a teszt megbízhatósága 1-hez, várhatóan annál kisebb lesz az $s\sqrt{1-r}$ kifejezés értéke, tehát annál kisebb lesz a teszt standard hibája.

Ezen a ponton szokás a szimulációkat végrehajtani. Ugyanis az egyes itemek megoldottsága természetesen nem rögzített (vannak a tesztekben könnyebb és nehezebb feladatok – egészségügyi tesztek esetében vannak súlyosabb és kevésbé súlyos tünetek). A teljes tesztre ránézve azt tudjuk szimulálni, hogy egyik-másik itemet elhagyva (vagy bevéve) az eljárásunkba, mennyivel tudunk gyorsabban célba érni – tehát relatíve kevesebb itemet felhasználva hasonló eredményre jutni.

Az adaptív tesztelés esetében ez nem egészen így történik [5], miként arra vonatkozóan Linden és Glas az elsők között tett említést: ők több itembank nagysággal és teszthosszal, mintanagysággal végeztek kísérletet – tehát nem egészen azt az utat járták be, mint amit mi fogunk választani.

Az egyik lehetséges cél tehát az, hogy rögzítve a teszt megbízhatóságát, rögzítve a teljes minta hibájának mértékét, azt szeretnénk megtudni, hogy milyen itemszámra van szükségünk, ha adaptív módon szeretnénk kérdezni. Magyarán: az esetszám, tehát a mintanagyság, melyet a becsléshez felhasználunk, változatlan marad.

Tekintettel arra, hogy e formula nem tartalmazza azon szinteket, melyekkel a nemzetközi [8], illetve például egy teljes népességet vizsgáló magyar mérés is dolgozik (Országos kompetenciamérés, [9]), így olyan formulával dolgoztunk helyette, mely ezt a sajátosságot is figyelembe veszi. A szintek esetében az alábbi példákra gondolhatunk:

1) Betegség esetében egy adott betegség súlyosságának fokozatai.

2) Iskolai teljesítmény esetében például nem egy teszt pontszámait értjük alatta, hanem az arra kapott osztályzatot.

3. Wright formulája – általános eset $(0 < p < 1)$

Általában a $p = 1/2$ esetet elemzik – magyarán, hogy egy teszt esetében ugyanolyan valószínűséggel oldja vagy nem oldja meg a diák a feladatot. Ettől mi eltérünk és azt mondjuk, hogy általánosságban könnyebb $\left(p > \frac{1}{2}\right)$, illetve nehezebb $\left(p < \frac{1}{2}\right)$ tesztek is górcső alá vehetők. Felhívjuk a figyelmet, hogy a formula p és q esetére szimmetrikus, mi most a $p \leq 0,5$ eseteket fogjuk formalizálni. Világos, hogy a $p = 0$ és a $p = 1$ esetek nem érdekesek.

Az előzőekben megismert jelölések tehát az alábbi alakban írhatók át. Tegyük fel, hogy b_i jelöli az adott személy képességét ($i = 1, \ldots, N$ kitöltővel számolva) és d_j jelöli az adott itemek nehézségét (továbbra is $j = 1, \ldots, L$ itemet használunk a papír-ceruza referencia tesztben).

Jelölje s_j azt a számot, ahányan az adott itemet jól megoldották (valamint n_w jelölje azt, hogy pontosan w darab feladatot hányan oldottak meg). Ha adaptív tesztelésben gondolkodunk, akkor a korábbi jelölésekkel $s_j = Np$ (illetve $s_j = \frac{N}{2}$ a teljesen klasszikus felállás esetében). Életszerű továbbá az a megközelítés, hogy azokat az alanyokat, akik minden itemet megoldanak vagy elrontanak, kihagyjuk a további elemzésekből (a kitöltők számát továbbra is N-nel jelölve). Hasonlóan, ha egy itemet mindenki/senki sem oldott meg, szintén elhagyhatjuk (jelölje a továbbiakban is a teszt hosszát L). Tehát N és L a valid teszthossz és kitöltők számát fogja jelölni.

Fontos megjegyezni, hogy a senki sem töltötte ki, illetve mindenki kitöltötte változatok elvi – intuitív – szinten nem szimmetrikus esetek (a legjobb és legrosszabbak elhagyása), azonban a tesztben lévő információk szintjén szimmetrikusak. Ugyanis annak, aki mindent megold az elvi maximumot kell adnunk teljesítményként (nem tudjuk, mennyivel van felette ennek az értéknek), míg aki semmit sem, őt az elvi minimumra helyezzük (és nem tudjuk, mennyivel teljesítene ez alatt a szint alatt).

Wright [7] az alábbi jelöléseket használja, alkalmazza (a képletek hátterét, levezetését itt nem közöljük, azokat [7] anyagai alapján alkalmazzuk):

$$x_j = \ln\left[\frac{N - s_j}{s_j}\right],$$

$$x = \sum_{j=1}^{L} \frac{x_j}{L},$$

$$U = \sum_{j=1}^{L} \frac{(x_j - x)^2}{L - 1}.$$

Az első három formula esetében azt láthatjuk, hogy ez lényegében az itemek relatív megoldottságát jelzik, illetve azok varianciáját, hibáját mutatják. Azaz, hogy a teljes mintán általában hányan oldják az adott feladatokat, itemeket.

Az x-ben rejlő információk tehát az itemekre vonatkozó információkat jelölik, azokból származtatott adatok x_j, x, illetve U.

$$y_w = \ln\left[\frac{w}{L-w}\right],$$

$$y = \sum_{w=1}^{L-1} \frac{n_w y_w}{N},$$

$$V = \sum_{w=1}^{L-1} \frac{n_w \left(y_w - y\right)^2}{N-1}.$$

Az x-ekkel szemben az y és V mutatók azt mutatják, hogy egy adott hosszúságú teszten a vizsgálati alanyok átlagosan hány feladatot tudnak megoldani – tehát ez a három összevont formula (y_w, y és V) a vizsgálati alanyok átlagos teljesítményére vonatkozó mutatóegyütteseket fedi, továbbra is Wright [7] jelöléseit és számításait használva:

$$X = \sqrt{\frac{1 + \frac{U}{2{,}89}}{1 - \frac{UV}{8{,}35}}},$$

$$Y = \sqrt{\frac{1 + \frac{V}{2{,}89}}{1 - \frac{UV}{8{,}35}}}.$$

Az X és Y értékek tehát a teszt és a vizsgálati alanyok teljesítményét mutatják.

$$d_j = Y\left(x_j - x\right),$$

$$SE\left(d_j\right) = y\sqrt{\frac{N}{s_j\left(N - s_j\right)}}.$$

A d_j paraméter tehát az összes megoldóra megoldásra/teljesítményre vetítve az itemek x_j nehézsége, tehát $SE\left(d_j\right)$ az adott itemek nehézségének standard hibája.

$$b_w = Xy_w,$$

$$SE\left(b_w\right) = X\sqrt{\frac{L}{w\left(L-w\right)}}.$$

Ezzel szemben b_w az adott megoldók, adott tesztet kitöltők átlagos teljesítménye lesz, illetve $SE\left(b_w\right)$ a teljesítményeken lévő átlagos (standard) hibaként kerül bevezetésre.

Látható tehát, hogy e mutatók segítségével megadható, hogy az itemeknek, illetve a teljes tesztnek (teljesítménynek) mi lesz a hibája, milyen biztonsággal tudunk item-paramétert vagy teljesítményt meghatározni.

Ha azt tekintjük, hogy továbbra is adaptív módon, de kicsit nehezítve vagy könnyítve (tehát mindenki a saját képességének megfelelő itemet kap, de rögzített p, illetve q valószínűséggel oldja meg/rontja el a feladatokat, valamint úgy kezeljük, hogy mondjuk K szinten akarjuk az eredményeket kezelni (Wright [7] 11 szinttel számolt, tehát -5 és 5 közötti képességértékekkel dolgozott), akkor további, szintén nem túlságosan életszerűtlen egyszerűsítések tehetők. Tudjuk továbbá, hogy

$$s_i = \frac{Np}{K},$$

hiszen ilyen esetben tudható, hogy az adott szinten lévő alanyok nem kapnak más szintekről itemeket (tehát azok adott hányadát fogják megoldani).

Ez utóbbi úgy is felfogható (ezért nem életszerűtlen a megkötés), hogy egy adaptív teszt esetében a nagyon jó nem kap nagyon könnyű feladatokat és a nagyon alul teljesítő sem kap megoldhatatlannak látszó példákat. Miként egy igen súlyos állapotban lévő pácienstől sem kérdezik az enyhe tüneteket – és az alapvetően enyhébb panaszokkal érkezőket sem a rendkívül súlyos esetekre jellemző tünetek mentén kezelik.

Ebből az egyszerűsítésből következik, hogy

$$x_i = \ln\left(\frac{N - \frac{Np}{K}}{\frac{Np}{K}}\right) = \ln\left(\frac{K-p}{p}\right).$$

A fenti formula azt mondja, hogy a jó és rossz válaszok aránya átlagosan tehát csak a szintektől függ (minden szinten lényegében állandó, hogy hányan, milyen arányban oldják meg jól vagy rosszul a feladatokat).

Ebben az esetben az is elmondható, hogy

$$x = \frac{\sum_{i=1}^{L} \ln\left(\frac{K-p}{p}\right)}{N} = \frac{L}{N}\ln\left(\frac{K-p}{p}\right),$$

ami az átlagos megoldottsági/elrontottsági kapcsolati mutatónk.

Ezek után a variancia:

$$U = \frac{\sum_{i=1}^{L}\left(\ln\left(\frac{K-p}{p}\right) - \frac{L}{N}\ln\left(\frac{K-p}{p}\right)\right)^2}{L-1},$$

$$y_w = \ln\left(L\frac{p}{q}\right).$$

Tehát $w = Lp$ (a teljes teszten hány jó megoldást adunk), amiből következik, hogy optimális adaptív teszt esetében $V = 0$ és $Y = 1$. Optimálisan adaptív egy teszt akkor, ha valóban minden vizsgálati alany folyamatosan a számára megfelelő szinten, tehát p valószínűséggel megoldható feladatokat kap.

Innen viszont azt is tudhatjuk, hogy

$$SE\left(d_i\right) = \sqrt{\frac{N}{s_i\left(N - s_i\right)}} = \frac{1}{\sqrt{N}} \frac{K}{\sqrt{p\left(K - p\right)}}.$$

Azaz az adott itemek hibája annál nagyobb, minél több szintet szeretnénk vele bemérni, viszont minél több kitöltővel rendelkezünk, annál jobban csökken. Ez egybevág azzal az intuícióval, hogy minél szélesebb tartományon szeretnénk, hogy egy kérdés jól mérjen, annál nagyobb bizonytalansággal tudjuk megtenni (specifikus kérdések pontosabban mérnek). Illetve, hogy a kitöltők számának növekedésével együtt jár az, hogy az itemek viselkedését egyre pontosabban fogjuk ismerni.

Szintén Wright formulái alapján [7] megadható a teljesítmény hibája is:

$$SE(b) = X\sqrt{\frac{L}{w\left(L - w\right)}} = X\sqrt{\frac{1}{Lpq}}$$

Amiből további behelyettesítéssel:

$$SE(b) = \sqrt{\frac{1}{Lpq}}\sqrt{1 + \frac{\left(\frac{L}{L-1}\right)\left(\frac{N-L}{N}\right)^2 \ln^2\left(\frac{K-p}{p}\right)}{2,89}}$$

A fenti formulából látható például, hogy adott teszt-hossz esetében az esetszám növelésével egy ideig csökkenthető a hiba mértéke – majd lényegében stagnálni fog, ha semmi más paraméteren nem változtatunk. Ahogy az is látható, hogy N-ben egy idő után nem fogunk tudni jobb eredményt mutatni – tehát más megközelítésben egy-egy kitöltő hibáját attól nem fogjuk tudni jobban megbecsülni, hogy rajta kívül még sokan kitöltik a tesztet.

Ez azt is jelenti, hogy ha a teszt hosszát nem növeljük, akkor az esetszám növelésével egy-egy alanyra pontosabb becsléseket nem fogunk tudni adni. Ezt felfoghatjuk úgy is, hogy az adott itemek egy idő után kellően pontosan bemérésre kerülnek, tehát a belőlük nyerhető információ lényegében stagnál, tehát újabb és újabb esetek hozzávételével már nem tudunk további információkhoz jutni.

Ez azt is jelenti, hogy a képességeket csak úgy tudjuk egyre pontosabban mérni, hogy a teszt hosszát növeljük, ha újabb és újabb itemeket veszünk hozzá a tesztünkhöz.

Felhívjuk arra a figyelmet, hogy ez alapvetően nem mond ellent annak az intuitív megfigyelésnek, melyet például az egészséggel kapcsolatos diagnosztikában

csinálnak, vagy akár a teljesítménymméréseknél tapasztalhatunk. Az egészséggel kapcsolatos teszteknél nem az történik, hogy újra és újra azonos tesztekkel vért vesznek (ha nem ismert a diagnózis), hanem újabb teszteket, másfajta információkat csatornáznak be. A teljesítménymérés esetében sem írja meg a diák újra és újra ugyanazt a tesztet (típusfeladatot), hanem másfajta típusokkal igyekszünk pontosabb képet kapni a tudásáról.

Valamint arra is kitérnénk, hogy alapvetően a kitöltők száma jellemzően 100 – 200-as minimális nagyságrendet jelent, a tesztek hossza pedig ritkán megy 100-as feladatszám fölé. Tehát az esetszám emelkedésével valódi, empirikus esetekben a többi paramétert fixen tartva folyamatosan javuló teszt eredményeket fogunk tapasztalni.

4. Szimulációs eljárás, eredmények

A szimuláció során szintén Wright [7] nyomvonalát követjük. Esetében a kitöltők száma 50 és 500 fő között alakult itemenként. Ez szintén nem elrugaszkodott a valóságtól, hiszen egy-egy standardizálás során a kérdőíveket hagyományosan legalább 500 fővel szokás kitöltetni – de az online felmérések során ennél jellemzően lényegesen nagyobb minták keletkeznek.

Fontos azonban kiemelnünk, hogy Wright [7] nem adaptív, hanem papír-ceruza tesztek esetében határozta meg ezeket a számokat – és esetünkben éppen az a kérdés, hogy ennél kevesebb kitöltővel is el lehet-e érni hasonló eredményeket egy adaptív tesztelés során.

Két kérdésre keressük tehát a választ:

1. Első lépésben a kérdés az, hogy egy-egy item megbízhatóságához (adott hibahatár eléréséhez) minimálisan hány kitöltőre van szükség.

2. Második lépésben a kérdés az, hogy ha megvan egy megfelelő méretű itembankunk, akkor ebből az itembankból minimálisan hány kérdésre van szükség ahhoz, hogy az egyes válaszadók teljesítményét meg tudjuk határozni.

Wright nyomán (a hiba $N(0,1)$ tehát) a $0,2$-es szintet határozzuk meg mint elérni kívánt minimumot. Ez nagyjából azt jelenti, hogy a teljes teszt esetében az itemek megbízhatósága a teljes teszt megbízhatóságának 20%-a alá kell, hogy csökkenjék. A kitöltőkre vonatkozó hibát/szórást ezzel szemben $0,5$-ös szinten határozzuk meg.

A szinteket a hagyományos OECD PISA [8], illetve Országos kompetenciamérés [9] szintjeihez szabjuk, azaz a szintek száma a szimulációkban 2 és 8 lesznek, azaz $K = 2, \ldots, 8$. Jellemzően e két felmérés esetében a diákok teljesítményén lévő hiba nagyságrendileg a szórás 40–50%-a is lehet. Ezért maradtunk a teljesítmények esetében a $0,5$-ös szint elérése mellett. A második esetben, amikor a teljesítmények teljes teszt szintjét fogjuk vizsgálni, $K = 3, 5, 8$ változatok esetére mutatjuk

be az elemzés eredményeit. Ezt úgy foghatjuk fel, mint az alacsony-közepes-magas ($K = 3$) esetben. Második esetben az iskolai osztályzatokat kezelhetjük szintekként. A harmadik, $K = 8$ esetben pedig egy részletesebb, betegségsúlyosság esetét vehetjük alapul.

A szimulációkban $N = \{10, 20, 30, 50, 100, 200, 300, 400, 500\}$ esetszámokkal fogunk dolgozni.

A teszt hosszát $L = \{10, 30, 60\}$ értékekre állítjuk be.

A megoldottsági szinteket $p = \{0,1; 0,2; 0,3; 0,4; 0,5\}$ esetekre állítjuk be.

Első lépésben azt vizsgáljuk meg, hogy az itemek hibája miként alakul a szintek, a megoldottsági valószínűségek és a kitöltők száma alapján.

Az eredményekből leolvasható (1. táblázat), hogy $p = 0,1$, illetve $p = 0,2$ (mely értelemszerűen megegyezik a $p = 0,9$ és a $p = 0,8$ esetekkel) legalább 400, inkább 500 fő kell ahhoz, hogy elérjük a 0,2-es alsó határt. A tapasztalatok egyébiránt azt mutatják, hogy ez jellemzően a nagyon nehéz/nagyon könnyű feladatok világa, amely esetben valóban azt láthatjuk, hogy több esetre van szükség a megfelelő minőség garantálásához.

Ráadásul ez a szintek emelkedésével még nehezebbé is válik – tehát minél több szintet kalibrálunk (minél árnyaltabban szeretnénk mérni), annál több kitöltőre van szükség a szélsőséges feladatok pontos bemérésére.

Ezzel szemben ha megnézzük a 0,4-es, illetve 0,5-ös szinteket, tehát a kiegyensúlyozottabb itemeket (nagyjából fele-fele arányban oldják vagy nem oldják meg), ilyen esetekben már jellemzően 100, illetve 200 kitöltő is elegendő a megfelelő szint biztosítására (ez alól csak a 8 szint esetében van kivétel).

Mit láthatunk akkor, ha mindezt kiegészítjük a teszt hosszával? Azt láthatjuk (2. táblázat), hogy egy 10 itemből álló teszt esetében lényegében nem tudjuk elérni a 0,4-es vagy 0,5-ös minimális szintet (és ezt egyébiránt az empirikus tapasztalatok is alátámasztják, ennyire rövid teljesítményt mérő tesztek általában nincsenek).

Ezzel szemben 30 item esetében már akár 20–30 kitöltővel is elérhető $p = 0,3$-mas megoldottsági nehézség esetében. Igaz ez $K = 3, 5, 8$ esetben egyaránt. Ez azt jelenti, hogy egy aránylag bonyolultabb teszt esetében is, akár még 8 szintet megkülönböztetve, a teljes teszt megbízhatósága már 30–40 kitöltővel is elérhető, amennyiben adaptív tesztet tudunk a kérdésben összeállítani megfelelő itembankkal rendelkezve.

5. Két példa gyakorlati felhasználásra

Jellemzően az OECD PISA [8] és az Országos kompetenciamérés [9] olyan felmérések, ahol $N >> L$, azaz lényegesen, nagyságrendekkel több kitöltő diák van, mint ahány item egy-egy felmérés során felhasználásra kerül egy tesztfüzetben. Jellemzően egy tesztfüzet egy-egy témakörre 50–60 itemet tartalmaz, míg a kitöltők

K	p	$N=10$	$N=20$	$N=30$	$N=50$	$N=100$	$N=200$	$N=300$	$N=400$	$N=500$
2	0,1	1,451	1,026	0,8377	0,6489	0,4588	0,3244	0,2649	0,2294	0,2052
2	0,2	1,0541	0,7454	0,6086	0,4714	0,3333	0,2357	0,1925	0,1667	0,1491
2	0,3	0,8856	0,6262	0,5113	0,3961	0,2801	0,198	0,1617	0,14	0,1252
2	0,4	0,7906	0,559	0,4564	0,3536	0,25	0,1768	0,1443	0,125	0,1118
2	0,5	0,7303	0,5164	0,4216	0,3266	0,2309	0,1633	0,1333	0,1155	0,1033
3	0,1	1,7617	1,2457	1,0171	0,7878	0,5571	0,3939	0,3216	0,2785	0,2491
3	0,2	1,2677	0,8964	0,7319	0,5669	0,4009	0,2835	0,2315	0,2004	0,1793
3	0,3	1,0541	0,7454	0,6086	0,4714	0,3333	0,2357	0,1925	0,1667	0,1491
3	0,4	0,9303	0,6578	0,5371	0,416	0,2942	0,208	0,1698	0,1471	0,1316
3	0,5	0,8485	0,6	0,4899	0,3795	0,2683	0,1897	0,1549	0,1342	0,12
4	0,1	2,0255	1,4322	1,1694	0,9058	0,6405	0,4529	0,3698	0,3203	0,2864
4	0,2	1,451	1,026	0,8377	0,6489	0,4588	0,3244	0,2649	0,2294	0,2052
4	0,3	1,2006	0,849	0,6932	0,5369	0,3797	0,2685	0,2192	0,1898	0,1698
4	0,4	1,0541	0,7454	0,6086	0,4714	0,3333	0,2357	0,1925	0,1667	0,1491
4	0,5	0,9562	0,6761	0,5521	0,4276	0,3024	0,2138	0,1746	0,1512	0,1352
5	0,1	2,2588	1,5972	1,3041	1,0102	0,7143	0,5051	0,4124	0,3571	0,3194
5	0,2	1,6137	1,1411	0,9317	0,7217	0,5103	0,3608	0,2946	0,2552	0,2282
5	0,3	1,3316	0,9416	0,7688	0,5955	0,4211	0,2977	0,2431	0,2105	0,1883
5	0,4	1,1656	0,8242	0,673	0,5213	0,3686	0,2606	0,2128	0,1843	0,1648
5	0,5	1,0541	0,7454	0,6086	0,4714	0,3333	0,2357	0,1925	0,1667	0,1491
6	0,1	2,4702	1,7467	1,4261	1,1047	0,7811	0,5523	0,451	0,3906	0,3493
6	0,2	1,7617	1,2457	1,0171	0,7878	0,5571	0,3939	0,3216	0,2785	0,2491
6	0,3	1,451	1,026	0,8377	0,6489	0,4588	0,3244	0,2649	0,2294	0,2052
6	0,4	1,2677	0,8964	0,7319	0,5669	0,4009	0,2835	0,2315	0,2004	0,1793
6	0,5	1,1442	0,809	0,6606	0,5117	0,3618	0,2558	0,2089	0,1809	0,1618
7	0,1	2,6649	1,8843	1,5386	1,1918	0,8427	0,5959	0,4865	0,4214	0,3769
7	0,2	1,8981	1,3422	1,0959	0,8489	0,6002	0,4244	0,3466	0,3001	0,2684
7	0,3	1,5613	1,104	0,9014	0,6983	0,4937	0,3491	0,2851	0,2469	0,2208
7	0,4	1,3624	0,9633	0,7866	0,6093	0,4308	0,3046	0,2487	0,2154	0,1927
7	0,5	1,2279	0,8682	0,7089	0,5491	0,3883	0,2746	0,2242	0,1941	0,1736
8	0,1	2,8463	2,0126	1,6433	1,2729	0,9001	0,6364	0,5197	0,45	0,4025
8	0,2	2,0255	1,4322	1,1694	0,9058	0,6405	0,4529	0,3698	0,3203	0,2864
8	0,3	1,6645	1,177	0,961	0,7444	0,5264	0,3722	0,3039	0,2632	0,2354
8	0,4	1,451	1,026	0,8377	0,6489	0,4588	0,3244	0,2649	0,2294	0,2052
8	0,5	1,3064	0,9238	0,7542	0,5842	0,4131	0,2921	0,2385	0,2066	0,1848

1. táblázat. Itemek megbízhatóságának táblázata kitöltők számának, szintek nagyságának és a teszt nehézségének függvényében

K	L	p	10	20	30	50	100	200	300	400	500
3	10	0,1	1,054	1,524	1,807	2,052	2,244	2,341	2,374	2,391	2,401
3	10	0,2	0,791	1,021	1,17	1,302	1,407	1,461	1,479	1,489	1,494
3	10	0,3	0,69	0,835	0,932	1,021	1,092	1,129	1,141	1,147	1,151
3	10	0,4	0,645	0,746	0,816	0,881	0,933	0,961	0,97	0,975	0,978
3	10	0,5	0,632	0,707	0,76	0,809	0,85	0,872	0,879	0,882	0,885
3	30	0,1	2,527	0,864	0,609	0,782	1,052	1,207	1,26	1,287	1,303
3	30	0,2	1,512	0,582	0,456	0,54	0,68	0,764	0,793	0,808	0,817
3	30	0,3	1,121	0,477	0,398	0,45	0,541	0,597	0,617	0,627	0,633
3	30	0,4	0,914	0,427	0,373	0,408	0,474	0,515	0,529	0,537	0,541
3	30	0,5	0,792	0,405	0,365	0,391	0,44	0,472	0,483	0,489	0,493
3	60	0,1	4,319	1,772	0,961	0,463	0,551	0,74	0,811	0,848	0,87
3	60	0,2	2,547	1,061	0,6	0,338	0,381	0,479	0,517	0,537	0,549
3	60	0,3	1,857	0,787	0,463	0,291	0,318	0,381	0,407	0,42	0,429
3	60	0,4	1,487	0,642	0,394	0,27	0,288	0,334	0,352	0,362	0,368
3	60	0,5	1,259	0,557	0,357	0,263	0,276	0,311	0,325	0,333	0,337
5	10	0,1	1,054	1,652	1,997	2,292	2,52	2,636	2,675	2,695	2,707
5	10	0,2	0,791	1,11	1,305	1,476	1,61	1,678	1,701	1,712	1,719
5	10	0,3	0,69	0,907	1,045	1,168	1,264	1,314	1,331	1,339	1,344
5	10	0,4	0,645	0,81	0,917	1,014	1,091	1,131	1,144	1,151	1,155
5	10	0,5	0,632	0,765	0,854	0,936	1,001	1,034	1,046	1,052	1,055
5	30	0,1	2,899	0,934	0,609	0,832	1,164	1,35	1,413	1,445	1,464
5	30	0,2	1,795	0,63	0,456	0,573	0,76	0,867	0,905	0,923	0,935
5	30	0,3	1,371	0,516	0,398	0,477	0,608	0,685	0,712	0,726	0,734
5	30	0,4	1,151	0,462	0,373	0,432	0,533	0,594	0,616	0,627	0,633
5	30	0,5	1,027	0,437	0,365	0,413	0,496	0,548	0,566	0,575	0,58
5	60	0,1	4,986	2,033	1,083	0,474	0,586	0,818	0,904	0,948	0,974
5	60	0,2	3,059	1,259	0,689	0,345	0,404	0,534	0,584	0,61	0,625
5	60	0,3	2,316	0,962	0,539	0,296	0,336	0,428	0,463	0,482	0,493
5	60	0,4	1,927	0,808	0,464	0,274	0,305	0,375	0,403	0,418	0,427
5	60	0,5	1,702	0,721	0,424	0,267	0,291	0,35	0,373	0,385	0,393
8	10	0,1	1,054	1,775	2,176	2,516	2,778	2,911	2,955	2,977	2,991
8	10	0,2	0,791	1,196	1,435	1,64	1,799	1,88	1,908	1,921	1,929
8	10	0,3	0,69	0,979	1,155	1,308	1,428	1,489	1,509	1,52	1,526
8	10	0,4	0,645	0,874	1,017	1,143	1,242	1,292	1,309	1,318	1,323
8	10	0,5	0,632	0,826	0,949	1,059	1,146	1,191	1,206	1,213	1,218
8	30	0,1	3,24	1,002	0,609	0,881	1,269	1,483	1,556	1,593	1,615
8	30	0,2	2,052	0,677	0,456	0,607	0,836	0,965	1,009	1,032	1,045
8	30	0,3	1,598	0,555	0,398	0,504	0,672	0,769	0,802	0,819	0,829
8	30	0,4	1,365	0,497	0,373	0,456	0,592	0,671	0,699	0,713	0,721
8	30	0,5	1,238	0,47	0,365	0,435	0,552	0,621	0,646	0,658	0,665
8	60	0,1	5,594	2,272	1,196	0,485	0,62	0,892	0,991	1,041	1,072
8	60	0,2	3,522	1,439	0,772	0,352	0,428	0,588	0,647	0,678	0,697
8	60	0,3	2,726	1,121	0,611	0,302	0,356	0,473	0,517	0,54	0,554
8	60	0,4	2,316	0,958	0,53	0,279	0,321	0,416	0,453	0,472	0,483
8	60	0,5	2,09	0,869	0,489	0,271	0,307	0,389	0,42	0,437	0,447

2. táblázat. Teljesítmények hibájának táblázata kitöltők számának, a teszt hosszának, szintek nagyságának és a teszt nehézségének függvényében

száma több tízezres nagyságrendet jelent (az OECD PISA esetében országonként is több ezer diák tölti ki a teszteket).

Mindkét teszt esetében próbateszteket tartanak (e próbateszteken mérik be a későbbiekben használatra kerülő itemeket), ami azt jelenti, hogy az itembank, ami rendelkezésre áll, nagyságrendileg 20–30-szorosa egy-egy évben a végül felhasználásra kerülő itemeknek. De ez azt is jelenti, hogy a korábbi évek itemjeivel együtt olyan gazdag itemállomány áll rendelkezésre, hogy akár adaptív módon is könnyen lehet mérést szervezni az itemek kimerülésének, korrumpálódásának kockázata nélkül [1].

A szimulációs eredmények megmutatták, hogy 200–300 kitöltő esetében érdemi különbségek a szintek ($K = 4$ és felette) és a megoldottságok között ($p = 0,3$ felett) már nincsenek, lényegében hasonlatos működéseket tapasztalunk. Az alábbi megkötések tehát gyakorlati szempontból életszerűek.

Tegyük fel tehát, hogy

1. $K = 5$: iskolai környezetben az 1-es és 5-ös közötti osztályzatokra asszociáló megkötések nem idegen megközelítések, ezeket mind a pedagógusok, mind a diákok megfelelő módon tudják értelmezni.

2. N legyen 100 és 500 között rögzített érték. Jellemzően ugyanis az OECD PISA [8] és a kompetenciamérés is [9] a próbamérések során az itemeket nagyjából ennyi diákkal tölteti ki. Az általános tapasztalatok alapján $N = 300$ megfelelő értéknek mutatkozik.

3. $p = 0,3$, $p = 0,4$ és $p = 0,5$ esetekkel dolgozunk, ugyanis ennél nehezebb vagy könnyebb teszteket jellemzően nem szokás íratni – legalábbis felmérés jelleggel azok a tesztek, amiket mindenki megold, illetve senki sem tud rajta érdemben jól teljesíteni, tömegesen nem alkalmazottak.

A fenti megkötések azért is érdekesek lehetnek, mert minket alapvetően nem az érdekel, hogy 100 vagy 500 kitöltőre van szükség. A mostani infrastruktúra mellett 300 diákkal egy teljesítményt mérő tesztet kitöltetni érdemi költségekkel nem jár. A $K = 5$ megkötés nem érdemi megkötés egy iskolai rendszerben. A tesztek nehézségének 30%–70% közötti rögzítése szintén kellően bő keretet szolgáltathat tesztek összeállításához.

Megfigyelhető (3. táblázat), hogy egy nehezebb tesztnél (30%-os megoldottság) 54 kérdéssel érhető el a megfelelő megbízhatósági szint. Könnyebb tesztnél (40%-os megoldottság) 43 itemes teszt mutatja az alsó határt, míg egy alapvetően igazságos, 50%-os megoldási szintre beállított teszt esetében 38 kérdésb?l álló tesztektől már elfogadható szintet érhetünk el.

A fenti táblázatokból látható, illetve az általános tapasztalatok [8, 9] azt mutatják, hogy papír-ceruza tesztek esetében akár 60–70 kérdés is megtalálható a

	0,3	0,4	0,5
30	0,712	0,616	0,566
31	0,699	0,604	0,555
32	0,686	0,593	0,545
33	0,673	0,582	0,535
34	0,661	0,572	0,526
35	0,650	0,563	0,517
36	0,639	0,553	0,509
37	0,629	0,544	0,501
38	0,618	0,536	**0,493**
39	0,609	0,527	**0,485**
40	0,599	0,519	**0,478**
41	0,590	0,512	**0,471**
42	0,582	0,504	**0,465**
43	0,573	**0,497**	**0,458**
44	0,565	**0,490**	**0,452**
45	0,557	**0,484**	**0,446**
46	0,550	**0,477**	**0,440**
47	0,542	**0,471**	**0,434**
48	0,535	**0,465**	**0,429**
49	0,528	**0,459**	**0,423**
50	0,522	**0,453**	**0,418**
51	0,515	**0,447**	**0,413**
52	0,509	**0,442**	**0,408**
53	0,502	**0,437**	**0,403**
54	**0,496**	**0,432**	**0,399**
55	**0,491**	**0,427**	**0,394**
56	**0,485**	**0,422**	**0,390**
57	**0,479**	**0,417**	**0,385**
58	**0,474**	**0,412**	**0,381**
59	**0,469**	**0,408**	**0,377**
60	**0,463**	**0,403**	**0,373**

3. táblázat. Könnyebb tesztek minimális itemszáma

tesztfüzetekben. Látható azonban, hogy adaptív módon ennek töredékével, 40 kérdéssel már egy 35–40%-os szórással, hibával rendelkező teszt is összeállítható. Egy ilyen adaptív teszt az alábbi előnyökkel rendelkezik:

1. Fele annyi itemmel, fele annyi idő alatt tudunk felmérést készíteni.

2. A diákok a nekik megfelelő szintű feladatokat kapják [3], tehát nem unják meg a feladatokat, jellemzően mindenki a számára még kihívást jelentő szinten dolgozik.

3. Ha van a felmérésre fennmaradó idő, akkor a próbafelmérés során használandó itemek azok, amelyeket a diákok a fennmaradó időben oldhatnak – így a következő mérés próbaidőszaka az előző időszakkal összerakható, párhuzamosan vezényelhető.

A fenti előnyök fenntartása mellett egy második példát is bemutatunk. Elsősorban felsőoktatásban vagy szakképzési területeken fordulnak elő olyan tesztek, melyek szintén adaptívvá tehetők és valójában $K = 2$ szintet követelnek meg (teljesített vagy nem teljesített). Ez esetben tehát nem feltétlenül osztályzatokat képzelünk el, hanem egy elvárt szint teljesítését tűzzük ki célként.

Ilyen esetben jellemzően nehezebbek a teljesítések, tehát $p = 0,2$, $p = 0,25$ és $p = 0,3$ eseteket fogjuk bemutatni (azaz a teljesítéshez egy 20%-os sikerességi határt veszünk, mint legnehezebb teljesítési szintet). Ezt értelemszerűen úgy is interpretálhatjuk, hogy a vizsga teljesítéséhez legalább 80%-os teljesítményre van szükség.

A megoldók száma továbbra is legyen $N = 300$ főben rögzítve, mely akár felsőoktatási, akár szakképzési keretek között életszerű feltételezés.

Bár a rendszer szigorúbb, itt is $0,5$-ös megbízhatóságot fogunk minimum határként megadni – tehát vizsgázónként hasonlóan pontos becslést szeretnénk az adaptív teszttől átlagosan elvárni.

E második esetben (4. táblázat) látható, hogy 30%-os, némileg megengedőbb teljesítési szint esetében már 37 kérdésnél elérhető a 0,5-ös megbízhatóság szint. Kicsit szigorítva, $p = 0,25$-nél ehhez minimum 46 itemre van szükség, illetve $p = 0,2$ esetében minimum 59 item lesz az alsó határ.

Azonban ez azt is jelenti, hogy adaptív módon egy teljesített/nem teljesített rendszer működtetésénél 60 item már elegendő annak kiderítésére, hogy az adott vizsgázó valóban megfelelő szinten helyezkedik-e el.

Abban az esetben ugyanis, ha 60 item segítségével egy nehezebb tesztet választunk, ebből már a teljesítményt mérő modellek segítségével a megfelelő képességpontja számítható a vizsgázónak – abból pedig láthatóvá válik, hogy eléri-e a számunkra elfogadható szintet vagy sem.

	0,2	0,25	0,3
30	0,726	0,625	0,557
31	0,713	0,614	0,547
32	0,700	0,603	0,538
33	0,688	0,593	0,529
34	0,677	0,584	0,520
35	0,666	0,575	0,512
36	0,656	0,566	0,504
37	0,646	0,557	**0,497**
38	0,637	0,549	**0,490**
39	0,628	0,541	**0,483**
40	0,619	0,534	**0,476**
41	0,610	0,527	**0,470**
42	0,602	0,520	**0,464**
43	0,594	0,513	**0,458**
44	0,587	0,506	**0,452**
45	0,579	0,500	**0,446**
46	0,572	**0,494**	**0,441**
47	0,565	**0,488**	**0,436**
48	0,558	**0,482**	**0,431**
49	0,552	**0,477**	**0,426**
50	0,546	**0,471**	**0,421**
51	0,539	**0,466**	**0,416**
52	0,533	**0,461**	**0,412**
53	0,528	**0,456**	**0,407**
54	0,522	**0,451**	**0,403**
55	0,517	**0,447**	**0,399**
56	0,511	**0,442**	**0,395**
57	0,506	**0,438**	**0,391**
58	0,501	**0,433**	**0,387**
59	**0,496**	**0,429**	**0,384**
60	**0,491**	**0,425**	**0,380**

4. táblázat. Nehezebb tesztek minimális itemszáma

6. Konklúzió

Megállapítható tehát, hogy még akár igen nehéz teszteket, a diákok számára kihívást jelentő kérdéseket alkalmazva is 60 kérdést összeállítva megfelelő pontosság érhető el ahhoz, hogy a megoldott feladatokból visszaszámítva a diák képességpontját, általánosságban elfogadható értékelést tudjunk biztosítani.

Ezzel szemben viszont, jellemző iskolai körülményeket szimulálva a tesztekhez ($K = 5$, tehát 5 fokozatú értékelést alkalmazva, $p = 0,4$, illetve $p = 0,5$, azaz átlagos tesztnehézséget feltételezve) ez az itemszám lényegesen kevesebb, 40–50 darabos szinten áll meg. Ez lényegében megegyezik azokkal a mutatókkal, melyekkel mind a nemzetközi [8], mind a hazai mérések [9] általánosságban találkoznak.

Továbbá ez azt is jelenti, hogy egy adaptív teszt esetében a mostani mérési idő jelent?sen csökkenthető (a hivatkozott mérések [8], [9] jellemzően tudásterületenként 50–60 itemet használnak) – és a fennmaradó idő a próbaidőszak itemeinek bemérésére fordítható.

Ez azt is jelenti, hogy egy-egy időszakban nem kell kétszeres logisztikát alkalmazni (próbaméréseket tartani az itemek tesztelésére), hanem az amúgy is mérésre fordított időben lehet a tesztelést megvalósítani. Tehetjük ezt akár úgy is, hogy a valós tesztitemek közé tesszük véletlenszerű helyekre a bemérendő próbafeladatokat [1] annak érdekében, hogy a kifáradást elkerüljük, a diákok érdeklődését folyamatosan fenntartsuk. Ez általánosságban is teljesülhet olyan tesztekkel, melyek rövidebbek és folyamatosan olyan kérdéseket adnak a diákoknak, amik a tudásszintjüknek éppen megfelelnek.

Megfigyelhető tehát, hogy az adaptív mérések során megfogalmazott általános nehézségek (itembank mérete, illetve az itemek kifáradásának kérdése) egy jól szervezett logisztikával, illetve a korábban már bemért itemek felhasználásával általánosságban orvosolható – és a tapasztalatok alapján még akár a szélsőségesebb megoldási mutatókkal bíró alanyok (20%-os megoldottság, tehát nehéz körülmények) is a mostani teszthosszok segítségével bemérhetők.

Köszönetnyilvánítás

A kutatás az Innovációs és Technológiai Minisztérium ÚNKP-20-3 kódszámú Új Nemzeti Kiválóság Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült.

Hivatkozások

[1] G.J. Cizek and J.A. Wollack: *Handbook of Quantitative Methods for Detecting Cheating on Tests*, Routledge, (2016). DOI: 10.4324/9781315743097

[2] P.M.C. Harrison, T. Collins and D. Müllensiefen: *Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation*, Scientific Report Vol. **7** No. **1**, article number: 3618 (2017). DOI: 10.1038/s41598-017-03586-z

[3] G.G. Kingsbury: *Adaptive Item Calibration: A Process for Estimating Item Parameters Within a Computerized Adaptive Test*, Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing, (2009).

[4] G.F. Kuder and M.W. Richardson: *The theory of the estimation of test reliability*, Psychometrika Vol. **2** No. **3**, pp. 151–160 (1937). DOI: 10.1007/BF02288391

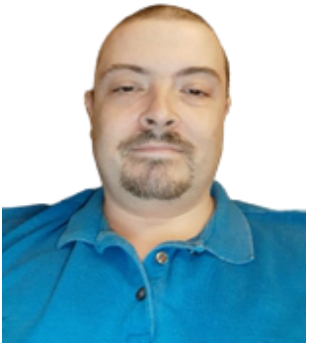
[5] J.W. van der Linden and C.A.W, Glas: *Capitalization on Item Calibration Error in Adaptive Testing*, Applied Measurement in Education Vol. **13** No. **1**, pp. 35–53 (2000). DOI: 10.1207/s15324818ame1301_2

[6] S.L. Wise: *A Critical Analysis of the Arguments for and against Item Review in Computerized Adaptive Testing*, ERIC, New York, (1996).

[7] B.D. Wright: *Solving Measurement Problems with the Rasch Model*, Journal of Educational Measurement, Vol. **14** No. **2**, pp. 97–116 (1977). DOI: 10.1111/j.1745-3984.1977.tb00031.x

[8] OECD, PISA http://www.oecd.org/pisa/

[9] Országos kompetenciamérés https://www.oktatas.hu/kozneveles/meresek/kompetenciameres



Takács Szabolcs 1980-ban született Marcaliban, 1998-ban a Berzsenyi Dániel Gimnáziumban érettségizett az utolsó 4 évfolyamos speciális matematika tantervű osztályban (megjegyzés: Vizvári Béla az első speciális matematika tantervű osztályban érettségizett, szintén a Berzsenyiben).
2004-ben végzett az ELTE TTK-n Alkalmazott matematikusként. Első munkahelyein statisztikusként dolgozott (Fővárosi Munkaügyi Központ, statisztikai referens, majd Sulinova Kht- később Oktatási Hivatal mérésértékelési osztály, statisztikus). 2014-ben doktorált az ELTE TTK Matematika Doktori Iskolájában, 2008 óta oktat a Károli Gáspár Református Egyetemen a Pszichológiai Intézet Általános Lélektani és Módszertani tanszékén statisztikát. Egyetemi oktatói állása mellett 2013 óta az ANIMA cégcsoporttal dolgozik, ahol biztonságtechnológiával, biztonságtechnikai fejlesztésekkel, ipari és nyomozati innovációval foglalkoznak.

2016 óta az Oktatási Hivatalban digitális mérések szakértőjeként dolgozik, itt a különböző országos mérések papír-ceruza tesztről digitális formára történő mérési átállását készítik elő.

2019 óta a CIVIL biztonsági szolgálatnál is dolgozik szakértőként, itt digitális oktató- és mérőterek megvalósításán, valamint munkahelyi beválás előrejelzésének rendszerein dolgoznak együtt egy több éves fejlesztési projekt munkatársaival (pszichológusokkal, mérnökökkel, biztonsági szakemberekkel).

TAKÁCS SZABOLCS

Károli Gáspár Református Egyetem, BTK, Pszichológiai Intézet, Általános Lélektani és Módszertani Tanszék,
1037, Budapest, Bécsi út 324.
takacs.szabolcs.dr@gmail.com

T. Kárász Judit 1979-ben született Budapesten, 2004-ben az ELTE TTK-n végzett alkalmazott matematikus szakon.
2004-től a Sulinova Kht. Értékelési Központjának, később Oktatási Hivatal, Köznevelési Mérés Értékelési Osztályának statisztikus munkatársa. Főbb feladatai az Országos kompetenciamérés statisztikusi teendőinek ellátása, a kiegészítő mérés elemzésétől a különböző szintű jelentések adatainak előállításáig. További feladatai az Idegennyelvi és Célnyelvi mérés, illetve a Középfokú felvételi központi írásbeli feladatsorai feladatainak elemzése. Munkája során részt vett a köznevelés keretrendszeréhez kapcsolódó mérési-értékelési és digitális fejlesztések, innovatív oktatásszervezési eljárások kialakítása, megújítása (EFOP 3.2.15.) és a Tematikus együttműködés erősítése a köznevelés és felsőoktatás terén a Kárpát-medence szomszédos országaival (EFOP 3.10.1.) projektekben. 2019-től az ELTE PPK Neveléstudományi Doktori Iskola hallgatója, kutatási témája az Adaptív teljesítménymérési algoritmusok kidolgozása az Országos kompetenciamérés adatainak felhasználásával. Eközben részt vett A köznevelés módszertani megújítása a végzettség nélküli iskolaelhagyás csökkentése céljából a köznevelési intézményekben (EFOP-3.1.2-16) és A felsőoktatás hozzáférhetőségének javítása, komplex fenntartható tanulástámogatási környezet kialakítása, az oktatás innovatív megújítása az ELTE telephelyein (EFOP-3.4.3-16) projektekben. 2020-ban elnyerte az Új Nemzeti Kiválóság Program ösztöndíját. Óraadóként oktat a KRE BTK Pszichológiai Intézetében statisztikát, az ELTE PPK Neveléstudományi Intézetében társoktatóként kutatásmódszertani alapokat.

T. KÁRÁSZ JUDIT

ELTE, PPK, Neveléstudományi Doktori Iskola,
1075, Budapest, Kazinczy utca 23-27.
t.karasz.judit@ppk.elte.hu

THE RELATIONSHIPS OF THE MINIMUM LENGTH, ERROR, EVALUATION LEVELS AND NUMBER OF RESPONDENTS OF ADAPTIVE TESTS WITH GENERAL SOLUTION PROBABILITY

Judit Kárász T., Szabolcs Takács

Correlations between the minimum length, error, evaluation level and number of solvers of adaptive tests - with overall solution rate Our article basically aims to create a general formalism that can be used to address one of the problems of digitization of paper-and-pencil-based tests. The developed formalism was constructed in a general way, but the results are also described in a practical example.

In our article, we present the limitations and benefits of a future adaptive version of a survey based essentially on a paper-and-pencil test, such as the OECD PISA [8] and the National Assesment of Basic Competencies [9]. In the case of international measurements, the development work that has prepared this direction has already begun, and the recent online educational experience has paved the way for the transition to digital and then adaptive measurement in the case of the National Assesment of Basic Competencies.

The necessary item-level logistics (task bank, modeling environment) are basically available. However, the literature generally covers these technical details - as well as how many people are needed for reliability. In our article, we were not interested in this question, because the itembank will be of sufficient size, or typically in a national measurement, the sufficient filling will be given by default. In contrast, the most important question for us is to what extent the test can be shortened to a test that is the same - or better, more reliable than the current one. So we are not interested in the number of respondents, but in the length of the test, we performed simulation calculations based on the commonly published indicators of the current [8], [9] measurements.