Magyar Pszichológiai Szemle, 2016, 71. 1/8. 143–163.
DOI: 10.1556/0016.2016.71.1.8

Az IKT-eszközök a tanulás közvetlen közegében

SZÁMÍTÓGÉPES, ADAPTÍV IQ-MÉRÉS: EGY GYAKORLATI PÉLDA

KOVÁCS KRISTÓF[1] – TEMESVÁRI ESZTER[2]

[1] Eszterházy Károly Főiskola Pszichológia Tanszéke
[2] Budapesti Műszaki és Gazdaságtudományi Egyetem
e-mail: kristof340@googlemail.com; esztertem@gmail.com

Beérkezett: 2015. október 25. – *Elfogadva:* 2015. december 10.

Az adaptív számítógépes tesztelés (CAT) a kognitív képességek mérésének legkorszerűbb formája, amely egyfelől a modern tesztelméletre, másfelől az informatikai lehetőségekre épül. Jelen tanulmányban egy gyakorlati példán, a Mensa HungarIQa adaptív IQ-tesztjén keresztül mutatjuk be a CAT elméletét és gyakorlatát. A cikk először áttekinti az intelligencia kutatásának legfontosabb eredményeit és az IQ-mérés történetét és módszereit. Ezt követően a pszichometria főbb területeit, a klasszikus tesztelméletet és az item-válasz elméletet tárgyalja, valamint a CAT alapjait. Végül bemutatja egy konkrét adaptív IQ-teszt készítésének folyamatát, az itemek készítésétől az item-paraméterek becslésén és az adaptív algoritmus összeállításán át a teszt validálásáig.

Kulcsszavak: *IQ, intelligencia, számítógépes adaptív tesztelés*

AZ INTELLIGENCIA MÉRÉSE

Az intelligencia mérésével Francis Galton és Alfred Binet próbálkoztak először, teljesen eltérő megközelítést alkalmazva. Galton 1884-ben – majd a nyomában James MacKeen Cattell 1890-ben – a kor filozófiai és pszichológiai ismereteire alapozva alakították ki tesztjeiket, amelyek elsősorban az érzékelés gyorsaságát és pontosságát mérték, vagyis elemi működésekre összpontosítottak (Fancher, 1985).

Binet-t ezzel szemben nem alaptudományi megfontolások vezették: a francia oktatási hatóság megbízásából készített olyan tesztet, amely képes kiszűrni az iskolai oktatásra alkalmatlan gyerekeket. Ugyanakkor technikai szempontból az ér-

telmességét a magasabb, gondolkodási működések közvetlen mérésével próbálta jellemezni. Ebből született meg a mai IQ-tesztek ősének tekinthető Binet–Simon-teszt. Az 1905-ben publikált teszt 30 fokozatosan nehezedő feladatból állt. 1908-ban dolgozták át és egészítették ki ezt a skálát annak érdekében, hogy a tételek a normál értelmi képességekkel rendelkező diákok számára se legyenek túlságosan egyszerűek. A vizsgálatot kétszemélyes helyzetben végezték el a gyerekkel, és ahol a gyerek képességei már nem voltak elegendők a következő szint megválaszolásához, befejezték a vizsgálatot. Ezt követően az addig elért eredmények alapján megállapították mentális életkorát, majd ezt életkorával összehasonlítva meghatározták, hol tart a fejlődésben. Világkarriert a teszt Lewis Terman (1916) által adaptált amerikai változata, a Stanford–Binet futott be.

Terman már Stern mentális hányadosával dolgozott, amelynek számításakor a mentális korból nem kivonják az életkort, hanem elosztják vele, valamint bevezette a 100-as szorzót, így alakult ki az IQ ismert képlete: (mentális kor/életkor) × 100. Az első világháború idején az amerikai hadseregben is felmerült az igény a mentális képességek mérésére a sorozottak alkalmassági vizsgálatának részeként. Itt azonban egyrészt felnőtteket kellett mérni, másrészt nagyon nagy számban, amire a Binet-teszt változatai nem voltak alkalmasak. A hadsereg megbízásából készítették el amerikai pszichológusok, Robert Yerkes (1921) vezetésével a US Army Alfa és Army Betha teszteket (utóbbit az írástudatlanok számára), amelyeket már csoportosan is kitölthettek, ezáltal kevesebb időt vettek igénybe és költséghatékonyabbak is voltak (Stern, Terman és Yerkes munkásságáról lásd Fancher, 1985).

A legismertebb ma is használatos tesztek egyik csoportját a Wechsler által kifejlesztett intelligenciatesztek jelentik. A ma használt Wechsler-skálákat mind különböző korcsoportokra tervezték, így a tesztnek három fő változata létezik: az iskoláskor előtti, az iskoláskortól 16 éves korig tartó célcsoportot mérő és a felnőtt változat. A tesztek profiltípusú eredményt adnak: a Wechsler-teszt különböző változatainak legújabb kiadásai tucatnyi altesztet tartalmaznak, amelyek eredménye önállóan és négy részképességre (verbális megértés, munkamemória, perceptuális következtetés és információfeldolgozási sebesség) bontva is eredményt adnak az átfogó IQ-eredményen kívül (Wechsler, 2008). Korosztálytól függetlenül a teszt felvétele kétszemélyes helyzetben történik, képzett tesztfelvevőt igényel, és viszonylag hosszú időt vesz igénybe. Ezért ezeket a teszteket elsősorban a klinikumban, nevelési tanácsadókban és más, alapos, profilalapú egyéni értékelést igénylő helyzetekben használják.

Az intelligenciatesztek spektrumának másik végén a gyors, egyetlen – vagy nagyon kevés – típusú feladatból álló, csoportosan felvehető és átfogó IQ-eredményt adó tesztek állnak. Ezek közül a legismertebb a Raven Progresszív Mátrix (Raven, Raven és Court, 2003) amelynek Színes változata gyermekek, idősek és alacsony képességűek mérésére alkalmas. A Sztenderd változat fedi le a népesség nagy részét, a Sztenderd Plusz teszt az átlagnál magasabb, a Haladó változat pedig a legmagasabb képességtartományt célozza. A tesztben minden feladatnál nyolc bemutatott elem után kell megtalálni a megfelelő kilencediket. Az egyes feladatokban különböző szabályok érvényesülnek, amelyek alapján az utolsó elem kikövetkeztethető (lásd később a módszerek leírásánál).

Az intelligencia modelljei

Az intelligencia mérésének kezdetei óta nagyszámú különféle tesztet fejlesztettek, amelyek számos különböző képességet mérnek, a szókincstől a numerikus gondolkodáson és a perceptuális sebességen át a mentális forgatásig. Ezeket a teszteket pedig rengetegszer használták iskolai, munkahelyi vagy katonai alkalmassági vizsgálatra. Különböző tesztek felvételekor minden esetben azt találták, hogy azok pozitívan korrelálnak egymással: függetlenül attól, hogy milyen speciálisnak tűnő képességet mérnek, aki az egyik teszten jobban teljesít, az várhatóan az összes többin is. Ez a „pozitív sokféleségnek" nevezett jelenség az intelligencia területének minden bizonnyal legfontosabb eredménye, és a sok száz publikált vizsgálat alapján valószínűleg nem túlzás kijelenteni, hogy az egész pszichológiában ez a legtöbbször replikált empirikus jelenség.

A pozitív sokféleség leírására a faktoranalízist használták, egy olyan statisztikai eljárást, amely egy számos változóból álló korrelációs mátrixot néhány alapdimenzióra, latens változóra egyszerűsít, feltételezve, hogy a mért változók egymással való korrelációja a mért változóknak a latens változókkal való korrelációjával magyarázható. Vagyis egy szókincsteszt és egy mentális forgatási teszt eredménye azért korrelál egymással, mert mindkettő korrelál egy közvetlenül nem mért (tehát latens) változóval.

A faktoranalízis módszerének kifejlesztése Charles Spearman nevéhez köthető, aki a pozitív sokféleség magyarázatára kétféle faktort határozott meg: egy általános faktort (*g*-faktort) és speciális faktorokat (*s*-faktorok). Elképzelése szerint minden teszt ugyanazt az általános faktort méri, az egyes tesztek csak abban különböznek, hogy milyen mértékben mérik a *g*-t, és milyen mértékben a saját specifikus faktorukat (Spearman, 1904, 1927). Thurstone (1938) Spearman kortársa vitatta a *g*-faktor létezését, és helyette egy hét, egymástól független faktorból (csoportfaktorból) álló modellt javasolt, megkülönböztetve egyebek között a nyelvi megértést, a számolási képességet és az emlékezés képességét.

Idővel mindkét modell tarthatatlanná vált az empirikus eredmények fényében: a csoportfaktorok nem függetlenek egymástól, hanem korrelálnak, ugyanakkor egyetlen, általános faktor nem elegendő a teljes változatosság leírására, ugyanis bizonyos fajta tesztek (mint például a szókincs és az olvasási készséget mérő) jobban korrelálnak egymással, mint másokkal (például a téri forgatással). Így egy átfogó modellben mind az általános faktornak, mind a specifikus képességeket reprezentáló csoportfaktoroknak helye van. Ezt idővel Spearman és Thurstone – valamint a követőik – is elismerték, vita már csak az általános és a csoportfaktorok viszonylagos jelentőségéről folyt. A leginkább elfogadott statisztikai modell több száz korábbi adathalmaz összesített faktoranalíziséből származik, és a faktorok három szintjét különbözteti meg: az első szinten a szűkebb képességek állnak, a második szinten az átfogó képességek, a harmadik szinten pedig a *g* (Carroll, 1993).

A faktorok értelmezése alapján számos különböző modell született az „intelligencia szerkezetéről", vagyis valójában az egyéni különbségek dimenzióiról (összefoglalásért lásd Conway és Kovacs, 2013). Ezeket e helyütt nem tekintjük át, érdemes azonban kiemelni a fluid-kristályos (Gf-Gc) modellt (Cattell, 1971; Horn,

1994). A modell különbséget tesz két alapképesség, a fluid és a kristályos intelligencia között. A fluid intelligencia azt a képességet jelöli, amelyet olyan, újszerű problémákkal szembesülve használunk, amelyek megoldásához nem áll rendelkezésünkre korábban elsajátított készség vagy ismeret. Rendszerint nem-verbális, induktív gondolkodást igénylő feladatokkal mérik, mint amilyen a Raven Progresszív Mátrixok vagy a számsorozatok. A kristályos intelligencia ezzel szemben a már megszerzett tudás és készségek alkalmazásának képessége, a mérésére alkalmasak például a szókincstesztek. A korábban említett Wechsler-tesztek próbáinak többsége is a kristályos intelligenciát méri. A két fő képességen kívül számos további faktor is helyet kap a modellben, például a sebesség, az emlékezet faktora és így tovább.

A fluid és kristályos képesség elkülönítése már Spearmanig visszavezethető, aki megkülönböztette a *g* két aspektusát, az *eduktív* (logikai következtetés) és a *reproduktív* képességeket (információk előhívása). A fluid-kristályos modell erőssége, hogy a faktorok, pontosabban a faktorok által reprezentált képességek, megkülönböztetésének igazolása túlmutat a statisztikai leírások és modellek világán. Így például a fluid és a kristályos képességek eltérő öregedési mintázatot mutatnak, fiataloknál a fluid, idősebbeknél a kristályos képességek jobbak (Horn és Cattell, 1967), a fluid intelligenciára sokkal inkább hat az IQ generációk közti növekedése, a Flynn-hatás (Flynn, 2007), és a frontális lebeny sérülése is a fluid képességek romlását eredményezi a kristályos képességek érintetlensége mellett (Duncan, Burgess és Emslie, 1995; Woolgar és mtsai, 2010). A klasszikus neuropszichológiában éppen a kristályos tesztek használata miatt gondolták sokáig, hogy a frontális lebeny nem játszik szerepet az intelligenciában (Weinstein és Teuber, 1957).

A fluid-kristályos elmélet és a hierarchikus modell ötvözetéből született meg az CHC (Carroll–Horn–Cattell-) modell (McGrew, 2009), amely Carroll háromszintű modelljének második szintjére illeszti a fluid-kristályos modell faktorait, ugyanakkor – Cattell eredeti elképzelésével szemben – helyet hagy az általános faktornak is.

Bár az IQ mérésére számos, kiváló pszichometriai tulajdonságokkal rendelkező eszköz készült, az intelligencia meghatározását és természetét illetően nincs átfogó konszenzus. A sokféle definíciót (Lásd Gottfredson, 1997; Sternberg és Detterman, 1986) e helyütt nem tekintjük át, érdemes azonban kiemelni közülük a talán leghírhedtebbet: „intelligencia az, amit a tesztek mérnek" (Boring, 1923). Ezt a sokszor a meghatározhatatlanság jellemzésére használt, körkörösnek tűnő definíciót újabban valódi tartalommal ruházták fel. Több új intelligenciaelmélet (Kovacs és Conway, megjelenés alatt; van der Maas és mtsai, 2006) is anélkül magyarázza meg a pozitív sokféleséget, és ezáltal az általános faktort, hogy feltételezne egy olyan pszichológiai folyamatot vagy idegrendszeri mechanizmust, amely megfeleltethető a *g*-nek. Vagyis, bár a *g* matematikai értelemben szükségszerű következménye a pozitív sokféleségnek (Krijnen, 2004), az általános faktor mint statisztikai konstruktum (pszichometriai *g)* nem feleltethető meg egy általános kognitív képességnek (pszichológiai *g).*

Ezek az elméletek újraértelmezik az általános faktor, és ezáltal az IQ és az intelligencia fogalmát is, amennyiben a *g*-t formatív, nem pedig reflektív latens változóként írják le (Conway és Kovacs, 2015). A reflektív modellekben – mint a szten-

derd intelligenciamodellekben is – az okság iránya latens változótól a mért változó felé halad, vagyis az egyes mérőeszközök a latens változót mérik, a mérőeszközök közti korreláció pedig megmagyarázható a mért változóknak a latens változóval való korrelációjával. A formatív modellekben azonban az okság iránya fordított: a latens változó nem létezik a mérőeszköztől függetlenül, vagyis a mérőeszköz nem a latens változót tükrözi, hanem éppen ellenkezőleg, a latens változó egyfajta súlyozott eredménye a méréseknek – ilyen változók például a szocioökonómiai státusz vagy a versenyképességi index. Az új elméletek tehát elutasítják, hogy a tesztek egy általános kognitív képességet mérnének, a formatív módon értelmezett intelligencia fogalomra pedig szó szerint érvényes Boring definíciója: intelligencia az, amit a tesztek mérnek (Van der Maas, Kan és Borsboom, 2014).

KLASSZIKUS ÉS MODERN TESZTELMÉLET

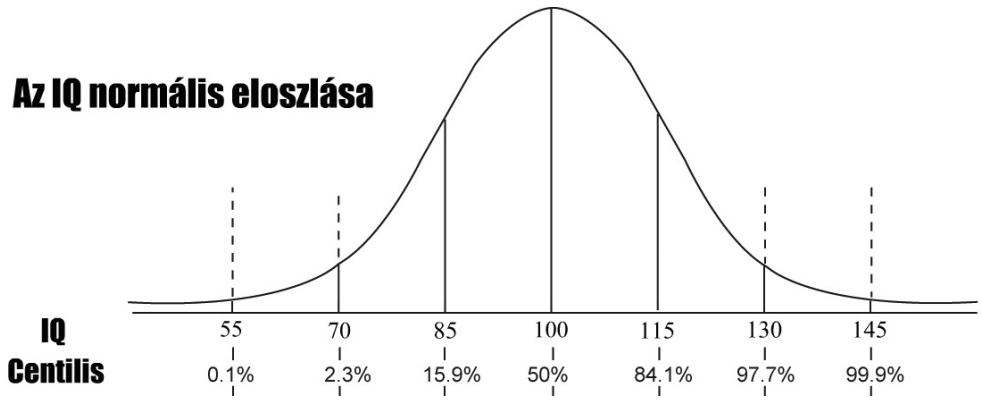
A klasszikus tesztelmélet alaptétele szerint a mért pontszám (X) minden esetben egy valós értékből (T) és egy hibaértékből (E) tevődik össze:

$$\mathbf{X = T + E}$$

A *T*-érték közvetlenül nem mérhető, elvi konstruktumként azt a pontszámot jelenti, amelyet akkor kapnánk, ha a vizsgálati személyünk végtelenszer kitöltené a tesztet úgy, hogy közben az előző kitöltések emléke valamilyen módon törlődne az emlékezetéből. A klasszikus tesztelmélet alapvető posztulátumokra épül. Az első és legfontosabb szerint az *E*, vagyis a hibaérték véletlenszerű és normális eloszlású, emiatt pedig a hibák középértéke nulla (emiatt van, hogy végtelen mérés esetén X=E). Egy másik posztulátum szerint a *T* és az *E* közti korreláció nulla, vagyis a hibaérték nagysága független a valódi érték nagyságától. Ilyen kikötések alapján levezethető a klasszikus tesztelmélet számos tétele a tesztek megbízhatóságáról, hibavarianciákról és így tovább (Horváth, 1993, 1997).

Annak érdekében, hogy sztenderd skálán mért értéket (Z, T, IQ, Sten, stb.) kapjunk a mért eredményből, a nyerspontszámot egy normacsoport korábban, a sztenderdizálás során mért eredményével hasonlítjuk össze. A sztenderdizálás során egy reprezentatív minta tölti ki a tesztet, majd kiszámításra kerül a pontszámaik átlaga és szórása, hogy később ezekhez tudjuk viszonyítani az egyedi eredményeket. Így a vizsgálati személy eredményét az átlaghoz képest el tudjuk helyezni, illetve a szórás alapján megmondhatjuk a percentilis értéket, vagyis azt, hogy a kitöltő a népesség hány százalékánál ért el jobb eredményt *(1. ábra)*.

A klasszikus tesztelméletben is vizsgálható a feladatok nehézsége, ezt a helyes válaszok aránya fejezi ki. A klasszikus tesztelmélet megfelelő elméleti keretet nyújt a pszichometriai méréshez, azonban számos hátulütője akad. Először is: noha a tesztek egésze rendelkezik megbízhatósági (reliabilitási) mutatóval, amely megadja a tesztek mérési pontosságát, az egyes teszteredmények pontossága nem megbecsülhető. Másrészt csak a teszt egésze képes mérni, hiszen az összpontszámot hasonlítjuk a normacsoporthoz, ennélfogva az egyes teszt-itemek nem felcserélhetők,



1. ábra. Az IQ normális eloszlása

a teszt pedig nem bővíthető, csak teljes újrasztenderdizálással együtt. Ez azt jelenti, hogy az egyes feladatok önmagukban semmilyen információt nem nyújtanak, és mindegyik ugyanannyit ér az összpontszám szempontjából. Harmadrészt, a teszt használhatósága nagyban függ a normához használt mintától, amelynek a normális eloszlás minden szegmensét arányosan le kell fednie.

Végül pedig, mivel a tesztek hosszúsága és így az itemek száma korlátozott, a klasszikus tesztelmélet alapján készült tesztek mérési tartománya szükségszerűen szűkös. Ez a gyakorlatban azt jelenti, hogy a legnagyobb hangsúly az átlagos nehézségű feladatokon lesz, kivéve azokat a teszteket, amelyek éppen a magas vagy alacsony képesség mérésére készültek, azonban a tartomány ebben az esetben is korlátozott.

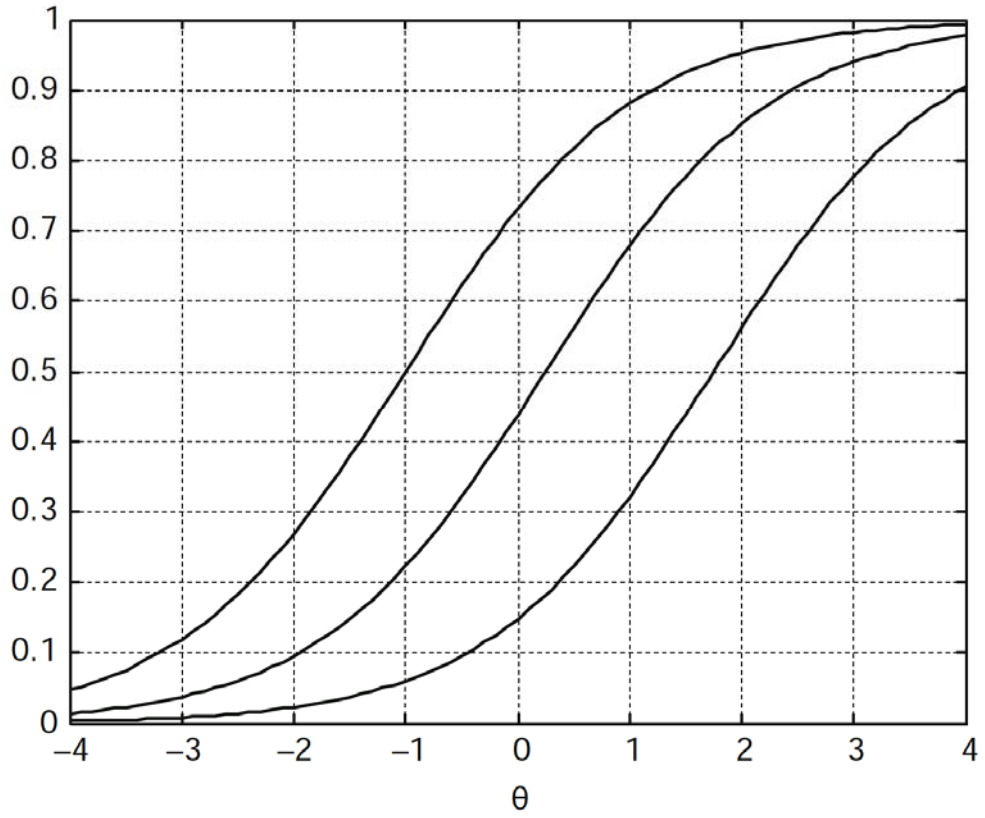
A klasszikust meghaladó modern tesztelmélet vagy item-válasz elmélet alapfeltételezése, hogy a jobb képességűek nagyobb valószínűséggel válaszolnak helyesen egy adott kérdésre, mint a rosszabb képességűek, függetlenül bármilyen más jellemzőjüktől. A modern tesztelmélet tehát probabilisztikus: az egyes feladatok saját item-paraméterekkel rendelkeznek, amelyek megjósolják, hogy a mérendő képesség egy adott szintjén mekkora valószínűséggel oldják meg az adott feladatot (Hambleton, Swaminathan és Rogers, 1991).

Így minden egyes item esetében külön meghatározható a helyes válasz valószínűsége a képességszint függvényében. A modell lehet egy- vagy többparaméteres. Az egyparaméteres modell (Rasch-modell) mindössze egy nehézségi paraméterből áll, ennek képlete:

$$P_i(\theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_1}}$$

Ahol *P* a helyes válasz valószínűségét jelenti egy adott *i* itemnél, *b* az item nehézségi paramétere, θ pedig a latens képesség szintje (a *b*-t és a θ-t *Z* értékben fejezik ki). A nehézség paramétere az az érték, ahol a helyes válasz valószínűsége 50%. Minél nagyobb a paraméter értéke, annál nehezebbnek bizonyul az item, minél

kisebb, annál valószínűbb, hogy helyes választ adnak rá, tehát annál könnyebbnek. Ugyanezt vizuálisan fejezi ki az itemjelleg-görbe. *(2. ábra)*.



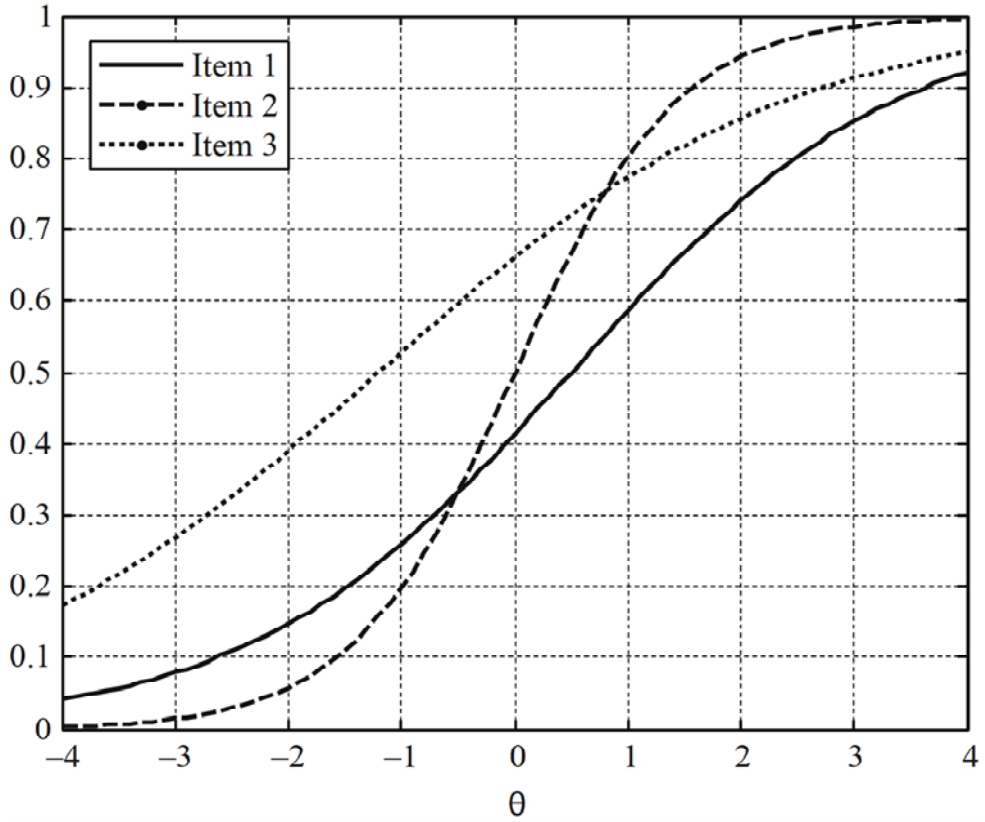
2. ábra. Egy példa az itemjelleg-görbére, amely három különböző nehézségű feladat (balról jobbra: b = –1, 0,25, 1,75) helyes megoldásának valószínűségét mutatja a képességszint (θ) függvényében *(Forrás:* Reckase, 2009, 20).

A kétparaméteres modell (Birnbaum, 1968) esetében az eddigiekhez hozzáadódik egy diszkriminációs paraméter, az *a:*

$$\mathbf{P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_1(\theta - b_1)}}}$$

Az *a* paraméter azt írja le, hogy az adott item mennyire differenciál, mennyivel nagyobb valószínűséggel válaszol helyesen egy magas képességszintű személy, mint egy alacsony képességszintű. Minél nagyobb a diszkriminációs paraméter, annál jobb a feladat, mivel a helyes vagy helytelen megoldás alapján annál ponto-

sabban becsülhető a képességszint. Vizuálisan ezt az itemjelleg-görbe meredeksége fejezi ki *(3. ábra)*.



3. ábra. Három különböző feladat itemjelleg-görbéje, a nehézségi és diszkriminációs paraméterek az egyes görbékre: folyamatos b=0.5, a=0.7, szaggatott: b=0, a=1,4, pontozott: b=–1,2, a=0,56 *(Forrás:* Reckase, 2009, 22).

Létezik háromparaméteres modell is, amely figyelembe veszi a találgatást, illetve négyparaméteres modell, amely e mellett még a figyelmetlenséget is. Ezek a modellek számolnak azzal, hogy a legnehezebb feladaton is van esélye egy alacsony képességű személynek a helyes válaszra, és a legkönnyebb feladatra is adhat helytelen választ egy magas képességű személy. Ennyiben tehát az egy- és kétparaméteres modell feltevései ezekhez a modellekhez képest nem tűnnek életszerűnek. Ugyanakkor a feladat paramétereinek kiszámításakor az egyes értékek annál pontosabbak lesznek, minél kevesebb paramétert kell megbecsülni az adatokból. Továbbá a legtöbb esetben a kétparaméteres modell pontosan működik az előfeltevések valószerűségétől függetlenül, ezért nem feltétlenül érdemes további paramétereket bevonni.

A modern tesztelmélet számos előnnyel rendelkezik a klasszikussal szemben. Egyrészt az egyes feladatokkal is lehet mérni, nem csak a teljes teszttel. Ebből következően nem minden válasz „ér ugyanannyit", szemben a klasszikus tesztekkel. Másodszor: a mérés pontossága minden egyes eredmény esetében megbecsülhető. Harmadszor: a paraméterek mintafüggetlenek, azok számítása több részminta segítségével is lehetséges, amelyek együtt sem kell, hogy lefedjék az összes képességtartományt. Végül: az item-válasz elmélet lehetővé teszi az adaptív tesztelést. A kétfajta megközelítés különbségeit az *1. táblázat* foglalja össze.

1. táblázat. A klasszikus és a modern (IRT) tesztelmélet összehasonlítása

	Klasszikus	IRT
Az értékelés alapja	Összpontszám (normacsoport összpontszám-eloszlásához viszonyítva)	Az egyes feladatok
Pontosság (hiba)	Csak a tesztnek van, az egyes eredményeknek nincs	Minden egyes eredménynek van
Itemek	Nem cserélhetők	Cserélhetők
Válaszok	Minden helyes válasz ugyanannyit ér	Az egyes itemekre adott válasz „értéke" az itemtől függ
Legfontosabb itemek	Átlagos nehézség	Bármilyen nehézség
Adaptivitás	Gyakorlatilag nem lehetséges	Lehetséges

SZÁMÍTÓGÉPES ADAPTÍV TESZTELÉS

A képességek mérésére új lehetőséget biztosít az informatikai fejlődés, amely lehetővé teszi a személyre szabott, adaptív tesztelést. Valójában a számítógépes adaptív tesztelés (Van der Linden és Glas, 2002; Weiner és Dorans, 2000) teljes egészében a modern tesztelméletre és a számítógépes környezetre épül. A számítógépes adaptív tesztelés (CAT) során, a papír-ceruza tesztekkel szemben nem egy kész feladatsort használnak, hanem egy item-bankot, amely akár több száz feladatból is állhat, és amelyből az adaptív algoritmus válogat. A teszt egyénekre szabása úgy valósul meg, hogy az algoritmus a kitöltő becsült képességszintje alapján választja ki az egyes itemeket, azok nehézsége alapján. Vagyis az adaptív algoritmus a következő kérdést mindig az előző kérdésekre adott válaszok alapján adja.

A folyamat a gyakorlatban azt jelenti, hogy elindul a teszt egy adott nehézségszinten, ami általában az átlagos képességnek felel meg. Az algoritmus kiválaszt egy feladatot az item-bankból, amelynek a nehézsége közel áll ehhez a szinthez. Regisztrálja, hogy a beérkező válasz helyes-e, majd ennek megfelelően ad egy likelihood-becslést a képességszintről, valamint megadja ennek a becslésnek a hibáját. A becsült képességszint alapján pedig az algoritmus kiválasztja a következő itemet.

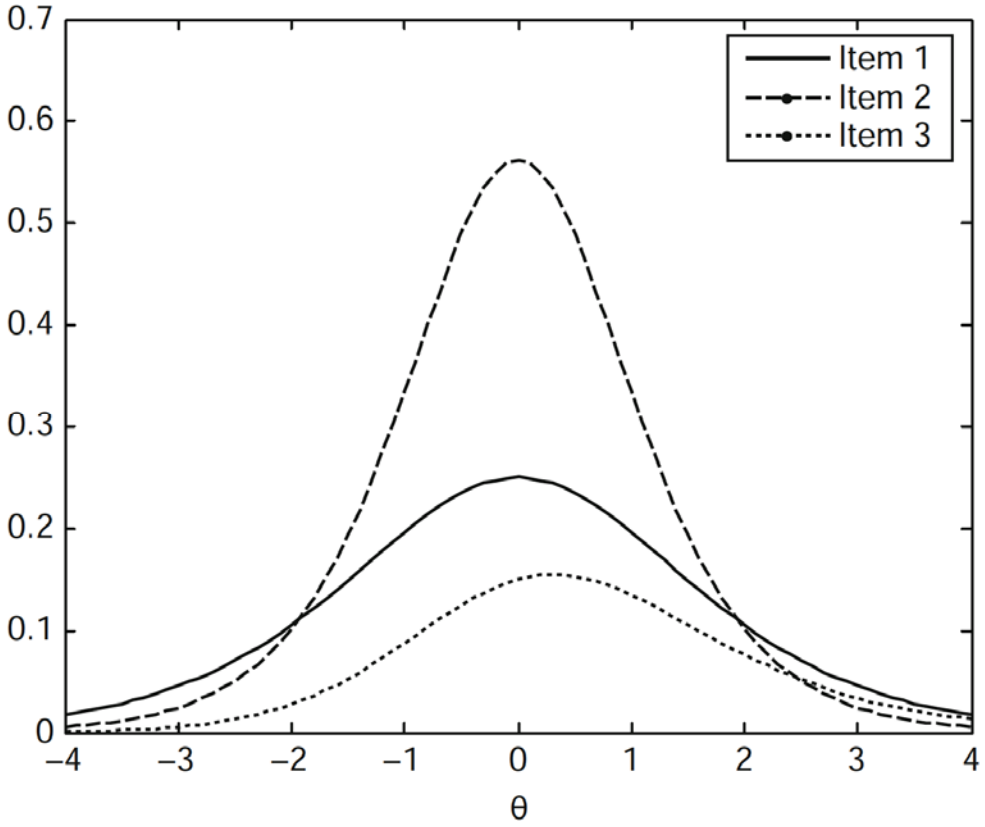
Mivel a modern tesztelmélet egyik alapfeltevése a „lokális függetlenség”, vagyis az, hogy az egyes feladatok helyes megoldásának valószínűsége kizárólag a képességszinttől függ, így több feladat helyes megoldásának valószínűsége egyenlő az egyes feladatok megoldási valószínűségeinek szorzatával. Az algoritmus minden egyes új item után újra megbecsüli, hogy az adott nehézségű feladatokra érkezett helyes, illetve helytelen válaszok együttes előfordulása mely képességszinten a legvalószínűbb.

Minél több itemre érkezett válasz, annál pontosabb lesz a képességszint becslése. Az item-információs funkció azt adja meg, hogy mennyi információra tehetünk szert az alapján, hogy valaki helyesen vagy helytelenül oldotta meg a feladatot. Ezen azt értjük, hogy mennyire pontosan következtethetünk a megoldás helyességéből a képességszintre: minél magasabb az adott képességszinten az item által nyújtott információ, annál nagyobb mértékben csökken a likelihood-becslés sztenderd hibája. A *4. ábra* mutatja az item-információs funkciót.

Minden feladat a nehézségparaméterének megfelelő képességszinten nyújtja a legtöbb információt, az egyes itemek tehát más-más képességszinten alkalmasak a mérésre. Ugyanakkor a két- vagy többparaméteres modell esetében a diszkrimináció is rendkívül fontos: minél nagyobb a diszkrimináció, annál több információ nyerhető ki egy feladatból. A teljes teszt esetében beszélhetünk teszt információs funkcióról is: a lokális függetlenség feltevéséből adódóan ez egyszerűen a tesztet alkotó feladatok által nyújtott információ összessége. A tesztinformációs funkció alapján megállapítható, hogy a teszt mely képességszinten milyen pontosan képes mérni *(12. ábra)*.

Az eljárás rendszerint addig tart, amíg a becsült érték hibatartománya le nem csökken egy előzetesen meghatározott szint alá, de megszabhatunk egy maximális kérdésszámot is, vagy egy képességszintet, amely alatt vagy felett a teszt már nem mér. Az adaptív tesztelés során tehát lényegében fordított valószínűséget számolnak az item-paraméterek alapján. Abból, hogy az egyes feladatokat különböző képességszinten milyen valószínűséggel oldják meg, megbecsülhető a legvalószínűbb képességszint, amely mellett a helyes és helytelen válaszok konkrét mintázata előfordulhatott.

A CAT a felhasználó szemszögéből tipikusan azt jelenti, hogy hibázás után könnyebb, helyes válasz után nehezebb feladat következik. A CAT lényege, hogy míg a hagyományos tesztben a kitöltő számára túl könnyű és túl nehéz feladatokkal is találkozik, az adaptív tesztelés alkalmazkodik a képességszintjéhez, így a tesztre szánt idő legnagyobb részében a saját képességszintjének megfelelő feladatokat kap. Ez egyrészt felgyorsítja a tesztelési folyamatot, másrészt növeli a mérés pontosságát, és megkímél a fölösleges frusztrációtól. A CAT további előnyei közé tartozik, hogy nincs kész – tehát ellopható és betanulható – megoldó kulcs, mert minden teszt „személyre szabott”, másrészt az item-bank cserélhető és bővíthető.



4. ábra. Három feladat item-információs funkciója
(Forrás: Reckase, 2009, 50).

Mindemellett meg kell említeni a módszer hátrányait is: a kifejlesztés folyamata bonyolultabb, költsége lényegesen nagyobb és a kérdéseket a legtöbb esetben a személy nem tudja ideiglenesen kihagyni és később visszatérni rájuk. Továbbá, mivel az adaptív teszt során mindenki a saját becsült képességszintjén kap feladatokat, a teszt szubjektív eredménye nem ad azonnali, reális képet a kitöltő képességszintjéről, mivel a papír-ceruza tesztekkel szemben egy alacsony és magas képességű személy is hasonló arányban érzi úgy, hogy helyesen oldotta meg a feladatokat – csak éppen teljesen különbözőeket. Összességében, megfelelő erőforrások és informatikai háttér megléte esetén az adaptív teszt előnyei túlszárnyalják a hátrányokat.

A MENSA HUNGARIQA ADAPTÍV PRÓBATESZTJE

A feladatok készítése

A Mensa HungarIQa (MH) adaptív próbatesztjének algoritmusát az első szerző (a teszt készítése idején a MH felügyelő pszichológusa) készítette, a feladatokat pedig a MH önkéntesei, többek között a második szerző, ezért az ő feladatait fogjuk használni illusztrációként a teszt készítésének bemutatásához.[1]

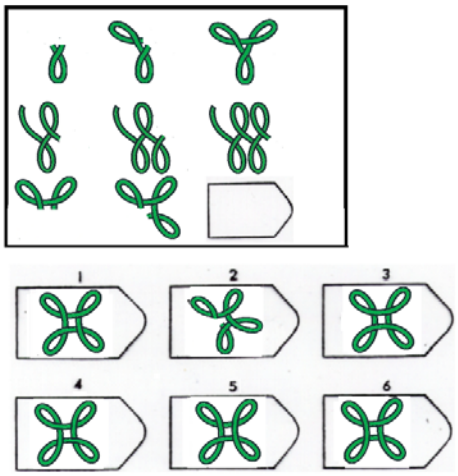
A teszt feladatai, a Raven Progresszív Mátrixokhoz hasonlóan, a fluid intelligenciát mérik, és nem-verbális, figuratív, induktív gondolkodást igénylő feladatokból állnak. Az egyes feladatok szintén mátrix formátumúak: egy 3×3-as elrendezésű elemeket tartalmazó ábrán kell megtalálni, hogy hat lehetőség közül melyik illik az utolsó, üresen maradt helyre (lásd az *5–8. ábrá*t).

Egy, a Haladó Progresszív Mátrixok teszt természetét vizsgáló kutatás során öt különböző szabályt határoztak meg, amelyekkel a Haladó teszt valamennyi feladata és a többi teszt feladatainak a többsége is megoldható: soron belüli állandóság, mennyiség változása, összeadás-kivonás, három érték kombinációja, két érték kombinációja (Carpenter, Just és Shell, 1990). A soron belüli állandóság azt jelenti, hogy az egyes sorokon belül valamely elem megmarad. Ez általában egy soron belül mindig ugyanott van, de olykor fordul. Ilyen logika alapján történik a mennyiség változása is. Ezen két szabály működése megfigyelhető az 1. feladaton *(5. ábra)*.

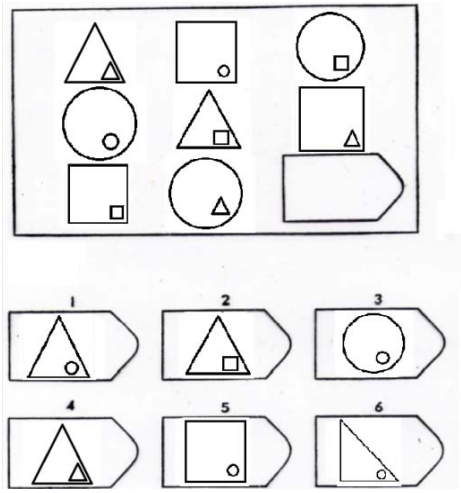
Az összeadás-kivonás esetében két alakzat elemei egymásra másolódnak vagy kivonódnak egymásból, így jön létre a harmadik. A három érték kombinációja szabály jellegzetesen a teljes mátrixra vonatkozik. Egy sorban egy változó három különböző értéke szerepel, a lényeg, hogy mindhárom. A megoldást mindig a hiányzó értékekből kapjuk meg. A szabály nemcsak mennyiségben, hanem bármilyen más tulajdonságban is megjelenhet (2. és 3. feladat, *6–7. ábra)*. A legnehezebb szabály a két érték kombinációja, ahol bizonyos elemek minden sorban és oszlopban pontosan kétszer fordulnak elő (4. feladat, *7. ábra)*. Ezeket a szabályokat egy feladatban egyszerre is lehet használni, és ugyanaz a szabály több különböző elemre is érvényes lehet. A feladatkészítők felhasználták a Raven-tesztek megoldásához szükséges szabályokat, ugyanakkor további szabályokat is alkottak.

A 1. feladat *(5. ábra)* egyszerre több szabályra épül. Soron belüli állandóság a csigavonal és annak kanyarodási iránya, ami a válaszok egy részét kizárja. A mennyiség változása a kunkorok számában és az adott sorban lévő csigavonalak számában jelenik meg. Így míg az első szabály kizárja az 1-es, 6-os, a második a 4-es és az 5-ös, addig a harmadik a 3-as lehetőséget, ezzel a 2-est hagyva jó válasznak. A feladat megoldásához az elsőre egyértelmű hurokszám-növekedésen kívül a helyes válasz megtalálásához figyelembe kell venni a hurkok szárainak átfedését, és hogy a sor előzetes hurkai milyen szögben kapcsolódnak egymáshoz.

[1] Az itt használt rajzok a második szerző eredeti munkái, a tesztbe ezeknek professzionálisan újrarajzolt változata került be.

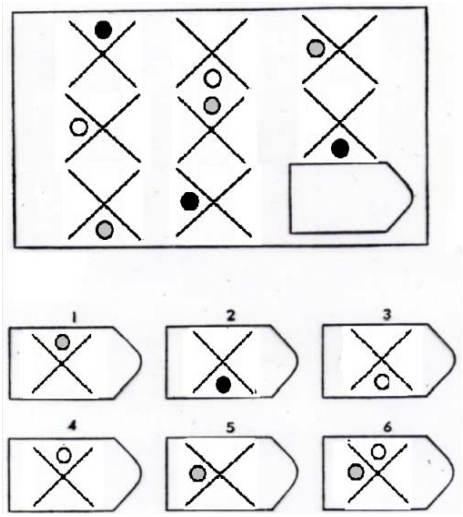


5. ábra. 1. feladat

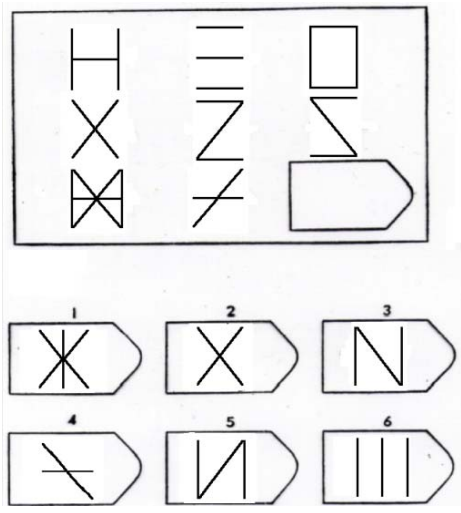


6. ábra. 2. feladat

A 2. feladat a három érték kombinációjára épít, minden sorban található ugyanis egy-egy nagy és egy-egy kis háromszög, kör és négyzet. Így ezek meglétét az utolsó sorban ellenőrizve megkapjuk az 1-est helyes válasznak. Az 3. feladat ugyanezt a szabályt használja fel, csak itt a változók a pötty színe és helyzete, így itt a 4-es válasz lesz a jó. A szabály mindkét esetben vizsgálható az oszlopokban és a sorokban is. A 4. feladatban a két érték kombinációja szabályt kell csak alkalmazni, és az ott megjelölt 3-as válasz a jó megoldás.



7. ábra. 3. feladat



8. ábra. 4. feladat

A teszt kialakítása

A feladatokat az első szerző válogatta, az első rostán kiestek azok a feladatok, amelyekhez nagyon apró részleteket kellett felfedezni; amelyek túlzottan bonyolult szabályokra épültek; amelyekhez nagyon egyszerű, alapműveleteket igénylő számolásoknál bonyolultabb számításra volt szükség; valamint amelyekben téri műveletekre (bonyolultabb forgatások és tükrözések) volt szükség. Így például kiesett a fentebb bemutatott 1. feladat is *(5. ábra)*. A válogatás végén 161 feladat maradt.

A vizsgálatban a Budai Középiskola és a Giorgio Perlasca Vendéglátóipari Szakközépiskola és Szakiskola 11–12. osztályos (17-18 éves) tanulói vettek részt, a helyszínt is az iskolák biztosították. A résztvevők először egy 45 perces tanóra alatt írták meg a Raven-féle Haladó Progresszív Mátrixot, 40 perces időkorláttal. A második szakaszban, körülbelül két héttel később a diákok számítógéptermekben megválaszolták az új feladatokat, erre 3×45 perc állt a rendelkezésükre. Adatvédelmi okokból nem rögzítettünk a pszichometriai számítások szempontjából felesleges adatokat (a diákok a Raven-teszt kitöltésekor kapott kóddal léptek az új feladatok megoldásához használt felületre), így a minta nemi eloszlása és pontos életkori szórása nem ismert. A diákok a feladatokat felügyelet alatt töltötték ki, a felügyeletet a Mensa HungarIQa Egyesület tagjai biztosították.

A vizsgálatban összesen 547 diák vett részt, azonban hiányzások miatt 17 diák nem töltött ki Raven-tesztet, 33 diák pedig az új feladatok megoldásában nem vett részt. A maradék 497 diákból 9-et kiszűrtünk, mert a Raven-teszt megoldása során véletlenszerűen vagy egyáltalán nem válaszoltak (2 diák üres válaszlapot adott be, 1 diák 3 perc után adta be a válaszlapot, 6 diák pedig véletlenszerűen válaszolt: 8 vagy kevesebb helyes válasza volt, a helyes válaszok nehézség szerinti eloszlása pedig nyilvánvalóvá tette a tippelést).

488 diák töltötte ki mind a Raven Haladó Mátrixokat, mind az új feladatokat. A diákok a feladatokat random sorrendben kapták. A rendelkezésre álló idő alatt a feladatoknak átlagosan nagyjából 80%-ára tudtak választ adni, így az egyes feladatokra a teljes mintának nagyjából ekkora arányában érkeztek válaszok. A véletlenszerű válaszadás, vagyis a random tippelő diákok a számítógépes felvétel során is problémát jelentettek, ennek megoldásához a reakcióidőket használtuk, amelyeket a szoftver rögzített. A reakcióidő-eloszlások elemzésekor látszott, hogy kiugróan nagy számban érkeztek néhány másodperc alatt válaszok, míg a 4-5 másodperces határ felett a válaszadás idejének eloszlása szabályossá vált. Itt tehát nem válaszadókat szűrtünk, hanem minden egyes olyan választ, amely 4 másodpercnél rövidebb idő alatt érkezett, hiányzóra cserétünk; ez átlagosan a válaszok 10%-ának kiesését jelentette.

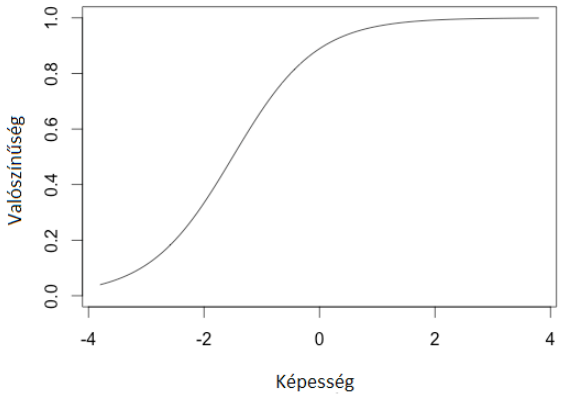
Az itemek elemzéséhez az R statisztikai programot használtuk, az LTM (Latent Trait Models) kiegészítő csomag segítségével. Figyelembe vettük a vizsgált minta (pontosabban a Raven-tesztet és az új feladatokat egyaránt megoldó 488 fő) IQ-átlagát, amely a Raven-teszt eredménye szerint 103,5 volt. Ez 0,23 szórással haladja meg a népesség átlagát, a feladatok nehézségparamétereit ezért ennyivel korrigáltuk, hogy az új teszt valóban a teljes népességhez viszonyítson. A számítás során a kétparaméteres modellt alkalmaztuk, így kiszámításra került az itemek nehézsége és diszkriminációja (a fentebb szereplő feladatok paraméterei a *2. táblázat*ban, az itemjelleg-görbék a *9–11. ábrá*n láthatók).

Az item-bankba 80 feladat került, a tesztinformációs funkciót a *12. ábra* mutatja. Az ábrán látható, hogy a teszt a legtöbb információt nagyjából a –2 és +1,5-ös Z értékek között adja, vagyis az átlagtól lefelé 2 és felfelé 1,5 szórás közti tartományon kívül nem mér pontosan. Ezért a 2. centilis alatt és a 95. centilis felett lévő eredményeket a teszt nem különbözteti meg, ami a gyakorlatban azt jelenti, hogy ezekben a tartományokban nem ad pontos értéket, az IQ-ban megadott eredmény itt „70 alatti” és „125 feletti”.

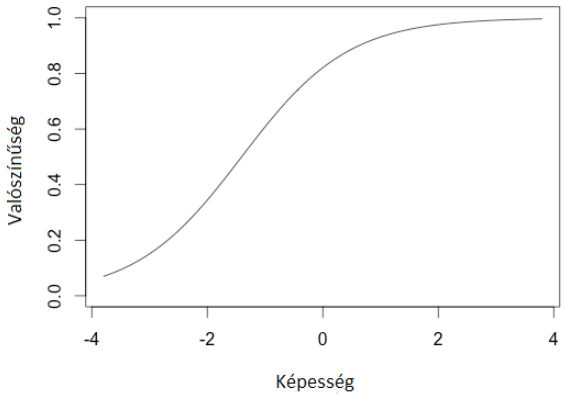
2. táblázat. A feladatok paraméterei

Feladat	Nehézség	Diszkrimináció
2. feladat	–1,272	1,386
3. feladat	–1,174	1,083
4. feladat	0,926	0,834

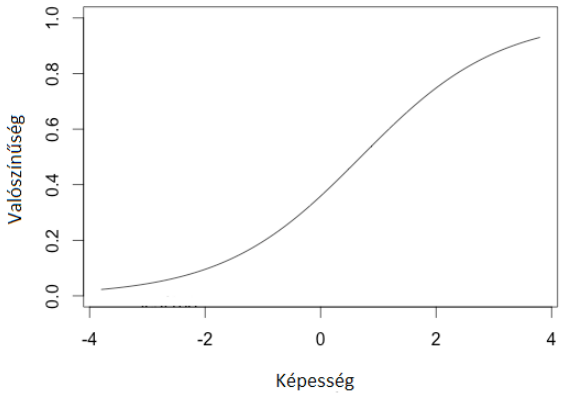
(A számozás a fenti ábrákét követi, az 1. feladat nem került be a felvettek közé.)



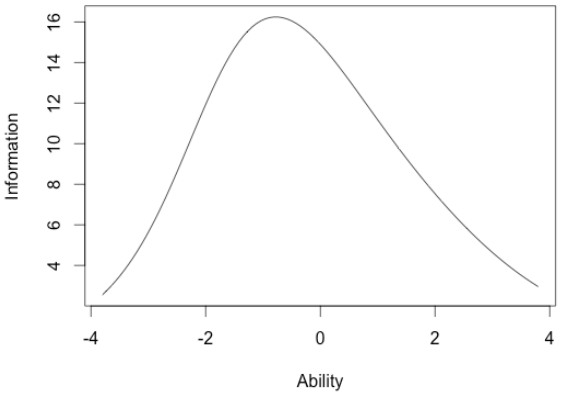
9. ábra. A 2. feladat item-karakterisztikus görbéje



10. ábra. A 3. feladat item-karakterisztikus görbéje



11. ábra. A 4. feladat item-karakterisztikus görbéje



12. ábra. A teszt információs funkciója

A teszt algoritmusa a „Concerto" szoftver 3.8.3-as változatán fut. Az algoritmus az első feladatot véletlenszerűen választja 18, átlag körüli nehézségű feladatból. Ezt követően a Maximum Fisher Information (MFI) módszert használja: a személy válaszát követő képességbecslést követően az algoritmus azt az itemet választja, amely a becsült képességszinten a lehető legtöbb információt nyújtja. Egy másik népszerű eljárás az Urry-módszer, amelynek során a becsült képességszinthez legközelebbi nehézségű feladat kerül kiválasztásra. Egyparaméteres modell esetén az Urry-módszer és az MFI azonos, kétparaméteres modell esetében elképzelhető, hogy az MFI alapján az algoritmus olyan feladatot választ, amely nem a legközelebb áll a nehézségét tekintve a becsült képességszinthez, azonban a legközelebbi feladat a rosszabb diszkriminációs paramétere miatt kevesebb információt nyújt, vagyis az arra adott válasz kevésbé csökkenti a mérési hibát az adott képességszinten, mint a nehézségben távolabb álló.

A teszt háromféle leállási kritériummal működik: befejeződik az algoritmus és a legutolsó becslés lesz a végeredmény akkor, ha *1*. a személy megoldott 25 feladatot, *2*. az utolsó becslése 2,5 szórással az átlag feletti (ez esetben az eredmény „125 feletti" az ebben a tartományban meglévő magas hibaérték miatt), *3*. ha a hiba (SEM) 0,35 alá csökken.

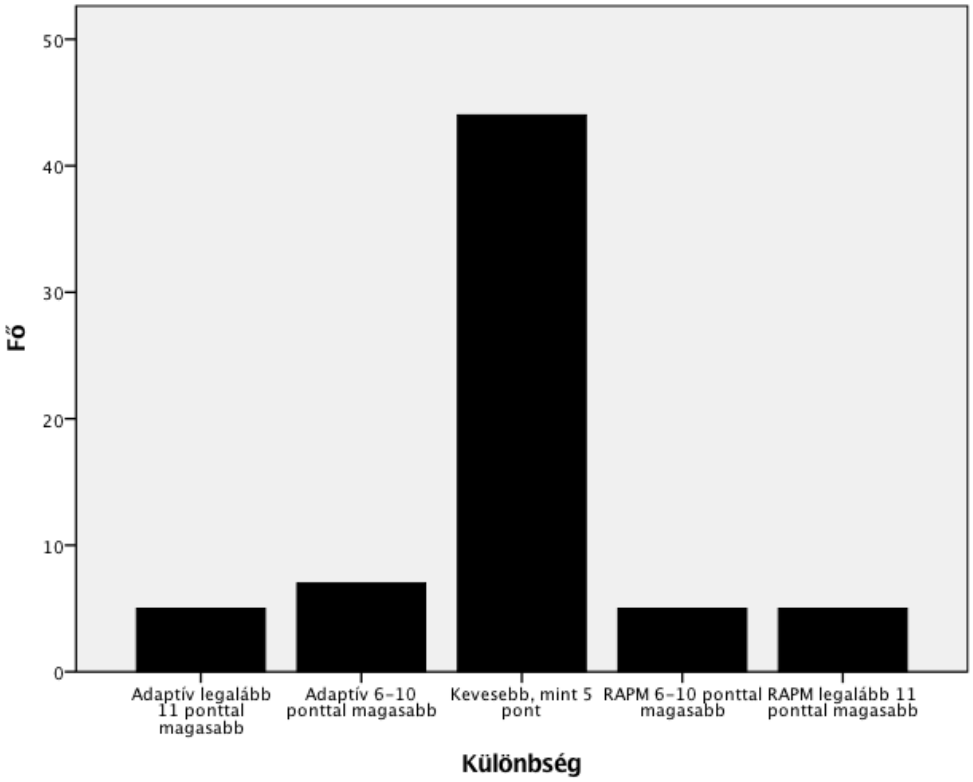
VALIDÁLÁS

Az elkészült adaptív teszt érvényességének alátámasztására konkurens validitási vizsgálatot végeztünk (Rust és Golombok, 1999). Mivel az új teszt ugyanazt a konstruktumot – a fluid intelligenciát – célozza mérni, mint a sokszorosan validált Raven Progresszív Mátrixok (Raven és mtsai, 2003), ezért ez utóbbi tesztet használtuk. A MH képviseltette magát a 2013-as Sziget Fesztiválon, az érdeklődők napközben egy „Mensa-sátorban" írhattak tesztet, és ismerkedhettek a szervezettel. 66 fő vállalkozott arra, hogy mind a Raven Haladó Progresszív Mátrix (RAPM) tesztet kitölti, mind pedig – tabletek segítségével – az új adaptív tesztet. (A pontosabb számítás érdekében a RAPM tartományokban kifejezett IQ-eredményt a nyerspontok alapján lineárisan módosítottuk a vizsgálathoz, és kiterjesztettük az átlag alatti tartományra is.) Véletlenszerűen dőlt el, hogy az egyes résztvevők melyik tesztet töltik ki először.

A RAPM nyerspontszáma és az adaptív IQ közti korreláció 0,803 ($p < 0{,}001$), a RAPM alapján számolt IQ és az adaptív IQ közti korreláció pedig 0,743 ($p < 0{,}001$). Ezek az értékek közel állnak a RAPM reliabilitási indexéhez, vagyis szinte ugyanazt méri a két teszt, csak az előbbi papír-ceruza módszerrel, az utóbbi pedig számítógépes, adaptív eljárással. A konstruktumvaliditáshoz használt korrelációkon kívül a tényleges IQ-eredmények is erősítik a két teszt egyezését: a minta átlaga az RAPM teszttel mérve 113 pont, az adaptív teszttel 110,8 pont (a medián mindkét esetben 113 pont).

A *13. ábra* mutatja a két teszt eredményének egybeesését. Látható, hogy az esetek többségében a két teszt 5 ponton belül azonos eredményt ad, továbbá a tesztek eredménye közti eltérések szimmetrikusak, vagyis nagyjából egyforma valószínű-

séggel értek el jobb eredményt a két teszt bármelyikén. Ez pedig arra utal, hogy a két teszt eredménye közti különbség mérési hiba, ami szintén megerősíti, hogy a két teszt nemcsak ugyanazt a konstruktumot méri, hanem szinte azonos módon is.



13. ábra. Az adaptív teszt és a RAPM által mért IQ megfeleltetése

Mivel a RAPM a felsőbb tartományok mérésére készült, és sztenderd adminisztrálás esetén nem differenciál 100-as IQ-érték alatt, ezért ellenőrzésképpen elvégeztük a fenti eljárást a kizárólag 100 és 125 közti eredmények figyelembevételével is. 47 fő ért el ilyen eredményt, az ő esetükben a minta átlaga az RAPM teszttel mérve 114 pont (medián: 115 pont), az adaptív teszttel 114,6 pont (medián: 116 pont).

A TESZT FELHASZNÁLÁSA

A teszt megtalálható a https://mensa.hu/tesztiras/online-iq-probateszt címen, 2015. december 22-ig 161 851 kitöltője volt (ugyanakkor egy személy többször is kitölt-

hette). A nagyszámú keresés és kitöltés eredményeként a teszt ugyanezen a napon a legelső nem fizetett Google-találat az „iq teszt" és „iq-teszt" keresésekre.

A tesztet felhasználták, illetve jelen cikk elkészültét követően is felhasználják a Magyar Templeton Program nevű tehetséggondozási program[2] kiválasztási szakaszában, három másik kognitív mérőeszközzel (szókincsteszt, divergens gondolkodási teszt és N-vissza) együtt. A program keretében közel 60 feladattal bővült az item-bank, és a cikk készültekor további 33 feladat bemérése zajlott. Az új feladatok a program kiválasztási szakaszának lezárultát követően a MH fenti linken elérhető próbatesztjének részévé válnak.

Bár egy adaptív teszt kifejlesztése erőforrás-igényesebb vállalkozás, mint egy hagyományos papír-ceruza teszté, az IKT alkalmazása a tesztek világában számos haszonnal jár. A már ismertetett előnyök (gyorsabb és pontosabb mérés; megoldó kulcs nélküli, tehát biztonságosabb teszt; egyedi hibaértékek stb.) mellett érdemes kiemelni a nagy mennyiségű online adat elérhetőségét és kutatási hasznosságát.

A világ első intelligenciamérése során Galton 1884-ben, a londoni South Kensington Múzeumban, a nemzetközi egészségügyi világkiállítás részeként állította fel „Antropometriai Laboratóriumát", ahol az érzékelési vizsgálatokat 3 pennyért végezhették el az érdeklődők. A kiállítás végére több mint 9000 személyről sikerült adatot gyűjteni. A MH adaptív IQ-tesztje iránti hatalmas érdeklődés azt mutatja, hogy a pszichológiai önismeret iránti igény az eltelt 130 évben szemernyit sem csökkent. Erre az igényre építve, a számítógépes tesztelés segítségével ma Galton szemléletét alkalmazva, de fizikai laboratórium nélkül is gyűjthetünk nagy mennyiségű adatot.

Köszönetnyilvánítás

Köszönetet mondunk Nagy Lantos Balázsnak, aki a Mensa HungarIQa (MH) szervezet elnöke volt a teszt készítésekor, és az önkéntes tagoknak, akik lelkesen segítették a munka különböző szakaszait. Hálásak vagyunk az OTP Fáy András Alapítványának a teszt készítéséhez nyújtott támogatásáért. Ferencz Endre és Bognár Péter az informatikai munkákat, Istvánfy Gergely az online teszt grafikai munkáit végezték. Végül kiemelt köszönet illeti Nádasi Tibort, aki a teszt készítésekor a MH tesztgondnoka volt, és aki fáradhatatlanul dolgozott a vizsgálat megszervezésén és lebonyolításán.

IRODALOM

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 395–479). Reading, MA: Addison-Wesley.
Retrieved from http://ci.nii.ac.jp/naid/10011856529/en/

Boring, E. G. (1923). Intelligence as the Tests Test It. *New Republic, 36,* 35–37.

[2] http://templetonprogram.hu/

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*(3), 404–431.
Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2381998

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge: Cambridge University Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=i3vDCXkXRGkC&pgis=1

Cattell, R. B. (1971). *Abilities: Their Structure, Growth, and Action*. Houghton Mifflin .
Retrieved from http://books.google.com/books?id=10EgAQAAIAAJ&pgis=1

Conway, A. R. A., & Kovacs, K. (2013). Individual Differences in Intelligence and Working Memory. In *Psychology of Learning and Motivation* (Vol. 58, pp. 233–270).

Conway, A. R. A., & Kovacs, K. (2015). New and emerging models of human intelligence. *Wiley Interdisciplinary Reviews. Cognitive Science, 6*(5), 419–26.

Duncan, J., Burgess, P., & Emslie, H. (1995). Fluid intelligence after frontal lobe lesions. *Neuropsychologia, 33*(3), 261–268.

Fancher, R. E. (1985). *The intelligence men: Makers of the IQ controversy*. New York: W. W. Norton.

Flynn, J. R. (2007). *What Is Intelligence?: Beyond the Flynn Effect*. Cambridge: Cambridge University Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=qvBipuypYUkC&pgis=1

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*(1), 13–23.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. London: Sage Publications.

Horn, J. L. (1994). Theory of fluid and crystallized intelligence. In R. Sternberg (Ed.), *Encyclopedia of Human Intelligence* (pp. 443–451). New York: MacMillan Reference Library.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26,* 107–129.

Horváth G. (1993). *Bevezetés a tesztelméletbe*. Budapest: Keraban Könyvkiadó.

Horváth G. (1997). *A modern tesztmodellek alkalmazása*. Budapest: Akadémiai Kiadó.

Krijnen, W. P. (2004). Positive loadings and factor correlations from positive covariance matrices. *Psychometrika, 69*(4), 655–660.

Kovacs, K., & Conway, A. R. A. (in press). Process Overlap Theory: A unified account of the general factor of intelligence. *Psychological Inquiry*,
DOI: 10.1080/1047840X.2016.1153946

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10.

Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Harcourt Assessment.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer-Verlag.

Rust, J., & Golombok, S. (1999). *Modern psychometrics. The science of psychological assessment*. New York: Routledge.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology, 15*(2), 201–292.

Spearman, C. (1927). *The Abilities of Man: Their Nature and Measurement*. New York: Macmillan. Retrieved from http://books.google.com/books?id=ws78nQEACAAJ&pgis=1

Sternberg, R. J., & Detterman, D. K. (Eds.) (1986). *What is Intelligence?: Contemporary Viewpoints on Its Nature and Definition*. Ablex Publishing Corporation.

Terman, L. M. (1916). *The measurement of intelligence*. Boston: Houghton Mifflin Company.

Thurstone, L. L. (1938). *Primary Mental Abilities*. Chicago: University of Chicago Press.

Van der Linden, W. J., & Glas, G. A. W. (2002). *Computerized Adaptive Testing: Theory and Practice*. New York: Kluwer Academic Publishers.

Van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review, 113*(4), 842–861.

Van der Maas, H. L. J., Kan, K.-J., & Borsboom, D. (2014). Intelligence Is What the Intelligence Test Measures. Seriously. *Journal of Intelligence, 2*(1), 12–15.

Wechsler, D. (2008). *Wechsler adult intelligence scale – Fourth Edition (WAIS–IV)*. San Antonio, TX: NCS Pearson.

Weiner, H., & Dorans, N. J. (2000). *Computerized Adaptive Testing: A Primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Weinstein, S., & Teuber, H.-L. (1957). Effects of penetrating brain injury on intelligence test scores. *Science, 125*, 1036–1037.

Woolgar, A., Parr, A., Cusack, R., Thompson, R., Nimmo-Smith, I., Torralva, T., … Duncan, J. (2010). Fluid intelligence loss linked to restricted regions of damage within frontal and parietal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 107*(33), 14899–902.

Yerkes, R. M. (1921). *Psychological Examining in the United States Army*. Washington, DC: US Government Printing Office.

COMPUTERIZED ADAPTIVE IQ-TESTING

KOVÁCS, KRISTÓF – TEMESVÁRI, ESZTER

Computerized adaptive testing (CAT) is the most up-to-date method for the measurement of cognitive abilities. CAT relies on modern test theory on the one hand and on information technology on the other. We present the theory and practice of CAT through a practical example, the adaptive IQ-test of Mensa HungarIQa. The paper surveys the most important results of intelligence research as well as the history and methodology of IQ-testing first. This is followed by a discussion of the main areas of psychometrics; classical test theory and item response theory, as well the basics of CAT. Finally the development of an adaptive IQ-test is presented, from the creation of items through the estimation of item parameters and the construction of the adaptive algorithm to the validation of the test.

Key words: *IQ, intelligence, computerized adaptive testing*