A STANAG nyelvi tesztelésben alkalmazott alapvető statisztikai módszerek

Bevezetés

A nyelvi tesztelés szubjektív dolog. Ez elkerülhetetlen, mivel emberek végzik. Ugyanakkor nem engedhető meg, hogy az emberi tényező az elfogadható határokon túl befolyásolja az eredményeket. Ennek érdekében ma már széles körben elfogadott eljárásokat, köztük statisztikai módszereket is alkalmaznak.

Mielőtt ezek részletes tárgyalásába belekezdenénk, fontos, hogy felvázoljuk a statisztikai módszerek helyét és szerepét a tesztek elkészítésében és lebonyolításában. Nem hitelesített teszt ma már nem számíthat arra, hogy komolyan vegyék és elismerjék. A hitelesítés folyamata azon eljárások összessége, melyek célja annak biztosítása és bizonyítása, hogy a teszt valóban azt (és csak azt) a készséget méri, amelynek mérésére szánták. Ennek alapfeltétele az, hogy a tesztet megfelelően kvalifikált emberek készítik. A megfelelő képzés amellett, hogy növeli a teszt minőségébe vetett bizalmat, biztosítja azt is, hogy a tesztelők megfelelő mélységben ismerik a teszthitelesítési eljárásokat, köztük a statisztikai módszereket is. Ezek kétféle módon játszanak szerepet a STANAG tesztek esetében: a többé-kevésbé objektíven mérhető készségek (beszédértés és olvasás) esetében elsősorban a kérdések[1] nehézségi fokának beállítására, illetve ellenőrzésére szolgálnak, míg a szubjektívabb módon mérhető készségek (beszéd és írás) esetében e szubjektív hatás minimalizálására szolgálnak.

[1] A *kérdés* itt a lehető legtágabb értelmezést kapja, tehát egy igaz/hamis állítást éppen úgy kérdésnek nevezünk, mint egy feleletválasztós teszt elemeit. A tesztelési nyelvben kérdés (angolul: *item*) az, amelyre a pontozás tovább már nem bontható vagy bontandó egységei adhatók. Általában véve: egy kérdés, egy pont.

Klasszikus módszerek

Beszédértés és olvasás

Mint már említettük, a beszédértés és az olvasás viszonylag objektívebb eljárással mérhető. Ez azt jelenti, hogy egy adott kérdésekre adott válasz a javító személy véleményétől függetlenül helyes, vagy helytelen, azaz objektívan pontozható. Míg azonban idáig eljutunk, több statisztikai adatot kell elemezni.

A kérdéseket rendszerint adott szint mérésére írják. A közhiedelemmel ellentétben azonban erre a feladatra hosszú tanári pályafutás sem tesz mindenkit alkalmassá. Sőt: még a szakképzetség sem. Ezért a kérdéseket ki kell próbál(tat)ni, az eredményeket pedig elemezni.

Az egyes kérdésekkel kapcsolatos két legfontosabb jellemző az úgynevezett diszkriminációs index[2] és a nehézségi érték[3]. Az előbbi azt mutatja meg, milyen mértékben különbözteti meg a kérdés a jobb képességű vizsgázókat a gyengébb képességűektől. Kiszámításához hagyományosan a kérdésre adott helyes válaszok számát hasonlítják össze a vizsgázói tartomány legjobb, illetve a leggyengébb harmadában. Minél közelebb van a kapott érték a +1-hez, annál jobb a kérdés. A 0 körüli értékek gyenge megkülönböztetést mutatnak, a negatív értékek pedig azt jelzik, hogy valami baj van a kérdéssel, hiszen a gyengébben szereplő vizsgázók közül többen válaszoltak rá helyesen, mint a jobban szereplők közül.

A tesztek hitelességének egyik nélkülözhetetlen alkotóeleme a megbízhatóság. Ez minden érésre igaz: amennyiben az adott körülmények között elvégzett mérések egy bizonyos hibahatáron kívül esnek, a mérés nem hiteles. A tesztek esetében több megbízhatósági formula is létezik (pl. Kronbach Alfa, KR20, stb.), melyek értékeit a széles körben elterjedt szoftverek automatikusan számolják és közlik. Ezek közül néhány (pl. SPSS) nemcsak az egész kérdéscsoport mint teszt összesített megbízhatóságára ad értéket, hanem igény esetén

[2] Az angolszász szakirodalomban: Discrimination Index (DI)
[3] Az angolszász szakirodalomban: Facility Value (FV)

arra is választ ad, hogyan befolyásolja a megbízhatóságot egy adott kérdés elhagyása, így nyújtva segítséget az optimális kérdésszám és -kombináció kialakításához.

A fentiek azonban egy nagy közös hátránnyal rendelkeznek: az eredmények mindig relatívak, azaz csak a vizsgált populáción értelmezhetők. Amennyiben ez a populáció reprezentatív, az eredmények elég jól általánosíthatók – ez azonban általában nehezen kivitelezhető. Hogy csak egy problémát említsünk: statisztikailag megfelelő számú résztvevő nagyban veszélyezteti a kérdések hitelességét, mivel egész egyszerűen túl sokan ismerik meg a kérdéseket azok közül, akik a későbbiekben esetleg éles helyzetben is találkozhatnak velük. Ezért e módszerek kiegészítésre sorulnak.

Beszéd és Írás

Az beszéd és az írás (az úgynevezett produktív készségek) általában nem értékelhetőek objektíven. Az értékelés mindig egy értékelő (beszéd) vagy javító (írás) személyes véleményén alapszik. E vélemény objektivizálásának legalapvetőbb tényezője a megfelelő végzettség. Az ember azonban a legjobb képzettséggel is hibázhat – ám ez nem mehet a pártatlanság rovására. A produktív készségek értékelése napszaktól, fáradtságtól – általában minden külső körülménytől függetlenül állandó magas színvonalon kell, hogy folyjék. Ennek két komponense van. Egyfelől biztosítani kell, hogy az értékelést végző minden tesztelő ugyanúgy értelmezi és alkalmazza az értékelés kritériumait. Ezt mintaanyagok értékeléseinek összehasonlításával lehet ellenőrizni: erre alkalmas a különböző személyek által adott értékelések korrelációja – például a sorrendiség vizsgálatán keresztül. A cél a +1-hez minél közelebbi értékek elérése. Senkinek nem múlhat azon a vizsgája sikere, kinél szóbelizett, vagy ki javította az írásbelijét. Másfelől minden értékelőnek folyamatosan hoznia kell egy adott pontosságot. Ennek ellenőrzésére adott minták ugyanazon tesztelő általi többszöri értékelésének korrelációja alkalmas.

Amikor a fenti vizsgálatok megfelelő eredményt hoznak – akkor sem dőlhetünk hátra. Előfordulhat ugyanis, hogy bizonyos értékelők ugyanolyan sorrendet állítanak fel a vizsgázók között, mint a többiek, de *következetesen* magasabb, vagy alacsonyabb pontszámmal. Erre a jelenégre a fenti módszerek nem világítanak rá – ezért van szükség újabb módszerek alkalmazására is (csakúgy, mint fentebb), mint például a Kérdés Válasz Elmélet[4] (IRT) eljárásai.

IRT

Az IRT mind a négy a STANAG vizsgákon mért készség esetében jól egészíti ki a klasszikus módszereket. Minden alkalmazásának közös vonása, hogy valószínűségekkel számol. Nézzük meg, mit jelent ez konkrétan a négy nyelvi készség esetében.

Beszédértés és olvasás

Bármilyen jól megírt tesztet is veszünk alapul, egy bizonyos mérési hibával számolnunk kell. Azonban, mivel a vizsgázók emberek, várhatóan az általuk éppen produkált teljesítmény is el fog térni attól, amire amúgy teljesen ideális körülmények között képesek. Ennek az a következménye, hogy az elért eredmények mind a vizsgázók, mind a kérdések szempontjából árnyaltabb képet adnak, mint amit a hogy a klasszikus módszerek eredményeivel leírhatunk. Ez azt jelenti, hogy lesznek olyan kérdések, amelyek összességében könnyebbek (nehezebbek) más kérdéseknél, mégis viszonylag kevesebben (többen) adtak rájuk helyes választ a jobb (gyengébb) összesített eredményt elért vizsgázók közül. Így két azonos pontszámot elért vizsgázó eredménye tulajdonképpen összehasonlíthatatlan. Nehéz ugyanis elhinni, hogy az, aki mondjuk 20 összességében könnyebb kérdésre adott jó választ, az ugyanolyan szinten van, mint az, aki 20 összességében

[4]Az angolszász szakirodalomban: Item Response Theory (IRT)

nehezebb kérdést válaszolt meg helyesen. (A kérdést tovább bonyolítja az elrontott kérdések esetleges vizsgálata.) Ezen segít az IRT. Itt ugyanis az adatmátrix (melynek elemei egy adott *vizsgázó* adott *kérdésre* adott helyes vagy helytelen válasza) minden elemét két szempontból vizsgálhatjuk. Az egyik szempont az adott vizsgázónak az összes többi kérdésre adott válaszainak helyessége. A másik szempont az adott kérdésre az összes többi vizsgázó által adott válaszok helyessége. E kettő alapján kiszámítható, mennyi annak a valószínűsége, hogy az adott vizsgázó az adott kérdésre helyes választ ad. Feltéve, hogy egy vizsgázó esélyei egy a képességeit *nem* meghaladó kérdés helyes megválaszolására nem kisebbek mint 50%, a helyes válasz valószínűsége ebben az esetben nagyobb lesz, mint 0.5 (ez az érték a program futtatása során módosítható). A Rasch analízis során ezt a számított valószínűséget vetjük össze valós válasszal, mely vagy megerősíti a nagyobb valószínűségű esemény bekövetkezését, vagy cáfolja azt. Ez utóbbi módosítást eredményez a korábbi becsült értékekben, s az algoritmus újra lefut – egészen addig, míg a becsült és valódi eredmények egy előre meghatározott értéknél kisebb eltérést mutatnak. Eddigre összeáll két lista: a kérdések nehézségi listája a rájuk adható helyes válaszok valószínűsége függvényében, illetve a vizsgázók listája az adott kérdések helyes megválaszolásának valószínűsége függvényében.

Fontos megjegyezni, hogy ez a módszer egy szempontból igencsak eltér a hagyományosan megszokottaktól, ezáltal alkalmazói némi ellenállásra számíthatnak az avatatlan érintettek részéről. A fentiekből következik ugyanis, hogy a hallgatók végső eredménye (azaz készségük megállapított szintje) nem csak (sőt nem elsősorban) a helyesen megválaszolt kérdések számától függ; sokkal inkább azok nehézségétől. Így előfordulhat, hogy valaki jobb minősítést kaphat másoknál, annak ellenére, hogy pontszámát tekintve mögöttük végzett.

Beszéd és Írás

A produktív készségek esetében az IRT lehetővé teszi a legnagyobb problémaforrás, a szubjektivitás kiküszöbölését. Az adatmátrix elemeinél itt ugyanis az a kérdés, mi a valószínűsége annak, hogy egy adott *vizsgázó* egy adott *értékelőtől* adott pontszámot (osztályzatot, fokozatot, stb.) kap. A fentiekkel megegyezően lefutó analízis végére a vizsgázók sorrendje mellett (mely azon alapul, mekkora valószínűséggel kapnak adott értékelést a különböző vizsgáztatóktól) előáll a vizsgáztatók sorrendje is. Ez megmutatja, ki milyen valószínűséggel jutalmazza adott értékeléssel a vizsgázókat – köznapi nyelven ki a szigorúbb. Mivel a valószínűségek számításánál itt is kétirányú a megközelítés, a vizsgázók végső sorrendje már nemcsak azon alapul, milyen értékelést kaptak – hanem azon is, hogy azt milyen szigorú értékelőtől érdemelték ki.

Megjegyzendő, hogy a produktív készségek ilyenfajta elemzése némi logisztikai problémát jelent. A másik két készség esetében az összehasonlíthatóság alapja az volt, hogy *ugyanazokat* a kérdéseket *ugyanazok* a vizsgázók válaszolták meg; ez kötötte őket össze, lehetővé téve az egy rendszeren belül váló kezelésüket. Ez a produktív készségek esetében azt jelenti, hogy az összehasonlítandó vizsgázóknak (értékelőknek) kell, hogy legyenek közös értékelőik (vizsgázóik). Ez lehetetlenné teszi a párhuzamos vizsgáztatást, ami időigényesebb lebonyolítást eredményez. De ki ne vállalná ezt a kis kellemetlenséget annak érdekében, hogy megbízhatóbb eredmény szülessen a vizsgán?