



Psühholoogiline testimine

Psühholoogiline test on mõnes mõttes sarnane joonlauaga – mõlemad on mõeldud mingi omaduse mõõtmiseks. Joonlaud mõõdab pikkust, test võib mõõta näiteks intelligentsust või ekstravertsust. Mõõtmise täpsus ei sõltu ainult joonlauast või testist endast, vaid ka sellest, kuidas seda kasutatakse. Kui me tahame mõõta A4 lehekülje laiust, siis tuleks joonlaud paigutada lehe lühema servaga paralleelselt nii, et skaala nullpunkt on lehe äärega täpselt kohakuti. Kui paigutame joonlaua diagonaalselt keset lehte, siis mõõdame juba midagi muud või üldse mitte midagi, hoolimata sellest, et mõõtevahend ise on piisavalt täpne. Sama kehtib ka testide kohta: ka kõige paremat testi saab kasutada vildakalt, nii, et selle tulemustel puudub igasugune väärtus. Kuid testi puhul on jaburat tulemust raskem avastada. Igaüks saab aru, et A4 lehekülje laius ei ole kindlasti mitte 6,8 ega 197 sentimeetrit; joonlaud, mis sellise tulemuse annab, on kas ise vigane, või on seda valesti kasutatud (või on skaalal sentimeetrite asemel midagi muud; muuseas, A4 lehe laius ongi ligikaudu 6,8 atoparsekit ehk 197 nanoversta). Selliste psühholoogiliste omaduste nagu isiksuseomadused või intelligentsus puhul on ilmselt jaburaid tulemusi raskem avastada: siin ei ole täpseid ja selgelt defineeritud mõõtühikuid, inimesed usaldavad vähem oma silmamõõtu ja tervet mõistust ning tahaksid uskuda asju, mis on küll meelitatavad, kuid väga ebatõenäolised (à la „minu kassi IQ on vähemalt 250“); lisaks sellele esitatakse testitulemusi teinekord viisil, mis petlikult imiteerib füüsikaliste mõõtmiste täpsust. Sellest hoolimata võib psühholoogilistest testidest mõistliku kasutamise korral olla palju kasu; testide hingeelu minimaalne tundmine aitab teha vahet testide mõisliku kasutamise ja bluffimise või lausa pettuse vahel.

Järgnevalt vaatleme paari mõistet, mis on testide kasutamisel kõige olulisemad.

1. Valiidsus. Klassikalise definitsiooni järgi näitab valiidsus, kas test mõõdab seda omadust, mida ta mõõtma peaks – näiteks, kas matemaatikavõimete või sõbralikkuse test mõõdavad matemaatikavõimeid või sõbralikkust, aga mitte midagi muud. On ilmselge, et ilma valiidsuseta ei ole ühegi testi kasutamine õigustatud. (Ja muidugi, igaüks, kes mõnd testi kasutab, peaks suutma näidata, millise info põhjal ta arvab, et test on valiidne – kui ta seda ei suuda, siis on tõenäoliselt tegu šarlataniga.)

On lihtne välja mõelda viise, kuidas teada saada joonlaua valiidsust: võiksime näiteks mõõta teadaoleva pikkuse või laiusega objekte (üks sellistest on A4 lehekülg, mille standardsed mõõtmed on 210 x 297 millimeetrit). Kui saame õigevõitu tulemused ning mõõtmistulemus varieerub sõltuvalt objekti mõõtmest, siis võime öelda, et joonlaud on valiidne. Seega on valiidsuse teada saamiseks vaja kriteeriumi – näiteks teadaoleva pikkusega objekte. Üks klassikaline meetod psühholoogiliste testide valiidsuse uurimiseks ongi **kriteeriumgruppide** meetod: testitakse inimeste grupe, mille puhul on teada, et nad erinevad üksteisest mingi omaduse poolest – näiteks intelligentsustesti puhul võiksime katsetada, kas insenerid saavad koristajatest paremaid tulemusi. Selle meetodi puuduseks on, et tegelikult ei ole olemas grupe, mis erineksid üksteisest ühe ja ainult ühe omaduse poolest. Inseneri töö nõuab arvatavasti küll rohkem taibukust kui koristaja oma, kuid neil gruppidel on ilmselt ka teisi erinevusi, ning koristaja puhul ei ole madal intelligentsus sugugi tööle saamise kriteeriumis. Seega võib

kriteeriumgruppide meetod anda kasulikku informatsiooni valiidsuse kohta, kuid sellest ei pruugi piisata.

Mõnikord nimetatakse valiidsuseks ka seda, kui õnnestub näidata, et test on küllalt sarnane mõne teise, üldtunnustatud testiga. Sellise lähenemise puhul eeldame, et selle teise, üldtunnustatud testi valiidsus on juba veenvalt tõestatud. Üldtunnustatud arvamust võib usaldada, aga tuleks kindlasti kontrollida – seega oleme algpunktis tagasi.

Teoreetiliselt kõige parem viis testi valideerimiseks oleks uurida **psühholoogilisi protsesse** testiküsimusele või -ülesandele vastamisel: kui õnnestub näidata, et mingile küsimusele õigesti (või jaatavalt) vastamiseks on tingimata vaja teatud tüüpi võimeid, teadmisi või omadusi, ning ilma nendeta ei saa sellele küsimusele nii vastata, siis see oleks kõige otsesem teave testi valiidsuse kohta. Sellise (range) lähenemisviisi puhul ei piisa valiidsuse tõestamiseks ühest või paarist uurimusest, vaid on vaja mitmeid erinevaid eksperimente. Ka on see raskemini rakendatav näiteks isiksuse- ja hoiakutestidele: vastaja võib mõne küsimusega nõus olla lihtsalt sellepärast, et ta tahab endast jätta paremat muljet (ta tahab seda ilmselt ka intelligentsustestis, kuid seal on endast hea mulje jätmiseks ainuke viis õige vastuse teadmine – eeldades muidugi, et spikerdamine on tehtud küllalt raskest).

Sisuväliidsus tähendab küsimust, kui hästi ja täielikult testiküsimused esindavad uuritavat omadust. Kõige lihtsam näide võiks siin olla kirjalik lõpueksam, millel on palju sarnaseid omadusi psühholoogilise testiga. Ajaloo lõpueksami sisuväliidsus on hea, kui eksamitöös on võrdselt esindatud kõik ajaloo õppekava olulisemad punktid; kui eksamiküsimused on kõik mingi ühe valdkonna (näiteks Jüriöö ülestõusu) kohta, siis on sisuväliidsus puudulik: hea tulemus näitab küll häid teadmisi ühe sündmuse kohta, kuid need ei pruugi esindada ajalooteadmiste koguhulka. Sellist testi ei saaks siis nimetada enam ajalooteadmiste testiks, vaid näiteks jüriöö-ülestõusu-teadmiste-testiks. Intelligentsuse või ärevuse puhul on sisuväliidsuse määramine keerulisem – ei ole olemas ühtset intelligentsuse või ärevuse õppekava, kus kõik oluline oleks kirjas. Sellise õppekava aset võiks täita teooria – intelligentsuse ja ärevuse teooriaid on mitmeid, ning hea test võiks toetuda neist ühele (soovitavalt muidugi kõige paremale!).

Sageli kasutatakse ka **ennustava valiidsuse** mõistet, mis küll otseses mõttes ei ole valiidsus, kuid on hädavajalik testide praktilisel kasutamisel. Oletame näiteks, et me tahame ärevusetesti kasutada personalivalikul – valida 1000 inimese hulgast need kümme, kes kõige paremini sobivad lenduriks. Selle ülesande puhul ei ole väga tähtis, kas test mõõdab tegelikult seda, mida ta peaks mõõtma (st ärevust), kuid kriitilise tähtsusega on see, kas ta ennustab lenduri tööga hakkama saamist.

2. Ühedimensioonilisus ja süstemaatiline mõõtmisviga. Peale selle, et test mõõdab seda, mida ta mõõtma peaks, oleks hea, kui ta mõõdaks *ainult* seda (st mõõtmistulemus sõltuks süstemaatiliselt ainult meid huvitavast omadusest). Kujutlege intelligentsustesti, kus suur osa küsimusi puudutab kriketimängu ning eeldavad baastadmisi sellest mängust. Sellist testi võiks edukalt kasutada maadel, kus enam-vähem igaüks kriketit tunneb (näiteks Sri Lankal). Kui aga vastajate teadmised kriketist tugevalt varieeruvad, siis hakkab test mõõtma juba kaht küllaltki erinevat omadust: üldintelligentsus ja kriketi-teadmised. Selline test ei ole enam ühedimensiooniline, ja kui me väidame, et tegu on intelligentsustestiga, siis on seal olemas süstemaatiline mõõtmisviga (tulemust mõjutab süstemaatiliselt peale intelligentsuse veel miski muu). Miks on süstemaatiline mõõtmisviga halb? Kui me kasutame testi teadusuuringus, siis võib see viia valedele järeldustele: näiteks võime leida, et testitulemus on seotud elueaga ja oletada, et taibukamad inimesed elavad pisut kauem. Tegelikuses võib see seos peegeldada hoopis kriketi mängimisega kaasnevat värskes õhus viibimist ja teatavat kehalist aktiivsust – me ei saa kunagi teada, kuidas tegelikult on, kui me ei saa süstemaatilist mõõtmisviga kontrolli alla. Kuid

süsteemaatiline mõõtmisviga on oluline ka praktiliselt: oletame näiteks, et tahame vastutusrikkale tööle (mis ei ole seotud kriketiga) leida mõned väga intelligentsed inimesed. Kui kasutame eelmainitud testi, siis diskrimineerime inimesi, kes on kriketimängus võhikud, ning anname teenimatu eelise neile, kes seda mängu hästi valdavad. Lõpptulemusena ei vasta test täita eesmärgile, milleks me teda püüame kasutada.

Süsteemaatiline mõõtmisvea näideteks isiksusetestides (peale kultuurilise, rahvusliku vm diskrimineerimise) on erinevad vastamiskalduvused: nt sotsiaalselt soovitatav vastamine ja kalduvus kasutada skaala äärmusi („täiesti nõus“, mitte lihtsalt „nõus“ või „peaaegu nõus“). Selliste vastamisstiilide olulisuse üle on rakenduslikus isiksusepsühholoogias palju vaieldud. Sotsiaalselt soovitatavat vastamist võib näiteks personalivalikult tõlgendada ka positiivsena: võib-olla me ei tahakski tööle võtta inimest, kes ei saa isegi sellest aru, et tööle kandideerimisel ei ole mõtet kõiki oma puudusi rõhutada.

3. Reliaablus, täpsus ja testiteooria. Reliaablus tähendab sõna-sõnalt mõõtmistulemuse usaldatavust; praktikas viitab see korratavusele (kui sarnased on tulemused, kui me mõõdame samal inimesel sama omadust mitu korda?). Joonlaud, mis igal mõõtmiskorral annab täiesti erineva tulemuse, ei saa olla kuigi hea (alternatiivne hüpotees on, et mõõtjal väriseb käsi). Joonlauri täpsus ei sõltu ainult reliaablusest: objektide mõõtmised võivad natuke muutuda (see asjaolu muudab reliaablust väiksemaks), samuti võivad millimeetreid tähistavad jooned asetseda üksteisest natuke kaugemal või lähemal kui täpselt millimeeter (see asjaolu ei mõjuta reliaablust, kuid vähendab mõõtmistäpsust). Ideaalselt täpse joonlauri puhul sõltub mõõtmistulemus ainult objekti tegelikest mõõtmetest, millele lisandub inimlik mõõtmisviga (kui täpselt joonlaud asetatakse objekti „algusega“ kohakuti). Oletame lihtsuse mõttes, et see mõõtmisviga on juhuslik, st mõnikord satub skaala nullpunkt objekti servast pisut vasakule, teine kord pisut paremale jne. Sel juhul võiksime saada täpsema tulemuse, mõõtes sama objekti mitu korda: mõni tulemus on tõelisest pisut suurem, teine pisut väiksem, kuid keskmises mõõtmistulemuses need juhuslikud kõikumised tasakaalustavad üksteist. (Võite proovida seda suuremas seltskonnas, paludes igaühel hinnata mingit vahemaad, näiteks kaugust aknast paistva puuni. Seejärel mõõtke tegelik kaugus nii täpselt, kui saate. Kümne inimese keskmine hinnang on kindlasti täpsem kui iga hinnang üksikult võetuna, ja võib olla küllalt lähedane tegelikule vahemaale.)

Sama asjaolu kasutatakse psühholoogilistes testides. Üksik küsimus või ülesanne ei pruugi olla kuigi usaldusväärne; kasutades mitut samateemalist ülesannet, on võimalik vähendada juhuslikku mõõtmisviga. Klassikalise testiteooria kõige lihtsam formulatsioon väidabki, et testitulemus koosneb „tõelisest skoorist“ ja juhuslikust mõõtmisveast, kusjuures juhusliku mõõtmisvea keskvärtus on null. (Pange tähele, et see väide eeldab ka testitulemuste ühemõõtmelisust, st süsteemaatilise mõõtmisvea puudumist!)

Samuti ei ole ükski küsimus ega väide universaalne: näiteks võimekustestis on lihtsamaid ja keerulisemaid küsimusi. Lihtsamad küsimused sobivad eristamiseks väga madala võimekusega inimesi keskmistest ja nutikamatest; keerulisemad küsimused eristavad väga taibukaid keskmistest ja päris rumalatest. Test ei saa olla eriti täpne, kui seal on ainult väga lihtsad või väga keerulised küsimused.

4. Normid ja ühikud. Testitulemusi väljendatakse tavaliselt **normskooridena**, – see tähendab, et ühe vastaja tulemusi tõlgendatakse populatsiooni keskmise suhtes. Kõige lihtsam versioon normskooridest on *z-skoorid*, mille keskvärtus on võrdne nulliga ja standardhälve ühega. See tähendab, et *z*-skoorides on keskmise inimese tulemus 0, negatiivsed arvud näitavad keskmisest madalamat ja positiivsed keskmisest kõrgemat tulemust. Kui testitulemused on normaaljaotusega, siis jääb -1 ja +1 vahele umbes 68% tulemustest, -2 ja +2 vahele juba üle 95% ning -3 ja +3 vahele üle 99% tulemustest. See tähendab, et *z*-skoorid, mille absoluutvärtus on suurem 3-st, on üsna

ebatõenäolised, ning absoluutväärtus üle 5 on juba üliharuldane (vähem kui üks miljonist). Tulemuste esitamisel kasutatakse z-skoore harva: kellelegi ei meeldiks kuulda näiteks, et tema IQ on nullilähedane, isegi kui see tähendab täiesti tavalist keskmist tulemust. IQ skooride puhul on tavaks võtta keskmiseks tulemuseks 100 ning standardhälbeks 15 (st 68% tulemusi jääb vahemikku 85...115 jne). Inimene, kes väidab oma IQ-ks näiteks 175, on kas tõepoolest väga haruldane eksemplar (tõenäoliselt kõige targem eestlane) või liiga kergeusklik (jäädes uskuma suvalist internetist leitud testi). Peale IQ-skaala kasutatakse normskooride väljendamiseks ka mitmesuguseid teisi kokkuleppelisi ühikuid, mis aga enamasti põhinevad z-skooridel: näiteks nn T-skooride keskvärtus on kokkuleppeliselt 50 ja standardhälve 10. Lisaks on olemas protsentskoorid, mis ei sõltu normaaljaotusest: näiteks 80 protsentskaalal tähendab, et vastaja sai pisut parema tulemuse, kui 80% vastajatest (kuid ülejäänud 20% on jälle temast „paremad“).

Normskoor tuleb tõlgendada sellena, mis nad on: võrdlus mingi grupi keskmisega. Tuleks arvestada, kas testi normgrupp esindab üldpopulatsiooni (näiteks kõiki Eesti elanikke) või on seal mingi valikuline seltskond (näiteks Tartu Ülikooli tudengid, keskkoolilõpetajaid või töötuid üle 50-aastaseid meesterahvaid).

Kuigi normskoorid on testitulemuste väljendamisel kõige sagedasemad, ei ole need sugugi ainus võimalus. Mõne testi puhul on võimalik ette kujutada absoluutskoor: kui testitav valdkond on väga selgelt piiritletud (nt Euroopa riikide pealinnad või USA presidendid), siis on tulemust mõistlik väljendada protsendina maksimumist. Praktilisel kasutamisel on võimalik testitulemust väljendada ka kriteeriumi ühikutes, näiteks tõenäosusena teha ära lenduriksam, või müügimehe oodatava läbimüügina kuus.

6. Testide praktiline kasutamine. Testide kasutamisel mingite praktiliste otsuste tegemiseks tuleks endalt (või testi kasutajalt) küsida vähemalt järgmised küsimused (on olemas ka palju pikemaid ja põhjalikumaid nimekirju):

1. Kas mul on piisavalt teadmisi selle testi kasutamiseks?
2. Kas ma saan seda testi legaalselt kasutada? (Kellele kuulub autoriõigus; kas ja kui palju testi kasutamine maksab?)
3. Kas see test mõõdab omadusi, mida ma tahan mõõta? (Valiidsus) Kui hästi ta neid mõõdab? Kas seda väidab lihtsalt testi autor, või on selle kohta empiirilisi tõendeid (näiteks teadusartikleid)?
4. Kas testist on olemas eestikeelne (vm keelne) versioon? Kas see on korralikult ja asjatundlikult kohandatud, või lihtsalt üle jala tõlgitud?
5. Kas test sobib kõigile gruppidele, kellega seda on plaanis kasutada? Kas ta on kõigi sihtgruppide suhtes õiglane?
6. Milliseid järeldusi saab testitulemuse põhjal teha? Kas test on piisavalt täpne selleks, et saaks teha indiviidi-taseme järeldusi (nt „see kandidaat sobib lenduriks“), või saab teha ainult grupi-taseme järeldusi (nt „need 10 kandidaati 200-st sobivad lenduriks tõenäoliselt kõige paremini“)?