

psykometriikan perusteet



Lauri Nummenmaa

Turun Yliopisto - Psykologian opintomonisteita 3 2002

Psykometriikan perusteet

Lauri Nummenmaa

Turun yliopiston psykologian laitos
ISBN 951-29-2030-1
ISSN 1238-8661

Whatever exists at all exists in some amount.

Edward Thorndike (1918)

Anything that exists in amount can be measured.

William McCall (1939)

Sisältö

Esipuhe	xi
I Psykologisten testien perusteet	1
1 Mitä on psykometriikka	3
1.1 Testaamisen historiaa lyhyesti	4
1.2 Psykologiset testit nykypäivänä	6
2 Psykologiset testit	9
2.1 Psykologisen testin ominaisuudet	9
2.1.1 Käyttätymisotos	10
2.1.2 Mittausten standardointi	10
2.1.3 Käyttätymisen muuntaminen symboleiksi	11
2.2 Pieni kontrolloitu koe	11
2.3 Erilaiset testit	11
2.3.1 Erilaiset testisuoritukset	12
2.3.2 Maksimaalinen vs. tyypillinen suoritus	12
2.3.3 Objektiiviset ja subjektiiviset testit	12
2.3.4 Testin mittauskohde	13
2.3.5 Testien jaottelu	13
3 Mittaaminen ja tilastollisten perusmenetelmien kertaus	15
3.1 Mittaaminen	15
3.2 Tilastotieteen ja mittaamisen peruskäsitteet	16
3.2.1 Havainnot ja mitta-asteikko	17
3.2.2 Summamerkintä	17

3.2.3	Keskiluvut	18
3.2.4	Hajontaluvut	18
3.2.5	Normaalijakauma	19
3.2.6	Z-pistemäärä	19
3.2.7	Kovarianssi	20
3.2.8	Korrelaatio	20
3.2.9	Regressio	22
3.2.10	Faktorianalyysin perusteet	22
3.3	Transformaatiot	23
3.3.1	Mitta-asteikoiden invarianssi	24
3.3.2	Prosenttimuunnos	24
3.3.3	Aluemuunnos	25
3.3.4	Transformaatio z-pistemäärien avulla	25
3.4	Normit	25
3.4.1	Mahdollisia ongelmia normatiivisessa tulkinnassa	26
II	Psykologisten testien teoria	29
4	Testiteoriat	31
4.1	Klassinen testiteoria	31
4.2	Klassisen testiteorian perusoletukset	33
4.2.1	Mittauksen arvioiminen klassisen testiteorian avulla	34
4.3	Muita todellisen pistemäärän teorioita	35
4.3.1	Tauekvivalenttien ja olennaisesti tauekvivalenttien mittareiden teoria	35
4.3.2	Kongeneeristen mittareiden teoria	35
4.4	Yleistettävyysteoria	35
4.4.1	Reliabiliteetin arvioiminen yleistettävyysteorian näkökulmasta	36
4.5	Latentin piirteen teoria	37
4.5.1	Latentti piirre	37
4.5.2	Itemin ominaisfunktio	37
4.5.3	Lokaali riippumattomuus	38
4.5.4	Latentin piirteen teorian sovelluksia	38
5	Reliabiliteetti	39
5.1	Reliabiliteettikerroin	40
5.2	Reliabiliteetin estimointi	43
5.2.1	Kahden testiversion käyttö	43

5.2.2	Test-retest -menetelmä	44
5.2.3	Split-half -menetelmä	44
5.2.4	Sisäisen konsistenssin menetelmä	44
5.3	Mittauksen keskivirhe	45
5.4	Reliabiliteettiin vaikuttavia tekijöitä	47
5.4.1	Testattaviin liittyvät tekijät	47
5.4.2	Testiin liittyvät tekijät	48
5.4.3	Testin käyttöön liittyvät tekijät	48
5.4.4	Testin reliabiliteetin määrittäminen	48
5.5	Korrelaation attenuaatio	49
5.6	Erotussuureet	50
5.7	Summamuuttujat	51
5.8	Kuinka reliaabeli testin pitäisi olla?	51
6	Validiteetti	53
6.1	Mittauksen validiteetti	53
6.1.1	Sisältövaliditeetti	54
6.1.2	Käsitevaliditeetti	54
6.2	Kriteerivaliditeetti	55
6.2.1	Ennustevaliditeetti	55
6.2.2	Rinnakkaisvaliditeetti	55
6.3	Face validity	56
6.4	Päätöksenteon laadun arviointi	56
6.5	Reliabiliteetin ja validiteetin välinen suhde	58
7	Itemianalyysi	61
7.1	Distraktorianalyysi	61
7.2	Itemin vaikeuskerroin	62
7.3	Itemin erottelukyky	63
7.4	Korrelaatioiden tarkasteleminen	64
7.4.1	Itemin ja testin kokonaispistemäärän välinen korrelaatio	64
7.4.2	Itemien väliset korrelaatiot	64
8	Testin rakentamisen perusteet	65
8.1	Testin laatimisen vaiheet	65
8.1.1	Mittauskohteen valinta ja analysointi	65
8.1.2	Mittauksen suunnittelu	66
8.1.3	Esitestaus ja itemianalyysi	66
8.1.4	Reliabiliteetin estimointi	67

8.1.5	Validiteetin estimointi	67
8.1.6	Testin viimeistely ja normitus	68
8.1.7	Testin päivittäminen	68
8.2	Pienempien testien ja mittareiden laatiminen	68
III Testiteorian sovelluksia		69
9	Älykkyytestit	71
9.1	Älykkyyden määrittely	71
9.2	Älykkysteorioita	71
9.3	Wechsler Adult Intelligence Scale Revised	72
9.3.1	Kielellinen osa	73
9.3.2	Suoritusosa	73
9.3.3	WAIS-R:n heikkoudet	74
9.4	Älykkyytestien ongelmia	74
9.4.1	Mittauksen harhaisuus	74
9.4.2	Ennusteen harhaisuus	75
9.4.3	ÄO:n muuttuvuus	75
10	Persoonallisuustestit	77
10.1	Persoonallisuuden määrittely	77
10.2	Miten persoonallisuustestejä voidaan suunnitella?	77
10.3	Sixteen Personality Factors Test (16PF)	78
10.4	Persoonallisuustestien ongelmia	80
10.4.1	Sosiaalinen suotavuus	80
10.4.2	Satunnainen vastaaminen	80
10.4.3	Tulkittavuuden ongelma	80
11	Yleisimmät Suomessa käytetyt testit	81
11.1	Kyky / suoritustestejä	82
11.1.1	Raven Progressive Matrices (Raven, RPM)	82
11.1.2	Stanford-Binet (SB)	82
11.1.3	WAIS-R	82
11.1.4	WISC-R	83
11.1.5	WPPSI-R	83
11.2	Persoonallisuustestejä	83
11.2.1	Children's Apperception Test (CAT) ja Thematic Apperception Test (TAT)	83

11.2.2	Cartoon Attribution Strategy Test (CAST)	83
11.2.3	Minnesota Multiphasic Personality Inventory (MMPI)	83
11.2.4	Personality Research Form (PRF)	84
11.2.5	Rorschach Inkblot Test (Ro)	84
11.2.6	Sixteen Personality Factors Questionnaire (16PF)	84
11.3	Muut testit	85
11.3.1	Bender Visual Motor Gestalt Test (Bender)	85
11.3.2	Ihmispöörros (Goodenoughin koe)	85
11.3.3	NEPSY	85
11.3.4	Pirkanmaan perhepsykologinen testi (PPT)	85
11.3.5	Wartegg Zeichen Test (WZT)	86
12	Testien käyttäminen tutkimus- ja asiakastyössä	87
12.1	Testien käyttö tutkimuksessa	87
12.1.1	Testien tutkiminen	87
12.1.2	Testit tutkimuksen mittalaitteina	88
12.2	Testit asiakastyössä	88
12.2.1	Testaamisen ongelmia	89
A	Kirjallisuutta	91
B	Kaavoja	93

Esipuhe

Tämä oppimateriaali on tarkoitettu lyhyeksi johdatukseksi psykometriikan teoriaan ja psykologisten testien käyttämiseen mittalaitteina. Monisteen tavoitteena on sekä esittää psykologisen mittaamiseen liittyvä teoria, että myös antaa jonkinlainen käsitys testeistä psykologin työvälineinä. Tämän vuoksi sisältö painottuu varsinaisen psykologisen mittaamisen teorian lisäksi myös testeihin. Testiteoriaa käsitellään siinä laajuudessa kuin se on testien käyttöperiaatteiden ymmärtämisen ja testien laatimisen kannalta tarpeellista, kaikkia kaavoja ja todistuksia ei tässä yhteydessä käydä läpi. Muutamia mielenkiintoisia todistuksia tosin esitetään, mutta niiden yksityiskohtainen ymmärtäminen ei ole välttämätöntä muun materiaalin ymmärtämiselle. Monistetta laadittaessa on yleensäkin pyritty siihen, että psykometriikan opiskeleminen ei vaatisi kovin laajoja pohjatietoja matemaatiikasta ja tilastotieteestä. Tilastollisten perusmenetelmien ja perusmatematiikan hallitseminen on kuitenkin välttämätöntä, sillä testiteorian johtaminen edellyttää tiettyjen matemaattisten perusmenetelmien soveltamista.

Psykometriikka antaa itsestään melko epäedullisen vaikutelman, koska suuri osa

teoriasta esitetään erilaisina kaavoina. Tätä kuitenkin seuraa eräs tämän psykologian osa-alueen erityispiirre. Suuri osa psykometriikan teoriasta on mahdollista esittää aksiomaattisesti, tietyistä lähtöoletuksista on mahdollista johtaa perustulokset ja niiden kehitelmät. Vastaavanlaista lähestymistapaa päästään soveltamaan hyvin harvoin psykologiassa. Tämänkaltaisen teoreettinen lähestymistapa poikkeaa psykologiassa yleisemmin sovelletuista menetelyistä, mutta loppujen lopuksi psykometriikan teorian perusteet eivät ole kovinkaan monimutkaisia. Kaavoista onkin tärkeää muistaa ennemminkin sisältö kuin ulkoasu, kaavathan vain tiivistävät tiettyjä asioita lyhyempään ja *matemaatikon mielestä* helpommin ymmärrettävään muotoon.

Psykometriikan tulosten välitön sovelletavuus on myös erittäin miellyttävä ominaisuus. Lähes kaikkia tässä esitettyjä tuloksia voidaan soveltaa psykologin työhön ja tutkimukseen sellaisenaan. Jo melko suppealla testiteorian tuntemuksella on mahdollista laatia käyttökelpoisia testejä ja kyselyitä, sekä soveltaa niitä tutkimukseen.

Käytännössä psykometriikan tietoja tarvitaan sekä testejä arvioitaessa että testejä laadittaessa. Varsinaisten testien laatimi-

nen on erittäin työlästä, ja harva psykologi osallistuu oikean testin laatimiseen työurallaan. Psykometriikan tuntemus on kuitenkin hyödyksi myös pienempiä mittareita ja kyselyitä laadittaessa. Siten psykologisen mittaamisen teorialle on useita sovellutuksia jo opiskeluvaiheessa. Esimerkiksi tutkielmissa mahdollisesti käytettäviä mittareita voi arvioida tämän monisteen tiedoilla aivan yhtä monipuolisesti kuin monet psykologian alan tieteelliset julkaisut edellyttäisivät.

Monet henkilöt ovat antaneet arvokkaita kommentteja tekstistä, mistä haluan tässä yhteydessä lausua kiitokseni. Jos sinulla on jotain kommentteja tästä monisteesta, ole hyvä ja ota yhteyttä minuun sähköpostitse osoitteeseen latan@utu.fi.

Turussa 8. helmikuuta 2004

Tekijä

Osa I

Psykologisten testien perusteet

Luku 1

Mitä on psykometriikka

William James (1890) esitti kuuluisan artikkelinsa nimessä klassisen kysymyksen "What is an Emotion", mitä on emotioni? Vaikka James ei kenties osannutkaan määritellä emotionia tavalla, joka tyydyttäisi tämän päivän tutkijoita, oli hänen lähestymistapansa tämän ilmiön tutkimiseen erinomainen. James aloitti emotionoiden tarkastelemisen *määrittelemällä* tutkittavan käsitteen. Tällä tavoin aloitetaan myös psykometriikka- nimiseen psykologian osa-alueeseen tutustuminen: Aivan ensimmäiseksi selvitetään, *mitä on psykometriikka*.

Sana psykometriikka muodostuu sanoista *psykhe* ja *metrika*. Psykhe merkitsee sielua, ja metrika mittaamista. Psykologia (*psykhe* + *logos*) tarkoittaa oppia ihmisen sielusta, toiminnasta ja käyttäytymisestä. Mittaaminen puolestaan on symbolien liittämistä ilmiöihin tiettyjen sääntöjen mukaan. Voitaisiin siis tiivistäen sanoa, että psykometriikalla tarkoitetaan *sielun mittaamista*. Nykykielellä on kenties parempi määritellä psykometriikka *psykologisen*

mittaamisen tutkimuksena.

Thorndike ja McCall toteavat yhdessä ensimmäisellä sivulla, että kaikki jossain määrässä esiintyvä asia on mitattavissa. Psykometriikka pyrkii selvittämään, millä tavalla psykologisten ilmiöiden mittaaminen voidaan suorittaa mahdollisimman virheettömästi. Monet psykologiassa tutkittavat ilmiöt ovat erittäin monimutkaisia ja vaikeasti mitattavia. Silti voimme usein olla helpostikin vakuuttuneita niiden olemassaolosta. Hyvä esimerkki tällaisesta ilmiöstä on esimerkiksi persoonallisuus tai älykkyys. Molemmat tuntuvat selkeästi maailmaa jäsentäviltä käsitteiltä ja olemassaolevilta ilmiöiltä. Suurin osa tutkijoita on yksimielisiä siitä, että esimerkiksi älykkyys ja persoonallisuus ovat olemassaolevia ihmisen ominaisuuksia. Mutta miten älykkyyttä tai persoonallisuutta voidaan mitata?

Jotkin mittauskohteet ovat paljon yksinkertaisempia. Esimerkki helposti määritellystä ja helposti mitattavasta käsitteestä on vaikkapa pituus. Pituuden mittaaminen ja määrittely on yksinkertaista. On sovittu, et-

tä pituuden mittayksikkö metri on sellainen matka, jonka valo etenee tyhjiössä $\frac{1}{299792458}$ sekunnissa. Tämän tiedon perusteella on mahdollista valmistaa pituuden mittaamiseen erilaisia laitteita, joiden avulla esineiden pituudet on helppo saada selville. Vastaavasti lämpötilan mittayksikkö kelvin on $\frac{1}{273.16}$ veden kolmoisasteen termodynaamisesta lämpötilasta, ja tämän tiedon avulla voidaan vastaavasti laatia mittareita esineiden lämpötilan mittaamiseksi.

Tästä esimerkistä käy varmaankin ilmi psykologisen mittaamisen ongelma. Käsitteet kuten älykkyys tai sosiaalisuus sisältävät erittäin paljon informaatiota, eivätkä ne ole *suoraan* mitattavissa. Jos sanomme henkilön olevan älykäs, niin määritämme hänen toimintaansa hyvin monella tavalla erilaisissa tilanteissa. Meillä on psykologeina jonkinlainen käsitys siitä, miten älykäs henkilö toimii esimerkiksi koulussa tai työpaikallaan. Älykkyys on siten sellainen ominaisuus, jonka määrä on jollain tavalla arvioitavissa ihmisen *käyttäytymisen* perusteella. Ei kuitenkaan voida määrittää mitään yksittäistä asiaa, jonka mittaaminen ilmoitaisi suoraan henkilön älykkyyden. Niinpä älykkyyden mittaamiseen täytyykin soveltaa toisenlaista lähestymistapaa. Sen sijaan, että mitattaisiin *suoraan* älykkyys -nimistä ominaisuutta, mitataankin henkilön käyttäytymistä ja tehdään havaintojen perusteella päätelmiä hänen älykkyydestään. Tällöin oletetaan, että älykkyys ilmenee tietynlaisena käyttäytymisenä, ja tämän 'älykkään' käyttäytymisen mittaamisella pyritään estimoimaan henkilön älykkyyttä. Psykometriikan tehtävänä onkin nimenomaan kehittää teoria ja välineet tällaisten mitta-

laitteiden luomiseen.

Tämä saattaa vaikuttaa melko itsestään selvältä. Fysiologisia mittauksia lukuunottamatta psykologinen tutkimus on lähes aina ihmisen käyttäytymisen mittaamista tavalla tai toisella. On kuitenkin muistettava, että useissa tapauksissa tutkija tai psykologi ei ole kiinnostunut käyttäytymisestä sinänsä, vaan ennemminkin niistä psykologisista prosesseista, jotka voidaan havaita tietynlaisena käyttäytymisenä. Psykologisten prosessien ja havaitun käyttäytymisen välillä ei kuitenkaan ole koskaan sataprosenttista vastaavuutta. Siksi onkin tärkeää tutkia, miten psykologisista ilmiöistä voidaan saada tietoa ihmisten käyttäytymistä mittaamalla.

Käytännössä psykometriikkaa tarvitaan usein erilaisten testien laatimisessa ja käyttämisessä. Testejä puolestaan käytetään useimmiten ihmisten välisien erojen selvittämiseen. Koska testit ovat monen psykologin keskeisiä työvälineitä, on ammattilaisen tunnettava tarkoin niiden laatimisessa ja soveltamisessa tarvittava teoria. Psykometriikka *ei* kuitenkaan ole pelkästään psykologisten testien teoriaa, vaan yleisemmin psykologisen mittaamisen teoriaa. Tässä yhteydessä käsiteltävät psykometriikan teoriat ovat siten sovellettavissa suureen osaan psykologisista mittauksista.

1.1 Testaamisen historiaa lyhyesti

Psykologia on tieteenä varsin nuori. Järjestelmällisen psykologisen testaamisen tutki-

muksen historia on vieläkin lyhyempi. Varsinaisia psykologisia testejä on ollut olemassa vain noin 150 vuotta. Huomattavimpia henkilöitä yksilöiden välisten erojen tutkimisen historiassa on eittämättä englantilainen **Sir Francis Galton** (1822-1911), joka Charles Darwinin luonnonvalintateorian innoittamana kehitti menetelmiä, joiden avulla hän saattoi mitata erilaisia ihmisten psyykkisiä ominaisuuksia. Galton oli eräs ensimmäisistä tutkijoista, joka käytti *kyselylomakkeita* tutkiakseen eroja ihmisten psyykkisissä toiminnoissa. Galtonin järjestelmälliset mittaukset olivat perustana myöhemmälle psykologisten testien kehittämiseksi. Hän oli myös eräs ensimmäisistä tutkijoista, joka sovelsi tilastollisia menetelmiä psyykkisten ilmiöiden tutkimiseen. Galtonin oppilas **James McKeen Cattell** (1860-1944) jatkoi edelleen psykologiassa tarvittavien tilastollisten menetelmien tutkimista. Englantilainen matemaatikko **Karl Pearson** (1857-1936) puolestaan kehitti suurta osaa nykyaikaisessa tilastotieteessä käytettävistä menetelmistä, kuten korrelaatiomenetelmiä ja regressioanalyysia. **Charles Spearman** (1863-1945) loi älykkyystudkimuksillaan edellytyksiä nykyaikaisten psykologisten testien suunnittelemiseksi. Spearmanin teorian mukaan ihmisen älykkyyteen vaikuttaa kaksi tekijää, yleinen älykkyyys (*g-faktori*) ja spesifi tilannekohtainen älykkyyys (*s-faktori*). Spearmanin panos *faktorianalyysina* tunnetun menetelmäjoukon kehittämiseen oli myös huomattava.

Käytännön tarpeet ovat usein vaikuttaneet hyvin voimakkaasti psykologisten testien kehittämiseen. Testejä on tarvittu suo-

rittamaan jokin erottelu- tai luokittelutehtävä. Nykyaikaisen psykologisen älykkyytestaamisen isänä voidaan pitää ranskalaisesta **Alfred Binet'a** (1857-1911). Ranskan opetusministeriö pyysi 1900-luvun alussa häntä kehittämään testin, jonka avulla heikkolahjaiset oppilaat pystyttäisiin tunnistamaan ja samalla arvioimaan heidän suorituskykynsä. Yhdessä työtoverinsa Simon'n kanssa Binet suunnitteli tarkoitukseen soveliaan testin, jota voidaan pitää ensimmäisenä systemaattisesti laadittuna älykkyytestinä. Tämä testi oli edeltäjä myöhemmin esiteltävälle *Stanford-Binet* -älykkyytestille. Testin varsinainen uutuus oli siinä, että se mittasi älykkyyttä usean erilaisen osatestin avulla, eli testillä voitiin mitata älykkyyden eri osa-alueita. Aikaisemmin suuri osa älykkyystesteistä oli perustunut jonkin yksittäisen ominaisuuden mittaamiseen. Binet'n ja Simon'n testi oli melko käyttökelpoinen ja luotettava tehtävässään, joskin sillä voitiin testata ainoastaan yksi henkilö kerrallaan. Lisäksi testiä käyttämään tarvittiin koulutettu ammattilainen.

Yhdysvaltain liittyessä ensimmäiseen maailmansotaan sotilaita värvättiin suuria määriä, ja niinpä tarvittiin jokin menetelmä, jolla rekrytoitujen miesten psyykkiset ominaisuudet pystyttiin edes summittaisesti arvioimaan nopeasti. **Robert M. Yerkes** (1876-1956) työtovereineen kehitti tarkoitusta varten *Army Alpha* ja *Army Beta* -testit, joiden perusteella voitiin arvioida värvättyjen henkilöiden sopivuutta erilaisiin tehtäviin. Testien etuna oli se, että ne voitiin esittää kerralla suurelle joukolle henkilöitä. Vaikka ne eivät olleetkaan

arvioissaan niin tarkkoja kuin mahdollista, testien avulla armeija sai nopeasti riittävän hyvän käsityksen kunkin miehen henkisistä kyvyistä. Army Alpha- ja Beta -testit olivat siten ensimmäisiä todella laajassa mittakaavassa käytettyjä *joukkotestejä*.

Itävaltalainen **Hermann Rorschach** (1884 - 1922) tunnetaan erityisesti hänen kehittämästään Rorschachin mustetahrates-tistä (*Ro*). Rorschach oli psykoanalyysista kiinnostunut lääkäri, joka tutki ihmisten mustetahroista tekemiä havaintoja. Hän pyrki selvittämään, miten ihmisen persoonallisuus vaikuttaa hänen mustetahroista tekemiinsä havaintoihin. Rorschachin tulokset eivät saavuttaneet suurta huomiota hänen elinaikanaan, mutta hänen kuoltuaan *Ro* -testi on levinnyt ympäri maailmaa. *Ro*:n alkuperäinen versio ei kuitenkaan ollut varsinaisesti psykoanalyttinen, vaan enemmänkin havaintopsykologinen testi. Psykodynaaminen tulkintatapa on kehitetty vasta Rorschachin kuoleman jälkeen.

Louis Leon Thurstone (1887-1955) tutki älykkyyden rakentumista faktorianalyysin avulla. Hän päätyi tulokseen, jonka mukaan älykkyyttä ei voida kuvata yhden yleisfaktorin (*g*) avulla, vaan älykkyys koostuu seitsemästä erillisestä osa-alueesta. Tutkimuksissaan Thurstone kehitti myös faktorianalyysia tutkimusmenetelmänä. **David Wechsler** (1896-1981) tunnetaan erityisesti hänen nimeään kantavista älykkyystesteistä. Wechslerin mukaan älykkyys ei ole yleinen ominaisuus vaan moniulotteinen käsite. Wechslerin testeissä älykkyyttä mitataan kymmenellä tai yhdellätoista erilaisella osatestillä. Wechsler esitteli myös *älykkyysosamäärän* käsitteen

(*ÄO*, engl. *IQ*, *Intelligence Quotient*).

1.2 Psykologiset testit nykypäivänä

Tänä päivänä testeihin liittyvä kaupallinen toiminta on laajentunut valtavasti, ja testejä on markkinoilla paljon. Tietokoneiden, tilastotieteen ja tutkimusmenetelmien kehittyminen on luonut testien suunnittelulle ja arvioimiselle aivan uudenlaisia mahdollisuuksia. Testien myynnin ja käytön valvonta ei aina ole kasvanut testikaupan ja menetelmien kehittymisen vauhdissa. Testejä on markkinoilla niin paljon, että nykyään voi olla vaikea sanoa, mikä testi on hyvä ja mikä on huono. Monet merkittävimmät testit ovat levinneet ympäri maailman eri kielisinä laitoksina (esim. WAIS-R), kun taas toiset pysyvät tiukasti pienen psykologijoukon harrastuksena (esim. WZT). Kriittinen suhtautuminen testeihin ja niiden käyttämiseen onkin tärkeää. Jotkin markkinoilla olevista testeistä ovat melko vaatimattomasti suunniteltuja ja niiden käyttökelpoisuus on objektiivisesti arvioituna vähintäänkin kyseenalainen. Näitäkin testejä ikävä kyllä käytetään. Vaikka jokin testi olisikin tutkimuksissa arvioitu käyttökelvottomaksi, jotkut psykologit käyttävät sitä edelleen, koska 'testi kuitenkin toimii käytännön työssä hyvin'. Tällaiset perustelut on kuitenkin unohdettava, jos testejä halutaan käyttää psykologin ammattieettisten ohjeiden mukaisesti. Käytännön kokemus antaa tietenkin tietoa testin soveltuvuudesta kliniseen työhön. Mittalaitteiden ominai-

suudet on kuitenkin pystyttävä *osoittamaan* empiirisesti. Jos testi todellisuudessa toimii hyvin käytännössä, toimivuus voidaan varmastikin osoittaa tutkimuksessa. Hyvä psykometriikan tuntemus onkin välttämätöntä, jos testien käyttökelpoisuudesta halutaan tehdä tieteellisesti päteviä päätelmiä.

Suomessa psykologisten testien välitystä hoitaa Psykologien Kustannus OY. Yhtiö kääntää testejä suomen kielelle, ja myy niitä käyttäjille. Osa Suomessa käytettävistä testeistä on katsottu sellaisiksi, että niiden käyttöön ja tulkintaan tarvitaan laillistettu psykologi. Kustannus myös huolehtii siitä, että testejä myydään ainoastaan ihmisille, jotka ovat päteviä niitä käyttämään.

Luku 2

Psykologiset testit

Psykologisia testejä käytetään useimmiten yksilöiden välisten erojen havaitsemiseen.¹ Erojen perusteella tehdään mm. soveltuvuutta ja terveydentilaa koskevia päätelmiä ja ratkaisuja. Erilaisia testejä ja ihmisen psyykkisten ominaisuuksien mittaamiseen käytettäviä menetelmiä on lukuisia. Osa menetelmistä on parempia, osa huonompia, eikä mikään yksittäinen testi sovel- lu käytettäväksi jokaisessa tilanteessa. Eräs psykometriikan sovelluksista on nimenomaan testien käyttökelpoisuuden arviointi.

2.1 Psykologisen testin ominaisuudet

Psykologi voi työssään joutua käyttämään hyvinkin erinäköisiä testisysteemeitä, mutta tässä yhteydessä keskitytään ainoastaan psykologisiin testeihin. Neuropsykologi voi työssään käyttää esimerkiksi motori-

sen suorituksen mittaamiseen laadittuja testejä. Tässä yhteydessä käsitellään kuitenkin ainoastaan sellaisia testejä, jotka mittaavat jotain *ihmisen psyykkistä ominaisuutta*. Psykologinen testi voidaan määritellä usealla eri tavalla, mutta useimmiten sen oletetaan täyttävän seuraavat kolme kriteeriä:

1. Psykologinen testi on otos käyttäytymisestä.
2. Mittaus suoritetaan standardoituissa olosuhteissa.
3. Mitattavan henkilön havaittu käyttäytyminen muunnetaan symboleiksi tiettyjen sääntöjen mukaan.

Jos nämä kolme kriteeriä täyttyvät, ja lisäksi testin mittaus on suoritettu vähintään välimatka-asteikolla, kyseessä on *psykometrinen testi*. Kaikki psykologiset testit eivät suinkaan ole psykometrisia: Esimerkiksi WZT:ssä ei ole täsmälleen standardoitua menetelmää testivastausten tulkitsemiseksi, ja sen mittaus suoritetaan laatu-

¹Ihmisten välisiä eroja tutkivaa psykologian osaluetta on joskus kutsuttu myös *differentiaalipsykologiaksi*

roasteikolla. Myöhemmin esitettävä *klassinen testiteoria* soveltuu kaikilta osin ai-noastaan psykometristen testien tarkastele-miseen. Ei kuitenkaan voida sanoa, että tes-tin olisi välttämättä oltava psykometrinen, jotta se olisi laadultaan korkeatasoinen. Jos testi on laadittu psykometrisin perustein, sen arvioiminen on mahdollista klassisen testiteorian avulla. Ei-psykometristen tes-tien käyttökelpoisuuden objektiivinen ar-viointi voi joissain tapauksissa olla hyvin-kin hankalaa, mutta esimerkiksi ns. *laten-tin piirteen teoria* mahdollistaa laatueroas-teikolla suoritettujen mittaustulosten käsit-telemisen.

2.1.1 Käyttäytymisotos

Tarve testien laatimiseen johtuu siitä, että monien ihmisen mielen sisäisten toiminto-jen tai ominaisuuksien suora mittaaminen on mahdotonta. Voidaan kuitenkin olettaa, että nämä psyykkiset ilmiöt vaikuttavat ih-misen käyttäytymiseen. Näin ollen voidaan väittää, että ihmisen käyttäytymistä ja toi-mintaa arvioimalla voidaan tehdä päätelmiä hänen mielensä sisällöistä, rakenteesta ja toiminnasta.

Psykologinen testi on siis eräänlainen *käyttäytymisotos*. Psyykkisten ominaisuuksien oletetaan ilmenevän tietynlaisena käyt-täytymisenä. Olisi kuitenkin mahdotonta mitata kaikkea ihmisen käyttäytymistä, jos-sa tutkittava ominaisuus ilmenee. Testiä varten poimitaan *otos* henkilön käyttäy-tymisestä. Sen, miten henkilö vastaa tai suoriutuu testissä, katsotaan edustavan hä-nen toimintaansa *kaikissa* niissä tilanteessa, joista ollaan kiinnostuneita. Tämän lähes-

tymistävän tarkoitus käy varmasti ilmi, jos ajatellaan esimerkiksi älykkyyden mittaa-mista. On mahdotonta asettaa 'älykkyys'-nimistä ominaisuutta mittanauhalle ja mi-tata sen painoa, pituutta tai tilavuutta. Tä-män vuoksi on mitattava henkilön sellaista käyttäytymistä, jonka voidaan ajatella joh-tuvan hänen älykkyydestään. Ei ole myös-kään mahdollista tutkia jonkin henkilön toi-mintaa *kaikissa* sellaisissa tilanteissa, jois-sa hänen oletetaan soveltavan älykkyyttään. Siksi testiä laadittaessa toimitaan samalla tavalla kuin empiiristä tutkimusta tehdes-sä. Sen sijaan että mitattaisiin kaikki po-pulaation jäsenet, poimitaan populaatiosta *otos* ja tehdään sen perusteella päätelmiä asioiden luonteesta. Älykkyyden tapauk-sessa mitataan siis henkilön toimintaa jois-sain sellaisissa tilanteissa joissa oletamme hänen älykkyytensä oletetaan tulevan ilmi. Otoksen katsotaan edustavan kaikkea hä-nen 'älykästä' käyttäytymistään. Käyttäy-tymisotoksen perusteella voidaan edelleen tehdä johtopäätöksiä henkilön älykkyydes-tä. Otos-populaatio -ajattelu on siis jok-seenkin analoginen tilastotieteen vastaavan määrittelyn kanssa.

2.1.2 Mittausten standardointi

Jotta eri ihmisten testituloksia voitaisiin verrata keskenään, on jokainen ihminen tes-tattava samalla tavalla. Testaus tilanteen on oltava samanlainen kaikille testattaville, jo-ten tilanne on *standardoitava*. Testin teke-misestä ja pisteyttämisestä laaditaan ohjeet, joita noudattamalla pyritään siihen, että jo-kainen testauskerta olisi mahdollisimman samanlainen. Menettelyllä koitetaan var-

mistaa, että testitilanne ei itsessään pääse vaikuttamaan eri tavalla eri henkilöiden testisuoritukseen. On kuitenkin huomattava, että kaikkia psykologisia testejä ei tietenkään voida standardoida yhtä hyvin. Monissa testeissä ohjeita annetaan suullisesti ja testin tulkinnessa psykologilla voi olla hyvinkin suuri subjektiivinen osuus. Tämä ei välttämättä tarkoita sitä, että tällaiset testit olisivat huonoja testejä. Ennemmin asia olisi nähtävä niin, että testit on standardoitu *niin hyvin kuin kulloinkin on mahdollista*.

2.1.3 Käyttäytymisen muuntaminen symboleiksi

Ihmisen informaation prosessointikyky on rajallinen. Suuri määrä tietoa ei välttämättä jäsenny ja tiivisty helposti tulkittaviksi kokonaisuuksiksi. Niinpä testisuoritukset² muunnetaan symboleiksi, useimmiten numeroiksi. Numerot tai symbolit tiivistävät testillä hankitun informaation pienempään, helpommin tulkittavissa olevaan muotoon kuin esimerkiksi videotallenne henkilön toiminnasta testitilanteessa. Kaikkea käyttäytymistä *ei* kuitenkaan muunneta numeroiksi. Symboleiksi muuntamista on myös se, että lomakkeellemme kirjoitetaan kuvauksia henkilön toiminnasta testaustilanteessa. Tällöin testissä mitatun henkilön käyttäytyminen muunnetaan kirjainsymboleiksi, ja informaatio tiivistyy

²Eli testissä ilmennyt käyttäytyminen

2.2 Pieni kontrolloitu koe

Edellä esitetyn perusteella voidaan tiivistäen todeta, että testi on ikäänkin *pieni kontrolloitu koe*. Testin ja testitilanteen suunnittelussa on noudatettava samoja periaatteita kuin kokeellisessa tutkimuksessakin. Aivan kuin kokeellisen psykologiankin piirissä, on myös 'testikokeiden' antamien mittaustulosten pätevyyttä tarkasteltava huolella. Tämä on tärkeää, koska testitilanne ei milloinkaan vastaa täsmälleen sitä käyttäytymistä, josta tutkija on kiinnostunut. Yleensä testin luotettavuutta ja käyttökelpoisuutta tarkastellaan kahden suureen, reliabiliteetin ja validiteetin avulla. *Reliabiliteetti* tarkoittaa testitulosten virheettömyyttä. Testi on reliaabeli, jos mittaustulos sisältää vähän mittausvirhettä ja paljon sitä tietoa, josta mittauksessa ollaan kiinnostuneita. *Validiteetti* taas tarkoittaa sitä, mittauko testi sitä asiaa tai ominaisuutta, mitä sen tuleekin mitata. Nämä kaksi käsitettä ovat erittäin tärkeitä testejä arvioitaessa, ja niihin palataan myöhemmin uudelleen. Käsitteet on kuitenkin hyvä pitää mielessä seuraavia kappaleita lukiessa.

2.3 Erilaiset testit

Psykologisia testejä on lukuisia erilaisia ja eri tavoin toimivia. Niitä voidaan myös luokitella useilla eri tavoilla. Seuraavissa kappaleissa on esitetty erilaisia tapoja testien luokitteluun.

2.3.1 Erilaiset testisuoritukset

Eräs mahdollisuus on jaotella testit niiden suoritustavan mukaan, ts. käyttää jaottelun perusteena testitilanteessa tehtävää suoritusta. Tällöin voidaan käyttää esimerkiksi seuraavaa jaottelua:

1. Suoritustestit, joissa testattavaa pyydetään tekemään jokin suoritus, jonka oikeellisuuden testaaja pystyy helposti arvioimaan.
2. Havainnointi, jossa testaaja arvioi testattavan käyttäytymistä jossain tilanteessa.
3. Itsearviointit, joissa testattava itse kuvailee esimerkiksi omia tuntemuksiaan, asenteitaan tai mielipiteitään.

2.3.2 Maksimaalinen vs. tyyppilinen suoritus

Testit voidaan jaotella myös sen mukaan, pyritäänkö niissä saamaan selville testattavan maksimaalinen vai tyyppilinen suoritus. Maksimaalisesta suorituksesta ollaan kiinnostuneita esimerkiksi älykkyys- tai kykytestauksen yhteydessä. Tällöin pyritään mittaamaan henkilön paras mahdollinen suoritus kyseisessä tehtävässä. Tällaiseen tehtävään suunnitellut testit ovat tyyppillisesti aikarajoitettuja, ja suoriutumisen kriteereinä ovat oikeat vastaukset. Mitä enemmän oikeita vastauksia tai oikeita suorituksia henkilö saa, sitä paremmin hänen katsotaan suoriutuvan testissä. Tällaisessa testissä testaajan harhauttaminen on vaikeaa. Jos

oikeaa vastausta ei tiedä tai oikeaan suoritukseen ei kykene, sen tuottaminen on luonnollisesti vaikeaa tai mahdotonta. On tietysti mahdollista, että testattava voi tällaisessa testissä yrittää antaa itsestään epäedullisen kuvan vastaamalla tarkoituksenmukaisesti väärin.

Käyttäytymisotos voidaan hankkia myöskin henkilön tyyppillisen käyttäytymisen arvioimiseksi. Tyyppillisestä käyttäytymisestä ollaan kiinnostuneita mm. persoonallisuus-, kiinnostus-, tai asennetesteissä. Tällöin ei pyritä arvioimaan esim. 'maksimaalista persoonallisuutta', vaan henkilön persoonallisuuden luonnetta tai laatua. Tällaisissa testeissä ei yleensä ole aikarajoitusta tai oikeita vastauksia. Testaajan harhauttaminen on kuitenkin mahdollista. Koska oikeita vastauksia ei useinkaan ole, henkilö voi tuottaa vastauksia, joiden hän olettaa miellyttävän testaajaa. Tämä voi olla ongelmallista esimerkiksi silloin, kun persoonallisuustestejä käytetään rekrytoinnissa. Jotkut testit sisältävätkin skaaloja, joiden avulla vastaajan rehellisyyttä voidaan arvioida.

2.3.3 Objektiiviset ja subjektiiviset testit

Edelleen testit voidaan jaotella objektiivisiin ja subjektiivisiin testeihin. Testin objektiivisuus, ulkopuolisuus, tarkoittaa sitä, että kuka tahansa testaaja saa ohjeita seuraten saman tuloksen. Testaajan omilla tulkinnoilla ei siis ole vaikutusta testitulokseen, vaan samoja tulkintatapoja sovelletaan kaikkiin testattaviin. Tämänkal-

taisten testien käyttäminen ja tulkitseminen on yleensä suhteellisen yksinkertaista, eikä testimenetelmän oppiminen vaadi käyttäjältä pitkää koulutusta.

Subjektiiiviset testit puolestaan *edellyttävät* testaajan omia tulkintoja testisuorituksesta. Testivastausten tai testisuorituksen tulkitsemiseksi ei ole tarkkoja ohjeita, vaan testaaja tekee johtopäätökset itse oman harkintakykynsä varassa. Tällaisten testien käyttäminen edellyttää yleensä korkeatasoista asiantuntemusta ja kokemusta testin käytöstä. Subjektiiivisuus ei siis välttämättä tarkoita epäluotettavuutta tai mieli-
vältaisuutta vaan sitä, että testaajan henkilökohtaisilla tulkinnoilla on suuri merkitys testisuoritusta arvioitaessa.

2.3.4 Testin mittauskohde

Testit voidaan luonnollisesti jaotella myös sen mukaan, mitä niillä pyritään mittaamaan. Ei ole olemassa mitään yleistä tapaa jaotella testejä niiden mittauskohteen mukaan, mutta karkeasti testit voidaan jakaa seuraaviin kolmeen ryhmään:

1. Kykytestit
2. Persoonallisuustestit
3. Muut testit

Kykytestit mittaavat tyypillisesti henkilön maksimaalista suoritusta jollakin toiminnan osa-alueella. Älykkyystestit ovat tyypillisiä esimerkkejä kykytesteistä. Persoonallisuustestit mittaavat nimensä mukaisesti testattavan persoonallisuutta jollakin tavalla. Kuuluisin persoonallisuustesti lienee Ro. 'Muut testit'-kategoria sisältää kaikki testit, jotka eivät kuulu kyky- tai

älykkyystesteihin. Tällaisia ovat esimerkiksi monet neuropsykologiset testit.

2.3.5 Testien jaottelu

Yhteenvedona voidaan todeta, että testit voidaan jaotella ainakin seuraavilla neljällä tavalla :

1. Testin suoritustavan mukaan (tehtävä-observointi-itsearviointi)
2. Mitattavan suorituksen laadun mukaan (maksimaalinen-tyypillinen)
3. Testin pisteytyksen mukaan (objektiivinen-subjektiiivinen)
4. Testin mittauskohteen mukaan (kyky-persoonallisuus-muu)

Jaottelu on sikäli hyödyllinen, että sen perusteella voidaan vertailla eri testien ominaisuuksia. Esimerkiksi persoonallisuutta voidaan mitata joko objektiivisella tai subjektiivisella testillä. Eri tavoin toimivat testit lähestyvät tutkittavaa ongelmaa eri näkökulmista, joten ne saattavat myös antaa erilaisia tuloksia. Tämän vuoksi onkin syytä tietää, millaiseen mittaamiseen kukin testi on suunniteltu.

Luku 3

Mittaaminen ja tilastollisten perusmenetelmien kertaus

Tämä luku esittelee psykologisen mittaamisen suorittamista ja saatujen mittaus tulosten käsittelemistä. Psykometriikan perusteoria ei edellytä kovinkaan monimutkaisten matemaattisten tai tilastollisten menetelmien tuntemusta. Perusasiat on kuitenkin syytä hallita, joten tässä yhteydessä esitetään myös lyhyt kertaus tilastollisista perusmenetelmistä.

3.1 Mittaaminen

Kuten edellä on jo mainittu, mittaamisella tarkoitetaan jonkin symbolin, usein numeroarvon, liittämistä johonkin ilmiöön sillä tavalla, että jokin ilmiön ominaisuus on kuvattuna symbolilla. Mittaaminen voi olla

a) *Laadullista*, jolloin havainnot jaotellaan toisensa poissulkeviin luokkiin jonkin ominaisuuden mukaan. Esimerkiksi silmien väri, sukupuoli ja ammatti ovat

laadullisia mittauskohteita.

b) *Määrällistä*, jolloin jokaiseen havaintoon liitetään jokin numero, joka kertoo jotain havainnon suuruudesta, järjestyksestä, paremmuudesta tms. Reaktioaika on tyypillinen määrällinen mittauskohde.

Mittaamisen määritelmään sisältyy oletus, että mittauskohteet eroavat toisistaan jollain tavalla mitattavan ominaisuuden suhteen. Mittaaminen on mielekästä vain silloin, jos voidaan olettaa eri kohteista suoritetuissa mittauksissa olevan eroja. Testejä käyttäessä ollaankin usein kiinnostuneita juuri löytämään eroja yksilöiden välillä. Sellaista mittauskohdetta, jonka arvot vaihtelevat yksilöiden välillä kutsutaan *muuttujaksi*. Muuttujat jaetaan *kvalitatiivisiin* eli laadullisiin ja *kvantitatiivisiin* eli määrällisiin muuttujiin sen mukaan, onko mit-

tauksen tulos lukuarvo vai jokin laadullinen ominaisuus. Mittauskohde, jonka arvo on sama jokaisella yksilöllä on *vakio*.

Mittaaminen ja sen tutkiminen on muun psykologisen tutkimuksen kannalta erittäin tärkeää, koska standardoitu mittaaminen mahdollistaa ilmiöiden objektiivisen arvioinnin. Jos mittaukset standardoidaan hyvin ja niiden tarkkuus kyetään määrittämään, toiset tutkijat voivat toistaa mittauksen ja arvioida osaltaan sen kelvollisuutta. Lisäksi mittausmenetelmien kehittäminen mahdollistaa uudenlaisten psykologisten tutkimusten tekemisen. Esimerkiksi älykkyystestien tutkimus on mahdollistanut älykkyyden käsitteen entistä tarkemman määrittelyn ja uudenlaisten älykkyysteorioiden rakentamisen.

Hyvä mittausta on edellytys hyvälle tulokselle. Suunnittelemalla mittausvaihe huolellisesti välttää tulosten tarkasteluvaiheessa monelta ongelmalta. Mittauskohteen tarkka määrittely onkin erittäin tärkeää. Jos mittauskohde on kovin epämääräinen, ovat saadut mittaukselkin luonteeltaan melko epämääräisiä. Mittaukset pyritään usein suorittamaan siten, että tuloksena on numeroita. Ihmisen käyttäytymistä ei ole kovinkaan yksinkertaista kuvata numeroiden avulla, mutta numeroiden käyttöön on useita hyviä syitä. Ensinnäkin, numerot ovat objektiivisia. Kymmenen tarkoittaa samaa asiaa lähes kaikille ihmisille, kun taas käsitteet kuten 'paljon' ja 'lievästi' voidaan tulkita hyvinkin monella tavalla. Numerot ovat myös helposti viestittävisiä ja niiden avulla on mahdollista ilmaista asioita tarkasti. Lisäksi numeroiden käsitteleminen ja analysoiminen on usein paljon

helpompaa kuin esimerkiksi tekstiaineistojen. Numeroiden käsittelyyn on myös erittäin paljon erilaisia menetelmiä. Käyttämisen muuntaminen numeroiksi on usein vaativa tehtävä. Kuitenkin hyvän mittarin suunnittelemiseen käytetty aika säästetään usein analyysivaiheessa. Jos mittarin avulla saadaan virheettömiä ja helposti käsiteltävissä olevia mittaustuloksia, säästytään analyysivaiheessa paljon vaivannäöltä.

3.2 Tilastotieteen ja mittaamisen peruskäsitteet

Seuraavissa kappaleissa esitellään muutamia keskeisimpiä tilastotieteen määritelmiä ja merkintöjä siten, kuin niitä tässä monisteessa käsitellään. Merkitsemiskäytäntöjä on useita erilaisia, ja tässä esitetty järjestelmä perustuu matematiikan ja tilastotieteen yleisiin käytäntöihin. Huomaa, että merkinnöistä on yleensä ainakin kolme versiota. Populaation *paramteri* kuvaa jotain jakauman ominaisuutta koko populaatiossa. Parametreja pyritään arvioimaan otoksesta laskettujen tuunuslukujen eli *estimaattien* avulla. Lisäksi psykologian alan julkaisuissa käytetään erilaista merkitsemistapaa. Tässä esitetään kaikki kolme merkintätapaa, ja kahta enismmäistä noudatetaan kaikissa kirjan kaavoissa.

3.2.1 Havainnot ja mitta-asteikko

Havainto (havaintoarvo) tarkoittaa yhden mittauksen tulosta. *Havaintojen lukumäärä* eli otoksen kokoa merkitään pienellä n -kirjaimella. Artikkeleissa otoksen kokoa merkitään suurella N :llä, ja pieni n merkitsee jonkin otoksen osajoukon kokoa. *Mitta-asteikoilla* tarkoitetaan sitä tarkkuutta ja järjestelmää, jolla mittaustulokset hankitaan. Mitta-asteikoita erotetaan yleensä viisi erilaista. Asteikot poikkeavat toisistaan nollapisteen, havaintojen järjestyksen ja havaintojen välisten erojen määrittelyn suhteen. Asteikot on seuraavassa esitetty epätarkimmasta tarkimpaan:

1. *Nominaaliasteikko* eli laatueroasteikko jakaa havainnot toisensa poissulkeviin luokkiin. Nominaaliasteikolla ei määritellä havaintojen järjestystä tai suuruutta.

2. *Ordinaali- eli järjestysasteikko* liittää laatueroasteikon luokkiin jonkin suuruusjärjestyksen. Luokkien väliset suuruuserot eivät välttämättä ole yhtä suuria.

3. *Välimatka-asteikko* on jatkuva mitta-asteikko, jolla on nollapiste. Havaintojen väliset etäisyydet tunnetaan, mutta asteikon nollapiste ei kuitenkaan ole absoluuttinen. Nollaa pienempiä havaintoja voi siis olla olemassa.

4. *Suhdeasteikko* on jatkuva mitta-asteikko, jolla on absoluuttinen nollapiste, ts. nollaa

pienempiä arvoja ei voi olla olemassa.

5. *Absoluuttinen asteikko* on käytössä silloin, kun mitattavaa ominaisuutta ei voida mitata kuin yhdellä asteikolla.

3.2.2 Summamerkintä

Summamerkintä eli sigma (Σ) esiintyy monissa kaavoissa, ja se tarkoittaa yksinkertaista yhteenlaskua. Tämän muistaminen riittää useimmissa tapauksissa. Yleensä summamerkinnot ovat muodossa

$$\sum_{k=1}^4 (k+1)$$

Perusperiaate on se, että Σ -merkin oikealla puolella oleva lause (tässä siis $k+1$) lasketaan tietyillä k :n arvoilla, ja saadut tulokset lasketaan yhteen. Indeksointi (alapuolella oleva $k=1$ ja yläpuolella oleva 4) ilmoittavat, mitä k :n arvoja käytetään. Alapuolinen merkintä kertoo aloitusluvun (tässä siis 1) ja yläpuolinen taas viimeisen luvun (4). Tältä väliltä lauseke lasketaan kaikilla kokonaislukuarvoilla, eli tämä summalauseke laskettaisiin näin:

$$(1+1) + (2+1) + (3+1) + (4+1) = 14$$

Jos indeksointia (siis Σ :n ylä- ja alapuolisia lukuja) ei ole merkitty, lasketaan summalauseke ykkösestä alkaen kaikilla kokonaislukuarvoilla niin pitkälle kuin havaintoja riittää. Tällaista merkintätapaa käytetään usein kaavoissa.

3.2.3 Keskiluvut

Keskiluvuilla tarkoitetaan sellaisia tilastollisia tunnuslukuja, jotka ilmoittavat jakauman sijainnin x -akselilla. Keskiluvut ilmoittavat suunnilleen jakaumassa olevien havaintoarvojen suuruusluokan.

Keskiarvo ($\bar{x}$, artikkeleissa M) saadaan, kun jonkin joukon luvut lasketaan yhteen ja saatu summa jaetaan yhteenlaskettujen lukujen määrällä. Keskiarvo ilmoittaa 'kuinka paljon jokainen yksilö saisi, jos mitattava ominaisuus laitettaisiin yhteen kasaan ja sen jälkeen tasan kaikkien kesken'. Keskiarvon kaava on siten

$$\bar{x} = \frac{\sum x}{n} \quad (3.1)$$

Näin laskettu keskiarvo on ns. *aritmeettinen keskiarvo*, mutta muunkinlaisia keskiarvoja voidaan määritellä. Otoksesta laskettavalla keskiarvolla pyritään yleensä esimoimaan populaation keskiarvoa eli *odotusarvoa* μ .

Mediaani (md , artikkeleissa Mdn) on jakauman suuruusjärjestyksessä keskimmäinen havaintoarvo. Mediaani saadaan määritettyä, kun havainnot laitetaan suuruusjärjestykseen ja niistä poimitaan keskimmäinen mediaaniksi. Jos havaintoja on parillinen määrä, ts. kun ei voida valita yhtä keskimmäistä arvoa, mediaani saadaan, kun valitaan kaksi keskimmäistä lukua ja lasketaan niiden keskiarvo.

Moodi (mo , artikkeleissa $mode$) on havaintoaineiston tyypillisin arvo, ts. se havaintoarvo jota on aineistossa lukumääräisesti eniten.

3.2.4 Hajontaluvut

Hajontaluvuilla mitataan havaintoarvojen jakautumista tai 'leviämistä' jonkin keskiluvun, yleensä keskiarvon ympärille.

Varianssi (σ^2, s^2) ilmoittaa, millä tavalla havainnot ovat levinneet keskiarvon ympärille. Varianssi saadaan laskettua, kun lasketaan ensin paljonko jokainen havaintoarvo poikkeaa keskiarvosta. Nämä poikkeamat korotetaan toiseen potenssiin ja lasketaan yhteen. Lopuksi saatu tulos jaetaan havaintojen lukumäärällä. Eli toisin sanoen, lasketaan havaintojen keskiarvosta poikkeamien neliöiden keskiarvo. Suuri varianssi merkitsee sitä, että havainnot ovat keskimäärin kaukana keskiarvosta. Pieni varianssi taas merkitsee sitä, että suurin osa havainnoista on melko lähellä keskiarvoa. Populaatiovarianssi σ^2 määritellään seuraavasti:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} \quad (3.2)$$

missä:

$$\begin{aligned} \sigma^2 &= \text{varianssi} \\ (x_i - \bar{x}) &= \text{henkilön testipistemäärän ja testipistemäärien keskiarvon erotus} \\ n &= \text{havaintojen (henkilöiden) lukumäärä} \end{aligned}$$

Tämä siis estimoii otoksen perusteella koko populaation varianssia σ^2 . Otoksen varianssi merkitään s^2 , ja se voidaan laskea samaan tapaan kuin populaatiovarianssi, paitsi että

kaavan nimittäjästä vähennetään 1:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} \quad (3.3)$$

Keskihajonta (σ, s , artikkeleissa *SD*) tarkoittaa yksinkertaisesti sitä, kuinka paljon havainnot 'keskimäärin poikkeavat keskiarvosta'. Keskihajonta ilmoittaa kuinka kauaksi keskiarvosta havainnot ovat levinneet. Teoreettisissa tarkasteluissa hajonnan tunnuslukuna käytetään yleensä varianssia, empiirisessä tutkimuksessa taas keskihajontaa, sillä keskihajonnan mittayksikö on sama kuin alkuperäisen muuttujankin. Populaatiokeskihajonta σ saadaan laskettua populaatiovarianssin neliöjuurena $\sqrt{\sigma^2}$, ja otoskeskihajonta s vastaavasti otosvarianssin neliöjuurena $\sqrt{s^2}$.

3.2.5 Normaalijakauma

Normaalijakauma on jatkuva todennäköisyysjakauma, joka on kenties kaikkein tärkein tilastollisessa päättelyssä käytetyistä jakaumista. Normaalijakauma on määritelty välillä $[-\infty, \infty]$ ja sen tiheysfunktio on muotoa

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2} \quad (3.4)$$

missä

- σ = populaatiokeskihajonta
- π = pii, yksikkösaateisen ympyrän pinta-ala
- e = ns. Napierin luku ≈ 2.7183
- μ = odotusarvo

Ei siis ole olemassa ainoastaan yhtä vaan useita normaalijakaumia. Normaalijakauman muoto riippuu parametreista μ ja σ . Näistä μ ilmoittaa jakauman huipun sijainnin x-akselilla, σ taas jakauman 'levinneyden'. Normaalijakauman tiheysfunktion kaavan muistaminen ei ole erityisen tärkeää. Sen sijaan on pidettävä mielessä, että normaalijakauma on *teoreettinen* jakauma, jonka avulla voidaan *mallintaa* jotain havaittua ilmiötä. Normaalijakauma on erittäin tärkeä, koska muuttujat, joiden arvoon vaikuttavat useat tekijät, noudattavat melko tarkasti normaalijakaumaa. Esimerkiksi lähes kaikki ihmisen psyykkiset ominaisuudet ovat normaalisti jakautuneita. Normaalijakaumaa voidaan siten käyttää tilastollisen päättelyn apuna jos voidaan olettaa, että tutkittava ominaisuus on normaalisti jakautunut.

3.2.6 Z-pistemäärä

Usein muuttujan arvot, tässä yhteydessä siis testipistemäärät, on tarpeen muuntaa toiselle mitta-asteikolle. Jos arvojen jakauma on kutakuinkin normaali, voidaan suorittaa ns. z-pistemuunnos. Se muuntaa jakauman sellaiseksi, että sen keskiarvo on 0 ja keskihajonta 1¹. Z-pistemuunnos voidaan laskea seuraavasti:

$$z = \left(\frac{x - \bar{x}}{s_x} \right) \quad (3.5)$$

missä

$$z = \text{z-pistemäärä}$$

¹ks. normaalijakauman määrittely edellä

- x = testipistemäärä
 $\bar{x}$ = testipistemäärien keskiarvo
 s_x = keskihajonta

Jokaisesta testipistemäärästä vähennetään testipistemäärien keskiarvo, ja jaetaan saatu luku testipistemäärien keskihajonnalla. Z-muunnokseen palataan myöhemmin *transformaatioiden* yhteydessä.

3.2.7 Kovarianssi

Usein on tarpeellista tarkastella myös kahden muuttujan samankaltaista vaihtelua. Tällaista tarkastelua varten voidaan laskea muuttujien välinen *kovarianssi*. Mitä suurempi on kovarianssi, sitä samankaltaisemmin muuttujat käyttäytyvät. Kovarianssi ilmoittaa siis eräällä tapaa muuttujien välisen yhteyden voimakkuuden. Yksiulotteisen jakauman (yhden muuttujan jakauman) varianssi ei voi koskaan olla negatiivinen. Kahden muuttujan välinen kovarianssi voi olla joko negatiivinen, positiivinen tai nolla. Jos kovarianssi on negatiivinen, toisen muuttujan arvojen kasvaessa toisen muuttujan arvot pienenevät. Jos kovarianssi on positiivinen, toisen muuttujan arvojen kasvaessa toisenkin muuttujan arvot kasvavat. Jos kovarianssi on lähellä nollaa, ei muuttujien välillä ole minkäänlaista yhteyttä: esimerkiksi toisen muuttujan arvojen kasvaessa toisen muuttujan arvot vaihtelevat täysin satunnaisesti. Kovarianssi lasketaan vähentämällä jokaisesta x -muuttujan arvosta x -muuttujan keskiarvo, ja vähentämällä jokaisesta y -muuttujan arvosta y -muuttujan keskiarvo. Kaikki erotusparit $(x - \bar{x})$ ja $(y - \bar{y})$ kerrotaan keskenään ja lasketaan yhteen.

Lopuksi saatu summa jaetaan havaintoparien lukumäärällä:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} \quad (3.6)$$

missä

- s_{xy} = x : n ja y : n kovarianssi
 x_i = x - muuttujan arvo
 $\bar{x}$ = x - muuttujan keskiarvo
 y_i = y - muuttujan arvo
 $\bar{y}$ = y - muuttujan keskiarvo
 n = havaintoparien lukumäärä

Edellä esitetty lauseke ilmoittaa populaatiokovarianssin suuruuden. Vastaavasti voitaisiin laskea otoskovarianssi, tällöin nimittäjäksi tulisi $(n - 1)$ ja kaava pysyisi muuten samana.

Kovarianssiin liittyy kuitenkin samankaltainen tulkinnallinen ongelma kuin varianssiinkin. Kovarianssin suuruus riippuu muuttujien mittayksiköistä. Siten eri mittayksiköillä mitattujen muuttujaparien kovarianssien vertailu on mahdotonta. Esimerkiksi aikuisten ihmisten pituuksien ja painojen välinen kovarianssi on paljon suurempi kuin vastasyntyneiden pituuksien ja painojen kovarianssi. Tämä johtuu siitä, että aikuiset ovat pidempiä ja painavampia kuin vastasyntyneet, ja siten kovarianssistakin tulee suurempi.

3.2.8 Korrelaatio

Mittayksiköstä päästään kuitenkin eroon jakamalla kovarianssi molempien muuttujien

keskihajonnoilla. Näin saadaan mittayksiköstä riippumaton yhteisvaihtelun tunnusluku, josta käytetään nimitystä *korrelaatiokerroin*. Laskettaessa otoskorrelaatiota, käytetään lähtökohtana otoskovarianssia ja populaatiokovarianssin tapauksessa populaatiokovarianssi. Korrelaatiokerroin voidaan laskea seuraavasti:

$$r = \frac{s_{xy}}{s_x \times s_y} \quad (3.7) \quad \text{missä}$$

missä

r = x :n ja y :n välinen korrelaatiokerroin
 s_{xy} = x :n ja y :n välinen kovarianssi
 s_x = x – muuttujan keskihajonta
 s_y = y – muuttujan keskihajonta

Korrelaatiokerroin² ilmoittaa kahden muuttujan välisen suoraviivaisen (lineaarisen) yhteyden voimakkuuden. Toisin kuin varianssi, korrelaatiokerroin vaihtelee aina välillä $[-1, +1]$, ja sen itseisarvo ilmoittaa muuttujien välisen yhteyden voimakkuuden. Positiivinen korrelaatiokerroin tarkoittaa sitä, että toisen muuttujan arvojen kasvaessa toisenkin muuttujan arvot kasvavat. Negatiivinen korrelaatio puolestaan tarkoittaa sitä, että toisen muuttujan arvojen kasvaessa toisen muuttujan arvot pienenevät. Jos korrelaatio on lähellä nollaa, ei muuttujien välillä ole lainkaan lineaarista yhteyttä. Itse asiassa korrelaatiokerroin voidaan määritellä

²Korrelaatiokertoimia on useita erilaisia, mutta tässä niitä käsitellään yhdessä.

x - ja y -muuttujien *z*-pistemäärien avulla laskettavana kovarianssina. Tämän perusteellinen todistaminen sivuutetaan. Edellinen korrelaatiokertoimen kaavaa voidaan kuitenkin *z*-pistemääriä käyttämällä muuttaa muotoon

$$r = \frac{\sum (z_x \times z_y)}{n} \quad (3.8)$$

r = korrelaatiokerroin

z_x = x -muuttujan *z*-pistemäärä

z_y = y -muuttujan *z*-pistemäärä

n = havaintoparien lukumäärä

Toinen muuttuja nimetään x :ksi ja toinen y :ksi. Seuraavaksi lasketaan kaikille x - ja y -muuttujien arvoille *z*-pistemäärä (ks. edellä). Seuraavaksi jokainen x -muuttujan *z*-pistemäärä kerrotaan vastaavalla y -muuttujan *z*-pistemäärällä, ja saadut tulot lasketaan yhteen. Lopuksi saatu summa jaetaan havaintoparien lukumäärällä.

Miksi tämä laskutoimitus sitten ilmaisee korrelaation, tai muuttujien välisen yhteyden suuruuden? Positiivinen korrelaatiohan tarkoitti sitä, että jos henkilön x -pistemäärä on suuri, on myös hänen y -pistemääränsä suuri. Negatiivinen korrelaatio tarkoittaa, että jos henkilön x -pistemäärä on suuri, niin hänen y -pistemääränsä on pieni³. *Z*-muunnos puolestaan muunsi pistemäärät siten, että pienimmistä pistemääristä tuli suuria negatiivisia pistemääriä, ja

³'pieni' tarkoittaa tässä yhteydessä samaa kuin 'paljon negatiivinen', suuri taas samaa kuin 'paljon positiivinen'

suurimmista pistemääristä suuria positiivisia pistemääriä. Korrelaatiossa osoittajassa oleva summa muodostuu x -ja y -muuttujien z -pistemäärien tuloista. Kaksi suurta positiivista pistemäärää kerrottuna toisillaan on suuri positiivinen pistemäärä. Kaksi suurta negatiivista pistemäärää kerrottuna toisillaan on myös suuri positiivinen pistemäärä. Korrelaatiokerroin on siis kaikkien näiden tulojen keskiarvo, joten jos pieniä x :n arvoja vastaa aina pieni y :n arvo, ja suuria x :n arvoja suuri y :n arvo, muodostuu korrelaatiokertoimesta suuri. Suuri negatiivinen korrelaatiokerroin taas johtuu siitä, että suurta negatiivista x :n arvoa vastaa aina suuri positiivinen y :n arvo (positiivinen luku $\times$ negatiivinen luku = negatiivinen luku), ja päinvastoin. Nyt summasta tulee suuri, mutta negatiivinen. Nollaa lähellä oleva korrelaatio taas syntyy, kun osoittajaan tulee sekä positiivisia että negatiivisia tuloja. Nämä kumoavat osin toisensa, ja osoittajaan tulee lähellä nollaa oleva negatiivinen tai positiivinen luku.

3.2.9 Regressio

Korrelaatio kuvaa kahden muuttujan välisen yhteyden voimakkuutta, mutta ei luokitella mallia yhteyden laadusta. Jos kuitenkin halutaan mallintaa yhteyttä ja ennustaa mallin avulla henkilön pistemääriä, on käytettävä regressiomenetelmiä. Regressioanalyysissä luodaan matemaattinen mallin kuvaamaan x - ja y -muuttujien välistä yhteyttä. Tämän avulla voidaan ennustaa henkilön y -pistemäärä kun hänen X -pistemääränsä tunnetaan, ja toisinpäin. X -muuttujasta käytetään nimitystä *selittäjä*,

y -muuttujaa kutsutaan vastaavasti *selitettäväksi*. Kaikkein yksinkertaisin regressiomalli on suora muotoa $\hat{y} = a + bx$. Regressiomallin tapauksessa käytetään kuitenkin seuraavanlaista merkintätapaa:

$$\hat{y} = \beta_0 + \beta_1 x \quad (3.9)$$

missä

$\hat{y}$ = y :n ennustettu arvo

β_0 = vakiotermi

β_1 = regressiosuoran kulmakerroin

x = x -muuttujan arvo

Näinollen y :n arvo saadaan selville kertomalla vastaava x :n arvo β_1 :llä, ja lisäämällä tulokseen vakiotermi β_0 . Regressioanalyysissä muuttujien välistä yhteyttä ei välttämättä tarvitse kuvata suoralla, myös muunlaisia käyriä voidaan käyttää. Selittävien muuttujien lukumäärääkään ei ole rajoitettu, niitä voi periaatteessa olla kuinka monta hyvänsä. Tällöin selittäjien täytyy kuitenkin olla keskenään korreloimattomia. Usean muuttujan regressiomalli on muotoa

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (3.10)$$

Yhden selittäjän regressiosuoria on vielä mahdollista laskea käsin melko pienellä vaivalla, mutta monimutkaisempien mallien muodostamiseen tarvitaan käytännössä tietokonetta.

3.2.10 Faktorianalyysin perusteet

Faktorianalyysi on menetelmä, jonka avulla voidaan tutkia useiden muuttujien väli-

siä korrelaatioita. Jos esimerkiksi jossakin testissä on kymmenen muuttujaa ja halutaan laskea niiden kaikkien väliset korrelaatiot, saadaan yhteensä sata korrelaatiokerrointa ($45 \text{ erilaista} \cdot 2 + 10 \text{ ykköstä}$). Näin suuren korrelaatiomatriisin tulkitseminen on hyvin työlästä. Tämänkaltaisissa tilanteissa saatetaan tarvita faktorianalyysia tulosten tulkitsemiseen. Faktorianalyysin tarkoituksena on tutkia, mitkä muuttujat liittyvät läheisesti toisiinsa ja mitkä eivät. Yhteydet kuvataan ns. *faktorien* avulla. Faktorit ovat eräänlaisia selittäviä yhdistelmämuuttujia: Kun alun perin aineisto kuvataan kymmenen muuttujan avulla, on faktorianalyysin tavoitteena kuvata sama aineisto - tosin hieman epätarkemmin - esimerkiksi kahden faktorin avulla. Jokainen faktori liittyy aina vähintään yhteen muuttujaan jollakin voimakkuudella. Faktorit voidaan tämän ominaisuuden perusteella jakaa kolmeen luokkaan. Yleinen faktori liittyy kaikkiin muuttujiin. Spesifi faktori liittyy ainoastaan yhteen muuttujaan, ja ryhmäfaktori liittyy tiettyyn muuttujaryhmään. Faktorianalyysi tuottaa jokaiselle faktorille ns. *lataukset* kaikkien muuttujien suhteen, joiden voi kuvitella esittävän faktorin ja eri muuttujien välisiä korrelaatioita. Mitä pienempi on faktorin lataus jollekin muuttujalle, sitä vähemmän faktori selittää vaihtelusta kyseisellä muuttujalla. Latauksen neliö puolestaan kertoo, kuinka suuren osan muuttujan vaihtelusta faktori selittää. On tärkeää huomata, että vaikka faktoreita usein nimetäänkin, ei faktorianalyysi itsessään tee muuta kuin kuvaa aineiston rakennetta. Faktoreiden nimeäminen ja ratkaisun mielekkyyden tulkitseminen on aina ihmi-

sen tehtävä.

3.3 Transformaatiot

Testistä suoraan saatavia suorituspistemääriä kutsutaan raakapisteiksi. Raakapistemäärä ei kuitenkaan ole välttämättä kovin informatiivinen. Siksi raakapisteet muunnetaan usein jollekin toiselle asteikolle. *Transformaatiolla* tarkoitetaan testipistemäärien muuntamista jonkin säännön mukaan. Hyvä esimerkki tästä on Wechslerin älykkyystestien antama $\dot{A}O$. Testistä suoraan saatu raakapistemäärä muutetaan tietyn säännön mukaan $\dot{A}O$:ksi, joka on helposti tulkittavissa: On sovittu, että kun testipistemäärät transformoidaan $\dot{A}O$:ksi, niin keskimääräinen $\dot{A}O$ on terveellä aikuisella tasan 100. Usein testin raakapisteet täytyy muuntaa toiselle asteikolle nimenomaan vertailun mahdollistamiseksi. Esimerkiksi jos halutaan verrata henkilön testisuoritusta kahdella samaa ominaisuutta mittaavalla testillä, on molempien testien pistemäärät muunnettava samalle asteikolle ennen kuin ne ovat vertailukelpoisia. Kaikille transformaatioille ovat tyypillisiä seuraavat kolme ominaisuutta:

1. Transformaatio ei sinänsä muuta testipistemäärää, vaan ilmaisee sen vain eri yksikössä (esim. fahrenheit- ja kelvin-asteikot).
2. Transformaatio ottaa huomioon informaatiota, joka ei sisälly raakapistemäärään (esim. populaation keskiarvo, keskihajonta, tai itemien määrä).

3. Transformoitu testipistemäärä on informatiivisemmassa ja helpommin tulkittavassa muodossa kuin raakapistemäärä.

3.3.1 Mitta-asteikoiden invariantianssi

On huomattava, että eri mitta-asteikoilla mitattuja muuttujia voidaan transformoida eri tavalla. Tämä johtuu asteikoiden nollapisteidä ja havaintojen välimatkojen määrittelystä. Suhdeasteikko on *invariantti* eli muuntumaton kaikkien sellaisten muunnosten suhteen, jotka ovat muotoa $x' = cx$, missä x' on transformoitu muuttuja. Jos suhdeasteikollisen muuttujan kaikki arvot kerrotaan tai jaetaan vakiolla c , saatu muuttuja on myös suhdeasteikollinen, koska (1) muuttujan arvojen järjestys säilyy samana, (2) arvojen väliset suhteet säilyvät samana ja (3) nollapiste säilyy samana. Tällainen muunnos on esimerkiksi senttimetrin muuttaminen tuumiksi. Jos suhdeasteikko olisi invariantti myös toisenlaisten muunnosten suhteen, monet mittaustulokset olisivat erittäin kummallisia. Olisi mahdollista, että yksi ihminen todettaisiin neljäsosan toista pidemmäksi, jos mittaus suoritettaisiin tuumissa, mutta ainoastaan viidesosan pidemmäksi, jos mittaus suoritettaisiin senttimetreissä. Esimerkki varmasti selvittää tämän suhdeasteikon ominaisuuden tärkeyttä. Oletetaan, että suhdeasteikko olisi invariantti myös muotoa $x' = cx + b$ olevien muunnosten suhteen. Lisäksi sovitetaan, että tällaisessa järjestelmässä senttimetrit voidaan muuttaa tuumiksi kaavalla

$in = 2 * cm + 100$. Henkilöiden a ja b pituudet ovat 150 cm ja 200cm. Tällöin heidän senttimetreissä mitattujen pituuksiensa suhde on $\frac{3}{4}$. Vastaavasti heidän pituutensa tuumissa olisivat $400in$ ja $500in$, ja tuumissa mitattujen pituuksien suhde olisikin $\frac{4}{5}$!

Välimatka-asteikko poikkeaa suhdeasteikosta siinä, että välimatka-asteikolla ei ole absoluuttista nollapistettä. Välimatka-asteikolliset muuttujat ovat invariantteja sellaisten muunnosten suhteen, jotka ovat muotoa $x' = cx + b$. Muuttujan arvojen kertomisen lisäksi jokaiseen muuttujan arvoon voidaan lisätä jokin reaaliluku. Tällainen muunnos on esimerkiksi celsius -asteiden muuntaminen fahrenheit -asteiksi. Huomaa, että suhdeasteikko ei ole invariantti $x' = cx + b$ -muotoa olevien muunnosten suhteen, koska vakion b lisääminen muuttaisi asteikon nollapistettä. Esimerkiksi lämpötilan muunnos kelvineistä fahrenheitteiksi on muotoa $x' = cx + b$, ja se muuntaakin mitta-asteikon välimatka-asteikolle.

Järjestysasteikolla muunnosten tekemiseen ei voida käyttää mitään peruslaskutoimituksia. Ei ole yksinkertaisesti mielekäästä kertoa tai laskea yhteen järjestyksiä. Järjestysasteikollisille muuttujille voidaan tehdä ns. monotonisia transformaatioita, joiden käsittely tässä yhteydessä ei ole mahdollista. Nominaaliasteikollisten muuttujien tapauksessa transformaation käsite ei ole enää edes mielekäs.

3.3.2 Prosenttimuunnos

Kaikista yksinkertaisin transformaatio on testipistemäärän muunnos prosenteiksi. Siinä lasketaan kuinka monta prosenttia testat-

tavan pisteet ovat täysistä pisteistä. Prosenttimuunnoksen vahvuus on siinä, että prosenttiluvut ovat lähes kaikille ihmisille tuttuja ja siten niiden tulkitseminen on helppoa.

3.3.3 Aluemuunnos

Aluemuunnos on erittäin yleisesti käytetty transformaatio. Siinä raakapistemäärät muutetaan yksinumeroisiksi kokonaisluvuiksi käyttäen hyväksi normaalijakautuman persentiilejä. Tiettyä persentiiliä vastaa aina tietty kokonaisluku pienimmän persentiilin ollessa numero 1. Tämänkaltaisen muunnos korostaa yksilöiden välisiä eroja jakautuman keskiosassa, mutta pienentää niitä jakauman laidoilla. Niinpä aluemuunnetuilla pistemäärillä pystytään tarkastelemaan erityisesti keskimääräisesti suorituneiden ihmisten välisiä eroja.

3.3.4 Transformaatio z-pistemäärien avulla

Z-pistemäärämuunnos eli z-muunnos on eräs käytetyimpiä transformaatioita. Se on erittäin käyttökelpoinen, koska z-muunnos yhdenmukaistaa eri testien pistemäärät samalle skaalalle. Näin testipistemäärät ovat keskenään vertailukelpoisia. Z-pistemäärä on tyypillinen esimerkki ns. lineaarisesta transformaatiosta. Kuten edellä on todettu, z-muunnoksessa pistemäärien mittayksiköksi otetaan keskihajonta seuraavalla tavalla:

$$z = \left(\frac{x - \bar{x}}{s_x} \right)$$

z-pistemäärä saadaan siis, kun henkilön pistemäärästä vähennetään pistemäärien keskiarvo, ja jaetaan saatu luku keskihajonnalla. Z-pistemäärä ilmoittaa siis kuinka monen keskihajonnan päässä keskiarvosta henkilön testipistemäärä sijaitsee. Z-muunnoksen käyttö edellyttää, että transformoitava muuttuja on normaalisti jakautunut. Z-pistemäärät voidaan muuntaa taulukon (normitetun normaalijakauman tiheysfunktio) avulla persentiileiksi. Lisäksi z-pistemäärillä voidaan tehdä laskutoimituksia, esim. ottaa keskiarvo useasta z-arvosta, vaikkapa usean testin z-pistemääristä. Tarkastellaan seuraavia z-pistemääriä:

z=1. Henkilön pistemäärä sijaitsee yhden keskihajonnan päästä keskiarvosta. Z-pistemäärän mittayksikkönä on aina keskihajonta, ei pituus, älykyys tai sosiaalisuus. Tämän vuoksi esim. eri testeissä saadut pistemäärät ovat keskenään vertailukelpoisia, jos niille tehdään z-muunnos.

z=1.5. Voidaan todeta, että 93 prosenttia testin suorittaneista on saanut tätä henkilöä huonomman pistemäärän.

3.4 Normit

Testauksen tarkoituksena on usein verrata testattavan suoritusta muiden ihmisten vastaaviin suorituksiin. Tämän vuoksi testajan tulee tietää miten ihmiset yleensä suoriutuvat testistä. Normeilla tarkoitetaan sel-

laista testipistemäärästä muodostuvaa vertailuaineistoa, johon yksittäisen henkilön testipistemäärää voidaan verrata. Henkilön testipistemäärää arvioidaan siis *suhteessa normeihin*. Tämänkaltaista testipistemäärän tulkintaa kutsutaankin normipohjaiseksi tai normatiiviseksi tulkinnaiksi. Erilaisia normityyppejä ovat esim seuraavat:

1. persentiilinormit
2. ikänormit (esim. WISC-R)
3. luokkanormit (esim. erilaiset koulutestit)
4. testiprofiilit (esim. 16PF)

Normiaineisto on aina testikohtainen, ts. se kerätään jokaiselle testille erikseen. Tarkoituksena luonnollisesti on, että normiaineisto olisi mahdollisimman kattava: Sen tulisi olla edustava otos siitä populaatiosta johon testiä aiotaan soveltaa. Normien kerääminen onkin aloitettava määrittämällä testin kohderyhmä, ja tämän jälkeen on kerättävä kohderyhmää hyvin edustavat normit.

Jos normitetaan lasten älykkyystestiä, jolla on tarkoitus tutkia kaikkia suomalaisia lapsia, ei kannata kerätä normiaineistoa mittaamalla pelkästään Helsingin eliittikoulujen oppilaiden suorituksia. Näin on itse asiassa tehty kerran erään suositun älykkyystestin normittamisessa Suomessa. Testin pistemäärien oletettu keskiarvo oli 100, mutta käytännössä havaittiin, että esim. turkulaislasten keskiarvo pysytteli jatkuvasti noin 90:ssä. Väärästä normiaineistosta johtuen turkulaiset arvioitiin siis järjestelmällisesti heikommiksi kuin mitä he todellisuudessa olivat.

On huomattava, että edellämainittu Helsingistä kerätty normiaineisto olisi pätevä, jos testillä tarkasteltaisiin pelkästään helsinkiläisten koululaisten suorituksia; Nythän normiaineisto olisi edustava otos siitä ihmisjoukosta (helsinkiläiset lapset), johon testiä ja normeja aiotaan soveltaa.

Kuten edellisestä varmaankin käy ilmi, normatiivisia ryhmiä voi olla hyvin monia erilaisia. Normit voidaan kerätä erikseen miehille ja naisille, tai erikseen eri koulutustason omaaville ihmisille. Edelleen normit voidaan kerätä paikallisesti, eli esimerkiksi turkulaisiin ja tamperelaisiin lapsiin sovellettaisiin eri normeja. Normit voidaan kerätä myös valtakunnallisesti, jolloin kaikkien suomalaisten tuloksia verrattaisiin samoihin normeihin. Eräs mahdollisuus normiaineiston keräämiseksi on myös yksinkertaisesti täydentää normeja sitä mukaa kun testiä käytetään, ts. jokaisen testin tehneen henkilön pistemäärä liitetään normiaineistoon.

3.4.1 Mahdollisia ongelmia normatiivisessa tulkinnessa

Normiaineiston laadun ja koon on oltava riittäviä. Voi olla, että jokin normatiivinen ryhmä ei olekaan riittävän edustava otos. Normiaineistoa laadittaessa ei tällöin otettu huomioon kaikkia niitä ihmisryhmiä, joihin testiä aiotaan soveltaa. Paikallisilla normeilla on kuitenkin mahdollista paikata tämänkaltaista ongelmaa. Tällöin on kerättävä uusi normiaineisto sitä ryhmää varten, jolta se puuttuu alkuperäisestä normiaineistosta.

Ikänormien käyttö saattaa helposti joutaa ylitulkintaan. Jos 22-vuotiaan henkilön älykkyystestissä saaman testipistemäärän havaitaan vastaavan keskimääräisen 16-vuotiaan testipistemäärää, voidaan ajatellaan testattavan toimivan 16-vuotiaan ikäta-solla myös muilla alueilla kuin varsinaisesti testatulla. Tämä johtopäätös ei enää välttämättä olekaan pätevä.

Normit vanhenevat usein nopeasti. Esim. ihmiset saavat tällä hetkellä WISC-R:sta keskimäärin korkeampia raakapistemääriä kuin 20 vuotta sitten. Jos tälläkin hetkellä käytettäisiin vanhaa normiaineistoa, ei testattujen henkilöiden ÄO:n keskiarvo olisi-kaan enää 100 vaan vähän enemmän. Koska normeeraus vie paljon aikaa, käytössä olevat normit eivät käytännössä ole koskaan ajan tasalla. Normeja on kuitenkin mahdollista täydentää jatkuvasti lisäämäl-lä jokaisen testin tehneen henkilön tulok-set normiaineistoon. Tämänkaltaisen me-nettely on mahdollinen sellaisissa paikois-sa, joissa testeillä testataan paljon ihmisiä. Tämä tietenkin edellyttää sitä, että testat-tava ryhmä on edustava otos siitä populaa-tiosta, johon normeja aiotaan soveltaa.

Jos normatiivinen ryhmä on pieni, niin tulkintavirheen mahdollisuus kasvaa. Esim. MMPI:ssa muutamien häiriöluokkien nor-mipopulaatio on huomattavan pieni, joten ei voida olla täysin varmoja, antaako tuo ryhmä edustavan kuvan ko. häiriön omaa-vien ihmisten suoriutumisesta testissä.

Normien käyttötarkoitus myös määrää, millaisia tulkintoja normien avulla voidaan tehdä. Esimerkiksi WAIS -testi on laadit-tu erottelemaan ihmisiä, joiden ÄO on *pie-nempi* kuin keskiarvo sata. Siten ei ole

mielekästä määritellä jotain ihmistä erittäin älykkääksi vaikka hän saisi WAIS:sta ÄO:ksi huomattavsti yli sata pistettä.

Osa II

Psykologisten testien teoria

Luku 4

Testiteoriat

Tässä monisteessa laajimmin läpikäytävä psykologisten testien ja psykologisen mittaamisen teoria on ns. *klassinen testiteoria*, toiselta nimeltään rinnakkaisten mitausten teoria. Vaikkakin mahdollisuuksiltaan hieman rajoitettu, klassinen testiteoria on kaikkein selkein psykologisten mittaustulosten arviointia käsittelevä teoria. On olemassa myös muita testiteorioita, jotka esitellään tässä yhteydessä vain lyhyesti. Näitä testiteorioita kutsutaan yhteisnimellä *modernit testiteoriat*. Kaikki tässä monisteessa myöhemmin esitetyt tarkastelut perustuvat nimenomaan klassiseen testiteoriaan. Muut testiteoriat käsitellään lähinnä esittelymielessä.

4.1 Klassinen testiteoria

Klassisessa - ja myös muissa testiteoriassa - mittauksen laatimiseen ja arviointiin liittyvät ongelmat voidaan jakaa neljään osaluokkaan:

1. Testi-itemien suunnittelu
2. Mittauksen suorittaminen

3. Mittaustarkkuuden arvioiminen

4. Mittaustulosten ennustearvon määrittäminen

Itemien suunnittelua, mittauksen suorittamista ja ennustearvon määrittämistä käsitellään *itemianalyysin* ja *validiteettitarkastelun* yhteydessä. Mittaustarkkuuden arviointia taas käsitellään *reliabiliteetin* yhteydessä.

Klassinen testiteoria sisältää tiettyjä implisiittisiä oletuksia, jotka on otettava huomioon jo ennen mittauksen suorittamista. Ensinnäkin, tutkittava mittari mittaa yhtä ominaisuutta kerrallaan. Toiseksi, mittaus tapahtuu vähintään välimatka-asteikolla. Kolmanneksi, mitta-asteikko oletetaan jatkuvaksi. Tästä seuraa, että klassisen testiteorian avulla voidaan käsitellä ainoastaan psykometrisia testejä. Esimerkiksi Ro:lla tehtyjä mittauksia ei voida tarkastella klassisen teorian avulla, koska testin käyttämä mitta-asteikko ei ole välimatka- tai suhteasteikko.

Klassisen testiteorian lähtökohta on, että lähes kaikki mittaaminen on jossain määrin virheellistä. Henkilön testissä saamaan pistemäärään vaikuttaa kaksi eri tekijää: itse mitattava ominaisuus ja mittausvirhe. Mittauksen tuloksena ei saada täsmälleen oikeaa tulosta, vaan jokin tekijä mittauksessa tuottaa virhettä tulokseen. Niinpä esim. WAIS-R:lla saatu henkilön älykkyysosamäärä ei ole täsmälleen hänen oikea älykkyysosamääränsä, vaan se sisältää mm. itse testistä ja testausilanteesta johtuvaa *satunaisvirhettä*.

Psykometriikan keskeisimpiä tutkimuskohteita on mittausvirheen tarkasteleminen. Suuri osa seuraavassa esitetystä liittyy toisaalta mittausvirheen suuruuden määrittämiseen, ja toisaalta keinoihin, joilla mittausvirheen suuruutta voidaan pienentää. Aluksi on kuitenkin tehtävä joitakin oletuksia tarkastelun pohjaksi. Ensinnäkin, jotta mittausvirhettä voidaan käsitellä ja tutkia, se on erotettava omaksi kokonaisuudekseen. Lisäksi sovitaan, että *havaitulla pistemäärällä* tarkoitetaan sitä pistemäärää, jonka testi antaa. Havaittu pistemäärä sisältää aina jonkin verran mittausvirhettä. *Todellisella pistemäärällä* taas tarkoitetaan henkilön todellista testissä suoriutumisen tasoa, johon mahdolliset mittausvirheet eivät ole vaikuttaneet. Klassisessa testiteoriassa havaitun pistemäärän oletetaan koostuvan todellisesta pistemäärästä ja mittausvirheestä, eli toisin ilmaistuna

$$X = T + e$$

missä

$$X = \text{havaittu testipistemäärä}$$

$$T = \text{todellinen pistemäärä ja}$$

$$e = \text{mittausvirhe}$$

Havaittua testipistemäärää ei siten käsitellä yhtenä kokonaisuutena, vaan se jaetaan kahteen erilliseen osaan, eli sovitaan että $X = T + e$.¹ Tässä yhtälössä on kaksi tuntematonta: T ja e . Jos henkilön havaittu pistemäärä tunnetaan, ei yhtälöstä voi ratkaista hänen todellista pistemääräänsä tai mittausvirhettä. Periaatteessa on mahdollista selvittää henkilön todellista pistemäärää ainoastaan yhden mittauksen perusteella. Mittauksesta saadaan ainoastaan havaittu pistemäärä, joka sisältää sekä todellisen pistemäärän että osan mittausvirhettä. Jotta todellista pistemäärää pystyttäisiin arvioimaan, on pyrittävä arvioimaan mittausvirheen osuutta havaitussa pistemäärässä. Yhden mittauksen perusteella ei kuitenkaan voida sanoa mitään mittausvirheen suuruudesta. Mittausvirhe ja todellinen pistemäärä eivät ole tällä tavoin erotettavissa toisistaan.

Klassinen testiteoria tarkastelee mittausvirhettä rinnakkaisten mittausten avulla, ts. käyttämällä kahta erilaista mittaria mittaamaan samaa ominaisuutta. Näistä mittalaitteista käytetään nimitystä *rinnakkaismittarit*. Rinnakkaismittarit määritellään siten, että kahta mittaria voidaan pitää toisilleen rinnakkaisina, jos mittaustuloksen kannalta ei ole merkitystä kumpaa niistä käytetään, eli

$$T_{ig} = T_{ih}$$

¹Merkintöjä X , T ja e käytetään yleisesti indeksoinneissa, esim. σ_X^2 tarkoittaa nimenomaan havaitun pistemäärän X varianssia

4.2 Klassisen testiteorian perusoletukset

Klassinen testiteoria perustuu kuudelle perusoletukselle. Nämä ovat luonteeltaan määritelmiä, eikä niitä voida osoittaa empiirisesti oikeiksi tai vääriksi. Klassisen testiteorian kuusi perusoletusta ovat:

1. Havaittu pistemäärä X koostuu todellisesta testipistemäärästä T ja mittausvirheestä e .
2. Virhekomponenttien² keskiarvo on nolla, jolloin mittausvirhe on harhaton eli se sekä laskee että nostaa havaittuja pistemääriä.
3. Mittausvirhe ja todellinen testipistemäärä ovat toisistaan riippumattomia.
4. Rinnakkaismittausten virhekomponentit ovat korreloimattomia.

Lisäksi rinnakkaismittarit määritellään seuraavasti:

5. Jokainen rinnakkaismittari antaa henkilölle saman todellisen pistemäärän.
6. Mittauksen keskivirhe on jokaisessa rinnakkaismittarissa sama.

Oletukset voidaan esittää myös lyhyemmässä muodossa käyttäen edellä esitettyjä merkintöjä. Huomaa, että seuraava esitystapa tarkoittaa täsmälleen samaa asiaa kuin

²virhekomponentilla tarkoitetaan mittausvirheen osuutta jossakin havaitussa pistemäärässä.

edellinenkin esitys:

1. $X = T + e$
2. $\bar{x}_e = 0$
3. $r_{Te} = 0$ ja $r_{Tge_h} = 0$
4. $r_{ege_h} = 0$
5. $T_{ig} = T_{ih} = \dots = T$
6. $SEM_g = SEM_h = \dots = SEM$

Perusoletuksista ensimmäinen merkitsee siis sitä, että kaikki mittaaminen on virheelistä, eli harhatonta mittausta on mahdoton suorittaa. Teoreettisesti havaittu testipistemäärä X voidaan kuitenkin purkaa todelliseksi (harhattomaksi) pistemääräksi T ja mittausvirheeksi e .

Virhekomponenttien keskiarvo oletetaan nolaksi sen vuoksi, että tällöin virheet eivät systemaattisesti vaikuta testipistemääriin. Jos virhekomponenttien keskiarvo olisi nollaa suurempi, virhekomponentit nostaisivat säännöllisesti testipistemääriä ja henkilöt saisivat testistä systemaattisesti liian korkeita pistemääriä. Sama pätee tietysti toiseen suuntaan: jos virhekomponenttien keskiarvo olisi nollaa pienempi, niin henkilöt saisivat säännöllisesti testistä liian pieniä pistemääriä.

Mittausvirheen ja testipistemäärän korreloimattomuus tarkoittaa siis sitä, että mittausvirheen ja testipistemäärän välillä ei ole suoraviivaista yhteyttä. Jos yhteys olisi olemassa, niin korrelaation ollessa positiivinen virhekomponentti kasvaisi suuressa testipistemäärällä, ts. suuret testipistemäärät olisivat enemmän virheellisiä kuin pienet. Vastaavasti korrelaation ollessa negatiivinen virhekomponentti pienenesi pie-

nillä testipistemäärillä, ts. suuret testipistemäärät olisivat vähemmän virheellisiä kuin pienet. Koska mittausvirhe kuitenkin oletetaan riippumattomaksi testipistemäärästä, ovat sekä suuret että pienet testipistemäärät yhtä virheellisiä tai virheettömiä. Myös rinnakkaismittarien virhekomponenttien oletetaan olevan korreloimattomia, eli toisen mittarin mittausvirhe ei ole systemaattisesti yhteydessä toisen mittarin virheeseen.

Rinnakkaismittareihin liittyy vielä kaksi määrittelyä. Ensinnäkin, mittauksen keskivirhe on jokaisessa mittarissa sama. Tämä tarkoittaa sitä, että mittausvirheen vaikutus testipistemääriin on samansuuruinen kaikissa rinnakkaismittareissa. Toiseksi, jokaisen rinnakkaismittarin oletetaan antavan henkilölle saman todellisen pistemäärän. Tämä siis tarkoittaa, että molemmat rinnakkaismittarit mittaavat samaa ominaisuutta samalla tavalla. Tämä on tärkeä määritelmä, sillä koko rinnakkaismittarien käyttäminen perustuu tähän: Se nimittäin selittää, miksi henkilö voi saada kahdella rinnakkaismittarilla eri pistemäärän. Koska oletuksen 5. mukaan todellinen pistemäärä on molemmissa rinnakkaismittauksissa sama, ei eron syy voi olla muu kuin *mittausvirhe*.

4.2.1 Mittauksen arvioiminen klassisen testiteorian avulla

Klassisessa testiteoriassa mittarin toimivuutta arvioidaan yleensä reliabiliteetti- ja validiteettikertoimien avulla. *Reliabiliteettikerroin* ilmaisee kuinka paljon todelliset pistemäärät selittävät havaittujen pistemää-

rien varianssista. Se on määritelmän mukaan havaitun ja todellisen pistemäärän välisen korrelaation neliö R_{XT}^2 , joka puolestaan on yhtä suuri kuin mittarin todellisen ja havaitun varianssin suhde $\frac{\sigma_T^2}{\sigma_X^2}$. On kuitenkin syytä huomata, että näiden määritelmien mukaan reliabiliteettikertoimen määrittäminen on mahdotonta yhden mittauksen havaittujen pistemäärien perusteella, koska todellisia pistemääriä ei tunneta.

Validiteettikertoimella puolestaan tarkoitetaan testipistemäärän ja sellaisen kriteerimuuttujan korrelaatiota, jonka tiedetään mittaavan hyvin testin mittaamaa piirrettä. Huomionarvoista onkin, että kriteerimuuttujana voidaan käyttää periaatteessa mitä hyvänsä muuttujaa, jonka valinnalle voidaan esittää tarpeeksi painavat perustelut.

Tässä esitetty johdanto klassiseen testiteoriaan on melko tiivis paketti, eikä missään nimessä täydellinen esitys. Esitettyjen oletusten perusteella on mahdollista johdattaa runsaasti erilaisia mittausten käsittelyssä tarvittavia lauseita, joista tärkeimmät käsitellään seuraavissa luvuissa. Omien mittausten laatiminen ja arvioiminen ei onneksi edellytä kovin monimutkaisten menetelmien tuntemista, ja perustiedoilla selviääkin melko pitkälle. Kattava esitys klassisesta testiteoriasta löytyy esimerkiksi Harold Gulliksenin (1950) teoksesta *Theory of Mental Tests*.

4.3 Muita todellisen pistemäärän teorioita

On olemassa myös joukko muita testiteorioita, jotka perustuvat oletukseen todellisesta pistemäärästä. Nämä ovat kaikki jollain tavalla klassisen testiteorian kehitelmiä, joissa rinnakkaismittarien määrittely poikkeaa klassisen teorian määrittelystä. Täten klassisen testiteorian tunteminen riittää useimmissa tapauksissa aivan mainiosti. Seuraavassa on kuitenkin lyhyt esittely muista todellisen pistemäärän testiteorioista.

4.3.1 Tauekvivalenttien ja olennaisesti tauekvivalenttien mittareiden teoria

Tauekvivalenttien ja olennaisesti tauekvivalenttien mittarien teoria (l. ekvivalenttien mittarien teoria) on oletuksiltaan klassista teoriaa väljempi, joten se soveltuu käytettäväksi useammissa tilanteissa. Ekvivalenttien mittarien teoria sisältää erilaiset oletukset rinnakkaismittareista. Klassisen testiteorian oletukset 1-4 ovat voimassa myös ekvivalenteille mittareille. Mittarit määritellään *tauekvivalenteiksi*, jos niiden mitaama todellinen pistemäärä on sama, mutta mittausvirheet erisuuruisia. *Oleellisesti tauekvivalentit mittarit* ovat sellaisia, joissa ensimmäisen mittarin antama todellinen pistemäärä saadaan, kun toisen mittarin antamaan todelliseen pistemäärään lisätään jokin vakio. Mittausvirheiden ei edelleenkään oleteta olevan samansuuruisia. Tällä tavoin väljentämällä rinnakkaismitta-

reilta edellytettäviä ominaisuuksia saadaan teoria, jonka soveltaminen on mahdollista useammissa käytännön tilanteissa. Tästä tosin seuraa, että monet tunnusluvut kuten reliabiliteettikerroin on määriteltävä eri tavoin kuin klassisessa testiteoriassa, koska rinnakkaismittarien todelliset pistemäärät ja mittausvirheet eivät olekaan enää samansuuruisia.

4.3.2 Kongeneeristen mittareiden teoria

Kongeneeristen mittausten teoriassa ovat myöskin voimassa klassisen testiteorian neljä ensimmäistä perusoletusta. Mittarit määritellään kuitenkin eri tavalla. Määrittelyn mukaan kaksi mittaria ovat kongeneerisiä, jos mittarin a todellinen pistemäärästä saadaan mittarin b todellisesta pistemäärästä jollain lineaarisella muunnoksella. Mittausten keskivirheet voivat olla erisuuruisia eri mittareissa. Reliabiliteettikertoimen tarkka määrittäminen ei nyt ole mahdollista, sillä mittausvirheen lisäksi todellinen pistemäärä vaihtelee eri mittareiden välillä. Kongeneeristen mittareiden tapauksessa onkin tyydyttävä arvioimaan reliabiliteettikertoimen alaraja. Tätä alarajaa voidaan soveltaa mittausvirheen arviointiin samalla tavalla kuin reliabiliteettikerrointakin.

4.4 Yleistettävyysteoria

Yleistettävyysteoria on eräällä tavalla klassisen testiteorian ja sen kehitelmien laajen-

nus. Yleistettävyysteoria³ luopuu klassisen teorian rinnakkaisten mittarien käsitteestä. Erilaisia mittariversioita ei oleteta toisilleen rinnakkaisiksi vaan yleistettävyysteoriassa oletetaan, että kukin mittari on *otos* kaikista kuviteltavissa olevista mittariversioista, eikä mittarien virhevariansseille aseteta yhtä tiukkoja oletuksia kuin klassisessa teoriassa. Tästä syystä tällaisia mittareita kutsutaan myös *nimellisesti rinnakkaisiksi*. Jokainen nimellisesti rinnakkainen mittari sisältää omalta osaltaan totuudenmukaisen arvion mitattavasta ominaisuudesta, mutta havaittava mittaluku on mittausvirheen johdosta epätarkka. Yhden mittarin mitaustulosten odotusarvoa nimitetään *spesifiksi todelliseksi mittaluvuksi*. Spesifi todellinen mittaluku siis saadaan, kun yksi henkilö tekee saman testin äärettömän monta kertaa, ja testipistemääristä lasketaan keskiarvo. Tämä kuvaa siis henkilön todellista suoriutumista tällä nimenomaisella mittariversiolla. Totuudenmukainen arvio mitattavasta ominaisuudesta sisältyy kuitenkin vasta *kaikkien* kuviteltavissa olevien mittarien antamien tulosten perusjoukkoon. Voidaankin ajatella, että yksittäiset testit ovat ikäänkuin otos kaikkien mahdollisten testien populaatiosta. Näiden kaikkien mahdollisten testien pistemäärien odotusarvoa kutsutaan *geneeriseksi todelliseksi mittaluvuksi* tai *universumpistemääräksi* μ_p . Tämän voidaan siis ajatella olevan analoginen klassisen testiteorian todellisen pistemäärän kanssa.

Yleistettävyysteoria ja klassinen testiteoria eroavat toisistaan myös virheen mää-

rittelyssä. Klassinen teoria olettaa kaiken virheen olevan *satunnaisvirhettä*, kun taas yleistettävyysteoria ottaa satunnaisvirheen lisäksi huomioon myös muut mahdolliset virheet kuten *systemaattisen virheen*. Tämä siis tarkoittaa sitä, että erilaiset mittaustilanteesta johtuvat tekijät kuten testaava psykologi saattavat aiheuttaa virhettä mitaustuloksiin. Yleistettävyysteorian avulla voidaan siis arvioida *missä tilanteissa* erilaiset testit ovat reliaabeleja. Yleistettävyysteoriassa reliabiliteetti riippuu siis siitä, millaisia *yleistyksiä* mittaustulosten perusteella aioitaan tehdä.

4.4.1 Reliabiliteetin arvioiminen yleistettävyysteorian näkökulmasta

Klassisen testiteorian näkökulmasta on mahdollista laatia täydellinen mittari: jos $e = 0$, niin silloin mittarin havaitut pistemäärät vastaavat todellisia pistemääriä, ja mittari mittaus on täysin virheetön. Mittaustulostenä käsitellään kuitenkin ainoastaan satunnaisvirhettä, joten virheettömän mittarin määritelmä on ongelmallinen. Toisin kuin klassisessa teoriassa, yleistettävyysteoriassa pyritään arvioimaan erikseen jokaisen virhelähteen vaikutusta mittaustuloksiin. Tämä tehdään *varianssianalyysin* avulla. Ensimmäiseksi määritellään *yleistettävyyysuniversumi* eli ne tilanteet, joihin mittaustulos halutaan yleistää. Tämän jälkeen määritellään kaikki mahdolliset virhetekijät, jotka voivat vaikuttaa henkilön mittaustulokseen. Tällaisia voivat olla esimerkiksi mittaajajankohda ja testaava psy-

³ks. esim. Cronbach (1970)

kologi. Sama testi voidaan tehdä esimerkiksi kahtena eri ajankohtana, ja molemmilla kerroilla kahden eri psykologin toimesta. Jokaiselta henkilöltä saadaan siten neljä mittaustulosta. Tämän jälkeen testituloksia voidaan tarkastella varianssianalyysin avulla. Päävaikutukset kertovat virhetekijöiden voimakkuuksista. Psykologi -muuttujan päävaikutus voidaan tulkita siten, että eri psykologit antavat testistä erilaisia pistemääriä ja tuottavat siten systemaattista virhettä. Jos testauskerta -muuttujalle ei tule päävaikutusta, voidaan testitulosten katsoa olevan yleistettävissä eri mittauskertoihin, mutta ei eri psykologien tekemiin testauksiin.

4.5 Latentin piirteen teoria

Latentin piirteen teoria poikkeaa todellisen pistemäärän teorioista ja yleistettävyysteorioista siinä, että latentin piirteen teoria ei tee oletusta todellisen pistemäärän olemassaolosta. Latentin piirteen teoria sisältää oletukset *latentista piirteestä*, *itemien ominaisfunktioista* sekä *lokaalista riippumattomuudesta*.

4.5.1 Latentti piirre

Latentin piirteen teoriassa tarkastelun lähtökohtana ovat itemien pistemäärät. Havaitut pistemäärät oletetaan yhden latentin piirteen Θ (Theta) ilmentymisiksi. Tällaisessa lähestymistavassa mittaustulokset oletetaan tulokseksi kulloinkin mitattavan

latentin piirteen ja mittarin välisestä vuorovaikutuksesta. Latenttia piirrettä ei voida suoraan havaita, mutta sen olemassaolosta voidaan tehdä päätelmiä, kun koehenkilö tuottaa vastauksen testin jossain itemissä. Vastauksen oletetaan heijastavan jotain sekä latentista piirteestä, että itemistä itsestään. Latentin piirteen ei suoraan oleteta määräävän henkilön vastausta itemiin, se ainoastaan määrittää millä *todennäköisyydellä* henkilö vastaa tietyllä tavalla. Henkilön vastaus itemiin riippuu siten latentista piirteestä ja mittausvirheestä.

4.5.2 Itemin ominaisfunktio

Latentin piirteen teoria olettaa, että jokaisen itemivastauksen todennäköisyys riippuu latentista piirteestä funktionaalisesti. Jokaiselle itemille on määriteltävissä ns. ominaisfunktio. Ominaisfunktion ja latentin piirteen voimakkuus yhdessä määräävät henkilön itemin vastauksen todennäköisyyden. Tämä siis tarkoittaa sitä, että henkilön vastaus itemiin riippuu sekä itemistä että latentista piirteestä. Yksilön mitattavan piirteen ja itemin vastauksen välinen yhteys esitetään tyypillisesti ns. itemin ominaiskäyränä, joka kuvaa henkilön oikein vastaamisen todennäköisyyttä. Tarkoituksena on luoda malli siitä, millä tavoin eri määrän piirrettä omaavat henkilöt vastaavat kuhunkin testi-itemiin. Eri tilanteissa tarvitaan hieman toisistaan poikkeavia malleja. Käytettävä malli riippuu itemin mitta-asteikosta, kuhunkin itemin ominaisfunktioon tulevista parametreista ja käytettävästä itemin ominaisfunktioista. Mallien muodostamista ja erilaisia malleja ei kä-

sitellä tässä yhteydessä sen tarkemmin, lisää aiheesta voi lukea esimerkiksi Konttisen (1980) teoksesta *Testiteoria*.

4.5.3 Lokaali riippumattomuus

Edellisten lisäksi latentin piirteen teoriassa oletetaan, että mittarin itemit ovat toisistaan riippumattomia. Käytännössä tämä merkitsee sitä, että henkilön vastaus yhteen itemiin ei vaikuta hänen vastauksiinsa muisa itemeissä ja toisaalta sitä, että vastauksiin vaikuttaa ainoastaan yksi latentti piirre. On tietenkin mahdollista muodostaa myös sellaisia malleja, joissa vastauksiin vaikuttaa useampi latentti piirre, mutta tällaisissa malleissa täytyy tehdä hieman toisenlaisia oletuksia. Yleensäkin kannattaa huomata, että latentin piirteen teoria soveltaminen käytäntöön edellyttää kohtuullisen monimutkaisia matemaattisia menetelmiä ja erilisiä ohjelmistoja, joten tämänkaltaisen testien käsittely vaatii melkoista paneutumista aiheeseen. Latentin piirteen teoria on kuitenkin erittäin tehokas ja monipuolinen menetelmä mittaustulosten arviointiin ja käsittelyyn, ja sitä voidaankin suositella käytettäväksi silloin, kun laaditaan laajempia ja monimutkaisempia mittareita.

4.5.4 Latentin piirteen teorian sovelluksia

Latentin piirteen teoria tarjoaa monia lupaavia käytännön sovelluksia, mutta ongelmana on tällaisten mallien laadinnassa vaadittavat matemaattiset valmiudet. Yleisluontoisesti voidaan todeta, että ainakin

seuraavat menettelyt mahdollistuvat latentin piirteen teorian avulla: Ensinnäkin, monivalintatehtävissä on mahdollista käyttää erottelukriteerinä myös väärää vastausvaihtoehtoja pelkän oikein-väärin luokittelun sijasta. Malli siis kertoo myös sen, millä todennäköisyydellä eri tasoiset henkilöt vastaavat väärin eri kysymyksiin.

Tämä taas mahdollistaa myös yksilöllisen testaamisen esim. tietokoneen avulla. Testin alkupään vastausten pohjalta tietokone voi estimoida yksilön kyvykkyyden ja valita seuraavaksi sopivan vaikeusasteen omaavia itemejä - ei liian vaikeita mutta ei liian helppojakaan. Tällä tavoin menettäessä testi lyhenee. Koska testattavan ei tarvitse vastata kohtuuttoman vaikeisiin tai aivan liian helppoihin kysymyksiin, arvio hänen suoriutumisestaan myöskin tarkentuu huomattavasti.

Luku 5

Reliabiliteetti

Testin reliabiliteetin avulla voidaan arvioida, kuinka paljon mittausvirhettä mitaustulos sisältää. Tässä yhteydessä reliabiliteetti tarkoittaa sitä, kuinka johdonmukaisesti testi toimii ja kuinka vähän mittausvirhettä testipistemääriin sisältyy. Täysin virheettömän mittarin laatiminen on käytännössä mahdotonta. Mittauksen virheellisyys ei kuitenkaan ole niin suuri ongelma kuin miltä se aluksi kuullostaa. Vaikka mittausvirhettä ei saataisikaan nollaksi, klassisesta testiteoriasta voidaan johtaa useita apuvälineitä, joiden avulla mittausvirheen vaikutusta mitaustuloksiin pystytään arvioimaan.

Klassisessa testiteoriassa reliabiliteettitarkastelun lähtökohtana on, että testipistemäärään vaikuttaa kaksi tekijäryhmää:

1. Tekijät, jotka myötävaikuttavat korkeaan yhdenmukaisuuteen eri mittauskertojen välillä. Tällaisia ovat yksilön pysyvät ominaisuudet, esim. verbaalinen älykyys tai introversio. Nämä tekijät vaikuttavat siis henkilöön *todelliseen*

liseen pistemäärään T .

2. Tekijät, jotka alentavat yhdenmukaisuutta. Näihin kuuluvat kaikki sellaiset satunnaistekijät, joilla ei ole tekemistä mitattavan piirteen kanssa, ts. nämä tekijät vaikuttavat *mittausvirheen* e suuruuteen. On myös hyvä huomattava, että mittaukseen saattaa myös liittyä systemaattista virhettä, esimerkiksi eri psykologit voivat käyttää järjestelmällisesti väärin jotain testiä. Klassisessa tällainen virhe oletetaan korjattavaksi siten, että testituloksiin lisätään virheen suuruinen vakio. Systemaattisen virheen erottamiseksi satunnaisvirheestä täytyykin käyttää esimerkiksi yleistettyä vyysteoriaa.

Virhelähteen ja mitattavan ominaisuuden vaikutus testipistemääriin on todettu toisessa muodossa jo aiemmin. Klassisessa testiteoriassahan mitatun pistemäärän oletetaan

taan siis koostuvan todellisesta pistemäärästä ja mittausvirheestä, eli toisin ilmaistuna $X = T + e$. Tämä yhtälö voidaan reliabiliteettitarkastelua varten saattaa myös muotoon $e = X - T$, eli mittausvirhe on testipistemäärän ja todellisen pistemäärän välinen erotus. Reliabiliteettia määritettäessä arvioidaan siis nimenomaan *mittausvirheen* suuruutta. Klassinen testiteoria olettaa, että mittausvirhe jakautuu satunnaisesti suuressa populaatiossa (oletukset 2-4). Näiden ja rinnakkaismittareihin liittyvien oletusten (5-6) avulla onkin mahdollista aloittaa testin reliabiliteetin estimointiin käytettyjen menetelmien tarkasteleminen. Tarvitaan kuitenkin vielä pari lisätyökalua, joista lisää seuraavassa.

5.1 Reliabiliteettikerroin

Mittausvirheen ja testitulosten johdonmukaisuuden arviointiin on useita erilaisia menetelmiä. Näistä kaikista yksinkertaisin on ns. reliabiliteettikerroin. Havaitut testipistemäärät sisältävät lähes aina vaihtelua, *varianssia*. Havaittu pistemäärään koostuu todellisesta pistemäärästä ja mittausvirheestä. Onkin luonnollista olettaa, että sekä todellinen pistemäärä että mittausvirhe vaihtelevat eri henkilöiden välillä. Ensimmäiseksi todetaan, että mittausvirhe vaikuttaa määritelmistä johtuen yksittäisiin pistemääriin, mutta ei pistemäärien keskiarvoihin! Voidaan nimittäin osoittaa, että

$$\bar{x}_T = \bar{x}_X \quad (5.1)$$

eli todellisten pistemäärien keskiarvo on yhtä suuri kuin havaittujen pistemäärien

keskiarvo. Tämä voidaan todistaa seuraavasti:

$$\bar{x}_X = \frac{\sum X}{n} = \frac{\sum (T_i + e_i)}{n} = \bar{x}_T + \bar{x}_e = \bar{x}_T$$

Todistuksen idea on siinä, että havaitut pistemäärät muutetaan ensin muotoon $T + e$. Tämän jälkeen huomataan, että havaittujen pistemäärien keskiarvo muodostuu todellisten pistemäärien keskiarvosta $\bar{x}_T$ ja virhekomponenttien keskiarvosta $\bar{x}_e$. Koska oletuksen 2 nojalla virhekomponenttien keskiarvo on 0, niin $\bar{x}_T = \bar{x}_X$. Täten testillä voidaan mitata jonkin tietyn ryhmän suoriutumista, ja testipistemäärien keskiarvo on harhaton todellisten pistemäärien estimaatti. Vaikka testi ei siis olisikaan reliabeli, sillä voidaan arvioida todellisten pistemäärien keskiarvoa

Reliabiliteettikertoimen määritelmä perustuu oletukseen, jonka mukaan havaittujen pistemäärien varianssi on todellisen pistemäärän varianssin ja virhevarienssin summa, eli

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2 \quad (5.2)$$

missä:

$$\begin{aligned} \sigma_X^2 &= \text{havaittujen pistemäärien} \\ &\quad \text{varienssi} \\ \sigma_T^2 &= \text{todellisten testipistemäärien} \\ &\quad \text{varienssi} \\ \sigma_e^2 &= \text{virhevarienssi} \end{aligned}$$

Tämän todistaminen on hieman monimutkaisempaa, mutta edelleen on tärkeintä ymmärtää ja muistaa todistuksen lopputulos. Itse todistus esitetään lähinnä kuriosio-

teettina:

$$\begin{aligned}
 \sigma_x^2 &= \frac{\sum (X_i - \bar{x}_X)^2}{n} \\
 &= \frac{\sum (T_i + e_i - (\bar{x}_T + \bar{x}_e))^2}{n} \\
 &= \frac{\sum [(T_i - \bar{x}_T) + (e_i - \bar{x}_e)]^2}{n} \\
 &= \frac{\sum (T_i - \bar{x}_T)^2}{n} + \frac{\sum (e_i - \bar{x}_e)^2}{n} + \\
 &\quad 2 \frac{\sum (T_i - \bar{x}_T)(e_i - \bar{x}_e)}{n} \\
 &= \sigma_T^2 + \sigma_e^2 + 2\sigma_T\sigma_e r_{Te}
 \end{aligned}$$

Todistus lähtee liikkeelle samalla tavalla kuin edellä. Kirjoitetaan ensin havaittujen pistemäärien varianssi σ_X^2 tavallista varianssin kaavaa hyväksi käyttäen. Seuraavaksi esitetään havaittu pistemäärä X_i todellisen pistemäärän T ja virheen e summana, ja vastaavasti havaittujen pistemäärien keskiarvo $\bar{x}_X$ todellisten pistemäärien keskiarvon $\bar{x}_T$ ja mittausvirheen keskiarvon $\bar{x}_e$ summana. Kirjoittamalla lauseke toisessa muodossa havaitaan, että havaittujen pistemäärien varianssi σ_X^2 todella muodostuu todellisten pistemäärien varianssin σ_T^2 ja virhevarienssin σ_e^2 summana. Jäljelle jäävä osa $2\sigma_T\sigma_e r_{Te}$ sisältää todellisten pistemäärien ja mittausvirheen välisen korrelaation. Koska määritelmän 3 nojalla tämä on 0, niin lausekkeen viimeinen osakin on 0.

Testipistemääriin vaikuttaa siten kaksi tekijää: vaihtelut todellisissa pistemäärissä ja vaihtelut mittausvirheessä. Tämän perusteella voidaan laskea ns. *reliabiliteettikerroin*. Reliabiliteettikerroin ilmoittaa todellisen varianssin ja mittausvirheen varianssin

välisen suhteen, eli kuinka suuri osa havaitusta varianssista on haluttua informaatiota, todellisen pistemäärän varianssia. Reliabiliteettikerroin on määritelty havaitun ja todellisen pistemäärän korrelaation neliönä

$$r_{xx} = R_{XT}^2 \quad (5.3)$$

Reliabiliteettikerroin on mahdollista ilmaista myös todellisten pistemäärien varianssin ja havaittujen pistemäärien varianssin välisenä suhteena

$$r_{xx} = \frac{\sigma_T^2}{\sigma_X^2} \quad (5.4)$$

missä:

$$\begin{aligned}
 r_{xx} &= \text{reliabiliteettikerroin} \\
 \sigma_T^2 &= \text{todellisten testipistemäärien} \\
 &\quad \text{varienssi} \\
 \sigma_X^2 &= \text{havaittujen testipistemäärien} \\
 &\quad \text{varienssi}
 \end{aligned}$$

Koska mitatun testipistemäärän varianssin oletetaan koostuvan todellisen pistemäärän varianssista ja mittausvirheen varianssista, voidaan nimittäjä muuttaa seuraavaan muotoon:

$$r_{xx} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}$$

missä

$$\begin{aligned}
 r_{xx} &= \text{reliabiliteettikerroin} \\
 \sigma_T^2 &= \text{todellisten testipistemäärien} \\
 &\quad \text{varienssi} \\
 \sigma_e^2 &= \text{mittausvirheen} \\
 &\quad \text{varienssi}
 \end{aligned}$$

Reliabiliteettikerroin siis ilmoittaa, kuinka suuri osa havaittujen pistemäärien vaihtelusta syntyy mitattavasta ominaisuudesta. Jos reliabiliteetti on .9, niin 90% mittaus- tulosten vaihtelusta aiheutuu mitattavasta ominaisuudesta ja mittausvirheen vaikutus on ainoastaan 10%. Erityisesti kannattaa huomata, että jos reliabiliteetti on .5, niin mitattava ominaisuus ja mittausvirhe vaikuttavat yhtä paljon havaittujen pistemäärien vaihteluun.

Pelkkään määritelmään nojautuen ei reliabiliteettikertoimen laskeminen kuitenkaan ole mahdollista yhdellä mittauksella saatujen testipistemäärien avulla. Voidaan kuitenkin osoittaa, että *reliabiliteettikerroin on mahdollista laskea kahden rinnakkaismittarin g ja h antamien pistemäärien korrelaationa R_{gh}* . Tämän todistaminen voidaan suorittaa näin:

$$r_{gh} = \frac{\sigma_T^2}{\sigma_x^2} = r_{xx}$$

Todistus: Oletetaan, että molemmat rinnakkaismittarit antavat koehenkilölle saman todellisen pistemäärän T . Rinnakkaisten mitausten havaitut pistemäärät X_g ja X_h ovat siten

$$\begin{aligned} X_g &= T + e_g \\ X_h &= T + e_h \end{aligned}$$

Lisäksi lyhennetään merkintöjä, ja merkitään havaittujen pistemäärien poikkeamat keskiarvosta eli ns. deviaatiopistemäärät

$$\begin{aligned} x_g &= X_g - \bar{x}_{X_g} \\ x_h &= X_h - \bar{x}_{X_h} \end{aligned}$$

Näistä pistemääristä saadaan z-pistemäärät, kun ne jaetaan keskihajonnoilla. Korrelaatio r_{gh} voidaan nyt kirjoittaa kaavan 3.8 avulla muodossa

$$r_{gh} = \frac{\sum x_g x_h}{n \times \sigma_{X_g} \sigma_{X_h}}$$

Molempien testien keskihajonnat ovat yhtä suuria, joten voidaan merkitä

$$\sigma_{X_g} \sigma_{X_h} = \sigma_X^2$$

Merkitään havaitut pistemäärät todellisen pistemäärän ja virheen summana:

$$\begin{aligned} r_{gh} &= \frac{\Sigma(T + e_g) + (T + e_h)}{n \times \sigma^2} \\ &= \frac{\Sigma T^2 + \Sigma T e_g + \Sigma T e_h + \Sigma e_g e_h}{n \times \sigma_X^2} \\ &= \frac{\sigma_T^2 + s_{T e_g} + s_{T e_h} + s_{e_g e_h}}{\sigma_X^2} \end{aligned}$$

Koska mittausvirhe ja todellinen pistemäärä sekä eri testien mittausvirheet eivät korreloi keskenään, kolme viimeistä kovarianssitermiä ovat nollija. Jäljelle siis jää ainoastaan todellisten pistemäärien varianssi jaettuna havaittujen pistemäärien varianssinä, eli

$$r_{gh} = \frac{\sigma_T^2}{\sigma_X^2}$$

Suuri osa reliabiliteettitarkastelusta perustuukin tähän tulokseen. Tämän vuoksi rinnakkaisten mitausten merkitys muodostuu varsin suureksi. Erilaisia rinnakkaismitareita voi periaatteessa olla ääretön määrä.

Koska rinnakkaiset mittaukset voidaan tehdä usealla eri tavalla, tämä lisää luonnollisesti mahdollisten mittausvirheiden lähteiden lukumäärää. Erilaisilla rinnakkaismittauksilla saadut reliabiliteettikertoimet kuvaavat erilaisia mittausvirheiden lähteitä. Näin ollen on ensiarvoisen tärkeää määrittää ensin mittaustilanteen kannalta oleelliset virhelähteet, ja tämän jälkeen valita sopiva rinnakkaismittauksen menetelmä. Seuraavassa käsitellään tarkemmin erilaisia lähestymistapoja rinnakkaismittarin ongelmiaan.

määrien korrelaation laskemiseksi. Klassisen testiteorian oletusten perusteella reliabiliteettiä voidaan kuitenkin arvioida muullakin tavalla, kuin laatimalla kaksi erilaista rinnakkaismittaria. Seuraavassa esitellään erilaisia tapoja rinnakkaisten mittausten suorittamiseen. Neljä yleisintä tapaa ovat:

1. kahden testiversion käyttö
2. test-retest -menetelmä
3. split-half- menetelmä
4. sisäisen konsistenssin menetelmät

5.2 Reliabiliteetin estimointi

Klassinen testiteoria perustuu rinnakkaismittarien käyttöön reliabiliteetin arvioinnissa. Ideaalitulanteessa testistä on olemassa kaksi täysin rinnakkaista versiota (*a* ja *b*), jotka antavat täysin saman todellisen pistemäärän. Nyt olisi mahdollista antaa koehenkilöjoukolle ensin testi *a* ja sitten testi *b*. Koska molemmat testit antavat koehenkilöille täysin saman todellisen pistemäärän, johtuu havaittujen pistemäärien ero ainoastaan mittausvirheestä. Mitä yhdenmukaisempi tulos, sitä vähemmän mittausvirhe on voinut vaikuttaa testipistemääriin. Koska reliabiliteetin arviointi on nimenomaan mittausvirheen arviointia voidaan myös sanoa, että mitä yhdenmukaisempia kahden mittauksen tulokset ovat, sitä reliabelimpi testi. Reliabiliteettikerroinhan on rinnakkaisten mittausten tapauksessa palautettavissa rinnakkaismittarien piste-

5.2.1 Kahden testiversion käyttö

Kahden rinnakkaisen testiversion käyttäminen on klassisen teorian suoraan tarjoama vaihtoehto reliabiliteetin arvioimiseen. Testistä laaditaan kaksi versiota, jotka mittaavat samaa ominaisuutta, ja molemmat versiot teetetään samalla koehenkilöjoukolla. Reliabiliteetti lasketaan tämän jälkeen ensimmäisen ja toisen testiversion pistemäärien korrelaationa. On ilmeistä, että tämänkaltaisen menettely on lähinnä klassisen testiteorian ideaalia samankaltaisista rinnakkaismittareista. Ongelmana on kuitenkin kahden mahdollisimman samankaltaisen version laatiminen. Lisäksi siirtovaikutuksen riski on olemassa myös kahta eri testiversiota käytettäessä, jos testit ovat liiaksi samanlaisia. Koska testien kuitenkin pitäisi antaa *täsmälleen* samat todelliset pistemäärät, voi siis olla vaikeaa tehdä kaksi testiä, jotka ovat tarpeeksi *samanlaisia*, mutta myös riittävän *erilaisia*. Lisäksi kahden erilaisen testiversion laatiminen on erittäin ai-

kaavievää ja kallista. Lisäksi klassinen testiteoria tekee melko tiukkoja oletuksia rinnakkaismittareista, joten klassisen teorian oletusten mukaisten mittareiden laatiminen voi olla erittäin vaikeaa.

5.2.2 Test-retest -menetelmä

Kahta testiversiota käytettäessä ongelmaksi muodostuu usein kahden tai useamman tarpeeksi samanlaisen rinnakkaismittarin laatiminen. Klassinen testiteoria olettaa kahden täsmälleen rinnakkaisen testiversion rakentamista. Käytännössä tämä on kuitenkin hyvin vaikeaa. Triviaaliratkaisu on luonnollisesti testata koehenkilöjoukko kaksi kertaa samalla testillä. Koska jokainen testi on varmasti täysin rinnakkainen itsensä kanssa, pitäisi kahden testauskerran mittaus tulosten erojen johtua pelkästään mittausvirheestä. Reliabiliteetti määritetään näiden kahden testauskerran pistemäärien korrelaationa. Asia ei kuitenkaan ole aivan näin yksinkertainen. Ensinnäkin, mitattava ominaisuus tai piirre saattaa muuttua testauskertojen välillä. Mitä suurempi aikaväli ja mitä muuttuvampi ominaisuus on sitä suurempi on tämänkaltaisen riski. Lisäksi ongelmana on myös ns. *reaktiivisuustekijä*. Testin tekeminen on saattanut muuttaa mitattavaa ominaisuutta tai piirrettä. Lisäksi siirtovaikutus on edelleen ongelmana. Testattavat saattavat muistaa, millä tavalla he vastasivat ensimmäisellä kerralla, mikä taas voi vaikuttaa heidän käyttäytymiseensä toisella testauskerralla. Yhden testin käyttäminen kaksi kertaa on myöskin aikaavievää. Siirtovaikutuksen vähentämiseksi kahden testauskerran välissä voi olla tarpeen

pitää pitkäkin tauko.

5.2.3 Split-half -menetelmä

Käytännössä edellisiä helpommin toteutettava tapa arvioida reliabiliteettia on ns. split-half -menetelmä. Siinä testi jaetaan kahteen eri osaan, ja testin puolikkaita käsitellään erillisinä testiversioina ts. rinnakkaismittareina. Yksi henkilö tekee aina testin kokonaisuudessaan, mutta testipistemäärät lasketaan erikseen kummallekin testipuolikkaalle. Nyt testin reliabiliteetti saadaan laskettua, kun lasketaan testin puolikkailla saatujen pistemäärien välinen korrelaatio. Tällainen menetelmä tarjoaa huomattavia etuja verrattuna edellä esitettyihin. Testauskertoja tarvitaan vain yksi, ja rinnakkaismittarien laatimisen ongelmasta päästään eroon. Reaktiivisuudesta ja siirtovaikutuksestakaan ei tarvitse välittää, koska split-half -menetelmässä ei ole kuin yksi testauskerta. Ongelmaksi muodostuukin lähinnä testin jakaminen kahteen osaan. Esimerkiksi vain kuudesta kysymyksestä muodostuva testi voidaan jo jakaa kahteen osaan kymmenellä eri tavalla. Jokainen jako tuottaa todennäköisesti erilaisen reliabiliteettikertoimen, eikä ole mahdollista perustella, mitä niistä tulisi käyttää.

5.2.4 Sisäisen konsistenssin menetelmä

Reliabiliteettia voidaan tarkastella myös pelkästään testin sisäisenä ominaisuutena. Sisäisen konsistenssin menetelmissä tarkastellaan testin reliabiliteettia lähtien liikkeelle yksittäisistä testikysymyksistä tai -

tehtävistä, *itemeistä*. Erityisesti tarkastellaan sitä, miten samanlaisia eri testi-itemit ovat keskenään. Jos testi-itemit mittaavat samaa ominaisuutta samalla tavalla, niin henkilön pitäisi saada noilla itemeillä suunnilleen sama pistemäärä. Sisäisen konsistenssin menetelmä onkin sukua split-half -menetelmälle; erona on vain analyysiyksikkö. Siinä missä split-half -menetelmä tarkastelee testin puolikkaita, niin sisäisen konsistenssin menetelmä tarkastelee yksittäisiä itemeitä. Tällöin testin reliabiliteettiin vaikuttaa kaksi tekijää: testi-itemien mittaustulosten samankaltaisuus ja testi-itemien määrä. Tästä seuraa, että mitä enemmän itemeitä testi sisältää, sitä reliabelimpi se on. Edelleen, mitä korkeampia ovat korrelaatiot testi-itemien välillä, sitä reliabelimpi testi on. Sisäisen konsistenssin menetelmä on hyvin käytännöllinen tapa estimoida reliabiliteettia; se ei vaadi ylimääräistä testaamiskertaa tai rinnakkaisittain mittaria. Käytännössä sisäisen konsistenssin menetelmät perustuvat siis siihen, että lasketaan korrelaatiot kaikkien testi-itemien välille (kaikki mahdolliset kahden itemin väliset kombinaatiot), ja lasketaan korrelaatioista keskiarvo. Tämän jälkeen sovelletaan reliabiliteetin laskemiseen kaavaa

$$\alpha = \frac{n(\bar{r}_{ij})}{1 + (n-1)\bar{r}_{ij}} \quad (5.5)$$

missä:

- α = reliabiliteetti (sisäinen konsistenssi)
- n = testi-itemien lukumäärä

$\bar{r}_{ij}$ = testi-itemien interkorrelaatioiden keskiarvo

Tätä tunnuslukua kutsutaan *Cronbachin α -kertoimeksi*. Reliabiliteettiin vaikuttaa siis testi-itemien interkorrelaatioiden lisäksi itemien lukumäärä. Reliabiliteetti tarkoittaaakin tässä yhteydessä sitä, kuinka hyvin testi-itemit mittaavat samaa asiaa. Tällaisesta reliabiliteetin määrittelystä seuraa, että täysin mieletön testi voi olla hyvinkin reliabeli. Vaikka itemien väliset korrelaatiot olisivat melko lähellä nollaa, kasvaa testin reliabiliteetti, kun itemien määrä lähestyy ääretöntä. Tämä on kuitenkin hyväksyttävä tulos, sillä reliabiliteetti mittaa tässä yhteydessä mittaustulosten yhdenmukaisuutta eikä mittauksen järkevyyttä. Tähän tarkasteluun palataankin *validiteetin* yhteydessä.

5.3 Mittauksen keskivirhe

Reliabiliteettikerroin mahdollistaa mittaustulosten suuruuden arvioinnin, mutta se ei itsessään kerro mitään testipistemäärien vaihtelusta. Vaikka testin reliabiliteetti olisi korkea, ei voida tietää, onko esim. 10 pisteen ero kahden testatun välillä todellinen vai johtuuko se mittaustulosten virheestä. Korkea reliabiliteetti ei siten takaa oikeaa mittaustulosta, vaikka lisääkin sen todennäköisyyttä.

Reliabiliteetin vaikutusta testipistemäärien vaihteluun voidaan arvioida *mittauksen keskivirheen* (Standard Error of Measurement, *SEM*) avulla. Havaituissa testipistemäärissä on aina vaihtelua, *varianssia*.

Mitä suurempi varianssi on, sitä laajemmalle jakautuvat henkilöiden testistä saamat pistemäärät. Reliabiliteetti taas ilmoittaa, kuinka mittausvirheettömiä pistemäärät ovat. SEM on määritelty näiden kahden tekijän tulona. SEM ilmoittaa kuinka paljon pelkät mittausvirheet vaikuttavat vaihteluun testipistemäärissä:

$$SEM = \sigma_x^2 \sqrt{1 - r_{xx}} \quad (5.6)$$

missä:

$$\begin{aligned} SEM &= \text{mittauksen keskivirhe} \\ \sigma_x^2 &= \text{varianssi} \\ r_{xx} &= \text{reliabiliteettikerroin} \end{aligned}$$

SEM muodostuu siis testipistemäärien varianssin ja $\sqrt{1 - \text{reliabiliteettikerroin}}$ tulona. Tulo kasvaa, jos reliabiliteetti pienenee tai varianssi kasvaa. Mitä suurempi varianssi ja mitä pienempi reliabiliteetti, sitä suurempi on mittauksen keskivirhe. Niinpä SEM antaa tietoa siitä, missä rajoissa testatun suoritus vaihtelee pelkän mittausvirheen vaikutuksesta. Mittauksen keskivirhettä voidaan käyttää hyväksi määrittäessä mittauksien luottamusväliä. Tämän avulla on mahdollista arvioida, millä välillä henkilön testipistemäärä on jollakin tietyllä todennäköisyydellä. Seuraava esimerkki osoittaa, kuinka melko pienelläkin SEM:lla voi olla suuri vaikutus testipistemääriin. Oletetaan, että jonkin tietyn mittarin SEM on 4,7. Haluamme tietää 95%:n todennäköisyydellä henkilön pistemäärän. Koska mittausvirhe on normaalisti jakautunut, tiedämme, että 95% mittausvirheistä sijaitsee 1.96:n keskihajonnan (+ tai -)

etäisyydellä keskiarvosta. Niinpä mittausvirheiden 95 prosentin luottamusväli on $4.7 \times 1.96 = 9.2$. Eli jos henkilö on saanut testissä 100 pistettä, hänen pistemääränsä on 95%:n todennäköisyydellä välillä (100 ± 9.2) , eli 90.8 – 109.2).

Koska mitattu pistemäärä ei koskaan ole identtinen todellisen testipistemäärän kanssa, luottamusväli olisi tietenkin optimaalisinta määrittää todellisen eikä havaitun pistemäärän ympärille. Tällöin ongelmana on todellisen pistemäärän estimointi. Todellista pistemäärää voidaan kuitenkin estimoida havaituista pistemäärästä lineaarisen regressiion avulla. Reliabiliteettikerroimen neliöjuuri $\sqrt{r_{xx}}$ ilmoittaa todellisen ja havaitun testipistemäärän välisen korrelaation. Näinollen sitä voidaan käyttää, kun halutaan estimoida todellisen pistemäärän alarajaa havaittujen pistemäärien avulla. Voidaan osoittaa, että

$$t' = r_{xx}x \quad (5.7)$$

Todelliset pistemäärät saadaan siten havaituista pistemäärästä, kun ne kerrotaan reliabiliteettikerroimella. Kannattaa kuitenkin huomata, että todellisten pistemäärien varianssi on pienempi kuin havaittujen pistemäärän varianssi on muodostunut eli havaitun pistemäärän varianssi muodostuu todellisen pistemäärän varianssista ja mittausvirheen varianssista :

$$r_{xx} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}$$

Tämä selittää ilmiön nimeltä *regression toward the mean*. RTM ilmenee, jos henkilöt esimerkiksi valitaan tutkimukseen sen

mukaan, kuinka he suoriutuvat jossain testissä. Kun heidät mitataan toisen kerran, jokaisen henkilön testipistemäärä on todennäköisesti lähempänä muuttujan keskiarvoa kuin aikaisempi mittausta *vaikka* siirtovaikutusta tai reaktiivisuutta ei ilmenisi! Tämä sinänsä paradoksaalinen ilmiö selittyy kuitenkin mittauksen keskivirheen avulla. Koska mittausvirheen vuoksi todellisten pistemäärien varianssi on havaittujen pistemäärien varianssia pienempi, ovat havaitut pistemäärät itseisarvoltaan suurempia kuin vastaavat todelliset pistemäärät. Tämä tarkoittaa siis sitä, että yleisesti ottaen havaitut pistemäärät sijaitsevat kauempana keskiarvosta kuin todelliset pistemäärät. Osa mitatuista henkilöistä saa siis liian suuria ja osa liian pieniä pistemääriä. Esimerkiksi WAIS-R:ssa pistemäärien keskiarvo on 100. Jos henkilö saa havaituksi pistemääräkseen esimerkiksi 160, tämä voi johtua kolmesta syystä:

1. 160 on henkilön todellinen pistemäärä
2. Henkilöllä oli onnea ja hän menestyi hyvin, joten havaittu pistemäärä on todellista pistemäärää korkeampi.
3. Henkilöllä oli epäonnea ja hän menestyi huonosti, joten havaittu pistemäärä on todellista pistemäärää pienempi.

Koska populaatiossa on erittäin vähän henkilöitä, joiden ÄO:n todellinen pistemäärä on yli 160, on vaihtoehto 2 paljon todennäköisempi kuin 3. Vastaava analogia pätee tietenkin henkilöön, joka saa havaituksi pistemääräkseen 40, tällöin on tietenkin

todennäköisempää, että henkilön todellinen pistemäärä on suurempi kuin 40. Onni ei kuitenkaan ole mittauksesta toiseen pysyvä ominaisuus. Täten nämä vaikutukset todennäköisesti poistuvat seuraavalla mittauskerralla, joten molempien henkilöiden havaitut pistemäärät todennäköisesti lähestyvät keskiarvoa.

5.4 Reliabiliteettiin vaikuttavia tekijöitä

Edellä on käsitelty erilaisia menetelmiä reliabiliteetin arvioimiseen ja mittaamiseen. Seuraavaksi käsitellään joitakin ilmiöitä mitkä vaikuttavat reliabiliteetin suuruuteen. Yleisesti ottaen nämä tekijät voidaan jakaa neljään luokkaan:

1. Testattaviin liittyvät tekijät
2. Testiin liittyvät tekijät
3. Testisuoritukseen liittyvät tekijät
4. Reliabiliteetin arviointiin käytetyt menetelmät

5.4.1 Testattaviin liittyvät tekijät

Testin reliabiliteettiin vaikuttavat ihmiset, joita sillä testataan. Se, kuinka paljon testattavaan piirteeseen liittyy yksilöllistä variaatiota, vaikuttaa testin luotettavuuteen. Testejä käytetään usein yksilöiden välisten erojen selvittämiseen. Jos henkilöt eivät eroa juurikaan jonkin ominaisuuden suhteen, on vaikea tehdä testiä, jolla tämä ero saataisiin

selville. Esimerkiksi älykkyystesti on reli-aabelimpi, mikäli testataan kaikki suoma-laiset kuin pelkästään yliopiston opiskeli-jat. Tähän ilmiöön sisältyy paradoksi. Mit-tauksen keskivirrehän saatiin eräänlaisena varianssin ja reliabiliteetin tulona $SEM = \sigma_x \sqrt{1 - r_{xx}}$. Jos yksilöiden väliset erot ovat pieniä, muodostuu reliabiliteetista todennä-köisesti pieni. Samalla varianssikin on pie-ni, koska sehän ilmoittaa nimenomaan sen kuinka laajalle jonkin ominaisuuden arvot ovat levinneet. Niinpä on mahdollista luoda testi, joka ei ole reliabeli, mutta jonka an-tamat testitulokset ovat SEM:lla arvioituna tarkkoja.

5.4.2 Testiin liittyvät tekijät

Itse testin rakenne vaikuttaa luonnollisesti sen reliabiliteettiin. Kuten sisäisen konsis-tenssin menetelmä osoittaa, testin reliabili-teettia voidaan kasvattaa

1. parantamalla testi-itemien vä-listä korrelaatiota eli itemien ho-mogeenisuutta ja
2. lisäämällä itemien määrää.

Testin pituutta on luonnollisesti helpom-pi lisätä kuin homogeenisuutta. On mel-ko helppoa keksiä testiin lisää samankal-taisia testi-itemeitä. Niinsanottu Spearman-Brownin kaava estimoi reliabiliteetin li-sääntymisen itemien määrää kasvattamalla.

$$\text{uusi } r_{xx} = \frac{n \times \text{vanha } r_{xx}}{1 + (n - 1) \text{vanha } r_{xx}} \quad (5.8)$$

missä:

$$n = \text{uuden testin itemien}$$

	lukumäärä
vanha r_{xx}	= alkuperäisen testin reliabiliteettikerroin
uusi r_{xx}	= pidennetyn testin reliabiliteettikerroin

Kuten α -kertoimen yhteydessä on todet-tu, pitkä testi voi olla reliabeli, vaikka itemien välinen korrelaatio olisikin varsin pieni. Testi-itemien määrää ei kuitenkaan kannata ruveta kasvattamaan suhteettomas-ti. Pitkä testi ei välttämättä ole miellyttävä tehdä, vaikka sen olisi tarkoitus olla kuinka reliabeli.

5.4.3 Testin käyttöön liittyvät tekijät

Myös se, mihin testipistemääriä aiotaan käyttää, vaikuttaa testin reliabiliteettiin. Yleisesti ottaen testit ovat reliabelimpia erottelemaan yksilöitä testaushetkellä kuin jollain pitemmällä aikavälillä. Eli, testit an-tavat yleensä reliabelimman kuvan testat-tavan ominaisuuksista tällä hetkellä kuin esimerkiksi kymmenen vuoden kuluttua. Myös silloin, kun testiä sovelletaan tes-tattavan ominaisuuden suhteen homogeeni-seen ryhmään, yksilöllisten erojen luotetta-va mittaaminen ei välttämättä onnistu.

5.4.4 Testin reliabiliteetin määrittäminen

Edellä on esitetty lukuisia menetelmiä tes-tin reliabiliteetin arvioimiseen. On hyvä

huomata, että myös käytetty reliabiliteetin arviointimenetelmä vaikuttaa reliabiliteettikertoimen suuruuteen, ts. eri menetelmät antavat erilaisen estimaatin reliabiliteetistä. Yleisesti ottaen reliabiliteetin arviointimenetelmät voidaan asettaa käyttökelpoisuuden mukaan paremmuusjärjestykseen seuraavasti :

sisäisen konsistenssin menetelmät > kahden testiversion käyttö > uusintatesti

Klassisen testiteorian näkökulmasta paras menetelmä ei siis sovellu käytännön tilanteisiin kovinkaan hyvin. Sisäisen konsistenssin menetelmien vahvuus onkin juuri siinä, että kahdesta mittauskerrasta voidaan luopua. Tämä tietenkin säästää aikaa, vaivaa ja rahaa. Arviointimenetelmää valittaessa on kuitenkin ensin mietittävä, mihin testituloksia aiotaan käyttää ja tehdä valinta sen pohjalta. Joskus voi olla perusteltua tehdä samasta testistä kaksi rinnakkaista versiota.

5.5 Korrelaation attenuaatio

Reliabiliteetti vaikuttaa mittaustulosten tarkkuuteen, ja samalla myös erilaisiin mittaustuloksista laskettuihin tunnuslukuihin, kuten korrelaatioihin. Pieni reliabiliteetti *attenuoi* eli heikentää kahden muuttujan välistä havaittua korrelaatiota, eli $r_{x_t y_t} > r_{x_X y_X}$. Tämän vuoksi Spearman kehitti ns. *attenuaatiokorjauksen*, minkä avulla voidaan estimoida todellisten piste-

määrien välistä korrelaatiota. Spearmanin attenuaatiokorjauksen kaava on

$$r'_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}} \quad (5.9)$$

missä

r_{xy} = x:n ja y:n korrelaatio

r'_{xy} = x:n ja y:n välinen korjattu korrelaatio

r_{xx} = x:n reliabiliteetti

r_{yy} = y:n reliabiliteetti

Attenuaatiokorjattu korrelaatio siis arvioi, mikä suuruinen muuttujien välinen korrelaatio olisi, jos molemmat mittarit olisivat täysin mittausvirheettömiä.

Attenuaation määritelmä on teoreettisesti mielekäs, mutta korjauksen käyttämiseen on liittyy kuitenkin käytännön ongelmia. Ensinnäkin, attenuaatiokorjauksen määritelmä lähtee oletuksesta, jonka mukaan voitaisiin rakentaa mittareita, jotka ovat täysin virheettömiä. Käytännössä tämä ei ole mahdollista. Toiseksi, lähes kaikissa tapauksissa tämänkaltainen korjaus arvioi muuttujien välisen korrelaation liian suureksi. Tämän vuoksi yleensä käytetäänkin toisenlaista lähestymistapaa. On nimittäin mahdollista arvioida, kuinka paljon kahden muuttujan välinen korrelaatio muuttuu, jos käytettyjen mittareiden reliabiliteettiä muutetaan. Tätä voidaan estimoida seuraavasti:

$$r'_{xy} = \frac{r_{xy} \sqrt{\text{uusi } r_{xx} \times \text{uusi } r_{yy}}}{\sqrt{\text{vanha } r_{xx} \times \text{vanha } r_{yy}}} \quad (5.10)$$

missä

r_{xy} = x :n ja y :n korrelaatio
 r'_{xy} = Uusi korrelaatio, kun
 reliabiliteetit muuttuvat
 vanha r_{xx} = x :n vanha reliabiliteetti
 vanha r_{yy} = y :n vanha reliabiliteetti
 uusi r_{xx} = x :n uusi reliabiliteetti
 uusi r_{yy} = y :n uusi reliabiliteetti

reiden reliabiliteetti voidaan määrittää seuraavasti:

$$r_{DD} = \frac{\frac{r_{xx} + r_{yy}}{2} - r_{xy}}{1 - r_{xy}} \quad (5.11)$$

missä

r_{DD} = erotussuureen reliabiliteetti
 r_{xx} = x -muuttujan reliabiliteetti
 r_{yy} = y -muuttujan reliabiliteetti
 r_{xy} = x :n ja y :n välinen korrelaatio

Nyt on siis mahdollista tarkastella myös reliabiliteetin *pienenemisen* vaikutusta muuttujien väliseen korrelaatioon. Kannattaa myös huomata, että kaava 5.10 on täysin identtinen Spearmanin attenuaation korjauskaavan (5.9) kanssa silloin, kun testien uusiksi reliabiliteeteiksi tulee 1 ($\sqrt{1 \times 1} = 1$). Kaavan avulla voidaan todeta, että reliabiliteetin vaikutus korrelaatioon on erittäin suuri. Jos alkuperäisten mittausten reliabiliteetit ovat .90 ja korrelaatio .50, niin korrelaatio olisikin vain .33 jos molempien testien reliabiliteetti olisikin vain .60!

Erotussuureiden reliabiliteetti on usein alhainen, ja niiden laskemiseen liittyy paradoksi. Mitä suurempi kahden testin tai testauskerran välinen korrelaatio on, sitä alhaisempi reliabiliteetti on erotussuureella! Tähän sinänsä järjettömältä vaikuttavaan ilmiöön on syynsä. Ensinnäkin, X - ja Y -muuttujien havaitut testipistemäärät X_x ja X_y koostuvat todellisesta pistemäärästä ja mittausvirheestä, eli

$$\begin{aligned}
 X_x &= T_x + e_x \\
 X_y &= T_y + e_y
 \end{aligned}$$

5.6 Erotussuureet

Erotussuureilla tarkoitetaan henkilön kahden eri hetkellä mitatun testipistemäärän välistä erotusta. Tällainen tarkastelu on kyseessä esim. silloin, kun halutaan selvittää kuinka paljon Tarjan ÄO on parantunut 6 vuodessa. Tällöin mitataan hänen suorituksensa vuonna 2000 (X) ja 2006 (Y), ja lasketaan testipistemäärien erotus. Erotussuu-

Siis jos Tarjan tapauksessa X ja Y korreloivat voimakkaasti, se tarkoittaa, että todelliset pistemäärät X_x ja X_y ovat pitkälle yhteneväiset. Näinollen X -ja Y -mittausten välillä ei ole juurikaan eroa. Täten eron X :n ja Y :n välillä täytyy johtua lähinnä mittausvirheestä. Niinpä voidaan sanoa, että mitä suurempi korrelaatio X :n ja Y :n välillä sitä vähemmän reliabeli on niiden erotus.

5.7 Summamuuttujat

Summamuuttujilla tarkoitetaan usean muuttujan yhdistämistä yhdeksi muuttujaksi. Psykometriikan piirissä summamuuttujia voidaan muodostaa esimerkiksi useamman erillisen testin pistemääristä tai saman testin osatestien pistemääristä. Summamuuttujat ovat reliabiliteettitarastelun kannalta merkityksellisiä, koska ne ovat yleensä reliabelimpia kuin ne yksittäiset testit, joiden pistemäärät on yhdistetty yhdeksi summamuuttujaksi. Mitä useamman testin pistemäärät yhdistetään ja mitä korkeampi testien välinen korrelaatio on, sitä reliabelimpi on summamuuttuja. Summamuuttujan reliabiliteetti voidaan laskea seuraavasti:

$$r_{ss} = \frac{k - k\bar{r}_{ii}}{1 - k + [(k^2 - k)\bar{r}_{ij}]} \quad (5.12)$$

missä

- r_{ss} = summamuuttujan reliabiliteetti
- k = yhdistettävien testien lukumäärä
- $\bar{r}_{ii}$ = testien keskimääräinen reliabiliteetti
- $\bar{r}_{ij}$ = testien välinen keskimääräinen korrelaatio

Summamuuttujan idea on siis tavallaan analoginen menetelmän kanssa, jossa testin reliabiliteettia kasvatetaan testi-itemien määrää kasvattamalla: mitä useammalla testillä ominaisuutta testataan, sitä luotettavampi tulos. Summamuuttujien muodosta-

misen haittapuoli kuitenkin on, että muuttujia yhdistettäessä hävitetään informaatiota. Vertaamalla erotuusuureita ja summamuuttujia voidaan tiivistäen todeta, että testipistemäärien *vähentäminen* toisistaan *pienentää* reliabiliteettia, kun taas testipistemäärien *summaaminen* kasvattaa reliabiliteettia.

5.8 Kuinka reliabeli testin pitäisi olla?

Edellä on käsitelty erilaisia menetelmiä testin reliabiliteetin arvioimiseen ja kasvattamiseen. Varoituksena todettakoon, että helpoimmillaankin reliabiliteetin määrittäminen ja kasvattaminen on melko työlästä. Usein onkin tarkoituksenmukaista miettiä kuinka korkeata reliabiliteettia tarvitaan. Koska reliabiliteetti ilmaisee mittausvirheen määrän, on selvää että tarkemmat mitaustilanteet vaativat reliabelimman testin. Eli, mitä tarkempia mittaustuloksia tarvitaan, sitä vähemmän mittausvirhettä halutaan mukaan tulokseen. Yleistykseenä voidaan sanoa, että korkea reliabiliteetti vaaditaan, kun

- a) testiä käytetään tehtäessä yksilön kannalta lopullisia päätöksiä tai
- b) yksilöitä luokitellaan useaan ryhmään melko pienten yksilölisten erojen perusteella.

Tällaisissa tilanteissa on luonnollisesti pyrittävä suurimpaan mahdolliseen mitaustarkkuuteen. Esimerkiksi kliinisessä

työssä tai rekrytoinnissa pienetkin vaihtelut testipistemäärissä saattavat aiheuttaa yksilölle vakavia seuraamuksia. Aina äärimmäisen korkea reliabiliteetti ei kuitenkaan ole tarpeen. Verraten matalallakin reliabiliteetilla pärjätään, kun:

- a) testiä käytetään vain tekemään alustavia päätöksiä tai
- b) yksilöitä luokitellaan vain muutamaaan ryhmään suurten yksilöllisten erojen pohjalta.

Yleensä 0.7ää pienempiä reliabiliteettikertoimen arvoja pidetään pieninä. Kliiniseen käyttöön tarkoitettujen testien reliabiliteetin pitäisi olla vähintään .9, ja tutkimuksessa käytettyjen mittareiden reliabiliteetti ei saisi alittaa .8aa. Jos reliabiliteetti on tasan .5, niin mittausvirhe ja todellinen pistemäärä vaikuttavat yhtä paljon havaittuihin pistemääriin, ja testin käyttöarvo alkaa olla melko kyseenalainen. Tästä syystä siis kovinkaan alhaisia reliabiliteettikertoimen arvoja ei pidä hyväksyä, vaan testin reliabiliteettia on pyrittävä parantamaan.

Tarkastelu voidaan päättää dilemmaan testin kattavuudesta ja tarkkuudesta. Tätä kutsutaan myös *bandwidth-fidelity -dilemmaksi*. Mitä laajempi on testin kattavuus eli mitä useampia asioita se mittaa, sitä pienempi on testin tarkkuus. Toisaalta, mitä suurempi on testin tarkkuus, sitä suppeampi on sen kattavuuden oltava. Tämän ongelman kanssa joudutaan luonnollisesti työskentelemään testejä suunniteltaessa: on vaikeaa luoda testi, joka olisi samaan aikaan hyvin kattava ja toisaalta hyvin tarkka.

Luku 6

Validiteetti

Kuten aiemmin todettiin, testin validiteetti ilmaisee eräällä tavalla testin käytökelpoisuutta. Reliabiliteettihan ilmaisee pelkästään testipistemäärien virheettömyyden, mutta validiteetista puhuttaessa erotetaan kaksi erilaista näkökulmaa:

1. Mittauksen validiteetti ilmaisee sen, mittaako testi sitä, mitä sen on ajateltu mittaavan. Mittauksen validiteetti voidaan jakaa
 - 1.1 sisältövaliditeettiin ja
 - 1.2 käsitevaliditeettiin
2. Kriteeripohjainen validiteetti ilmaisee voidaanko testiä käyttää luotettavasti päätöksenteon pohjana. Kriteerivaliditeetti voidaan jakaa
 - 2.1 ennustevaliditeettiin ja
 - 2.2 rinnakkaisvaliditeettiin

Kuumemittarin mittausvaliditeetti on siis hyvin korkea mitattaessa henkilön kuumetta, mutta erittäin matala, jos sillä mitataan henkilön pituutta. Samoin kuumemittarin kriteerivaliditeetti on korkea, jos sillä pyritään erottelemaan kuumeiset henki-

löt ei-kuumeisista, mutta hyvin matala, jos sillä pyritään erottelemaan skitsofreenikot terveistä koehenkilöistä. Testi voi siten olla mittauksen osalta validi, mutta päätöksenteon välineenä epävalidi. Kuumemittari siis mittaa validisti kuumetta, mutta saatua mittaustulosta ei voida käyttää skitsofrenia -diagnoosin tekemiseen. Myös käänteinen on mahdollista. Testi voi myös olla päätöksenteon pohjana validi, mutta mittauksen osalta epävalidi.

6.1 Mittauksen validiteetti

Mittauksen validiteetti tarkoittaa yksinkertaisesti mittausmenetelmien laadun tarkastelua. Mittauksen validiteetilla arvioidaan, kuinka todennäköisesti testi mittaa sitä, mitä sen oletetaan mittaavan. Mittauksen validiteettia voidaan tarkastella kahdella tavalla, sisältövaliditeetin ja käsitevaliditeetin avulla.

6.1.1 Sisältövaliditeetti

Testin sisältövaliditeettia määritettäessä arvioidaan, kuinka edustavan otoksen testin mittaamat käyttäytymispiirteet muodostavat mittauksen kohteena olevasta piirteestä. Toisin sanoen, onko testin käyttäytymisotos edustava otos siitä käyttäytymisestä, jota testillä halutaan mitata. Keskeisenä ongelmana sisältövaliditeettia määritettäessä on mitattavan sisältöalueen käsitteellisen määrittäminen: mitkä käyttäytymispiirteet kuuluvat mitattavaan ominaisuuteen ja mitkä taas eivät. Lisäksi sisältöalueelle voidaan usein määrittää hierarkkinen rakenne. Sisältöalue jaetaan osa-alueisiin ja määritetään osa-alueiden suhteellinen relevanssi mitattavan ominaisuuden kannalta.

Sisältövaliditeetti määritetään arvioimalla, kuinka suuri osa testi-itemeistä kuuluu sisältöalueeseen ja kuinka edustavasti testi-itemit heijastavat sisältöalueen hierarkista rakennetta. Keskeisten osa-alueiden pitäisi siis olla paremmin edustettuina kuin vähemmän keskeiset. Esimerkiksi WAIS-R:n sisältövaliditeettia voitaisiin tutkia arvioimalla, kuinka hyvin eri osatestit edustavat eri älykkyyden osa-alueita. Lisäksi voitaisiin tutkia kielellinen vs. suoritusosa - jaottelun mielekkyyttä. Sisältöalueen täsmällinen määrittäminen on usein vaikeaa. Ajattele esimerkiksi älykkyyttä tai verbaalista lahjakkuutta: ei ole itsestään selvää millaiset toiminnot liittyvät älykkyyteen ja miten niitä voidaan mitata. Tällaisessa tapauksessa täytyy usein arvioida testin *käsitevaliditeettia*.

Reliabiliteetti ja sisältövaliditeetti mittaavat jossain määrin toisiaan. Näillä kah-

della käsitteellä on kuitenkin selvä ero. Reliabiliteettitarkastelussa lähtökohtana on, että testi-itemit mittaavat jotain tiettyä ominaisuutta. Tämän jälkeen arvioidaan, kuinka tarkasti tämä mittaus onnistuu. Sisältövaliditeetti taas nimenomaan arvioi, missä määrin testi-itemit mittaavat tuota haluttua ominaisuutta.

6.1.2 Käsitevaliditeetti

Toinen tapa tarkastella mittauksen validiteettia on ns. käsitevaliditeetin tutkiminen. Lähtökohtana on, että käsitteet ovat abstraktioita, joilla on aina yhteys - joko suora tai epäsuora - käyttäytymiseen tai kokemukseen. Käsitevaliditeetti on läheistä sukus sisältövaliditeetille: jälkimmäinen sopii tilanteisiin, joissa sisältöalue on tarkkaan selvillä (esim. soveltuvuus tiettyyn työtehtävään), edellinen taas tilanteisiin, joissa mitataan enemmän abstrakteja piirteitä tai ominaisuuksia kuten älykkyyttä. Tällaiset abstraktiothan ovat yleisiä psykologiasa. Käsitevaliditeetin määrittäminen alkaa mitattavan käsitteen operationalisoinnilla. Tällöin määritellään käsitteen (1) tyypillisimmät käyttäytymisvasteet, (2) sellaiset, jotka liittyvät käsitteeseen väljemmin, sekä (3) sellaiset, joilla ei ole mitään yhteyttä käsitteeseen. Aluksi tutkitaan, millaisia käyttäytymisen muotoja mitattavaan käsitteeseen liittyy, ja kuinka kiinteässä yhteydessä ne ovat käsitteeseen. Tämän jälkeen mitataan suoraan näitä käyttäytymispiirteitä, sekä korreloidaan ne käsitettä mittaavan testin kanssa. Mikäli korrelaatiot ovat sen suuruisia kuin niiden teoreettisen tarkastelun perusteella tulisi olla, voidaan olettaa, että tes-

tillä on korkea käsitevaliditeetti. Eräs tapa arvioida käsitevaliditeettia on ns. *multitrait-multimethod*- lähestymistapa. Siinä jokaista käsitettä arvioidaan *usealla* eri menetelmällä samanaikaisesti. Esimerkiksi koululaisten epäsuoraa aggressiota voitaisiin mitata opettajan arviointien avulla, toveriarvioinnilla ja oppilaan itsearviointilla.

6.2 Kriteerivaliditeetti

Toinen vaihtoehtoinen lähestymistapa validiteettitarkasteluun on tarkastella testillä tehtyjen päätösten laatua jonkin *kriteerin* avulla. Kriteerillä tarkoitetaan sellaista mittausta, jonka perusteella voidaan arvioida tehdyn päätöksen laatua. Tällaisessa tapauksessa on kyse testin *kriteerivaliditeetin* arvioimisesta. Testillä on korkea kriteerivaliditeetti, jos sen avulla voidaan tehdä onnistuneita päätöksiä. Esimerkiksi rekrytointissa käytettävän testin kriteerivaliditeetti on korkea, mikäli sen avulla voidaan valita työtehtävässä parhaiten menestyvät työntekijät. Kriteeripohjainen validiteetti voidaan jakaa kahteen alatyypin: ennustevaliditeettiin ja rinnakkaisvaliditeettiin.

6.2.1 Ennustevaliditeetti

Ennustevaliditeetin määrittäminen tehtäisiin ideaalitalanteessa seuraavasti: Aluksi testataan joukko ihmisiä (esimerkiksi työnhakijoita) testillä, mutta ei käytetä testituloksia valinnan perusteena, vaan valitaan henkilöt tehtävään satunnaisesti. Seuraavaksi arvioidaan jollain tavoin valittujen työmenestys. Nyt lasketaan korrelaatio työ-

menestyksen ja testipistemäärän välille, ja saadaan arvio ennustevaliditeetille.

Tämänkaltaisen tarkastelu ei kuitenkaan useimmiten ole mahdollista. Se edellyttää satunnaista päätöksentekoa valintatilanteessa, mikä ei yleensä tule kysymykseen. Menettely tulisi myös kalliiksi ja olisi eettisesti varsin arveluttavaa. Niinpä seuraava järjestely on lähempänä mahdollista: Valintatilanteessa testiä käytetään päätöksen teon pohjana, ja valitaan tehtävään testissä parhaiten menestyneet. Ennustevaliditeetti määritetään tällöin niiden henkilöiden suoritusten perusteella, jotka päätetään valita. Ongelmallista on kuitenkin se, että validointia ei suoriteta koko populaation pohjalta. Joku testin perusteella hylätty ihminenhan olisi voinut menestyä tehtävässä erinomaisesti. Useinhan testillä halutaan erottaa huonoimmin suoriutuvat henkilöt muista, mutta ennustevaliditeetin arvioinnissa ei 'huonoja' ole mukana laisinkaan. Tässä tapauksessa arvioidaan pikemminkin sitä, kuinka 'hyvät erottuvat erittäin hyvistä'.

6.2.2 Rinnakkaisvaliditeetti

Todellisen ennustevaliditeetin määrittäminen testille on usein mahdotonta. Käytännön sanelema vaihtoehto onkin usein ns. rinnakkaisvaliditeetti. Menetelmä on rinnakkaisvaliditeetin tapauksessa muuten sama kuin ennustevaliditeetin kohdalla paitsi, että kohdejoukkona jokin ennalta valikoitu ryhmä kuten esim. yhtiön työntekijät tai koulun oppilaat. Tämä ryhmä tekee testi, ja heidän tehtävässä suoriutumisensa arvioidaan. Näiden mittausten pohjalta lasketaan

testin rinnakkaisvaliditeetti. Rinnakkaisvaliditeetin etuna on käytännöllisyys, mutta ongelma on kuitenkin siinä, että validiteetin määrittämiseen käytetty ryhmä ei ole sama kuin se, johon testiä aiotaan soveltaa.

Yleisesti ottaen voidaan todeta, että kriteeripohjaisen validiteetin estimaatit yleensä verraten alhaisia (.3 – .5). Parhaimmillaankin ennustevaliditeetti nousee harvoin suuremmaksi kuin 0.7. Tämä ei kuitenkaan kuvaa suoraan päätöksenteon hyvyttä, vaan ennemminkin sitä, kuinka vaikeaa on laatia testi, joka mittaisi suoriutumista jossain nimenomaisessa tehtävässä. Tästä huolimatta testin validiteetti on pyrittävä saamaan mahdollisimman hyväksi. Kuten seuraavissa kappaleissa esitetään, tietynlaisissa mittauslaitteissa kuitenkin melko matalankin kriteerivaliditeetin omaavat testit ovat käyttökelpoisia päätöksenteon apuvälineitä.

6.3 Face validity

Testin laadinnassa voidaan lisäksi puhua *face validitysta*. Tällä tarkoitetaan sitä, kuinka mielekkäänä ja järkevänä itse testattavat pitävät testiä. Face validity tarkoittaa siis testin *katu-uskottavuutta*. Kovin kaan monet ihmiset eivät esimerkiksi suostuisi siihen, että oppilaitoksen oppilaat valitaan sen perusteella, mitä kukin testattava piirtää paperilla oleviin ruutuihin. Täten onkin hyvä koittaa luoda sellainen testi, joka olisi myös testattavan näkökulmasta järkevä ja tarkoituksenmukainen. Monien käytettävien testien face validity saatalla olla melko alhainen. Testattavat saat-

tavat epäillä esimerkiksi Ro:n käyttämistä, koska persoonallisuuden arvioiminen mustetahroja näyttämällä ei vaikuta heidän mielestään pätevältä menettelyltä.

Face validityta voidaan tietenkin käyttää myös väärin. Melko epämääräisestä testistä voidaan koittaa laatia mahdollisimman vakuuttava ja 'psykologisen' näköinen, etteivät ihmiset epäilisi sen toimivuutta. Ikävä kyllä myös tällaisia testejä on markkinoilla paljon. Koska asiakkailla ei useimmiten ole edellytyksiä arvioida testien käyttökelpoisuutta, psykologilla on suuri vastuu sopivien mittauslaitteiden valinnassa. Kauniin ulkokuoren ei saakaan antaa hämätä. Testin on tietenkin oltava miellyttävän näköinen ja huolellisesti laadittu, mutta ulkoasu ei missään tapauksessa voi korvata laadusta testikonstruktiota.

6.4 Päätöksenteon laadun arviointi

Testeillä tehdään päätöksiä ihmisistä. Tämä on todettu jo monta kertaa. Millaista sitten on hyvä päätöksenteko? Tehtävien päätösten laatuun voidaan olettaa vaikuttavan kolme seikkaa: perustaso, valintasuhde ja testin validiteetti. Perustaso *BR* tarkoittaa sitä, kuinka suuren osa hakijoista oletetaan suoriutuvan tehtävästä. Valintasuhde *SR* taas merkitsee sitä, kuinka suuri prosentti kaikista hakijoista valitaan. Tätä asiaa voidaan tarkastella erilaisten mahdollisten päätösten avulla. Kun suoritetaan henkilövalintaa, niin yksittäinen henkilö voidaan joko 1) valita tai 2) jättää valitsematta. Lisäksi pää-

tös valita tai olla valitsematta voi olla joko 1) oikea tai 2) väärä. Eri mahdollisuudet on esitetty taulukossa 6.1. Todellinen positiivinen päätös tarkoittaa, että henkilö todetaan testin avulla kyvykkääksi, ja hän myös on sitä. Väärä positiivinen taas tarkoittaa, että henkilö arvioidaan kyvykkääksi vaikka hän ei sitä ole. Sama periaate pätee myös negatiivisiin valintoihin.

onnistuu	Väärät negatiiviset	Todelliset positiiviset
epäonnistuu	Todelliset negatiiviset	Väärät positiiviset
	hylätään	hyväksytään

Taulukko 6.1: Vaihtoehtoiset ratkaisut valintatilanteessa

Jos valinta suoritetaan täysin satunnaisesti ja valintasuhde sekä perustaso tunnetaan, niin todellisen positiivisen valinnan todennäköisyys P_{tp} voidaan laskea helposti:

$$P_{tp} = BR \times SR \quad (6.1)$$

Todennäköisyys saadaan siis yksinkertaisesti perustason ja valintasuhteen tulona. Vastaavasti voidaan määrittää todennäköisyydet myös väärälle negatiiviselle (P_{fn}), väärälle positiiviselle (P_{fp}) ja todelliselle negatiiviselle (P_{tn}), ja valinnalle:

$$P_{fn} = BR - P_{tp} \quad (6.2)$$

$$P_{fp} = SR - P_{tp} \quad (6.3)$$

$$P_{tn} = 1 - P_{tp} - P_{fp} - P_{fn} \quad (6.4)$$

Näiden laskemiseen sovellettava periaate on yksinkertainen. Ensin lasketaan todellisen positiivisen valinnan todennäköisyys. Väärän negatiivisen todennäköisyys saadaan, kun perustasosta vähennetään todellisen positiivisen valinnan todennäköisyys. Väärän positiivisen valinnan todennäköisyys saadaan valintasuhteen ja todellisen positiivisen valinnan todennäköisyyden erotuksena, ja väärän negatiivisen valinnan todennäköisyys saadaan, kun kaikki edelliset todennäköisyydet vähennetään ykköses-tä. Yhteensä kaikkien todennäköisyyksien summa on 1.

Jos valinta tehdään täysin satunnaisesti, niin puolet päätöksistä on todennäköisesti oikeita, eli $P_{tp} + P_{tn} = .5$. Jos testin validiteetti tunnetaan, on mahdollista arvioida miten testin käyttäminen parantaa valintaprosessia. Jos testin validiteetti on r_{xy} , niin testin käyttäminen valinnassa kasvattaa todellisen positiivisen valinnan todennäköisyyttä

$$P_{tp} = BR \times SR + r_{xy} \times \sqrt{BR(1-BR)SR(1-SR)} \quad (6.5)$$

Kannattaa erityisesti huomata, että jos testin validiteetti on 0, niin silloin todellisen positiivisen valinnan todennäköisyys on sama kuin satunnaisesti valittaessa, eli $BR \times SR$. Lisäksi testi, jonka validiteetti on nollaa suurempi, parantaa aina päätöksentekoa, joskaan ei välttämättä kovinkaan paljon. Jos perustaso on .5 ja valintasuhde .2, niin satunnaisesti valittaessa todellisen positiivisen valinnan todennäköisyys on .12, ja oikean valinnan todennäköisyys¹ $T_{tp} +$

¹Todennäköisyys tehdä joko todellinen positiivinen tai negatiivinen päätös

T_{tn} on .5. Jos valinnassa käytetään testiä, jonka validiteetti on .5, niin todellisen positiivisen valinnan todennäköisyydeksi tulee .2, ja vastaavasti oikean valinnan todennäköisyydeksi .7. Jos valintasuhde olisikin ollut .5 ja testi sama, niin testin avulla tehtävän oikean päätöksen todennäköisyys olisi .75.

Testiä laadittaessa pyritään tietenkin siihen, että sen avulla voitaisiin tehdä mahdollisimman paljon todellisia positiivisia ja todellisia negatiivisia johtopäätöksiä. Edellinen tarkastelu osoittaakin, että validiteettikertoimen tarkasteleminen sinällään ei ole mielekästä, vaan validiteetin merkitystä arvioitaessa on aina otettava huomioon perustaso ja valintasuhde. Jos testattavien perustaso on korkea, on kuitenkin vaikea laatia testi, joka parantaisi merkittävästi päätöksentekoa. Mikäli perustaso taas on matala - esimerkiksi jos haluamme erottaa skitsofreenikot muusta populaatiosta - väärin päätösten tekemisen todennäköisyys kasvaa. Yleensä ottaen voidaankin todeta, että testin merkitys päätöksenteossa on suurimmillaan silloin, kun perustaso on noin 0.5:n luokkaa.

Valintasuhteen merkitys taas on tämänkaltaisen: Mitä suurempi on valintasuhde, sitä pienempi on testin merkitys. Eli jos jostain työpaikkaa hakevasta sadasta ihmisestä valitaan 99, ei testillä ole enää paljoakaan väliä. Mikäli valintasuhde taas on matala, testin merkitys kasvaa; tällöin verraten alhaisenkin validiteetin omaava testi voi parantaa päätöksentekoa.

6.5 Reliabiliteetin ja validiteetin välinen suhde

Reliabiliteetti- ja validiteettitarkastelun päätteeksi on vielä syytä tutkia näiden kahden käsitteen välistä suhdetta. Reliabeli testi on mittausvirheetön, ja validi testi mittaa sitä ominaisuutta, mitä sen oletetaan mittaavan. Määritelmistä seuraa, että jos testin reliabiliteetti on alhainen, testi ei voi koskaan olla validi. Tällöinhän testi ei voi *luotettavasti* mitata sitä mitä sen on tarkoitettu mittaavan, ts. suuri osa mittaustuloksesta on mittausvirhettä eikä sitä ominaisuutta, jota haluamme mitata. Mutta, jos testin reliabiliteetti on korkea, testi ei välttämättä ole validi. Palauta mieleen esimerkiksi kuumemittari: Se mittaa luotettavasti ja virheettömästi kuumetta, mutta ei juurikaan muiden sairauksien esiintymistä. Korkea reliabiliteetti on siis validiteetin välttämätön muttei riittävä ehto. Tämä sääntö onkin muistettava aina, kun testejä käytetään päätöksenteon välineinä. *Testi ei koskaan ole validi mittaamaan muuta kuin sitä ominaisuutta, mitä se on suunniteltu mittaamaan.* Testejä ei siis tule käyttää muuhun kuin niiden alkuperäiseen tarkoitukseen, vaikka houkutus olisikin suuri². Monet sinänsä käyttökelpoiset testit ovat saaneet kyseenalaisen maineen, kun niitä on käytetty alkuperäisestä poikkeavaan tarkoitukseen. Testin manuaalissa täytyy *aina* lukea, mitä ominaisuutta testi mittaa. Jos sitä ei lue, testiin kannattaa suhtautua

²*Eikä* näin saa menetellä, vaikka muut työpaikalla niin tekisivätkin. Oma arvostelukyky on säilytettävä myös sosiaalisen paineen alla.

suurella varauksella.

Luku 7

Itemianalyysi

Monet psykologiset testit koostuvat joukosta testi-itemeitä¹, joihin testattavan tulee vastata. Testin arvioinnissa ei usein riitä, että arvioidaan kuinka reliabeli tai validi testi on. On pyrittävä löytämään syitä siihen, miksi testin käyttökelpoisuus on havaitun kaltainen. Jos testi koostuu useista itemeistä, on helppo päätellä jonkin noiden itemien ominaisuuden vaikuttavan testin reliabiliteettiin ja validiteettiin. Itemianalyysillä tarkoitetaan sellaista menetelmäjoukkoa, joiden avulla testi-itemien käyttökelpoisuutta voidaan arvioida. Itemianalyysin tarkoituksena on poistaa ja korvata huonosti toimivat itemit testistä. Testiitemit voivat olla huonoja monista eri syistä: Ne voivat olla huonosti muotoiltuja, ne voivat sisältää tarpeettomia kompieja tai siten ne yksinkertaisesti eivät mittaa kiinnostuksen kohteena olevaa sisältöaluetta.

Itemianalyysissa arvioidaan testi-itemeistä kolmea seikkaa:

1. vääriä vastausvaihtoehtoja

¹ Itemeistä käytetään myös nimitystä *osio*.

2. itemien vaikeutta sekä
3. itemien erottelukykä

7.1 Distraktorianalyysi

Väärin vastausvaihtoehtojen arviointimenetelmä tulee useimmiten kysymykseen monivalintatestien yhteydessä. Tyypillisessä monivalintatehtävässä vastausvaihtoehtoja vain yksi on oikea. Jos testiitemi on hyvin laadittu, niin niiden henkilöiden, jotka tietävät oikean vastauksen pitäisi myös valita oikea vastaus. Niiden jotka eivät tiedä taas tulisi valita vastauksensa täysin satunnaisesti kaikkien vastausvaihtoehtojen joukosta. Jos vastausvaihtoehtoja on 4 niin ideaalilanteessa $\frac{1}{4}$ 'tietämättömistä' valitsee oikean vastauksen, ja $\frac{3}{4}$ väärän vastauksen. Väärin vastausvaihtoehtojen arvioinnissa onkin kiinnitettävä huomiota vastausten jakautumiseen. Jos johonkin väärään vaihtoehtoon tulee huomattavan vähän vastauksia, se vähentää itemin erottelukykä. Samoin käy, jos jokin vastausvaihtoehto on huomattavan suosittu.

Ideaalitilanteessa vastauksen tietävä henkilö valitsee aina oikean vastauksen, ja tietämättömien vastaukset jakautuvat tasaisesti väärin vaihtoehtojen kesken. Distraktorianalyysissä asiaa tarkastellaan oikeiden ja väärin vastausten näkökulmasta. Kun tiedetään, kuinka monta henkilöä vastaa itemiin oikein ja kuinka monta vastausvaihtoehtoa itemissä on, voidaan laskea ns. väärin vastauksen odotusarvo p_i . Odotusarvo ilmoittaa, kuinka monen henkilön pitäisi valita kukin vääristä vaihtoehtoista:

$$p_i = \frac{\text{väärin vastanneiden lukumäärä}}{\text{väärin vaihtoehtojen lukumäärä}} \quad (7.1)$$

Esimerkki selvittää varmasti tilannetta: Oletetaan, että koehenkilöitä on 80, vastausvaihtoehtoja on yhteensä neljä ja p_i :ksi saadaan 20. Näinollen 20 henkilön pitäisi valita väärä vaihtoehto 1, 20 henkilön väärä vaihtoehto 2 ja 20 henkilön väärä vaihtoehto 3. Seuraavaksi tarkastellaan vastaus-ten todellista jakaumaa. Jos johonkin väärään vaihtoehtoon on tullut vastauksia huomattavasti vähemmän kuin 20, on itemi ollut selvästi muita vääriä vaihtoehtoja epäsuosituampi. Se on voinut olla esimerkiksi liian ilmeisesti väärä. Tämä ei ole toivottavaa, koska jos testattava ei tiedä oikeaa vastausta mutta tietää yhden vastauksen olevan varmasti väärä, hänen todennäköisyytensä arvata oikea vastaus kasvaa. Tällaisessa tilanteessa itemi muodostuu siis helpommaksi. Vastausvaihtoehtojen jakaumien tarkastelu voidaan suorittaa esimerkiksi χ^2 -testin avulla.

Toisaalta jokin väärä vaihtoehto on voi-

nut olla suhteettoman suosittu; silloin kyseessä on voinut olla esim. kompa tai harhauttava kysymys. Joskus - esimerkiksi hahmottaessa erotella ihmisiä jakauman yläpäässä - tällaisia kysymyksiä voidaan käyttää, mutta yleistykseenä todettakoon, että sekä suositut että epäsuositut väärät vaihtoehdot heikentävät testin luotettavuutta.

7.2 Itemin vaikeuskerroin

Itemien vaikeuden tarkasteleminen ei ole mitenkään yksinkertaista. Jotkin asiat ovat helppoja toisille ihmisille, jotkin muut taas toisille. Psykometriikassa itemien vaikeutta tarkastellaankin laskemalla, kuinka moni henkilö osaa vastata itemiin oikein. Itemin vaikeuskerroin voidaan laskea helposti:

$$\text{itemin vaikeuskerroin} = \frac{n_o}{N_k} \quad (7.2)$$

missä

n_o = oikein vastanneiden
lukumäärä

N_k = testattujen lukumäärä

Itemin vaikeuskerroin vaihtelee nollan ja ykkösen välillä. Huomaa, että itemin vaikeuskerroin mittaa tiettyjä ominaisuuksia sekä itemistä että populaatiosta joka tekee testin. Saman itemin vaikeuskertoimet voivat olla erilaisia eri populaatioille. Sama itemi voi esimerkiksi olla paljon helpompi aikuisille kuin lapsille. Itemien vaikeusker-
toimilla on suuri vaikutus testitulosten variaanssille, ts. testin erottelukyvylle. Mikäli

kaikkien itemien vaikeuskertoimet ovat lähellä nollaa, saa suurin osa testatuista henkilöistä matalan pistemäärän. Jos taas vaikeuskertoimet ovat lähellä ykköstä, saavat kaikki henkilöt testistä hyvin korkeat pisteet. Jos siis itemin vaikeuskerroin on lähellä ykköstä tai nollaa, itemiin vastaamisen suhteen ei ole olemassa minkäänlaista varianssia, ja itemi on testattavien erotte-
lun kannalta hyödytön. Käytännössä hajontaa syntyy eniten silloin, kun kaikkien itemien vaikeuskerroin on tasan 0.5.

Koska kaikkia testejä ei ole tarkoitettu mittaamaan yhtä hyvin koko skaalalla, joskus on tarpeen tehdä testi jossa itemien vaikeuskertoimet ovat joko hyvin pieniä tai hyvin suuria. Esimerkiksi tilanteessa, jossa halutaan ainoastaan *erotella* populaatiosta jokin erityisen huonosti suoriutuva joukko, kannattaa tehdä paljon helppoja itemeitä, jotka kuitenkin todennäköisesti tuottavat vaikeuksia tuolle erityisjoukolle. Sama pätee tietysti myös toiseen suuntaan. Jos testillä halutaan erotella parhaiten suoriutuva keskitasoisesti ja huonosti suoriutuvista, testiin pitää luonnollisesti laatia paljon vaikeita itemeitä, joihin ainoastaan parhaiten suoriutuvat osaavat vastata. Huomaa kuitenkin, että tällaisilla testeillä ei saada tietoa toisen ääripään suorituksista juuri lainkaan. Jos suurin osa ihmisistä osaa vastata kaikkiin itemeihin oikein, ei heidän osaamisistaan saada kovin tarkkaa kuvaa. Sama pätee myös silloin, jos juuri kukaan ei osaa vastata itemeihin oikein. Jos halutaan mitata koko populaation ominaisuuksia, itemien vaikeuskerroin tulee asettaa lähellä 0.5:a.

7.3 Itemin erottelukyky

Itemien käyttökelpoisuutta voidaan tarkastella myös sen perusteella missä määrin yksittäinen itemi erottelee "hyvät huonoista", ts. testissä parhaiten menestyvät huonoimmin menestyvistä tai korkeita pistemääriä saavat matalia pistemääriä saavista. Tätä itemin ominaisuutta tarkastellaan erotte-
luindeksin D avulla. Indeksien laskemiseksi määritetään testatuista kaksi ääriryhmää, esimerkiksi parhaimmat 25% ja huonoimmat 25%². Seuraavaksi määritetään, kuinka monta prosenttia kummastakin ryhmästä vastaa oikein kussakin testi-itemissä. Nyt D voidaan laskea kullekin itemille vähentämällä ylärühmän pistemäärä alaryhmän pistemäärästä:

$$D = \frac{U}{n_u} - \frac{L}{n_l} \quad (7.3)$$

missä:

- U = ylärühmässä oikein vastanneita
- n_u = henkilöitä ylärühmässä
- L = alaryhmässä oikein vastanneita
- n_l = henkilöitä alaryhmässä

Mitä suurempi D on, sitä enemmän oikeita vastauksia on ylärühmän jäsenillä. Pieni tai negatiivinen D taas tarkoittaa sitä, että alaryhmä on vastannut yhtä hyvin tai paremmin kuin ylärühmä. Täten voidaan todeta, että jos D on selvästi positiivinen, itemillä on vahva erottelukyky. Jos taas D on lähellä nollaa tai negatiivinen, itemin erottelukyky on huono.

²Myös muut persentiilit sopivat hyvin tähän tarkoitukseen

7.4 Korrelaatioiden tarkasteleminen

Kuten reliabiliteettitarkastelun yhteydessä esitetty sisäisen konsistenssin menetelmä osoittaa, on testi-itemien välisillä korrelaatioilla huomattava merkitys testin reliabiliteetille. Itemianalyysi lähestyy samaa problematiikkaa yksittäisen itemin näkökulmasta. Korrelaatioita tarkastelemalla pyritään selvittämään, mitkä testi-itemit eivät sovi testiin. Itemianalyysi tarjoaa tähän kaksi vaihtoehtoa. Voidaan joko tarkastella yksittäisen itemin ja koko testipistemäärän välistä korrelaatiota tai testi-itemien interkorrelaatioita.

7.4.1 Itemin ja testin kokonaispistemäärän välinen korrelaatio

Erotteluindeksi D kertoo, kuinka hyvin yksittäinen itemi mittaa samaa asiaa kuin koko testi, ts. kuinka hyvin itemi erottelee yksittäiset testattavat toisistaan. Samaa asiaa on mahdollista tutkia myös korrelaatiotarkastelun avulla: Lasketaan korrelaatio yksittäisen testi-itemin³ ja testin kokonaispistemäiden välille. Jos korrelaatio on suuri ja positiivinen, niin itemin voidaan katsoa mitaavan samaa asiaa kuin koko muukin testi, ja itemin erottelukyky on hyvä. Jos korrelaatio on lähellä nollaa tai negatiivinen, itemin erottelukyky on huono ja se voidaan pudottaa pois lopullisesta testistä.

³Monivalintatestin yhteydessä oikea vastaus on 1 piste ja väärä 0. Tällöin korrelaatioiden laskemiseen on käytettävä ns. *biseriaalisia korrelaatioita*

7.4.2 Itemien väliset korrelaatiot

On myös mahdollista tutkia yksittäisen itemin korrelaatioita muiden itemien kanssa. Tällä tavalla voidaan selvittää, miksi tietyn itemin korrelaatio testin kokonaispistemäärän välillä on heikko: Mikäli itemi ei korreloi minkään toisen itemin kanssa, voidaan olettaa että se ei mittaa samaa asiaa kuin muut itemit. Tällöin itemi on syytä poistaa tai muuttaa. Mikäli itemi taas korreloi muutamien toisten -mutta ei kaikkien- itemien kanssa, nämä itemit saattavat mitata muuta kuin testi kokonaisuutena. Tällöin on syytä miettiä, mittaavatko kaikki itemit loppujen lopuksi samaa ominaisuutta. Itemien välisten korrelaatioiden tarkastelulla voidaan siis selvittää *miksi* joidenkin itemien D on suuri ja joidenkin ei. Korrelaatioita voidaan kuitenkin joutua laskemaan suuri määrä, ja niiden tarkastelu voi siten olla monimutkaista. Käytännössä faktorianalyysi sopiikin yksittäisiä korrelaatioita paremmin laajojen testien tarkastelemiseen.

Luku 8

Testin rakentamisen perusteet

Koska testin laadintaan liittyy monia toisistaan irrallaankin olevia vaiheita, on ehkä hyvä esittää jonkinlaisena yhteenvetona tiivistelmä testin laadintaan liittyvistä työvaiheista. Seuraavassa on esitetty yksinkertaisen testin tai mittarin laatimisen työvaiheet siinä järjestyksessä kun ne on tarkoituksenmukaista suorittaa. Näitä ohjeita noudattamalla pitäisi saavuttaa melko hyvä lopputulos. Ohjeet on laadittu enemmänkin varsinaisia testejä silmällä pitäen, mutta ne soveltuvat hyvin myös yksinkertaisempien, esimerkiksi gradussa tarvittavien mittareiden laadintaan. Tässä esitettyjä testin laatimisen vaiheita voidaan luonnollisesti suorittaa myös rinnakkain.

- ta.
2. Suunnitellaan mittaus.
3. Suoritetaan esitestaus, jonka perusteella poistetaan ja korvataan toimimattomat testi-itemit.
4. Suoritetaan reliabiliteetin arviointi tarkoitukseen sopivalla menetelmällä. Tehdään tarvittavat korjaukset.
5. Suoritetaan validiteetin arviointi. Tehdään tarpeelliset korjaukset.
6. Tehdään testistä lopullinen versio ja normitetaan se.
7. Päivitetään testiä tarpeen mukaan.

8.1 Testin laatimisen vaiheet

1. Valitaan mittauskohde ja tehdään mahdollisimman kattava sisällöllinen käsiteanalyysi mittauksen kohteena olevasta asias-

8.1.1 Mittauskohteen valinta ja analysointi

Testin laatiminen on aloitettava mittauskohteen valinnalla ja teoreettisella määrittelyllä. Mittauskohteeksi pyritään valitsemaan jokin mahdollisimman selkeä teoreettisesti määriteltävissä oleva käsite, jonka

mittaamisen voidaan olettaa olevan mahdollista. Tässä vaiheessa tehdään mahdollisimman kattava käsitteellinen analyysi mitattavasta kohteesta. Samalla analysoidaan käsitteiden mahdollisesti hierarkkinen rakenne. Teorian avulla päätellään, millaisia käyttäytymisvasteita mitattavaan kohteeseen voi liittyä.

Alustavassa suunnittelussa on myös päätettävä, mihin testituloksia aiotaan käyttää. Tämä päätös vaikuttaa esimerkiksi laadittavien itemien vaikeuskertoimiin, testiltä edellytettävään reliabiliteettiin ja validiteettiin sekä normiaineiston kokoon ja laatuun.

8.1.2 Mittauksen suunnittelu

Kun mittauskohteen teoreettinen määrittely on suoritettu, on aloitettava varsinaisen mittalaitteen suunnittelu. Kun tiedetään, millaisia käyttäytymisvasteita mittauskohteeseen liittyy, voidaan laatia soveliaita testi-itemeitä käytännössä tapahtuvan mittauksen suorittamiseksi. Käytännössä itemeitä kannattaa suunnitella paljon enemmän kuin mitä lopulliseen testiin halutaan, koska kaikki laaditut itemit eivät välttämättä toimikaan halutulla tavalla. Tässä vaiheessa on myös päätettävä, mikä on testin kohdepopulaatio, koska eri kohdepopulaatioille voidaan tarvita eri tavoin toimivat testit. Itemeitä voidaan tehdä hyvin monenlaisia: esimerkiksi yksinkertaisia monivaikeuslaskutehtäviä (16PF) tai monimutkaisempia suorituksia vaativia (WAIS:n kuutiotehtävät). Taustalla oleva teoria määrääkin minikäytyyppejä testi-itemeitä on suunniteltava. Myös testin face validity on hyvä arvioida jo tässä vaiheessa.

Mittarin vaikeus on pyrittävä laatimaan siten, että testi on herkin erottelemaan juuri niitä henkilöitä, joita sen toivotaankin erottelevan. Koska itemin vaikeuskerrointa on mahdotonta määrittää etukäteen, kannattaa aluksi laatia eritasoisia itemeitä ja valita esitestauksen perusteella niistä parhaat lopulliseen testiin.

Mittauksen suunnittelussa on otettava huomioon, millä mitta-asteikolla mittaus halutaan suorittaa. Klassinen testiteoria edellyttää vähintään välimatka-asteikolla tapahtuvaa mittausta ja jatkuvia muuttujia. Jos mittaustulosten käsittelyssä halutaan soveltaa klassista teoriaa, ei mittausta voida suorittaa laatuer- tai järjestysasteikolla.

8.1.3 Esitestaus ja itemianalyysi

Kun testi-itemit on laadittu, niistä kootaan jonkinlainen esitesti. Esitestauksen tarkoituksena on pääasiassa tarkastella testin rakennetta ja eri itemien toimivuutta. Tässä vaiheessa suoritetaan siis itemianalyysi, jonka perusteella poimitaan testiin parhaiten toimivat itemit. Heikosti toimivia itemeitä voidaan muuttaa tai ne voidaan jättää kokonaan pois. Tavoitteena on saada johdonmukainen ja mittaustulokseltaan yhdensuuntaisista itemeistä koostuva testi. Samalla voidaan laskea alustavia ennusteita reliabiliteetille esimerkiksi α -kertoimen avulla. Jos α näyttää jäävän kovin alhaiseksi, pitää testi itemien homogeenisuutta parantaa tai itemien määrää kasvattaa.

Esitestauksen tavoitteena on muodostaa testistä melko lopullinen versio reliabiliteetin ja validiteetin arvioimista varten. Jos

itemeistä ei saada koottua edes kohtuullisen toimivalta vaikuttavaa testiä, kannattaa miettiä josko mittauksen voisi suorittaa jollain toisella tavalla. Lisäksi on mahdollista, että asian teoreettinen jäsentely ei ole onnistunut. Kaikki mahdolliset vaihtoehdot kannattaa tietenkin tutkia ennen kuin testin laatimisen aloittaa kokonaan alusta.

8.1.4 Reliabiliteetin estimointi

Reliabiliteetin arviointi alkaa olla mielekästä vasta siinä vaiheessa kun testi alkaa saada lopullisen muotonsa. Itemien olisi sitten oltava jo melko valmiita ja testin rakenteen tulisi muistuttaa melko pitkälle lopullista. Kun tämänkaltaisen testiversio on laadittu, voidaan reliabiliteetille muodostaa estimaatti.

Reliabiliteetin arvioimiseksi suoritetaan esitestaus, johon valitaan koehenkilöiksi edustava otos testin kohdepopulaatiosta. Käytettävä reliabiliteetin estimointimenetelmä riippuu tietenkin testistä ja sen käyttötarkoituksesta, mutta sisäisen konsistenssin menetelmät ja split-half -menetelmä ovat melko yleisesti sovellettavissa. Kahden testiversion käyttöön ei kannata turvautua ainakaan reliabiliteetin estimoinnin alkuvaiheessa.

Testiltä edellytettävä reliabiliteetin taso riippuu testillä tehtävien päätösten luonteesta. Tämä on otettava huomioon testin reliabiliteettia arvioitaessa. Kaikkien testien reliabiliteetin ei välttämättä tarvitse olla lähellä ykköstä. Mittauksen keskivirheen avulla on mahdollista arvioida, kuinka paljon mittausvirhe vaikuttaa testipistemäärien vaihteluun.

Jos mittausvirheestä aiheutuva vaihtelu todetaan liian suureksi, on testin suunnittelussa otettava askel taaksepäin. Esimerkiksi Spearman-Brownin kaavalla voidaan laskea, paljonko testiä täytyisi pidentää, jotta sen reliabiliteetti olisi halutun kaltainen. Jos näyttää siltä, että testiin tarvitaan kohtuuttoman suuri määrä uusia itemeitä, on kenties parempi koettaa laatia kokonaan uudentyyppisiä, johdonmukaisemmin toimivia itemeitä.

8.1.5 Validiteetin estimointi

Koska reliabiliteetti on validiteetin välttämätön mutta ei riittävä ehto, ei lopulliseen validiteettitarkasteluun kannata ryhtyä ennen kuin testistä on saatu riittävän reliabeili. Alustavaa validiteettitarkastelua on tehty jo aivan testin laadinnan alkuvaiheessa kun mitattavan sisältöalueen käsitteellistä rakennetta on hahmoteltu. Kun testistä on olemassa jo melko lopullinen versio, on tähän sisältövaliditeettitarkasteluun hyvä palata. On mahdollista, että itemien karsimisen ja muuttamisen takia nyt käsillä olevan testiversion sisältövaliditeetti ei olekaan enää kovin korkea.

Jos sisältövaliditeetti todetaan riittäväksi, voidaan siirtyä tarkastelemaan käsitevaliditeettia esimerkiksi multitrait-multimethod -lähestymistavalla. Jos laaditaan esimerkiksi älykkyystestiä, käsitevaliditeettia voidaan arvioida korreloimalla koehenkilöiden saamat pistemäärät nyt laadittavasta testistä ja esimerkiksi WAIS:sta.

Kriteerivaliditeetin arvioimisessa voidaan testin käyttötarkoituksesta riippuen käyttää joko ennuste- tai rinnakkaisvalidi-

teettä. Kriteerivaliditeetin arvioimisessa on otettava huomioon, että sen estimaattien ei oletetakaan nousevan kovin korkeiksi. Toisaalta taas pienessä aineistossa melko suurikin korrelaatio voi esiintyä sattumalta, joten turvallisinta on käyttää arvioinnin pohjana suurehkoa otosta.

8.1.6 Testin viimeistely ja normitus

Kun testi on läpäissyt itemianalyysin sekä reliabiliteetti- ja validiteettitarkastelun, voidaan testistä tehdä lopullinen versio. Mitä tavat standardoidaan lopulliseen muotoonsa ja testimateriaalin ulkoasu viimeistellään. Tämän jälkeen testi normitetaan kohdepopulaatiolle. Tarvittaessa kerätään esimerkiksi ikä- tai alueelliset normit.

Kun normiaineisto on valmiina, testille voidaan laatia normeihin ja johonkin transformaatioon perustuva pisteytysjärjestelmä. Lopuksi testistä laaditaan yksiselitteinen ja kattava ohjekirja, jossa esitetään testin käyttötarkoitus, laatimisperusteet, testiä koskevat tutkimukset, tärkeimmät psykometriset tunnusluvut, testin käyttöohjeet ja normit. Valmis testi laitetaan markkinoille, ja tarvittaessa testin käytöstä suunnitellaan koulutusta.

8.1.7 Testin päivittäminen

Kun testi on ollut käytössä jo jonkin aikaa, voi ilmetä tarvetta päivittää testi. Käyttäjäpalautteen perusteella voidaan tehdä pienempiä korjauksia seuraaviin painoksiin. Uusi normiaineisto voidaan myös kerätä

tarvittaessa. Lisäksi psykologian ja psykometriikan teorian edistymien saattaa aiheuttaa paineita uudistaa testi.

Hyvin laadittu testi voi olla käytössä kymmeniä vuosia¹, jos sitä pidetään jatkuvasti ajan tasalla. Testin laatimisessa on niin suuri työ, että kerran laadittu hyvä testi kannattaa ehdottomasti pitää käyttökunnossa mahdollisimman pitkään. Jossain vaiheessa testi tietenkin muuttuu auttamatta vanha-aikaseksi. Testi onkin ymmärrettävä vetää pois markkinoilta hyvissä ajoin ennenkuin se saavuttaa kyseenalaisen statuksen.

8.2 Pienempien testien ja mittareiden laatiminen

Edellä esitetty työjärjestys testien laatimisesta sopii parhaiten laajempien testien suunnittelemiseen ja rakentamiseen. Pienempiä projekteja varten ei tietenkään ole tarkoituksenmukaista käydä kaikkia vaiheita lävitse. Kuitenkin validiteetin ja reliabiliteetin tarkastelu, samoin kuin itemianalyysi on lähestulkoon välttämätöntä suorittaa kaikille laadittaville mittareille ja kyseilyille, sillä hyvin suoritettut mittaukset ovat edellytys luotettaville päätelmille.

¹Ikävä kyllä myös huonot testit saattavat pysyä markkinoilla pitkään

Osa III

Testiteorian sovelluksia

Luku 9

Älykkyystestit

Älykkyystestien kehittäminen liittyy läheisesti psykologian ja psykometriikan historiaan, ja tänä päivänä älykkyystestejä käytetään kenties enemmän kuin koskaan aikaisemmin. Älykkyuden tutkiminen ja testaaminen on kuitenkin erittäin haastavaa, ja tähän psykologisen testaamisen osaluokkaan liittyy edelleen lukuisia ongelmia.

9.1 Älykkyuden määrittelyminen

Psykologiassa on useita käsitteitä, joiden tarkasta määrittelystä ei ole saavutettu yksimielisyyttä. Tällaisia ovat esimerkiksi emootio, persoonallisuus ja *älykkyys*. Älykkyudesta on edelleen monia erilaisia teorioita, mutta tietyistä asioista lähes kaikki tutkijat ovat yksimielisiä. Ensinnäkin, älykkyys on *käsite*, ei *asia*. Siten on tarkkaan ottaen mahdotonta määritellä, millainen asia älykkyys on. Sen sijaan on pyrittävä määrittelemään, millaisena käyttäytymisenä älyk-

kyys ilmenee. Tämän käyttäytymisen perusteella voidaan luoda teorioita älykkyuden rakenteesta. Toiseksi, älykkyuden katsotaan liittyvän jollain tavalla yleiseen kognitiiviseen suorituskkyyn. Kolmanneksi, älykkyuden määrän oletetaan vaihtelevan eri ihmisillä.

Älykkyteen liittyvästä käyttäytymisestä ei kuitenkaan ole saavutettu yksimielisyyttä, kuten ei myöskään älykkyuden rakenteesta. Älykkyuden mittaaminen onkin erittäin monimutkaista. Koska älykkyuden yleensä oletetaan ilmenevän monenlaisina käyttäytymisen muotoina, mitataan älykkyyttä yleensä laajaan käyttäytymisotokseen perustuvilla älykkyystesteillä.

9.2 Älykkysteorioita

Älykkyuden rakenteesta on psykologian historian aikana esitetty lukuisia teorioita, ja useat teoriat ovat johtaneet erilaisen älykkyuden mittaamiseen tarkoitettujen testien kehittämiseen. Seuraavassa on esitelty muutamia psykometriikan kannalta

keskeisimpiä teorioita viimeisen sadan vuoden ajalta, vuosiluvut viittaavat teorian julkaisuvuoteen eivätkä mihinkään tiettyyn artikkeliin.

Charles Spearmanin *kahden faktorin teorian* (1904) mukaan ihmisen älykkyyteen vaikuttaa kaksi tekijää, yleinen faktori *g* ja spesifit faktorit *s*. Yleinen faktori vaikuttaa voimakkaasti kaikkeen älykkääseen käyttäytymiseen, kun taas spesifi faktori liittyy aina johonkin tiettyyn toimintaan. Koska spesifit faktorit liittyvät erilaisiin toimintoihin, johtuvat yksilölliset erot älykkyystesteissä yleisestä faktorista. Siten älykkyystestin pitäisi mitata mahdollisimman paljon yleistä faktoria.

Louis Leon Thurstonen *ryhmäfaktorien teoria* (1935) perustuu oletukseen, jonka mukaan älykkyys koostuu seitsemästä erilaisesta tekijästä, *verbaalisen ymmärtämisen, fluenssin, lukujen käsittelyn, spatiaalisen havaitsemisen, assosiativisen muistin, visuaalisen havaitsemisen ja järjestyksen* faktoreista. Faktorit ovat jossain määrin - mutta eivät suoraan - yhteydessä yleisfaktoriin *g*. Faktorit ovat myös melko riippumattomia toisistaan, joten tämän teorian pohjalta laaditun älykkyystestin on mitattava jokaisista faktoria erikseen.

David Wechslerin (1969) mukaan älykkyys on erityisesti kokonaisvaltainen kyky, sillä se vaikuttaa ihmisen käyttäytymiseen yhtenä, jakamattomana kokonaisuutena. Älykkyys koostuu erilaisista osataidoista, mutta älykkyyttä ei voida käsittää pelkästään osataitojen summana. Älykkyys riippuu myös osataitojen ryhmittymisestä, sillä tietyt osataidot ovat toisia merkityksellisempiä. Wechsler liittää älykkyyteen

myös muita kuin kognitiivisia ominaisuuksia, esimerkiksi sosiaalisen sopeutumiskyvyn. Wechsler työtovereineen on laatinut teorian pohjalta älykkyystestin nimeltä WAIS, jonka uudistettu laitos WAIS-R on tällä hetkellä Suomen käytetyin älykkyystesti. Nykyinen *älykkyysosamäärän* käsite on peräisin juuri WAIS-testistä. Tästä syystä WAIS-R esitellään tässä yhteydessä laajemmin.

Robert Sternbergin (1984) *triarkkisesa älykkyysteoriassa* älykkyys määritellään huomattavan laajasti. Perinteisten älykkyystestien mittaamaa älykkyyttä kutsutaan teoriassa *komponentiaaliseksi älykkyudeksi*. Se tarkoittaa siten kykyä ongelmanratkaisuun ja tiedonkäsittelyyn, eli hyvin pitkälti niitä ominaisuuksia, joita WAIS-R mittaa. *Kontekstuaalinen älykkyys* on yksilön kyky sopeutua ympäristöönsä, ja kokeuksellinen älykkyys tarkoittaa kykyä soveltaa opittua tietoa jokapäiväisessä toiminnassa. Älykkyys ei siis Sternbergin mukaan ole pelkästään koulumenestystä, vaan laajemmin kykyä selviytyä elämässä. Sternbergin teoriaan perustuvia älykkyystestejä on olemassa, mutta toistaiseksi niistä ei ole olemassa kattavia suomenkielisiä versioita.

9.3 Wechsler Adult Intelligence Scale Revised

WAIS-R on Suomen käytetyin aikuisten yleisälykkyuden mittari. Vastaava lapsille tarkoitettu testi on nimeltään WISC-R, ja esikouluikäisten vastaava testi on nimeltään WPPSI-R. Älykkyystesti- nimestään huoli-

Kielellinen osa	Suoritusosa
Yleistietous Numerosarjat Sanavarasto Laskutehtävät Yleinen käsityskyky Samankaltaisuudet	Kuvien täydentäminen Kuvien järjestäminen Kuutiotehtävät Kokoamistehtävät Merkkikoe

Taulukko 9.1: WAIS-R:n osatestit

matta WAIS-R ei ole mikään Mensa-testi. Sitä ei käytetä huippulahjakkuuksien seulontaan, vaan pääasiassa heikompien suoritusten löytämiseen. WAIS-R soveltuukin siis ainoastaan kliiniseen käyttöön.

WAIS-R koostuu 11 osatestistä, joista 6 kielellisiä (ns. kielellinen osa) ja 5 ei-kielellisiä (ns. suoritusosa). Testistä voidaan laskea erikseen suoritusosan ja kielellisen osan pisteet, sekä yhteenlaskettu yleisälykkyuden pistemäärä. Tämä summa voidaan muuntaa *älykkyysosamääräksi* eli *ÄO*:ksi käyttäen hyväksi ikänormeja. ÄO on standardoitu pistemäärä, jonka keskiarvo on 100 ja keskihajonta 15. Eri ihmisten älykkyysosamääriä on siten helppoa vertailla.

Testi esitetään osatesti kerrallaan, ja testaaminen voi tapahtua useammalla kerralla. Kunkin osatestin tehtävät esitetään vaikeutuvassa järjestyksessä; kun testattava epäonnistuu tarpeeksi monessa peräkkäisessä suorituksessa, osatesti lopetetaan ja siirrytään seuraavaan osatestiin. Taulukossa 9.1 on esitetty WAIS-R:n osatestit.

9.3.1 Kielellinen osa

Yleistietous-osatestissä on kysymyksiä, joihin vastataan yhdellä tai muutamalla sanalla; esim. kuinka monta kiloa on tonnissa?

Numerosarjat ovat ns. digit span-tehtäviä, jotka mittaavat työmuistia. Niissä tulee muistaa numerosarjoja annetussa tai päinvastaisessa järjestyksessä.

Sanavarasto-osatestissä testattavan tulee selittää erilaisten sanojen merkityksiä, esim. mitä tarkoittaa sana *jänishousu*.

Laskutehtävät ovat sanallisia päässä-laskutehtäviä, joiden suorituksessa on aikarajoitus.

Yleinen käsityskyky-osio sisältää erilaisia miksi - ja mitä merkitsee -kysymyksiä; esim. Miksi ihmisten on maksettava veroja?

Samankaltaisuudet-osioissa testattavan pitää keksiä sanapareille yhteinen nimittäjä eli yläkäsite; esim. vadelma ja mansikka ovat marjoja.

9.3.2 Suoritusosa

Kuvien täydentämisessä tehtävänä on puuttuvan osan löytäminen kuvasta annetussa ajassa.

Kuvien järjestäminen-osatestissä tehtävänä on sekoitetun kuvasarjan asettaminen loogiseen järjestykseen annetussa ajassa.

Kuutiotehtävät-osiossa testattavalle näytetään kuvio, joka hänen tulee muodostaa puna-valkoisista kuutioista annetussa ajassa.

Kokoamistehtävät ovat erilaisten palapelin rakentamisesta koostuvia tehtäviä.

Merkkikokeessa testattavan pitää korvata lomakkeella olevat numerot toisilla symboleilla annetun järjestelmän mukaan. Tarkoituksena on tehdä annetussa ajassa mahdollisimman monta oikeaa suoritusta.

Testin tekeminen kestää testattavasta riippuen noin reilun tunnin. WAIS:sta on olemassa myös vastaavat versiot lapsille ja esikoululaisille. WAIS:n osatestien reliabiliteetit ovat yleisesti ottaen korkeita. WAIS:n sisältö- ja käsitevaliditeetit on myös arvioitu hyviksi useissa tutkimuksissa. Henkilöiden WAIS:lla mitatut älykkyysosamäärät ovat myös yhteydessä mm. koulu- ja työmenestykseen, joten testin kriiterivaliditeetti on kohtuullisen korkea.

9.3.3 WAIS-R:n heikkoudet

Vaikka WAIS on psykometrisilta ominaisuuksiltaan erinomainen, testin käyttämiseen liittyy kuitenkin tiettyjä ongelmia. Suurin näistä on kenties testitulosten käyttö. WAIS toimii parhaiten erotteludiagnostisena välineenä, mutta sillä saatujen tulosten käyttökelpoisuudesta voidaan olla monta mieltä. Jos jonkun henkilön älykkyysosamäärä WAIS:lla mitattuna on alhainen, niin mitä tällä tiedolla voidaan tehdä? WAIS ei sinällään kerrokaan mitään yksilön ongelmien laadusta. WAIS:n ja muiden älykkyystestien käyttäminen onkin useimmiten perusteltua ainoastaan muun diagnostiikan yhteydessä. Lisäksi WAIS-R -testissä älykkyys on määritelty melko suppeasti verrattuna esimerkiksi Sternbergin määritelmään. Tulokset eivät siis välttämättä ole yleistettävissä kovinkaan laajasti yksilön toimintaan arkipäivän tilanteissa.

9.4 Älykkyystestien ongelmia

Huolellisesta suunnittelusta huolimatta älykkyystestit eivät välttämättä ole tasapuolisia arvioidessaan erilaisia ihmisiä. Joidenkin testien tulokset saattavat sisältää systemaattista virhettä. Tästä testien ominaisuudesta käytetään nimitystä *harhaisuus* (engl. *bias*).

9.4.1 Mittauksen harhaisuus

Yleisesti ottaen on havaittu, että älykkyystestien tapauksessa

- 1) Lapsen sosioekonominen tausta on yhteydessä testisuoritukseen
- 2) Valkoihoiset menestyvät testeissä etnisiä vähemmistöjä paremmin
- 3) Tietyillä testeillä arvioituna naiset saavat miehiä korkeampia pistemääriä, toisilla testeillä arvioituna taas tilanne on päinvastainen.

On huomattava, että testin harhaisuus on ainoastaan mittalaitteen *tilastollinen ominaisuus*. Mittaus on suunniteltu huonosti, koska se tuottaa systemaattista virhettä mitaustulokseen. Monet älykkyystestit on laadittu yhteiskunnan kulttuurisen enemmistön näkökulmasta, ja ne mittaavat tämän kulttuurin arvostamia taitoja. Vastaavat taidot eivät kuitenkaan välttämättä ole samalla tavalla arvostettuja joissain vähemmistökulttuureissa. Tällaisen testin sanotaan ole-

van *kulttuurisidonnainen*, sillä siinä menestymisen edellyttää valtakulttuurin tapojen ja arvojen tuntemista. Testien kulttuurisidonnaisuutta voidaan vähentää mm. non-verbaaleilla itemeillä, suullisilla vastauksilla ja täysin uudentyyppisiin ongelmanratkaisuun perustuvilla tehtävillä. Esimerkiksi WAIS on melko kulttuurisidonnainen, kun taas Raven Progressive Matrices taas vähemmän kulttuurisidonnainen. Ravenin ongelma on kuitenkin siinä, että se mittaa älykkyyttä vain yhdellä tavalla.

9.4.2 Ennusteen harhaisuus

Mittauksen harhaisuuden lisäksi voidaan käsitellä myös *ennusteen harhaisuutta*, jolla tarkoitetaan testin perusteella tehtyjen ennustusten systemaattista virheellisyyttä. Esimerkiksi lasten älykkyystesti saattaa ennustaa keskiluokkaisten lasten koulumenestyksen järjestelmällisesti paremmaksi kuin alemman sosioekonomisen taustan omaavien lapsien. Ennusteen harhaisuus ei ole välttämättä yhteydessä mittauksen harhaisuuteen. Myös täysin harhattomien mitaustulosten perusteella voidaan tehdä harhaisia ennusteita.

9.4.3 ÄÖ:n muuttuvuus

Ihmisen psyykkisen kapasiteetin muuttuvuus tai muuttumattomuus ei ole yksinkertaisesti selvitettävissä oleva asia. Sekä ympäristö että perimä vaikuttavat tämänkaltaisten ominaisuuksien kehittymiseen, mutta on mahdotonta arvioida tarkkaan missä määrin kumpikin vaikuttaa. On selvää, että kaikista ihmisistä ei kehity esimerkiksi

huippuluokan kapellimestareita edes äärimmäisen intensiivisellä harjoittelulla. Tästä huolimatta harjoittelulla voidaan saavuttaa huomattavaa kehitystä: lähes kaikki ihmiset oppivat riittävän harjoittelun tuloksena soittamaan auttavasti esimerkiksi pianoa.

Myöskään älykkyysosamäärä ei ole mikään pysyvä ominaisuus, vaan yhdellä tietyllä hetkellä saatu mitaustulos. Pitkällä aikavälillä henkilön ÄÖ saattaa muuttua suurestikin. Tämä aiheuttaa kuitenkin päätöksentekoon liittyviä ongelmia, sillä yhdellä hetkellä mitatun ÄÖ:n perusteella saatetaan tehdä hyvinkin suuria yksilöä koskevia päätöksiä. Yleensä ottaen pitkittäistutkimuksissa eri ikäisinä suoritettut älykkyysosamäärän mittaukset kuitenkin korreloivat melko voimakkaasti toistensa kanssa. Älykkyyden muunneltavuus onkin kenties nähtävä siten, että perimä asettaa rajat, joiden puitteissa ympäristötekijät voivat kehittää älykkyyttä.

Luku 10

Persoonallisuustestit

Persoonallisuuden tutkiminen on vähintäänkin yhtä vaikeaa kuin älykkyydenkin. Persoonallisuus on kuitenkin erityyppinen käsite kuin älykkyys, joten persoonallisuustestien laatimiseen ja käyttämiseen liittyykin erityyppisiä ongelmia.

10.1 Persoonallisuuden määritteleminen

Persoonallisuuden määritelmästä ei ole päästy tutkijoiden kesken yksimielisyyteen. Kuten älykkyyden määrittelyssä, niin myös persoonallisuuden määrittelemisessä on tiettyjä seikkoja, joista ollaan melko yksimielisiä. Ensinnäkin, jokaisen henkilön persoonallisuus on *ainutlaatuinen*. Kahta samanlaista persoonallisuutta ei ole olemassa. Silti ihmiset muistuttavat persoonallisuudeltaan toisiaan, joten *samantyyppisiä* persoonallisuuksia on olemassa. Toiseksi, persoonallisuus oletetaan ainakin jossain määrin pysyväksi ominaisuudeksi. Kolmanneksi, persoonallisuus määrittelee hyvin laajasti yksilön toimintaa. Silti sama

henkilö voi toimia eri tavalla samanlaisissakin tilanteissa. Persoonallisuudella ei siten ole determinististä roolia ihmisen toiminnan ohjaajana. Useimmat persoonallisuusteoriat sijaitsevat jossain kahden ääripään, piirreteorioiden ja situationaalisten teorioiden välillä. Puhtaat piirreteoriat eivät ota lainkaan huomioon tilanteen vaikutusta ihmisen käyttäytymiseen, kun taas puhtaasti situationaaliset teoriat perustuvat pelkästään tilanteen vaikutusten arvioimiseen. Erilaisista teoreettisista lähtökohdista voidaan siis muodostaa erityyppisiä persoonallisuustestejä.

10.2 Miten persoonallisuustestejä voidaan suunnitella?

Teoreettisesti muodostetut testit on laadittu puhtaasti jonkin persoonallisuuden teorian pohjalta. Tällaiset testit pyrkivät mittaamaan sellaista persoonallisuuteen liittyvää käyttäytymistä, jota teorian perusteella

voitaisiin olettaa havaittavan. Testien validiteetti riippuukin siten hyvin paljon käytetyn persoonallisuusteorian laadusta.

Empiirisesti muodostetut testit perustuvat kriteeriryhmien käyttöön testin laadinnassa. Testiin valitaan sellaisia itemeitä, jotka erottelevat tehokkaasti eri ryhmät toisistaan. Tällaista lähestymistapaa on sovellettu esimerkiksi MMPI:ssa. Näiden testien avulla ei kuitenkaan voida tutkia *miksi* eri kriteeriryhmät vastaavat testissä eri tavoin. Kliinisessä diagnostiikassa tällaiset testit ovat kuitenkin käyttökelpoisia.

Faktorianalyttiset testit on laadittu suoriittamalla faktorianalyysi testi-itemeille. Faktorit (skaalat, persoonallisuuden piirteet) nimetään sen mukaan, millaisia testi-itemeitä niille latautuu. Jos testin suunnittelussa ollaan huolellisia, saadaan faktorianalyttisesti laadituista testeistä psykometrisiltä ominaisuuksiltaan erinomaisia. Faktorianalyttisten testien ongelma on kuitenkin se, että niissä skaalat luokitellaan enemmän tilastollisten menetelmien kuin psykologisen teorian avulla. Näinollen testien soveltaminen kliniseen käyttöön voi olla vaikeaa.

Persoonallisuustestit voivat olla joko subjektiivisia¹ tai objektiivisia. Subjektiivinen mittaamenetelmä on ollut suosittu nimenomaan persoonallisuustestien kehittämisessä, koska on katsottu, että objektiivisilla testeillä ei voida mitata niin monimutkaisia käsitteitä kuin persoonallisuus. Objektiivisilla testeillä saadaan kuitenkin helpommin *harhattomia* mittauksia kuin pro-

jektiivisilla testeillä. Lisäksi objektiiviset testit ovat helpompia ja nopeampia käyttää. Esimerkiksi Ro:n (subjektiivinen testi) pisteyttäminen saattaa kestää yli tunnin. 16PF (objektiivinen testi) taas on pisteytettävissä muutamassa minuutissa. 16PF on malliesimerkki hyvin laaditusta persoonallisuustestistä, joten sitä käsitellään seuraavassa laajemmin.

10.3 Sixteen Personality Factors Test (16PF)

Vuonna 1949 julkaistu Sixteen Personality Factors eli 16PF on Raymond Cattellin kehittämä persoonallisuustesti. 16PF onkin esimerkkinä juuri sen eleganttien psykometristen ominaisuuksien vuoksi. 16PF perustuu oletukseen ns. persoonallisuuden piirteiden olemassaolosta. Jokaisen ihmisen persoonallisuuden oletetaan koostuvan tietyistä määrästä piirteitä, ja jokaisen yksittäisen piirteen voimakkuus vaihtelee eri ihmisillä.

Cattellin tutkimusryhmä aloitti 16PF:n laatimisen etsimällä mahdollisimman suuren määrän adjektiiveja, jotka kuvaavat ihmisen persoonallisuutta. Tarkoituksena oli faktorianalyysin avulla redusoida tämä adjektiivien joukko pienemmäksi määräksi persoonallisuutta kuvaavia adjektiiveja. Aluksi adjektiivit - joita oli yli 4000 - jaettiin 45 luokkaan niiden lingvististen ominaisuuksien perusteella. Faktorianalyysien lopputuloksena oli 16 toisistaan riippumattonta faktoria, persoonallisuuden piirrettä, joiden Cattell katsoi vaikuttavan suurim-

¹Subjektiivisista testeistä käytetään myös nimitystä *projektiiviset* tai *impressionistiset* testit

Ensimmäisen tason faktorit		Toisen tason faktorit
varautuneisuus	luottavuus	introversio
älykkyys	käytännöllisyys	ahdistuneisuus
emotionaalinen stabiilitetti	suoruus	emotionaalinen tasapaino
kuuliaisuus	itsevarmuus	riippuvuus
maltillisuus	konservatiivisuus	luonnollisuus
tunnollisuus	itsenäisyys	realistisuus
uskaliaisuus	kontrolli	
hentomielisyys	rentoutuneisuus	

Taulukko 10.1: 16PF:n mittaamat persoonallisuuden piirteet

paan osaan ihmisen käyttäytymistä. Nämä ovat ns. ensimmäisen tason faktorit. Suorittamalla faktorianalyysia muille persoonallisuustesteille, löytyi vielä kuusi merkittävää faktoria, ns. toisen tason faktorit. Tämän jälkeen laadittiin monivalintakysymyksistä muodostuva testi, jonka jokainen kysymys liittyy johonkin persoonallisuuden piirteeseen. Cattell on myös testin manuaalissa kuvannut, miten eri piirteet ilmenevät psykologisina ominaisuuksina. Taulukossa 10.1 on esitetty 16PF:n mittaamat persoonallisuuden piirteet.

16PF on itsearviointitesti. Testattava vastaa lomakkeella oleviin kysymyksiin, joiden oletetaan mittaavan noita kuuttatoista eri persoonallisuuden piirrettä. Lomakkeessa esitetään erilaisia väitteitä, joihin on kolme kyllä - ei - en osaa sanoa - tyyppistä vastausvaihtoehtoa. Jokaiseen kysymykseen tulee vastata, ja EOS- vaihtoeht-

don valitsemista ei suositella. Jokainen vastaus liittyy aina johonkin persoonallisuuden piirteeseen, ja vastausten avulla henkilön persoonallisuudesta voidaan piirtää eräänlainen profiili. Testin manuaalissa on kuvattu joitain keskiarvoprofiileja eri ammattiryhmiin kuuluville henkilöille.

16PF:ta on käytetty sekä kliinisessä arvoinnissa, että henkilöstövalinnoissa. Ongelma on kuitenkin testin tulosten vaikea sovellettavuus. On vaikea sanoa, millä tavoin testin antama persoonallisuuden profiili on yhteydessä henkilön todelliseen käyttäytymiseen. Teknisesti 16PF on kuitenkin erinomaisesti laadittu, joten se onkin ollut suosittu nimenomaan tutkimuskäytössä.

10.4 Persoonallisuustestien ongelmia

Persoonallisuustestien käyttämiseen liittyy monia samoja ongelmia kuin älykkyyks-testeihin. Myös persoonallisuustestien mit-taukset ja estimaatit voivat olla harhai-sia. Persoonallisuuden testaamiseen liittyy myös tiettyjä ongelmia, jotka ovat ominai-sia tämäntyyppisille testeille.

10.4.1 Sosiaalinen suotavuus

Persoonallisuustestejä tekevät ihmiset ovat yleensä tietoisia siitä, mihin mittaustulos-ta aiotaan käyttää. Koska persoonallisuus-testeissä ei ole varsinaisia oikeita vastauk-sia testattava saattaa antaa sellaisia vastauk-sia, joiden hän olettaa olevan *sosiaalisesti suotavia*. Persoonallisuustestit saattavat siis mitata persoonallisuuden lisäksi myös hen-kilön oletuksia 'toivottavasta' persoonalli-suudesta. Joihinkin testeihin on rakennettu tällaista vastaamistapaa arvioivia skaaloja, mutta yritykset eivät toistaiseksi ole olleet erityisen onnistuneita.

10.4.2 Satunnainen vastaami-nen

Toinen testivastauksiin liittyvä ongelma on *satunnainen vastaaminen*. Testattava saat-taa tuottaa vastauksensa täysin satunnaises-ti miettimättä kysymyksiä ollenkaan. Täl-laiset testivastaukset on huomattavasti hel-pompi tunnistaa kuin sosiaalista suotavuut-ta tavoittelevat. Testiin on melko helppo laatia kysymyksiä, joista jokin vaihtoeh-

to on täysin mieletön. Jos henkilö valitsee useita tällaisia vaihtoehtoja, on kyseessä to-dennäköisesti satunnainen vastaaminen ja testin tekemistä voidaan yrittää uudelleen.

10.4.3 Tulkittavuuden ongelma

Persoonallisuustestien klinisen käytön on-gelma on usein niiden antamien tulosten soveltaminen. Kriteeriryhmien erotteluun käytettävien klinisten testien kuten MM-PI:n kohdalla tällaista ongelmaa ei ole, mutta esimerkiksi 16PF:n ja Ro:n tulosten tulkitseminen onkin jo paljon monimutkai-sempaa. Ei ole itsestään selvää, miten tie-tynlainen persoonallisuuden piirteiden voi-makkuuksien yhdistelmä ilmenee yksilön toiminnassa. Lisäksi persoonallisuustestit antavat yksinkertaisesti niin yleisellä tasol-la olevaa tietoa, että sen käyttöarvo voi ol-la melko olematonta. Tämän vuoksi per-soonallisuustestien käyttäminen esimerkiki-si rekrytoinnissa voi olla arveluttavaa. Klii-nisessä työssä diagnostikkaan käytettävät testit puolustavat kuitenkin edelleen paik-kaansa psykologin työvälineinä.

Luku 11

Yleisimmät Suomessa käytetyt testit

Psykologi joutuu monilla työpaikoilla käyttämään mitä erilaisimpia testejä. Koska erilaisia- ja eri taseisia testejä on erittäin runsaasti, on alustava tutustuminen hyvä aloittaa jo tässä vaiheessa. Klinikakursilla testeihin tutustutaan laajemmassa mitakaavassa. Seuraavassa on esitelty lyhyesti muutamia yleisiä Suomessa käytettyjä testejä. Testeistä on esitetty seuraavat ominaisuudet:

1. Käyttötarkoitus (kliininen arviointi, kykykartoitus tms.)
2. Testaukseen kuluva aika
3. Testityyppi (tehtävä - observointi - itsearviointi / maksimaalinen suoritus - tyypillinen suoritus / subjektiivinen - objektiivinen)
4. Kohderyhmä
5. Testin rakenne lyhyesti

Kaikki seuraavassa esiteltävät testit¹ löyty-

¹Stanford-Binet- testiä ei ole suoraan käännet-

vät tällä hetkellä psykologian laitoksen testiarkistosta, missä niihin voi käydä tutustumassa. Eräs asia on kuitenkin syytä huomioida jo tässä vaiheessa: kaikki seuraavista testeistä *eivät* ole ns. psykometristen periaatteiden mukaan rakennettuja, eivätkä perustu klassisiin tai moderneihin testiteorioihin. Joidenkin testien kohdalla esim. reliabiliteetin ja validiteetin tarkastelu on joko jätetty tekemättä, tai testit on rakennettu niin että se on mahdotonta. Vaikka jokin testi esiteltäisiinkin tässä yhteydessä, ei kyseistä testiä silti välttämättä kannata käyttää. Tämän monisteen tietojen perusteella on kuitenkin mahdollista päätellä, mitkä seuraavista testeistä sopivat käytettäväksi tutkimus- ja asiakastyössä.

Kaikki eivät myöskään välttämättä ole kaikkein eleganteimpia ja tarkimpia mahdollisia mittareita. Kaikkia testejä *ei* siis millään muotoa suositella käytettäväksi.

ty Suomeen. On kuitenkin olemassa suomenkielinen *Stanford-Binet-tyyppinen testistö*.

Suomessa käytetään niin lukuisia erilaisia testejä, että niiden kaikkien esitteleminen tässä yhteydessä on mahdotonta. Seuraava valikoima antaakin lähinnä käsityksen testaamiskäytäntöjen monipuolisuudesta.

11.1 Kyky / suoritustestejä

Tähän kategoriaan kuuluvat testit, joiden avulla voidaan mitata henkilön suorituksen laatua tai määrää jollain osa-alueella. Suurin osa testeistä on jonkinlaisia älykkyyss-testejä ta niihin rinnastettavissa olevia testejä.

11.1.1 Raven Progressive Matrices (Raven, RPM)

Käyttötarkoitus: älykkyyss testi

Tyyppi: tehtävä / maksimaalisen suorituksen arviointi / objektiivinen

Kesto: noin 45 minuuttia

Kohderyhmä: lapset ja aikuiset

Rakenne: Raven mittaa henkilön yleistä älykkyyttä (g-faktoria). Testivihossa on erilaisia tehtäviä, joissa pitää joko valita kuvioon sopiva pala tai kuvio, joka jatkaa loogisesti edellisten kuvioden sarjaa. Ravenin testitulokset korreloivat melko voimakkaasti WAIS:n tulosten kanssa, joten Raven sopii käytettäväksi tutkimuksissa, joissa halutaan nopeasti karkea arvio henkilön älykkyydestä.

11.1.2 Stanford-Binet (SB)

Käyttötarkoitus: älykkyyss testi

Tyyppi: tehtävä / maksimaalisen suorituksen

sen arviointi / objektiivinen

Kesto: noin 90 minuuttia

Kohderyhmä: lapset ja nuoret

Rakenne: SB mittaa yleisälykkyyttä, jonka oletetaan jakautuvan lyhytkestoiseen muistiin, analyyttisiin taitoihin ja erityisiin taitoihin. Jokaisesta osa-alueesta on erilaisia osatestejä, joista esitetään testattavan kykyjen mukaan tarvittava määrä. SB:n vahvuus on sen ikä: testin perusluonne on pysynyt muuttumattomana jo lähes sata vuotta, joten sitä on tutkittu paljon. SB on ensimmäinen varsinainen psykologinen älykkyyss testi.

11.1.3 WAIS-R

Käyttötarkoitus: älykkyyss testi

Tyyppi: tehtävä / maksimaalisen suorituksen arviointi / objektiivinen

Kesto: noin 60 minuuttia

Kohderyhmä: yli 16-vuotiaat

Rakenne: Testi koostuu 11 osatestistä, joista 6 kuuluu kielelliseen ja 5 ei-kielelliseen skaalaan. WAIS-R:lla voidaan mitata erikseen henkilön kielellinen ja ei-kielellinen älykkyyssosamäärä (ÄO), sekä näistä laskettu yhteinen yleinen älykkyyssosamäärä ($\bar{x} = 100, s = 15$). Osatestien suoritusprofiilin perusteella voidaan lisäksi kartoittaa erityisiä heikkoja tai vahvoja alueita. Henkilön ÄO:ta voidaan verrata normeihin, ja siten tutkia hänen kehitysvajaavuuksiaan. WAIS-R:sta on olemassa kattavat kotimaiset ikänormit. Huomaa: kaikki Wechslerin testit ovat rakenteeltaan ja periaatteiltaan melko samanlaisia.

11.1.4 WISC-R

Käyttötarkoitus: lasten älykkyystesti

Tyyppi: tehtävä / maksimaalisen suorituksen arviointi / objektiivinen

Kesto: noin 90 minuuttia

Kohderyhmä: 6-16-vuotiaat

Rakenne: Testi on pääosin samanlainen kuin WAIS-R, mutta tehtävät on sovittettu nuoremmille.

arviointi / subjektiivinen

Kesto: noin 60 minuuttia

Kohderyhmä: CAT lapset, TAT aikuiset

Rakenne: Testit koostuvat tauluista, joissa on kuvattu erilaisia tapahtumia. Testattava pyydetään kertomaan jokaisesta taulusta mitä kuvassa tapahtuu. Vastauksissa ilmi tulleiden teemojen perusteella tehdään päätelmiä testattavan sosiaalisten tilanteiden tulkinnasta.

11.1.5 WPPSI-R

Käyttötarkoitus: esikouluikäisten älykkyystesti

Tyyppi: tehtävä / maksimaalisen suorituksen arviointi / objektiivinen

Kesto: noin 90 minuuttia

Kohderyhmä: 4-6½-vuotiaat

Rakenne: kuin WAIS-R ja WISC-R, mutta tehtävät on sovitettu nuoremmille.

11.2.2 Cartoon Attribution Strategy Test (CAST)

Käyttötarkoitus: persoonallisuuden sekä ajattelu- ja tulkintatapojen arvioiminen.

Tyyppi: tehtävä / Tyypillisen suorituksen arviointi / subjektiivinen

Kesto: noin 30 minuuttia

Kohderyhmä: yli 12-vuotiaat

Rakenne: Testi koostuu kolmesta neljän kuvan kuvasarjasta, joissa esitetään jokin tapahtuma. Puolet sarjoista päättyy päähenkilön onnistumiseen, puolet epäonnistumiseen. Sarjat näytetään testattavalle yksi kuva kerrallaan, ja häntä pyydetään kertomaan mitä kuvassa tapahtuu, mitä henkilö ajattelee. Vastausten perusteella arvioidaan henkilön ajattelu- ja tulkintatapojen rakennetta. CAST on Suomessa laadittu testi.

11.2 Persoonallisuustestejä

Tämäntyyppiset testit ovat yleisesti käytettyjä kliinisessä asiakastyössä. Mittaaminen on useammin laadullista kuin määrällistä, ja monet näistä testeistä ovat subjektiivisesti tulkittavia.

11.2.1 Children's Apperception Test (CAT) ja Thematic Apperception Test (TAT)

Käyttötarkoitus: persoonallisuuden dynamiikan arvioiminen

Tyyppi: tehtävä / tyypillisen suorituksen

11.2.3 Minnesota Multiphasic Personality Inventory (MMPI)

Käyttötarkoitus: psykopatologian tutkiminen

Tyyppi: itsearviointi / tyypillisen suorituksen

sen arviointi / objektiivinen

Kesto: noin 60 minuuttia

Kohderyhmä: nuoret ja aikuiset

Rakenne: MMPI koostuu 566 pitää paikkansa - ei pidä paikkaansa - en osaa sanoa -tyyppisestä väittämästä. Tiettyjen vastausten on todettu tilastollisesti liittyvän tiettyihin psyykkisiin häiriöihin. Vastauksista lasketaan summapistemäärät seuraaville kymmenelle häiriöskaalalle: hypokondria, depressio, hystertia, patologinen poikkeavuus, maskuliinisuus-feminiinisyys, paranoia, psykastenia, skitsofrenia, hypomania ja sosiaalinen introversio.

11.2.4 Personality Research Form (PRF)

Käyttötarkoitus: persoonallisuuden rakenteen arvioiminen.

Tyyppi: itsearviointi / tyypillisen suoritusten arviointi / objektiivinen

Kesto: noin 25 minuuttia

Kohderyhmä: aikuiset

Rakenne: Testissä on 256 kyllä-ei -tyyppistä väittämää, joihin testattavan pitää vastata. Vastausten perusteella arvioidaan henkilön persoonallisuutta kuudella toista asteikolla. Toisin kuin 16PF, PRF ei ole empiirisesti ja faktorianalyttisesti laadittu persoonallisuustesti, vaan se perustuu teoreettiseen malliin persoonallisuuden rakenteesta.

11.2.5 Rorschach Inkblot Test (Ro)

Käyttötarkoitus: persoonallisuuden dynamiikan ja psykopatologian tutkiminen

Tyyppi: suoritus+observointi / tyypillisen suoritusten arviointi / subjektiivinen

Kesto: noin 60 minuuttia

Kohderyhmä: aikuiset, joskus myös lapset

Rakenne: 10 taulua (5 mustavalkoista, 3 värillistä ja 2 punamustaa), joissa on mustat tahrojen kuvia. Taulut näytetään yksitellen testattavalle, ja pyydetään häntä kertoamaan mitä hän näkee kuvassa. Vastausten ja testitulanteesta toimimisen perusteella tehdään päätelmiä henkilön persoonallisuuden rakenteesta ja dynamiikasta. HUOMAA : Ro:n tulkintaan on ainakin viisi erilaista menetelmää. Suomessa näistä käytetyimpiä ovat Klopfer-Davidsonin ja Exnerin, joista jälkimmäisempi lienee suositeltavampi, jos Ro:ta haluaa käyttää.

11.2.6 Sixteen Personality Factors Questionnaire (16PF)

Tyyppi: itsearviointi / tyypillisen suoritusten arviointi / objektiivinen

Kesto: noin 40 minuuttia

Käyttötarkoitus: persoonallisuuden rakenteen arviointi

Kohderyhmä: aikuiset

Rakenne: Testattava vastaa lomakkeella oleviin kyllä - ei - en osaa sanoa -tyyppisiin väittämiin. Faktorianalyysin avulla on selvitetty erilaisten vastausten liittymistä kuuteentoista persoonallisuuden peruspiirteeseen. Henkilön vastausten perusteella pyri-

tään arvioimaan näiden persoonallisuuden piirteiden voimakkuutta.

11.3 Muut testit

Tähän osastoon kuuluvat testit, jotka eivät ole varsinaisia älykkyys- tai persoonallisuustestejä. Kokoelma on siis hyvin sekalainen.

11.3.1 Bender Visual Motor Gestalt Test (Bender)

Käyttötarkoitus: visuaalisen ja motorisen suorituskyvyn arviointi

Tyyppi: tehtävä / maksimaalisen suorituksen arviointi / objektiivinen

Kesto: noin 15 minuuttia

Kohderyhmä: lapset

Rakenne: Testattavalle näytetään kortteja, joissa on erilaisia viivoista muodostettuja kuvioita. Testattavan tulee jäljentää kuviot, ja jäljennyksen laatua arvioimalla pyritään päättämään visuaalisen ja motorisen suorituskyvyn taso.

11.3.2 Ihmisiirros (Goode-noughin koe)

Käyttötarkoitus: kehitystason arviointi

Tyyppi: tehtävä / tyypillisen suorituksen arviointi / subjektiivinen

Kesto: noin 10 minuuttia

Kohderyhmä: lapset

Rakenne: Vapaamuotoinen, lasta pyydetään piirtämään paperille ihminen. Piirroksen detaljien rikkaudesta ja yleisestä tasosta tehdään päätelmiä lapsen kehitystasosta.

Tämäntyyppisistä piirrosteisteistä on kerätty paljon aineistoa, joten lapsen kehitystason karkea arviointi onnistuu hyvin. Testiä on joskus käytetty myös persoonallisuuden dynamiikan arviointiin, mutta tulokset eivät ole olleet kovinkaan yhtenäisiä.

11.3.3 NEPSY

Käyttötarkoitus: neurokognitiivisten toimintojen kehityksen ja toiminnan kartoitus

Tyyppi: tehtävä / maksimaalisen ja tyypillisen suorituksen arviointi / objektiivinen

Kesto: riippuu testattavasta

Kohderyhmä: lapset, nuoret

Rakenne: NEPSY koostuu 38 osatestistä, jotka jakautuvat seuraaviin alueisiin: 1) tarkkaavuus ja toiminnanohjaus 2) kielelliset toiminnot 3) sensomotoriset toiminnot 4) visuospatiaaliset toiminnot ja 5) muisti ja oppiminen. Osatesteistä voidaan valita vain osa riippuen siitä minkälaisen häiriön arvelaan olevan kyseessä. NEPSY on kotimainen testi.

11.3.4 Pirkanmaan perhepsykologinen testi (PPT)

Käyttötarkoitus: perheen toiminnan, rakenteen ja vuorovaikutuksen arvioiminen

Tyyppi: tehtävä / tyypillisen suorituksen arviointi / subjektiivinen ja objektiivinen

Kesto: vaihtelee huomattavasti

Kohderyhmä: perheet

Rakenne: PPT on monipuolinen testi - ja kyselyjärjestelmä, jolla pyritään selvittämään perheen rakennetta ja toimintaa sekä siihen liittyviä ongelmia. PPT on Suomessa laadittu testi.

11.3.5 Wartegg Zeichen Test (WZT)

Käyttötarkoitus: lähes mitä tahansa

Tyyppi: suoritus / tyypillisen suorituksen arviointi / subjektiivinen

Kesto: noin 30 minuuttia

Kohderyhmä: aikuiset, joskus myös lapset

Rakenne: Testilomake, jossa on kahdeksan ruutua joissa jokaisessa on jokin alkumerkki. Testattava piirtää jokaiseen ruutuun vapaavalintaisen kuvan. Piirrosten perusteella yritetään tehdä päätelmiä henkilön kyvyistä ja persoonallisuudesta.

Luku 12

Testien käyttäminen tutkimus- ja asiakastyössä

Testit ovat mittalaitteita, joita käytetään pääasiassa asiakastyössä ja tutkimuksessa. Niinpä testiteoreettisen tarkastelun lisäksi on hallittava myös monia testien käyttämiseen liittyviä taitoja. Seuraavassa esitellään lyhyesti testien käyttömahdollisuuksia ja käytön ongelmia tutkimuksessa ja asiakastyössä.

12.1 Testien käyttö tutkimuksessa

Psykologisia testejä voidaan käyttää hyväksi tutkimuksessa monella tavalla. Testiteorian kehittyminen on osaltaan mahdollistanut monipuolisempien psykologisten tutkimusten tekemisen. Testiteoriahan on alunperin syntynyt tarpeesta arvioida psykologisen mittaamisen tarkkuutta.

12.1.1 Testien tutkiminen

Psykologisia testejä tutkitaan hyvin paljon - sekä valmiita testejä että alustavia versioita. Esimerkiksi WAIS-R:n toimivuutta puoltavia tutkimuksia on runsaasti, samoin kuin Ro:n toimivuuden kyseenalaistavia. Testien tutkiminen onkin erittäin tärkeä osa psykologiaa, sillä testien laatijat eivät itse välttämättä kykene suhtautumaan omaan tuotokseensa täysin objektiivisesti. Testit ovat myös kauppatavaraa, joten tutkimuksen etiikka on joskus uhrattava markkinoinnin nimessä. Tiedeyhteisö onkin velvollinen tutkimaan käytössä olevien testien ominaisuuksia. Tämäkään ei ikävä kyllä aina riitä. Monia testejä käytetään, vaikka ne olisi objektiivisissa tutkimuksissa todettu vähintäänkin epäilyttäviksi.

12.1.2 Testit tutkimuksen mittalaitteina

Useissa tutkimuksissa ihmisten psyykkisiä ominaisuuksia on mitattava erilaisten testien tai niihin rinnastettavien mittalaitteiden avulla. Monille kliinisessä käytössäkin oleville testeille on sovellusmahdollisuuksia psykologisessa tutkimuksessa.

Kokeellisissa asetelmissa on usein tarpeellista muodostaa koe- ja kontrolliryhmät siten, että molempien ryhmien lähtötaso jonkin ominaisuuden suhteen on sama. Tällöin voi olla tarkoituksenmukaista testata molemmat ryhmät ennen interventiota ja suorittaa ryhmiin jakaminen testin tulosten perusteella. Jos lähtömittausta ei käytetä ryhmien muodostamiseen, sitä voidaan käyttää myös *kovariaattina* eli ominaisuutena, joka vaihtelee ihmisillä, mutta jota tutkimuksessa ei ole kontrolloitu. Kovarianssianalyysin avulla voidaan kontrolloida tällaisen muuttujan vaikutus mittauksiin. Testejä voidaan myös käyttää erilaisissa interventioissa muutoksen arvioimiseen. Esimerkiksi psykoterapian vaikutavuutta voidaan arvioida testaamalla henkilöiden depressiivisyys ennen ja jälkeen terapiaa. Tällöin on luonnollisesti käytettävä testejä, joilla voidaan helposti havaita muutokset mitattavassa ominaisuudessa.

Tutkimuskäytössä voidaan myös käyttää sellaisia testejä, joiden käyttö asiakastyössä ei ole mahdollista. Tällaisia ovat esimerkiksi alustavat testiversiot. On luonnollisesti toivottavaa, että tutkimuksessa käytettävät mittalaitteet olisivat mahdollisimman tarkkoja. Tutkimuskäytössä testeihin voidaan kuitenkin suhtautua *hieman* - mutta

vain hieman - vapaammin kuin asiakastyössä, koska tutkimuksissa testien perusteella ei tehdä mitään ihmisiä koskevia päätöksiä.

Monia tutkimuksia varten joudutaan myös laatimaan erilaisia testityyppisiä mittalaitteita. Näiden mittareiden psykometristen ominaisuuksien arvioiminen on luonnollisesti tärkeä osa koko tutkimusta. Tutkimusmittareiden kehittämistä voidaan jatkaa myös varsinaisen tutkimuksen päätyttyä. Monet kliinisessä käytössä olevat testit ovatkin alunperin syntyneet tutkimuksessa käytettäväksi mittareiksi, ja vasta myöhemmin testit on sovitettu asiakastyön tarpeisiin sopiviksi.

12.2 Testit asiakastyössä

Hyväkään psykometriikan tuntemus ei yksinään riitä testien käyttämiseen. Myös itse testien käyttämisen opetteleminen on osa testaamisen taitoa. Käyttäjän tulee tuntea sekä testin käyttämisen mekaniikka, että testin mahdollisuudet ja rajat. Koska testejä käytetään paljon sekä tutkimuksen että asiakastyön parissa, käsitellään vielä lopuksi psykologisten testien käyttämisen yleisiä periaatteita.

Kuten edellisten lukujen tarkastelussa on käynyt ilmi, testit eivät todellakaan ole mitään huipputarkkoja instrumentteja. Niinpä niiden antamiin tuloksiin onkin suhtauduttava varauksella. Testi itsessään ei ole kuin työväline muiden joukossa. Se ei korvaa muita ammattitaitoisen psykologin tiedonhankintamenetelmiä, vaan ennemminkin tukee niitä. Koska eri testit mittaavat eri asioita, yksi testi ei sovellu jokaiseen

tilanteeseen. Rutiininomainen samojen testien käyttäminen kaikissa tilanteissa ei ole alkuunkaan tarkoituksenmukaista. Testejä olisikin käytettävä omien tilanteesta muodostettujen hypoteesien hyväksymiseen tai hylkäämiseen. Jos epäilet haastattelun perusteella, että lapsella saattaa olla neurologisista syistä johtuvia oppimisvaikeuksia, et tietenkään tee hänen kanssaan esimerkiksi Ro:ta, vaan tarpeelliset osat NEPSY:stä. Kaikessa asiakastyössä testejä tulisikin käyttää haastattelun ja muiden informaationhankintamenetelmien tukena, ja johtopäätökset pitäisi tehdä muodostuneen kokonais kuvan perusteella.

12.2.1 Testaamisen ongelmia

Testeihin - niinkuin kaikkiin muihinkin psykologin työvälineisiin - on suhtauduttava asiaankuuluvalla varauksella. Jokaisella testillä on omat kannattajansa, ja he luonnollisesti kannustavat oman menetelmän käyttöön. Työpaikalle saapuva uusi psykologi on tietysti hankalassa tilanteessa. Vanhoilla työntekijöillä on jo valmiit totut työtavat, joita he koettavat siirtää uusille psykologeille. Tähän ei kuitenkaan kannata lähteä välittömästi mukaan, juuri valmis-tuneella psykologillahan pitäisi olla uusinta tietoa psykologian teoriasta ja sovelluksista.

Monet Suomessa käytetyistä testeistä ovat erittäin kyseenalaisia. Tässä yhteydessä ei oteta kantaa minkään erityisen testin käyttämiseen tai käyttämättä jättämiseen, mutta seuraavien yleisten ohjeiden avulla on helppo välttää käyttökeltottomat testit. Ennen kuin alkaa käyttää mitään testiä, on

syitä tarkistaa ainakin seuraavat asiat:

1. Kuka testin on tehnyt ja koska? Onko tämä testi käytössä muualla maailmassa (tai edes muualla Suomessa), vai onko se vain jokin pienen piirin harraste?

2. Onko testiin olemassa *testin tekijän* laatima käsikirja? Jos on, niin kuinka kattavan käsityksen sillä saa testin käyttämisestä, ominaisuuksista sekä testillä tehtävien päätösten laadusta?

3. Mitä tarkoitusta varten testi on alunperin tehty? Onko tämä testi alunperin suunniteltu sitä tarkoitusta varten, johon sitä aiotaan käyttää?

4. Jos testi on laadittu psykometristen periaatteiden mukaan, niin millainen on testin reliabiliteetti ja validiteetti? Miten niitä on tarkasteltu?

5. Jos testistä on normit, niin ovatko ne ajanmukaiset ja tarpeeksi kattavat?

6. Onko testistä asiallista, mielellään kansainvälistä tutkimustietoa?

Jos näihin kysymyksiin ei löydy tyydyttäviä vastauksia, voi testin melko huolelta unohtaa. Eri asia on tietenkin, jos testiä

ei käytetä ollenkaan päätöksentekoon vaan ennemminkin esimerkiksi kontaktin luomiseen asiakkaaseen. Tämä voi olla mahdollista, mutta se on kuitenkin eettisesti melko arveluttavaa. Miksi suotta käyttää testiä ja kerätä asiakkaasta sellaista tietoa, jota ei päätöksenteossaan tarvitse? Ennen kuin alkaa käyttää *mitään* testiä, on aina hyvä kysyä itseltään: mitä sellaista lisäinformaatiota tämä testi antaa, jolla on merkitystä tämän ongelman ratkaisemisen kannalta?

Testit eivät ole haastattelun korvikkeita, eikä ole mitenkään perusteltua väittää, että testaamalla saisi aina parempaa informaatiota kuin haastattelulla. Testauksen pitäisi aina olla ainoastaan osana muuta tiedonkeruuta, haastatteluita, observointeja ja muita tiedonhankintatapoja. Eri tutkimusmenetelmien käytön pitäisikin olla joustavaa ja vaihteista. Kun tutkimuksissa tulee ilmi joihtakin hypoteeseja, niin niiden tutkimiseen voidaan tarpeen vaatiessa valita sopiva menetelmä. Tämän tiedon perusteella hypoteesi voi taas muuttua, ja uutta hypoteesia voidaan kenties tutkia jonkin muun testin avulla. Tämä on paljon mielekkäämpi tapa toimia kuin se, että asiakkaalle teetetään kaikki 17 testiarkistosta löytyvää enemmän tai vähemmän soveliaista testiä, ja tehdään päätös tämän informaationsilpun perusteella. Pelkkä testaaminen - tai pelkkä haastatteleminen - ei välttämättä anna kovinkaan laajaa kuvaa ihmisestä. Jokaista tiedonhankintamenetelmää käytettäessä on aina pidettävä mielessä niiden rakenteelliset heikoudet ja vahvuudet: mitä kullakin menetelmällä saadaan selville, ja ennen kaikkea mitä sillä *ei* saada selville.

Lopuksi esitetään vielä kerran varoituk-

sena usein toistettu seikka: testeillä tehdään usein jotain päätöksiä ihmisestä. Niinpä psykologin tulee tarkkaan tuntea käyttämiensä menetelmien mahdollisuudet. Väärillä päätöksillä voi olla yksilön kannalta vakavia seurauksia. Psykologin on aina oltava valmis myös juridisesti puolustamaan käyttämiään menetelmiä. Voit arvioida, että menestys oikeussalissa ei ole kovinkaan kummoista, kun psykologi selittää valinneensa työnhakijat sellaisen testin perusteella, jonka ennustevaliditeettia ei ole tutkittu ollenkaan, ja jonka reliabiliteettia on mahdotonta tutkia. Asiakkaan oikeusturvan kannalta onkin siis välttämätöntä, että psykologi

1. tuntee käyttämänsä menetelmät rajoituksineen
2. osaa tulkita eri menetelmillä hankkimaansa tietoa asianmukaisesti
3. tietää milloin mikäkin menetelmä on käyttökelpoinen ja milloin käyttökelvoton
4. käyttää ainoastaan sellaisia menetelmiä, jotka hallitsee
5. kerää asiakkaastaan ainoastaan sellaista informaatiota, joka on merkityksellistä ongelman ratkaisemiseksi ja
6. osaa suhtautua kaikkeen hankkimaansa tietoon kriittisesti.

Liite A

Kirjallisuutta

Seuraavassa luettelossa on kirjoja, joita on käytetty lähdemateriaalina tämän monisteen kirjoittamisessa. Kaikki luettelon kirjat ovat suositeltavia oheislukemistoksi.

Aiken, L.R. (1985). *Psychological Testing and Assessment*. Boston: Allyn and Bacon.

Cattell, R.B. (1952). *Factor Analysis*. New York: Harper And Brothers.

Cronbach, L. (1970). *Essentials of Psychological Testing*. New York: Harper And Row.

Gulliksen, H. (1950). *Theory of Mental Tests*. New York: John Wiley And Sons.

Helenius, H. (1995). *Tilastolisten menetelmien perustiedot*. Salo: Statcon.

Holopainen, M. & Pulkkinen, P. (1994). *Tilastolliset menetelmät*. Porvoo: WSOY.

Konttinen, R. (1980). *Testiteoria: Johdatus kasvatus- ja käyttäytymistieteellisen mittauksen teoriaan*. Helsinki: Gaudeamus.

Magnusson, D. (1961). *Testteori*. Stockholm: Alqvist Och Wiksell Forlag AB.

Murphy, K.R. & Davidshofer, C.O. (1988). *Psychological Testing: Principles and Applications*. Englewood Cliffs: Prentice-Hall.

Nummenmaa, T. (1990). *Psykometristen ja tulkinnallisten testien käyttö tutkimuksessa*. Tampere: Tampereen yliopiston jäljennepalvelu.

Nummenmaa, T., Konttinen, R., Kuusinen, J. & Leskinen, E. (1997). *Tutkimusaineiston analyysi*. Porvoo: WSOY.

Nunnally, J. (1967). *Psychometric Theory*. New York: McGraw-Hill.

Vahervuo, T. (1956). *Psyko-metriikan metodeja 1*. Helsinki: WSOY.

Vahervuo, T. (1958). *Psyko-metriikan metodeja 2*. Helsinki: WSOY.

Liite B

Kaavoja

Keskiarvo

$$\bar{x} = \frac{\sum x}{n}$$

Populaatiokeskihajonta

$$\sigma = \sqrt{\sigma^2}$$

Populaatiovarianssi

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Otoskeskihajonta

$$s = \sqrt{s^2}$$

missä:

$$\begin{aligned}\sigma^2 &= \text{populaatiovarianssi} \\ (x_i - \bar{x}) &= \text{henkilön testipistemää-} \\ &\quad \text{rän ja testipistemäärien} \\ &\quad \text{keskiarvon erotus} \\ n &= \text{havaintojen lukumäärä}\end{aligned}$$

Z-pisteet

$$z = \left(\frac{x - \bar{x}}{s_x} \right)$$

missä

Otosvarianssi

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

$$\begin{aligned}z &= \text{z-pistemäärä} \\ x &= \text{testipistemäärä} \\ \bar{x} &= \text{testipistemäärien keskiarvo} \\ s_x &= \text{keskihajonta}\end{aligned}$$

Populaatiokovarianssi

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

missä

- s_{xy} = x : n ja y : n kovarianssi
 x_i = x – muuttujan arvo
 $\bar{x}$ = x – muuttujan keskiarvo
 y_i = y – muuttujan arvo
 $\bar{y}$ = y – muuttujan keskiarvo
 n = havaintoparien lukumäärä

Otoskovarianssi

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Populaatiokorrelaatio

$$r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y}$$

missä

- r = x :n ja y :n välinen korrelaatio-
kerroin
 s_{xy} = x :n ja y :n välinen
populaatiokovarianssi
 σ_x = x – muuttujan
populaatiokeskihajonta
 σ_y = y – muuttujan
populaatiokeskihajonta

Otoskorrelaatio

$$r = \frac{\sum (z_x \times z_y)}{n}$$

missä

- r = korrelaatiokerroin
 z_x = x -muuttujan z -pistemäärä
 z_y = y -muuttujan z -pistemäärä
 n = havaintojen lukumäärä

Regressiosuoran yhtälö

$$\hat{y} = \beta_0 + \beta_1 x$$

missä

- $\hat{y}$ = y :n ennustettu arvo
 β_0 = vakiotermi
 β_1 = regressiosuoran kulmakerroin
 x = x -muuttujan arvo

SEM

$$SEM = \sigma_x \sqrt{1 - r_{xx}}$$

missä:

- SEM = mittauksen keskivirhe
 σ_x = varianssi
 r_{xx} = reliabiliteettikerroin

Cronbachin α

$$\alpha = \frac{k(\bar{r}_{ij})}{1 + (k-1)\bar{r}_{ij}}$$

missä:

- α = reliabiliteetti(sisäinen konsistenssi)
 k = testi-itemien lukumäärä
 $\bar{r}_{ij}$ = testi-itemien interkorrelaatioiden keskiarvo

Spearman-Brownin kaava

$$\text{uusi } r_{xx} = \frac{n \times \text{vanha } r_{xx}}{1 + (n-1)\text{vanha } r_{xx}}$$

missä:

- n = uuden testin itemien lukumäärä
 vanha r_{xx} = alkuperäisen testin reliabiliteettikerroin
 uusi r_{xx} = pidennetyn testin reliabiliteettikerroin

Attenuaation korjauskaava

$$r'_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}}$$

missä

- r_{xy} = x:n ja y:n korrelaatio

- r'_{xy} = x:n ja y:n välinen korjattu korrelaatio
 r_{xx} = x:n reliabiliteetti
 r_{yy} = y:n reliabiliteetti

Attenuaation muutos

$$r'_{xy} = \frac{r_{xy} \sqrt{\text{uusi } r_{xx} \times \text{uusi } r_{yy}}}{\sqrt{\text{vanha } r_{xx} \times \text{vanha } r_{yy}}}$$

missä

- r'_{xy} = Uusi korrelaatio, kun reliabiliteetit muuttuvat
 r_{xy} = x:n ja y:n korrelaatio
 vanha r_{xx} = x:n vanha reliabiliteetti
 vanha r_{yy} = y:n vanha reliabiliteetti
 uusi r_{xx} = x:n uusi reliabiliteetti
 uusi r_{yy} = y:n uusi reliabiliteetti

Erotussuureen reliabiliteetti

$$r_{DD} = \frac{\frac{r_{xx} + r_{yy}}{2} - r_{xy}}{1 - r_{xy}}$$

missä

- r_{DD} = erotussuureen reliabiliteetti
 r_{xx} = x-muuttujan reliabiliteetti
 r_{yy} = y-muuttujan reliabiliteetti
 r_{xy} = x:n ja y:n välinen korrelaatio

Summamuuttujan reliabiliteetti

$$r_{ss} = \frac{k - k\bar{r}_{ii}}{1 - k + [(k^2 - k)\bar{r}_{ij}]}$$

missä

r_{ss} = summamuuttujan reliabiliteetti

k = yhdistettävien testien
lukumäärä

$\bar{r}_{ii}$ = testien keskimääräinen
reliabiliteetti

$\bar{r}_{ij}$ = testien välinen keskimääräinen
korrelaatio