RELIAABELIUS JA VALIDIUS KIELITAIDON TESTAAMISESSA

Reliaabelius

- luotettavuus, johdonmukaisuus
- ei oikeastaan voida arvioida yhden mittauksen perusteella (mutta yhden mittauksen mittarin johdonmukaisuus/sisäinen konsistenssi/homogeenisuus on kyllä järkevä likiarvo)
- laskennallinen kerroin mahdollinen
- vaihteluväli normaalisti 0.00 – 1.00
- kyseessä on ensisijaisesti tulosten luotettavuus (ei mittavälineen)
- on tilannekohtainen – ei yleispätevä, yleinen - ominaisuus; arvioitava joka tilanteessa ja aineistossa erikseen
- välttämätön mutta ei riittävä edellytys validiudelle

Validius

- pätevyys, "osuvuus"
- ei yleensä voi ilmaista laskennallisesti (tosin reliaabeliuden neliötä voidaan pitää validiuden maksimaalisena arvona), vaan edellyttää validiusväitettä tukevien perusteiden pohdiskelua; validiusdiskurssi on parhaimmillaan uskottavaa monipuolista argumentointia testauksen osuvuudesta/kyvystä testata mitä halutaan
- kyseessä on tuloksien pohjalla tehtävien johtopäätösten, ratkaisujen ja päätösten pätevyys; Arviointi on validia, jos sen perusteella voidaan tehdä oikeaan osuvia tulkintoja, päätöksiä tai ratkaisuja. Validiuspäätelmä edellyttää koko arviointijärjestelmän laadun pohtimista.
- validius on reliaabeliuden tapaan kontekstisidonnainen, ei yleispätevä/yleinen ominaisuus; se riippuu arviointitarkoituksesta, kielitaitokäsityksestä, kohderyhmästä ja mittaustavasta

Erottelukyky

- jos kaikkien tulisi osata jokin asia ja näin voidaan kohtuudella odottaakin, erottelukyky ei ole keskeinen vaatimus; ihannetilanne olisi että kaikki osaisivat kaiken; koska tämä on harvoin – tuskin koskaan – realistinen odotus, erottelukyky on käytännössä – ja ylioppilaskokeessa – hyvin tärkeä vaatimus
- erottelukyky tarkoittaa osion/tehtävän kykyä erotella kokeensuorittajia eri suoritustasoilla
- tehtävien kyky erotella kokeen suorittajia (erottelukyky) on PAIKALLINEN ominaisuus: helppo mutta hyvin tehty tehtävä pystyy erottelemaan heikoimmat muista ja hyvin tehty vaikea tehtävä puolestaan erottelee parhaat muista: keskivaikea osio (ratkaisuprosentti 50) erottelee tehokkaimmin KOKO taitotasolla.
- Siksi yleensä on järkevää sijoittaa kokeeseen eniten keskivaikeita tehtäviä ja kohtalainen määrä helppoja ja vaikeita tehtäviä.
- erotteluindeksi voi vaihdella teoriassa –1/+1; se perustuu tavallisesti osion ja kokeen summapistemäärän väliseen korrelaatioon, ts. siihen missä määrin osio erottelee samansuuntaisesti kuin kaikki muut tehtävät yhdessä
- erotteluindeksin tulee olla positiivinen ja mielellään ainakin .20 mutta tyytyväisiä voidaan olla vain osioihin, joiden erotteluindeksi on vähintään .30

Kokeet ja arviointi ovat siten aina enemmän tai vähemmän reliaabeleja ja valideja, ja niihin sisältyy väistämättä tiettyä virhettä. Arvioinnin seuraamusten merkittävyyden perusteella voidaan ratkaista, kuinka tarkkaan arviointiin olisi pyrittävä ja toisaalta kuinka suuri virhemarginaali voidaan hyväksyä.

Reliaabeliuteen vaikuttavia tekijöitä

- tehtävien/osioiden määrä
  ⇒ yleensä suurempi tehtävämäärä on eduksi reliaabeliudelle, koska saadaan parempi näyte mitattavasta taidosta; satunnaisvaikutus vähenee. Tähän liittyy kysymys monivalintatehtävien vaihtoehtojen määrästä. Jos "päätön" arvaus on yleistä, vaihtoehtojen määrä on eduksi, koska se merkitsee, että päättömällä arvaamisella ei saada kovin korkeaa pistemäärällä. Koska päätön arvaaminen on kuitenkin aika harvinaista, vaihtoehtojen määrä ei ole yhtä tärkeää, kuin niiden toimivuus, oikeansuuntainen erottelevuus. Vaihtoehtojen laatu on tärkeämpää kuin niiden määrä. Huonosti toimivat vaihtoehdot itse asiassa alentavat osion laatua ja siten kokeen reliaabeliutta.
- osioiden vaikeustaso
  ⇒ koko taitotasoaluetta ajatellen paras tulos saavutetaan, kun noin puolet osaa ratkaista sen
- pistemäärien hajonta
  ⇒ yleensä - mitä suurempi hajonta, sitä korkeampi reliaabelius. Jos testi järjestettäisiin uudelleen, vähäisiä eroja sisältävässä testissä yksilöiden järjestys vaihtelee enemmän suurihajontaisessa kokeessa, jossa yksilöiden järjestys säilyy paremmin samana, ja tämä näkyy edellisessä tapauksessa alhaisempana reliaabeliutena
- objektiivisuus, eli arvioinnin yksimielisyys
  ⇒ avoimissa tehtävissä voi objektiivisuutta lisätä tehtävän huolellisella ja selkeällä muotoilulla ja selkeillä arviointiohjeilla; testausolosuhteiden vakioiminen

Mitä johtopäätöksiä edellä esitetystä voidaan tehdä kielikokeen laadinnalle?

1) Ratkaisevan tärkeää on yksittäisten osioiden laatu: niiden erottelukyvyn tulisi olla mahdollisimman korkea. Jokaisen osion laadintaan, arviointiin ja muokkaamiseen (ja jos suinkin mahdollista – esitestaamiseen) kannattaa kiinnittää suurta huomiota. Hyvä esityö kannattaa aina.
2) Osioita tulisi olla mahdollisimman paljon, koska pidempi koe on lyhyttä luotettavampi ja virheen osuus pienenee yleensä pituuden kasvaessa ja näytön kasvaessa. Kannattaa miettiä keinoja, jolla saadaan mahdollisimman paljon näytteitä osaamisesta. Sellainen arviointi ei ole järkevää, jossa joudutaan käyttämään paljon aikaa ja kuitenkin saadaan vähän näytettä osaamisesta. Kannattaa harkita (mahdollisimman) lyhyitä ja monia tehtäviä laajojen ja harvojen tehtävien sijasta.

Testi on, laajasti ymmärrettynä:

- joukko tehtäviä, jotka "elisitoivat" testattavien käyttäytymistä tietyllä sisältöalueella, tai
- skaala (esim. asenneväittämälista), joka kuvaa käyttäytymistä tietyllä sisältöalueella, tai
- järjestelmä, jolla kootaan näytteitä yksilön työskentelystä tietyllä sisältöalueella.

Pisteitys/arviointimenetelmä: voidaan kvantifioida (antaa numeeriset arvot), arvioida ja tulkita osaamisnäytteet.

Reliaabelius viittaa tällaisten menetelmien johdonmukaisuuteen, kun testaus toistetaan yksilöillä tai ryhmillä.

- yksilöiden ja ryhmien käyttäytymisen oltava tietyssä määrin pysyvää, vaikka tuloksissa esiintyykin luonnollisesti jossakin määrin vaihtelua
- vaihtelu johtuu yksilöistä itsestään, testeistä ja testausolosuhteista.

Perinteellisesti on käytetty 3 reliaabeliuskerrointa:

1. rinnakkaistesteihin perustuva reliaabelius
2. uudelleentestaukseen perustuva reliaabelius (stabilisuus)
3. testin osioiden sisäinen johdonmukaisuus

Usein sisäinen konsistenssi (alfa) antaa varsin samanlaisen tuloksen kuin rinnakkaistestaus ja uudelleentestaus, ja on siksi varsin käyttökelpoinen tapa arvioida reliaabeliutta yhdellä mittauksella.

Arviointia sisältävässä mittauksessa käytetään tavallisesti 4) arvostelijoiden johdonmukaisuutta kuvaavia indikaattoreita.

Mittaukseen sisältyy aina virhettä. Mittausvirhe voi olla:

- satunnaista ja ennustamatonta, mikä on tavallinen tapa hahmottaa mittausvirhettä, ja se alentaa reliaabeliutta

Koska satunnaisvirheet ovat epäjohdonmukaisia ja ennustamattomia, niitä ei saada erotettua havaituista arvoista, mutta niiden suuruutta voidaan kuvata eri tavoilla.

- systemaattista: systemaattisen virheen katsotaan yleisesti merkitsevän konstruktin kannalta irrelevanttia tulosten vaihtelua ja siten heikentävän validiutta (opetus, oppiminen, kypsyminen)

Mittausvirhe:

- heikentää mittalukujen hyödyllisyyttä
- rajoittaa tulosten yleistettävyyttä
- heikentää luottamusta yksittäiseen mittaukseen
- voi vaihdella asteikon eri kohdissa
- tieto siitä on oleellisen tärkeätä mittavälineen arvioinnille

Mittausvirhe on usein mielekkäämpi kuin reliaabeliuskerroin, kun tuloksia tulkitaan. Siksi reliaabeliuskerrointa olisi hyvä - ainakin tutkintojärjestelmien puitteissa - täydentää ilmoittamalla kuinka tarkasti mittaus tapahtuu eri taitotasoilla.

Pistemäärien tulkinta voi olla:

- suhteellista
- absoluuttista

Mittauksen tarkkuus ja johdonmukaisuus ovat aina toivottavia:

- Tarkkuuden vaatimus lisääntyy, kun mittauksen seuraukset kasvavat. Jos aiheettomia seurauksia ei voida helposti korjata, mittaustarkkuudelle asettavat vaatimukset kasvavat.
- Silloin kun mittaaminen merkitsee luokittelua, mittausvirheen tärkeys vaihtelee. Taitotasorajojen lähivyöhykkeessä mittaustarkkuus on paljon merkittävämpää kuin etäällä niistä. Tästä syystä olisi hyvä olla tietoinen mittausvirheestä juuri rajojen kohdalla.

Yleiset tutkinnoille asetettavat vaatimukset:

- Reliaabelius- ja mittauksen keskivirhe tulee aina ilmoittaa.
- Testin suorittajien tulisi tietää, missä määrin suoritusnopeus vaikuttaa arvioon.
- Kun käytetään subjektiivista arviointia, tulee raportoida arvioitsijoiden välinen johdonmukaisuus.
- Jos mittaustarkkuuden ei voida olettaa olevan vakio, tulee raportoida mittausvirhe useilla suoritustasoilla. Erityisesti tulee kiinnittää huomiota katkaisukohtien mittaustarkkuuteen.

RELIAABELIUS JA VALIDIUS KIELITAIDON TESTAAMISESSA

Reliaabelius

- luotettavuus, johdonmukaisuus
- ei oikeastaan voida arvioida yhden mittauksen perusteella (mutta yhden mittauksen mittarin johdonmukaisuus/sisäinen konsistenssi/homogeenisuus on kyllä järkevä likiarvo)
- laskennallinen kerroin mahdollinen
- vaihteluväli normaalisti 0.00 – 1.00
- kyseessä on ensisijaisesti tulosten luotettavuus (ei mittavälineen)
- on tilannekohtainen – ei yleispätevä, yleinen - ominaisuus; arvioitava joka tilanteessa ja aineistossa erikseen
- välttämätön mutta ei riittävä edellytys validiudelle

Validius

- pätevyys, "osuvuus"
- ei yleensä voi ilmaista laskennallisesti (tosin reliaabeliuden neliötä voidaan pitää validiuden maksimaalisena arvona), vaan edellyttää validiusväitettä tukevien perusteiden pohdiskelua; validiusdiskurssi on parhaimmillaan uskottavaa monipuolista argumentointia testauksen osuvuudesta/kyvystä testata mitä halutaan
- kyseessä on tuloksien pohjalla tehtävien johtopäätösten, ratkaisujen ja päätösten pätevyys; Arviointi on validia, jos sen perusteella voidaan tehdä oikeaan osuvia tulkintoja, päätöksiä tai ratkaisuja. Validiuspäätelmä edellyttää koko arviointijärjestelmän laadun pohtimista.
- validius on reliaabeliuden tapaan kontekstisidonnainen, ei yleispätevä/yleinen ominaisuus; se riippuu arviointitarkoituksesta, kielitaitokäsityksestä, kohderyhmästä ja mittaustavasta

Erottelukyky

- jos kaikkien tulisi osata jokin asia ja näin voidaan kohtuudella odottaakin, erottelukyky ei ole keskeinen vaatimus; ihannetilanne olisi että kaikki osaisivat kaiken; koska tämä on harvoin – tuskin koskaan – realistinen odotus, erottelukyky on käytännössä – ja ylioppilaskokeessa – hyvin tärkeä vaatimus
- erottelukyky tarkoittaa osion/tehtävän kykyä erotella kokeensuorittajia eri suoritustasoilla
- tehtävien kyky erotella kokeen suorittajia (erottelukyky) on PAIKALLINEN ominaisuus: helppo mutta hyvin tehty tehtävä pystyy erottelemaan heikoimmat muista ja hyvin tehty vaikea tehtävä puolestaan erottelee parhaat muista: keskivaikea osio (ratkaisuprosentti 50) erottelee tehokkaimmin KOKO taitotasolla.
- Siksi yleensä on järkevää sijoittaa kokeeseen eniten keskivaikeita tehtäviä ja kohtalainen määrä helppoja ja vaikeita tehtäviä.
- erotteluindeksi voi vaihdella teoriassa –1/+1; se perustuu tavallisesti osion ja kokeen summapistemäärän väliseen korrelaatioon, ts. siihen missä määrin osio erottelee samansuuntaisesti kuin kaikki muut tehtävät yhdessä
- erotteluindeksin tulee olla positiivinen ja mielellään ainakin .20 mutta tyytyväisiä voidaan olla vain osioihin, joiden erotteluindeksi on vähintään .30

Kokeet ja arviointi ovat siten aina enemmän tai vähemmän reliaabeleja ja valideja, ja niihin sisältyy väistämättä tiettyä virhettä. Arvioinnin seuraamusten merkittävyyden perusteella voidaan ratkaista, kuinka tarkkaan arviointiin olisi pyrittävä ja toisaalta kuinka suuri virhemarginaali voidaan hyväksyä.

TESTIEN LAADINTA JA MUOKKAAMINEN

Testin laadinta on prosessi, jonka kuluessa tuotetaan mittaluku henkilön jostakin tiedon, taidon, asenteiden, persoonallisuuden jne. jostakin piirteestä kehittämällä osioita ja yhdistelemällä niitä testiksi tietyn suunnitelman mukaisesti.

Testien laadinnassa on tavallisesti neljä vaihetta:

1) testauksen tarkoituksen ja testattavan asian ("konstruktin") rajaaminen
2) testispesifikaatioiden laadinta ja arviointi
3) osioiden/tehtävien laadinta, esitestaaminen, arviointi ja valinta; arviointiohjeiden ja –menetelmien laadinta
4) käyttöön tulevan testin koostaminen ja arviointi

Testin viitekehyksen laadinnassa käytetään hyväksi teoreettista tietoa sekä sisältöalueen tai työtehtävien analysointia. Viitekehys on kaiken muun toiminnan lähtökohta ja ohjenuora.

Testityyppien valinnassa on kolme päätyyppiä:

- vastauksen valinta (selected response)
- lyhyet vastaukset (tuotokset)
- pitkät vastaukset (tuotokset)

Kaikkiin on kehitettävä arviointiohjeet (pisteitysohjeet). Arviointi voi olla holistista tai analyyttistä.

Suoritusarvioinnissa (performance assessment) tehtävät muistuttavat läheisesti todellisen elämän tilanteita. Sen koespesifikaatioissa on kiinnitettävä paljon huomiota tehtävien analysointiin. Tässä käytetään sekä loogista että empiiristä analyysia.

Portfoliot ovat erityinen suoritusarvioinnin muoto. Portfoliot ovat systemaattinen kokoelma: työnäytteitä, "opinnäytteitä" pitkähkölta tai pitkältä ajanjaksolta. Portfolioiden tarkoituksena voi olla:

- dokumentoida työssä tai opinnoissa tapahtunut kehitys/edistyminen
- auttaa valinta- ja ylennystilanteessa
- dokumentoida tutkinnon valmistumista/pätevyyden saavuttamista

Porftfolio voi sisältää:

- edustavan näytteen suorituksia
- parhaat näytteet
- edistyksen näytön
- sisällön valintojen perusteleminen

Portfolion koostumuksena voi olla: kirjallista, kuvallista, äänitettyä, videoitua aineistoa, demonstraatioita, simulaatioita jne.

Portfoliolle asetettavia vaatimuksia:

- niiden pohjana tulee olla selkeät spesifikaatiot.
- niiden muotoutumiseen vaikuttavien osapuolten osuudet ja vastuut tulee määritellä spesifikaatioissa. Niille asetetaan samat testaustekniset laatuvaatimukset kuin muille testeille.
- koska portfoliot tavallisesti perustuvat pitkille vastauksille/tuotoksille, niiden arviointikriteereiden laadintaan tulee kiinnittää paljon huomiota
- portfolioiden laatijat tarvitsevat riittävästi tietoa ennen työhönsä ryhtymistä.

TASAPUOLISUUS

Tasapuolisuus on tärkeää kaikissa testausvaiheissa, mutta on mahdotonta saavuttaa täydellistä tasapuolisuutta jo pelkästään siitä syystä, että mittaamisen reliaabelius ja validius ei koskaan ole täydellistä missään tilanteessa.

Testaustilanteet ovat vuorovaikutustilanteita. *Testaajan vuorovaikutuksen testattavien kanssa tulee olla:*

- ammatillista
- kohteliaan asiallista
- testattavista välittävää
- testattavia kunnioittavaa

Tämä on erityisen tärkeätä, koska testaajan ja testattavien statukset eivät ole samantasoisia.

Tasapuolisuutta on hahmotettu usein neljällä eri tavalla:

1) Harhaisuuden (bias) – ei ole "puolueellisuutta"
2) Testattavien tasapuolinen kohtelu testaustilanteessa
3) Testitulosten yhtäläisyys eri osaryhmissä
4) Saavutustestauksessa kaikilla on ollut tasapuolinen tilaisuus oppia testattavat asiat.

Harhaisuus (bias) on tekninen termi: testituloksilla on erilainen merkitys eri osaryhmissä.
Osioharhaisuus (DIF) : esimerkiksi jotkut tehtävät suosivat erityisesti naisia tai miehiä.

Kohtelu on tasapuolista, testaustarkoituksesta riippumatta, kun kaikilla testattavilla on tasapuoliset mahdollisuudet osoittaa miten he hallitsevat testattavat asiat. Tämä merkitsee, että:

- testausolosuhteet ovat asianmukaiset
- on ollut samanlaiset mahdollisuudet tutustua testaustapaan tai muuhun materiaaliin (esim. koemallit)
- ideaalisessa tilanteessa testattavilla on myös ollut tasapuoliset mahdollisuudet valmistautua testiin.

Tasapuolisuus koskee myös tulosten mahdollista julkisuutta.

Testauskirjallisuudessa EI yleensä katsota, että tasapuolisuus edellyttää samankaltaisia tuloksia eri ryhmissä. Jos on valintatilanne, tasapuolisuus edellyttää, että samanlaisen tuloksen saaneilla on kaikilla yhtäläiset mahdollisuudet tulla valituksi riippumatta siitä, minkä ryhmän jäseniä he ovat.

Useimmat testausasiantuntijat ovat sitä mieltä, että jos harhaisuutta ei esiinny ja kohtelu testaustilanteessa on ollut tasapuolista, tasapuolisuuden vaatimukset on täytetty.

Tasapuolinen mahdollisuus oppia testiin sisältyvät asiat koskee saavutustestausta. Testaustulos saattaa heijastaa tarkasti osaamistasoa, mutta tilanne on erilainen, jos ei ole ollut tilaisuutta oppia ja jos on ollut tilaisuus oppia mutta ei ole oppinut. Sertifioinnissa ja työhön valinnassa puolestaan katsotaan tasapuolisuuden tulleen täytettyä, kun testi kattaa hyvin tarvittavat tiedot ja taidot.

Testin harhaisuus voi johtua *sisällöstä* ja *vastausprosessista* (konstruktin kannalta irrelevantit seikat). Testin sisällön asianmukaisuutta voidaan varmistaa asiantuntijapaneelin avulla, tutkimalla saatuja tuloksia, opetussuunnitelmaa, testiohjeiden selkeyttä. Jotkut osiot voivat saada aikaan odottamattomia vastauksia. Testitulos voi esimerkiksi johtua osittain kielellisestä kyvykkyydestä,

hienomotoriikasta jne. Näitä mahdollisuuksia voidaan tutkia asiantuntijoiden arvioilla, DIF:illä, korrelaatiota tarkastelemalla jne.

Testiin osallistuvien tasapuolinen kohtelu ei merkitse vain tasapuolisuuden edistämistä vaan edistää myös testisuorituksista tehtävien johtopäätösten *reliaabeliutta ja validiutta.*

Testiin osallistujat voivat oikeutetusti edellyttää monia asioita:

- heillä on oikeus tulla arvioiduksi testeillä, jotka täyttävät nykyaikaiset ammatilliset vaatimustasot: reliaabelius, validius, tasapuolisuus, testijärjestelyt ja tulosten raportointi.
- heillä on oikeus etukäteen tietoonsa muutamia tärkeitä asioita: testin luonne ja sisältö, testin käyttötarkoitus, tulosten julkisuus/luottamuksellisuus
- heillä tulee olla mahdollisuus kysyä ja ilmaista huolenaiheensa ja saada kohtuullisessa ajassa vastaukset kysymyksiinsä.
- heillä voi olla mahdollisesti myös oikeus saada tietoa ja materiaalia, joka auttaa testin suorittamisessa: opinto-opas, testausvaihtoehdot, mallikysymyksiä, mallikoe, ohjeita testin suorittamiseksi (esim. ajankäyttö, arvaaminen, vastaamatta jättäminen, uusintamahdollisuus.