# Validiusargumentin rakentuminen suullisen kielitaidon arviointiprojektissa

Raili Hildén
raili.hilden@helsinki.fi
Helsingin yliopisto
Soveltavan kasvatustieteen laitos

Validiusargumentin rakentuminen suullisen kielitaidon arviointiprojektissa

Artikkelissa kuvataan aluksi kielitaidon arvioinnin validiustutkimuksen keskeisiä suuntauksia, joilla kehystetään käynnissä olevan suullisen kielitaidon arviointihankkeen tutkimustehtävä. Tutkimustehtävä kohdentuu kansallisiin opetussuunnitelmaperusteisiin sisältyvän taitotasoasteikon validiuteen, jota tarkastellaan validiusargumentin rakentumisen näkökulmasta. Väitettä asteikon ja sen pohjalta laadittujen puhetehtävien pätevyydestä suullisen taidon mittareina koetellaan Toulminin mallin avulla asettamalla tätä johtopäätöstä tukevat empiiriset lähtötiedot ja niitä tukevat perusteet vastakkain väitettä horjuttavan näytön kanssa. Perusteet ja varaukset koskevat väitetyn johtopäätöksen relevanssia, hyödyllisyyttä, tarkoitettuja seurauksia ja riittävyyttä. Hankkeen tutkijoiden spesifit tutkimusongelmat kohdentuvat johonkin mainituista piirteistä.

Validiusargumenttia sovelletaan suullisen kielitaidon arviointihankkeen, Hy-Talkin, kontekstissa. Ongelmanasettelussa ja menetelmävalinnoissa hyödynnetään validiustutkimuksen perinteisempiä sisältö- ja kriteeriperustaisia lähestymistapoja, mutta tehdään myös uusia avauksia tiedon syventämiseksi niistä tulkinnoista ja näkemyksistä, joita suullisten tehtävien suorittajilla ja arvioijilla ilmenee.

## 1. Validiuden vaiheita: Tieteellisestä tiedonhankinnasta hyödylliseen tulkintaan

### 1.1 Ongelmattomuudesta systematiikkaan: kriteeri-, sisältö- ja konstruktiperustainen tarkastelu

Varhaisimpien määritelmien mukaan testin validiudella tarkoitettiin testin kykyä mitata sitä, mitä se oli tarkoitettu mittaamaan (Kelley, 1927, s. 14). Perinteinen testaus ei ollut teoriaperustaista, vaan piti sekä reliaabeliutta että validiutta itsestään selvinä (Davies, 2003, s. 356). Arviointikäytännöt noudattivat opetuskäytäntöjä ja esitieteellinen kielitaitotestauskin keskittyi mittaamaan kielitietoa käyttötaidon sijasta. (Spolsky, 1995.)

1950-luvulta lähtien testauksessa alettiin pyrkiä selvemmin tieteellisen tiedonhankinnan ihanteeseen luonnontieteellisten esikuvien tapaan (McNamara & Roever, 2006, s. 10). Käyttäytymistieteissä validius koski tällöin jonkin latentin ja melko staattisen yksilömuuttujan (persoonallisuuspiirteen, ominaisuuden, taidon) mittaamista (Kane, 2002, s. 320). Tilastoanalyysien ja niiden suorittamiseen tarvittavan välineistön kehittyminen vauhdittivat erityisesti reliaabeliutta kohtaan tunnettua mielenkiintoa.

Ensimmäiset yritykset varmistaa validiutta kanavoituivat reliaabeliuden kautta, koska mittauksen virheettömyyttä pidettiin mittarin validiuden välttämättömänä ehtona. Näkemys, että riittävä ehto se ei ollut, sai enemmän huomiota vasta myöhemmin. Amerikkalaisten Cronbachin ja Meehlin (1955) kriteeriperustainen validiustarkastelu pohjautui testipistemäärien rinnastukseen johonkin toiseen, samaa piirrettä edustavaan muuttujaan eli kriteeriin. **Kriteeriviitteinen** validiustarkastelu voi olla **ennusteista**, jolloin testipistemäärien avulla pyritään ennustamaan jotain muuttujaa tulevaisuudessa

(testisuoritusta, tosielämän tehtävää tms.) tai **samanaikaista**, jolloin testipistemääriä rinnastetaan testin suoritusajankohtana hankittaviin kriteerimuuttujiin (testin osiin tai toisiin testeihin). Näitä harkinnanvaraisesti valittuja kriteerimuuttujia pidettiin lähtökohtaisesti todempina kuin itse testitulosta eikä niiden laatua useinkaan kyseenalaistettu (Kane, 2002, s. 320).

Sittemmin kriteeriperustaisen tarkastelun rajallisuus validiuden yksinomaisena takeena huomattiin, mutta edelleen tämä usein melko perustasoisiin tilastollisiin operaatioihin ja päättelyyn nojautuva lähestymistapa on validiustarkastelussa keskeinen. Laskennalliset välineet ovat kehittyneet, mutta perusidea elää esimerkiksi asetelmissa, joissa testissä menestymistä verrataan tosielämässä suoriutumiseen samassa tehtävässä (Cronbach 1971) tai asiantuntija-arvioihin perustuvissa mallinnuksissa (Angoff 1988).

Toinen varhaisimmista validiuslajeista, joka kehitettiin kriteeriviitteisen tarkastelun vaihtoehdoksi ja sitä täydentämään liittyy testin **sisältöön** (content-based validity) ja yritykseen hankkia testillä mahdollisimman edustava näyte niistä piirteistä tai suorituksista, joita testillä on tarkoitus mitata (Fulcher & Davidson, 2007, s. 4). Carrollin (1980, s. 67) mukaan kielitaitotestauksen sisältövalidius määriteltiin analysoimalla ensin testattavien viestintätarpeet ja spesifioimalla testin sisältö niiden mukaiseksi. Testitulos tulkitaan testisisältöjen valossa ja riittävän samankaltaisia testejä voitiin käyttää toistensa kriteereinä (Ebel 1961).

1950-luvulta on peräisin myös **konstrukti- eli käsitevalidius**, jonka kehitys meidän aikaamme saakka on osa arviointiteorian keskeisintä sovellushistoriaa. Sillä ei ollut tarkoitus korvata aiempia validiustarkasteluja, vaan ainoastaan täydentää niitä teoreettisemmin selvittämällä havaintojen ja teoreettisen mallin yhteensopivuutta. Mikäli havainnot sopivat malliin, sekä malli (konstruktit ja niiden väliset suhteet) että tiedonhankintapa validoituvat. (Cronbachin & Meehl, 1955.) Luonnontieteellisiä esikuvia mukaileva tukeutuminen vankkoihin teoriamalleihin ei kuitenkaan toiminut optimaalisesti ihmistieteissä, joissa tällaisia aksiomeja on kovin vähän (Kane, 2001, ss. 325 – 327). Konstruktilähtöisestä validiusajattelusta jäi kuitenkin voimaan se seikka, että mitattavan käsitteen määrittelyä pidetään edelleen laatunormit täyttävän testinlaadinnan lähtökohtana. Edelleen pidetään tärkeänä myös erilaisen näytön eli evidenssin hankkimista testituloksesta tehtävien tulkintojen tueksi (Fulcher & Davidson, 2007, s. 10). Aika on sen sijaan kumonnut monet tuolloiset oletukset havaintojen ja teoreettisten uskomusten keskinäisestä riippumattomuudesta.

1950-luvulta aina 1980-luvun lopulle ulottuvana ajanjaksona validiustutkimus myötäili siis tieteellisen tiedonhankinnan ja teorianmuodostuksen ihanteita. Tuolloin vakiintui kolme yhä ajankohtaista periaatetta. Ensinnäkin validiuden todentaminen nähdään monivaiheisena toimintana (validointina) (ks. myös Weir, 2005), jonka lähtökohtana on teoria, josta valitaan tietyt ulottuvuudet ja suunnitellaan niihin sopivat mittausmenetelmät. Mittausta ohjaamaan asetetaan hypoteesit, joita testataan suhteessa havaintoihin. Toiseksi testituloksesta tehtävä tulkinta tulee spesifioida ja kirjoittaa selkeään muotoon hypoteesiksi, ennen kuin sen validiutta voidaan arvioida. Tämä on oleellinen laajentuma käsitykseen, että validoinnin kohteena olisi itse testi tai testitulos

sellaisenaan. Kohteita ovat sen sijaan testituloksesta tehtävät johtopäätökset. Kolmas tekijä, joka sittemmin on noussut yhä tärkeämmäksi, on vaihtoehtoisten tulkintojen ja ristiriitaisen näytön vakava huomioon ottaminen validiustarkastelussa. (Kane 2001, ss. 323 – 324.)

## 1.2 Validius kokonaisvaltaisessa tarkastelussa

Edellä kuvatut muutokset validiuskäsityksessä ennakoivat perusteellista siirtymää komponentiaalisesta rakenteesta yhtenäisyyteen. Messickin (1989) teoreettisen mallinnuksen myötä käsitevalidius vakiintui kattamaan kaiken validiusevidenssin lähtien kriteeri- ja sisältöviitteisestä näytöstä ja reliaabeliudesta ulottuen aina testituloksesta tehtävien johtopäätösten arvoperustaan ja testin toimeenpanon ja testitulosten käytön sosiaalisiin seurauksiin (Messick, 1989, s. 20).

Messickin ajattelun mukaisen validiusnäkemyksen omaksuivat mm. American Educational Research Association ja American Psychological Association ohjeistoissaan, mikä merkitsi sille läpimurtoa arviointitutkijayhteisössä. Niissä esitettiin kaksi psykologisen, kasvatustieteellisen ja kielitestauksen pitkään luovuttamatonta periaatetta. Näistä ensimmäinen oli validiuden kehitys komponentaalisesta rakenteesta kohti yhtenäisyyttä ja toinen näkemys siitä, että validius ei ole itse testin eikä testituloksen vaan testiin perustuvien johtopäätösten ja niiden seurausten ominaisuus.

Kielitaitotestauksessa ja arvioinnissa yleisemminkin Messickin perintö on ollut ja on edelleen hyvin elinvoimainen ajoittaisesta kritiikistä huolimatta (Boorsbom ym., 2004). Tässä yhteydessä mainittakoon pari päälinjaa. Ensimmäinen liittyy yhtenäisen validiusmallin käytännölliseen sovellukseen testauksessa, ja sen yleisin sovellus lienee Bachmanin (1990) ja Bachman & Palmerin (1996) operationaalinen validiusnäkemys. Siinä arvioinnin kysymykset kiteytyvät testin *hyödyllisyyden* (usefulness) käsitteeseen, joka kattaa seuraavat osa-alueet: *reliaabelius, käsitevalidius* (construct validity), *vuorovaikutteisuus* (interactiveness), *vaikuttavuus* (impact) ja *käytännöllisyys* (practicality).

## 1.3 Validius tulkinnallisena toimintana

### 1.3.1 Validiuden etiikka

Toinen Messickin ajattelusta kumpuava juonne on vähemmän käytännöllinen ja vasta etsimässä muotojaan, mutta josta näyttäisi olevan nousemassa arviointitutkimuksen uusi makroparadigma, joka näkee arvioinnin sosiaalisena, kulttuurisena ja poliittisena toimintana.

Perinteinen validiuskäsitys liittyi kognitiiviseen psykologiaan, psykometriikkaan ja automaattisen tietojenkäsittelyn malleihin. Validiuden laajentuminen taas liittyy sosiokonstruktivismiin ja matemaattisten mallien ja tiedonkäsittelyn sofistikoitumiseen. Testisuorituksen rakentuminen sosiaalisena toimintana ilmeni etenkin suullisen kielitaidon testauksen yhteydessä. Osallistujien kokonaisvaltainen panos testattavaan

konstruktiin on tiedostettu jo pitkään ja yksi osuvimmista kysymyksistä olikin McNamaran (1997) esittämä ”Whose performance” koskien suullisen kokeen arviointikohdetta.

Sittemmin arvioinnin eettisyyttä on punnittu laajemmassa katsannossa yhteiskunnan ja vallankäytön näkökulmasta. Taustalla on usein postmoderni kriittinen teoria, jota ovat kielitaidon arviointiin soveltaneet ainakin Shohamy (2001), Hamp-Lions (1997), Lynch (2001), McNamara ja Roever (2006) ja Kunnan, (2003). Heidän ajatteluunsa perustuvat modernin validiusteorian monet juonteet. Niitä ei sovelleta merkittävästi käsillä olevassa tutkimuksessa, joka suuntautuu vasta pilotoimaan mahdollisesti myöhemmin laajempaan käyttöön tarkoitettua suullista koetta. Koetuloksen seurauksien tutkimus ajankohtaistuu, kun tai jos suullisen kielitaidon koe saa vaikuttavamman virallisen aseman.

### 1.3.2 Validius pragmaattisena argumenttina

Viimeaikaisen pragmaattisen suuntauksen vahvuutena on teoreettisen päättelyn ja käytännöllisten näkökohtien yhdistäminen (Fulcher & Davidson, 2007, ss. 18 – 21) ja tätä kautta myös menetelmien joustava integroituminen tilanteen ja testauskontekstin edellytysten mukaisesti. Kielitaidon arvioinnin validiustutkimuksessa lupaavaksi on osoittautunut Kanen, Crooksin ja Cohenin (1999) sovellus Toulminin (2003) argumentaatiologiikasta, jota myös Bachman (2005) on täsmentänyt. Perusajatus on yksinkertainen: testituloksen validius todentuu siinä määrin kuin tuloksen tarkoitettua tulkintaa tukeva näyttö päihittää sitä kyseenalaistavan näytön.

**Argumentoinnin perusvaiheet**

Pragmaattinen tulkinta-argumentointi soveltuu erityisen hyvin suoritusarviointiin, koska sen lähtöoletus on hankkia suoria näytteitä mitattavasta taidosta teettämällä tehtäviä, joiden on määrä olla edustavia suhteessa niihin tehtäviin ja kielenkäyttöalueisiin, joihin havaitun koesuorituksen tulos on tarkoitus yleistää. Yleistyksen pätevyydelle on ratkaisevaa, miten vakaasti on tositettavissa kolmivaiheinen kytkös lähtien havaitusta suorituksesta ja päätyen odotettavaan suoriutumiseen kyseisellä kielenkäyttöalueella (Crooks, Kane, & Cohen, 1996, s. 6).

Validiusargumenttiketjun ensimmäinen lenkki on *pisteitys* (scoring), joka tapahtuu arviointiohjeiden mukaisesti ja jonka tulisi olla mahdollisimman yhdenmukaista ja virheetöntä. Tämän vaiheen pätevyyttä koetellaan pragmaattisessa validoinnissa tarkastelemalla kriittisesti arviointiohjeita ja – menettelyjä sekä koetilanteen toimeenpanoa. (Crooks, Kane, & Cohen, 1996, ss. 9 - 10).

Validiusargumentin rakentamisen toinen vaihe on *yleistys* (generalization), joka tapahtuu havaitusta testituloksesta samankaltaisiin tehtäviin tosielämässä. Tämän vaiheen tuloksena on nk. universaalitestitulos tai universaalipistemäärä (universal score). Oletetaan, että henkilö joka saa havaitusta suorituksesta tietyn tuloksen, saisi saman tuloksen myös koetehtävän kaltaisista tai sen kanssa vaihtoehtoisista tehtävistä, joihin

koetulos halutaan yleistää. Tilastollisesti perusteltu yleistys edellyttää edustavaa otosta yleistysavaruudesta, mikä kompleksien taitosuoritusten arvioinnissa on harvoin mahdollista. Yleistysehtoja pyritään parantamaan tarjoamalla useampia tehtävävaihtoehtoja ja toisaalta rajoittamalla kriittisten koetehtäväpiirteiden määrää. (Crooks, Kane, & Cohen, 1996, s. 10). Tyypillisiä heikkouksia ovat esimerkiksi suullisen kielitaidon arvioinnissa otoksen (usein pikemminkin näytteen) koko ja edustavuus sekä monet mitattavan taidon kannalta epärelevantit tekijät, kuten tehtävien ja arvioijien piirteet, testin toimeenpano ja suorituskontekstiin kuuluvat tekijät. Reliaabeliutta kohennetaan tavallisesti lisäämällä riippumattomien havaintojen määrää, mutta suoritusarvioinnissa tähän vaaditaan mittavat aika- ja henkilöresurssit. Siksi Crooks, Kane ja Cohen (1999, s. 10) suosittavat ensisijaisesti tehtäväpiirrekimppujen standardointia tehtävävaihtoehtojen yhteismitallisuuden turvaamiseksi ja arviointimenettelyjen tiukkaa standardointia.

Päättelyketjun kolmas vaihe on *ekstrapolointi* (extrapolation). Edellisessä vaiheessa yleistyksen kautta saatiin testitulos, joka on tarkoitus yleistää koetehtävän kaltaisiin tehtäviin. Ekstrapoloinnissa tämä testitulos yleistetään edelleen koko sille kielenkäyttöalueelle (kohdealueelle), jolta testitehtävä on poimittu. Kohdealue on määritelmällisesti laajempi ja rajoiltaan häilyvämpi kuin yleistysavaruus, ja etenkin yleissivistävässä kielikoulutuksessa kohdealueet voivat olla erittäin monimuotoisia (kuten arkielämä tai aikuiskoulutus). Varmuuden aste riippuu siitä, miten samankaltaisia yleistysavaruus ja kohdealue ovat keskenään. Tosielämän kielenkäytöstä irrotettujen simuloitujen tehtävien voisi ajatella vastaavan kohdealuetta heikommin kuin vaikkapa autenttisessa työtilanteessa annetun suoritusnäytön. (Crooks, Kane, & Cohen, 1996, s. 10) Koska koetehtävien ja tosielämän tehtävien verrannettavuutta on yleensä vaikea todentaa, kokeenlaatijoita neuvotaan yleensä pyrkimään siihen, että koesuoritus vaatisi suunnilleen samoja tietoja ja taitoja kuin tosielämän kriteerisuoritus.

Tulkinta-argumentin viimeinen rakennusvaihe on arviointituloksen *käyttö* (utilization), mikä viittaa suoraan arvioinnin sosiaalisiin ja poliittisiin kytköksiin edellä esiteltyjen modernien validiusnäkemysten hengessä. Kysymys on testitulokseen perustuvan päätöksenteon monitahoisista seurauksista sekä yksilön että yhteiskunnan kannalta. Kane (2004) puhuu arvioinnin käyttöargumentista, joka muodostuu tulkinta-argumentista ja validiusargumentista. Validiusargumentti kattaa suunnilleen perinteisen, psykometrisen validiustarkastelun, ja sen menettelyistä vallitsee melko vakiintunut käsitys arvioinnin asiantuntijayhteisössä. Tuoreempi tulokas tulkinta-argumentti sen sijaan näyttää jakavan käsityksiä jonkin verran sen suhteen, missä määrin kielitaidon arvioijien voidaan katsoa olevan vastuussa testien käytön sosiaalisista ja yhteiskunnallisista seurauksista (Bachman, 2005, s.28). Hy-Talk –projektissa keskitytään lähinnä validiusargumenttiin, koska laaditulla puhekokeella on tässä vaiheessa vain tutkimuksellisia seurauksia.

**Validiusargumentin rakenne**

Argumentointiperustaisessa validiustarkastelussa jokainen vaihe kielennetään selkeästi siten, että sekä taustaoletukset että tulkintavaihtoehdot muotoillaan mahdollisimman yksiselitteisiksi lausumiksi (Crooks, Kane, & Cohen, 1996, s. 6). Lähtökohtana on väite

(claim), perusteltava johtopäätös, jota tuetaan ja koetellaan erityyppisen evidenssin valossa. Evidenssi kytketään väitteeseen puoltolauseen (warrant) avulla (”koska on näin, niin siitä seuraa, että väite on tältä osin pätevä”). Näyttö nojaa dataan (data), joka koostuu väitteen tukena esitettävästä informaatiosta, kuten koehenkilöiden suorituksista. (Toulmin, 2003, s. 90; Bachman, 2005, s. 9.) Tarkoitettua johtopäätöstä tukeva näyttö (backing) on aiemman tutkimuksen tuottamaa tietoa, joka tukee puoltolausetta tai varta vasten arviointitutkimusta varten kerättyä informaatiota kuten suorittajapalautteita tai arviointikokousmuistioita (Bachman, 2005, s. 10; Fulcher & Davidson, 2007, s. 165). Data ja muu puoltava näyttö on yhdistettävissä johtopäätöstä kannattaviksi perusteiksi (grounds), kuten Fulcher ja Davidson ehdottavat (2007, ss. 164 – 165).

.

Vastaavasti väitettä horjuttamaan on olemassa rinnasteinen, jollei sitten –lausumille perustuva rakennelma, vasta-argumentti. Vasta-argumentin pohjana oleva kyseenalaistava näyttö nojautuu horjuttavaan dataan (rebuttal data), joka kytkeytyy vastaväitteeseen varauksien (qualifier) ja vastaväitteiden (counterclaim) kautta. Näillä pyritään horjuttamaan väitettä tarjoamalla sille vaihtoehtoisia selityksiä tai osoittamalla sen rakenteessa loogisia tai empiirisiä aukkoja (Bachman, 2005, s. 10). Puutteet koskevat mitattavan taidon tai ominaisuuden kannalta asiaankuulumatonta vaihtelua (construct-irrelevant variance) testituloksissa.

Puoltolauseet (koska-lausumat) ja niitä kiistävät vastaväitteet (jollei –lausumat) koskevat sisällöltään jotakin seuraavista neljästä ulottuvuudesta: näytön nojalla tehtävän johtopäätöksen relevanssia, hyödyllisyyttä, tarkoitettuja seurauksia tai informaation riittävyyttä. Relevanssilausuman ytimessä on se, missä määrin arvioitu taito vastaa tosielämän kielenkäyttöä eli arviointitehtävän ja tosielämän kielenkäyttötehtävien samankaltaisuudesta (Bachman, 2005, s. 18). Relevanssilausumat ovat perua etenkin sisältö- ja käsitevalidiudesta ja sivuavat myös tehtävän autenttisuutta. Hyödyllisyyslausumin ilmaistuja koetuloksesta tehtävän johtopäätöksen ominaisuuksia voidaan tarkastella esimerkiksi rinnastamalla tulokset toisen, samaa ominaisuutta mittaavan kokeen tuloksiin. Hyödyllisyystarkastelu liittyy kokeen käytännöllisyysnäkökohtiin mutta myös sen seurauksiin. Tarkoitettujen seurauksien punnitseminen on sopusoinnussa arvioinnin oikeudenmukaisuutta ja eettisyyttä sekä arvioijien yhteiskunnallista vastuuta tähdentävien näkemysten kanssa. Riittävyyslausumat koskevat käytettävissä olevan arviointitiedon määrää ja kattavuutta tehtävän johtopäätöksen pohjaksi. Riittävyyden käsite sivuaa testin sisällön kattavuutta ja käsitevalidiutta sekä myös kielitaidon ja muiden kompetenssien suhteita koesuorituksessa. (Bachman, 2005, s. 19.) Tämän alueen ongelmat ovat erityisen tuttuja vuorovaikutteisessa puheen arvioinnissa.

## 2 Validiustarkastelu Hy-Talk –projektissa

### 2.1 Projektin tausta ja nykytila

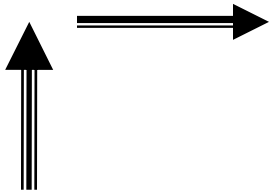
Hy-Talk projekti on Helsingin yliopiston tutkimusvaroista rahoitettu hanke, jonka toimijat edustavat kielitieteitä ja kielididaktiikkaa (ks. lähemmin http://blogs.helsinki.fi/hy-talk/). Hankkeen tutkimustavoite on kielten valtakunnallisiin

opetussuunnitelmaperusteisiin kuuluvan kielitaidon tasojen kuvausasteikon validointi (ks. http://www.oph.fi/SubPage.asp?path=1,17627,1558; LOPS, 2003; POPS, 2004). Asteikko on sovellus Eurooppalaisen viitekehyksen taitotasoasteikoista (CEFR, 2001), joihin sillä on empiirisesti todennettu kytkös (Hildén ja Takala, 2007). Asteikolla määritellään perusopetuksen ja lukion oppilaiden keskimääräinen tavoitetaso kussakin kielessä kolmessa ns. nivelvaiheessa eli perusopetuksen 6. ja 9. vuosikurssin sekä lukion päättyessä. Näissä kuvattua suullista kielitaitoa mittaamaan laadittiin kolme koetta, joissa kaikissa on yksi monologinen ankkuritehtävä ja 2-4 tasonmukaista keskustelutehtävää. Noin 200 opiskelijaa suoritti nämä tehtävät ruotsin, englannin, saksan tai ranskan kielellä, ja suoritukset tallennettiin tarkoitusta varten perustettuun tietokantaan. 5-10 kunkin kielen asiantuntijaa arvioi näytteet opetussuunnitelmaperusteisiin sisältyvän taitotasoasteikon ulottuvuuksien mukaan. Arviot yhdistetään matriiseiksi tietokantaan, jonne niiden lisäksi tallennetaan suoritusten ja niiden valmistelutuokioiden litteroinnit tutkimuskäyttöön projektikaudella ja sen jälkeen.

## 2.2 Validiusargumentti projektin tutkimuskehyksenä

Taulukosta 1 ilmenee, miten validiuden eriaikaisia kerrostumia on sovellettu tutkimusprojektin ongelmanasetteluun.

Taulukko 1. Validointiargumentin sovellus Hy-Talk –projektiin (Fulcher & Davidson, 2007, ss. 164 – 174; Bachman, 2005 mukaillen)

| | |
|---|---|
|  | **Väite:**<br>Kansallisiin opetussuunnitelmaperusteisiin sisältyvän puhumisen taitotasoasteikon kuvausasteikot ja käytetty puhekoe mahdollistavat pätevät johtopäätökset yleissivistävän koulutuksen opiskelijoiden suullisesta kielitaidosta. |
| **… koska**<br>**Puoltava näyttö: Puoltolauseet ja niitä tukeva näyttö** | **… jollei sitten**<br>**Kiistävä näyttö:**<br>**Varaukset ja vastaväitteet** |
| Saatu näyttö ja sen yhteydet aikaisempaan tutkimukseen tukevat käytetyn arviointiasteikon ja laaditun puhekokeen sisältöjä pätevinä välineinä mitata suullista kielitaitoa. | Kiistävä näyttö kumoaa puoltavan näytön ja osoitetaan, että käytetty arviointiasteikko ja laadittu puhekoe eivät ole päteviä välineitä mitata suullista kielitaitoa. |
| **Puoltolauseet** ↑<br>(esimerkkejä) | **Vastaväitteet** ↑<br>(esimerkkejä) |
| **1. relevanssi** | |

| | |
|---|---|
| Koetehtävät vastaavat opetussuunnitelman tavoitteita ja tosielämän kielenkäyttöä.<br><br>Arviointiasteikon tasokuvaukset ovat oleellisia puhumisen arvioinnin välineitä.<br><br>Analyyttiset piirrearviot mukailevat johdonmukaisesti holistista kokonaisarviota puhetaidon tasosta. | Koetehtävät sisältävät opetussuunnitelman tavoitteiden tai tosielämän kielenkäytön kannalta epärelevantteja piirteitä.<br><br>Arviointiasteikon tasokuvaukset sisältävät suullisen kielitaidon käsitteelle epärelevantteja aineksia.<br><br>Analyyttiset piirrearviot tai jotkut niistä eivät liity johdonmukaisesti kokonaisarvioon puhetaidon tasosta. |
| **2. hyödyllisyys** | |
| Puhekoe katsotaan mahdolliseksi panna toimeen koulun arjessa.<br><br>Arviointiasteikon käyttö koetaan resursseihin nähden kohtuullisen vaivattomaksi.<br><br>Puhenäytteiden tasoarvioista vallitsee riittävä yksimielisyys arviojien kesken.<br>Eri arvioijat käyttävät asteikkoa johdonmukaisesti ja yhtenevin painotuksin. | Puhekoe tai jotkut sen osat nähdään epäkäytännöllisinä tai kohtuuttomina koulun arjessa toteutettaviksi.<br>Arviointiasteikon käyttö ja/tai tulkinta koetaan kohtuuttoman työlääksi.<br><br>Arvioijat ovat erimielisiä puhenäytteiden tasoarvioista.<br><br>Eri arvioijat painottavat arvioinnissaan eri asioita ja tulkitsevat asteikkoa hyvin eri tavoin. |
| **3. tarkoitetut seuraukset** (ei sovelleta) | |
| Hyvä menestys puhekokeessa auttaa opiskelijaa opinnoissa tai työllistymisessä<br><br>Heikko menestys puhekokeessa huonontaa opiskelijan mahdollisuuksia opinnoissa tai työllistymisessä. | Puhekokeessa menestymisellä ei ole merkitystä opintojen tai työllistymisen kannalta.<br><br>Puhekokeessa menestyminen on ristiriidassa myöhemmän opinto- tai työmenestymisen kanssa. |
| **4. riittävyys** | |
| Tehtävät ja arviointiasteikon tasokuvaukset kattavat arviointidatasta ilmenevät olennaiset puheviestinnän piirteet.<br><br>Arvioinnin ulkopuolelle ei jää olennaisia ilmiöitä. | Suoritusanalyysi ilmentää tekijöitä, joita tehtävä- tai tasokuvaukset eivät kata.<br><br>On perusteltua olettaa, että em. tekijät tulisi ottaa huomioon puheviestinnän arvioinnissa. |

| | |
|---|---|
| **Data ↑** | |
| Väitettä tukeva data / Väitettä horjuttava data:<br><br>- Suoritustallenteet, tallenteet suunnittelu- ja palautetuokioista sekä arviointisessioista, kirjallinen suorittaja- ja arvioijapalaute, tallennelitteroinnit, tehtäväkuvaukset<br>- Aiemmat tutkimustulokset | |

Kysessä on teoreettinen ongelmakehikko, johon kunkin tutkijan valitsemat varsinaiset tutkimusongelmat ja –kysymykset sijoittuvat. Jäsennys on alustava ja muuttuva sikäli, että samakin tutkimuskysymys voisi sijoittua useampaan kategoriaan tutkijan intentiosta riippuen. Validiusargumentin relevanssia voidaan punnita rinnastamalla arviointiasteikon sisältöjä teoriatietoon suullisen kielitaidon relevanteista ulottuvuuksista ja niiden arvioinnista. Tähän liittyen asetetaan kysymyksiä, jotka palautuvat suullisen kielitaidon käsitteeseen osana kielenkäyttäjän kompetensseja sekä tämän taitoalueen sisäiseen koherenssiin mitattuna arviointiasteikoissa kuvatuilla kriteeripiirteillä. Relevanssin saralle kuuluisivat ne muiden muassa kysymyksenasettelut, jotka koskevat virhetaajuuden, kielen rikkauden tai erilaisten sujuvuusindikaattorien suhdetta tasoarvioon. Projektimme relevanssisidonnaisten tutkimusongelmien käsittelyssä sovelletaan sisältö- ja kriteerivalidiuden perinteisiä välineitä, kuten rinnastettavien piirteiden korrelointia, monimuuttuja-analyysejä sekä asteikkojen loglineaaristen ominaisuuksien tarkastelua. Määrällisen datan rinnalla kerätään ja analysoidaan laadullista, toistaiseksi tai luonteeltaan kokonaan skaalautumatonta dataa ja kuvataan mahdollisia tasosidonnaisia säännönmukaisuuksia. Relevanssia voidaan kartoittaa myös vertaamalla testitehtävien piirteitä opetussuunnitelman tavoitteisiin tai tosielämän kielenkäyttötehtäviin esimerkiksi sisällönanalyysin keinoin.

Koetuloksesta tehtävien johtopäätösten hyödyllisyysnäkökohtien tarkastelu pohjautuu kokeen ja arviointiasteikon käyttäjien tekemiin taitotasopäätöksiin ja heidän raportoimiinsa kokemuksiin ja näkemyksiin. Arvioijien keskinäistä yksimielisyyttä tarkastellaan reliaabeliuden todentamiseen vakiintunein tilastollisin keinoin, ja laadullista dataa saadaan eri osapuolten (testattavien oppilaiden, kokeen toimeenpanijoiden ja arvioijien) suullisista ja kirjallisista raporteista. Erityisesti olemme kiinnostuneita koetehtävien ja asteikon toimivuudesta koulukontekstissa.

Kuten aiemmin tuli todetuksi, projektissa järjestettävä puhekoe palvelee ainoastaan tutkimustarkoitusta eikä siitä suoriutumisella siten ole vaikuttavuuden käsitteen edellyttämiä seurauksia yksilö- eikä järjestelmätasolla.

Arviointievidenssin riittävyyttä punnittaessa kysytään, tarvitaanko itse testin antaman informaation ohella vielä jotain muuta tietoa tukemaan tehtävää johtopäätöstä testatun henkilön suullisesta kielitaidosta. Riittävyys liittyy siis yksilötason seurauksiin, joiden puuttuessa tämäntyyppisten puoltolauseiden ja vastaväitteiden laatiminen ei ole kovin

luontevaa. Tässä yhteydessä koeteltava väite koskee käytetyn kokeen ja arviointiasteikon pätevyyttä, ei yksittäisten suorittajien kielitaitoa sinänsä. Siksi myös riittävyys voidaan nähdä suhteessa projektin tavoitteeseen, joka on käytetyn arviointiasteikon validointi. Silloin riittävyyden alueella merkityksellisiä voivat olla ne havainnot ja tutkimustulokset, jotka saadaan relevanssin ja hyödyllisyyden analyysien yhteydessä ja jotka niissä luokittuvat vasta-argumenteiksi, koska eivät vastaa ennakko-oletuksia. Suullisen kielitaidon arvioinnin ongelmallisuudesta on jo paljon tutkimustietoa (esim. Brown, 2003; Chalhoub-Deville, 2003; Luoma, 2004; Fulcher, 2003), joka ei kuitenkaan ole ratkaissut vuorovaikutuksessa elävän taidon arviointihaasteita.

Hy-Talk –projektissa näytön riittämättömyyden alueelle sijoittuvat tulokset edistävät tietämystä suorituksen rakentumisesta puhumista mittaavissa kokeissa ja arvioinnin eri vaiheissa ja sitä kautta myös viitteitä suullisen kielitaidon arvioinnin ilmeisistä kehittämiskohteista.

Lähteet


Angoff, W. H. (1988). Validity: An evolving concept. Teoksessa H. Wainer & H. I. Braun (toim.), *Test validity* (ss. 19 – 33). Hillsdale, NJ: Lawrence Erlbaum.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Bachman. L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly,* 2(1), 1–34.

Borsboom, D., Mellenbergh, G.J. & van Heerden, J. (2004). The concept of validity. *Psychological Review* 111, 1061 - 1071.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20,* 1–25.

Carroll, B. J. (1980). Specifications for an English language testing service. Teoksessa J.C. Alderson & A. Hughes (toim.), *Issues in Language Testing. ELT Documents 111* (ss. 66 – 110). London. British Council.

CEFR. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Council of Europe: Cambridge University Press.

Chalhoub-Deville, M. (2003). L2 interaction: current perspectives and future trends. *Language Testing* 20 (4), 369 – 383.

Cronbach, L. J. (1971). Test validation. Teoksessa R. L. Thorndike (toim.), *Educational Measurement* (ss. 443 – 507).Washington, DC: American Council of Education,

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281 – 302.

Davies, A. (2003). Three heresies of language testing research. *Language Testing, 20* (4), 355 – 368.

Ebel, R. L. (1961). The relation of scale fineness to grade accuracy. *Journal of Educational Measurement, 6* (4), 217 - 221.

Fulcher, G. 2003. Testing second language speaking. London: Longman.

Fulcher, G. & Davidson, F. (2007). *Language Testing and Assessment. An advanced resource book*. Abington & New York: Routledge.

Hamp-Lions, L. (2001) Ethics, fairness(es) and developments in language testing. Teoksessa C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. McNamara & K. O'Loughlin (toim.), Experimenting with uncertainty: Essays in honour of Alan Davies. Studies in language testing 11 (ss. 222 – 227). Cambridge: Cambridge University Press.

Hildén, R. & Takala, S. 2007. Relating Descriptors of the Finnish School Scale to the CEF Overall Scales for Communicative Activities. Teoksessa Koskensalo, A., Smeds, J., Kaikkonen, P. & Kohonen, V. (toim.) *Foreign languages and multicultural perspectives in the European context; Fremdsprachen und multikulturelle Perspektiven im europäischen Kontext.* Dichtung, Wahrheit und Sprache (ss. 73 – 88). LIT-Verlag.

Kane, M. D. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38* (4), 319 – 342.

Kane, M. D. (2006).Validity. In R. L. Brennan, (toim.), *Educational Measurement* (4th edition), (pp. 17 – 64). Westport, CT: Praeger.

Kane, M., Crooks, T. & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice* 18 (2), 5 – 17.

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York. Macmillan.

Kunnan, A. (2003). Fairness and justice for all. Teoksessa A. J. Kunnan (toim.) *Fairness and validation in language assessment.* Studies in language testing 9, ( ss. 1 – 14). Cambridge: Cambridge University Press.

LOPS (2003). Lukion opetussuunnitelman perusteet 2003. Helsinki: Opetushallitus.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. L*anguage Testing, 18* (4), 351–372.

Luoma, S. (2004) *Assessing speaking*. Cambridge: Cambridge University Press

McNamara, T. (1997). ´Interaction` in second language performance assessment: Whose performance? *Applied Linguistics*, 18 (4), 446–465.

McNamara, T. & Roever, C. (2006). *Language testing: The social dimension*. Language Learning: Monograph Series. London: Blackwell.

Messick, S. (1989). Validity. Teoksessa R. L. Linn (toim.), *Educational measurement* (13 – 103). NY: McMillan.

POPS (2004). Perusopetuksen opetussuunnitelman perusteet 2004. Helsinki: Opetushallitus.

Shohamy, E. (2001). *The power of tests: a critical perspective on the uses of language tests.* London: Longman.

Spolsky, B. (1995). *Measured words. The development of objective language testing.* Oxford: Oxford University Press.

Toulmin, S. E. (2003). *The uses of argument.* Cambridge: Cambridge University Press.

Weir, C. (2005). *Language testing and validation. An evidence-based approach.* New York: Palgrave Macmillan.