KRITEERIMITTAAMISEN KÄSITTEESTÄ JA KÄYTÄNNÖN SOVELLUKSISTA

On the concept and practical applications of criterion-referenced measurement

Sauli Takala

Kasvatustieteiden tutkimuslaitos. Selosteita ja tiedotteita 146/1980
ISBN 951-678-337-6
ISSN 0357-122X

Sauli Takala

Kriteerimittaamisen käsitteestä ja käytännön sovelluksista

Kasvatustieteiden tutkimuslaitos. Selosteita ja tiedotteita 146/1980
ISBN 951-678-337-6
ISSN 0357-122X

Tässä raportissa kuvataan kriteerimittaamisen käsitettä, kehitystä ja käytännön sovelluksia. Julkaisu on osa kirjoittajan laajempaa tutkimusohjelmaa, joka käsittelee opetussuunnitelmien, erityisesti vieraiden kielten opetussuunnitelmien, laadinnan teoreettisia ja käytännöllisiä ongelmia. Raportti täydentää kirjoittajan aikaisempaa julkaisua, jossa käsitellään vaatimustasojen asettamisen ongelmia opetussuunnitelmia laadittaessa. Se liittyy läheisesti myös Kasvatustieteiden tutkimuslaitoksen evaluaatio-osastolla käynnissä olevaan laajaan evaluaatiotutkimushankkeeseen nimeltä "Peruskoulun tilannekartoitus I".

Kriteerimittaamisen voidaan sanoa lähteneen liikkeelle varsinaisesti vasta 1970-luvulla. Osittain samaan aikaan tapahtui vilkasta kehittämistyötä ns. modernin testiteorian alalla. Moderni testiteoria on antanut tilastomatemaattisia työkaluja kriteerimittaamisen ongelmien ratkaisemiseen, jotka ovat tyypillisiä evaluaatiotutkimuksen ongelmia.

Kriteerimittaamisen alalla eivät käsitteet ymmärrettävästi ole vielä vakiintuneet. Käynnissä on kriteerimittaamisen periaatteiden, käyttö- ja menettelytapojen, tulosten tulkinnan ja niiden validiteetin ja reliabiliteetin tutkiminen ja määrittäminen. Selvitystä odottavat monet ratkaisemattomat ongelmat ja ongelmalliset ratkaisut.

Vaikka kriteerimittaamisen käsite on vielä jonkin verran vakiintumaton, voidaan sen erityisenä ansiona pitää sitä, että se antaa tarkan kuvauksen henkilön suoritustasosta tietyllä tarkasti määritellyllä käyttäytymis- ja sisältöalueella. Siinä missä perinteellinen normimittaaminen antaa tietoa henkilön suoriutumisesta muihin nähden, kriteerimittaaminen antaa kuvan henkilön suoriutumisesta tiettyä sisältöaluetta edustavista tehtävistä. Normikoetta voitaisiinkin nimittää *myös* erottelukokeeksi ja kriteerikoetta ehkä sisältökokeeksi tai suoritustasokuvailukokeeksi.

Keskeisiä tehtäviä kriteerimittaamisessa on aluetäsmennyksen laatiminen. Tässä on käytetty erilaisia menetelmiä, joista tunnetuimpia ovat kielitieteeseen perustuvat menetelmät (Bormuth), osiolomake (Hively),

piirreanalyysi (Guttman), lavennetut tavoitelauseet (Popham) ja koetäsmennys (Popham). Ne käsittävät kokeen ärsyke- ja reaktio-osan täsmennyksen sekä vastausten pisteistyssysteemin määrittämisen.

Osioita laadittaessa ja valittaessa tai karsittaessa on erityisesti pidettävä huolta siitä, että osiot todella vastaavat määriteltyä aluetta. Kokeen sisällön validiteettin (edustavuuden) eli kuvauksen validiteetin arvioinnissa voidaan käyttää apuna sekä asiantuntijoiden harkintaa että empiiristä osioanalyysiä. Erityisen tärkeätä on huomata, ettei empiiristä osioanalyysiä kuitenkaan voida käyttää osioiden karsimiseen vaan lähinnä puutteellisten osioiden paljastamiseen. Empiiriseen osioanalyysiin perustuva osioiden karsinta saattaisi kohtalokkaalla tavalla heikentää kokeen sisällön validiteettia.

Muita kriteerimittaamiseen liittyviä keskeisiä ongelmia ovat mm. seuraavat: 1) Millainen tulisi olla mitattavan osa-alueen koko? 2) Miten läheisesti kokeen osioiden tulisi liittyä opetukseen? 3) Miten osion sisältö ja muoto vaikuttavat sen vaikeuteen? 4) Millainen vaatimustaso tulisi asettaa hyväksyttävälle suoritukselle? 5) Miten kokeen reliabiliteetti ja validiteetti tulisi arvioida? Näissä kysymyksissä vallitsee jonkin verran erilaisia näkemyksiä kriteerimittaamisen asiantuntijoiden parissa.

Kriteerimittaamisella on useita käyttötapoja. Sitä voidaan käyttää apuna tarveanalyysejä suoritettaessa ja opetuksen yksilöllistämisessä. Erityisen lupaavalta tuntuu kriteerimittaamisen anti opetussuunnitelmien ja koulujärjestelmän laadullisen tuotoksen arvioinnille. Kriteerimittaamisen menetelmiä voidaan käyttää kokelaiden aluepistemääriä eli suoritustasoa arvioitaessa ja tarvittavan kokeenpituuden arvioinnissa.

Kriteerimittaamisen kehityksessä on ollut havaittavana tieteen kasvulle ehkä yleinen piirre, että alussa korostetaan eroja aikaisempaan lähestymistapaan ja myöhemmin etsitään yhteisiä piirteitä. Aluksi torjutaan monet aikaisemman lähestymistavan menetelmät, mutta ajan kuluessa niiden käyttömahdollisuudet ollaan valmiit myöntämään. Tällainen eroja korostava alkudogmaattisuus saattaa olla paitsi edullista myös välttämätöntä, jotta uuden lähestymistavan mahdollisuudet ja rajoitukset tulisivat riittävän perusteellisesti tutkituksi. Esimerkkeinä tällaisista piirteistä mainittakoon suoritustason hajontaa, empiiristä osioanalyysiä ja normitietoja koskevat käsitykset.

Moderni testiteoria ja kriteerimittaaminen edustavat uudenlaista ajattelutapaa. Ne tuovat terävästi esille ongelmia, mutta ne tarjoavat myös uusia mahdollisuuksia. Näitä mahdollisuuksia ei saa ilmaiseksi,

vaan ne on ostettava runsaan ajattelu- ja kehittelytyön kautta. Moderni testiteoria ja kriteerimittaaminen ovat selvästi tuoneet ilmi mittaamisen vaativuuden ja vaikeuden. Mekaaniset keittokirjareseptit eivät tule kysymykseen. Kriteerimittaaminen edellyttää tietoa, ymmärtämystä, älykkyyttä ja luovuutta. Näin kriteerimittaaminen korostaa asiantuntemusta ja sen hyväksikäyttöä ennen varsinaista mittaamista. Kun perinteellisessa mittaamisessa kiinnitettiin paljon huomiota empiiristen mittalukujen selvittämiseen osioanalyysin avulla, kriteerimittaamisessa suurin osa työstä ja tiukimmat vaatimukset kohdistuvat kokeen ja sen osioiden tuottamiseen. Yksinkertaistaen voidaan sanoa, että jälkikäsittelystä on siirretty painopistettä ennakkoajatteluun.

Hakusanat:

- evaluaatio
- koetoiminta
- mittaaminen
- mittaustekniikka
- metodologia
- kriteerimittaaminen
- normimittaaminen
- moderni testiteoria

Descriptors:

- evaluation
- test
- measurement
- measurement technique
- methodology
- criterion-referenced measurement
- norm-referenced measurement
- test theory

Sauli Takala

On the concept and practical applications of criterion-referenced measurement

Institute for Educational Research. Bulletin 146/1980
ISBN 951-678-337-6
ISSN 0357-122X

The report deals with the concept, development and applications of criterion-referenced measurement (CRM). It is part of the author's larger research programme dealing with some of the theoretical and practical problems related to FL syllabus construction. The report is a supplement to the author's previous report which focussed on the problem of setting standards in curriculum construction. It has been carried out within the framework of the first national assessment of educational progress in the comprehensive school, in progress at the Institute.

Criterion-referenced measurement can be said to have been launched first in the 1970's. Concurrently with CRM there was intensive developmental work on behavioral measurement, which has resulted in the so-called modern test theory. Modern test theory has given statistical tools for solving problems related to criterion-referenced measurement, which are typically problems of evaluation research.

It can be expected that the concepts within CRM are not fully established yet. At the moment there is intensive work on the study and definition of the concepts, procedures, uses, interpretation, validity and reliability of criterion-referenced measurement. There are still a number of unsolved problems and problematic solutions that need to be addressed by researchers.

Although the concept of CRM is still not fully settled, it can be said that a particular advantage of CRM is that it provides a good description of a person's status with respect to a well-defined behavioral domain. Whereas norm-referenced measurement (NRM) describes a person's status relative to other persons, CRM describes a person's level of performance within a well-defined domain. Norm-referenced tests (NRT) might be called differentiating tests and criterion-referenced tests (CRT) content tests.

One of the most Central tasks in CRM is domain specification. Several approaches have been used for this purpose, including linguistic methods (Bormuth), item form (Hively), facet analysis (Guttman), amplified objectives (Popham) and test specification (Popham). They all specify the stimulus and response attributes of the test carefully and provide specific coding instructions.

In the generation and selection of items special attention must be given to the fact that items really match the domain. The descriptive or content validity of items is based on the judgement of content specialists and can be supplemented by empirical item analyses, It is important to note that empirical item analysis cannot be used to select items. Its proper function is to reveal flawed items. The screening of items on the basis of empirical item analysis might seriously jeopardize the content validity of the test.

Other Central questions related to CRM are, inter alii, the following: (i) What should be the proper size of the domain? (ii) How closely should the items be related to the curriculum and instruction? (iii) How do the content and wording of the items affect its difficulty? (iv) How should the passing score be set? (v) How should the reliability and validity of CRTs be determined? There are somewhat divergent opinions about these questions among the experts of CRM and they need further research.

There are several applications of criterion-referenced measurement in education. It can be used with benefit in needs assessment and in the individualizing of instruction. CRM promises to be a very useful contribution to the evaluation of curricula and of the quality of the whole educational system. CRM techniques can also be used in the estimation of domain scores, allocation of students to mastery classes and in determining the required length of the test.

The history of CRM exhibits the presumably common characteristic of the growth of science, that at first the divergent features of a new approach are stressed but later common features shared by the new and older approaches are sought. First the methods of the older approaches are rejected and their potential merits are acknowledged only at a later stage. This initial dogmatism may not only be useful but even necessary, so that the possibilities and limitations of the new approach are investigated thoroughly enough. Three instances may be cited to illustrate this point: the question of variance in the performance scores, the role of item analysis, and the role of norm data in criterion-referenced measurement.

Modern test theory and criterion-referenced measurement represent new lines of thought. They highlight problems but they also offer new possibilities. Thece possibilities cannot be obtained free; they require a lot of thought and developmental work. One of the merits of modern test theory and CRM is that they have clearly brought out the difficulty of measurement. Nobody can escape seeing how demanding measurement is. Cookbook recipes will not do; a lot of knowledge, understanding, intelligence and creativity is needed. Criterion-referenced measurement emphasizes expertise and its utilization before measurement is carried out. Whereas traditional measurement devoted a lot of attention to empirical item analyses, the greatest effort in CRM is devoted to the definition and generation of test items. With some simplification it could be said that a posteriori analysis has given way to *a* priori analysis.

Descriptors:

- evaluation
- test
- measurement
- measurement technique
- methodology

- criterion-referenced measurement
- norm-referenced measurement
- test theory

ESIPUHE

Tämä julkaisu on osa kirjoittajan laajemmasta tutkimusohjelmasta, jonka puitteissa käsitellään opetussuunnitelmien, erityisesti vieraiden kielten opetussuunnitelmien, laadintaan liittyviä teoreettisia ja käytännöllisiä ongelmia. Se liittyy läheisesti *myös* Kasvatustieteiden tutkimuslaitoksessa pääasiassa Kouluhallituksen rahoituksella toteutettavaan projektiin nimeltä "Peruskoulun tilannekartoitus 1". Raportilla on liittymäkohtia *myös* professori Raimo Konttisen uutta testiteoriaa käsittelevään teokseen. Se myös täydentää eräiltä osin kirjoittajan aikaisempaa julkaisua, joka käsittelee vaatimustasojen asettamista opetussuunnitelmia laadittaessa (Selosteita ja tiedotteita No. 145).

Raportin tarkoituksena on esittää katsaus kriteerimittaamisen synnystä, kehittymisestä ja nykytilanteesta. Katsauksen laadinta on tunnetusti vaikea ja vaativa tehtävä ja tässä tapauksessa vaikeutta on lisännyt kriteerimittaamisen teorian ja käytännön vakiintumattomuus. Raportissa on jonkin verran toistoa. Tämä on osittain tarkoituksellista sikäli, että toistosta on etua uudenlaista näkemystä sisältävän asian esittelyssä. Osittain toisto johtuu kuitenkin aiheen jäsentämisen puutteista. Kriteerimittaaminen ei vielä ole siinä vaiheessa, että siitä voitaisiin helpolla laatia kovinkaan selkeätä ja johdonmukaista yleisesitystä. Terminologian ja kriteerimittausta käsittelevien ilmaisujen vakiintumattomuus englannin kielessä – suomen kielestä puhumattakaan – ovat *myös* omiaan vaikeuttamaan raportin luettavuutta. Käsillä olevaa raporttia ei olekaan tarkoitettu miksikään definitiiviseksi alan esitykseksi vaan johdannoksi.

Professori Raimo Konttinen on lukenut käsikirjoituksen ja tehnyt siihen eräitä hyödyllisiä tarkennusehdotuksia.

Jyväskylässä toukokuun 7 päivänä 1980

Sauli Takala

SISÄLTÖ

1. Johdanto 1
2. Katsaus kriteerimittaamisen kehittymisestä 3
2.1. Kriteerimittauksen kehittymisen syistä 3
2.2. Kriteerimittaamisen käsitteistä ja kehittymisestä 4
2.2.1. Kriteerimittaamisen lähtökohtia 4
2.2.2. Kriteerimittaamisen käsitteestä 6
2.2.2.1. Kriteerimittauksen määritelmiä 6
2.2.2.2. Vakiintumassa oleva käsitys kriteerimittaamisesta 7
2.2.2.3. Kriteerimittaamisen lähisukulaisia 10
2.3. Normimittaamisen ominaisuuksia 14
2.3.1. Normimittaamisen käsitteestä 14
2.3.2. Normimittaamisen rajoituksia 16
2.4. Kriteerimittaamisen ja normimittaamisen vertailua 17
2.5. Kriteerimittaamisen ja normimittaamisen käyttötapojen vertailua 20
3. Keskeisiä vaiheita kriteerimittareiden laadinnassa 22
3.1. Alkukatsaus 22
3.2. Mitattavan osa-alueen määrittäminen 25
3.2.1. Yleisiä näkökohtia 25
3.2.2. Kielitieteeseen perustuvat menetelmät 26
3.2.3. Osiolomake 28
3.2.4. Piirreanalyysi 30
3.2.5. Lavennetut tavoitelauseet 32
3.2.6. Koetasmennys 33
3.2.7. Mitattavan osa-alueen määrittämisen ongelmia 46
3.2.7.1. Alueen koko ja sitä määrittelevät piirteet 46
3.2.7.2. Osioiden ja opetuksen välinen yhteys 48
3.3. Osioiden tuottaminen ja valikoiminen 50
3.4. Osion sisällön ja muodon merkitys 52
4. Vaatimustasojen asettaminen 54
4.1. Vaatimustasojen asettamisen perusteista 54
4.2. Hyväksymisrajan asettamisen menetelmiä 56
4.2.1. Muiden suoritustasoon vertaaminen 56

4.2.2. 100 %:sta tinkiminen 57
4.2.3. Muuhun kriteeriin vertaaminen 57
4.2.4. Osioiden sisällön arviointi 58
4.2.5. Päätösteoreettiset menetelmät 63
4.2.6. Operaatiotutkimukseen perustuvat menetelmät 65
4.2.7. Vaatimustasojen seurausvaikutuksia 70
4.2.7.1. Opiskelua koskevat seuraukset 70
4.2.7.2. Psykologiset ja taloudelliset kustannukset .. 70
5. Affektiivisen alueen mittaaminen kriteeripohjaisin mittavälinein . 71
6. Reliabiliteetti kriteerimittaamisessa 76
6.1. Yleisiä näkökohtia 76
6.2. Varianssin merkitys reliabiliteetin estimoinnissa 76
6.3. Kokeen käyttötavan ja siitä tehtävien johtopäätösten merkitys reliabiliteetin estimoinnille 77
6.4. Erilaisia lähestymistapoja kriteerikokeen reliabiliteetin estimoinnissa 78
6.5. Aluepistemäärien estimaattien reliabiliteetti 79
6.6. Johtopäätösten reliabiliteetti 80
7. Validiteetti kriteerimittaamisessa 82
8. Kriteerimittaamisen käytännön sovelluksia 86
8.1. Yleisiä näkökohtia 86
8.2. Tarveanalyysit 87
8.3. Opetuksen yksilöllistäminen 87
8.4. Opetusohjelman arviointi 90
8.5. Opetuksen kehittäminen 92
8.6. Aluepistemäärien estimointi 93
8.7. Kokeen pituuden määrittäminen 97
8.8. Monitasomittaaminen 100
9. Kriteerikokeiden arviointiperusteista 102
10. Ratkaisemattomia ongelmia ja ongelmallisia ratkaisuja 106
11. Diskussio 108
Lähteet 111

1. JOHDANTO

Tämän raportin tarkoituksena on kuvata kriteeriin perustuvan mittaamisen (criterion-referenced measurement, CRM) eli lyhyemmin kriteerimittaamisen (KRM) käsitettä ja kehitystä. Julkaisu on osa kirjoittajan laajempaa tutkimusohjelmaa, joka käsittelee opetussuunnitelmien (erityisesti vieraiden kielten opetussuunnitelmien) laadinnan teoreettisia ja käytännöllisiä ongelmia. Se liittyy läheisesti Kasvatustieteiden tutkimuslaitoksen evaluaatio-osastolla käynnissä olevaan laajaan evaluaatiotutkimushankkeeseen nimeltä "Peruskoulun tilannekartoitus 1. Tilannekartoituksessa on tiettävästi ensimmäistä kertaa maassamme pyritty soveltamaan kriteerimittaamiseen läheisesti liittyvää matriisiotantaa, jossa otostetaan oppilaiden lisäksi *myös* opetustavoitteita ja -sisältöjä. Raportti täydentää tekijän aikaisempaa julkaisua, jossa käsiteltiin vaatimustason asettelun ongelmia opetussuunnitelmia laadittaessa.

Kriteerimittaamisen (KRM) ja kriteerimittareiden eli -kokeiden (KRK) kehittämisen voidaan sanoa lähteneen liikkeelle varsinaisesti vasta 1970-luvun alussa. Pari vuotta sitten käsitti alan bibliografia varovaisesti arvioiden jo ainakin 600 nimikettä. Osittain samanaikaisesti kriteerimittaamisen kanssa, joka liittyy läheisesti evaluaatiotutkimuksen ongelmiin, tapahtui ns. modernin testiteorian alalla vilkasta kehitystoimintaa (ks. esim. Konttinen 1981). Uusi testiteoria, jonka erään haaran – yleistettävyysteorian – kehittäminen vei Cronbachin (Cronbach et. al 1972) mukaan yli kymmenen vuotta, onkin luonut teoreettista pohjaa myös evaluaatiotutkimukseen liittyvän kriteerimittaamisen kehittymiselle. On hyödyllistä ja opettavaista panna merkille Cronbachin toteamus, ettei lyhyemmässä ajassa voitane kehitellä ja testata uuden lähestymistavan perusteita ja sen erilaisia aspekteja.

Kriteerimittaamisen alalla on edelleen todettavissa käsitteiden vakiintumattomuutta. Käynnissä on kriteerimittaamisen periaatteiden, käyttötapojen, menettelytapojen, tulosten tulkinnan ja niiden validiteetin ja reliabiliteetin selvittäminen ja määritteleminen. Vaikka selvää edistystä on tapahtunut, jäljellä on Popham:ia (1978) lainataksemme monia ratkaisemattomia ongelmia ja ongelmallisia ratkaisuja.

Vaikka kriteerimittaamisen käsite on vielä jonkin verran vakiintumaton, ja se täsmentyy raportin lukemisen kuluessa, on syytä jo tässä

vaiheessa todeta, että kriteerimittaamisen suurimpana ansiona pidetään yleisesti sitä, että se antaa tarkan kuvauksen kokelaan suoritustasosta tietyllä tarkasti määritellyllä käyttäytymis- ja sisältöalueella.

Tässä julkaisussa ei käsitellä kriteerimittaamisen testiteoreettisia eikä tilastomatemaattisia perusteita. Niihin voi perehtyä tutustumalla mm. Millmanin (1974), Glassin (1978), Hambletonin (Hambleton et. al. 1978) ja suomeksi Konttisen (1981) julkaisemien artikkelien ja teosten avulla. Myöskään osionlaadinnan menetelmiä ei käsitellä, koska nämä eivät kriteerimittaamisessa sanottavasti poikkea perinteellisista normimittaamisen (norm-referenced measurement, NRM) osionlaadinnan ohjeista. Niitä on kriteerimittaamisen kannaltakäsitellyt seikkaperäisesti mm. Popham (1978).

Tämän raportin painopisteet ovat toisaalla. Tarkoituksena on ensisijaisesti valaista kriteerimittaamisen perusteita ja käyttömahdollisuuksia. Tämän vuoksi raportissa keskitytään ensisijaisesti mm. seuraaviin kysymyksiin:

1) Kriteerimittaamisen käsitteen ja sen kehityksen selvittäminen.
2) Kriteerimittaamisen keskeisten periaatteiden selvittäminen.
3) Kriteeri- ja normimittaamisen yhtäläisyyksien ja erojen sekä edellisen etujen selvittäminen.
4) Kriteerimittaamisen keskeisten vaiheiden selvittäminen.
5) Kriteerimittaamisen käyttötapojen selvittäminen.
6) Kriteerimittaamisen reliabiliteetin ja validiteetin selvittäminen.

Tämän katsauksen laadinnassa on käytetty lukuisia eri lähteitä. Keskeisimpiä niistä ovat olleet Kleinin ja Kosecoffin (1973) Millmanin (Millman 1974) ja Hambletonin (Hambleton, Swaminathan, Algina & Coulson 1978) korkeatasoiset artikkelit, seka Pophamin (Popham 1978) informatiivinen ja samalla hauskasti kirjoitettu kirjamuotoinen esitys kriteerimittaamisesta. Käytetyt lähteet ovat pääasiassa amerikkalaisia, koska kriteerimittaamisen kehityksen painopiste on selvästi ollut Yhdysvalloissa. Varsin aikaisin kiinnitettiin kriteerimittaamiseen huomiota myös Ruotsissa (Wedman, 1973; Henrysson & Wedman, 1974; Jansson 1975).

Tässä raportissa keskitytään erityisesti tiedollisten koulusaavutusten mittaamiseen kriteerikokeiden avulla. Myös affektiivisen alueen kriteerimittaamisesta esitetään lyhyt kuvaus. Psykometrisen alueen mittaamista ei käsitellä, vaikka kriteerimittaamista voidaan soveltaa ja on sovellettu myös sillä alueella.

Raportissa on tiettyä päällekkäisyyttä ja toistoakin, mutta se lienee jopa toivottavaa, koska käsiteltävänä on osittain uusi asia, jonka omaksumisessa toistosta on ilmeistä etua.

2. KATSAUS KRITEERIMITTAAMISEN KEHITTYMISESTÄ

2.1. Kriteerimittauksen kehittymisen syistä

Pophamin (1978, 1-2) mukaan kokeilla oli perinteellisesti yleensä vain yksi selvä tehtävä: niitä käytettiin oppilasarvostelun pohjana ja kokeita käytettiin oppilaiden edistymisen arvioimiseksi. Diagnostinen mittaaminen oli varsin vähäistä. Tällä hetkellä kokeita kuitenkin käytetään yhä useammin koululaitoksen laadun arvioimiseen. Opettajien toimintaa saatetaan arvioida oppilaiden koesuoritusten perusteella. Samaten koko maan koululaitosta arvioidaan oppilaiden koesuoritusten valossa. Koska kokeita käytetään uusiin merkittäviin tarkoituksiin, on varsin ymmärrettävää, että myös kokeiden sisältö ja muoto ovat muuttuneet. Uutta vaihetta oppimistulosten arvioinnissa voidaan kutsua kriteeripohjaisen mittaamisen aikakaudeksi. Kriteeripohjaisella mittaamisella 1. kriteerimittaamisella pyritään saamaan aikaan selkeä kuvaus siitä, mitä oppilaan koesuoritus itse asiassa merkitsee. Koesuoritusta ei tulkita niinkään vertaamalla oppilaan suoritusta muiden suoritukseen, kuten tavallisesti meneteltiin perinteellisissa kokeissa, vaan koetulos sinänsä antaa paremman kuvan siitä, mitä oppilas osaa tehdä tai ei osaa tehdä. Kriteerimittaus on tällä hetkellä vasta kehityksensä alkuvaiheessa.

Eräs syy mittauksen uusille kehityssuunnille on Pophamin mukaan (Popham 1978, 3) se, että koulua on yhä useammin vaadittu selkeästi osoittamaan, että se toimii odotetulla tavalla. Mielipiteiden ja vakuuttelujen sijasta halutaan konkreettista osoitusta, että koululaitos todella toimii tyydyttävästi. Koulun tehokkuuden pääasiallisena indikaattorina pidetään varsin ymmärrettävistä syistä oppilaiden koesuorituksia. Koulun laadullisten tuotosten arvioimiseksi tarvitaan juuri tätä tarkoitusta varten laadittuja kriteerikokeita.

Koululaitoksen kehittymiseksi on kaikkialla käynnissä monenlaisia kokeiluohjelmia. Yhdysvalloissa näihin on Pophamin mukaan (3-5) alettu säännönmukaisesti liittää evaluaatiokomponentti. Julkista rahoitusta saavien kokeiluprojektien on täytynyt osoittaa, millaisia vaikutuksia niillä on ollut. Tällöin on asetettu suuria vaatimuksia evaluaatiossa tarvittavien mittavälineiden kehittämiselle. Joissakin paikoin, mm. Kaliforniassa, on kaavailtu ryhtyä arvioimaan yksityisen opettajan toiminnan tuloksellisuutta oppilaiden koulusaavutusten avulla. Mikäli tällainen suuntaus yleistyisi, olisi *myös* opettajien etujen mukaista tietää riittävästi mittaamisesta, jotta he olisivat vakuuttuneita siitä, että heidän toimintaansa arvioidaan kunnollisilla mittavälineillä.

Yhdysvalloissa on ollut useamman vuoden ajan käynnissä kehityssuunta, joka asettaa kyseenalaiseksi automaattisen siirtymisen luokalta toiselle ja myöskin automaattisen päästötodistuksen saamisen viimeisellä kouluvuodella (Takala 1980). Koulun penkillä kulutettu aika ("seat time") ei enää olisi ensisijainen todistuksen antamisen kriteeri vaan kokeissa osoitettu minimitaitojen hallinta. Tämä merkitsee suurta muutosta perinteelliseen menettelyyn ja edellyttää huolellisesti laadittuja kokeita.

2.2. Kriteerimittaamisen käsitteistä ja kehittymisestä

2.2.1. Kriteerimittaamisen lähtökohtia

Vaikka kriteeripohjainen mittaaminen on saanut runsaasti huomiota osakseen vasta aivan viime vuosina, Thorndike esitti jo vuonna 1913 selvästi suhteellisen ja absoluuttisen mittaamisen välisen eron. Suomessa Toivo Vahervuo (1948, 1951, 1958) on tutkinut arvostelua ja arvosanojen antamista ja selvittänyt "A-systeemin" ja "S-systeemin" eroja. Kuitenkin Robert Glaserin vuonna 1963 julkaiseman artikkelin voidaan Pophamin mukaan (Popham 1978, 9-18) katsoa tuoneen esille käsitteet "normiviitteinen mittaaminen" ja "kriteeriviitteinen mittaaminen". On mielenkiintoista todeta, että vaikka Glaserin artikkeli herätti varsin runsaasti myönteistä huomiota, kesti lähes kymmenen vuotta, ennenkuin hänen esittämänsä ideat johtivat testitekniseen kehitystoimintaan.

Mikä sai Glaserin kirjoittamaan artikkelinsa? Aiheen antoi lähinnä ohjelmoitu opetus ja sen tulosten arvioiminen. Perinteisesti oletetaan ihmista ja hänen käyttäytymistään koskevien muuttujien jakautuvan normaalijakautuman mukaisesti. Muutama oppilas saa esimerkiksi koulusaavutuskokeessa alhaisia pistemääriä ja muutama korkeita pistemääriä, kun taas suurin osa saavuttaa keskitasoisia tuloksia. *Tämä* ei välttämättä päde ohjelmoidussa opetuksessa, jossa pyritään opetettavan asian huolellisella jäsentämisellä siihen, että oppiminen edistyisi tasaisesti. Niinpä normaalijakautuman olettamukseen perustuvat kokeet, jotka mittaavat oppilaiden suhteellista asemaa jossakin oppilasjoukossa, eivät toimi hyvin ohjelmoidussa opetuksessa. Ohjelmoidun opetuksen koko ideanahan on opettaa oppilaille jotkut tietyt asiat ja näin ollen *myös* sen lopussa pidettävällä kokeella halutaan katsoa, missä määrin oppilas on oppinut opetellut asiat. Jos opetus on ollut todella tehokasta, koetulokset eivät jakaudukaan normaalijakautuman mukaisesti, vaan hyviä tuloksia on runsaasti.

Mittausmetodiikan kehittymiseen vaikutti *myös* 1960-luvulla runsaasti huomiota saavuttanut liike, joka korosti käyttäytymistavoitteiden määrittelemisen tärkeyttä. Magerin (1962) ohjelmoitu kirjanen käyttäytymistavoitteiden (behavioral objectives) laatimisesta saavutti runsaasti vastakaikua vaikkakaan ei yksimielistä hyväksymistä. Käyttäytymistavoitteiden korostuksesta oli kuitenkin se hyöty, että varsin pian kävi ilmi, etteivät perinteelliset standardikokeet soveltuneet mittaamaan tarkoin määriteltyjä oppimistavoitteita.

Uranuurtajia kriteerimittauksen alalla on epäilemättä myös Hively (1968), joka kehitteli matematiikan ja luonnontieteen koulusaavutuskokeiden sisällön määrittelemisen menetelmiä. Kleinin ja Kosecoffin (1973) katsaus osoitti, että jo 1970-luvun alussa Yhdysvalloissa oli käynnissä useita hankkeita, joiden puitteissa tehtiin työtä kriteerimittauksen kehittämiseksi.

2.2.2. Kriteerimittaamisen käsitteestä

2.2.2.1. Kriteerimittauksen määritelmiä

Glaser (1963, 519) tarkoitti kriteerimittauksella sellaista mittausta, jossa "mittaustulokset riippuvat absoluuttisesta laatuvaatimuksesta" kun taas normimittaamisessa mittaluvut "riippuvat suhteellisesta vaatimustasosta". Glaserin näkemyksen mukaan saavutusmittaamisen pohjana on käsitys, että tiedon hankkiminen on jatkumo, joka lähtee täydellisestä tietämättömyydestä ja päätyy asian täydelliseen hallintaan. Yksilön saavutustaso sijoittuu jatkumon jollekin kohtaa, jonka paikallistamiseen käytetään apuna mm. kokeessa osoitettua "käyttäytymistä". Tavoitteena olleen suoritustason saavuttamista arvioidaan kriteerikokeiden avulla. Oppilaan suoritustasoa verrataan "vaatimustasoon (standard), jota kuvaa tietynlainen käyttäytyminen saavutusjatkumon eri kohdissa".

Popham ja Husek (1969) määrittelivät kriteerikokeet "sellaisiksi kokeiksi, joita käytetään saamaan selville yksilön asema tiettyyn kriteeriin, ts. suoritustasoon nähden (status with respect to some criterion, i.e., performance standard). Haluamme tietää, mitä yksilö osaa tehdä, ei miten hän suoriutuu suhteessa muihin".

Nitko (1971, 653) tarkoittaa kriteerikokeella sellaista koetta, "joka laaditaan tietoisesti antamaan mittalukuja, jotka voidaan suoraan tulkita ennalta tarkasti määriteltyinä suoritukselle asetettavina vaatimustasoina". Nitkon määritelmässä ei mainita selvästi suoritustasojatkumoa, mutta se sisältyy kyllä implisiittisesti hänen esitykseensä.

Glaser ja Nitko (1971) määrittelivät kriteerikokeen tyypilliseksi piirteeksi sen, että "se laaditaan tarkoituksellisesti antamaan mittalukuja, jotka voidaan suoraan tulkita spesifeinä suoritustasoina (performance standards)".

Millmanin (1973) mukaan kysymys on dikotomisesti (nolla tai yksi pistettä, oikein-väärin) pisteittävän osioperusjoukon käsitteellisestä hahmottamisesta. Osioiden ei kuitenkaan tarvitse olla "fyysisesti" olemassa, saatavilla. Tärkeätä sen sijaan on, että osioperusjoukko on kuvattu niin hyvin, että voidaan hyvin yksimielisesti todeta kuuluvatko tietyt osiot siihen vai ei. Useimmissa tapauksissa tarvitaan vain edustava osio-otos osioperusjoukosta (osiouniversumista).

Millmanin (1973) mukaan osa-alueeseen kuuluvat osiot voivat olla melko heterogeenisia sisällön, muodon ja vaikeustason suhteen. Käytännössä niiden

tulisi mitata suppeaa määrää taitoja ja tietoja, jotta olisi järkevää asettaa esim. tietty yksi vaatimustaso suoritukselle eikä useita erillisiä vaatimustasoja. Millman lainaa Davisia, joka on ehdottanut, että osa-alueena voitaisiin pitää selvästi homogeenista käyttäytymisklusteria, joka opetetaan yhtenä yksikkönä.

Ebel (1970, 35) suhtautuu jossakin määrin kriittisesti kriteerimittaukseen. Hänen mukaansa olennaisin ero kriteerimittauksen ja normimittauksen välillä koskee henkilön suorituksia kuvaavia kvantitatiivisia asteikkoja. Normimittauksessa asteikko ankkuroidaan tavallisesti tietyn ryhmän keskimääräiseen suoritustasoon. Asteikon yksiköt kuvaavat suoritustason jakautumista keskitason ala- ja yläpuolelle. Kriteerimittauksessa asteikko ankkuroidaan puolestaan ääripäihin, jolloin ylemmässä ääripäässä oleva pistemäärä osoittaa asian täydellistä hallintaa ja alemmassa ääripäässä oleva pistemäärä täydellistä osaamattomuutta. Täten Ebel perustaa käsityksensä Glaserin tavoin suoritusjatkumoon. Hänen epäilyksensä kohdistuukin lähinnä kriteerimittauksen absoluuttisuustulkintaan. "Kriteerimittaustulokset kertovat mielekkäällä tavalla mitä henkilö osaa tehdä ja mitä ei. Ne eivät kerro miten hyvä tai huono hänen tiedon tai taidon tasonsa on. Erinomainen tai puutteellinen hallinta ovat välttämättä suhteellisia käsitteitä. Niitä ei voida määritellä absoluuttisesti (36)".

Kriteerimittaus on ollut suosittu suhteellisen lyhyen aikaa, mutta sitä käsittelevä kirjallisuus on laajaa ja näyttää kasvavan kiihtyvällä vauhdilla. Uudella alalla eivät käsitteet ja termit ole kovinkaan vakiintuneet, ja siksi aluksi on syytä määritellä mitä kriteerimittauksella tarkoitetaan.

2.2.2.2. Vakiintumassa oleva käsitys kriteerimittaamisesta

Ehkä tavallisin käsitys kriteerikokeista on, että sellaiset kokeet antavat tarkkaa tietoa tutkittavien spesifeistä tiedoista ja taidoista seka tuottavat pistemääriä, jotka on tulkittavissa tehtävinä tai suorituksina. Termi "kriteeri" on sekava. Kriteerikokeen pistemäärä ei viittaa kriteeriin normatiivisen standardin mielessä, vaan pikemminkin kyseessä on "spesifit" tehtävät, jotka oppilaan tulee pystyä suorittamaan, ennenkuin hän saavuttaa yhden "määritellyistä" tiedon tasoista. Tässä mielessä suoritustason mittaustulokset ovat kriteeripohjaisia (Glaser, 1963). Kriteerikokeen perinteisessä määritelmässä kriteeri merkitsee enemmän kuin

suoritusstandardi: Se osoittaa, että tehtävien suoritus voidaan tulkita kunkin yksilön kohdalta ilman että viitataan muiden vastaavaan suoritukseen.

Kriteerikokeella tarkoitetaan koetta, joka koostuu osioista otetusta satunnaisotoksesta tai ositetusta satunnaisotoksesta, kun perusjoukkona on huolellisesti määritelty tehtäväluokka tai -ryhmä (alue). "Huolellisesti määritelty" merkitsee selvästi määriteltyä osioaluetta, tehtäväjoukkoa tai osioiden tuottamismenettelyä. Osiot poistetaan vain, jos ne eivät vastaa alueen määrittelyä. Niitä ei jätetä pois, jos kaikki oppilaat osaavat ratkaista ne oikein. Itse asiassa tällainen tulos on odotettu, mikäli opetus on ollut todella tehokasta. Täten ei edellytetä osioeikä testivarianssia. Kriteerikokeen tärkeä etu on, että se sallii erityisen kriteeritulkinnan eli arvion kokeen suorittajan aluepistemäärästä tai toimintakyvyn tasosta. Tämä määritellään prosenttina siitä koko osiopopulaatiosta, jonka kokeen suorittaja osaisi vastata oikein tai tietyssä suunnassa.

Kriteerikokeen reliabiliteettina voidaan pitää oppilaan suoritustasoa koskevien arvioiden konsistenssia. Kriteerikokeen validiteettia voidaan parhaiten arvioida analysoimalla loogisesti alueen määrittelyä, osioiden tuottamissuunnitelmaa ja yksityisiä osioita. Tämä analyysi muistuttaa sisällön validiteettia, mutta sisällön validiteetti on sallinut paljon väljemmän osa-alueen sisällön määrittelyn kuin kriteerikoe.

Kriteerikokeen pistemäärä voidaan tulkita esim. seuraavasti: Oppilas kirjoitti oikein kahdeksan sanaa kymmenestä, jotka oli valikoitu satunnaisesti kuudennen luokan sanastolistasta. Arvioidaan että hän osaa kirjoittaa oikein 80 % kaikista listan sanoista. Tavanomaisesta pistemäärän tulkinnasta poikkeavasti tämä tulkinta kuvaa oppilaan asemaa suhteessa joukkoon suoritustehtäviä. Se kuvaa miten hyvin oppilas osaa suoriutua suhteessa tehtäviin pikemminkin kuin suhteessa yleisiin kuvaussanoihin (esim. tyydyttävä) tai suhteessa persentiileihin.

Usein esitetään kriteerimittauksen rajoituksena, että kyseessä on harhakuvitelma, koska kaikki mielekkyys tulee suhteellisesta arvioinnista: suoritusta ei voida tulkita ellei mittaajalla ole jotakin käsitystä millaisia pistemäärän arvojen tulisi olla. Tässä näkökannassa on tietenkin osittain perää. Ilmaisu "2000 lyöntiä kymmenessä minuutissa" tulee mielekkäämmäksi, kun sitä verrataan siihen, miten nopeasti samanlaiset testisuorittajat pystyvät kirjoittamaan koneella, mutta pistemäärällä 2000 lyöntiä kymmenessä minuutissa on myös oma mielekkyytensä. Voidaanhan sen avulla

arvioida kauanko esim. jonkin artikkelin puhtaaksikirjoittaminen tulee viemään aikaa.

Tämän näkökohdan lisäksi kriteerimittauksella on epäilemättä ollut muitakin positiivisia vaikutuksia. Se on saanut kasvattajat kiinnittämään huomiota siihen, mitä oppilaat osaavat ja mitä he eivät osaa tehdä sekä siihen missä suhteessa opetus on ollut tehokasta ja missä ei. Millmanin (1974) mukaan käsite kriteerimittaus onkin kuitenkin yleisemmin hyväksytty yleisenä tulkintatyyppinä kuin mittausmenettelynä. Kriteerimittauksen perinteelliset määritelmät ovat osaltaan syynä sekaannukseen, koska ne sallivat hyvinkin erilaisen näkemyksen siitä, miten tällaisia kokeita tulisi kehittää ja valikoida. Seuraavassa kappaleessa esitetään kaksi erilaista lähestymistapaa.

Kriteeritulkinnan luonne ja validiteetti riippuu siitä missä määrin osioiden sisältö, muoto ja valinta on ennalta määritelty (specified). Ns. tavoitteisiin perustuvat kokeet (objectives-based tests) eivät välttämättä mahdollista kriteeritulkintoja. Monille opetuksen tavoitteille on kuitenkin mahdollista kuvailla hyvinkin tarkasti osiopopulaation sisältö, josta kokeeseen valitut osiot on valittu satunnaisesti tai ositetun otannan mukaisesti. Tällä tavalla syntynyttä koetta mm. Hively (1974) kutsuu osa-aluekokeeksi (domain-referenced test, DRT).

Kriteerikoe sallii kaikkein parhaiten kriteeritulkintoja, koska kokeen pistemäärät voidaan tulkita suorimmin suoritustehtävinä, ja voidaan arvioida kuinka suuren prosentin tehtäväpopulaatiosta oppilas osaisi ratkaista oikein tai tietyssä suunnassa. On huomattava, että sisällön spesifisyys ja homogeenisuus ovat eri käsitteitä. Vaikka sallittujen osioiden tarkempi määrittely tavallisesti johtaa rajatumpaan sisältöön kuin epämääräinen spesifiointi, asia ei välttämättä ole näin. Kriteerikokeen mittaama osa-alue voi olla laaja tai yksi ainut kapea tavoite, mutta sen täytyy olla hyvin määritelty, mikä merkitsee että kokeen sisällön ja muodon rajat täytyy määritellä huolellisesti. Käsitettä "alue" tai "osa-alue" (domain) ei pidä ymmärtää laajana käsitteellisenä alana kuten "affektiivinen alue" tai "lukemistaidon alue". Kyseessä on paljon rajatumpi käsite.

Vaikka termillä "osa-aluekoe" on eräitä etuja (mm. se että kyseessä on selvästi määritellyn käyttäytymisalueen hallinnan mittaus), käytetään tässä esityksessä Pophamin suosituksen mukaisesti termiä "kriteerikoe". (Popham 1978, 94).

Kriteerikokeiden rajoituksena on Ebelin (1970, 5) mukaan, että ne "saattavat olla käyttökelpoisia vain niillä harvoilla alueilla, jotka

koskevat korkea-asteisten taitojen käyttöä rajoitetuilla toiminta-alueilla. Kriteerikokeiden käyttö näyttää vähemmän todennäköiseltä alueella, jossa korostetaan tietoa ja ymmärtämistä. Huolenaiheena on siis, että emme kykene laatimaan kokeita, jotka sallisivat kriteeritulkinnan tärkeitä kasvatustavoitteita arvioitaessa". Epäilemättä tehtävä on vaikea, mutta meillä ei ole ollut riittävästi kokemusta kriteerikokeiden laatimisesta käyttäen jäljempänä kuvattuja menetelmiä, jotta osaisimme arvioida tämän kritiikin paikkansapitävyyttä.

Kriteerikoe poikkeaa selvästi niistä mittareista, joita kasvatustieteilijät ovat käyttäneet tällä vuosisadalla. Sen lähtökohtia ovat ainakin Ebelin (1962) esittämä ajatus sisältöstandardikoepistemääristä (content standard scores) ja Cronbachin ym. (1963, 1972) ajatus osioista satunnaisena muuttujana, joka esiintyy heidän yleistettävyysteoriassaan (theory of generalizability). Hively (1962) ja Osburn (1968) olivat ensimmäisiä osa-aluekokeiden (= kriteerikokeiden) puolestapuhujia.

2.2.2.3. Kriteerimittaamisen lähisukulaisia

Useita käsitteitä on käytetty kuvaamaan kokeita, joiden väitetään antavan tietoa tutkittavien spesifeistä tiedoista ja taidoista. Käytännössä tällaiset kokeet eroavat toisistaan siinä suhteessa, kuinka eksplisiittisesti nämä tiedot ja taidot on määritelty. Ebel (1962, 15) on ehdottanut termiä sisältöstandardikoe (content-standard test) kokeille, jotka tuottavat pistemääriä, jotka osoittavat "kuinka monta prosenttia systemaattisesta otoksesta määriteltyjä tehtäviä yksilö on ratkaissut oikein". Pistemäärä pohjautuu suoraan kokeeseen sisältyviin tehtäviin. Keskeinen piirre tässä määrittelyssä on, että pistemäärien saamiseksi käytetyt prosessit - kokeen laadinta, kokeen pitäminen ja pisteistys - ovat kylliksi eksplisiittisiä ja objektiivisia, jotta eri tutkijat saisivat olennaisesti samat pistemäärät samoille henkilöille. Ebel havainnollisti käsitettään sanan merkityksiä koskevan kokeen avulla.

Osburn (1968, 96) käytti termiä universumimäärittelty koe tai osiouniversumikoe (universe-defined test) kuvaamaan "koetta, joka on laadittu ja joka pidetään siten, että koehenkilön pistemäärä antaa harhattoman estimaatin hänen pistemäärästään jossakin eksplisiittisesti määritellyssä osiosisältöuniversumissa." Kuten sisältöstandarditesteissä, määritellään osiouniversumikokeessa eksplisiittisesti kriteerit, joiden avulla osiot

valitaan tai suljetaan pois kokeesta, ja kokeeseen tulevat osiot valitaan jollakin systemaattisella tavalla osiouniversumista (= osioperusjoukosta). Keskeisin ero Ebelin sisältöstandardikokeiden ja Osburnin universumikokeiden välillä on, että Osburn käyttää Hivelyn kehittämää systemaattista menetelmää, jota kutsutaan osiolomakkeeksi ja jonka avulla tuotetaan koeosiouniversumi ja kuvataan osioiden keskeisiä ominaisuuksia. Osiolomaketta esitellään tarkemmin kuviossa 1. Näyttää silta, että Osburnilla oli mielessään heterogeenisempi koeosiokokoelma univerumissaan kuin Ebelillä, mutta kyseessä on pikemminkin aste- kuin tyyppiero. Hively suosittelee käytettäväksi käsitettä "osa-aluekoe" itseasiassa universumimääritellyistä kokeista, koska hähen mukaansa "universumi" tuo mieleen yrityksen määritellä kaikki tiedot tietyllä oppiainesalueella, mikä merkitsisi valtavaa testiosioiden joukkoa. Tämä on Hivelyn mukaan käytännössä mahdotonta. Voidaan määritellä vain joitakin keskeisiä ydinosia, joista muiden käyttäytymismuotojen odotetaan yleistyvän. Hively suosii käsitettä "alue" tai "osa-alue", koska sillä on vähemmän kunnianhimoisia mielleyhtymiä.

On esitetty myöskin käsitettä tavoitepohjainen koe, tai tavoitekoe (objective-based test). Tämä käsite on paljon epämääräisempi kuin aikaisemmin mainitut käsitteet, koska tässä tapauksessa ei ole sovittu mistään testinlaadinta- eikä validointimenettelyistä. Joidenkin kokeiden käyttäjien mielestä kyseessä on tavoitepohjainen koe, mikäli se on laadittu käyttäytymistavoitteiden pohjalta. Tällaiset kokeet eivät kuitenkaan takaa, että tyydyttävä kriteeritulkinta on mahdollinen. Tavoitepohjaisuudesta huolimatta tällaisten kokeiden lopullinen muoto riippuu suurelta osin osioiden kirjoittajien henkilökohtaisista ominaisuuksista. Ei ole esimerkiksi olemassa mitään syytä uskoa, että kaksi osionkirjoittajaryhmää, joille annetaan samat käyttäytymistavoitteet, laatisivat kokeet, jotka näyttäisivät samanlaisilta tai antaisivat vastaavanlaisia pistemääriä, jos samat koehenkilöt suorittaisivat molemmat kokeet. On olemassa useita tapoja laatia osioita tavoitteiden pohjalta. Osion muotoon, vaihtoehtojen valintaan, ärsykeaineiston valintaan, vaikeustasoon ym. seikkoihin liittyy liian suuri tulkintamahdollisuus.

Kokeen suorittajan pistemäärä riippuu paljolti testin laatijasta, ja näin ollen on vaikea tulkita sitä suhteessa absoluuttisiin suoritusstandardeihin. Esimerkkinä tästä vaikeudesta voidaan mainita projektissa National Assessment of Education Progress (NAEP) käytetyt tavoitepohjaiset kokeet. Vaikka projektissa työskennelleet tutkijat yrittivät ryhmittää osioita ymmärrettävyyden parantamiseksi, testitulosten mielekäs tulkinta edellyttää yksityisten koeosioiden tarkastelua. Kokeenlaadinnan pohjana käytetyt tavoitteet

eivät johda selkeään määrittelyyn kokonaisesta suoritusten luokasta, jolla olisi mieltä riippumatta mittauksessa käytetyistä osioista.

Millmanin (1974) mukaan olemme tässä tekemisissä merkittävän käytännöllisen ongelman kanssa. Tavoitteisiin pohjautuvien kokeiden pistemäärät eivät pysty optimaalisesti ilmoittamaan, mitä tutkittava tietää tai osaa tehdä. Toisaalta hyvin määriteltyjen koeosioalueiden tuottaminen, jota vaaditaan kriteerikokeissa, ei ole usein mahdollista tai on ainakin hyvin vaikeata ja aikaa vievää. Lisäksi on vaarana aluemäärittelyn kapeus, koska pyrittäessä tarkkaan kuvaan sisällöstä ja muodostaonluonnollisena pyrkimyksenä rajoittaa ja supistaa määriteltävän alueen laajuutta. Jäljempänä selostetaan mahdollisia kompromissiratkaisuja.

Hallintakokeet (mastery tests) on määritelty "kriteerireferenssikokeiksi, jotka pidetään opetuksen lopussa selvittämään pystyvätkö henkilöt suorittamaan kaikki opetusohjelmassa määritellyt tehtävät" (Cleary, 1971, 7). Bloomin (1968, 1971) "formatiivinen koe" vastaa yllä olevaa määritelmää ja hänen "summatiivinen kokeensa" vastaa perinteellistä koulusaavutuskoetta. Mikään edellä esitetyssä määritelmässä ei estä hallintakokeita olemasta joko kriteerikokeita tai kriteerierottelukokeita. Määrittely kuvaa pikemminkin kokeiden tehtävää kuin niiden luonnetta. Tavoiteoppimisessa ei ole, eikä ole väitettykään, pyrittävän edistämään ensisijaisesti uudenlaista kokeiden kehittämistä, vaikka Bloom suositteleekin käytettäväksi kriteerikokeita. Pikemminkin tavoiteoppimisen ensisijaisena tarkoituksena on rohkaista kasvattajia asettamaan tavoitteekseen sen, että käytännöllisesti katsoen kaikkien oppilaiden suoritukset vastaavat opetukselle asetettuja tavoitteita, ja tämän lisäksi pyritään antamaan neuvoja miten tällainen tavoite voidaan toteuttaa, Harris (1974, 99) on sitä mieltä, että ihanteellinen hallintakoe on kriteerikoe sillä rajoituksella, että koeosiot ovat "sekä käsitteellisesti homogeenisia siinä mielessä että yksityiset tehtävät ovat keskenään vaihdettavissa määriteltäessä hallintaa että vastauksellisesti homogeenisia, ts. on olemassa pysyvä ehdollinen todennäköisyys ratkaista tietty osio oikein edellyttäen että oppilas on ratkaissut toisen osion oikein".

Donlon (1974) on ehdottanut, että "on hylättävä yritys käyttää vain yhtä käsitettä, kriteeriviitteinen, koska se yksinkertaisesti ei voi kantaa niin suurta semanttista kuormaa". Hän ehdottaa 10 käsitettä: käyttäytymisviitteinen (behavior-referenced), päätösviitteinen (decision-referenced), jakautumaviitteinen (distribution-referenced), skaalaviitteinen (scale-referenced), vaatimustasoviitteinen (standard-referenced), käsittelyviit-

teinen ja neljä muuta jo aikaisemmin mainittua (kriteeri-, normi, osa-alue- ja universumiviitteinen). Donlonin esittämät käsite-erottelut ovat johdonmukaisia hänen teesinsä kanssa, jonka mukaan testin määrittelevä piirre on se tulkinta, joka sen pohjalla halutaan tehdä. Millmanin (1974) mukaan erilaiset tulkinnat vaativat kuitenkin vain toista kahdesta yleisestä testien laadintastrategiasta, jotka johtavat kriteerikokeeseen ja normikokeeseen.

On mielenkiintoista todeta, että äskettäin on Gray (1978) vertaillut kriteerimittaamista ja Piagetin teoriaa ja todennut tiettyjä yhtäläisyyksiä.

Carver (1974) on kiinnittänyt huomiota psykologiseen testaamisen ja opetukseen liittyvän mittaamisen eroihin. Psykologisessa testaamisessa ollaan Carverin mukaan ensisijaisesti kiinnostuneita henkilöiden välisistä eroista (between-individual differences). Opetuksen puitteissa tapahtuvassa mittaamisessa puolestaan halutaan erityisesti saada selville kunkin yksilön kehitys (within-individual growth). Edellistä Carver nimittää mittaamisen ja kokeiden psykometriseksi dimensioksi ja jälkimmäistä edumetriseksi dimensioksi. Psykometrisissä testeissä ollaan tyypillisesti kiinnostuneita ensisijaisesti lahjakkuudesta, kyvyistä (ability) kun taas edumetrisissa kokeissa ollaan kiinnostuneita suoritustasosta (competence). Koska edumetristen kokeiden tarkoituksena on mitata kunkin yksilön tiedon, taidon yms. kasvua, koeosioiden tulee herkästi paljastaa tällainen kasvu. Edumetrisen kokeen validiteetti testataan esittämällä se kahdessa eri tilanteessa, joiden välillä olisi tullut ilmetä kasvua, jonka mittaus paljastaa. Sen reliabiliteetti todetaan samalla tavalla ja kokeen luotettavuutta osoittaa se, että koe tuo yksilöiden sisäiset erot ilmi johdonmukaisesti kummallakin kerralla.

Carver toteaa, että mm. Cronbach ja Furby (1970) pitävät muutospistemääriä (gain scores) ongelmallisina ja harvoin todella hyödyllisinä. Carverin mielestä Cronbachin ja Furbyn kritiikki koskee kuitenkin vain psykometrisiä testejä. Sen sijaan kokeellista tutkimusta suorittavat tutkijat käyttävät luonnostaan muutospistemääriä tutkiessaan kokeellisen käsittelyn vaikutuksia. Koska opetuksessa on luonnostaan kysymys "käsittelystä", edumetrinen lähestymistapa on luontevampi kuin psykometrinen, koska ollaan kiinnostuneita yksilöiden kehityksestä eikä heidän välisistään eroista. Carverin mukaan saattavat erilaisten käsittelyjen vähäiset vaikutuserot oppimistuloksissa osittain selittyä sillä, että on käytetty pääasiassa lahjakkuuseroille sensitiivisiä (älykkyyskoetyyppisiä) psykomet-

risiä testejä edumetristen kokeiden sijasta. Jos olisi käytetty edumetrisiä kokeita, käsittelyerot olisivat Carverin mukaan saattaneet osoittautua suuremmiksi.

2.3. Normimittaamisen ominaisuuksia

2.3.1. Normimittaamisen käsitteestä

Kriteeri- ja normimittaustuloksia verrataan usein keskenään. Normimittauksessa pistemääriä arvioidaan suhteuttamalla ne usein jonkin ulkopuolisen viiteryhmän saavuttamiin tuloksiin. On vaikeata, ellei mahdotonta, tulkita tyypillisen standardikokeen raakapistemäärää siten, että se osoittaisi oppilaan tietoja tai taitoja. Haluttuja oppimistuloksia määritellään harvoin suorituksina ennen kokeen laatimista, ja tästä syystä pistemäärä on mielekäs vain verrattaessa sitä muiden, saman kokeen suorittaneiden pistemääriin. Tätä tarkoitusta varten käytetään normitaulukoita ja tästä johtuu nimitys normikokeet. Koska minkä tahansa kokeen pistemääriä voidaan verrata jonkun ulkopuolisen ryhmän saavuttamiin tuloksiin, mikä tahansa koe voi tuottaa normitulkintoja.

On tehtävä selvä ero kriteerikokeen ja erottelukokeen välillä. Kriteerikokeet sallivat tarkkojen tulkintojen suorittamisen ja erottelukokeet maksimoivat mahdollisuutta suorittaa erotteluja. Välimaastossa ovat kriteeri-erottelukokeet. Ne ovat erottelukokeita, jotka mahdollistavat kriteeritulkinnan.

Normikokeilla 1. erottelukokeilla (differential assessment devices, DAD, Millman 1974, 316) on paljon pidempi psykometrinen historia kuin kriteerikokeilla, ja niitä käytetään useimmiten mittaamisessa. Ne on laadittu mittaamaan yksilöiden ja ryhmien välisiä eroja ja antamaan normitulkintoja. Kriteeri-erottelukokeet ovat tätä tyyppiä, paitsi että niiden osioiden tulee liittyä tarkasti määriteltyyn tavoitteeseen tai taitoon, minkä vuoksi kriteeri-erottelukokeet myöskin sallivat kriteeritulkinnan (differential assessment devices capable of criterion-reference inferences, CRDAD). Jotkut yksilöiden välisiä eroja mittaavat kokeet sallivat lähes minkä tahansa osion sisältyvän kokeeseen, mikäli se lisää ennustettavuutta tai valinnan tehokkuutta. Kriteerierottelukokeen osioiden täytyy kuiten-

kin olla sopusoinnussa tavoitteiden kanssa, joten on ainakin jonkin verran mieltä sanoa, että tällaisen kokeen pistemäärä on mielekäs riippumatta muiden kokeen suorittajien suorituksesta. Kriteerierottelukoe poikkeaa kuitenkin selvästi kriteerikokeesta. Edelliseen valittujen osioiden odotetaan tuottavan mahdollisimman korkeita perinteellisiä reliabiliteetti-indeksejä ja erottelevan mahdollisimman tehokkaasti ryhmiä toisistaan. Kokeesta poistetaan osiot, jotka kaikki oppilaat osaavat vastata oikein, koska ne eivät lisää sen erottelukykyä. Millmanin mukaan ei ole mielekästä muuttaa kriteerierottelukokeen raakapistemäärää arvioksi tutkittavan suoritustasosta, koska jotkut osioaltaan osiot (= huonosti erottelevat) suljetaan tarkoituksella pois kokeesta. Ei ole perusteltua odottaa, että kriteerierottelukokeeseen jäävät osiot ovat joko satunnainen tai edustava otos määritellystä suoritustehtäväluokasta. Kriteerierottelukokeen mittaama kyky riippuu vertailtavien ryhmien valinnasta ja siitä mitkä osiot satuttiin laatimaan ja sijoittamaan alkuperäiseen osioaltaaseen. Kriteerikokeen, kriteerierottelukokeen ja erottelu- 1. normikokeen eroja käsitellään tiivistetysti taulukossa 1.

Jotkut testiasiantuntijat ovat väittäneet, että kriteerikokeiden laatimismetodologiassa ei ole mitään uutta. On helppo ymmärtää tätä näkemystä, erityisesti jos kriteerikokeet samaistetaan kriteerierottelukokeisiin. Koska kriteerierottelukoe on yksilöiden ja ryhmien eroja havainnollistava mittari, jossa on osioita tietyltä määritellyltä tavoite- tai taitoalueelta, ei ole yllättävää että kriteerierottelukokeen laadintaan soveltuvat ne testikonstruktion ja evaluaation menetelmät, jotka ovat niin hyvin palvelleet testaajia aikaisemmin. Tällä hetkellä useimmat testausekspertit turvautuvat kriteerierottelukokeeseen mitatessaan henkilön suoritustasoa. Huolimatta sisällön standardikokeen käsitteestään esim. Ebel (1973, 278) esittää perinteellistä näkemystä, jonka mukaan "jokaisen kokeen ja testin tehtävänä on mitata. Tämä merkitsee, että sen täytyy erotella ne joilla on tietty kyky niistä joilta se puuttuu. Sen täytyy erotella ne joilla on enemmän tätä kykyä niistä joilla on vähemmän.... Testin reliabiliteetti ja osion erottelukyky ovat täysin yhtä tärkeitä sekä kriteeri- että normikokeille". Millmanin mielestä on kuitenkin kyseenalaista edellyttääkö henkilön kyvyn tarkka mittaaminen erottelukykyä. Yksi täysin sallittava tapa kuvata henkilön kykyä tietyn tehtäväluokan suhteen on esittää hänelle ratkaistavaksi satunnainen otos näistä tehtävistä. Tällaisen kokeen osioiden ei tarvitse erotella, koska kokeen suorittajat voisivat todella olla täysin samanvertaisia mitattavalla alueella. Lisäksi on huomattava, että monien tärkeiden

kasvatuksellisten sovellusten kannalta henkilön kykyjen kuvaus kriteerityyliin on asianmukaisempaa kuin erottelutyyliin.

2.3.2. Normimittaamisen rajoituksia

Yhdysvalloissa on käytetty normitettuja standardisoituja kokeita useisiin tarkoituksiin viime vuosikymmenien aikana. Pophamin mukaan (Popham 1978, 74-78) niitä on kuitenkin viime aikoina voimakkaasti arvosteltu. Arvostelu on osittain oikeutettua mutta osittain vailla perusteita. Arvostelu on oikeutettua sikäli, että normitettuja standardikokeit On kritiikittömästi käytetty mitä erilaisempiin tarkoituksiin, vaikka niiden varsinainen oikea käyttötapa on yksityisten oppilaiden suoritustason ilmaiseminen edustavan oppilasjoukon suoritustasoon verrattuna. Kun on kyseessä valintatilanne, jossa vain osa hakijoista voidaan ottaa johonkin oppilaitokseen, normiviitteiset standardikokeet ovat täysin paikallaan. Normikokeilla on kuitenkin selviä rajoituksia, joita käsitellään tarkemmin seuraavassa.

Normikokeita käytetään usein, paitsi oppilaiden suorituserojen selville saamiseen, myös opetuksen tulosten arvioimiseen. Tähän tarkoitukseen normikokeet eivät kuitenkaan kovin hyvin sovellu, koska ne eivät kerro selvästi, mitä oppilaan suoritustaso itse asiassa merkitsee. Normikokeita ei myöskään voida käyttää apuna opetuksen suunnittelemiseen, koska niiden avulla ei voi saada tarkkaa diagnostista tietoa oppilaiden asianhallinnan hyvistä ja heikoista puolista. Normikokeiden avulla ei voi myöskään saada tietoa koko koululaitoksen toiminnan tuloksellisuudesta ja tehokkuudesta. Kun koululaitokseen käytetään vuosittain varsin suuria rahasummia, seka päätöksentekijät että veronmaksajat haluavat entistä ponnekkaammin positiivista näyttöä siitä, miten koulu suorittaa tehtävänsä. Pulmana Pophamin mukaan on, että yleensä koulutuksen laatua arvioidaan väärän todistusaineiston perusteella. Yhdysvaltojen tapauksessa tämä virhepäätelmä johtuu siitä, että käytetään epätarkoituksenmukaista todistusaineistoa, ts. standardisoitujen normikokeiden tuloksia.

Pophamin mukaan (Popham 78-85) voidaan esittää ainakin neljä syytä, minkä vuoksi normikokeet eivät ole asianmukaisia koululaitoksen tuotoksia arvioitaessa. Ensinnäkin on olemassa vaara, ettei kokeen sisältö todellisuudessa vastaa annettua opetusta. Tällöinhän koe antaa virheellisen kuvan opetuksen tehokkuudesta, Toinen puute on se, että normikokeista

ei hevillä saa vihjeitä opetuksen suunnittelmiseen eikä parantamiseen. Kolmas normikokeen heikkous on luonteeltaan tekninen. Hyvän normikokeen täytyy erotella oppilaat tehokkaasti. Koetuloksissa tulee esiintyä hajontaa: mitä enemmän sen parempi. Tehokkaasti erottelevat oppilaita sellaiset osiot, joiden ratkaisuprosentti on viidenkymmenen prosentin tienoilla. Liian helpot tai liian vaikeat osiot karsitaan kokeesta pois heikon erottelukyvyn vuoksi riippumatta siitä mitä sisältöalueita ne mittaavat. Tällöin saattaa olla tuloksena, että kokeesta karsitaan pois niitä alueita mittaavat osiot, joita opettajat ovat pitäneet tärkeinä ja opettaneet huolellisesti. Onkin vaarana, että normikokeilla mitataan enemmänkin sitä mitä kouluissa ei ole opetettu kuin mitä siellä on opetettu. Normikoe saattaa siten muistuttaa lahjakkuustestiä enemmän kuin koulusaavutuskoetta. Neljäntenä puutteena, joka ei ole yhtä väistämätön kuin edellä mainitut puutteet, on mahdollinen erilaisten ryhmien asiaton suosiminen tai syrjiminen (cultural bias) erityisesti kokeen kielen ja sisältöjen suhteen.

2.4. Kriteerimittaamisen ja normimittaamisen vertailua

Gronlundin (1977) mukaan normi- ja kriteerikokeiden kesken on enemmän yhtäläisyyksiä kuin eroavaisuuksia. Gronlundin mielestä erot ovatkin lähinnä painotuseroja, kuten seuraavasta vertailusta käy ilmi:

1. Molemmissa koetyypeissä on tarpeen, että tavoitteet eli aiotut oppimistulokset on määritelty ennakolta kokeenlaadinnan pohjaksi.
   - NRK: tavoitteet on saatettu ilmoittaa melko yleisesti tai melko yksityiskohtaisesti
   - KRK: tavoitteet on yleensä ilmaistu hyvin tarkasti ja yksityiskohtaisesti
2. Molempien koetyyppien avulla pyritään mittaamaan edustavaa otosta tavoitteena olevista oppimistuloloksista (käyttäen apuna esim. koesisällön tarkennustaulukkoa).
   - NRK: koe kattaa tavallisesti laajan sisältöalueen, jolloin vain muutama tehtävä mittaa kulloinkin tiettyä oppimistulosta
   - KRK: koe kattaa tavallisesti rajoitetun osa-alueen ja kutakin oppimistulosta mittaa usea osio

3. Molemmissa koetyypeissä käytetään monenlaisia koetyyppejä.

   NRK: koetyyppinä on usein monivalintakoe

   KRK: valintakokeen asema ei ole yhtä keskeinen kuin normikokeissa

4. Molemmissa koetyypeissä sovelletaan samantapaisia periaatteita osioita laadittaessa.

   NRK: korostetaan osioiden kykyä erotella oppilaita

   KRK: korostetaan osioiden kykyä kuvata oppilaan suoriutumista edustavassa osiojoukossa

5. Molemmissa koetyypeissä on kiinnitettävä huomiota koetulosten luotettavuuteen.

   NRK: perinteelliset reliabiliteetin arviointimenetelmät ovat paikallaan, koska koesuoritusten vaihtelu on tuntuva

   KRK: perinteelliset reliabiliteetin arviointimenetelmät eivät ole soveliaita, koska koesuoritusten hajonta voi olla vähäinen

7. Molemmat koetyypit laaditaan aina tiettyä käyttöä varten.

   NRK: käytetään erityisesti valintakokeina ja summatiivisessa arvioinnissa

   KRK: käytetään erityisesti mittaamaan valmiuksia ja edellytyksiä formatiivisessa ja diagnostisessa arvioinnissa

Gronlundin kuvaus kriteerikokeen käyttötavoista on yllättävän suppea, kuten edellä ja jäljempänä esitetystä käy ilmi.

Myös Millman (1974) on vertaillut kriteeri- ja normikokeiden ominaisuuksia. Hänen käsityksensä esitetään taulukossa 1, johon on lisätty normi- 1. erottelukokeita käsittelevä sarake.

TAULUKKO 1. Kriteeri- ja normikokeiden ominaisuuksien vertailu (Millmania 1974 mukaillen)

OMINAISUUS	KRITEERIKOE (CRT)	KRITEERIEROTTELUKOE (CRDAD)	NORMI. L. EROTTELUKOE (DAD)
OSIOIDEN SISÄLLÖN RAJOITUKSET	LAADITAAN ALUEEN MÄÄRITTELYN POHJALTA: SISÄLLÖN RAJOJEN MAHDOLLISIMMAN TARKKA MÄÄRITTELY	"SOPEUTETAAN" OPETUSTAVOITTEI-SIIN: SISÄLLÖN RAJAT VAIN OSITTAIN MÄÄRITELTY	SISÄLLÖN RAJAT MELKO VÄLJÄT
OSIOIDEN VALINTA	SATUNNAISOTOS OSIOALTAASTA	VALITAAN EMPIIRISESTI SITEN, ETTÄ MAKSIMOITAISIIN EROTTE-LUKYKY	VALITAAN EMPIIRISESTI, SITEN ETTÄ MAKSIMOITAISIIN EROTTE-LUKYKY
ONKO OSIO- JA KOEVARI-ANSSI VÄLTTÄMÄTÖNTÄ?	EI OLE	ON	ON EHDOTTOMASTI
RELIABILITEETTI-TYYPPI	OSAAMISTASOARVIOINTIEN JOHDONMUKAISUUS	TAVANOMAISET MENETTELYT	TAVANOMAISET MENETTELYT
ENSISIJAINEN VALIDI-TEETTITYYPPI	SISÄLLÖN VALIDITEETTI	ULKOPUOLINEN KRITEERI	ULKOPUOLINEN KRITEERI
KRITEERIRYHMIEN LUONNE	EI KÄYTETÄ KOKEEN LAATIMISESSA	RYHMIÄ KÄSITELLÄÄN ERI TA-VOILLA TAI YKSILÖIDEN USKO-TAAN EROAVAN ARVIOITAVAN OMINAISUUDEN SUHTEEN	RYHMIÄ KÄSITELLÄÄN ERI TA-VOILLA TAI YKSILÖIDEN USKO-TAAN EROAVAN ARVIOITAVAN OMINAISUUDEN SUHTEEN
ASIANMUKAISIIN JOHTO-PÄÄTÖS KOSKEE	TUTKITTAVIEN OSAAMISTASOA	TUTKITTAVIEN KYKYJEN (SUORI-TUSTEN JNE.) EROJA	TUTKITTAVIEN KYKYJEN (SUORI-TUSTEN JNE.) EROJA

2.5. Kriteerimittaamisen ja normimittaamisen käyttötapojen vertailua

Kriteerikoe on suositeltava, kun halutaan arvioida oppilaiden tasoa suhteessa suoritustehtäviin. Sitä vastoin kriteerierottelukoe näyttää parhaiten sopivan oppilaiden vertailemiseen, kun pyritään määrittämään heidän keskinäistä suoritustasoaan tai ennustamaan mihin ryhmään oppilaat kuuluvat. Yleensä sama koe ei voi toisaalta antaa pistemääriä, jotka on tulkittavissa suoraan suhteessa suoritustehtäväjoukkoon (ts. tulkittavissa aluepistemäärän estimaattina), Ja toisaalta olla optimaalisen käyttökelpoinen erotteluarviointien tekemiseen. Määritellystä osiouniversumista satunnaisesti valitut osiot eivät välttämättä ole samoja osioita, jotka ovat optimaalisia vertailtavien ryhmien erottelemiseksi.

Seuraavassa esitetään kahdeksan esimerkkiä, joiden tarkoituksena on valaista kriteerikokeiden ja erottelukokeiden asianmukaista käyttöä.
Esimerkki 1: Kokeen tarkoituksena on maksimoida eroa, joka on ennen opetusta suoritetun mittauksen ja opetuksen jälkeen suoritetun mittauksen välillä. Esimerkin muotoilusta voitaisiin päätellä että erottelukoe on asianmukaisempi, koska kokeen käyttäjä haluaa koepistemäärien erottelevan toisistaan opetusta saaneet ryhmät niistä jotka eivät opetusta ole saaneet. Glaser (1963) kannatti maksimaalisia erotteluja aikaansaavia kokeita. Hän kritisoi standardikokeiden käyttöä tutkimuksissa, koska nämä sisältävät osioita, jotka eivät vastaa opetuksen tavoitteita eivätkä näin ollen todennäköisesti paljasta opetuksen vaikutusta. Kriteerikokeita voitaisiin puolestaan kritisoida, koska ne eivät tehokkaasti osoittaisi opetuksen vaikutuksia. Voidaan kuitenkin kysyä onko viisasta pyrkiä valitsemaan osioita, jotka maksimoivat opetuksen muutospistemääriä tai eroja. Tällaiset osiot mittaavat Millmanin (1974) mukaan todennäköisesti yleisä taitoja tai koskevat vain osaa opetuksen sisällöstä. Ulkopuolelle voivat jäädä ne tavoitteet, joita ei saavutettu, ne jotka osattiin jo ennen opetusta tai jotka ovat yhteisiä kahdelle erilaiselle käsittelylle. Millman kysyy onko kasvatuksessa tarkoituksenmukaista pyrkiä saamaan mäet näyttämään vuorilta? Kriteerikokeet, jotka voivat antaa realistisen kuvan oppilaiden kyvyistä tietyillä kiinnostavilla sisältöalueilla, antavat terveemmän pohjan opetusohjelmien korjaamiselle kuin on mahdollista, jos käytetään erottelukokeita. Erottelukokeiden edellytykset opetusohjelmien evaluoinnissa voivat parantua, mikäli edellytetään että koe sisältää kaikkia opetuksen tavoitteita mittaavia osioita, vaikka niillä saattaa olla vähäinen erottelukyky.

Millman kysyy miksi ei kuitenkaan voitaisi jatkaa pysähtymättä puolitiehen ja määritellä, mitä oppilaiden käyttäytymisen luokkia tulee sisällyttää kokeeseen ja valita satunnainen otos näistä alueista ts. käytettäisi kriteerikoetta.

Esimerkki 2: On arvioitava osaako oppilas ratkaista sanallisia tehtäviä, jotka käsittelevät voitto- tai tappioprosenttia. Näyttää siltä että kriteerikoe täyttäisi tehokkaammin tämän tarkoituksen, koska ei edellytetä, että tulisi suorittaa erotteluja oppilaiden kesken. Miksi kasvattaja haluaisi määritellä opiskelijan suoritustason? Eräs mahdollisuus on, että hän haluaa tietää suoritustason päättääkseen tarvitaanko tukitoimenpiteitä. Tätä tarkoitusta varten on annettava etusija kriteerikokeelle, joka sisältää osioita huolellisesti määritellyltä käyttäytymisalueelta. Se antaisi kasvattajalle tietoa oppilaan kyvystä ratkaista edustava otos kyseisiä tehtäviä ja täten antaisi tietoa siitä tarvitaanko lisäopetusta.

Esimerkki 3: Tehtävänä on valita ne oppilaat, jotka voivat menestyksellisesti seurata seuraavaan oppimisyksikköön sisältyvää opetusta. Tarvitaan koe, joka pystyy määrittelemään esim. pystyykö kaksi eri opiskelijaryhmää menestyksellisesti seuraamaan seuraavan yksikön opetusta vai ei. Mittauksen tarkoituksena ei ole kuvata, kuinka hyvin opiskelijat pystyvät ratkaisemaan samantapaisia tehtäviä kuin koe sisältää, vaan erottelemaan ryhmä oppilaita toisista. Tätä tarkoitusta varten Millman suosittelee erottelukoetta. Tämänlaisen kokeen laatimiseksi tarvitaan kriteeriryhmiä. Kyseessä olevassa esimerkissä on pystyttävä identioimaan kaksi ryhmää, joiden kyky onnistua seuraavassa opiskeluyksikössä eroaa toisistaan.

Esimerkki 4: Arvosanojen antaminen peruskoulussa tai ylioppilaskokeessa. Erottelukoe on todennäköisempi, koska kummassakin toimitaan normaalijakautuman pohjalla. Koska yksityisen oppilaan saama arvosana riippuu hänen menestymisestään muihin oppilaisiin verrattuna, tarvitaan koe joka oppiaineksen puitteissa erottelee oppilaita tehokkaasti. Erottelukoetta käytetään tyypillisesti silloin, kun tietylle osalle oppilaista tulisi antaa tietty arvosana. Tärkeä kysymys on: erottelevatko osiot tehokkaasti. Osioita karsitaan niiden heikon erottelukyvyn vuoksi tai ne yritetään ennakkoon saada mahdollisimman erotteleviksi.

Esimerkki 5: Selvittää kuinka hyvin oppilas osaa muodostaa englannin kielen kysymyslauseita. Kriteerikoe on sopivampi, koska on kyseessä oppilaan suoritustason määrittäminen tietyllä sisältöalueella. Keskeinen kysymys ei ole erottelevatko osiot tehokkaasti vaan ovatko osiot sellaisia, että ne kuuluvat oleellisesti mitattavan sisältöalueen piiriin.

Esimerkki 6: Suositella oppilaille tasokursseja peruskoulussa. Erottelukoe on todennäköisempi, koska kokeen tulisi nykyisten hallinnollisten jatko-opintokelpoisuusrajoitusten vuoksi erotella erityisen selvästi ne oppilaat, joilla ei ole edellytyksiä kuin suppeimpien kurssien käymiseen.

Esimerkki 7: Koulun laadullisen tuotoksen varmistaminen vähimmäisvaatimusten tai perustavoitteiden saavuttamisen kontrolloinnin avulla. Kriteerikoe on paikallaan, koska on selvästi kyseessä tavoitteiden saavuttamisen arviointi eikä oppilaiden valikointi tai ohjailu tai suhteellinen arvostelu.

Esimerkki 8: Tukiopetuksen tarpeen määrittely. Kriteerikoe on paikallaan, koska tukiopetuksen ollakseen tuloksellista täytyy kohdistua juuri oppimisen heikkoihin kohtiin. Siksi kokeen sisällön tulee olla edustava otos tietystä tarkasti spesifioidusta sisältöalueesta.

3. KESKEISIÄ VAIHEITA KRITEERIMITTAREIDEN LAADINNASSA

3.1. Alkukatsaus

Glaser esitti jo vuonna 1963, että siinä missä normimittaaminen pyrkii saamaan selville oppilaan suhteellisen aseman (relative status), kriteerimittaaminen pyrkii selvittämään oppilaan absoluuttisen aseman (absolute status). Kriteeriviitteisiä kokeita kehitettäessä on Pophamin mukaan (Popham 1978, 89-111) kiinnitettävä erityistä huomiota oppilaiden suorituksen kuvaamiseen, osioiden kirjoittamiseen, osioiden parantamiseen ja tarvittavan osiomäärän määrittämiseen.

Tärkein ero normi- ja kriteerikokeiden välillä koskee koetulosten antamaa kuvaustiedon laatua. Normikoe kuvaa sitä, miten oppilaiden suoritukset hajoavat ja kuka suoriutuu paremmin tai heikommin kuin joku muu. Normimittauksessa validiteetin arvioimiseen on käytetty sekä sisällön validiteettia että erityisesti ulkopuolista kriteeriä. Kriteerimittaamisessa sisällön validiteetti on ylivoimaisesti tärkein validiteetin muoto.

Tavoitteena on kuvata mahdollisimman selkeästi sitä, mitä oppilaan koesuoritus tarkoittaa,

Kriteerimittaamiseen liittyy läheisesti kysymys suoritustasolle asetetuista vaatimuksista. Näitä asioita ei kuitenkaan pidä sekoittaa keskenään. Erityisesti tulee varoa pitämästä kriteerikokeena jokaista koetta, jossa asetetaan jokin tietty vaatimustaso (esim. 85 % oikein tehtävistä) hyväksymisen ehdoksi. Kriteeriviitteisinä kokeina ei voi myöskään automaattisesti pitää ns. tavoitepohjaisia kokeita, jotka laaditaan käyttäytymistavoitteiden pohjalta. Tavoitepohjaista koetta voidaankin Pophamin mukaan pitää lähinnä kriteerikoeperheen varsin heikkona jäsenenä.

Mitä sitten tarkoitetaankaan kriteeriviitteisellä mittauksella? Pophamin mukaan kriteeriviitteistä koetta käytetään määrittelemään yksilön asema hyvin määritellyn käyttäytymisalueen suhteen. Jotkut kriteerimittaamisen kehittelijöistä (mm. Hively) kutsuvatkin tällaista mittaamista alueviitteiseksi tai osa-alueviitteiseksi mittaamiseksi (domain-referenced testing, DRT), koska siinä on keskeistä tietyn käyttäytymisalueen (domain, behavioral domain) tarkka määritteleminen. Pophamin mukaan mittaamisesta voitaisiin käyttää tätä nimitystä, mutta saattaa olla kuitenkin tarkoituksenmukaista pitäytyä jo varsin yleiseen tietoisuuteen tulleeseen nimitykseen (= kriteerimittaaminen). Myös Hambleton et al. (1978) ovat tästä asiasta samaa mieltä.

Tärkeintä kriteerimittaamisessa on kokeen kuvaus (descriptive scheme). Kokeen kuvauksen tehtävänä on antaa mahdollisimman seikkaperäinen ja valaiseva käsitys siitä, mitä oppilaan suoritus merkitsee. Kokeen kuvauksen tehtävänä on kommunikoida kokeen käyttäjille, mitä koe mittaa. Sen tehtävänä on myös kommunikoida osioiden laatijoille, millaisia osioita voidaan liittää kriteerikokeeseen. Tarkkaa käyttäytymisalueen määrittelemistä Popham kutsuu koenormistoksi tai koetäsmennykseksi (test specifications). Riittävän yksityiskohtainen koetäsmennys ei ole mahdollista tyypillisen lyhyen käyttäytymistavoitelauseen pohjalta. Käyttäytymisalueen huolellinen määritteleminen edellyttää tuntuvasti seikkaperäisempää ja laveampaa kuvausta.

Koetäsmennyksen jälkeen on tehtävänä laatia koeosiot. Mikäli koetäsmennys on tehty huolellisesti, osioiden laadinta sujuu varsin kitkattomasti. Tavoitealueiden osioiden tulee olla tietyssä määrin homogeenisia siinä mielessä, että ne voidaan katsoa johdetun loogisesti käyttäytymisalueen koetäsmennyksestä. Koeosioiden tulee olla sopusoinnussa koetäsmennyksen kanssa, mutta tämä ei Pophamin mukaan tarkoita sitä, että oppilaiden

pitäisi osata vastata johdonmukaisesti jokaiseen osioon joko oikein tai väärin. Täydellisen homogeenisuuden vaatimus johtaisi suppeiden käyttäytymisalueiden koetäsmennysten pohjattomaan suohon. Käytännössä kannattaa menetellä siten, että määritellään koetäsmennys siten, että se sallii eri vaikeustasoa olevien osioiden laatimisen. Jälkikäteen voidaan sitten arvioida, missä määrin osiot on johdettu homogeenisesti koetäsmennyksestä. Osioiden laatijat voivat käytännön työssään *myös* havaita puutteita koetäsmennyksessä ja antaa vihjeitä sen parantamiseksi. Näin voi olla hyödyllistä vuorovaikutusta koetäsmennyksen laatijoiden ja osioiden kirjoittajien välillä.

Kuinka monta osiota tulisi laatia mittaamaan kutakin huolellisesti määriteltyä käyttäytymisaluetta? Kysymykseen ei Pophamin mukaan voida antaa mitään täysin yksiselitteistä vastausta. Osioiden määrä riippuu osittain siitä, miten tärkeätä on välttää tekemästä virheellisiä johtopäätöksiä koetuloksen perusteella. Tärkein vältettävä virhelähde on liian vähäinen osiomäärä. Tämän hetkisten kokemusten mukaan kutakin käyttäytymisaluetta mittaamaan tulisi Pophamin mukaan laatia noin 10-20 osiota. Kymmentä osiota voidaan yleensä pitää yleisenä miniminä. Osiomäärää tulisi tästä lisätä, jos joudutaan tekemään tärkeitä ratkaisuja, ja osiomäärää voidaan vähentää mikäli päätöksillä ei ole yhtä ratkaisevaa merkitystä.

Miten sitten voidaan parantaa heikoksi osoittautuneita koeosioita? Tässä voidaan Pophamin mukaan käyttää seka ennakkotarkistusta että esikokeilun antamia empiirisiä tuloksia. Empiirisessä analyysissä on perinteellinen erotteluindeksi edelleenkin käyttökelpoinen. Toinen mahdollinen menetelmä on laskea alkukokeen ja loppukokeen tuloksissa esiintyvä eroindeksi vähentämällä loppukokeen ratkaisuprosentista alkukokeen ratkaisuprosentin. Kolmas mahdollisuus on vertailla tavoitetason saavuttaneiden tuloksia sellaisten oppilaiden tuloksiin, jotka eivät ole saavuttaneet vaadittua tasoa. Tämän lisäksi on mahdollista käyttää bayesilaista tai Raschin mallia.

Pophamin mukaan on kuitenkin todennäköistä, että asiantuntijoiden ennakkotarkistus tulee vastaisuudessa näyttelemään yhä huomattavampaa osaa. Tällöin on huolehdittava siitä että arviot tehdään systemaattisesti. Arvioitsijoita pyydetään päättelemään, missä määrin kukin osio on sopusoinnussa koetäsmennyksen kanssa. Arvioitsijoiden johdonmukaisuutta voidaan empiirisesti testata siten, että arvioitavien osioiden joukkoon sijoitetaan tahallaan kaksi tai kolme asiaan kuulumatonta osiota. Tämä tehdään luonnollisesti siten, etteivät arvioitsijat ole tietoisia tästä. Näin mene-

tellen voidaan Rovinellin ja Hambletonin (1976) mukaan saada varsin johdonmukaisia ja yksimielisiä arvioita koeosioiden validiteetista. On huomattava, että empiiriset tulokset saattavat osoittaa, että osiossa näyttäisi olevan jotakin vikaa, mutta asiantuntijat eivät pysty löytämään kuitenkaan osiosta mitään virhettä. Tällöin on perusteltua pitää asiantuntijoiden arviota ratkaisevana kriteerinä. Joko osio on johdonmukainen testitasmennyksen kanssa tai se ei ole. Empiiriset indeksit eivät voi tätä tosiasiaa muuttaa miksikään.

3.2. Mitattavan osa-alueen määrittäminen

3.2.1. Yleisiä näkökohtia

Seuraavassa käsitellään kokeita, joiden tarkoituksena on kuvata oppilaan tämänhetkistä asemaa suhteessa huolellisesti määriteltyihin suoritustehtäviin, joita kutsutaan alueeksi (domain). Tältä alueelta valittua satunnaista tai ositettua satunnaisotosta osioita kutsutaan kriteerikokeeksi. Seuraavassa käsitellään erikseen osiopopulaation määrittelemistä, koeosioiden valintaa, hyväksyttävän pistemäärän määrittelemisestä, aluepistemäärän arviointia, kokeen pituuden määrittelemistä sekä kriteerikokeen arviointia.

Seuraavassa esityksessä lähdetään siitä olettamuksesta, että on olemassa ainakin yleinen määrittely mitattavista tiedoista, taidoista tai asenteista. Tässä yhteydessä ei ole tärkeätä korostetaanko transfertehtäviä, hyödyllistä tietoa, affektiivisia tavoitteita jne. Perinteelliset osionlaadintatavat käyttävät koesuunnitelmaa, joka usein luettelee oppiaineksen riviotsikoina ja sellaiset käsitteet kuin tieto ja soveltaminen sarakeotsikkoina. Näin syntyneet ruudukot täytetään numeroilla, jotka osoittavat kuinka monta osiota laaditaan mittaamaan tiettyä kognitiivista taitoa ja käytettäessä tiettyä sisältökohtaa. Tällaista koetäsmennystä (table of specifications, master chart, matrix of content and behaviors) ovat suositelleet ja käyttäneet mm. Tyler (Smith & Tyler 1942) ja Bloom (Bloom et al. 1971). Koesuunnitelma voi esimerkiksi määritellä, että kokeeseen tulee kaksi Shakespearen näytelmiä koskevaa osiota, jotka mittaavat oppilaan kykyä ymmärtää pääasioita. "Nämä käsitteet (esim. ymmärtäminen) viittaavat henkisiin prosesseihin, ei havaittaviin tapahtumiin. Kun

kokeenlaatija valitsee tällaisen käsitteen, hän käyttää sitä viittaamaan johonkin, mikä tapahtuu vain hänen omassa henkisessä toiminnassaan" (Bormuth, 1970). Tällaiset testisuunnitelmat antavatkin vain vähän ohjetta osion kirjoittajalle ja siksi hänen tuottamansa spesifit osiot heijastavat suurelta osin hänen omaa kokemustaustaansa ja kirjoitustyyliään. Kaksi osionlaatijaa, joille annettaisiin sama testisuunnitelma ja jotka työskentelisivät toisistaan riippumatta, tuottaisivat kokeita, jotka todennäköisesti korreloisivat vain kohtalaisesti keskenään ja joilla olisi erilainen vaikeustaso.

Kriteerikokeiden laadinta alkaa yleensä käyttäytymistavoitteiden määrittelystä. Hyvät tavoitteet eivät ole kuitenkaan osiospesifejä, mikä merkitsee että on yleensä olemassa lukuisa joukko mahdollisia osioita, jotka liittyvät tiettyyn tavoitteeseen. Hyviäkin tavoitteita voidaan mitata lukuisalla määrällä oppilaiden käyttäytymisen muotoja. Alueen määrittämiseen tarvitaan enemmän kuin käyttäytymisen muotoina ilmaistut tavoitteet. Tavoitteena on sisäänrakentaa mielekkyys kokeeseen siten, että kun ilmoitetaan että oppilas on vastannut 95 % osiosta oikein, tämä pistemäärä voidaan tulkita kokonaisena joukkona tehtäviä, joissa oppilas on osoittanut suurta kyvykkyyttä. Tällainen tulkinta ei ole mahdollista, ellei osiopopulaatiota ole selvästi identifioitu.

Popham on todennut, että kriteerikokeen laatijalla on ongelmana saattaa tasapainoon selkeyden ja käytännöllisyyden kaksi kriteeriä. Vaikeana tehtävänä on laatia suunnitelma, jonka avulla eristetään tavoitteena olevien oppilaan käyttäytymismuotojen luokan tärkeät ulottuvuudet, jonka jälkeen kuvataan näitä ulottuvuuksia yksityiskohtaisesti niin, että kuvaus viestii selvästi eikä kuitenkaan ole niin pitkä, että kasvattajat välttävät sen käyttämistä. Popham toteaa, ettei tämä ole mikään vähäpätöinen temppu.

3.2.2. Kielitieteeseen perustuvat menetelmät

Esimerkkinä kielitieteeseen perustuvista järjestelmistä mainittakoon Bormuthin (1970) lähestymistapa, joka käyttää operationaalisia määritelmiä: "Näitä operaatioita pitäisi voida käyttää systemaattisesti opetusohjelmaan sillä tavalla, että tuotetaan kaikki sen tyyppiset osiot, jotka on johdettavissa näiden operaatioiden avulla. Kun joukko operaatioita täyttää tämän vaatimuksen, tämä ei ainoastansa takaa osiopopulaatioiden määriteltävyyttä vaan se myöskin takaa sen, että näiden operaatioiden avulla tehdyt kokeet

ovat itsenäisesti reprodusoitavia, kunhan vain tiedetään mitä operaatioita käytettiin johtamaan osiot ja niihin annettavat reaktiot" (s. 35).

Operaatioilla Bormuth tarkoittaa osioiden transformaatioita (item transformations). Olettakaamme että lause "Vanhempi sisar sammutti tulipalon" oli osa opetussekvenssiä. Seuraavassa taulukossa esitetään joitakin osiotransformaatioita, jotka ovat sovellettavissa tähän lauseeseen.

TAULUKKO 2. Osiotransformaatioita lauseesta "Vanhempi sisar sammutti tulipalon" Millmani (1974, 329) mukaillen

Transformaatio	Kielellinen ilmaisu
Väittämä sellaisenaan (ei muunnosta, vaan kyllä-ei väittämä)	Vanhempi sisar sammutti tulipalon.
Varmistuskysymys / liitekysymys	Vanhempi sisar sammutti tulipalon, eikö niin?
Vaihtoehtokysymys (joko /tai, kyllä / ei)	Sammuttiko vanhempi sisar tulipalon?
Hakukysymys	
- subjektin poisto	Kuka sammutti tulipalon?
- objektin "	Minkä vanhempi sisar sammutti?
- subjektin määritteen poisto	Kumpi sisarista sammutti tulipalon?

Bormuth ja Anderson (1972) antavat useita esimerkkejä transformaatioista, joita voidaan käyttää generoimaan kysymyksiä kirjoitetusta tekstistä. Bormuthin ansiona on tämän raportin kirjoittajan mielestä se, että hän on kiinnittänyt huomiota loogis-kielellisiin mahdollisuuksiin sisältökysymyksiä laadittaessa. Kysymyksiä laadittaessa ei tarvitse pitäytyä pelkästään intuitioon, vaan voidaan menetellä systemaattisesti, kuten ylläoleva esimerkki osoittaa.

3.2.3. Osiolomake

Osiolomake (item form) on toinen tekniikka, jota voidaan käyttää määrittelemään osioalueita hyvin tarkasti. Osiolomakkeella on Osburnin mukaan (Osburn 1968, 97) seuraavat ominaisuudet:

1. se tuottaa osioita, joilla on ennalta määrätty syntaktinen rakenne
2. se sisältää yhden tai useamman vaihdettavissa olevan elementin, ja
3. se määrittelee osiolauseiden luokan spesifioimalla sen joukon, josta voidaan valita jokin elementti korvaamaan vaihdettavia elementtejä.

Osiolomaketta havainnollistaa oheinen kuvio 1. Hively ym. (1973) esittävät myöskin yksityiskohtaisen selostuksen aluereferenssikokeiden laatimisesta ja käytöstä laajassa opetussuunnitelman kehittämisprojektissa (MINNEMAST).

Osiolomakkeiden käyttäminen sallii osiopopulaation tarkan kuvauksen osa-aluekokeessa. Osiolomakkeiden ei kuitenkaan tarvitse olla niin kompleksisia kuin esimerkki osoittaa. Tällaisen yksityiskohtaisen lomakkeen etuna on, että osiopopulaatio tunnetaan ja voidaan kuvata tarkasti. Lisäksi on mahdollista arvioida kuinka suuren osan sisältöalueesta oppilas osaa, mikäli hänelle esitetään satunnainen tai edustava otos näistä osioista. Tätä selvyyttä ei saada ilman kustannuksia, vaan joudutaan käyttämään paljon aikaa ja energiaa osa-alueen tarkkaan määrittelemiseen tai vastaavasti joudutaan supistamaan alueen rajoja. Edellä kuvattujen lähestymistapojen toteuttaminen vaatii joko paljon työtä tai se pakottaa osion kirjoittajan tyytymään hyvin homogeeniseen tehtäväjoukkoon. Lausetransformaatiot ja osiolomakkeet eivät välttämättä kata tyydyttävästi erilaisia tehtäviä, joita osionkirjoittaja haluaisi käyttää mittaamaan tietoa, taitoa tai asennetta. Durnin ja Scandura (1973) ovatkin kritisoineet osiolomakkeen käyttöä ja ehdottavat tilalle algoritmista lähestymistapaa, joka perustuu sääntöjen ja prosessien intuitiiviseen analyysiin. Millmanin mielestä algoritminen lähestymistapa näyttäisi sopivan parhaiten matematiikan alalla. Heidän kritiikkinsä ei näytä kohdistuvan niinkään osiolomakkeen käsitteeseen kuin tapaan jolla tätä tekniikkaa toteutetaan.

Osiolomakementelmällä on eräitä ilmeisiä etuja:

1) Se tuottaa (generoi) osioita, joilla on tietty syntaktinen rakenne.
2) Se sisältää yhden tai useampia vaihdeltavia elementtejä.
3) Se määrittelee joukon osiolauseita määrittelemällä substituutiojoukot kohdassa 2 mainituille vaihteleville elementeille.

KUVIO 1. Osiolomakkeen malli (Hively et ai. 1973, hieman mukaillen)

OSIOLOMAKE 16.14

Oppilaan tehtävänä on verrata kahta esinettä vaa'an avulla ja valita symboli, joka täydentää oikealla tavalla painosuhdetta koskevan väittämän.

Yleiskuvaus:

oppilasta pyydetään vertailemaan kahden esineen painoa, jota (1) ei ehkä voida todeta pelkästään käsin punnitsemalla, (2) ei ehkä voida erottaa toisistaan edes vaa'an avulla. Kussakin tilanteessa esineiden koko vaihtelee irrelevanttina piirteenä. Tasavartinen vaaka on käytettävissä mutta oppilasta ei määrätä käyttämään sitä. Oppilasta kehoitetaan valitsemaan yksi kolmesta symbolista (>, <, =) ja sijoittamaan se tyhjään tilaan kahden painosymbolin väliin.

Ärsyke- ja reaktio-ominaisuudet:

Vakiot kaikissa ruuduissa. Tasavartinen vaaka (Tinkertoys). Vertailtavat esineet ovat läpikuultavia, sylinterin muotoisia pulloja, jotka ovat samanlaisia lukuunottamatta painoa (joko 23 tai 75 gr) ja kokoa (joko 2"x5/8" tai 2 1/2 x 1 3/4 tuumaa). Kumpaakin esinettä kuvataan satunnaisesti määräytyvällä pikkukirjaimella.

Oppilasta pyydetään täydentämään väittämä valitsemalla esineiden välistä painosuhdetta kuvaava symboli.

Ruutujen välinen erottelu
Kolme painosuhdetta (jotka voidaan saada selville vain vaa'an avulla, ei nostamalla tai tunnustelemalla) jotka määrittyvät sen perusteella missä asemassa ne ovat oppilaan edessä:

vasen > oikea; vasen < oikea; vasen=oikea

Kolme painosuhdetta:

vasen > oikea, vasen < oikea, vasen=oikea

Ruutumatriisi

Painosuhteet

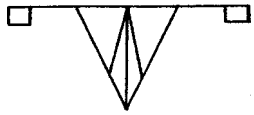
Kokosuhteet	Pvas>Poik	Pvas<Poik	Pvas=Poik
Kvas>Koik	(1)	(4)	(7)
Kvas<Koik	(2)	(5)	(8)
Kvas=Koik	(3)	(6)	(9)

Osiolomakekehikko

Aineisto Vipuvaaka 2 esinettä Ärsyke-reaktio lomake Lyijykynä	
Kokeenpitäjän ohjeet: Sijoita koeaineisto oppilaan eteen ylämainitussa järjestyksessä vaaka esineet Ä-r-lomake Oppilas	Koeteksti: Tässä on kaksi esinettä. Ne on merkitty kirjaimin. Vertaa esineiden painoa ja kirjoita jokin näistä kolmesta (näyttää) merkistä tähän tyhjään kohtaan (näyttää), jolloin saadaan täydellinen vertailua kuvaava lause. Voit halutessasi käyttää apuna vaakaa.

Vastauksen kirjaaminen

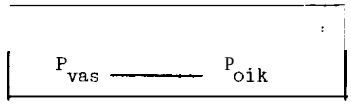
Liitä ärsyke-reaktiolomake tähän kohtaan. Kuvaa mitä oppilas teki. Jos oppilas käytti apuna vaakaa, merkitse esineiden kirjaintunnukset vaa'an kuppeihin ja merkitse missä asennossa pystysuora osoitin oli ratkaisuhetkellä.



Koeaineiston kuvaus:

Kynä.

Koko	23 gr	25 gr
Pieni	a	m k
Suuri	b	o n



vas ja oik tilalle sijoitetaan oikeat kirjainsymbolit substituutiotaulukosta.

Substituutiosuunnitelma (SS)

(vas,oik) esineet

Ruiitu 1:	(o,a)	
Ruutu 2:	(m,b)	
Ruutu 3:	valitse	SS 16:13:sta
Ruutu 4:	(b,m)	
Ruutu 5:	(a,o)	
Ruutu 6:	valitse	SS 16:14:sta
Ruutu 7:	valitse	SS 16:15:sta
Ruutu 8:	valitse	SS 16:16:sta
Ruutu 9:	valitse	SS 16:17:sta

Substituutioparit

SS 16.13	Järjestetyt parit (m,a);(o,b)
SS 16.14	" " (a,m);(b,o)
SS 16.15	" " (b,a);(o,m)
SS 16.16	" " (a,b);(m,o)
SS 16.17	" " (m,k);(o,n)

Pisteistysohjeet

Oikea vastaus syntyy, kun oppilas kirjoittaa oikean symbolin (> , < tai =) tyhjään tilaan ja täydentää näin vertailulauseen. Symbolin tulisi olla > ruuduissa 1,2 ja 3 < ruuduissa 4,5 ja 6 sekä = ruuduissa 7,8 ja 9.

KUVIO 1. Osiolomakkeen malli (Hively et al. 1973, hieman mukaillen)

OSIOLOMAKE 16.14

Oppilaan tehtävänä on verrata kahta esinettä vaa'an avulla ja valita symboli, joka täydentää oikealla tavalla painosuhdetta koskevan väittämän.

Yleiskuvaus:

Oppilasta pyydetään vertailemaan kahden esineen painoa, jota (1) ei ehkä voida todeta pelkästään käsin punnitsemalla, (2) ei ehkä voida erottaa toisistaan edes vaa'an avulla. Kussakin tilanteessa esineiden koko vaihtelee irrelevanttina piirteenä. Tasavartinen vaaka on käytettävissä mutta oppilasta ei määrätä käyttämään sitä. Oppilasta kehoitetaan valitsemaan yksi kolmesta symbolista (>, <, =) ja sijoittamaan se tyhjään tilaan kahden painosymbolin väliin.

Ärsyke- ja reaktio-ominaisuudet:

Vakiot kaikissa ruuduissa. Tasavartinen vaaka (Tinkertoys). Vertailtavat esineet ovat läpikuultavia, sylinterin muotoisia pulloja, jotka ovat samanlaisia lukuunottamatta painoa (joko 23 tai 25 gr) ja kokoa (joko 2"x5/8" tai 2 1/2 x 1 3/4 tuumaa). Kumpaakin esinettä kuvataan satunnaisesti määräytyvällä pikkukirjaimella.

Oppilasta pyydetään täydentämään väittämä valitsemalla esineiden välistä painosuhdetta kuvaava symboli.

Ruutujen välinen erottelu
Kolme painosuhdetta (jotka voidaan saada selville vain vaa'an avulla, ei nostamalla tai tunnustelemalla) jotka määrittyvät sen perusteella missä asemassa **ne** ovat oppilaan edessä:

vasen > oikea; vasen < oikea; vasen=oikea

Kolme painosuhdetta:

vasen > oikea, vasen < oikea, vasen=oikea

Ruutumatriisi

Painosuhteet

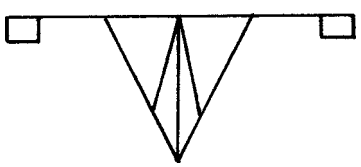
Kokosuhteet	Pvas>Poik	Pvas<Poik	Pvas=Poik
Kvas>Koik	(1)	(4)	(7)
Kvas<Koik	(2)	(5)	(8)
Kvas=Koik	(3)	(6)	(9)

Osiolomakekehikko

Aineisto Vipuvaaka 2 esinettä Ärsyke-reaktio lomake Lyijykyna	
Kokeenpitäjän ohjeet: Sijoita koeaineisto oppilaan eteen yllämainitussa järjestyksessä vaaka esineet Ä-r-lomak Oppilas	Koeteksti: Tässä on kaksi esinettä. Ne on merkitty kirjaimin. Vertaa esineiden painoa ja kirjoita jokin näistä kolmesta (näyttää) merkistä tähän tyhjään kohtaan (näyttää), jolloin saadaan täydellinen vertailua kuvaava lause. Voit halutessasi käyttää apuna vaakaa.

Vastauksen kirjaaminen

Liitä ärsyke-reaktiolomake tähän kohtaan. Kuvaa mitä oppilas teki. Jos oppilas käytti apuna vaakaa, merkitse esineiden kirjaintunnukset vaa'an kuppeihin ja merkitse missä asennossa pystysuora osoitin oli ratkaisuhetkellä.



Koeaineiston kuvaus:

Kynä.

Tasavartinen vipuvaaka.
Joukko esineitä painon vertailemista varten: joukko läpinäkyviä sylinterinmuotoisia muovisia pulloja, joissa tiivis korkki. Pullot ovat kahdenlaisia. Pieni pullo on 2 tuuman pituinen ja halkaisijaltaan 5/8 tuumaa. Ison pullon pituus on 2 1/2 tuumaa ja halkaisija 1 3/4 tuumaa. On valittu kaksi painoarvoa siten, ettei esineitä voi helposti erottaa käsin punnitsemalla mutta kyllä vaa'an avulla. Kunkin esineen tunnuksena on satunnaisesti annettu pieni kirjain.

Koko	Paino 23 gr	25 gr
Pieni	a	m k
Suuri	b	o n

Ärsyke-reaktio- lomake '(liitetään osioon): n. 6X4 tuumainen paperiarkki, jonka ulkoasuna on seuraava:

Kirjoita aukkoon > ,< tai =

P_{vas} ———— P_{oik}

vas ja oik tilalle sijoitetaan oikeat kirjainsymbolit substituutiotaulukosta.

Substituutiosuunnitelma (SS)

(vas,oik) esineet		
Ruutu 1:	(o,a)	
Ruutu 2:	(m,b)	
Ruutu 3:	valitse	SS 16:13:sta
Ruutu 4:	(b,m)	
Ruutu 5:	(a,o)	
Ruutu 6:	valitse	SS 16:14:sta
Ruutu 7:	valitse	SS 16:15:sta
Ruutu 8:	valitse	SS 16:16:sta
Ruutu 9:	valitse	SS 16:17:sta

Substituutioparit

SS 16.13	Järjestetyt parit	(m,a);(o,b)
SS 16.14	"	(a,m);(b,o)
SS 16.15	" "	(b,a);(o,m)
SS 16.16	" "	(a,b);(m,o)
SS 16.17	" "	(m,k);(o,n)

Pisteistysohjeet

Oikea vastaus syntyy, kun oppilas kirjoittaa oikean symbolin (> , < tai =) tyhjään tilaan ja täydentää näin vertailulauseen. Symbolin tulisi olla > ruuduissa 1,2 ja 3 < ruuduissa

Hively et al. osoittivat, että on mahdollista laatia toimivia osiontuottamissääntöjä. Hambleton et al. (1978) mukaan osiolomake soveltuu parhaiten sellaisiin sisältöalueisiin, joilla on selvä hierarkkinen rakenne (esim. matematiikka). Vaikka kaikille sisältöalueille ei voitaisikaan yhtä helposti laatia osiontuottamissääntöjä, on ilman muuta selvää, että vaikeuksista huolimatta jokaisella sisältöalueella olisi aineksen jäsentämisessä ja kuvauksessa päästävä nykyistä tarkempaan kuvaan.

3.2.4 Piirreanalyysi

Piirreanalyysin (facet analysis) kehittelijänä voidaan pitää Luis Guttmania (1969). Piirreanalyysiä voidaan käyttää samalla tavoin kuin edellä kuvattuja lähestymistapoja kuvaamaan testausolosuhteiden alueen rajoja ja rakennetta. Piirteet ovat tilanteen dimensioita, joiden uskotaan olevan mielekkäitä mittaamisessa. Piirteitä voidaan osa-aluemittauksessa pitää ulottuvuuksina tai piirteinä, joiden suhteen alueen osioiden tulee vaihdella. Piirteet valitaan ennenkuin osioalue kuvataan ja täten ne eroavat faktoreista, jotka saadaan esiin faktorianalyysissa sen jälkeen kun tiedot on kerätty ja analysoitu. Piirteet liitetään usein toisiinsa semanttisesti "kartoituslauseessa" (mapping sentence). Kuvio 2 havainnollistaa kartoituslausetta. Kun analysoidaan kartoituslauseita, käy Millmanin mukaan ilmeiseksi, että mahdollisten koeosioiden luonne jää suhteellisen epämääräiseksi. Onkin huomattava, että Guttman ja monet hänen työtovereistaan eivät käytä piirreanalyysiä tuottamaan testejä, jotka antavat tietoa jonkin asian hallinnan asteesta. He pyrkivät identifioimaan dimensiot, jotka ovat tärkeitä testin osioiden reaktioiden ja näiden dimensioiden rakenteen selvittämisessä. Millmanin arvion mukaan joitakin piirreanalyysin ideoita voidaan käyttää hyväksi rakennettaessa osa-aluekokeita.

A. KOHDE

Henkilö arvioi
1. itsellään
2. musiikinopettajilla
3. suurella yleisöllä
4. oppilailla

olevan sellaisen

B. KOHTEEN TUNTEMUSTILA että
1. uskomuksen
2. tuntemuksen

musiikkikasvatuksen opetussuunnitelmaa

C. TASO
1. peruskoulun
2. keskiasteen koulujen

D. TOIMIJAN (so. opetussuunnitelman) KÄYTTÄYTYMINEN
1. opettaa
2. tulisi opettaa

E. OPETTAJA
1. luokanopettaja
2. aineenopettaja tavoitteena

F. OPPILAAN YLEISET TARPEET
1. rentoutuminen
2. ilmaisu
3. vaihtelu lukuaineille
4. tunne-elämän kehittäminen
5. itsekuri
6. huvittaminen
7. kontaktit muihin ihmisiin
8. ryhmätoiminta
9. saada ilmi piileviä lahjakkuuksia
10. esiintyminen julkisuudessa
11. luova ilmaisu
12. menestyminen

KWIO 2. Kartoituslauseen malli (Millman 1974)

3.2.5. Lavennetut tavoitelauseet

Pophamin mielestä on saatava aikaan tasapaino selkeyden ja käytännöllisyyden välillä. Pophamin esittämä lavennettu tavoitelause (amplified objective) pyrkii tämän tasapainon aikaansaamiseen. Oheinen esimerkki valaisee lavennettua tavoitelausetta. Lavennettu tavoite on kasvatustavoitteen tarkennus, joka määrittelee rajat testaustilanteille, vastausvaihtoehdoille ja oikeellisuuden kriteereille. Lavennettu tavoite antaa osion kirjoittajalle paljon selkeämmän kuvan millaisia osioita heidän tulee kirjoittaa kuin mitä alkuperäinen tavoitelause antaisi. Millman mukaan jaa

---

Tavoite: Kun oppilaalle esitetään lause, josta on jätetty pois substantiivi tai verbi, hän osaa valita kahdesta vaihtoehdosta sen sanan, joka tarkemmin tai konkreettisemmen täydentää lauseen.

Malliosio

Ohjeet: Vedä viiva sen sulkujen sisällä esitetyn sanan alle, joka saa lauseen kuvauksen selkeämmäksi.

~~Esimerkki~~: Maastojuoksija (kompuroi, juoksi) mäkeä alas.

Lavennettu tavoite

Koetilanne

1. Oppilaalle annetaan yksinkertaisia lauseita, joista on jätetty pois substantiivi tai verbi, ja häntä pyydetään vetämään viiva sen vaihtoehdon alle, joka tarkemmin tai konkreettisemmin täydentää lauseen.

2. Kussakin kokeessa jätetään pois substantiiveja ja verbejä suurinpiirtein samassa suhteessa.

3. Sanaston tulee olla tuttua x-luokkalaiselle oppilaalle.

Vastausvaihtoehdot

1. Oppilaalle esitetään substantiivipareja tai verbipareja, joissa sanojen kuvaustarkkuudessa on selvä ero.

2. Verbipareissa toinen verbi on joko kopula tai yleistä toimintaa kuvaava verbi (esim. on, menee) ja toinen kuvailee toiminnan tapaa tai luonnetta (esim. kompuroi, hyppeli).

3. Substantiivipareissa toinen substantiivi on abstraktinen tai epämääräinen (esim. henkilö, esine) ja toinen konkreettinen tai tarkkamerkityksinen (esim. puuseppä, tietokone).

Oikean vastauksen kriteeri

Oikea vastaus on se, kun oppilas on kussakin parissa alleviivannut konkreetimman tai tarkemman substantiivin tai tarkemmin kuvaavan verbin.

KWIO 3. Lavennetun tavoitelauseen malli (Millman 1974)

kuitenkin varsin epävarmaksi, millainen olisi täydellinen tehtävien luokka, Mitkä ovat yksinkertaisia lauseita? Mikä on tuttu sanasto? Kuinka erotetaan abstraktit ja konkreetit substantiivit? Kuitenkin nämä epäselvyydet voivat olla pieni hinta siitä suhteellisesta helppoudesta jolla joitakin alueita voidaan kuvata.

Pophamin (1975) mukaan esimerkillisessa lavennetussa tavoitteessa täytyy olla perusteellinen kuvaus siitä, mitkä tilanteet tai ärsykkeet voivat muodostaa osion mukaanlukien mahdollinen sisältö, josta osiot voidaan generoida. Tämä voidaan saada aikaan joko laatimalla sisällön generoimissääntöjä (algoritmeja) tai luettelemalla kaikki aihepiirit (esim. romaanit, periaatteet, vuodet, kirjailijat tai muut alueet], joita osiot voivat käsitellä. Pophamin mielestä on yksilöitävä ei ainoastansa oikeat tai parempana pidetyt vastaukset (kuten joissakin affektiivisissa testeissä) vaan myöskin virheelliset tai vähemmän hyväksyttävät vastaukset. Tämä tarve on erityisen suuri esimerkiksi monivalinta- ja oikein-väärin -tehtävissä. Vapaita tuotoksia edellyttävien alueiden kuvauksen tulee Pophamin mielestä tuoda esiin selvä suppea kriteerijoukko, jonka perusteella voidaan arvioida vastausten asianmukaisuus. Popham on myöhemmin (1978, 137) todennut, että lavennetut tavoitelauseet eivät ole riittävän tarkkoja kriteerimittaukselle. Eva Baker (1974) on myöskin käsitellyt alueenmäärittelyn komponentteja. Hänen listaansa kuuluvat: 1. vastauksen kuvaus, ts. käyttäytymisen muotoina esitetty tavoite, 2. sisällön rajat, ts. joukko laadintacääntöjä tai luettelo sisällöistä, jotka voitaisiin liittää testiin, 3. vapaasti tuotettujen vastausten arviointikriteerit tai virheellisten vaihtoehtojen luonteen määritteleminen osioissa, joissa on annettu valmiiksi vastausvaihtoehtoja, 4. muoto jossa osiot esitetään oppilaille, 5. koeohjeet.

3.2.6. Koetäsmennys

Kriteerikokeen tärkein ominaisuus on, että se antaa selvän kuvauksen mitä oppilas osaa tehdä tai ei osaa tehdä tietyllä käyttäytymisalueella. Itse asiassa "kriteeri" Pophamin mukaan tarkoittaa juuri mitatun käyttäytymisen kuvausta.

Mitattavaa käyttäytymistä määriteltäessä on ratkaistava ainakin kaksi tärkeää kysymystä: toinen koskee mitattavan käyttäytymisen yleisyystasoa ja toinen mitattavan käyttäytymisen muodon valintaa. Yleisyystasoa määri-

teltäessä on löydettävä järkevä kompromissi hyvin tarkan määrittelyn ja kovin yleisen määrittelyn välillä. Pophamin mukaan on järkevää valita "puolikarkea" mittausstrategia (limited-focus measurement strategy), jossa otetaan mittauksen kohteeksi pieni määrä tärkeitä käyttäytymisen muotoja, vaikka ne osoittautuisivatkin melko kompleksisiksi. Tämä merkitsee, että mitataan todella tärkeitä käyttäytymisen piirteitä, joihin sisältyy joukko pienempiä osatekijöitä. Tämän jälkeen on valittava varsinainen mittaustapa. Lähes jokaista käyttäytymisen muotoa voidaan mitata perustellusti usealla eri tavalla. On syytä käydä huolellisesti läpi kaikki kyseeseen tulevat mittaustavat ja valita niistä sellainen, joka antaa parhaimmat mahdollisuudet tulosten yleistämiseen. Pophamin mukaan useinkaan ei ole käytännössä järkevää mitata samaa käyttäytymisen muotoa useilla kyseeseen tulevilla mittaustavoilla, koska tämä vaikeuttaa tulosten tulkintaa. Mittaustapojen yleistettävyyttä voidaan mitata empiirisesti katsomalla, miten eri mittaustapojen antamat tulokset korreloivat keskenään. Käytännössä on kuitenkin usein tyydyttävä asiantuntijaharkintaan.

Popham suosittelee, että kriteeriviitteisessä koetäsmennyksessä olisi ainakin seuraavat viisi komponenttia: 1) kokeen mittaavan käyttäytymisen yleinen kuvaus, 2) mittaustapaa havainnollistava esimerkkiosio, 3) sallittavan ärsykeaineiston ominaisuuksien kuvaus ja rajaus, 4) edellytetyn vastauksen ominaisuuksien kuvaus: a) valittavien vaihtoehtojen ominaisuuksien kuvaus tai b) tuotettavan vastauksen arviointikriteerien määrittely, 5) joissakin tapauksissa saattaa olla suositeltavaa tai jopa välttämätöntä esittää erillisessa liitteessä lisätäsmennyksiä. Liite saattaa sisältää esimerkiksi yksityiskohtaisia luetteloita tai kyseeseen tulevan sisällön tarkkaa selostusta.

Carrollin (1968, 1976) työtä kokeiden dimensioiden systematisoinniksi voidaan käyttää hyväksi myös kriteerimittauksessa. Carrollin hiljattain julkaisema analyysi kokeista kognitiivisina tehtävinä on perusteellinen erittely kokeen ominaisuuksista (Carroll 1976, 27-56).

Carrollin systeemissä käsitellään ärsykeaineistoa, reaktioaineistoa, tehtävärakennetta, operaatioita ja strategioita seka muistin rakennetta seuraavalla tavalla:

I. ÄRSYKEAINEISTO (TEHTÄVÄN ALUSSA]

1A Ärsykeluokkien määrä

1. yksi ärsykeluokka (sana, kuva tms.)
2. kaksi ärsykeluokkaa (esim. useissa monivalintatyypeissä, pari assosiaatiotehtävissä)

Tietyn arsykeluokan i kuvaus:

1B Täydellisyys/kokonaisuus

1. kokonainen/täydellinen
2. osittainen/vajavainen (visuaalista tai auditiivista "hälyä")

1C Tulkittavuus

1. selkeä/yksiselitteinen (suoraan tulkittavissa)
2. moniselitteinen (koodattavissa usealla tavalla)
3. epäselvä (ei suoraan tulkittavissa]

Tulkinnassa käytettävä muisti

6A Muistinkesto (ks. kohta 6A)
6A Sisältö (ks. kohta 6B)
6C Yksilöllisten erojen relevanssi ko. muistivarastossa (ks. lista 6C)

II. ILMIREAKTIO (TEHTÄVÄN LOPUSSA)

2A Määrä ja tyyppi

1. valita reaktio esitetyistä vaihtoehdoista
2. tuottaa yksi oikea vastaus tarvittavien operaatioiden perusteella
3. tuottaa mahdollisimman monta erilaista reaktiota
4. tuottaa tietty määrä erilaisia reaktioita

2B Vastaustapa

1. osoittaa valittu vaihtoehto
2. tuottaa yksi symboli (kirjain, numero]
3. kirjoittaa sana
4. kirjoittaa fraasi tai lause
5. kirjoittaa kappale tai enemmän
6. vastata suullisesti
7. vetää viiva tai tehdä yksinkertainen piirustus

2C Reaktion hyväksyttävyyden kriteeri

1. identtisyys
2. samanlaisuus (tai erilaisuus) yhden tai useamman piirteen suhteen
3. semanttinen vastakohta
4. sisältyminen/kuuluminen johonkin luokkaan tms.
5. oikea tulos operaatiosarjan perusteella
6. esimerkkitapaus (arsykeluokkaan kuuluva tapaus)
7. yläkäsite
8. oikea vastaus kysymykseen ("Ilmaista mikä, kuka, miksi jne.")
4. vertaileva arvio
10. mielivaltainen assosiaatio
11. semanttinen ja/tai kieliopillinen hyväksyttävyys ("tajuttava", "järkevä")
12. viivojen tai polkujen yhteenliittyvyys/jatkuvuus

III TEHTÄVÄRAKENNE

3A 1. yhtäjaksoinen (kaikki tehtävät suoritetaan samalla kertaa)

2. ärsykkeet esitetään tiettynä ajankohtana ja reaktiot myöhempänä ajankohtana

Tätä kohtaa on tarkennettava, jotta saataisiin katettua erilaiset kokeelliset tilanteet.

IV OPERAATIOT JA STRATEGIAT

4A Operaatioiden ja strategioiden määrä

Tietyn operaation i kuvaus:

4B Kuvauksen tyyppi

1. identifioida/tunnistaa ja tulkita ärsyke
2. todeta samanlaisuuksia kahden tai sitä useamman ärsykkeen välillä
3. palauttaa mieleen nimi, kuvaus tai esimerkki
4. varastoida tehtävä muistiin
5. palauttaa muistista assosiaatioita tai yleistä tietoa
6. palauttaa mieleen tai muodostaa hypoteeseja
7. tutkia muistin eri osia
8. suorittaa sarja operaatioita muistitiedon perusteella
9. kirjata/varastoida välitön tulos
10. visuaalinen tarkistusstrategia (tarkastella visuaalisten ärsykkeiden eri osia)
11. tulkita uudelleen mahdollisesti moniselitteinen tehtävä
32. luoda mielikuvia tai käyttää jotakin muuta tapaa ärsykkeiden abstraktiseen edustamiseen
13. rotatoida mielessään spatiaalista kuviota/hahmoa
14. ymmärtää ja analysoida verbaalinen ärsyke
15. arvioida ärsykettä tietyn ominaisuuden suhteen
16. jättää vaille huomiota asiaankuulumaton/irrelevantti ärsyke
37. käyttää apuna tiettyä muistitekniikkaa (tarkenna mitä)
18. kertailla mielessään assosiaatioita
19. kehittää erityinen (visuaalinen) hakutekniikka
20. ryhmitellä/jaksotella ärsykkeitä muistista laajemmiksi kokonaisuuksiksi

4C Onko operaatio määritelty tehtävän ohjeistossa?

1. määritelty selkeästi
2. käy epäselvästi ilmi
3. ei määritelty eikä käy ilmi

4D Kuinka riippuvainen hyväksytty suoritus on ko. operaatiosta tai strategiasta?

1. ratkaisevasti riippuvainen
2. on avuksi mutta ei ehdottoman välttämätön
3. epävarma (voi olla avuksi tai haitaksi)
4. mahdollisesti jopa haitaksi

Ko. operaatiossa tarvittava muisti:

6A Muistin kesto (ks. kohta 6A)
6B Sisältö (ks. kohta 6B)
6C Yksilöllisten erojen reievanssi ko muistivarastossa (ks. lista 66)

V OPERAATION TAI STRATEGIAN AIKA-ASPEKTIT

(jos 5A = 0 eli "irrelevantti, 5B viittaa siihen miten todennäköisesti henkilö valitsee tietyn strategian)

5A Kesto (keston keskimääräinen vaihteluväli)

0. irrelevantti
1. hyvin lyhyt (esim. < 200 msek)
2. keskikestoinen (esim. < 1 sek)
3. pitkähkö (esim. 1-5 sek)
4. pitkä (esim. > 5 sek)

5B Yksilölliset kestoerot (tai strategian todennäköisyys)

1. todennäköisesti vailla merkitystä
2. mahdollisesti relevantti seikka
3. todennäköisesti suuri yksilllinen vaihtelu

5C Operaation lopettamisen kriteeri

0. irrelevantti
1. itselopetus (oikeaan vastaukseen päädyttyä)
2. itselopetus (arvaus, virheelliseen vastaukseen tyytyminen)

VI MUISTIVARASTO

6A Kesto

1. sensorinen puskurimuisti
2. lyhytaikaismuisti (muutama sekunti)
3. työmuisti (muutama minuutti]
4. kestomuisti

6B Sisältö

0.5. ei spesifi
1.0. visuaalinen (yleinen, ei-spesifinen)
1.1. pisteitä, pisteiden paikat
1.2. viivoja (1-dimensionaalinen)
1.3. viivoja ja käyriä (2-dimensionaalinen)
1.4. geometrisia kuvioita
1.5. kuvallinen esitys (esim. esine)
1.5.1. --- alaluokka (esim. työkaluja)
1.6. todellisia 2-dimensionaalisia esineitä
1.7. karttoja ja kuvioita
1.8. 3-ulotteisia geometrisia hahmoja
1.9. 3-ulotteisia esineitä
2. auditiivinen (ei spesifioida tässä tarkemmin)
3. grafeeminen (yleinen)
3.1. kirjaimia
3.2. sanoja (ei oteta huomioon niiden semanttista informaatiota)
3.3. aakkosjärjestykseen järjestettyä informaatiota
4.0. lingvistinen (äidinkieli)
4.01.--- alaluokat (esim. alan terminologia)
4.1. sanastollinen (leksikaalinen)
4.33.--- alaluokat
4.2. kieliopillinen (syntaktinen)
4.21 sanastollis-kieliopillinen (esim. sanojen kieliopillinen luokittelu)

4.3. kieliopillisia sääntöjä ja piirteitä
4.4. semanttinen (sanojen merkityksiä, syntaktisia piirteitä jne.)
4.5. ei-kielellisiä merkityksiä (esim. kuvasymbolien merkityksiä)
5.0. numeerinen, matemaattinen
5.1. numerosymbolit merkityksineen
5.2. alkeelliset numero-operaatiot ja symbolit
5.3. algoritmit kvantitatiivisten suhteiden käsittelemiseksi
6.0. logiikka
6.1. erilaiset abstraktit kuviot (vaihtelu, sekvenssi jne.)
6.2. ominaisuudet, joiden suhteen ärsykkeet voisivat vaihdella
7.0. liikkeet "liikekäsitteet"
8.0. "todellisen elämän" kokemukset, tilanteet, faktat, informaatio
8.1. ---alaluokat (esim. mekaniikkaa ja sähköä koskeva informaatio)
9.0. mielivaltaiset uudet koodaukset ja assosiaatiot, jotka syntyvät koetilanteessa

6C Yksilöllisten erojen relevanssi tässä muistivarastossa

3. useimmilla henkilöillä on tarvittava muisti(varasto)
2. epävarmaa *onko* kaikilla tarvittava muisti(varasto)
3. suuria yksilöllisiä eroja kyseisessä muistivarastossa

Tämän raportin kirjoittaja laati IEA:n Grannan kesäseminaarissa v. 1971 kielenopetuksen tavoitteiden spesifointiyrityksen, joka julkaistiin viisi vuotta myöhemmin (Takala 1976). Se perustui Carrollin (1968) ja Valetten (1971) esityksiin. Keväällä 1979 toimeenpannun projektin "Peruskoulun tilannekartoitus I" yhteydessä toteutetussa englannin kielen projektissa, jossa kirjoittajan ensisijaisen mielenkiinnon kohteena oli sanaston hallinnan tutkiminen, laadittiin sanaston kokeiden tuotamissäännöstön malli. Kyseessä ei ole siis varsinainen koetäsmennys, vaan yleisempi koetasmennysjärjestelmä (Carrollin, 1976, tyyliin), jonka avulla voidaan määritellä yksityisten sanastokokeiden spesifi sisältö.

Sanaston kokeen tuottamissäännöt

Oppilaan tehtävä: a) osoittaa, että hän tunnistaa englanninkielisen sanan merkityksen(set)

b) osoittaa, että hän osaa käyttää oikeassa merkityksessä englanninkielistä sanaa

Yleinen kuvaus

a) Sisältöalueen (domain) kartoitus

On kartoitettava kaikissa oppikirjoissa esiintyvä yhteinen sanasto ja kussakin oppikirjassa esiintyvä lisäsanasto. Tämän lisäksi on selvitettävä, onko sanasto peruskappaleisiin kuuluvaa, vai mahdollisesti valinnaiseen ainekseen kuuluvaa. Luetteloa verrataan opetussuunnitelmassa esiintyvään sanastoa koskevaan määrälliseen ja sisällölliseen määrittelyyn.

b) Sisältöalueen otanta

On selvitettävä, onko tarkoituksenmukaista jakaa sisältöalue osa-alueisiin, esim. sanaluokkiin. Otanta suoritetaan satunnaisperiaatteella joko yksinkertaisena satunnaisotantana tai suhteellisena ositettuna otantana. Harkintaotantaa ei käytetä.

Esimerkkiosiot

Tehtävä a: hundred _____

Tehtävä b: harmaa _____

Ärsyke- ja reaktiotilanne

A. Ärsyketilanne

1. Kieli

1.1. Ärsyke esitetään opiskeltavalla kielellä — x (tehtävä a)

1.2. Ärsyke esitetään oppilaan äidinkielellä — x (tehtävä b)

2. Ärsykkeen esitysmuoto

2.1. Ärsyke esitetään pelkän puheen välityksellä

2.2. Ärsyke esitetään pelkästään kirjoitettuna _____
2.3. Ärsyke esitetään pelkästään kuvallisesti _____
2.4. Ärsyke esitetään pelkästään toiminnan avulla (ostensiivisesti) _____
2.5. Ärsyke esitetään puheen ja kuvan avulla _____
2.6. Ärsyke esitetään puheen ja toiminnan avulla _____
2.7. Ärsyke esitetään puheen ja kirjoituksen avulla _____
2.8. Ärsyke esitetään kirjoituksen ja kuvan avulla _____
2.9. Ärsyke esitetään kirjoituksen ja toiminnan avulla _____
2.10. Ärsyke esitetään kuvan ja toiminnan avulla _____
2.31. Ärsyke esitetään puheen, kirjoituksen ja kuvan välityksellä _____
2.12. Ärsyke esitetään puheen, kirjoituksen ja toiminnan avulla _____
2.13. Ärsyke esitetään puheen, kuvan ja toiminnan avulla _____
2.14. Ärsyke esitetään kirjoituksen, kuvan ja toiminnan avulla _____
2.15. Ärsyke esitetään puheen, kirjoituksen, kuvan ja toiminnan avulla _____

Tarkempi kuvaus:

_____

_____

_____

_____

_____

3. Ärsykkeen kompleksisuus

3.3. sana (yksin) X

3.2. Fraasi/lauseke

Tarkempi kuvaus:

4. Kontekstuaalisuuden määrä

4.1. irralliset sanat/fraasit vailla kontekstia X

4.2. sanat/fraasit irrallisen lauseen osana

4.3. sanat/fraasit yhtenäisen, vähintään kaksi lausetta sisältävän diskurssin (dialogi tai monologi] osana

4.4. sanat/fraasit yhtenäisen tekstin osana (cloze)

Tarkempi kuvaus:

5. Kielenkäyttötilanteen ilmitulon aste

5.1. ei käy ilmi X

5.2. voidaan epäsuorasti päätellä

5.3. voidaan selvästi päätellä

5.4. käy selvästi ilmi

Tarkempi kuvaus:

B. Reaktiotilanne

1. Reaktion tuottamisen aste

1.3. Reaktio koostuu vaihtoehtojen kesken tapahtuvasta valinnasta _____

1.2. Reaktio koostuu vaihtoehtojen ja ärsykkeiden yhdistelemisestä _____

1.3. Reaktio koostuu ohjatusta (erilaisiin vihjeisiin perustuvasta) tuottamisesta (prompted recall) X

1.4. Reaktio koostuu ohjaamattomasta tuottamisesta (unaided recall) _____

Tarkempi kuvaus:

_____

_____

_____

_____

_____

2. Reaktio-osan kielimuoto:

2.1. opiskeltava kieli X(tehtävä b)

2.2. äidinkieli X(tehtävä a)

Tarkempi kuvaus:

_____

_____

_____

_____

_____

3. Reaktion esitystapa

3.3. puhuminen _____

3.2. kirjoittaminen X

3.3. piirtäminen _____

3.4. toiminta _____

3.5. merkin tekeminen _____

3.6. puhuminen + kirjoittaminen _____

3.7. puhuminen + piirtäminen _____
3.8. puhuminen + toiminta _____
3.9. kirjoittaminen + piirtäminen _____
3.10. kirjoittaminen + toiminta _____
3.11. piirtäminen + toiminta _____
3.12. puhuminen + kirjoittaminen + piirtäminen _____
3.13. puhuminen + kirjoittaminen + toiminta _____
3.14. puhuminen + piirtäminen + toiminta _____
3.15. kirjoittaminen + piirtäminen + toiminta _____
3.16. puhuminen + kirjoittaminen + piirtäminen + toiminta _____

Tarkempi kuvaus:

_____

_____

_____

_____

_____

4. Reaktion pisteistys _____
4.1. kaksiportainen pisteistys: 0 - 1 (oikein - vaarin) X
4.2. moniportainen pisteistys (erotetaan useampia osaamisen asteita) _____

Tarkempi kuvaus:

_____

_____

_____

_____

_____

5. Reaktiolle asetettava suoritusaikavaatimus

5.1. vastausaika täysin rajoittamaton (puhdas tasokoe)	
5.2. vastausaikaaperiaatteessa riittävästi (tasokoe)	X
5.3. vastausaika selvästi rajoitettu (nopeuskoe)	

Tarkempi kuvaus:

6. Reaktiolle asetettava vaihtelevuus/monipuolisuusvaatimus

6.1. tuotettava irrallinen sana/fraasi	X
6.2. tuotettava sana/fraasi osana vähintään kokonaisia 1ausetta	
6.3. ei relevantti	

Tarkempi kuvaus:

7. Reaktiolle asetettava korrektisuusvaatimus

7.1. oltava täysin edellytetyn vastauksen kaltainen maksimipisteen samiseksi	
7.2. kaikki hyväksyttävät virheettömät vastaukset tuottavat maksimipisteen	
7.3. ymmärrettävä mutta osin virheellinen vastaus tuottaa maksimipisteen	X
7.4. virheitä sisältävästä vastauksesta vähennetään tietty osa maksimipisteestä	

Tarkempi kuvaus:

_____

_____

_____

_____

_____

Yhteenvetona voimme todeta, että Ebelin (Ebel 1962) ja Hivelyn (Hively, Patterson & Page 1968) ja Pophamin (1978) työ sisältöalueiden tarkennuksen alalla merkitsee, että kriteerikokeiden osioille asetetaan seuraavat vaatimukset:

1) Kaikki osiot, jotka voidaan kirjoittaa kyseiseltä sisältöalueelta, on laadittava (tai ainakin tunnettava) ennen kuin osiot valitaan kokeeseen.
2) Osioita valittaessa on käytettävä satunnaista tai ositettua satunnaisotantaa.

3.2.7. Mitattavan osa-alueen määrittämisen ongelmia

3.2.7.1. Alueen koko ja sitä määrittelevät piirteet

Kaikki tässä jaksossa käsiteltävät osioiden generointisysteemit jättävät määrittelemättä kysymyksen, miten määritellään mitkä mahdollisista tehtävistä tulisi sisällyttää alueeseen. Tässä käsitellään pikemminkin sääntöjä joiden perusteella generoidaan osioallas sen jälkeen kun alueen piirteet tai dimensiot on identifioitu. Tässä näyttää olevan kaksi toistensa kanssa yhteydessä olevaa ongelmaa: ensimmäinen ongelma koskee alueen kokoa eli kuinka laaja tehtävien populaatio tulisi kysymykseen ja toinen koskee niitä spesifejä piirteitä ja elementtejä, jotka tulisi valita määrittelemään kyseistä aluetta,

Millaista "käyttäytymistä" mitattavan osa-alueen tulisi edustaa? Tulisiko mitata päätekäyttäytymistä (terminal behaviors) vai etappikäyttäytymistä (enroute behaviors)? Käsitteet "osa-alueen koko" (domain size), "yleisyystaso" (terminality, degree of generality) ja "saavutustasot" (levels of achievement) liittyvät kaikki mitattavan osa-alueen koon ongelmaan. Nitkon (1974) mukaan osa-alueen koko voi vaihdella koulutusjärjestelmän tuloksista tietyn kurssin tuloksiin.

Scrivenin (1967) ehdottamat käsitteet formatiivinen ja summatiivinen arviointi, jotka mm. Bloom omaksui mastery learning -järjestelmäänsä, ovat käyttökelpoisia tässä yhteydessä. Formatiivisia kriteerikokeita voidaan käyttää välietappievaluoinnissa.

Popham (1972) on ollut enemmän kiinnostunut päätekäyttäytymisen evaluoinnista, koska hän ei ole kiinnittänyt sanottavaa huomiota oppimishierarkioiden validointiin. Myös Eva Baker (1974) on esittänyt argumentteja kovin suppeita osa-alueita vastaan, koska ne saattavat johtaa kovin triviaalien asioiden mittaamiseen ja näin ollen liialliseen testaamiseen. Popham (1975) on suositellut vähintään viikon opintoaikaa osa-alueen kooksi, mikä Yhdysvaltojen opetusjärjestelmän luonteen mukaisesti merkinnee vähintään viittä oppituntia.

Millmanin (1974) mukaan tietynlainen hyötysuhde (trade-off) vallitsee kysymyksessä, joka koskee alueen kokoa. Kokeenlaatija haluaa arvioida laajasti määritellyn alueen, joka kattaa hyvin opetussuunnitelman tavoitteet. Kuitenkin kokeenlaatija haluaa myöskin tietää, miten hyvin oppilas suoriutuu tehtävistä, ja laajasti määritellyt alueet (mikä merkitsee vastaavasti heterogeenisia osioita) ovat taipuvaisia antamaan testin tulkinnan,

että opiskelija osaa tehdä joitakin asioita mutta ei joitain muita. Esimerkiksi on liian ylimalkaista sanoa, että oppilas ymmärtää uutta matematiikkaa. Joukko-opin alkeellisten käsitteiden ymmärtäminen on parempi kuvaus. Joukoilla suoritettavien keskeisten operaatioiden ymmärtäminen on sopivan tuntuinen kuvaus. Samanlaisten ja vastaavien joukkojen erotteleminen saattaa olla liian kapea. Jos kokeenlaatijat pelkäävät että alueiden määrittely tulee liian laveaksi, Millmanin mukaan on syytä harkita sisältöalueen jakamista osa-alueisiin ja arvioida suoritustasoa kunkin osalta erikseen.

Alueen koon optimi riippuu kokeen tarkoituksesta. Laajempi aluekoko voidaan sallia kokeessa, joita käytetään tärkeitä päätöksiä tehtäessä tai yleistuloksia raportoitaessa. Pienempi alueen koko on suositeltava esim. tukiopetusta ajatellen, jossa tarvitaan tietoa spesifeistä taidoista. Popham (1975) antaa kolme käytännön suositusta alueen koon määrittelemiseksi. Ensinnäkin hän kehoittaa kokeenlaatijaa arvioimaan, kuinka paljon opetusaikaa tarvittaisiin kyseiseen alueeseen sisältyvien taitojen omaksumiseen. Yhden oppitunnin aikana opittava asia on liian kapea alueeltaan; sellaiset asiat jotka vaativat kokonaisen lukukauden ovat liian laajoja. Toiseksi hän ehdottaa, että kokeenlaatija arvioisi alueen koon jakamalla mitattavaksi haluttujen tärkeiden taitojen käyttäytymismuotojen kokonaismäärän niiden kokeiden määrällä, jotka hän on valmis laatimaan. Kolmas näkökohta on tehdä alue niin laajaksi kuin mahdollista ja samalla kuitenkin säilyttää tietty sisällön homogeenisuus.

Kun arvioitavaksi tarkoitetut tiedot, taidot tai asenteet on määritelty, kokeenlaatijan tulisi valita käytännön rajoitukset huomioonottaen ne piirteet ja elementit näiden piirteiden sisällä, joiden arvioidaan maksimoivan osiovarianssia. Humphreys (1962) esitti, että kokeenlaatijan tulee tarkoituksellisesti laatia koe niin heterogeeniseksi kuin mahdollista ottaen huomioon mittauksen tarkoituksen. Alue kannattaa Millmanin (1974) mukaan määritellä keskeisten piirteiden ja elementtien suhteen, jotka vaikuttavat oppilaan vastaukseen (make a difference is how the examiner responds). Jos esimerkiksi väri ei vaikuta millään tavalla oppilaiden vastauksiin, väri voidaan pitää vakiona tai se voidaan jättää kokonaan huomioon ottamatta. Ei ole tiedossa mitään syytä, miksi alueen piirteiden pitäisi välttämättä olla hierarkkisesti järjestyneitä. Mitkä piirteet sitten ovat tärkeitä? Tässä täytyy Millmanin mukaan käyttää hyväksi intuitiivista tietoa. Joillakin aloilla on empiiristä tietoa käytettävissä tästä kysymyksestä. Tarvitaan kuitenkin enemmän tutkimusta sisältö- ja menetel-

mämuuttujista, jotka vaikuttavat testin suorittamiseen. Tätä ongelmaa voidaan tutkia multippeliregressioanalyysillä ja selvittämällä varianssikomponentteja, joiden avulla voidaan tutkia yleistettävyysongelmia. Millman (1978) on äskettäin itse tehnyt esitutkimuksen, joka käsittelee osion muotoilun vaikutusta sen ratkaisuprosenttiin.

3.2.7.2. Osioiden ja opetuksen välinen yhteys

Kokeen mittaustarkoituksen määrittely on kaiken mittaamisen perusläh-tökohtia. Tällöin nousee Smithin (1978) mukaan ongelmaksi mm. seuraava kysymys: kuinka opetussidonnaisen tulisi kriteerikokeen olla?

Opetussidonnaisuudella (instructional dependence) tarkoitetaan sitä, kuinka läheisesti osa-alue ja osiot riippuvat tai ovat yhteydessä opetussuunnitelmaan, oppimateriaaleihin ja opetukseen. Kriteerimittauksen alalla ei vallitse, ehkä odotetusti, yksimielisyyttä tässä asiassa.

Toista äärimmäisyyttä edustaa Bormuth (1970), jonka mukaan kokeiden sisältö tulee johtaa suoraan opetuksesta tarkasti tunnettujen operaatioiden (= kielellisten transformaatioiden) avulla. Koe voidaan kyllä linkittää opetukseen ilman ennaltamäärättyjä tavoitteita. Bormuthin mielestä opetuksen evaluoinnin tarkoituksena ei ole pelkästään saada selville mitään niin tiukasti rajattua seikkaa, kuin kuinka hyvin opetus saa aikaan opetussuunnitelman, tai ehkä paremminkin opetuksen suunnitelman, tavoitteet. Bormuthin mukaan opetuksen tavoitteet saattavat jättää vaille huomiota opetuksen positiiviset ja negatiiviset oheisvaikutukset ja opetusohjelmat voivat sisältää paljonkin sellaista opetusta, jota ei selvästi ilmaista tavoitteissa. Bakerin (1974) Southwest Regional Laboratoryssä (SWRL) kehittämässä evaluaatiojärjestelmässä mittausmenettelyt kytketään tiukasti tavoitteisiin ja niiden oppilailta vaatimiin reaktioihin.

Robert Baker (1974) on, kuten edellä mainittiin, esittänyt että osiot tulisi johtaa suoraan opetuksesta. Eva Baker (1974) ja Popham (1975, 1978) puolestaan kannattavat transfertyyppisten osioiden sisällyttämistä kriteerikokeeseen. Popham (1975) ja Nitko käyttävät käsitettä "likitavoitteet" (proximate goals) kuvaamaan sellaisia tärkeitä käyttäytymisen muotoja, jotka oppilaiden odotetaan omaksuvan opetuksen tuloksena vaikkakaan niitä ei olisi suoraan harjoitettu opetuksen aikana. Tässä on kuitenkin ongelmana Bormuthin (1970) mainitsema seikka: mikä yhdelle oppilaalle on transfertehtävä, voi toiselle olla pelkkä muistitehtävä. Kriteerikokeen laatijan tulisi siksi tuntea oppilaan kokemustausta (Harris 1974).

Robert Bakerin työtoveri Schutz (1978) korostaakin, että SWRL:ssä ollaan enemmän kiinnostuneita mittaamisesta opetuksen puitteissa (measurement in instruction) kuin opetuksen (tulosten) mittaamisesta (measurement of instruction). Kasvatuksessa ja opetuksessa tavoitteet, opetus ja mittaaminen tulee liittää funktionaalisesti toisiinsa. Schutzin mukaan Yhdysvalloissa on korostettu ulkopuolisen evaluoinnin merkitystä, koska opetuksen on arveltu vaihtelevan voimakkaasti. Hän väittää tarkemman analyysin osoittavan, että n. 90 % matematiikan ja lukemisen alkeisopetuksesta on tavoitteeltaan samansuuntaista ja myös samansisältöistä.

Schutzin mukaan on siten mahdollista kehittää apuvälineitä, jotka vastaavat opetusohjelmia ja joita voidaan käyttää luokan ja koulun tasolla seka sitä yleisemmälläkin tasolla. SWRL kutsuu tällaista mittavälineistöä "opetuksen tulosten informaatiojärjestelmäksi" (Instructional Accomplishment Information, IAI). Mittareiden lisäksi SWRL on kehitellyt opetustulosten aikaansaamista edistäviä välineitä, joita nimitetään "kokonaisvaltaisiksi opetusohjelmiksi" (Comprehensive Programs for Instruction, CPI). Näiden tarkoituksena on tehdä "korkeatasoisesta opetuksesta" (high quality instruction) todellisuutta eikä pelkkää retoriikkaa.

Jonkin verran toisenlaista lähestymistapaa edustavat tavoitteiden ja sisältöjen analyysimenetelmät. Glaserin ja Nitkon (1971) käsityksen mukaan osa-alueen valinnassa on *myös* tarpeen analysoida opetettavaa sisältöaluetta, jotta saataisiin esiin mielekkäitä sisältöyksiköitä, sekä *myös* suoritus- eli käyttäytymisdimensiota kunkin yksikön puitteissa. Sisällön analysointi on luonnollisesti tärkeätä erityisesti silloin, kun opetusohjelma itse perustuu tietynlaiseen tavoitetaksonomiaan tai teoriaan. Hivelyn (1973) MINNEMAST-projekti on hyvä esimerkki tutkimuksesta, jossa tavoitteena ei ole pelkästään oppimistulosten arviointi vaan myös oppimishierarkioita koskevien hypoteesien tutkiminen.

Pophamin organisoimassa projektissa Instructional Objectives Exchange (IOX), josta tarvitsijat voivat tilata osa-aluemäärityksiä ja malliosioita, käytetään apuna oppiaineasiantuntijoiden arvioita. Samoin on menetelty laajassa kansallisessa evaluaatiotutkimusprojektissa (National Assessment of Educational Progress, NAEP) Yhdysvalloissa.

3.3. Osioiden tuottaminen ja valikoiminen

Alueen määritteleminen ei ole yksinkertainen tehtävä, kuten kävi ilmi edellisessä jaksossa. Huolellisesti määritelty alue helpottaa kuitenkin osioiden laatimista. Osioiden laatijan tehtävänä on tuottaa osioita, jotka pitäytyvät alueen määrittelyn luomiin rajoituksiin. Tämä tehtävä on lähes mekaaninen joissakin tilanteissa. Joskus tarvitaan älykkyyttä pikemminkin kuin luovuutta.

Yritettäessä laatia osioita saattaa herätä kysymyksiä aluemäärittelystä. Mitkä testityypit ovat hyväksyttäviä? Ovatko transfertehtävät asianmukaisia? Kuinka sidottuja opetuksen sisältöön ja menetelmiin tehtävien tulisi olla? Alueen määritteleminen ja osion kirjoittaminen ovat toisiaan täydentäviä toimintoja. Yleisesti ottaen kriteerikokeen osioiden kelvollisuus riippuu siitä, kuinka hyvin ne heijastavat mitattavaa aluetta., Osioiden sisällön validiteetti voidaan todeta kahdella eri tavalla: toisessa sisällön asiantuntijat arvioivat mitattavan alueen ja osioiden vastaavuutta (match) ja toisessa käytetään empiirisiä keinoja. Edellistä voitaisiin ehkä kutsua harkintapohjaiseksi osioanalyysiksi, jota käytetään tyypillisesti ennen kokeen lopullista kokoamista. Jälkimmäinen edustaa perinteellistä empiiristä osioanalyysia.

Rovinelli ja Hambleton (1977) ovat kuvanneet kolmea eri tapaa arvioida osioiden sisällön validiteettia harkintapohjaisen osioanalyysin keinoin. Ensimmäisessä sisällön asiantuntijoita pyydettiin arvioimaan osioita suhteessa tiettyihin tavoitelauseisiin. Arviointiasteikko oli seuraava:

+1 = varma käsitys, että osio mittaa tavoitetta
0 = epävarma, mittaako osio tavoitetta vai ei
-1 = varma käsitys, että osio ei mittaa tavoitetta

Toisessa menetelmässä sisällön asiantuntijoita pyydettiin arviointiasteikkoa käyttäen arvioimaan, kuinka sopivia osiot ovat mittaamaan tiettyä tavoitetta. Kolmannessa menetelmässä asiantuntijoita pyydettiin yhdistelemään osiolistassa esitetyt osiot tavoitelistassa esitettyihin tavoitteisiin (matching task).

Myös Cronbach (1971) on esittänyt erittäin hyödyllisen tavan arvioida osioiden edustavuutta. Kaksi yhtä pätevästi arvioitua ryhmää saa tarkat kokeenlaadintaohjeet (sisällön määrittely, otantaohjeet, ohjeet osioiden ennaltatarkistamisesta, ohjeet esikokeen pitämisestä ja tulosten tulkinnasta), ja heidän tehtävänsä on laatia kriteerikoe. Jos ohjeet ovat selkeät,

tuloksena tulisi olla kaksi ekvivalenttia koetta. Rinnakkaisuus voitaisiin empiirisesti kokeilla esittämällä kokeet samalle henkilöjoukolle.

Millmanin (1974) mukaan on suositeltavaa että kokeen laatija esittää osiot riippumattomien (= ulkopuolisten) tarkastelijoiden arvioitavaksi, jotta saataisiin kuva osion ja alueen keskinäisestä yhteensopivuudesta. Tämä menettely on erityisen tärkeätä, kun käytetään vähemmän tarkkoja alueenmäärittelyn lähestymistapoja, kuten lavennettuja tavoitelauseita. On suositeltavaa, etteivät arvioitsijat ole aikaisemmin olleet mukana alueen määrittelemisessä eikä osioiden laatimisessa.

Vaikka varianssi ei olekaan keskeinen tekijä kokeenlaadinnassa, sillä on kuitenkin tietty käyttönsä. Suorituksissa esiintyy tiettyä varianssia silloin, kun kokelasjoukko on ainakin jossakin määrin heterogeeninen. Empiirisiä keinoja voidaan käyttää apuna kokeenlaadinnassa.

Henrysson & Wedman (1974) esittävät, ettei hyvin huolellisesti laaditut koetäsmennykset eikä tarkat osiontuottamissäännötkään voi täysin eliminoida tiettyä subjektiivisuutta kokeenlaadinnasta. Siksi on syytä käyttää apuna empiirisiä keinoja osioiden analysoimiseksi. Millman (1974) toteaa myös, että osioanalyysin tuloksia voidaan käyttää apuna puutteellisten osioiden löytämisessä. Kriteerimittaamisen henkeen kuuluu kuitenkin olennaisena sekä Millmanin että Hambletonin (Hambleton et al. 1978) korostama näkökohta, jonka mukaan osioanalyysin tuloksia ei tule käyttää ainoana perusteena osiojoukon parantelemisessa eikä niiden valinnassa. Osioiden valitseminen näiden kriteereiden perusteella saattaa johtaa kokeeseen, jossa osiot eivät ole edustava näyte alueesta vaikeustason ja mitattavien ominaisuuksien suhteen. Kokelaan asema suhteessa hyvin määriteltyyn alueeseen voidaan parhaiten saada selville hänen vastauksistaan osiouniversumista valittuun edustavaan osio-otokseen. Empiirisin menetelmin valitut osiot todennäköisesti ovat vaikeustasoltaan keskimääräisiä ja homogeenisempia kuin kaikki osiot.

Osiota koskevien tilastotietojen käyttö mitätöi satunnaisvalinnan periaatteen, mikä on keskeinen kriteerikokeen ominaisuus. Jos osioita ei valita satunnaisesti, henkilön suoritustason arvio menettää merkityksensä ja koepistemäärän tulkittavuus heikkenee. Osiota koskevia tilastotietoja voidaan kuitenkin käyttää paljastamaan puutteellisia osioita. Jos osiot eivät käyttäydy ennakko-odotusten mukaisesti, ne on syytä ottaa lähempään tarkasteluun. Myös oppilaita voidaan haastatella jotta saataisiin lisävalaistusta asiaan. Osioita koskevat tilastotiedot voivat myöskin auttaa havaitsemaan, missä suhteessa alueen määrittelyä tulisi tarkistaa.

Jos esimerkiksi havaitaan että tietty ryhmä osioita korreloi matalasti muiden osioiden kanssa, voi Millmanin (1974) mukaan olla harkinnan arvoista jakaa alue kahtia. Osioanalyysin tulokset voivat myöskin tukea loogista analyysiä, kun arvioidaan sekä alueen määrittelyä että koetta. Vaikka empiiriset tiedot ovat relevantteja alueen määrittelemiseksi ja arvioimiseksi ja myöskin kokeen määrittelemiseksi ja arvioinniksi, niitä ei tulisi käyttää perusteena kun valitaan osioita kriteerikokeeseen. Se että tässä ei vaadita laajoja empiirisiä esikokeiluja ei merkitse, etteikö kriteerikokeen kehittäminen ole yhtä tiukkaa kuin muidenkin kokeiden laatiminen. Millmanin mukaan se yksinkertaisesti merkitsee että tiukkuus on suurimmaksi osaksi siirretty alueen määrittelyyn. Ikävä kyllä hyvin määriteltyjä aluekuvauksia ei juurikaan ole olemassa.

Kun osioiden voidaan katsoa olevan peräisin tietyltä suhteellisen homogeeniselta sisältöalueelta, voidaan osioiden vaikeustaso- ja erotteluindeksejä tarkastelemalla löytää osioita, jotka vaativat tarkkaa katsastamista. Jos sisällön asiantuntijat eivät löydä osiosta mitään vikaa, se voidaan siirtää takaisin osioaltaaseen. Osioiden vaikeustason yhtäläisyyttä voidaan testata esim. Cochranin Q-testin avulla. Osion vaikeustasoindeksiä voidaan käyttää myös hyödyksi silloin, kun osio on esitetty ennen opetusta ja sen jälkeen. On kuitenkin syytä huomata, että toisin kuin Carver (1974) esittää, suuri ero vaikeustasossa ei välttämättä todista että osio mittaa tavoitetta. Samoin pieni ero voi toki johtua siitä, ettei opetus ole ollut tehokasta, eikä siitä, ettei osio mittaisi opetustavoitetta.

3.4. Osion sisällön ja muodon merkitys

Millman (1978) on pyrkinyt selvittämään osion vaikeustason determinantteja. Hän toteaa, että näennäisesti pienetkin muutokset osion sisällössä ja muodossa voivat aiheuttaa tuntuvia muutoksia osion oikein vastanneiden suhteellisessa määrässä. Normimittaamisessa on osion erottelukyky tärkein osioiden valinta- ja karsintaperuste. Erottelevia osioita tarvitaan tehtäessä vertailevia arvioita kokelaiden keskinäisestä paremmuudesta. Kriteerikokeilla puolestaan pyritään pikemmin absoluuttisiin kuin vertaileviin osaamistasotulkintoihin. Tällöin on syytä huomata, että

päätelmät siitä, kuinka suuren osan kaikista mahdollisista osioista kokelas osaa ratkaista, riippuu suoraan kokeeseen valittujen osioiden vaikeudesta.

Millman (1978) käytti laatimaansa tietokoneohjelmaa tuottamaan 133 osiolomaketta, jotka olivat tyypiltään seuraavanlaisia:

Osiolomake (6007)

Jos {N} koepistemäärää, joiden keskiarvo on {M} ja keskihajonta {S}, jakautuvat normaalisti, {kuinka moni / kuinka monta prosenttia}

pistemääristä olisi {pienempi kuin / suurempi kuin} {x / X + / +} ?

Huom! N = satunnainen (10, 200, 10)
M = satunnainen (10, 100)
S = satunnainen (. 2 M, . 3 M)
x- = satunnainen (M-3 S, M) x- ≠ M
x+ = satunnainen (M, M+ 3 S) x+ # M

Kaarisulut edustavat paikkoja, joissa voidaan arvoja korvata toisilla. Osiolomakkeen avulla voitaisiin tuottaa esim. seuraavat kaksi osiota:

1) Jos 80 pistemäärää, joiden keskiarvo on 47 ja keskihajonta 10, jakautuvat normaalisti, kuinka monta prosenttia pistemääristä olisi pienempiä kuin 26?

2) Jos 140 pistemäärää, joiden keskiarvo on 75 ja keskihajonta 19, jakautuu normaalisti, kuinka moni pistemääristä olisi suurempi kuin 71?

Millman käytti erilaisia kysymystyyppejä (mm. oikein - väärin väittämiä ja yhdistelemistä), erilaista kielellistä muotoilua (erilaisia synonyymisia ilmaisuja, erilaisia symboleja, erilaista sanajärjestystä). Koska koehenkilöitä oli vainyhden yliopiston tilastotieteen alkeisopetusryhmän verran (N = 30), ei tuloksista voi tehdä pitkälle meneviä johtopäätöksiä. Tulokset antavat kuitenkin tukea näkemykselle, että osioiden vaikeustasoa voidaan vaihdella kysymysten erilaisella muotoilulla.

4. VAATIMUSTASOJEN ASETTAMINEN

4.1. Vaatimustasojen asettamisen perusteista

Keskeinen ongelma koulutuksessa ja opetuksessa on hyväksyttävän suoritustason määritteleminen. Tätä kysymystä on tarkasteltu yksityiskohtaisemmin tekijän aikaisemmassa julkaisussa (Takala 1980). Vaatimustason asettamisessa voidaan käyttää apuna sekä kokemusta että aikaisempia tutkimustuloksia. Erityisesti psykomotorisen ja affektiivisen käyttäytymisen arvioinnissa on luontevaa kerätä näytteitä oppilaiden suorituksista ennen opetusta ja sen jälkeen. Nämä voidaan esittää asiantuntijoille arvioitavaksi ilman, että he tietävät mitkä suoritukset kuvaavat tilannetta ennen opetusta ja mitkä sen jälkeen. Samalla tavalla voidaan arvioida esimerkiksi kirjoitelman laadinnan taidon kehittymistä. Kolmas tapa on esittää asiantuntijoille ja joissakin tapauksissa myös oppilaille osa-alueen kuvauksia ja pyytää heitä arvioimaan, mikä olisi hyväksyttävän suorituksen alaraja.

Vähimmäisvaatimustasoa määriteltäessä on kiinnitettävä erittäin paljon huomiota siihen, kuinka vakavia erilaiset virheluokitukset ovat. Päätös siitä, tulisiko oppilaan suoritus hyväksyä vai ei, ei riipu ainoastaan hänen testipistemäärästään suhteessa vaatimustasoon, vaan myöskin siitä mitä menetetään tehtäessä jompikumpi päätös. Kumpi on vakavampi virhe: mittauksen tuloksen perusteella oppilas saa aiheetta edetä eteenpäin tai joutuu aiheetta kertaamaan? Pophamin (1978) mukaan aivokirurgien koulutuksessa etenemisen salliminen puutteellisin tiedoin olisi kohtalokasta. Joissakin muissa tapauksissa taas aiheeton etenemisen estäminen saattaisi olla haitallisempaa kuin aiheeton etenemisen salliminen.

Tekijän aikaisemmasta julkaisusta (Takala 1980) käy selvästi ilmi, että vähimmäisvaatimustason asettaminen perustuu aina viime kädessä harkintaan. Popham (1978) korostaa, että on täysin perusteetonta olettaa, että on olemassa jokin todellinen ja täysin selvä vähimmäisvaatimustaso, kunhan vain pystyisimme sen jollakin konstilla samaan selville. Vähimmäisvaatimustason asettaminen ei kuitenkaan ole mielivaltaista. Erilaisin menettelytavoin voidaan vaatimustaso asettaa käyttäen kulloisessakin tilanteessa todellisten asiantuntijoiden analyyttisiä kykyjä hyväksi par-

haalla mahdollisella tavalla. Tällöin voidaan hyödyntää sekä aikaisempaa tietoa että systemaattisia menettelytapoja. Mm. Nedelsky (1954) ja Ebel (1972) ovat kehittäneet hyvinkin yksityiskohtaisia ja systemaattisia menetelmiä vaatimustason asettamiseksi. Kehittämistyö tällä saralla on vasta alussa. Se, valitaanko suhteellisen yksinkertainen vai suhteellisen monimutkainen tapa vähimmäisvaatimustason asettamiseksi, riippuu mittauksen tarkoituksesta ja sen pohjalla tehtävistä ratkaisuista.

On huomattava, että osioiden vaikeus luonnollisesti vaikuttaa vaatimustason osuvuuteen. On aivan eri asia, asetetaanko 90 % hyväksymisrajaksi helpossa vai vaikeassa kokeessa. Tämän vaikean ongelman ratkaisemisessa on suurta apua kriteerimittauksen tarkasta alueen määrittelystä ja koetäsmennyksestä.

Vaikka joissakin tapauksissa kriteeriviitteinen mittaaminen antaa mielekästä tulkintapohjaa sinänsä, useimmissa tapauksissa normitieto on hyödyllistä. Useimmissa tapauksissa olisi Pophamin (1978) mukaan hyödyllistä tietää, miten saman alueen muissa kouluissa suoriudutaan vastaavasta kokeesta tai mikä on koko maan keskimääräinen suoritustaso. Tieto siitä, että omassa koulussa saavutetaan keskimääräistä heikompia tuloksia, antaa aina aihetta toimenpiteisiin. Sen sijaan tieto siitä, että koulun taso on keskimääräistä parempi, ei välttämättä anna aihetta laakereilla lepäämiseen. Keskimääräistä parempi suoritustaso ei välttämättä vielä ole kyllin hyvä taso.

Jotkut kriteerimittauksen kannattajat katsovat, että normitietojen hankkiminen ja esittäminen eivät ole sopusoinnussa kriteerimittaamisen perusajatusten kanssa. Pophamin (1978) mielestä asia ei ole näin. Hänen mukaansa normitiedot eivät millään tavalla vähennä tarkan aluekuvauksen ja koekuvauksen arvoa, tarpeellisuutta ja hyötyä. Päin vastoin normitiedot auttavat vaatimustason asettamisessa ja auttavat tulkitsemaan sitä, onko tarkasti kuvattu oppilaiden suoritustaso riittävän hyvä.

Miksi yleensä tulisi määritellä hyväksyttävän suoritustason pistemäärä? Jos kriteerikokeen tarkoituksena on arvioida kokelaan tämänhetkistä suoritustasoa, miksi yleensä tarvitaan hyväksyttävää suorituspistemäärää tai standardia? Jos pistemäärän perusteella ei tehdä mitään tärkeätä päätöstä, miksi asetetaan hyväksyttävän suoritustason pistemäärä? Millmanin (1974) mukaan kasvattaja saattaa haluta määritellä hyväksyttävän suorituksen pistemäärän, ei välttämättä koska jokin päätös riippuisi kummalla puolella tätä pistemäärää oppilas on, vaan pikemminkin koska hän haluaa identifioida ja viestiä muille, mitä suoritusta pidetään hyväksyt-

tävänä. Jos koetta käytetään päätöksenteon pohjana, silloin hyväksyttävä pistemäärä voidaan asettaa sen tiedon pohjalta mitä tiedetään erilaisten pistemäärien ja erilaisten päätösvaihtoehtojen välisistä yhteyksistä. Hyvaksyttavan pistemäärän asettamiseksi on erittäin hyödyllistä, jos on olemassa tietoa mitä tapahtuu eritasoisille oppilaille, kun he opiskelevat erilaisissa olosuhteissa.

4.2. Hyväksymisrajan asettamisen menetelmiä

Hyväksyttävän pisterajan (cut-off score, cutting score, passing score) asettamista ovat käsitelleet mm. Millman (1973), Meskauskas (1976) ja Glass (1978). Millman käsittelee artikkelissaan muiden suoritukseen vertaamista, osioiden sisällön analysointia, opetukseen kohdistuvia vaikutuksia, psykologisia ja taloudellisia kuluja sekä arvaamisesta ja osio-otannasta aiheutuvia virheitä. Kuten kirjoittajan aikaisemmassa julkaisussa todettiin (Takala 1980), Glass on vaatimustasoja ja kriteereitä koskevassa artikkelissaan (Glass 1978) erottanut kuusi erilaista menetelmää hyvaksyttavan pisterajan määrittelemisessä. Seuraavassa käsitellään näitä menetelmiä seikkaperäisemmin kuin aikaisemmassa julkaisussa.

4.2.1. Muiden suoritustasoon vertaaminen

Hyväksyttävää vaatimustasoa määriteltäessä on joskus käytetty pohjana tietoa sopivien vertailuryhmien suoritustasosta. Californian high schoolin tutkintojärjestelmässä (California High School Proficiency Examination) käytettiin 50:nnettä persentiiliä yli 16-vuotiaiden päästötutkinnon saavuttamisen kriteerinä (Glass 1978). Toisaalta voidaan käyttää vertailukohtana jo hyväksyttyjen henkilöiden suoritustasoa hyväksymisrajan määrittelemisessä. Heille voidaan esittää koe ja hyväksymisrajaksi asetetaan esimerkiksi kymmenes persentiili (Millman 1973), ts. tullakseen hyväksytyksi varsinaisen kokelaan tulee suoriutua ainakin yhtä hyvin kuin heikoin kymmenes jo aikaisemmin läpäisseistä. Glass katsoo, että turvautuminen normitietoihin on hieman noloa kriteerimittaamisen puolestapuhujille, koska se syntyi juuri osittain normimittaamisen vastapainona. Hambleton (Hambleton,

Swaminatham, Algina & Coulson 1978) onkin esittänyt, että tällainen menettely ei ole sopusoinnussa kriteerimittauksen alkuperäisten perusteiden kanssa. Kuten edellä mainittiin Popham (1978) on kuitenkin todennut, ettei tällainen kielteinen kannanotto ole aiheellinen.

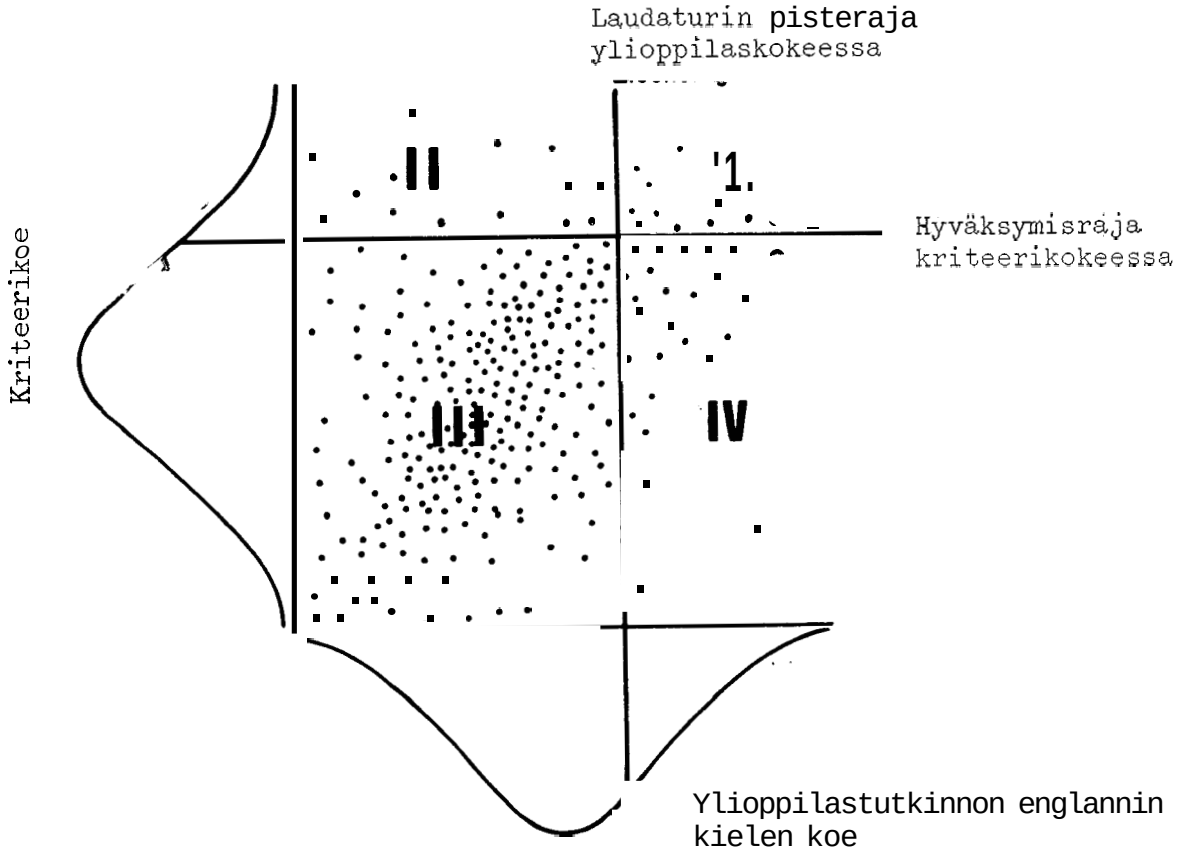
4.2.2. 100 %:sta tinkiminen

Tämä menetelmä ("Counting Backwards from 100 %") on Glassin (Glass 1978) mukaan varsin yleinen. Kutakin tavoitetta mittaamaan laaditaan osio tai mieluummin luonnollisesti useampia kuin yksi osio. Itseasiassa useimmat kokeenlaatijat ovat sitä mieltä, että periaatteessa kaikkien tulisi osata ratkaista osiot (ratkaisuprosentti = 100 %), koska kyseessä ovat keskeiset perustavoitteet. Tätä pidetään kuitenkin yleensä mahdottomana. Täten esimerkiksi opetussuunnitelman laadinnan ja evaluaation kehittämisen uranuurtaja Ralph W. Tyler (Tyler 1973) esittää 85 % tehtävistä oikein hallinnan rajaksi, koska on "varattava pieni marginaali meille itse kullekin sattuville lapsuksille". Glass kritisoi menettelyä mielivaltaiseksi. Joissakin tapauksissa saattaa hyväksyttyjen osuus vaihdella 10-50 %:iin, jos vaatimusraja muuttuu esimerkiksi 95 %:sta 90 %:iin oikeinratkaistuja osioita.

4.2.3. Muuhun kriteeriin vertaaminen

Tätä menetelmää, josta Glass (1978) leikillisesti käyttää nimikettä "Bootstrapping on Other Criterion Scores", ei ole sanottavasti käytetty. Oletetaan, että käsillä oleva kriteerikoe, esim. yliopiston englannin kielen laitoksen valintakoe, esitetään ylioppilastutkinnon englannin kielikokeen suorittaneille. Oletetaan edelleen, että vain siinä laudaturin saaneilla olisi realistisia mahdollisuuksia menestyä englannin kielen yliopisto-opiskelussa. Tällöin voi syntyä kuvion 4 esittämä tilanne.

Koska valintakoe ei varmaankaan korreloi täydellisesti ylioppilastutkinnon kielikokeen kanssa, suuri osa laudaturia heikommin kirjoittaneista todennäköisesti epäonnistuisi valintakokeessa (III), mutta osa läpäisisi (II). Osa laudaturin saaneista epäonnistuisi valintakokeessa (IV) ja osa läpäisisi (I). Valintakokeesaa hyväksyttävän pistemäärän kohtaa ei voida sijoittaa niin, että se osuu täysin yksiin laudaturpiste-



KUVIO 4. Kriteerikokeen ja ulkopuolisen tutkinnon välinen yhteys (Glass 1978).

rajan kanssa. Glassin mielestä kummassakin tapauksessa (sekä kriteerikokeessa että ulkopuolisessa tutkinnossa) käytetään tavallisesti varsin mielivaltaista menettelyä hyväksyttävien määrää ratkaistaessa. "Muusta kriteeristä" ei hänen mielestään siten ole paljonkaan apua.

4.2.4. Osioiden sisällön arviointi

Millman (1973) puhuu tässä yhteydessä osioiden sisällön (item content) analysoinnista ja Glass (1978) minimikompetenssin arvioinnista (judging minimum competence). Minimitason hallinta todetaan yksityisiä osioita tarkastelemalla. Nedelsky (1954), Angoff (1971) ja Ebel (1972) ovat kehitelleet menetelmiä osiotarkastelun pohjalta.

Nedelsky (Nedelsky 1954, 4-7) antaa seuraavanlaiset ohjeet:

Opettajille annettavat ohjeet

Ennen kuin koe pidetään opettajille annetaan seuraavat ohjeet:

Vedä yli kussakin osiossa ne vastausvaihtoehdot, jotka juuri ja juuri viitosen saaneenkin oppilaan tulisi tietää vääriksi. Kirjoita osion vasemmalle puolelle jäljelle jäävien vaihtoehtojen käänteisluku. Jos esim. vedät yli yhden viidestä vaihtoehdosta, koodiksi tulee 1/4.

Alustava sopimus vaatimustasosta

Kun opettajat ovat käyneet läpi 5-6 osiota edellä kuvatulla tavalla, on suositeltavaa, että he pitävät lyhyen neuvottelun ja keskustelevat asettamistaan vaatimustasoista. Saattaa olla myös aiheellista, että he tässä vaiheessa sopivat vakion k alustavasta arvosta (ks. kohtaa hyväksymisraja). Tämän neuvottelun jälkeen kunkin opettajan tulisi jatkaa arviointia omin päin.

Terminologia

Kuvattaessa metodologiaa, jota käytetään heikkoa "viitosta" vastaavaa pistemäärää laskettaessa, käytetään seuraavaa terminologiaa:
(a) Niitä vaihtoehtoja, jotka heikoimmankin viitosen oppilaan tulisi osata hylätä, ja jotka siten ovat houkuttelevia vain nelosen oppilaille, nimitetään nelosreaktioiksi tai -vaihtoehdoiksi (F-responses 1. failure-responses).
(b) Oppilaita, jotka omaavat juuri ja juuri riittävästi tietoa nelosvaihtoehtojen hylkäämiseksi ja joiden täytyy siis valita sattumanvaraisesti jäljelle jääneiden vaihtoehtojen kesken, kutsutaan nelos-viitosoppilaiksi (F-D students).
(c) Nelos-viitosoppilaiden todennäköisintä pistemäärien keskiarvoa nimitetään nelos-viitosoppilaiden arvauspistemääräksi (M_{FD}). Kuten myöhemmin osoitetaan M_{FD} vastaa muiden kuin nelosvaihtoehtojen lukujen käänteislukujen summaa.
(d) M_{FD}:ää vastaavaa todennäköisintä keskihajonnan arvoa merkitään & FD.

On selvää, että "nelos-viitosoppilaat" ovat tilastollinen abstraktio. Sellaista oppilasta, joka osaa hylätä nelosvaihtoehdon mutta valitsee sattumanvaraisesti yhden jäljellejääneistä vaihtoehdoista, ei todennäköisesti ole olemassa. Itse asiassa M_{FD}:ää vastaavia pistemääriä saavat oppilaat, joiden vastaukset vaihtelevat suuresti.

Alin hyväksyttävä pistemäärä

Heikointa viitosta vastaavaa pistemäärää merkitään $M_{FD} + k_{FD}$, jossa M_{FD} on eri opettajien saama M_{FD} ja k on useilla eri perusteilla määriteltävä vakio. Nelos-viitosoppilaille ei ole niinkään tyypillistä varma tieto kuin kyky välttää tiettyjä virhearviointeja. Useimmat nelos-viitospistemääriä käyttäneet opettajat ovat sitä mieltä, että tietämättömyyden puuttumista merkitsevä vaatimustaso on liian lievä ja että siksi hyväkcymicpistemäärän tulisi olla sellainen, että suurin *osa* nelos-viitosoppilaista tulisi hylättyä. Kun k:lle annetaan arvot -1, 0, 1 tai 2, nelos-viitosoppilaista epäonnistuu kokeessa vastaavasti 16 %, 50 %, 84 % ja *98 %*. Järkevä päätös k:n arvosta on mahdollista sen jälkeen kun opettajat ovat valinneet nelosvaihtoehdot, koska tuolloin he pystyvät parem-

min arvioimaan käyttämiensä vaatimustasojen tiukkuutta. Jotta pitäydyttäisiin absoluuttisten vaatimustasojen hengessä, k:n arvo tulisi kuitenkin sopia ennen kuin M_{FD}:n arvot on laskettu ja ehdottomasti ennen kuin oppilaiden pistemäärät tiedetään.

Ehdotetussa tekniikassa on olennaista, että saavutustasolle asetettava vaatimustaso määritellään tutkistelemalla tarkasti yksityisiä osioita. Vain vähäisiä tarkistuksia tulisi tehdä k:n arvoa varioimalla. Syy siihen, että otetaan käyttöön vakio k, mistä aiheutuu tiettyä joustavuutta mutta myös moniselitteisyyttä, on nelosvaihehtojen vaihtelu kahden äärimmäisyyden välillä useimmissa kokeissa: ne ovat joko täysin vääriä ja niiden valinta osoittaa suurta tietämättömyyttä tai kohtalaisen virheellisiä, jolloin niiden valinta merkitsee melko hyväksyttävää suoritustasoa. Jos tietyssä kokeessa on ensisijaisesti täysin vääriä nelosvaihtoehtoja, tätä kokeen erikoispiirrettä voidaan korjata antamalla k:lle korkea arvo. Matala k:n arvo korjaa puolestaan runsasta vain kohtalaisesti väärien nelosvaihtoehtojen esiintymistä kokeessa. On odotettavissa, että useimmissa tapauksissa tarvittava korjaus saadaan lisäämällä tai vähentämällä alustavasti sovitusta k:n arvosta 0.5. Olisi vaikea löytää teoreettisia perusteita niin korkealle k:n arvolle kuin 2, ja toisaalta 0 olisi liian alhainen. Vaihteluväliksi jää näin ollen 0.5 - 7.5, ja k = 1 on hyvä lähtökohta.

Ebelin (Ebel 1972) tekniikka poikkeaa jonkin verran Nedelskyn menetelmästä. Ebelin mukaan hyväksymisrajan (passing score) ilmoittamisella tiettynä prosenttina maksimipistemäärästä, on puutteena se, että tällöin jää vielä paljon puhtaan sattuman varaan. Osiot saattavat olla aiottua vaikeampia, helpompia tai heikommin erottelevia. Kokeen hyväksyttävästi suorittaminen saattaa riippua enemmän kokeen tehtävistä kuin osaamistasosta. Tätä puutetta voidaan kuitenkin vähentää johtamalla hyväksyttävä prosenttiraja (passing percentage) kokeeseen sisältyvien osioiden relevanssin ja vaikeuden subjektiivisen arvioinnin perusteella. Taulukossa3 esitetään neljä relevanssikategoriaa ja kolme vaikeustasokategoriaa sekä odotettu ratkaisuprosenttimäärä kunkin kategorian osioille. Odotetut prosenttiluvut edustavat juuri ja juuri hyväksyttävälle tasolle päässeeltä kokelaalta (minimally qualified, barely passing) odotettavaa tulosta.

TAULUKKO 3. Koeosioiden relevanssi, vaikeus ja odotettu ratkaisuprosentti (Ebel 1972, taulukko 19.7)

Relevanssi-kategoriat	Vaikeustasot		
	Helppo	Keskivaikea	Vaikea
Välttämätön	100 %	–	–
Tärkeä	90 %	70 %	–
Hyväksyttävä	90 %	60 %	40 %
Kyseenalainen	70 %	50 %	30 %

Oletetaan, että 100-osioisessa kokeessa osiot jakautuvat viiden arvioitsijan arviointien mukaan taulukon 4 osoittamalla tavalla. Kertomalla osioiden määrällä odotetut ratkaisuprosentit ja jakamalla summa 500:lla saadaan asianmukainen alin hyväksymispistemäärä.

TAULUKKO 4. Osioiden ominaisuuksien perusteella arvioitu hyväksymisraja (Ebel 1972, Taulukko 19.8)

Osiokategoria	Osioiden määrä	Odotettu ratkaisu-prosentti	Osimäärä x odotettu ratkaisuprosentti
Välttämätön	94	100	9400
Tärkeä			
Helppo	106	90	9540
Keskivaikea	153	70	10710
Hyväksyttävä			
Helppo	24	80	1920
Keskivaikea	49	60	2940
Vaikea	52	40	2080
Kyseenalainen			
Helppo	4	70	280
Keskivaikea	31	50	55
Vaikea	7	30	210

Hyväksymisraja on täten 37135/500 = 74.27 eli 74 %.

Angoffin (Angoff 1971, 514-515) mukaan pistemääriä halutaan joskus kuvata asteikolla, jossa tietyillä pistemäärillä on ennaltamäärätty normatiivinen merkitys - normatiivinen siinä mielessä, että ne kuvaavat todellista suoritustasoa mutta myös siinä mielessä, että ne kuvaavat asetettuja vaatimustasoja. Usein menetellään empiirisesti niin, että hyväksyttäväksi pisterajaksi asetetaan skaalattu pistemäärä 70 ja läpäistään esim. 65 % kokelaista. Tällöin 65 %:ia vastaava raakapistemäärä määräytyy empiirisen aineiston perusteella. Skaalaa ei Angoffin mukaan kuitenkaan tarvitse perustaa empiiriseen aineistoon, vaan se voidaan johtaa tarkastelemalla yksittäisiä osioita huolellisesti, josta päädytään alimpaan hyväksyttävään pistemäärään (lowest acceptable, or passing raw score). Systemaattinen menettely voisi olla seuraavanlainen: arvioitsija pitää mielessä hypoteettista, juuri ja juuri hyväksyttävälle tasolle yltävää kokelasta (minimally acceptable person), käy läpi kokeen osio osiolta ja päättää osaisiko tällainen kokelas ratkaista osiot oikein. Jos osattavaksi arvioidusta osiosta annetaan yksi piste ja osaamattomasta nolla pistettä, osioiden summapistemäärä edustaa juuri ja juuri hyväksyttävällä tasolla olevan kokelaan raakapistemäärää. Kun joukko arvioitsijoita tekisi itsenäisesti tällaiset arviot, olisi mahdollista saada yhteinen käsitys skaalamuunnoksesta ilman että koetta pidetään. Haluttaessa voidaan tätä arviota verrata empiiriseen tulokseen.

Glassin (1978, 248) mukaan tällaisissa menetelmissä on kaksi ongelmaa: 1) Pystyvätkö arvioitsijat tekemään tällaisia arvioita johdonmukaisesti ja luotettavasti? 2) Mikä on käsitteen "vähimmäisvaatimustaso" (minimal competence) loogis-psykologinen asema?

Ensimmäistä ongelmaa on tutkittu varsin vähän. Meskauskas ja Webster (1975) käyttivät tutkimuksessaan Nedelskyn menetelmää. Kuusi arvioitsijaa, joiden vastuulla oli lääkäreiden toimivaltuuden uusimiseen liittyvän tutkinnon laatiminen, vaihtelivat arvioinneissaan niin paljon, että hyväksymispistemäärä vaihteli välillä 36 % - 80 % tehtävistä ratkaistava oikein. Andrew ja Hecht (1976) vertailivat empiirisesti Nedelskyn ja Ebelin menetelmiä. Kahdeksan arvioitsijaa, jotka olivat laatineet 180 osiota eräaseen monivalintakokeeseen, asettivat vaatimustason ensin Nedelskyn menetelmällä ja eri kerralla Ebelin metodilla. Muistin ja järjestyksen osuutta pyrittiin huolellisesti eliminoimaan kokeellisen asetelman vakiomenetelmillä. Ebelin menetelmällä alimmaksi hyväksyttäväksi pistemääräksi saatiin 68 % tehtävistä oikein ja Nedelskyn menetelmällä 49 % oikein. Tämä on kovin suuri ero ja sen vakavuutta lisää se tosiasia, että Glassin

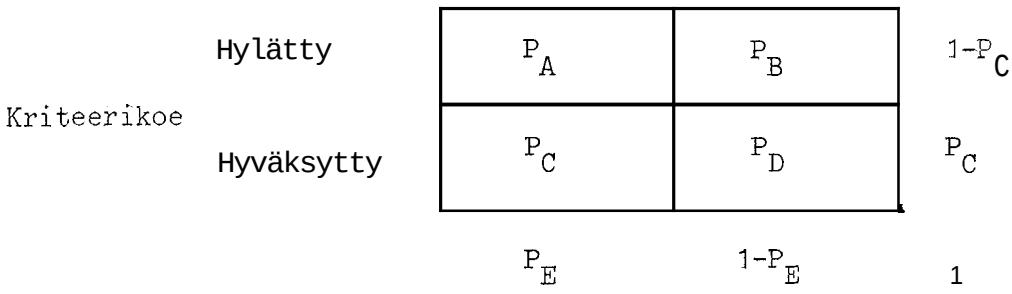
mukaan 95 % kokelaista olisi läpäissyt Nedelskyn metodia noudatettaessa ja vain 50 % Ebelin metodin mukaisesti toimittaessa.

Glass (Glass 1978) asettaa kyseenalaiseksi myös vähimmäissuoritustason loogisen ja psykologisen statuksen. Hän siteeraa Trevantia, joka v. 1927 kirjoitti, ettei toksikologiassakaan käsite "pienin tappava määrä" ole loogisesti perusteltavissa. Aivan kuten elämässä on monia ilmiöitä, joille ei ole maksimia (esim. korkeushypyn maailmaennätys, henkilön saksan sanavaraston määrä), ei kasvatuksessa eikä edes ammattisuuntautuneessa koulutuksessa voida Glassin mukaan perustellusti ilmoittaa minimivaatimustasoa.

4.2.5. Päätösteoreettiset menetelmät

Kriteerimittaamisen matemaattiset kehittämis- ja soveltamismahdollisuudet ovat saaneet jo tässä vaiheessa paljon huomiota osakseen. Kriteerimittaamisessa on kiinnitetty paljon huomiota juuri hyväksymispisterajan (cut-off score) tarkkuuden selvittelemiseen. Täten mm. Hambleton ja Novick (Hambleton ja Novick 1973) ovat esittäneet, että keskeinen ongelma kriteerimittauksessa on oppilaan luokitteleminen yhteen useista toisensa poissulkevista asianhallinnan luokista. Tyypillinen tehtävä on luokitella oppilaat jompaan kumpaan kahdesta toisensa poissulkevasta luokasta - asian hallitseviin ja puutteellisesti hallitseviin (masters vs. nonmasters). Kriteerimittaamisen validiteetti ja reliabiliteetti määritellään tällaisen luokittelun 1. päätöksen johdonmukaisuutena kahdessa rinnakkaismittarissa tai kahdella eri mittauskerralla. Ongelmana on todeta, ylittääkö oppilaan todellinen hallintataso asetetun vaatimustason.

Päätöksentekoteoriaa sovelletaan kriteerimittauksessa tyypillisesti seuraavalla tavalla: Kokelaat luokitellaan kahteen luokkaan jonkin ulkopuolisen kriteerin perusteella, esim. ylioppilastutkinnossa hyväksyttyihin vs. hylättyihin, opiskelemaan hyväksyttyihin vs. karsiutuneisiin. Näiden kahden kategorian suhteellisia henkilömääriä merkitään P_E ja $1-P_E$. Jos näille henkilöille oli esitetty etukäteen kriteerikoe (esim. "preliminäärikoe" lukion toisella luokalla tai oppilaitoksen päättötutkinto) ja siinä on asetettu hyväksymisraja C_x, syntyy nelikenttä (ks. kuvio 5).



KUVIO 5. Kriteerikokeen ja ulkopuolisen kriteerin perusteella tehtyjen ratkaisujen yhteyksien todennäköisyydet (Glass 1978).

P_A osoittaa virheellisten negatiivisten tapausten eli virheellisesti hylättyjen ("false negatives") suhteellista määrää, ts. niiden osuutta, jotka hylätään kriteerikokeessa mutta jotka läpäisevät ulkopuolisen kriteerin. P_D osoittaa virheellisten positiivisten tapausten eli virheellisesti hyväksyttyjen ("false positives") osuutta, ts. niiden osuutta, jotka läpäisevät kriteerikokeen mutta tulevat hylätyksi ulkopuolisessa kriteerissä. Ulkopuolisen kriteerin hyväksymisraja oletetaan yleensä annetuksi. Sen sijaan päätösteoreettisessa menettelyssä annetaan kriteerikokeen hyväksymisrajan vaihdella, joten nelikentän ruutujen suhteelliset osuudet voivat myös vaihdella. On mahdollista johtaa hyväksymispistemäärän C_x hyvien seurausten eli oikeiden luokitusten (P_B ja P_C) ja huonojen seurausten eli virheellisten luokitusten (P_A ja P_D) yhdistetty funktio, jossa minimoidaan virheitä ja maksimoidaan oikeita luokituksia. Jos virheellisten hyväksymisten ja hylkäämisten "kustannukset" ovat yhtäsuuret, virheiden minimoiminen voidaan ilmaista seuraavasti:

$$f(C_x) = (P_A + P_D)/ (P_B + P_C)$$

Jos taas hylkäämiskustannukset ovat α ja hyväksymiskustannukset β, mimimoimisfunktio voidaan ilmaista seuraavasti:

$$f(C_x) = (\alpha P_A + \beta P_D)/ P_B + P_C)$$

Funktio on täten sensitiivinen alfan ja betan arvoille, joiden määrittely Glassin mukaan on mielivaltaista ja saattaa vaihdella tuntuvasti henkilöstä toiseen. Tärkeitä kysymyksiä ovat Glassin mukaan: "Mistä hyväksymispisteraja C_x tulee?" ja "Miten tiettyä hyväksymispisterajaa voidaan perustella

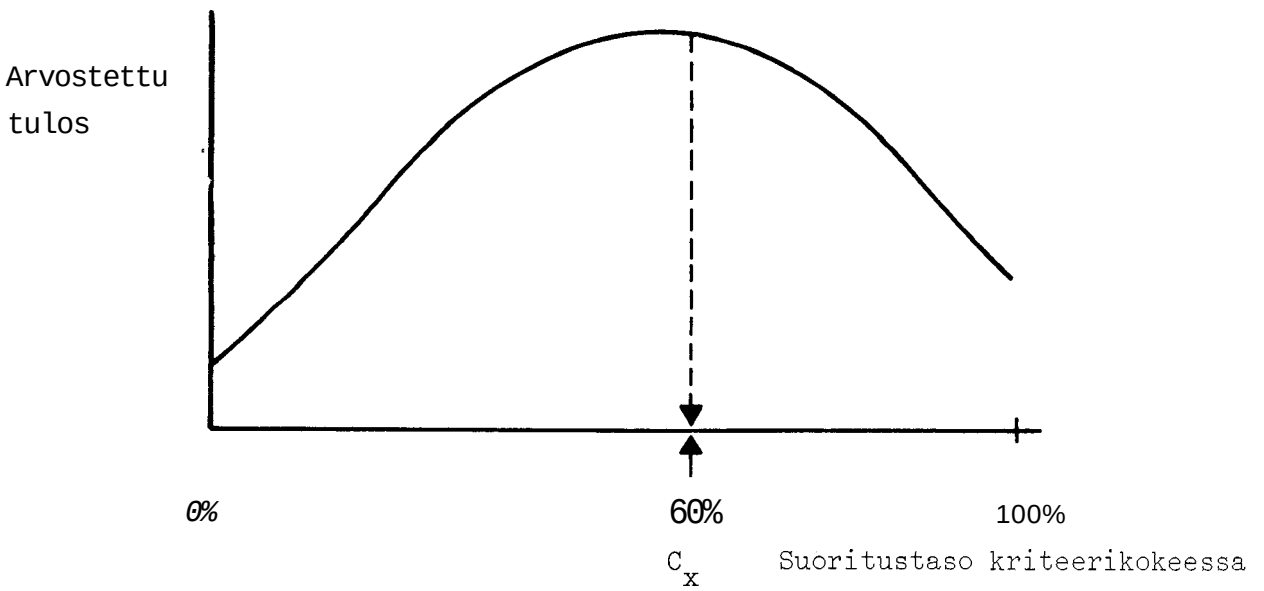
muihin mahdollisiin rajoihin verrattuna?" Päätösteorian menetelmät ja tilastomatematiikka tuottavat kyllä oikeita tuloksia, mikäli toiminnan premissit ovat oikeita. Glassin mukaan käytännössä menetellään kovin kritiikittömästi ja tiedostamatta hyväksymisrajan asettamiseen liittyviä periaatteellisia vaikeuksia.

4.2.6. Operaatiotutkimukseen perustuvat menetelmät

Tätä menetelmää hyväksymispisterjan määrittelemiseksi kutsutaan operaatiotutkimusmetodiksi, koska se perustuu operaatiotutkimuksessa käytettävään menettelytapaan jonkin arvostetun seikan maksimoimiseksi etsimällä optimikohta matemaattiselta käyrältä tai kuviolta.

Block (1972) sovelsi operaatiotutkimuksen strategiaa pyrkiessään määrittelemään rationaalisesti perusteltavaa hyväksymispisterajaa kriteerikokeille. Tätä voidaan tutkia mm. siten, että satunnaisesti muodostetuille ryhmille opetetaan jotakin asiaa, kunnes he saavuttavat erilaisia osaamistasoja, esim. 30 %, 40 %, 50 % .. 85 %, 90 %, 95 %, 100 %. Tämän lisäksi oppilaille esitetään jokin muu ulkopuolinen mittari, jota mittaa arvostettua tulosta (esim. muistissasäilymistä tai siirtovaikutusta). Tämän jälkeen piirretään kuvio, jossa kriteerikokeen tulokset suhteutetaan ulkopuolisen kriteerin tuloksiin (ks. kuvio 6). Se kriteerikokeen taso, jolla ulkopuolinen arvostettu tulos maksimoituu, on rationaalinen hyväksymisraja. On helppo huomata, että tämä metodi ei ratkaise tyydyttävästi hyväksymisrajan ongelmaa, ellei käyrä ole ei-monotoninen, ts. se ei saa tasaisesti nousta vaan jossakin kohdalla 0 %:n ja 100 %:n välillä käyrän nousun tulee pysähtyä ja mahdollisesti kääntyä alaspäin. Ilman käyrän taipumista arvostettu tulos tulee maksimoiduksi kriteerikokeen täydellisellä ratkaisulla (100 %), joka Glassin mukaan on joko triviaalinen tai mahdoton vaatimustaso. Glass arvelee, että kuvion 6 kaltaiset käyrät ovat harvinaisia silloin, kun kriteerikoe ja arvostettu tulos mittaavat kumpikin kognitiivista aluetta.

Toinen mahdollisuus on ottaa arvostetuksi tulokseksi sellainen muuttuja, joka korreloi negatiivisesti kriteerikokeen tulosten kanssa. Saattaa olla, että oppilaat suhtautuvat kielteisemmin johonkin asiaan mitä pitempään heidän tulee sitä opiskella. Arvostettuna tuloksena voisi siis olla asennoituminen opiskeltavaa ainetta kohtaan.



KUVIO 6. Hypoteettinen yhteys kriteerikokeen ja arvostetun tuloksen välillä (Glass 1978)

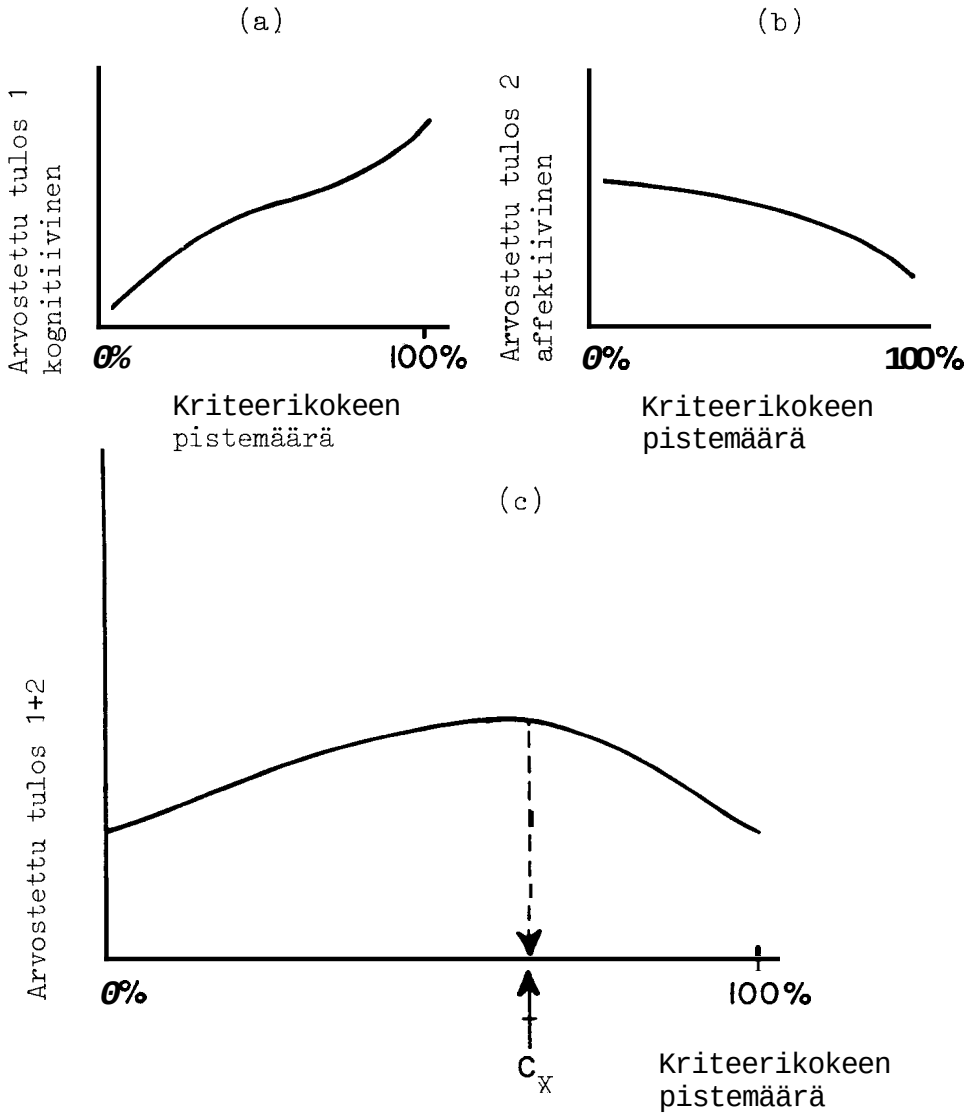
Kuviossa 7 (a) kuvaa tilannetta, jossa arvostettuna tuloksena on jokin muu kognitiivinen tulos ja (b) tilannetta, jossa arvostettuna tuloksena on suhtautuminen ko. oppiaineeseen, ja (c) kuvaa tilannetta, jossa arvostetut tulokset on yhdistetty.

Menetelmän avulla näyttäisi olevan mahdollista löytää kriteerikokeelle hyväksymisraja, joka maksimoi arvostetut tulokset (1+2) yhdessä. Rationaalisuus on Glassin mukaan kuitenkin näennäistä, koska molempia arvostettuja tuloksia on painotettu yhtä paljon, vaikka tämä on vain eräs erikoistapaus yleisestä kaavasta:

Yhdistetty arvostettu tulos = α(tulos 1) + β (tulos 2)

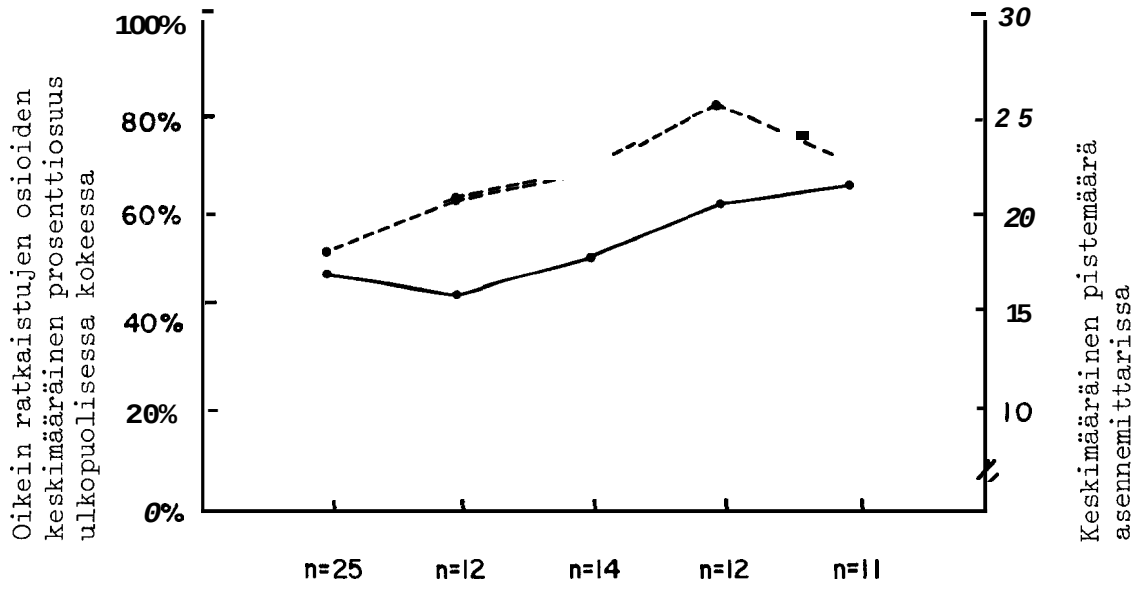
Jos ensimmäistä arvostettua tulosta painotettaisiin kahdella ja toista yhdellä, hyväksymispisteraja siirtyisi kuviossa tuntuvasti oikealle.

Mastery learning-järjestelmää kehitellyt Bloomin oppilas Block (Block 1972) on soveltanut operaatiotutkimusmenetelmää. Kahdeksasluokkalaisille (N=91) opetettiin matriisialgebran alkeita. Oppilaat jaettiin viiteen ryhmään: kontrolliryhmä, ja ryhmät, joilta vaadittiin kriteeripohjaisissa formatiivisissa kokeissa: 65 %, 75 %, 85 % ja 95 % asian hallintaa. Ryhmät (kontrolliryhmää lukuunottamatta) eivät saaneet edetä, ennen kuin vaadittava suoritustaso oli saavutettu. Arvostettua tulosta mitattiin ulkopuoli-



KUVIO 7. Erilliset ja yhdistetyt yhteydet kriteerikokeen pistemäärän ja kahden eri arvostetun tuloksen välillä (Glass 1978)

sella kokeella, joka esitettiin kun vaadittava suoritustaso oli saavutettu. Opetuksen jälkeen esitettiin myös algebraan asennoitumista mittaava kysely-lomake. Kuviossa 8 kuvataan Blockin tutkimuksen tuloksia.

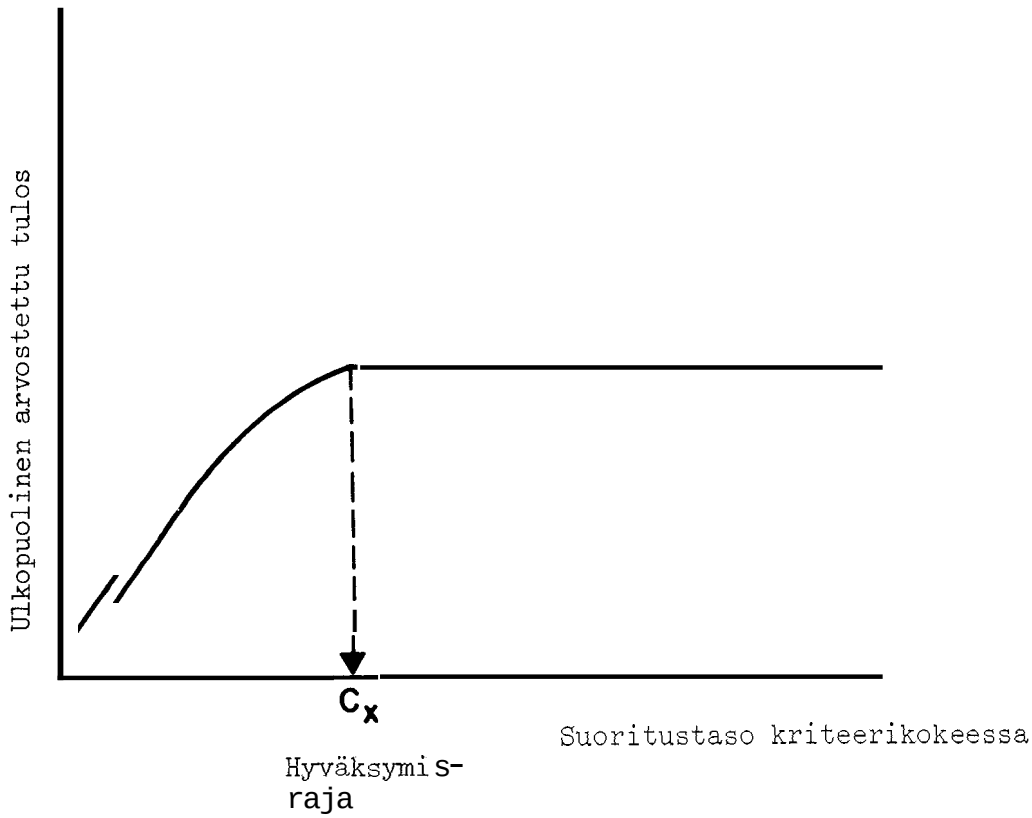


KUVIO 8. Asennoituminen algebraa kohtaan ja algebran oppimistulokset erilaisen vaatimustason sisältyneissä ryhmissä (Glass 1978).

Algebran oppimiskäyrä on selvästi monotoninen (tasaisesti nouseva) mutta asennekäyrä taipuu alaspäin. **Block** (Block 1972) toteaa, että 95 %:n vaatimustason ylläpitäminen maksimoi kognitiiviset oppimistulokset, mutta 85 %:n taso maksimoi affektiiviset "oppimistulokset". Glass (Glass 1978) pitää Blockin tuloksia kyseenalaisina ja katsoo, että hyväksymispistemäärän asettamisen mielivaltaisuus on vähentynyt pikemminkin näennäisesti kuin todellisuudessa.

Glassin mukaan (Glass 1978) on olemassa vaatimattomampi versio operaatiotutkimuksen käytöstä hyväksymistasoja määriteltäessä. Oletetaan, ettei kriteerikokeen tietyn tason tultua ylitettyä tapahdu yhtään edistystä ulkopuolisessa arvostetussa tuloksessa (ks. kuvio 9). Tämä antaisi perusteita asettaa tämä kohta vaatimustasoksi (C_x), kotka tämän jälkeen edistyminen kriteerikokeessa ei tuo mitään positiivista tulosta ulkopuolisessa kriteerissä.

Glassin mukaan (Glass 1978) Robert Glaser on sanonut, että 6-8 -vuotiaat lapset eivät tyypillisesti saavuta 70 %:a korkeampaa tasoa yksinumeroisten lukujen yhteenlaskussa. Lisäopetus on paljolti tuloksetonta. Parempi suoritustaso tulee iän mukana. Glassin mukaan myös John Tukey on todennut,



KWIO 9. Hypoteettinen yhteys kriteerikokeen tuloksen ja arvostetun ulkopuolisen tuloksen välillä (Glass 1978).

etteivät puhelinvälittäjät millään harjoituksen määrällä pääse 98 %:a tarkempaan suoritukseen. Tämä osoittaisi, että on olemassa psykofyysisiä rajoja tarkkaavaisuudelle ja tarkkuudelle. Vaikka nämä ovat houkuttelevia mahdollisuuksia hyväksymisrajan asettamista ajatellen, Glassin mukaan niiden anti on monimutkainen ja pitkälti epäselvä.

Popham, Scriven, Block ja Hambleton ovat esittäneet vastineensa Glassin kritiikkiin. Vastineiden yhteisenä piirteenä voidaan esittää, ettei vaatimustasojen asettamista voida pitää mielivaltaisena vaan perusteltavaan harkintaan perustuvana. Vastineista on esitetty laajahko yhteenveto tekijän aikaisemmassa julkaisussa (Takala 1980), joten tässä ei ole aihetta toistaa siinä esi'tettyä.

4.2.7. Vaatimustasojen seurausvaikutuksia

4.2.7.1. Opiskelua koskevat seuraukset

Millmanin (Millman 1973) mukaan vaatimustasoja asetettaessa on kiinnitettävä erityistä huomiota siihen vaikutukseen, joka niillä on oppilaiden vastaiseen opiskeluun. Jos vaatimustaso on liian alhainen, oppilaille tulee eteen sellaista ainesta, jota he eivät ole valmiit omaksumaan. Jos vaatimustaso on liian kova, opiskelun tehokkuus kärsii, kun käytetään tarpeettomasti aikaa kertaamiseen ja tukiopetukseen. Keskeinen kysymys onkin, millainen vaatimustaso maksimoi oppimistuloksia (educational benefits).

Edellä on jo viitattu Blockin tutkimukseen, jossa 85 %:n vaatimustaso näytti maksimoivan oppimistuloksia, kun otettiin huomioon seka kognitiiviset että affektiiviset vaikutukset. Bormuth (1971) on tutkinut cloze-tekniikalla määritellyn oppimateriaalin vaikeustason optimaalista yhteyttä oppimistuloksiin.

Millman (Millman 1973) suosittelee seuraavia periaatteita noudatettaviksi vaatimustasoja asetettaessa, kun tarkasteltavana on niiden vaikutukset opetukseen. Jos loogisen analyysin ym. tiedon perusteella tietyt tiedot ja taidot katsotaan välttämättömiksi ennakkotiedoiksi myöhemmälle opetukselle, vaatimustaso on syytä asettaa korkeaksi. Matalampi vaatimustaso on mahdollinen, jos kyseessä ei ole kompleksisen taidon tietty osalenkki. Niissä kohdissa, joissa ei ole kyse välttämättömistä ennakkotiedoista tai -taidoista, ei vaatimustasoja tarvinne asettaa. Hambleton (Hambleton, Swaminathan, Algina & Coulson 1978) on käsitellyt tarkemmin ongelmia ja matemaattisia perusteita, jotka liittyvät kokelaiden sijoittamiseen erilaisiin hallintoluokkiin (allocation of examinees to mastery states).

4.2.7.2. Psykologiset ja taloudelliset kustannukset

Millmanin (Millman 19731 mukaan muiden tekijöiden pysyessä samana tulisi käyttää matalaa vaatimustasoa, kun tukiopetusohjelman psykologiset ja taloudelliset kustannukset ovat suhteellisen korkeat. Toisin sanoen, suorituksia tulisi hylätä suhteellisesti vähemmän, jos nämä kustannukset ovat korkeat. Tällaisia "kustannuksia" voivat olla mm. alentunut motivaatio ja kyllästyminen, itsetunnon heikkeneminen, tehoton ajankäyttö ja kor-

keat taloudelliset kulut. Korkeaa vaatimustasoa voidaan käyttää, kun ko. kustannukset ovat alhaiset ja kun aiheeton hyväksyminen itse saattaa aiheuttaa ikäviä seurauksia (vaikeus seurata opetusta jne.)

5. AFFEKTIIVISEN ALUEEN MITTAAMINEN KRITEERIPOHJAISIN MITTAVÄLINEIN

Affektiivisen alueen mittaamista on yleensä harrastettu huomattavasti vähemmän kuin kognitiivisen alueen. Tämä johtunee ennen kaikkea affektiivisen alueen mittauksen vaikeudesta. Vaikka affektiivisten reaktioiden mittaaminen on vaikeata, kriteeriviitteinen mittaaminen antaa Pophamin (1978) mielestä runsaasti vihjeitä myös sen kehittämiseksi.

Popham (1978) esittää, että ensiksi on tarpeen selvittää mitattavaa affektiivista ominaisuutta, jotta sen luonne ymmärrettäisiin tarkemmin. Tämän jälkeen on hyödyllistä kuvitella mielessään henkilö, joka olisi tällaisen ominaisuuden läpitunkema. Vastapainoksi on hyvä kuvitella henkilö, jolla ei olisi hiventäkään kyseistä ominaisuutta. Seuraavassa vaiheessa yritetään kuvitella mahdollisimman monia tilanteita, missä affektiivisen ominaisuuden erot voisivat käydä ilmi. Tässä on Pophamin mukaan syytä käyttää mielikuvitusta ja vapaata assosiointia. Vasta tämän jälkeen on syytä miettiä, mikä mittaustavoista olisi validi ja käytännöllinen. Erityisesti tulee huolehtia siitä, ettei mittari johda sosiaalisesti toivottaviin vastauksiin.

Pophamin (1978) mukaan tavallisesti affektiivisen alueen mittaaminen kohdistuu pikemminkin ryhmiin kuin yksilöihin. Täten usein arvioidaan, kuinka myönteisesti tai kielteisesti oppilasryhmä suhtautuu tietynlaiseen opetusohjelmaan. Mittauksessa käytetään tyypillisesti kolmenlaista menettelyä: 1) oppilaat vastaavat suoriin kysymyksiin tai väittämiin, jolloin tarvitaan vain vähän tulkintaa (low-inference self-report), 2) oppilaat vastaavat kysymyksiin ja väittämiin, mutta niistä ei käy yhtä selkeästi ilmi mitä asioita mitataan ja täten tuloksia tulee tulkita (high-inference self-report), 3) havainnoidaan oppilaiden käyttäytymistä luonnollisissa tai varta vasten luoduissa tilanteissa.

Kun affektiivista aluetta mitataan, on Pophamin (1978) mukaan tärkeätä määritellä mitattava alue ja mittaustapa yhtä tarkasti kuin tiedollista aluetta mitattaessa. Täten tulee ensiksi yleisesti kuvata mitä affektiivista piirrettä tai käyttäytymistä halutaan mitata. Tämän jälkeen esitetään esimerkki mittaustavasta, joka sisältää vastausohjeet, esitettävien ärsykkeiden ominaisuuden tarkan kuvauksen, reaktioiden rekisteröimisen tarkan kuvauksen ja reaktioiden pisteistyksen määrittelyn. Näin menetellen voidaan myös affektiivisen alueen mittausta edistää ja tulosten tulkittavuutta parantaa.

Seuraava esimerkki on Suomen oloihin mukailtu versio Pophamin (1978, 200-203) esittämästä havaintoesimerkistä.

Tarkoitus: selvittää missä määrin henkilöt suhtautuvat ihmisiin yksilöinä (ei ryhmän jäseninä]

Yleiskuvaus

Henkilöt osoittavat taipumusta arvioida ihmisiä yksilöinä valitsemalla saman vastausvaihtoehdon - arvioidessaan tietyn yksityisen henkilön moraalisesti problemaattista käyttäytymistä - riippumatta siitä, ilmoitetaanko ko. henkilön kuuluvan johonkin etniseen vähemmistöryhmään, yhteiskunnalliseen kerrostumaan tai ryhmään tms. vai ei. Kussakin versiossa henkilö ilmoittaa missä määrin hän hyväksyy tai ei hyväksy ko. henkilön käytöstä. Yhdessä versiossa kuvataan tarkemmin identifioimattomien henkilöiden toimintaa, toisessa ilmoitetaan henkilöiden "tausta". Jos henkilöiden vastauksissa ei ole merkitseviä eroja näissä kahdessa eri versiossa, tulkitaan tuloksen osoittavan, että vastaajat arvioivat ihmisiä yksilöinä eikä niinkään jonkin ryhmän jäseninä.

Esimerkkitehtävä

Ohjeet: Lue seuraavat kaksi kertomusta ja ilmoita miten suhtaudut siihen toimintaan, mikä alleviivatussa kohdassa kuvataan. Valitse yksi annetuista vaihtoehdoista ja merkitse se vastauslomakkeelle. Ei ole oikeita tai vääriä vastauksia, joten vastaa rehellisesti miltä sinusta tuntuu. Voit vastata nimettömästi.

~~Versio-A~~ (henkilön taustaa ei ilmoiteta):

1. Kymmenvuotias Matti löysi 50 markan setelin kaupan lattialta. Hän mietti, **voisiko** hän pitää rahan vai tulisiko hänen antaa se esim. kassaneidille. Koska Matti tarvitsi pyöräänsä uuden kumin, hän päätti pitää rahan,

A. hyväksyn täysin menettelyn
B. hyväksyn menettelyn
C. paheksun menettelyä
D. paheksun jyrkästi menettelyä

~~Versio-B~~ (henkilön tausta ilmoitetaan)

1. Kymmenvuotias mustalaispoika Matti.../Kymmenvuotias saamelaispoika Matti/...

Ärsykepiirteiden määrittely

1. Kussakin kertomuksessa kuvataan lyhyesti se tilanne, jossa henkilöllä on ratkaistavana moraalinen ongelma, ja hänen menettelytapansa.
2. Kussakin moraalisessa ongelmassa on kyseessä ristiriita, jonka toisena puolena on yleisluonteinen sosiaalinen arvo (rehellisyys, lainkunnioitus tms.) ja toisena tietyn henkilökohtaisen tarpeen tyydyttäminen (rahan, vaatteiden, ajan tarve tms.).
3. Kussakin tilannekuvauksessa ovat seuraavat osat:
   a) Nimeltä mainittu fiktiivinen henkilö, jonka ikä ilmoitetaan. Iän tulee olla lähellä kohdejoukon ikää. Nimi ei saa olla jonkin ryhmän tyypillinen nimi. Esim. Oula-nimeä ei saa käyttää, koska sen yleisesti katsotaan olevan lappalaisnimi.
   b) Fyysisen ympäristön tulee olla vastaajille tuttu: koti, koulu, lähiympäristö jne.
   c) Ongelmatilanteen tulee edustaa sellaisia tilanteita, joihin vastaajat saattavat itsekin joutua.
   d) Esitellään kaksi eri menettelytapaa ja menettelyn syy(t). Kertomuksen henkilön valitsema menettelytapa alleviivataan. Menettelytapa edustaa moraalisesti kyseenalaista vaihtoehtoa. Se kuvastaa oman edun ajamista pikemmin kuin yleisiä arvoja. Jotta vastaajille ei tulisi automaattinen taipumus "paheksua" tai "paheksua jyrkästi" kertomusten henkilöiden käytöstä, on syytä liittää mukaan joukko tilanteita, joissa käyttäytyminen on moraalisesti hyväksyttävää, ja se heijastaa yleisiä yhteiskunnallisia arvoja. Kahta pisteitettävää tehtävää kohti tulee olla vähintään yksi täytetehtävä, jota ei siis pisteistetä. Tehtävien järjestys satunnaistetaan.

4. Kustakin tilanteesta laaditaan kaksi versiota. Ne ovat muussa suhteessa identtisiä paitsi että versiossa B ilmoitetaan henkilön kuuluvan johonkin ryhmään, jota kohtaan voidaan olettaa tunnettavan ennakkoluuloja, A-versiossa ei henkilön taustaa ilmoiteta. tilanteet esitetään kummassakin versiossa samassa järjestyksessä.
   Vastaajien tulee tuntea taustaryhmät joko henkilökohtaisesti tai joukkotiedotusvälineiden kautta.
5. Koko mittarissa voidaan viitata yhteen tai useampaan ryhmään, mutta kussakin tehtävässä vain yhteen ryhmään. Jos mittari rajataan vain yhteen ryhmään, tulisi pisteistettäviä tehtäviä olla ainakin viisi. Jos viitataan yhtä useampaa ryhmää, tulee kutakin ryhmää käsitellä ainakin kolmessa tehtävässä.
6. Versio A esitetään satunnaisesti valitulle henkilöjoukolle (esim. luokan puoliskolle) ja versio B lopuille.
7. Tilannekuvausten kielen täytyy olla hyvin yksinkertaista (lauserakenteet ja sanasto), jotta heikommankin lukutaidon omaavat ymmärtävät mistä on kyse. Todella heikoille lukijoille tai nuorille lapsille tehtävät voidaan lukea ääneen.

Vastauspiirteiden määrittely

1. Vastaajat valitsevat yhden neljästä vaihtoehdosta, joka heijastaa heidän suhtautumistaan valittuun menettelytapaan.
2. Vastausvaihtoehdot ovat samat kaikissa ongelmatilanteissa.
   A. hyväksyn täysin menettelyn
   B. hyväksyn menettelyn
   C. paheksun menettelyä
   D. paheksun jyrkästi menettelyä
3. Arviointi perustuu versioihin A ja B annettujen vastausten vertailemiseen, jolloin menetellään seuraavasti:
   a) Aineisto jaetaan kahtia: A-versiot ja B-versiot. Kummassakin tulee olla yhtä paljon vastaajia. Ylimääräiset tapaukset poistetaan satunnaisperiaatteen mukaisesti. Kumpikin versio pisteistetään erikseen.
   b) Vain varsinaiset tehtävät pisteistetään (täytetehtävät jätetään huomioonottamatta) seuraavasti:

      hyväksyn täysin menettelyn: 4 pistettä
      hyväksyn menettelyn: 3 pistettä
      paheksun menettelyä: 2 pistettä
      paheksun jyrkästi menettelyä: 3 piste

c) A-ryhmän ja B-ryhmän kokonaispistemääriä verrataan keskenään käyttäen jompaa kumpaa seuraavista menetelmistä siitä riippuen käytettiinkö B-versiossa yhtä tai yhtä useampaa ryhmää:

(1) Jos B-versiossa on käytetty vain yhtä taustaryhmää:
Laske yhteen kaikki vastaajien pistemäärät erikseen ryhmissä A ja B. Näiden kahden summapistemäärän vertailu osoittaa suhtautuvatko vastaajat henkilöihin yksilöinä taustaryhmistä riippumatta. Pelkästään sattuman vaikutuksesta voidaan odottaa vähäisiä eroja, mutta tuntuvasti alemmat pistemäärät B-versiossa osoittavat, että vastaajat saattavat arvioida käyttäytymistä ankarammin, jos henkilö kuuluu johonkin "vähemmistöryhmään". Tilastollista merkitsevyyttä voidaan testata tilastotieteen oppikirjoissa esitetyille vakiomenetelmille (esim. t-testi tai Mann-Whitney U-testi).

(2) Jos B-versiossa on käytetty useampaa kuin yhtä taustaryhmää:
Kutakin ryhmää käsittelevien tehtävien pisteet summataan erikseen seka versiossa A että B (huom! versioon A saadaan ryhmäkohtaiset osapistemäärät version B vastaavien tehtävien perusteella). Jos B-versiossa on käytetty esim. taustaryhminä mustalaisia ja suomen-ruotsalaisia, lasketaan kummallekin ryhmälle pistemäärät erikseen kummassakin versiossa. Vertailu perustuu osapistemäärien vertailuun versioiden A ja B kesken samaan tapaan kuin kohdassa (1). Eri taustaryhmien vertailun lisäksi voidaan A- ja B-ryhmiä vertailla myös siten, että lasketaan summapistemäärät yli kaikkien tehtävien, kuten kohdassa (1) meneteltiin.

Edellä olevasta esimerkistä voitaneen päätellä, että kriteerimittaamisen periaatteiden soveltaminen myös affektiivisella alueella toisi tervetullutta jäntevyyttä tälle vaikealle ja paljolti laiminlyödylle alueelle.

6. RELIABILITEETTI KRITEERIMITTAAMISESSA

6.3. Yleisiä näkökohtia

Kaksi keskeistä ominaisuutta, joiden suhteen kokeita useimmiten arvioidaan, ovat reliabiliteetti ja validiteetti. Vaikka normi- ja kriteerimittaamisessa on reliabiliteetti- ja validiteettikäsitteillä paljon yhteisiä piirteitä, on myös selviä eroja.

Reliabiliteettina pidetään yleisesti sitä, kuinka johdonmukaisia tuloksia koe antaa. Mittaa koe mitä tahansa, se on sitä luotettavampi, mitä johdonmukaisempia tuloksia se antaa. Kriteerimittaamisen tulee täyttää reliabiliteetin vaatimukset siinä missä normimittauksen. Pophamin (1978) mukaan kokeen reliabiliteettia arvioidaan tyypillisesti neljän tunnusluvun avulla jotka mittaavat pysyvyyttä, ekvivalenssia, ekvivalenssia ja pysyvyyttä sekä sisäistä konsistenssia. Sisäistä konsistenssia käytetään useimmin erityisesti siitä syystä, että se on helpoimmin laskettavissa. Ennen kuin näitä neljää reliabiliteetin mittaustapaa käsitellään tarkemmin, on syytä käsitellä kahta kriteerimittauksen erityisongelmaa: vähäistä varianssia sekä kokeen käyttötapaa ja siitä tehtäviä johtopäätöksiä.

6.2. Varianssin merkitys reliabiliteetin estimoinnissa

Popham ja Husek (1969) sekä Hambleton ja Novick *(3973)* ovat esittäneet, että perinteelliset tavat arvioida näitä ominaisuuksia ovat todennäköisesti vähemmän soveliaita kokeille, joilla on huolellisesti spesifioitu sisältöalue ja joille on määritelty menettelytavat asianmukaisten koeosioiden tuottamiseksi. Koska Hambletonin ja Novickin mukaan kriteerikokeen laatija ei ole kiinnostunut erottelemaan oppilaita, ei pyritä valitsemaan osioita siten että tuotettaisiin mahdollisimman tehokkaasti erotteleva koe ja täten varianssi on tyypillisesti vähäinen, mikäli opetus on ollut tehokasta.

Kriteerikokeet esitetään tavallisesti joko ennen lyhyitä opetusjaksoja tai niiden jälkeen. Täten ei ole hämmästyttävää, että esi- ja jälkikokeissa testipistemäärät jakautuvat homogeenisesti, mutta ne sijoittuvat toisaalta saavutusskaalojen alapäähän ja toisaalta niiden yläpäähän. Perinteellisestä testiteoriasta tiedetään hyvin, että kun koepistemäärien varianssit ovat suppeat, korrelaatioihin perustuvat reliabiliteetin ja validiteetin estimaatit ovat matalia. Täten näyttää selvältä, että klassisia menetelmiä reliabiliteetin ja validiteetin arvioimiseksi on tulkittava varovaisemmin tai ne on jopa hylättävä analysoitaessa kriteerikokeita.

Reliabiliteetin estimaatit perinteisessä normimittaamisessa perustuvat korrelaatioanalyysiin. Muuttujissa tulee esiintyä tuntuvaa varianssia jotta korrelaatioiden laskeminen olisi mielekästä. Usein saatetaan virheellisesti luulla, että kriteerimittaamisessa todetaan vain hyvin vähän varianssia oppilaiden pistemäärissä. Tämä olettamushan liittyy läheisesti mm. tavoiteoppimisen teoriaan, mutta Pophamin (1978)mukaan käytännöllisesti katsoen aina esiintyy varianssia oppimissuorituksissa.

6.3. Kokeen käyttötavan ja siitä tehtävien johtopäätösten merkitys reliabiliteetin estimoinnille

Kokeen tuloksista tehtävien päätöksien luonne vaikuttaa myös reliabiliteetin arviointimenettelyyn. Normikokeen tuloksia käytetään lähinnä apuna tehtäessä yksilöitä koskevia ratkaisuja. Käytännössä normikokeen tuloksia on käytetty myöskin tehtäessä ryhmiä koskevia päätöksiä. Popham (1978) korostaa kuitenkin, että koe, joka antaa luotettavia tietoja ryhmän tulosten pohjalta tehtäville päätöksille, ei välttämättä anna luotettavaa pohjaa yksilöitä koskeville ratkaisuille. Sitä vastoin koe, joka antaa luotettavaa pohjaa yksilöitä koskeville päätöksille, on luotettava myöskin ryhmää koskevia päätöksiä tehtäessä.

Jos kriteerikokeen tuloksia käytetään yksilöitä koskevien ratkaisujen tekemiseen, tulee tietää, onko olemassa vain kaksi vai useampia kuin kaksi vaihtoehtoa. Voidaanko esimerkiksi kokeen tulosten perusteella järjestää hyväksyttävän suoritustason alle jääneille oppilaille useampi kuin yksi tukiopetusohjelma? Tätä kysymystä on käsitellyt myös Glaser (1976) hyvin mielenkiintoisella tavalla.

Perinteellisessa normimittauksessa reliabiliteettia ja validiteettia tarkasteltaessa lähdettiin tyypillisesti yksilöä koskevien päätösten pohjalta. Kriteerikokeen tuloksia käytetään kuitenkin usein tehtäessä päätöksiä (1) vaihtoehtoisista opetusohjelmista, (2) yksilöistä tilanteessa, jossa on kaksi vaihtoehtoa, ja (3) yksilöistä tilanteessa, joissa on useampia kuin kaksi vaihtoehtoa. Kriteerimittaamisessa reliabiliteettia ja validiteettia on tulkittava siten, että kiinnitetään huomiota seka pistemäärien johdonmukaisuuteen että niiden pohjalta tehtävien päätösten johdonmukaisuuteen.

6.4. Erilaisia lähestymistapoja kriteerikokeen reliabiliteetin estioiminnissa

Livingston (1972) on ehdottanut perinteellisestä poikkeavan reliabiliteetin arviointimenetelmää. Hänen lähtökohtanaan on, että kriteerimittaamisen tarkoituksena on diskriminoida kunkin kokelaan estimoitu aluepistemäärä hyväksymispistemäärästä. Livingstonin menetelmässä ei olla kiinnostuneita keskimääräisestä aluepistemäärästä, kuten klassisessa testiteoriassa, vaan määritellään variaatio estimoiduissa aluepistemäärissä ja niiden variaatio hyväksymispistemäärän ympärillä. Livingstonin estimaatti antaa korkeampia arvoja kuin klassinen reliabiliteettiestimaatti. Mitä etäämmällä ryhmän aluepistemäärän keskiarvo on hyväksymispisterajasta, sitä luotettavampia pistemäärien sanotaan olevan.

Hambleton ja Novick (1978) ovat eri mieltä Livingstonin kanssa kriteerimittaamisen tarkoituksesta. Heidän mielestään yksityisen kokelaan aluepistemäärän poikkeama hyväksymispistemäärästä ei ole läheskään yhtä tärkeätä kuin se, luokitellaanko kokelas hyväksymispistemäärän samalle puolelle saman kokeen rinnakkaisversioissa tai sen uudelleen esittämisessä.

Harris (1972) on todennut, että mittauksen keskivirhe on sama riippumatta siitä, mitä reliabiliteetin estioimintimenetelmää käytetään. Siksi Hambletonin (Hambleton et al. 1978) mukaan kriteerimittaamisessa ei ole syytä hylätä kaikkia klassisen testiteorian käsitteitä. Mittauksen keskivirhe on siten eräs, Hambletonin ym. mukaan, melko konservatiivinen, menetelmä asettaa luottamusrajoja (confidence bands) aluepistemäärien estimaattien ympärille.

Brennan ja Kane (1977) ovat esittäneet, että virheen neliön menetysfunktiolla (squared error loss function) Livingstonin tapaan on etuna, että se on sensitiivinen virheiden suuruudelle, mutta haittana, että se on sensitiivinen kaikille mittausvirheille mukaan lukien sellaiset virheet, jotka eivät johda kokelaiden virheluokituksiin. Valinta tulee perustaa käytännön vaatimuksiin. Kynnysmenetysfunktio (threshold loss function), jota Hambleton ja Novick suosittelevat käytettävän, on paikallaan silloin, kun on olemassa selvä ja terävä hyväksymiskohta ja kun kaikilla virheluokituksilla on lähes samanlainen vaikutus. Virheen neliön menestysfunktiota voidaan Brennanin ja Kanen mukaan suositella, kun ko. olettamukset eivät pidä paikkaansa.

6.5. Aluepistemäärien estimaattien reliabiliteetti

Hambleton (Hambleton et al. 1978) toteaa, että silloin kun koepistemäärissä esiintyy varianssia, on mahdollista estimoida kriteerikokeen mittauksen keskivirhe. Lord ja Novick (1968) ovat todenneet, että vaikka kokeen reliabiliteettiestimaatit vaihtelevat henkilöjoukosta toiseen, mittauksen keskivirhe pysyy yleensä muuttumattomana henkilöotoksesta toiseen. Kun on käytettävissä aidosti paralleeliset kokeet (strictly parallel tests, ks. Konttinen 1981), voidaan käyttää perinteellisiä menetelmiä mittauksen keskivirheen arvioimiseksi. Koska kriteerimittareissa usein esitetään satunnaisotos osioaltaasta, kyseessä on nimellisesti paralleeliset kokeet (randomly or nominally parallel tests), joten ei voida soveltaa aidosti rinnakkaisten testien olettamuksia. Cronbach (Cronbach et al. 1972) on yleistettävyysteoriassaan käsitellyt erilaisia mittausvirheiden tyyppejä. Niitä on esitellyt modernin testiteorian yleisesityksessään myös Konttinen (1981).

Kuten edellä todettiin Millman (1974) ehdotti käytettävän binomimallia aluepistemäärien estimaattien luotettavuusrajojen määrittämiseen. Hambletonin (Hambleton et al. 1978) mukaan tällä menettelyllä on seuraavat edut: 1) Virheen estimaatti on aluepistemäärän funktio. 2) Se on konservatiivisempi kuin mittauksen keskivirheen antama virheen estimaatti. 3) On mahdollista tutkia kokeen pituuden vaikutusta estimaattien tarkkuuteen. 4) Estimaatti on suhteellisen helppo laskea.

Hambleton et al. (1978) toteavat, että binomimallin ohella, jossa käytetään hyväksi havaittuja pistemääriä, voidaan käyttää bayesilaista metodia, jossa on etukäteen määriteltävä aluepistemäärän tiheysfunktio subjektiivisesti.

6.6. Johtopäätösten reliabiliteetti

Carver (1970) on ehdottanut käytettäväksi kahta menetelmää kriteerimittareiden reliabiliteettia arvioitaessa. Toisessa menetelmässä on sama koe esitettävä kahdelle vertailukelpoiselle ryhmälle ja tämän jälkeen verrataan,kuinka monta prosenttia kokelaista luokiteltiin asian hallitseviin (masters) kummassakin kokeessa. Toisessa menetelmässä esitetään kaksi rinnakkaiskoetta samalle ryhmälle ja verrataan kuinka monta prosenttia luokiteltiin asian hallitseviksi kummassakin kokeessa. Koetta pidetään sitä reliaabelimpana mitä lähempänä toisiaan prosenttiluvut ovat. Carver ei siis pidä korrelaatiomenetelmää sopivana, koska hänen mielestään reliabiliteetti riippuu toistettavuudesta (replicability), mutta toistettavuus ei ole riippuvainen varianssista. Carverin suosittelemat menetelmät perustuvat jakautumien replikoitavuuteen kun taas perinteelliset menetelmät perustuvat yksityisten pistemäärien jakautumiin. Hambleton (Hambleton et al. 1978) on sitä mieltä, että Carverin kriteerit antavat vain hyvin heikon näytön kriteerimittauksen reliabiliteetista, ts. Carverin ehdot ovat välttämättömiä mutteivat riittäviä osoittamaan kokeen reliabiliteettia.

Carverin ehdotusten tilalle Hambleton ja Novick (1973) ovat ehdottaneet, että hallintaluokittelupäätösten reliabiliteettia tulisi arvioida sen perusteella, kuinka johdonmukaisia päätöksiä tehdään saman kokeen kahden esittämisen tai rinnakkaiskokeiden perusteella. Reliabiliteettiindeksi kuvastaisi sitä, kuinka suuri osa päätöksistä käy yksiin. Tämä on Hambletonin (Hambleton et al. 1978) mielestä intuitiivisesti hyvän tuntuinen indeksi, joka on myös helppo laskea. Sillä on kuitenkin heikkoutena se, että se ei ota huomioon sattumalta yhteenkäyviä luokitteluja. Siksi he ehdottavat käytettäväksi Cohenin (Cohen 1960) kerrointa k, joka ottaa huomioon useita päätöksentekoon vaikuttavia tekijöitä: hyväksymispistemäärän, kokelasryhmän heterogeenisuuden ja hallintaluokituksessa käytettävän menetelmän.

Shavelson et al. (1972), Hambleton ja Novick (1973) ja Swaminathan et al. (1975) ovat korostaneet sitä, että reliabiliteettitieto on ilmoitettava kunkin mitattavan tavoitteen tai osa-alueen osalta erikseen olipa kyseessä osa-aluepistemäärä tai hallintaluokittelu.

Mikäli opetusohjelma todella olisi ollut erittäin tehokas, josta olisi seurauksena vähäinen pistemäärien hajonta, joudutaan kokeen reliabiliteetin arvioimisessa käyttämään muita kuin korrelaatiotekniikoita. Eräs mahdollisuus on pitää tietyin välein kaksi loppukoetta ja jakaa oppilaat neljään ryhmään sen perusteella, olivatko he mediaanin alapuolella vai yläpuolella kummassakin kokeessa. Luokittelun konsistenssin määrää voidaan tämän jälkeen arvioida käyttämällä phi-kerrointa, joka teoreettisesti voi vaihdella välillä -1.0 - +1.0. Käytännössä phi-kerroin on tavallisesti korkeintaan .55. Toinen tapa arvioida kokeen reliabiliteettia olisi ilmoittaa, kuinka monta prosenttia oppilaista säilytti asemansa samanlaisena mediaanin suhteen kummassakin kokeessa.

Edellä kuvatut reliabiliteetin estimointikeinot ovat vähemmän tarkkoja kuin monimutkaiset tilastolliset analyysit, mutta joissakin tilanteissa ne saattavat olla riittäviä. Popham (1978) korostaa, että käytämme mitä tahansa analysointitekniikkaa arvioimaan kriteerikokeen reliabiliteettia, on välttämätöntä, että oppilaiden suorituksissa esiintyy vähintään jonkin verran varianssia tai muuten analyysien tulokset ovat suurelta osin vailla mieltä. Koetäsmennyksessä on selvästi jotain pielessä, mikäli sen pohjalta tehdyt tehtävät ovat joko niin helppoja, että kaikki osaavat ratkaista kaikki tehtävät oikein, tai niin vaikeita, ettei kukaan oppilas osaa ratkaista yhtään niistä. Poikkeuksena on tietenkin sellainen tilanne, jossa koe pidetään ennen opetusta, jolloin kaikki saavat nolla pistettä, ja opetuksen jälkeen, jolloin jotkut tai kaikki saavat maksimipistemäärän.

Lopuksi on syytä kiinnittää huomiota Millmanin (1974) mainitsemaan seikkaan, joka liittyy hallintaluokittelun ongelmiin. Kun on päätetty millä tavalla esimerkiksi hyväksymis- ja hylkäämisraja toteutetaan, voidaan yksinkertaisesti laskea yhdenmukaisuuden prosenttimäärä, ts. kuinka suuren osan kummatkin kokeet sijoittavat samaan luokkaan. Jos käytetään hyvin alhaista hyväksymispistemäärää, ja kaikki kokelaat helposti ylittävät tämän arvon, koetta pidetään hyvin luotettavana vaikka rinnakkaistestien pistemäärissä saattaa esiintyä huomattavaakin vaihtelua. Johtopäätösten reliabiliteetti on korkea ja tämä indeksi osoittaakin päätösten johdonmukaisuutta. Hyväksymispistemäärä vaikuttaa täten reliabiliteettiindeksin kokoon. Koska yhdenmukaisuuden prosentuaalinen osuus riippuu

päätössäännöstä (ja kokeen suorittajien kyvystä suhteessa kriittisiin kokeen arvoihin), reliabiliteetti-indeksiä tulisi täydentää tekemällä selkoa päätössäännöstä. Swaminathan ym. (Swaminathan et al. 1977) ovat esittäneet kaavan, jolla arviointien yhdenmukaisuutta voidaan korjata niin että otetaan huomioon arviointien satunnainen yhdenmukaisuus.

7. VALIDITEETTI KRITEERIMITTAAMISESSA

Kriteerimittaamisessa validiteettina voidaan Millmanin (1974) mukaan parhaiten pitää testipistemääristä tehtyjen johtopäätösten tarkkuutta. Mitkä johtopäätökset voidaan asianmukaisemmin tehdä kriteerikokeen perusteella? Kuten aikaisemmin esitettiin, asianmukaisin johtopäätös kriteerikokeesta koskee kokelaan suoritustasoa. Tämä johtopäätös koskee oppilaan asemaa suhteessa aikaisemmin määriteltyyn alueeseen. Millmanin mielestä tätä voidaan parhaiten arvioida analysoimalla loogisesti alueen määritystä, osioiden tuottamisen järjestelmää ja yksityisiä osioita.

Kriteerimittaamisessa on Pophamin (1978) mukaan kiinnitettävä huomiota kolmeen validiteettiaspektiin: 1) kuvauksen validiteettiin, jossa yritetään selvittää, missä määrin kriteerikoe todella mittaa sitä, mitä koekuvaus väittää sen mittaavan, 2) funktionaaliseen eli käyttövaliditeettiin, jossa selvitetään, missä määrin kriteerikoe vastaa alkuperäistä tarkoitustaan, ja 3) mitattavan osa-alueen valinnan validiteettiin, jossa selvitetään, kuinka onnistunut käyttäytymisalueen valinta on ollut. Kaikkien kriteerikokeiden tulee täyttää kuvauksen ja osa-alueen valinnan validiteetin vaatimukset. Käytön validiteetin tärkeys riippuu kokeen käyttötavasta.

Kuvauksen validiteetilla (descriptive validity) Popham tarkoittaa suurin piirtein samaa kuin perinteellisellä sisällön validiteetilla (content validity), mutta pitää sitä parempana terminä yleistettävyytensä vuoksi. Perinteisessä normikokeessa ei koskaan anneta mitään systemaattisesti johdettua kvantitatiivista sisällön validiteetti-indeksiä. Sen sijaan kriteerimittaamisessa käytetään usein asiantuntijoita arvioimaan, missä määrin koeosiot vastaavat koetäsmennystä. Toinen mahdollisuus arvioida koetäsmennyksen kuvausvaliditeettia on pyytää muutamia asiantuntijoita kuta-

kin laatimaan esim. kolme osiota ja tämän jälkeen tarkastella, kuinka johdonmukaisia osiot ovat keskenään. Yksimielisyys voidaan ilmaista tavallisena prosenttilukuna. Kriteerikokeen kuvauksen validiteetista on aina syytä laatia lyhyt selostus, jossa selvitetään menettelytapa ja saadut numeeriset indeksit.

Kuvauksen validiteetti on välttämätön edellytys muille validiteetin muodoille. Jos emme ole varmoja kokeen kuvauksen validiteetista, emme voi mielekkäästi tulkita koepistemääriä. Kuvauksen validiteetti voidaan arvioida toisaalta pyytämällä asiantuntijoita arvioimaan, missä määrin osa-alueen määrittely rajaa riittävän tarkasti koeosioiden luonteen ja toisaalta missä määrin laaditut koeosiot ovat sopusoinnussa osa-alueen määrittelyn kanssa,

Cronbach (1971) ja Messick (1975) ovat analysoineet validiteetin ongelmia. Messick toteaa, että sisällön validiteetissa on keskeisenä ongelmana se, että se kohdistuu ensisijaisesti koeversioihin eikä koepistemääriin, mittareihin pikemmin kuin mittauksiin. Empiirisissä mittauksissa tehdään johtopäätöksiä pistemääristä ja ne heijastavat henkilöiden reaktioita. Kokeen sisältö on tärkeä näkökohta, mutta Messickin mukaan se ei ole samaa kuin validiteetti. Messickin mukaan sitä voitaisiin kutsua "sisällön relevanssiksi" tai "sisällön edustavuudeksi", mutta ei "sisällön validiteetiksi", koska se ei anna pohjaa vastausten ja pistemäärien tulkinnalle.

Hambletonin (Hambleton et al. 1978) mukaan sisällön validiteetti on kokeen ominaisuus. Se ei vaihtele sanottavasti henkilöjoukosta toiseen eikä ajan mukana myöskään. Sen sijaan koepistemäärien tulkinnan validiteetti vaihtelee tilanteesta toiseen. Näin ollen kokeen sisällön validiteettia koskeva tieto ei ole riittävä varmistamaan koepistemäärien tulkinnan validiteettia.

Funktionaalisella eli käyttövaliditeetilla (functional validity) Popham tarkoittaa suurin piirtein samaa kuin perinteellisellä tavalla arvioida validiteetti korreloimalla kokeen tulos ulkopuoliseen kriteeriin. Kriteerimittauksen tuloksia voidaan käyttää moniin tarkoituksiin. Jotkut näistä voivat edellyttää ulkopuolista kriteeriä, mutta eivät kaikki. Siksi käsite "käyttövaliditeetti" on yleisempi kuin "kriteerivaliditeetti" (criterion-related validity). Pophamin mielestä ei käyttövaliditeetin lisäämiseksi voida tinkiä kuvauksen validiteetista. Kuvauksen validiteetti on välttämätön edellytys käyttövaliditeetille.

Kriteerikokeen käyttövaliditeettia voidaan arvioida empiirisin keinoin. Tämä on Pophamin (1978) mukaan kuitenkin suhteellisen harvoin tarpeellista, koska useimpia käyttötarkoituksia varten riittää kuvauksen validiteetti, ts. tarkka kuvaus siitä, mikä on oppilaan asema tarkoin määritellyllä käyttäytymisalueella.

Jos tietyltä alueelta voitaisiin esittää hyvin laaja osiomäärä useille henkilöille. silloin n-osiota sisältävän kokeen validiteetti-indeksi voisi Millmanin (1974) mukaan esittää sitä, missä määrin kunkin henkilön oikein ratkaistujen osioiden prosenttimäärä hyvin laajassa osiomäärässä vastaa hänen pistemääräänsä pienemmässä n-osiota sisältävässä kokeessa. Jos laskettaisiin korrelaatiot lyhyemmän ja pidemmän kokeen osioiden välillä, tällöin kriteerikokeen validiteetti vastaisi Millmanin (1974) mukaan perinteellisen testiteorian reliabiliteetti-indeksiä (havaittujen ja todellisten pistemäärien välinen korrelaatio], mikä vuorostaan on testin validiteetin yläraja. Ellei toisin sanoen oteta huomioon otantavariaatiota, teoriassa kriteerikokeen ja samasta alueesta laaditun hyvin laajan kokeen välisen korrelaation tulisi olla korkeampi kuin kriteerikokeen ja minkä tahansa muun havaittavan kriteerin välinen korrelaatio.

Millmanin (1974)mukaan voidaan myös suorittaa ennusteita kokeen yhteydestä kriteerimuuttujiin (esim. pistemäärien erot ennen ja jälkeen opetusta seka opettajan arviot). Tällaiset ennusteet kuuluvat olennaisena osana erottelukokeiden laadintaan ja validointiin. Millmanin mukaan kaikki tällainen empiirinen näyttö on hyödyllistä kriteerikokeen validoimiseksi vain sikäli, että oletetut yhteydet ovat järkeviä, koska evidenssi on osoitus sekä kokeen laadusta että sen teorian laadusta johon ennusteet perustuivat. Voidaan olla väärässä sen suhteen miten osioiden pitäisi käyttäytyä, miten menestyksellistä opetus oli tai millaiset korrelaatiot koepistemäärien välillä tulisi olla, Sitä, ettei saada empiirista vahvistusta kriteerikokeen validiteetista, ei tule tulkita lopullisena näyttönä kriteerikokeen epävalidiudesta.

Mitattavan osa-alueen valinnan validiteetissa (domain-selection validity) on keskeisenä kysymyksenä tulosten yleistettävyys. On valittava sellainen osa-alue, jonka tulokset on yleistettävissä mahdollisimman monelle muullekin osa-alueelle. Osa-alueen valinnan validiteetti muistuttaa Pophamin (1978) mukaan jossakin määrin perinteellistä käsitevaliditeettia (construct validity). Käsitevaliditeetti on kuitenkin teoreettisempi käsite kuin osa-alueen valinnan validiteetti.

Osa-alueen valinnan validiteetissa olemme kiinnostuneita siitä, missä määrin osa-alueen valinta antaa kuvan jonkin yleisemmän tavoitteen saavut-

tamisesta, Kuinka hyvin valitun osa-alueen sisältö ja mittaustapa ovat yleistettävissä koskemaan laajempaa käyttäytymistavoitetta? Osa-alueen valinnan validiteettia voidaan Pophamin (1978) mukaan selvittää empiirisesti laatimalla osioita vaihtoehtoisten aluetäsmennysten pohjalta ja vertaamalla saavutettuja tuloksia keskenään. Helpompi tapa on kuitenkin käyttää asiantuntijoita osa-alueiden valinnassa ja kuvata tarkasti noudatettuja menettelytapoja.

Millmanin (1974) mukaan on olemassa myös muita empiirisiä menetelmiä kriteerikokeen käsitevaliditeetin arvioimiseksi. Tällöin on olemassa hypoteesi kriteerikokeen psykometrisistä ominaisuuksista. Kokeen laatijalla on teoria siitä miten pistemäärien tulisi käyttäytyä. Validointimenettely on tällöin sellainen, että selvitetään missä määrin ennustettu ja empiirinen yhteys vastaavat toisiaan. Yksi tällainen ennustettu yhteys Harrisin (1974) mukaan olisi, "että kaikki osiot ovat yhtä vaikeita oppilaalle ja että todennäköisyys = 1, että hän läpäisisi satunnaisesti valitun osion mikäli hän on läpäissyt jonkun muun saman osiolomakkeen osion". Jotkut kokeenlaatijat saattavat toisaalta haluta määritellä alueensa laajemmin kuin mm. Harris niin, että yllämainittu vastaushomogeenisuus ei ole järkevä tai todennäköinen vaatimus. Toinen ennuste saattaisi olla, etta kokeen osiot mittaavat hierarkkista taitojärjestelmää niin, että tietyn osion läpäisseiden oppilaiden tulisi osata suorittaa myös osio, joka mittaa edellisen osion edellyttämää ennakkotietoa, mutta päinvastoin ei voisi olla. Tätä ennustetta varten voidaan laskea Yulen Q, joka päinvastoin kuin phi-kerroin on riippumaton kahden osion vaikeustasosta.

Kokeen sisällön validiteettia koskevat pulmat ovat keskeisiä silloin kun koetta laaditaan. Kokeen sisällön validiteetti myös vaikuttaa koepistemäärien tulkinnan validiteettiin. Koepistemäärien käsitevaliditeetin selvittäminen parissa olisi Hambletonin (Hambleton et al. 1978) mukaan tehtävä työtä, jotta kokeen tulkinnan validiteetti voitaisiin varmistaa. Konttinen (1981) esittelee yksityiskohtaisesti erilaisiin mittaustilanteisiin ja -tarkoituksiin liittyvien päätelmien validiteettia ja niihin liittyviä mittausvirheitä.

8. KRITEERIMITTAAMISEN KÄYTÄNNÖN SOVELLUKSIA

8.1. Yleisiä näkökohtia

Kuten useaan kertaan on edellä todettu, kriteerimittaamisen keskeisimpiä anteja on oppilaan suoritustason tarkka kuvaaminen. Tästä syystä siitä voi olla paljon apua seka opetukselle että oppimistulosten mittaamiselle ja arvioinnille.

Kriteerimittaamiseen olennaisesti kuuluva käyttäytymisalueiden tarkka kuvaus saattaa huomattavasti helpottaa myös opetussuunnitelmaa koskevia ratkaisuja, koska se sisältää tarkan kuvauksen taidoista ja niiden osataidoista. Tarkka tavoitekuvaus on eduksi myös opetuksen suunnittelemiselle ja toteuttamiselle. Kun oppilaat ovat tietoisia opetuksen tavoitteista, he yleensä saavuttavat parempia tuloksia kuin jos he eivät tiedä tavoitteista.

Kriteerimittaamisen avulla voidaan myöskin arvioida opetusohjelmien ja koulujärjestelmän tehokkuutta. Vaikka kriteerimittaaminen ei ole samaa kuin opetuksen evaluointi, kokeet näyttelevät huomattavaa osaa opetusohjelman laadun arvioinnissa. Laajamittaisissa evaluaatio-ohjelmissa on järkevää käyttää otantaa. Viime aikoina on kehitetty ns. matriisiotantamenetelmä, jossa suoritetaan seka koehenkilö- että osio-otantaa. Laajasta osiomäärästä laaditaan useita koeversioita, jotka esitetään luokassa eri oppilaille. Täten kullekin oppilaalle tulee vastattavaksi vain suhteellisen vähäinen määrä osioita, mutta kuitenkin saadaan testattua suuri määrä osioita. Tällaista menettelyä sovellettiin Kasvatustieteiden tutkimuslaitoksessa, kun keväällä 1979 kerättiin ns. peruskoulun ensimmäisen tilannekartoituksen aineisto.

Aikaisemmin tehtiin selvä ero kriteerikokeiden ja normikokeiden välillä. Kriteerikokeet vaativat hyvin tarkasti määritellyn kuvauksen suoritustehtäväpopulaatiosta, josta koeosiot voidaan valita satunnaisesti, ja ne antavat tarkimman kuvauksen oppilaan statuksesta. Sitä vastoin normikokeet laaditaan maksimoimaan erotteluja yksilöiden tai ryhmien välillä. Seuraavassa käsitellään sovellutuksia, jotka vaativat suoritustason mittaamista tai yksilöiden ja ryhmien erottelua.

8.2. Tarveanalyysit

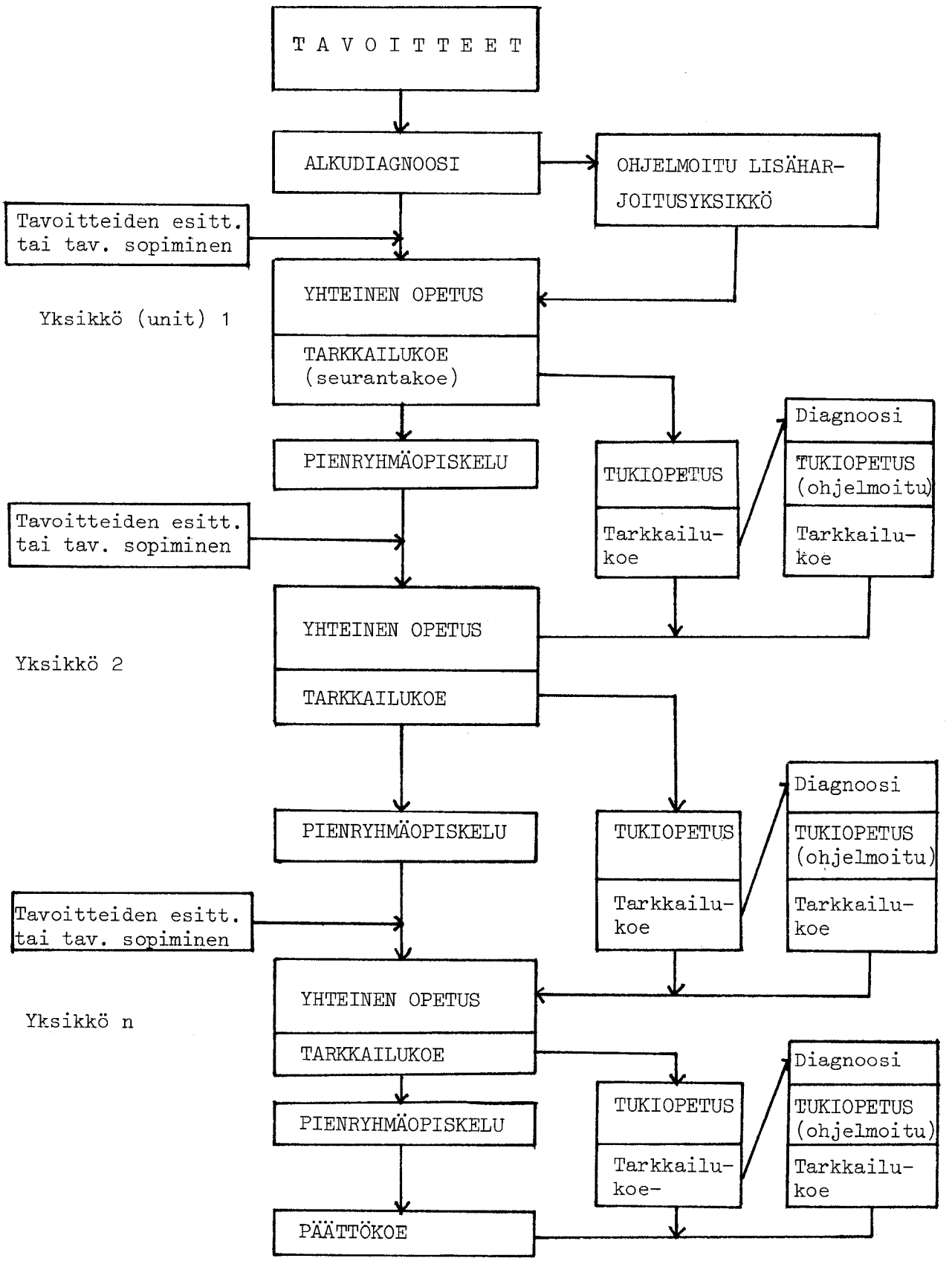
Yhteiskunta ja koulut huolehtivat kasvatuksen ja opetuksen prioriteettien asettamisesta. Tarve on eräs kriteeri, jonka avulla valitaan ne opetustavoitteet, joita tulisi korostaa. Tarve voidaan määritellä erona odotetun ja todellisen tilanteen **välillä**. Täten voimme puhua tarpeesta parantaa tai laajentaa kriittisen ajattelun opettamista, kun on havaittavissa diskrepanssi tavoitteena olevan ja todetun käyttäytymistason välillä. Tämän hetkistä suoritustasoa (= statusta) määrittämään voidaan epäilemättä parhaiten käyttää kriteerikoetta.

8.3. Opetuksen yksilöllistäminen

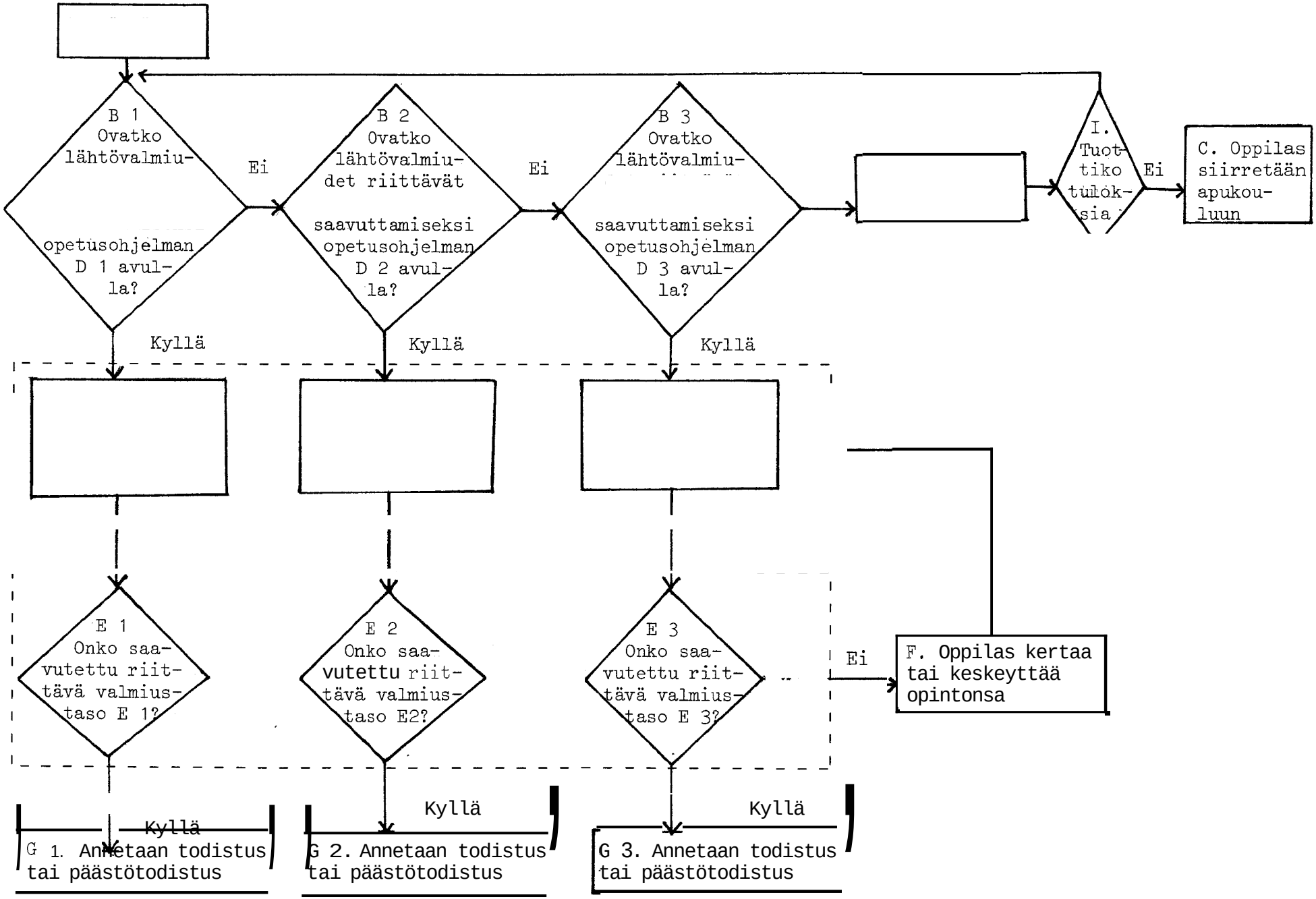
Yksi tapa opetuksen yksilöllistämiseen on antaa oppilaiden edetä omaa vauhtiaan. Oppimisyksikön sisältöä kattavan (formatiivisen) kokeen pistemäärät ovat kriteerinä sille, saavatko oppilaat edetä vai ei. Kuvio 10 havainnollistaa tällaista järjestelmää. Glaserin (1976) esittämää pitkälle yksilöidyn opetusjärjestelmän hahmotelma esitetään kuviossa 11.

Kriteerikokeet

Tällaiset yksilöidyt opetusohjelmat vaativat kunkin yksikön sisällön hallintaa riippumatta siitä, onko tämä saavutus todella välttämätön edellytys ohjelman päätetavoitteitteiden hyväksyttävälle suorittamiselle. Täten kriteerikokeet soveltuvat määrittelemään, hallitseeko oppilas oppimisyksikön vai tarvitseeko hän tukiopetusta. Yksityisten osioiden tulokset voivat antaa diagnostista tietoa, joka voi olla hyödyllistä suunniteltaessa tukiopetusta. Kokeiden diagnostista arvioimista helpottaa, jos virheelliset vaihtoehdot on muodostettu systemaattisesti. Kriteerikokeiden osioiden tuottamisjärjestelmät tekevät mahdolliseksi vaatia huomattavan määrän sisällöltään rinnakkaisia kokeita, joita tarvitaan testaamaan samoja oppilaita eri aikoina. Useat kirjoittajat (mm. Millman 1970, Airasian ja Madaus 1972) ovat ehdottaneet kriteerikokeisiin pohjautuvaa mittausta ja todistusarvosanoja. Normitiivisen arvostelun sijasta ehdote-



KUVIO 10. Malli tavoiteoppimisen soveltamisesta vieraiden kielten opiskeluun (Takala 1976).



KUVIO 11. Lähtövalmiuksien kehittämisen ja erilaiset oppimistyylit huomioonottava opetusjärjestelmä, jossa tavoitteet ja saavutukset voivat vaihdella (Glaser 1976)

taan käytettäväksi arviointia, jossa oppilaan asema (status) suhteutetaan kurssin tavoitteisiin. Suomessa Korpinen (1976, 1977a,b, 1978a,b, 1979) on kehittänyt ja tutkinut tällaista tavoitepohjaista arvostelu- ja tiedottamismenetelmää (= sanallinen tiedottaminen). Siinä on mahdollista kuvata oppilaan edistymistä paremmin kuin suhteellisen arvostelun vallitessa, jolloin arvosanojen parantuminen voi ainoastansa tapahtua muiden oppilaiden kustannuksella.

Normikokeet

Silloin kun eriytetään oppimisen tavoitteita, ts. silloin kun kaikki oppilaat eivät opiskele samoja asioita, erottelukokeilla on Millmanin (1974) mukaan ilmeisen selvä tehtävä osoittaa miten erilaiset opetussisällöt vastaisivat oppilaiden harrastusta ja edellytyksiä. Tällä kokeiden käytöllä on pitkä historia eikä sitä käsitellä enemmälti tässä yhteydessä. Kun opetusta yksilöidään varioimalla aikaa, joka on käytettävissä yhteisten tavoitteiden saavuttamiseen, erottelukokeet näyttävät olevan erityisen hyödyllisiä erottelemaan ne, jotka todennäköisesti pystyvät käyttämään hyväksi tulevaa opetusta niistä, jotka eivät onnistuisi seuraamaan tulevaa opetusta. Kriteeriryhmien tulisi olla kaksi ryhmää, jotka eroavat kyvyssään onnistua seuraavassa opetusyksikössä. Erottelukokeen hyödyllisyyttä tässä yhteydessä on käsitelty edellä jaksossa 2.5. esimerkissä kolme.

8.4. Opetusohjelman arviointi

Kriteerikokeet

Keskeisiä näkökohtia arvioitaessa opetusohjelmia on, missä määrin ohjelman tavoitteet on saavutettu. Kriteerikokeilla pyritään saamaan tällaista tietoa. Opetuksen tavoitteena olevien käyttäytymismuotojen saavuttamista mittaavat spesifit kriteerikokeet antavat normikokeisiin verrattuna paljon paremman mahdollisuuden havaita alueita, joissa ohjelma on ollut tuloksekas, ja alueita joissa on muutoksen tarvetta. Ei ole välttämätöntä heittää koko ohjelmaa yli laidan. Tehostetut toimenpiteet voidaan kohdistaa niihin ohjelman tavoitteisiin, joita ei ole saavutettu.

Jotkut kirjoittajat (mm. *Cox* ja Sterrett, 1970) ovat suositelleet, että kutakin standardikokeen osiota arvioidaan sen suhteen, missä määrin se vastaa koulun opetussuunnitelmaa, ja tämän jälkeen verrataan oppilaan suoritusta näissä osioissa verrattuna osioihin, joiden ei katsota vastaavan opetussuunnitelmaa. Vaikka tällaiset vertailut saattavatkin olla kiinnostavia, ne eivät Millmanin (1974) mukaan anna kasvattajalle tietoa siitä mitkä tavoitteet on opetettu menestyksellisesti ja mitä tavoitteita ei. Tässä on tarpeen viitata edellä esitettyihin näkemyksiin siitä, miten vaikeata on saada kriteeritulkintoja tavoitepohjaisista (objectives-based) kokeista.

Kun mittauksen tarkoituksena on tehdä päätöksiä opetusohjelmasta, ei ole välttämätöntä että jokainen oppilas suorittaa samat osiot tai edes että jokainen oppilas testataan. Voidaan saavuttaa huomattavia taloudellisia säästöjä otostamalla sekä oppilaita että osioita ja arvioimalla koko alueen suoritustasoa koko oppilasjoukossa.

Otoksen aluepistemäärä on harhaton estimaatti populaation aluepistemäärästä. Mikäli kussakin kokeessa on yhtä paljon osioita ja yhtä paljon koehenkilöitä kussakin ryhmässä, voidaan käyttää Lordin ja Novickin laatimia kaavoja arvioimaan aluepistemäärien estimaattien otosvarianssia.

Kouluoloissa on suositeltavaa, että luokassa esitetään samanaikaisesti useita testejä (ts. satunnaisotoksia osioista) yhdellä kertaa. Tämä ei ole kuitenkaan mahdollista, jos osiot on luettava ääneen, mikä usein on välttämätöntä nuorten lasten ollessa kyseessä.

Normikokeet

Opetusohjelman arvioitsija saattaa kysyä esimerkiksi: Kuinka hyvin ohjelman tavoitteet hallitaan? Onko ohjelmalla ollut minkäänlaista vaikutusta? Edellä jaksossa 2.5. esitetty esimerkki 1 osoittaa, että kriteerikokeet pystyvät paremmin vastaamaan ensimmäiseen näistä kysymyksistä. Kun esim. poliittisista syistä pyritään osoittamaan, että on tapahtunut edistymistä, erotteleva normikoe on asianmukainen. Tämä siitä syystä ettei ole niinkään tärkeätä ottaa satunnaisotosta hyvin määritellyltä alueelta kuin järjestää koe, joka erottelee erilaisen käsittelyn saaneet ryhmät toisistaan eli osoittaa muutosta, joka on tapahtunut uudenlaisen opetuksen ansiosta.

8.5. Opetuksen kehittäminen

Opettajan tärkeimpiä ominaisuuksia on kyky tuottaa käyttäytymisen muutoksia oppilaissa ennalta määriteltyjen tavoitteiden suunnassa. Kun oppilaiden suoritusta mitataan kriteerikokeiden avulla, toivottu oppilaan käyttäytyminen tulee eksplisiittisesti näkyviin. Määritellään tarkasti käyttäytymisen rajat ja samoin määritellään tarkasti kriteerit, joiden perusteella päätellään oppilaiden vastausten asianmukaisuus. Tällainen tieto antaa opettajalle mahdollisuuden laatia parempia opetustilanteita ja se antaa paremman arvion opettajan oman toiminnan tuloksellisuudesta.

Käyttäytymistavoitteita arvostellaan usein siitä syystä, että niiden ansiosta opetus typistyy liiaksi. Kriteerikokeet voivat auttaa välttämään tätä ongelmaa, koska ne mittaavat selvästi määriteltyjä käyttäytymisen luokkia. Tämä tarkennus voi auttaa opettajaa laajentamaan käsitystään opetuksen tavoitteista.

Merkitseekö tämä että sallitaan se että opetetaan täsmälleen mitä kokeessa tullaan mittaamaan? Millmanin (1974) mukaan vastaus on "ei". Kuitenkin Hively on Millmanin mukaan todennut, että testausohjelmaa ympäröivän salaisuuden määrä on opetuksen tavoitteita koskevan tietämättömyyden mittari. Huolellisesti määritellyt alueet tulisi Hivelyn mielestä saattaa julkiseen tietoon ja vain tiettyihin mittauksiin tarkoitetut osiot tulisi pitää salaisina.

Kun tässä esityksessä on käsitelty kokeiden käyttötapoja, kriteerierottelukokeita ei ole paljonkaan käsitelty. Saattaa herätä kysymys: Miksi kokeen täytyy olla joko kriteerikoe tai koe joka maksimoi ryhmien välisiä eroja? Miksi ei voisi laatia kriteerierottelukoetta käyttämällä osioiden valinnan empiirisiä menetelmiä, jotka maksimoivat erottelua samalla kun rajoitetaan alkuperäinen osioallas niihin osioihin, jotka ovat sisällöltään valideja?

Kun käytetään kompromissina kriteerierottelukoetta, ei Millmanin (1974) mukaan ole enää mahdollista muuttaa näitä testipistemääriä aluepistemääriksi (koska ei ole noudatettu osioiden valinnassa satunnaista tai ositettua otantaa) ja on olemassa vaara, että tuloksena on alentunut validiteetti. Vaikkakin jotkut kompromissit sallivat näiden kahden menettelyn hyvin puolien hyväksikäytön, kompromissi saattaa johtaa siihen, ettei käyttäjällä ole minkäänlaista vankkaa pohjaa jalkojensa alla.

8.6. Aluepistemäärien estimointi

Oletetaan että koe, joka koostuu tietyltä alueelta satunnaisesti valituista osioista, on esitetty yhdelle tai useammalla henkilölle. Tässä kappaleessa kuvataan pääasiassa kahta lähestymistapaa koehenkilön aluepistemäärän (domain score, level of functioning score, true proportion-correct score) estimoinniksi, ts. sen arvioimiseksi millainen hänen toimintakykytasonsa olisi eli kuinka suuren prosentin osioista hän todennäköisesti vastaisi oikein, jos kaikki alueeseen kuuluvat osiot esitettäisiin hänelle. Nämä kaksi lähestymistapaa ovat binomimalli ja bayesilainen malli. Kummatkin mallit sisältävät rajoituksia siitä, mitä yleistyksiä voidaan tehdä henkilön aluepistemäärästä laajempaan osiouniversumiin, josta on valittu satunnaisotanta kokeeseen. Osioiden sisällön homogeenisuudesta ja osioiden vaikeudesta ei ole tehty mitään olettamuksia. Kummatkin menetelmät vaativat, että kukin osio on pisteistetty joko oikein tai väärin (0-1). Kummatkin menetelmät edellyttävät vastausten riippumattomuutta toisistaan eli itse asiassa kokeen suorittajan odotettu suoritustaso ei todennäköisesti muutu kun osioihin on vastattu. Vastoin viimeksimainittua olettamusta on täysin mahdollista, että tapahtuu oppimista tai interferenssiä kokeen suorittamisen aikana, jopa ilman että kokeensuorittajalle annetaan minkäänlaista palautetta. Mikäli näin tapahtuu, kummatkaan menetelmät eivät anna täysin oikeata kuvaa henkilön aluepistemäärästä.

Oikeinratkaistujen tehtävien suhteellinen määrä

Yksinkertaisin ja perinteellisin tapa estimoida kokelaan osa-aluepistemäärää on hänen empiirinen oikein ratkaistujen osioiden suhteellinen määrä (proportion-correct score). Tämä saadaan jakamalla kokelaan pistemäärä (= oikein ratkaistujen osioiden määrä) kokeeseen sisältyneiden osioiden kokonaismäärällä. Vaikka tämä pistemäärä on osa-aluepistemäärän harhaton estimaatti, se on Hambletonin (Hambleton 1974, Hambleton et al. 1978) mukaan hyvin epäreliaabeli silloin kun osioiden kokonaismäärä on vähäinen. Varsinkin ohjelmoidun opetuksen välikokeissa osioiden määrä jää vähäiseksi (ehkä puoli tusinaa). Tällöin on hyödyllistä käyttää apuna aikaisempaa tietoa kokelaista: hänen aikaisempaa menestymistään kokeissa, koko oppilasryhmän keskimääräistä suoritustasoa kokeessa jne. Tätä menetelmää käytetäänkin seuraavassa mallissa.

TAULUKKO 5. Väärin luokiteltujen oppilaiden prosentuaaliset osuudet binomimallissa, kun alin hyväksytty ratkaisuprosentti on 70 tehtävistä oikein (Millman 1974, Table 6-2)

Hyväksymis-	Koeosioiden	Oppilaan todellinen suoritustaso						
pistemäärä	määrä	40 %	50 %	60 %	65 %	75 %	80 %	90 %
1	1	40	50	60	65	25	20	10
2	2	16	25	36	42	44	36	19
3	3	6	13	22	27	58	49	27
3	4	18	31	48	56	26	18	5
4	5	9	19	34	43	37	26	8
5	6	4	11	23	32	47	34	11
5	7	10	23	42	53	24	15	3
6	8	5	14	32	43	32	20	4
7	9	3	9	23	34	40	26	5
7	10	5	17	38	51	22	12	1
9	12	2	7	23	35	35	21	3
11	15	1	6	22	35	31	16	1
14	20	1	6	25	42	21	9	-
18	25	-	2	15	31	27	11	-
21	30	-	2	18	36	20	6	-
28	40	-	1	13	31	18	4	-
35	50	-	-	10	28	16	3	-
42	60	-	-	7	25	15	2	-
53	75	-	-	4	18	16	2	-
70	100	-	-	2	17	10	1	-

a) Oppilaan todellinen suoritustaso tarkoittaa sitä, kuinka monta prosenttia hän osaisi ratkaista kaikista osa-alueeseen kuuluvista tehtävistä. Jos oppilaan todellinen suoritustaso on alempi kuin 70 %, hänen tulisi epäonnistua osa-aluetta mittaavassa kokeessa. Kuitenkin osa tällaisistakin oppilaista saattaa mittausvirheiden vuoksi saada havaitun koepistemäärän, joka on hyväksymispistemäärää korkeampi. Tällä tavalla väärin luokiteltujen (= aiheetta läpäisseiden) oppilaiden odotettu prosenttiosuus esitetään katkoviivan vasemmalla puolella. Vastaavasti 70 % yläpuolella olevien oppilaiden tulisi läpäistä koe, mutta osa käytännössä epäonnistuisi. Tällä tavalla väärin luokiteltujen (aiheetta hylättyjen) oppilaiden prosentuaaliset osuudet käyvät ilmi katkoviivan oikealla puolella.

Klassinen malli II

Hambletonin (Hambleton et al. 1978) mukaan jo vuonna 1927 Kelly kehitteli menetelmiä, joiden avulla voitiin käyttää hyväksi tietoa koko kokelasryhmästä arvioitaessa yksityisen kokelaan todellista pistemäärää. Kyseessä on todellisen pistemäärän regressioestimaatti, joka on kahden komponentin - kokelaan havaitun pistemäärän ja koko kokelasryhmän keskiarvon - painotettu summa. Menetelmän matemaattisiin perusteisiin voi tutustua mm. Lordin ja Novickin (Lord & Novick 1968), Novickin ja Jacksonin (Novick & Jackson 1974) ja Hambletonin (Hambleton, et al. 1978) avulla.

Binomimalli

Millmanin (Millman 1970, 1974) mukaan tässä mallissa tarkkaillaan vain kunkin yksilön suoritustasoa eikä ryhmän pistemääriin kiinnitetä mitään huomiota. Koehenkilön oikein ratkaisemien osioiden prosenttimäärää pidetään arviona hänen aluepistemäärästään. Jos oppilas ratkaisee oikein kuusi kahdeksasta kysymyksestä, silloin mallin mukaan paras arvio on, että koehenkilön oikeiden vastausten pistemäärä olisi 6/8 eli 75 %, mikäli alueen kaikki osiot esitettäisiin hänen ratkaistavakseen. Oheinen taulukko 5 havainnollistaa binomimallia. Se osoittaa kuinka suuri osa oppilaista luokiteltaisiin virheellisesti joko hylättyihin tai hyväksyttihin erilaisella osiomäärällä. Taulukossa on alin hyväksyttävä määrä oikein ratkaistuja osioita 70 %. Jos kokeessa olisi 40 osiota, oppilaan tulisi osata vastata oikein vähintään 70 % 40:stä osiosta eli 28 osioon. Jos oletetaan että oppilas osaisi todellisuudessa ratkaista vain 60 % osioista, mikäli hänelle esitettäisiin kaikki alueen osiot, on olemassa 13 %:n todennäköisyys että hän osaa vastata oikein 28:aan tai useampaan osioon ja voi täten saada 70 %:n hyväksyttävän pistemäärän tai sitä paremman pistemäärän. Luku 33 löytyy sarakkeen 60 % ja rivin 28 - 40 leikkauskohdasta. Jos henkilön aluepistemäärä, ts. hänen toimintatasonsa on 80 %, todennäköisyys on 96 % että hän läpäisee hyväksyttävästi 40 osiota sisältävän kokeen. Rivin 28 - 40 ja sarakkeen 80 leikkauskohdassa oleva luku 4 osoittaa, että hyväksytyn suorituksen todennäköisyys tulla luokitelluksi hylätyksi on tässä tapauksessa 4 % (joten hyväksymistodennäköisyys on yllä esitetty 96 %).

Taulukko on tietyssä suhteessa mielenkiintoinen, vaikka ei olisi määriteltykään hyväksyttävän prosenttimäärän rajaa. Oletetaan että oppilas

saa 14 pistettä 20-osioisesta kokeesta, ts. ratkaisee 70 % oikein. Taulukosta näemme, että mikäli oppilaan aluepistemäärä olisi 50 %, todennäköisyys on 6 % että hän olisi osannut ratkaista 20:stä osiosta oikein 14 osiota tai sitä enemmän. Täten on järkevää olettaa, että oppilas erittäin todennäköisesti osaa vastata oikein vähintään 50 %:iin alueen osioista. Binomimalli ei kuitenkaan anna tämän todennäköisyyden numeerista estimaattia. Sen sijaan bayesilainen malli antaa kyseisen estimaatin.

Binomimallia voidaan Millmanin (1979) mukaan käyttää antamaan estimaatteja myös muunlaisissa tilanteissa: kun on valittu osiot ositetusti (esim. piirteittäin), kun halutaan keskimääräinen aluepistemäärä joukolle henkilöitä ja on käytetty matriisiotantasuunnitelmia, tai kun ollaan kiinnostuneita muutoksista aluepistemäärissä.

Bayesilainen malli

Edellä kuvattu binomimalli käsittelee kutakin koehenkilöä erikseen, ikäänkuin ei olisi olemassakaan muita kokeensuorittajia. Sen sijaan jäljempänä kuvattava bayesilainen malli II pitää etukäteisinformaationa tietoa muista oppilaista ja kokeenpitäjän ennakkokäsityksiä. Ennakkotiedot yhdessä kriteerikokeiden tulosten kanssa antavat tarkennetun jälkikäteisarvion aluepistemääristä. Bayesilaista mallia käytettäessä saadaan tarkempia estimaatteja tai, mikäli niin halutaan, saavutetaan sama tarkkuus kuin binomimalleilla mutta käyttämällä lyhyempiä kokeita. Bayesilaisen mallin suurempi tarkkuus ostetaan sillä hinnalla, että edellytetään binomimallia enemmän rajoituksia. Mallit eivät eroa pelkästään tarkkuudeltaan, vaan ne eroavat myöskin aluepistemäärien estimaattien suhteen. Mikäli kokeenpitäjästä tuntuu, etta ennakko-olettamukset ovat järkeviä, on Millmanin (1974) mukaan syytä antaa etusija bayesilaiselle mallille.

Millmanin (1974) antamassa esimerkissä oletetaan että on esitetty kymmenosioinen koe oppilaille. Mallin mukaan henkilöllä, joka ratkaisi oikein vain 50 % kymmenosioisesta kokeesta, ennustetaan olevan pistemäärän, joka vastaa 67 %:a osiouniversumista. Yleensä ennustetaan, että oppilailla jotka saavat ryhmän keskiarvoa alemman pistemäärän, tulisi olemaan korkeampi pistemäärä ja vastaavasti ryhmän keskiarvoa paremman tuloksen saaneella oppilaalla tulisi olemaan alhaisemmat aluepistemäärät (vrt. regressio kohti keskiarvoa). Lisäksi on huomattava, että korjaus on varsin huomattava sellaisen henkilön kohdalla, joka on saanut ääripistemäärän. Itse asiassa jos henkilö on saanut 50 % oikein kymmenosioisesta kokeesta, bayesilainen malli

arvioi, että on olemassa 40 %:n todennäköisyys, että hänen aluepistemäärän-sä on korkeampi kuin 70 %.

Seuraava Millmanin (1974) esittämä esimerkki havainnollistaa binomi-mallin ja bayesilaisen mallin eroa. Binomimallin mukaisesti odotetaan, että henkilö, jolla on 70 %:n aluepistemäärä, saa pistemäärän viisi tai sitä alhaisemman pistemäärän kymmenosioisessa kokeessa vain 15 %:ssa kaikista tapauksista. Bayesilaisen mallin mukaan henkilön, joka on saanut viisi pistettä kymmenosioisessa kokeessa, odotetaan 40 %:ssa kaikista tapauksista saavan aluepistemäärän, joka on 70 % tai enemmän. Edellä kuvatut lauseet on luettava huolellisesti, koska ne eivät ole suoraan keskenään verrattavissa.

Kun kokeenpitäjän luottamus ennakkokäsitysten paikkansapitävyydestä lisääntyy arvosta t = 2,75 (.05) arvoon t = 17.75 (.01), 50 % oikein ratkaisseen koehenkilön estimoitu aluepistemäärä muuttuu 67 %:sta 77 %:iin ja todennäköisyys että henkilön aluepistemäärä on yli 70 % lisääntyy 40 %:sta 79 %:iin. Täten puolet lyhyen kokeen tehtävistä oikein ratkaisseen henkilön ajatellaan omaavan neljä mahdollisuutta viidestä saavuttaa aluepiste-määrän, joka on 70 % tai parempi.

Millmanin (1974) mukaan bayesilainen malli on hyvin lupaava. Hänen käsityksensä mukaan tarvitaan kuitenkin lisää tutkimuksia siitä, miten paljon poikkeamia olettamuksesta malli sallii ilman että tulokset vääristyvät. Niin kauan kuin tällaista tietoa ei ole, tuntuu järkevältä käyttää estimaatteja, jotka ovat suhteellisen konservatiivisen binomimallin ja bayesilaisen mallin välimaastossa.

Bayesilaisen mallin matemaattisia perusteita ovat selvitelleet mm. Lewis, Wang & Novick (1973), Novick, Lewis & Jackson (1973), Millman (1974) ja Hambleton, Swaminathan, Algina & Coulson (1978).

8.7. Kokeen pituuden määrittäminen

Kokeenlaatijalla on mielessä tietynlainen hyötysuhdeajattelu (trade-off), kun tehdään päätöstä kriteerikokeen osioiden määrästä. Käytännön rajoitukset, mm. käytettävissä oleva testausaika, edellyttävät että käytetään vain suhteellisen vähän osioita. Tarkka tieto edellyttää kuitenkin pitkää koetta. Hyötysuhde on erityisen selvästi esillä, kun kokeen käyttäjä haluaa

mitata koehenkilöiden asemaa samanaikaisesti useilla alueilla. Monien sovellutusten kohdalla tämä tilanne ei esiinny tai sitten voidaan käyttää mm. matriisotantasuunnitelmaa, joka lyhentää testausaikaa yksityisten oppilaiden osalta.

Monet kokeet sisältävät varsin vähäisen määrän osioita. Niiden tarkkuus onkyseenalaista, ja sellaiset kokeet antavat tietoa vain miten oppilas osaa ratkaista tietyn osion eikä miten hän osaisi ratkaista yleisemmän tehtavaluokan. Tällaisen tiedon arvo ei ole aina kovin suuri ja yhden osion indikaattoreiden kehittelijät ovat huomanneet välttämättömäksi yrittää jonkinlaista profiilianalyysia, jossa käytetään vastauksia useampiin samantapaisiin koeosioihin. Olisi kuitenkin hyödyllistä, jos osiot valittaisiin huolellisesti ja tarkasti määritellyltä alueelta.

Binomimalli

Edellä esitetty taulukko 5 on avuksi kokeen pituuden määrittelemisessä. Se osoittaa kuinka estimaatin tarkkuus vaihtelee kokeen pituuden funktiona, jos kokeenpitäjä esimerkiksi sallii 15 %:ssa kaikista tapauksista oppilaan, jonka todellinen aluepistemäärä (suoritustaso) on 60 %, saavan 70 %:n testipistemäärän. Voimme sarakkeesta 60 % todeta, että 25 osion koe antaa tämän tarkkuuden. Tämä merkitsee siis sitä, että on olemassa vain 15 %:n todennäköisyys, että oppilas jonka aluepistemäärä on 60 %, saisi tuloksen 70 % tai sitä paremman kokeessa, jossa on 25 osiota.

Bayesilainen malli

Novick ja Lewis (1974) ovat esittäneet bayesilaisen mallin kokeen pituuden määrittelemiseksi. Heidän esittämiensä menettelytapojen soveltaminen merkitsee yleensä, että tarvitaan lyhyempiä kokeita kuin mitä jouduttaisiin käyttämään sovellettaessa binomimallia. Suurempi tarkkuus saavutetaan osittain sillä kustannuksella, että kokeen pitäjän on tehtävä monia lisäolettamuksia. Oheinen taulukko havainnollistaa bayesilaista mallia.

TAULUKKO 6. Kriteerikokeen pituutta koskevia suosituksia (Lord & Novick 1974) olettaen beta-jakautumat, joiden sisältämä ennakkotieto vastaa 8–15 -osioisesta kokeesta saatavaa tietoa (Millman 1974, Table 6–6)

π_0 a)	$\varepsilon(\pi)$ b)	Menetyssuhde c)							
		1.5		2.0		2.5		3.0	
70 %	70 %	6/8	(75 %) d)	10/13	(77 %)	11/14	(79 %)	12/15	(80 %)
75 %	75 %	8/10	(80 %)	16/20	(80 %)	17/21	(81 %)	18/22	(82 %)
80 %	80 %	6/7	(86 %)	7/8	(88 %)	17/20	(85 %)	19/22	(86 %)
80 %	85 %	8/10	(80 %)	9/11	(82 %)	10/12	(83 %)	11/13	(85 %)
85 %	85 %	7/8	(87 1/2%)	9/10	(90 %)	17/19	(89 %)	19/21	(90 %)

a) π_0 on kriteeritaso eli suoritustaso, joka osoittaa asian hallintaa (usein 70–85 %).

b) $\varepsilon(\pi)$ on testattavien odotettu keskimääräinen suoritustaso, joka arvioidaan ennen koesuorituksia. Jos opetuksen kuluessa oli formatiivisissa kokeissa vaatimustasona 70 %, voidaan perustellusti odottaa, että oppilaiden keskimääräinen todellinen suoritustaso olisi 70 tai parempi. Sarakkeessa π_0 olevat 2 riviä, joilla molemmilla on arvo 80 %, osoittavat kuinka herkkiä kokeen pituutta koskevat suositukset ovat $\varepsilon(\pi)$:n arvoille.

c) Menetyssuhde = $\dfrac{\text{menetys, joka aiheutuu sellaisen oppilaan hyväksymisestä, jolla } \pi < \pi_0}{\text{menetys, joka aiheutuu sellaisen oppilaan hylkäämisestä, jolla } \pi > \pi_0}$

Esim. Menetyssuhde 2 merkitsee, että katsotaan 2 kertaa vakavammaksi virheeksi hyväksyä asiaa hallitsematon oppilas kuin hylätä oppilas virheellisesti. On syytä panna merkille kokeen pituutta koskevan suorituksen sensitiivisyys suhteessa valittuun menetyssuhteeseen.

d) Luetaan seuraavalla tavalla: Suositellaan 8-osioista kriteerikoetta ja alinta hyväksymisrajaa 6 (75 %), kun hallinnan tason kriteerinä on 70 %, opetusryhman odotettu keskimääräinen suoritustaso on 70 % ja kun menetyssuhde on 1.5.

Sen lisäksi että oletetaan binomimallin mukaisesti toisistaan riippumattomia havaintoja satunnaisella otoksella oikein-välin pisteistettäviä osioita, taulukossa annettavat arvot ovat tuloksena kokeenpitäjän ennakkokäsityksestä, että oppilaan suoritustasolla on betajakautuma. Tämä oletettamus ei Millmanin (1974) mukaan ehkä ole asianmukainen, mikäli kokeen alue kattaa yhden ainoan spesifin taidon, jota oppilaat eivät joko osaa lainkaan (pistemäärät ovat lähellä 0 %) tai osaavat täysin (100 %:n suoritustaso). Bayesilainen malli myöskin edellyttää, että kokeenpitäjä ilmaisee missä määrin hän luottaa ennakkotietoihinsa. Taulukossa esitetty luottamus eli informaation määrä vastaa sitä informaatiota mikä saataisiin kokeesta, jossa on 8-15 osiota. Tämän lisäksi kokeenpitäjän tulee määritellä kriteeritaso, keskimääräinen suoritustaso, ja menetyssuhde, kuten taulukon alaliitteessä selostetaan.

Taulukosta ei käy ilmi, että kyse on itse asiassa hyötysuhteesta. Kokeenpitäjä saattaisi määritellä hyväksyttävän pistemäärän lähemmäksi kriteeritasoa pitämällä pidemmän kokeen tai päinvastoin voisi käyttää lyhyempiä kokeita, mikäli hän edellyttäisi hyvin korkeita suorituspistemääriä hyväksyttävän suorituksen ehtona. Novickin ja Lewisin (1974) mukaan taulukossa esitetyt kokeen pituudet edustavat järkevää tasapainoa sen välillä, että hyväksyttävät prosenttimäärät ovat lähellä hallintakriteeriä ja samalla on valittu tehokas kokeen pituus. Taulukko antaa täten ohjeita siitä minkälaiseen hallintaan tulisi pyrkiä sekä siitä kuinka pitkä kriteerikokeen tulisi olla. Täten taulukko antaa mahdollisuuden tutkia neljän variaabelin: 1) ryhmän keskimääräisen suoritustason, 2) menetyssuhteen, 3) määritellyn vaatimustason ja 4) kokeen pituuden vaikutuksia opetusta koskevien päätösten tekemiseen.

8.8. Monitasomittaaminen

Hambletonin (1974) mukaan erityisesti silloin kun mitattavalla sisältöalueella on todettavissa oppimishierarkioita, mittaamiseen tarvittavaa aikaa voidaan lyhentää haarautuvan mittaamisen (branched testing, branch testing) avulla. Ferguson (1969, 1971) ja Spineti (1973) ovat käyttäneet tietokonetta apuna sopivantasoisten osioiden esittämisessä kokelaille ja todenneet, että koeaikaa on voitu tuntuvasti lyhentää (Spinetin mukaan jopa puoleen) ilman, että päätösten tarkkuus olisi lainkaan kärsinyt. Myös Lord (1976,

24-30) on sitä mieltä, ettei samalla kokeella voida kovin hyvin mitata sekä kovin heikkojen että kovin hyvien kokelaiden suoritustasoa. Lisäämällä muutamia hyvin helppoja ja muutamia hyvin vaikeita osioita ei ongelmaa voida tyydyttävästi ratkaista. Lordin mielestä ei saavutettaisi tyydyttävää tulosta, vaikka oppilaat jaettaisiin kolmeen tai neljään tasoryhmään ja osioita painotettaisiin sopivalla kertoimella (esim. osion erotteluindeksillä). Jos vain kolmas- tai neljäsosa osioista on sopivantasoisia, ei tilastollisella manipuloinnillakaan Lordin mielestä mittarista saada aikaan hyvää koetta kyseiselle oppilasryhmälle.

Lord pitää tarpeellisena kehittää uudenlaisia menettelyjä, joita voidaan nimittää monitasomittaamiseksi (multilevel testing), kaksivaiheiseksi mittaamiseksi (two-ctage testing), "mittatilausmittaamiseksi" eli "räätälintyömittaamiseksi" tai proosallisemmin ehkä yksilölliseksi mittaamiseksi (tailored testing). Tällaista mittausta on nimitetty myös tietokonepohjaiseksi mittaamiseksi (computer-based testing). Tämä on kirjoittajan mielestä huonohko nimitys, koska siinä kiinnitetään huomiota mittauksen apuvälineeseen eikä kokeen sisältöön tai käyttötapaan. Kuvaavampia nimityksiä ovat edellisten lisäksi hierarkkinen testaaminen (sequential item testing) ja joustava mittaaminen (adaptive testing, flexilevel testing). Yhdysvalloissa on Civil Service Commission kokeillut monitasomittaamista positiivisin tuloksin. Koetekniikasta on Lordin (1976) mukaan pidetty enemmän kuin tavanomaisesta menettelystä ja osiomäärä on voitu pudottaa 100:sta 20:een.

Lord (1976) kuvaa monitasomittaamista seuraavalla tavalla. Oletetaan, että on laadittu 50 osiota, jotka mittaavat samaa osa-aluetta (taitoa, kykyä piirrettä, tms). Osiot on jaettu viiteen tasoon a,b,c,d ja e vaikeustason mukaisesti. Kaikki kokelaat saavat vastattavakseen c-version ensiksi. Jos osiot tuntuvat vaikeilta, kokelasta kehoitetaan siirtymään b-versioon ja tarvittaessa edelleen a-versioon. Jos taas tehtävät tuntuivat helpoilta, suositellaan ensin siirtymistä d-versioon ja tämän jälkeen mahdollisesti vielä e-versioon. Kukin kokelas suorittaa korkeintaan 30 tehtävää ja osioblokit ovat joko abc, bcd tai cde. Versiopistemäärät tulee verrantaa (equate) samalle asteikolle joko perinteellisin menetelmin tai osion ominaiskäyrää (item characteristic curve) koskevan teorian avulla (ks. Konttinen 1981). Kun pistemäärät on verrantamisen avulla saatu samalla asteikolle, monitasotestauksen avulla on mittaaminen ollut tehokkaampaa kuin perinteellisin menetelmin, koska kukin kokelas on todennäköisesti suorittanut suoritustasoaan paremmin vastaavat tehtävät. Tämän raportin kirjoittajan mielestä monitasotestillä on tehokkuuden lisäksi se etu, että

kokelaat välttyvät tarpeettomilta shokeilta saadessaan suoritustasoaan vastaavia tehtäviä. Lisäksi menetelmä siirtää osan vastuusta oppimistulosten seurannasta sinne minne se ihanteellisesti kuuluukin eli oppilaille itselleen.

9. KRITEERIKOKFIDEN ARVIOINTIPERUSTEISTA

Walker (1978) on CSE:n (Center for the Study of Evaluation) julkaisusarjassa ilmestyneessä raportissa pyrkinyt määrittelemään perusteita laadittujen kriteerikokeiden arvioinnille. Alan kirjallisuudesta ja testikustantajien mainosjulkaisuista saatiin eristettyä 70 mahdollista arviointiperustetta, jotka ryhmiteltiin uudelleen 21:ksi, kolmeen paaluokkaan jakautuvaksi kriteeriksi. Luokat koskevat kriteerikokeiden mittausominaisuuksia, niiden sopivuutta kokelaille ja niiden käytännöllisyyttä. Seuraavassa esitetään lyhyt yhteenveto kriteereistä.

IA. MITTAUSOMINAISUUDET: KÄSITEVALIDITEETTI

1. Kuvaus: Kuinka hyvät (ts. perusteelliset ja ymmärrettävät) ovat mitattavien alueiden kuvaukset?
2. Vastaavuus: Kuinka hyvin osiot vastaavat opetuksen tavoitteita?
3. Edustavuus: Kuinka edustava otos osiot ovat tavoitteista?

IB. MITTAUSOMINAISUUDET: EMPIIRINEN VALIDITEETTI

4. Herkkyys: Kuinka herkästi opetuksen tulokset tulevat ilmi koepistemäärissä?
5. Osioiden yhdenmukaisuus: Kuinka yhdenmukaisia ovat samaa tavoitetta mittaavien osioiden pistemäärät?
6. Divergentti validiteetti: Missä määrin kutakin tavoitetta mittaavien osioiden pistemäärät ovat vapaita ulkopuolisten tekijöiden vaikutuksilta?
7. Ryhmäsyrjinnän puuttuminen: Missä määrin koetulokset ovat riippumattomat erilaisista sosiaaliryhmiin liittyvistä tekijöistä?

8. Koepistemäärien johdonmukaisuus: Ovatko yksityisiä tavoitteita mittaavien osioiden pistemäärät johdonmukaisia eri aikoina ja eri koeversioissa?

II. SOPIVUUS KOKELAILLE

9. Koeohjeiden selkeys: Kuinka selkeät ja täydelliset koeohjeet ovat?
10. Osioiden tarkistus: Onko osiot arvioitettu asiantuntijoilla ja esitestattu?
11. Ulkoasu: Onko ulkoasu selkeä ja helppolukuinen?
12. Vastaamisen helppous: Kuinka helppo tai vaikea on vastaustekniikka?

III. KÄYTÄNNÖLLISYYS

13. Tiedottavuus: Kuinka paljon annetaan käyttäjälle tietoa kokeesta?
14. Opetussuunnitelmaan liittyminen: Kuinka selkeästi koe liittyy voimassa olevaan opetussuunnitelmaan ja käytettävissä oleviin oppimateriaaleihin?
15. Joustavuus: Voidaanko samaa tavoitetta mitata useammalla hallinnan tasolla ja voidaanko erillisiä tavoitteita mitata helposti erikseen?
16. Rinnakkaisversiot: Onko rinnakkaistestiä saatavana?
17. Kokeen esittäminen: Ovatko kokeen pitäjälle annetut ohjeet selkeät ja riittävät?
18. Pisteistys: Voidaanko koe korjata sekä koneellisesti että käsin?
19. Koetulosten kirjaaminen: Onko saatavilla lomakkeita koetulosten kirjaamiseksi?
20. Päätössuositukset: Annetaanko perusteltuja ja selkeitä ohjeita koetulosten pohjalta tehtäville ratkaisuille?
21. Vertailutiedot: Annetaanko edustaviin otoksiin perustuvia normitietoja?

Myös Hambleton ja Eignor (1978) ovat laatineet ohjeiston kriteerikokeiden ja niiden käsikirjojen arvioimiseksi. Ohjeet on kirjoitettu kysymysten muotoon ja ne on luokiteltu kymmeneen ryhmään: tavoitteet, koeosiot, kokeen järjestäminen, kokeen ulkoasu, reliabiliteetti, hyväksytyn suorituksen pisterja, validiteetti, normit, koepistemäärien raportointi ja koepistemäärien tulkinta. Kustannus- ja aikanäkökohdat on jätetty luettelosta pois, vaikka ne tulee luonnollisesti myös ottaa huomioon.

A. Tavoitteet

A.1. Onko kokeen tarkoitus tai tarkoitukset ilmaistu selvästi ja lyhyesti?
A.2. Onko kukin mitattava tavoite ilmaistu niin selvästi, että on mahdollista tunnistaa "osioallas"?
A.3. Käykö tavoiteluettelosta ilmi mitä koe mittaa?
A.4. Esitetäänkö asianmukaiset perustelut kunkin tavoitteen sisällyttämiselle kokeeseen?
A.5. Voidaanko koe muokata paikalliseen käyttöön siten, että voidaan selvittää mitä kohtia laajemmasta tavoitelistasta koe kattaa?
A.6. Ovatko kokeen sisältö ja kokeen käyttötilanne keskenaan sopusoinnussa?
A.7. Tiedetäänkö, ketkä määrittelivät mitattavat tavoitteet?
A.8. Ovatko mitattavat tavoitteet edustava otos kiinnostuksen kohteena olevasta sisältöalueesta?

B, Koeosiot

B.1. Kuvataanko miten osioita arvioitiin ennen kokeen pitämistä?
B.2. Ovatko koeosiot valideja indikaattoreita mitattavaksi tarkoitetuilta tavoitealueilta?
B.3. Ovatko osiot edustava otos tavoitetta mittaavasta "osioaltaasta"?
B.4. Ovatko osiot teknisesti virheettömät?
B.5. Ovatko koeosiot ulkoasultaan sopivat mittaamaan ko. tavoitteita?
B.6. Ovatko koeosiot harhattomia (puolueettomia sukupuolen, kulttuuritaustan ym. suhteen)?
B.7. Käytettiinkö heterogeenista koehenkilöjoukkoa osioita esitestattaessa?
B.8. Käytettiinkö osioanalyysia vain heikkojen osioiden paljastamiseksi?

C. Kokeen järjestäminen

C.1. Esitetäänkö koeohjeissa tietoa kokeen tarkoituksesta, aikarajoituksista, harjoitustehtävistä, vastauslomakkeista ja pisteistyksestä?
C.2. Ovatko koe-ohjeet selkeät?
C.3. Onko koe helppo pisteistää?
C.4. Esitetäänkö koeohjeistossa kokeenpitäjän tehtävät?

D. Kokeen ulkoasu

D.1. Onko kokeen ulkoasu miellyttävä
D.2. Onko kokeen ulkoasu helppokäyttöinen?

E. Reliabiliteetti

E.1. Onko koeohjeistossa ilmoitettu reliabiliteettitieto asianmukainen kokeen käyttötarkoitusta tai -tarkoituksia ajatellen?

E.2. Oliko reliabiliteettitieto hankittu sellaisella koehenkilöotoksella, joka on asianmukainen otoksen koolle ja edustavuudelle asetettavien vaatimusten suhteen?

E.3. Ovatko kokeet riittävän pitkät, jotta saavutettaisiin tarvittava reliabiliteetti?

E.4. Ilmoitetaanko koeohjeistossa reliabiliteettitietoa erikseen kutakin koepistemäärien käyttötarkoitusta varten?

F. Hyväksyttävän suorituksen pisteraja

F.1. Esitetäänkö perustelut hyväksymisrajan määrittelemisessä käytetylle menetelmälle?

F.2. Selitetäänkö menetelmän eri vaiheet ja ovatko ne asianmukaiset?

F.3. Esitetäänkö näyttöä tai perusteluja valitun menetelmän validiteetista?

G. Validiteetti

G.1. Onko esitetty validiteettitieto sopusoinnussa kokeen käyttötarkoituksen kanssa?

G.2. Esitetäänkö koeohjeistossa katsaus koepistemäärien validiteettia koskevista tekijöistä?

H. Normit

H.1. Esitetäänkö normitiedot asianmukaisessa muodossa?

H.2. Kuvataanko normeerauksessa käytetyt koehenkilöotokset?

H.3. Esitetäänkö asianmukaiset ohjeet ja varoitukset koepistemäärien asianmukaisesta tulkinnasta?

1. Koepistemäärien raportointi

1.1. Ilmoitetaanko koepistemäärät kustakin mitatusta tavoitteesta erikseen?

1.2. Voidaanko koetulokset raportoida monella eri tavalla (esim. kouluttain, luokittain, luokkatasoittain jne)?

1.3. Voidaanko kokeet korjata kätevästi myös käsin ja onko tulosten yhteenvetolomake käytettävissä?

J. Koepistemäärien tulkinta

J.1. Esitetäänkö koeohjeistossa sopivia varauksia ja varoituksia yksilöitä ja ryhmiä koskevien pistemäärien tulkitsemiseksi?

J.2. Esitetäänkö koeohjeistossa viitteitä koepistemäärien käyttämisestä oppimistulosten kuvaamiseen, opetusta koskevien ratkaisujen tekemiseen, opetusohjelmia koskevien ratkaisujen tekemiseen tai muihin vastaaviin tarkoituksiin.

Edellä esitetyistä arviointikriteereistä on epäilemättä hyötyä myös kriteerikokeiden laatijoille.

10. RATKAISEMATTOMIA ONGELMIA JA ONGELMALLISIA RATKAISUJA

Pophamin (1978) mukaan eräs tärkeimpiä ongelmia kriteerimittaamisen kehittämisessä on sen vaikeus ja työläys. Mistä saadaan varat toimintaan, joka edellyttää korkeasti koulutettua päätoimista mittausasiantuntemusta ja runsaasti aikaa? Toinen keskeinen ongelma on käsitteiden vakiintumattomuus. Joku mittausasiantuntija puhuu "kriteerimittaamisesta" kun taas toinen nimittää sitä "osa-aluemittaamiseksi". Jotkut puhuvat "kuvauksen validiteetista", kun taas toiset käyttävät samasta tai suurinpiirtein samasta asiasta nimityksiä "sisällön validiteetti", "opetussuunnitelmallinen validiteetti", "ilmeisvaliditeetti" tai pelkästään "validiteetti". Tästä syystä on Pophamin mielestä suositeltavaa, että validiteetista ja reliabiliteetista puhuttaessa mainitaan, mitä näillä käsitteillä kussakin yhteydessä tarkoitetaan. Eräs huomioonotettava seikka on myös se, etteivät läheskään kaikki mittausasiantuntijat (mm. Ebel ja Glass) ole suinkaan vakuuttuneita kriteerimittaamisen eduista ja käyttömahdollisuuksista. Kriteerimittaaminen on alkuvaiheessaan ja tarvitaan runsaasti työtä sen periaatteiden ja testimetodiikan kehittämiseksi.

Hambleton (Hambleton et al. 1978) on todennut, ettei ole käytettävissä riittävää teoriaa eikä käytännön ohjeita kriteerimittaamisen toteuttamiseksi niinkin erilaisissa olosuhteissa kuin luokan ja valtakunnan tason arviointitoiminnassa.

Millaista kehitystä on tapahtunut ja millaista tutkimusta on tehty kriteerimittaamisen alalla? Seuraava katsaus perustuu pääosin Hambletonin (Hambleton et al. 1978) artikkelissa esitettyihin näkökohtiin.

Ensimmäisenä kehityspiirteenä voidaan mainita se, että käyttäytymistavoitteiden tilalle tai pikemminkin niitä täydentämään ovat tulleet lavennetut tavoitelauseet ja osa-aluetäsmennykset eli koetäsmennykset. Vasta yksityiskohtaiset täsmennykset tekevät mahdolliseksi koepistemääriin perustuvat yleistykset ja suoritustason tarkan kuvauksen.

Toiseksi on havaittavissa, että osioanalyysin rooli kokeita kehitettäessä on selkeytynyt. Perinteellistä empiiristä tilastotieteeseen perustuvaa osioanalyysiä on tullut täydentämään oppiainesasiantuntijoiden harkintaan perustuva osiotarkastelu. Oppiainesasiantuntijat arvioivat erityisesti sitä, onko osa-aluetäsmennys selkeä ja ovatko osiot edustava otos täsmennyksen rajaamalta alueelta. Empiiristä osioanalyysiä käytetään apuna mahdollisten puutteellisten osioiden (flawed items) paljastamiseen, mutta se ei ole pohjana osioiden valinnalle eikä karsinnalle, koska tästä voisi aiheutua sisällön edustavuuden heikkenemistä.

Kolmanneksi on tiedostettu, että kriteerikokeiden käsitevaliditeetti vaatii tutkimusta. Kokeelliset asetelmat, faktorianalyysi ja alhaisen validiteetin syiden tutkimus ovat esimerkkejä mahdollisista tutkimusaiheista.

Neljänneksi voidaan todeta, että on jo olemassa melko paljon testiteknistä tietoa kokeen pituuden ja reliabiliteetin asettamista vaatimuksista. Sen sijaan hyväksyttävän pisterajan määrittämisessä on vielä paljon ongelmia ratkaistavana.

Viidenneksi voidaan mainita, että bayesilaisia tilastomenetelmiä on kehitelty osa-aluepistemäärien estimointia, hallintatasoluokittelua ja kokeen pituuden selvittelyä varten. Päätöksentekoon liittyvien menetysten sisällyttäminen prosessiin tuo mukaan päätöksentekijän arvostukset. Bayesilaiset menetelmät käyttävät hyväksi aikaisempaa tietoa kokelaasta ja tietoa kaikista kokelaista, ja täten ne lisäävät päätöksentekijän käytettävissä olevaa informaatiota ilman että osiomäärää tarvitsisi kasvattaa tuntuvasti. Bayesilaisten menetelmien testaaminen on kuitenkin vasta alulla.

Kuudenneksi on todettavissa, ettei kriteerimittaamisessa ole vielä juuri lainkaan käsitelty luotettavuusrajojen ongelmaa. Koska mittausaika on aina rajallinen, miten eri tavoitteita mitattaessa saataisiin aikaan paras mahdollinen ajankäyttö?

Seitsemänneksi voidaan mainita, että haarautuva mittaaminen näyttäisi antavan mahdollisuuksia mittausajan tuntuvaan supistamiseen etenkin sellaisilla sisältöalueilla, joilla on todettavissa selviä oppimishierarkioita.

Lisää huomiota tulee jatkossa kiinnittää mm. seuraaviin kysymyksiin:

1) kriteerimittareihin perustuvien pistemäärien raportoiminen
2) kriteerikokeiden laadinta ja käyttöohjeistojen laatiminen
3) normitietojen hyväksikäyttö kriteerimittauksessa
4) latenttien piirteiden mallien käyttö kriteerikokeiden laadinnassa, arvioinnissa seka kokeiden ja koepistemäärien tulkinnassa, seka
5) kriteerikokeiden laatijoiden ja käyttäjien koulutus

Kriteerimittaamisen alalla on siten useita tärkeitä alueita, joista tarvitaan tutkimuspohjaista tietoa.

11. DISKUSSIO

Kriteerimittaamisella ja normimittaamisella on ymmärrettävästi paljon yhteisiä piirteitä. Kuten monella muullakin alalla, mm. opetussuunnitelmien kehittämisessä, uudet lähestymistavat merkitsevät usein vain uudenlaisia painotuksia, mikä merkitsee että joitakin näkökohtia korostetaan aikaisempaa enemmän ja jotkut muut jäävät vähemmälle huomiolle tai todetaan jopa irrelevanteiksi. Alussa ollaan helposti taipuvaisia liioittelemaan eroja. Saattaa olla mahdollista, että eroavaisuuksien korostaminen on tiedon kasvamisen kannalta tarpeellista, ehkä suorastaan välttämätöntä, jonkin uuden idean vesillelaskun yhteydessä. Dogmaattisuudella saattaa siten olla oma tärkeä tehtävänsä tietyssä vaiheessa tiedon kasvun edistämisessä. Tästä löytyy esimerkkejä myös kriteerimittaamisessa. Aluksi esitettiin varsin kategorisesti, ettei kriteerimittaamisessa tarvita sellaisia käsitteitä kuin suorituksen hajonta tai varianssi, ettei empiirisiä osioanalyysejä juurikaan tarvita ja ettei normitietoja tarvitse tai tule kerätä. Kaikissa näissä kohdissa ollaan alkuajan tietynlaisen dogmaattisuuden jälkeen käsityksiä tarkistettu ja todettu, etteivät mainitut käsitteet ja menetelmät heikennä vaan pikemminkin tukevat kriteerimittaamista. Empiiriset menetelmät on hyväksytty loogis-rationalististen menetelmien tueksi ja vastapainoksi.

Mikä on kriteerimittauksen erityisansio? Mihin sillä erityisesti pyritään? Edellä esitetystä lienee käynyt ilmi, että kriteerimittaamisen keskeinen tarkoitus ja samalla ansio on, että se antaa tarkan kuvauksen henkilön asemasta tietyllä sisältö- ja käyttäytymisalueella, olipa kyseessä kognitiivinen, affektiivinen tai psykomotorinen alue. Täten se tiedollisella alalla antaa tarkan kuvauksen henkilön suoritustasosta koko tietyllä osa-alueella eikä pelkästään esitetyissä tehtävissä. Kriteerikoetta voisi luonnehtia sisältökeskeiseksi kokeeksi ja kriteerimittaamista sisältökeskeiseksi mittaamiseksi ja normikoetta puolestaan henkilökeskeiseksi kokeeksi. Kriteerikoe kuvaa mitä henkilöt osaavat tehdä ja tätä tietoa voidaan käyttää mm. henkilöiden asettamiseen keskinäiseen paremmuusjärjestykseen. Normikoe 1. erottelukoe asettaa henkilöt paremmuusjärjestykseen, mutta se ei kuvaa kunnolla henkilöiden suoritustasoa. Vain kunnollisella kriteerikokeella voidaan kuvata sitä, mitä oppilaat osaavat (kvalitatiivinen kuvaus). Vain kunnollisella kriteerikokeella voidaan kuvata myös sitä, kuinka paljon oppilaat osaavat (kvantitatiivinen kuvaus). Normikokeella voidaan kuvata vain osaamisen suhteellista määrää (A osaa enemmän kuin B: järjestys, erottelu), kun taas kriteerikokeella voidaan saada myös osaamisen määrän absoluuttisia arvioita (esim. peruskoulun päättyessä oppilaat osaavat englannin kielessä keskimäärin X sanaa aktiivisesti ja Y sanaa passiivisesti).

Mihin kriteerimittausta voidaan käyttää? Missä siitä voisi olla konkreettista hyötyä? Edellä esitetystä käy ilmi, että sitä voidaan käyttää yksityisen oppilaan suoritustason tarkkaan kuvaamiseen. Täten se luo pohjaa opetuksen eri vaiheissa tehtävillä päätöksille (mm. tarveanalyysit ja opetuksen yksilöllistäminen). Kun se antaa kelvollisen perustan yksilöä koskeville ratkaisuille, se luo pohjaa myös ryhmää koskeville päätöksille. Näin ollen kriteerimittaamisen tuloksia voidaan käyttää opetussuunnitelmien ja koko koulujärjestelmän evaluointiin.

Mitä kriteerimittaaminen edellyttää? Ensinnäkin on laadittava tai oltava käytössä mitattavan alueen tarkka kuvaus. Toiseksi on oltava käytettävissä tarkka mittarin kuvaus, joka sisältää ärsyke- ja reaktio-osan erittelyn ja reaktioiden pisteistyssysteemin. On laadittava tehtävät, jotka ovat valideja ja samalla edustava satunnaisotos tai ositettu satunnaisotos mitattavalta alueelta. Mikäli on kyseessä opetussuunnitelman tai koulujärjestelmän evaluointi, on lisäksi valittava edustava oppilasotos. Tällöin on edullista käyttää matriisiotantaa. Monitasotestaamista voidaan myös käyttää tehokkaasti hyväksi mittausajan supistamiseksi.

Kriteermittaaminen liittyy läheisesti tavoitteellisten opetusohjelmien kehitttämiseen ja koulukokeilujen yleistymiseen. Oli kehitettävä menetelmiä uusien opetusohjelmien tulosten arvioimiseen. Kriteerimittaaminen osuu yksiin myös ns. modernin testiteorian kehittymisen kanssa. Eräs selitys voi olla siinä, että monet opetuksen ja oppimisen ongelmien parissa työskennelleet asiantuntijat (Glaser, Cronbach, Bloom jne.) ovat olleet myös evaluoinnin ja mittaamisen asiantuntijoita. Modernin testiteorian kehittäjät ovat tuottaneet tehokkaita tilastomatemaattisia välineitä, jotka palvelevat myös kriteerimittaamista.

Kriteerimittaamiseen liittyvä, mitattavan alueen täsmentämistä koskeva voimakas korostaminen tukee myös tämän raportin kirjoittajan useassa yhteydessä esiintuomaa ajatusta, että ainekohtainen kehittämistoiminta on välttämätöntä koulun kehittämiseksi. Ainekohtainen jäsennystyö, jota kriteerimittaamisen termein kutsutaan aluetäsmennykseksi tai osa-aluetasmennykseksi (domain specification), luo tiedollista pohjaa opetuksen tavoitteiden ja oppisisältöjen määrittelylle, oppimateriaalin laadinnalle, opetuksen järjestämiselle ja lopulta *myös* oppimistulosten arvioinnille. Oppiaineksen jasentäminen ja täsmentäminen ei siten ole samaa kuin opetussuunnitelmien laatiminen tai niiden evaluoiminen. Se kyllä palvelee niitä, mutta sen tavoitteet ovat paljon laajemmat ja monipuolisemmat.

Moderni testiteoria ja kriteerimittaaminen edustavat uudenlaista ajattelutapaa. Ne tuovat terävästi esille ongelmia mutta antavat myös uusia mahdollisuuksia. Nämä mahdollisuudet joudutaan kyllä ostamaan, niitä ei saada ilmaiseksi. Kriteerimittaaminen ja modernin testiteorian soveltaminen on vaikeampaa ja vaativampaa kuin normimittaaminen ja klassisen testiteorian soveltaminen. Mitattava alue on tunnettava ja ymmärrettävä syvällisesti. Vaativuudesta ja vaikeudesta saattaa olla positiivisena seikkana se, ettei ole mahdollista menetellä mekaanisesti "keittokirjareseptejä" noudattaen. On *myös* ajateltava. Mittaamisen vaativuus tajutaan ja tästä lienee seurauksena mittauksen tason kohoaminen. Samalla on mahdollisuus osoittaa mittaamisen kriitikoille entista perustellummin, missä kohden heidän kritiikkinsä on kritiikitöntä ja missä kohdin aiheellista.

LÄHTEET

Airasian, W., and Madaus, F. (1972) Criterion -referenced testing in the classroom. Measurement in Education, 3, 1-8.

Anderson, C. (1972) How to construct achievement tests to assess comprehension. Review of Educational Research, 42, 145-170.

Andrew, B.J., & Hecht, J.T.A. (1976) A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-50.

Angoff, W.H. (1971) Scales, norms and equivalent scores. Teoksessa R.L. Thorndike (toim.) Educational measurement. Washington, D.C.: American Council of Education.

Baker, E.L. (1974) Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. Educational Technology, 14, 10-16.

Baker, R. (1974) Measurement considerations in instructional product development. Teoksessa Harris, C.W., Alkin, M.C. & Popham, W.J. Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.

Berk, R.A. (1976) Determination of optional cutting scores in criterion-referenced measurement. Journal of Experimental Education, 45, 4-9.

Block, J.H. (1972) Student evaluation: Toward the setting of mastery performance standards. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Bloom, B.S., Hastings, J.T., Madaus, G.F. (1971) Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill.

Bormuth, R, (1970) On the theory of achievement test items. Chicago: University of Chicago Press.

Bormuth, J.R. (1971) Development of standards of readability: Toward a rational criterion of passage performance. Final report, U.S. Office of Education, Project No. 9-0237. Chicago: University of Chicago.

Brennan, R.L. & Kane, M.T. (1977) An index of dependability for mastery tests. Journal of Educational Measurement 14, 277-289.

Brennan, R.L. & Kane, M.T. Signal noise ratios for domain-referenced tests. Psychometrika.

Carver, R.P. (1970) Special problems in measuring change with psychometric devices. Teoksesta Evaluative research: Strategies and methods. Washington: American Institutes for Research.

Carver, R.P. (1974) Two dimensions of tests: Psychometric and edumetric. American Psychologist, 29, 512-518.

Cohen, J.A. (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cox, C., and Sterrett, G. (1970) A model for increasing the meaning of standardized test scores. Journal of Educational Measurement, 7, 227-228.

Cronbach, L.J. (1975) Dissent from Carver. American Psychologist, 30, 602-603.

Cronbach, L.J. & Furby, L. (1970) How we should measure change - or should we? Psychological Bulletin, 74, 68-80.

Cronbach, L.J. (1971) Test validation. Teoksessa R.L. Thorndike, Educational Measurement. Washington, D.C.: American Council of Education.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972) The dependability of behavioral measurements. Theory of generalizability for scores and profiles. New York: Wiley.

Denham, C.H. (1975) Criterion-referenced, domain-referenced and norm-referenced measurement: a parallax view. Educational Technology, 15, 9-13.

Durnin, J. & Scandura, J.M. (1973) An algorithmic approach to assessing behavior potential. Journal of Educational Psychology, 65, 262-272.

Ebel, (1962) Content standard test scores. Educational and Psychological Measurement, 22, 15-25.

Ebel, R.L. (1971) Criterion-referenced measurements: limitations. School Review, 69, 282-288.

Ebel, R.L. (1972) Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R. (1973) Evaluation and educational objectives. Journal of **Educational Measurement**, 10:273-279.

Ferguson, R.L. (1969) The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh.

Ferguson, R.L. (1971) Computer assistance for individualizing measurement. Learning Research and Development Center, University of Pittsburgh.

Glaser, R. (1963) Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 18:519-521.

Glaser, R. & Nitko, A. (1971) Measurement in learning and instruction. Teoksessa R.L. Throndike (toim.), Educational Measurement. Washington D.C.: American Council on Education, 652-670.

Glaser, R. (1976) The processes of intelligence and education. Teoksessa L.B. Resnick (toim.) The nature of intelligence. Hillsdale, N.J.: Lawrence Erlbaum.

Gray, W.M. (1978) A comparison of Piagetian theory and criterion-referenced measurement. Review of Educational Research, 48, 2, 223-249.

Gronlund, N.E. (1977) Constructing achievement tests. Englewood Cliffs, N.J.: Prentice-Hall.

Guttman, Louis (1969) Integration of test design and analysis. Teoksessa Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service.

Hambleton, R.K., Swaminathan, H., Algina, J. & Coulson, D.B. (1978) Criterion-referenced testing and measurement: a review of technical issues and developments. Review of Educational Research, 48, 1, 1-47.

Hambleton, R.K. & Eignor, D.R. (1978) Guidelines for evaluating criterion-referenced tests and test manuals. Journal of Educational Measurement, 15, 4, 321-327.

Hambleton, R.K. (1974) Testing and decision-making procedures for selected individualized instructional programs. Review of Educational Research, 44, 371-400.

Hambleton, R.K. & Novick, M.R. (1973) Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, 159-170.

Harris, C.W. (1972) An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, 9, 27-29.

Harris, C.W. (1974) Problems of objectives-based measurement. Teoksessa Harris, C.W., Alkin, M.C. & Popham, W.J. (toim.), Problems in criterion referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.

Henrysson, S. & Wedman, 1. (1974) Some problems in construction and evaluation of criterion-referenced tests. Scandinavian Journal of Educational Research, 18, 3-12.

Hively, E., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. (1973) Domain-referenced curriculum evaluation: A technical handbook and a case study from the Minnemast Project. CSE monograph series in evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California.

Hively, W., Patterson, H.L. & Page, S.A. (1968) A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 5, 275-290.

Janson, S. (1975) Undervisningsmål som utgångspunkt vid konstruktion av målrelaterade prov: några teoretiska och empiriska problem. Umeå universitet. Pedagogiska institutionen.

Klein, P. & Kosecoff, J. (1973) Issues and procedures in the development of criterion-referenced tests. ERIC Clearinghouse on Sests, Measurement & Evaluation, TM Report 26. Princeton, N.J.: Educational Testing Service.

Konttinen, R. (1981) Testiteoria. Johdatatusta kasvatus- ja käyttäytymistieteellisen mittauksen teoriaan (Gaudeamus, painossa).

Korpinen, E. (1976) Sanallisen tiedotteen kehittäminen peruskoulun ala-asteella. Jyväskylän yliopisto. Kasvatustieteiden tutkimuslaitoksen julkaisuja No. 264.

Korpinen, E. (1977) Oppilaiden huoltajien käsityksiä sanallisen tiedotteen kehittämisestä peruskoulun ala-asteelle. Jyväskylän yliopisto. Kasvatustieteiden tutkimuslaitoksen julkaisuja No, 271.

Korpinen, E. (1977) Sanallinen arviointi peruskoulun 1.-3. luokilla lukuvuonna 1976-77 1. Jyväskylän yliopisto. Kasvatustieteiden tutkimuslaitoksen julkaisuja No. 281.

Korpinen, E. (1977) Sanallinen arviointi peruskoulun 1.-3.-luokilla lukuvuonna 1976-77 II. Jyväskylän yliopisto. Kasvatustieteiden tutkimuslaitoksen julkaisuja No. 282.

Korpinen, E. (1978) Sanallisen arvioinnin seka kodin ja koulun yhteistyökokeilu 1. Kokeiluohjelma. Jyväskylän yliopisto. Kasvatustieteiden tutkimuslaitoksen selosteita ja tiedotteita No. 108.

Korpinen, E. & Lång, A. (1979) Sanallisen arvioinnin seka kodin ja koulun yhteistyökokeilu II. Jyväskylän yliopisto. Kasvatustieteiden tutkimuslaitoksen selosteita ja tiedotteita No. 132.

Korpinen, E. (1979) Sanallisen arvioinnin seka kodin ja koulun yhteistyökokeilu III. Jyväskylän yliopisto. Kasvatustieteiden tutkimuslaitoksen selosteita ja tiedotteita No. 334.

Lewis, C., Wang, M., & Novick, R. (1973) Marginal distributions for the estimation of proportions in m groups. Technical Bulletin, No. 13. Iowa City: American College Testing Program.

Livingston, S.A. (1972) Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 9, 13-26.

Lord, F.M. & Novick, M.R. (1968) Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Lord, F.M. (1976) Testtheoryand the public interest. Proceedings of the 1976 ETS invitational conference. Princeton, N.J.: Educational Testing Service, 17-30.

Macready, G.B. (1975) The structure of domain hierarchies found within a domain referenced testing system. Educational and Psychological Measurement 35, 583-598.

Meskauskas, J.A. (1976) Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 46, 133-158.

Meskauskas, J.A. & Webster, G.W. (1975) The American Board of Internal Medicine recertification examination process and results. Annals of Internal Medicine, 82, 577-581.

Messick, S.A. (1975) The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Millman, J. (1970) Reporting student progress: A case for a criterion-referenced marking system. Phi Delta Kappan, 52, 226-230.

Millman, J. (1973) Passing scores and test lengths for domain-referenced tests. Review of Educational Research, 43, 205-216.

Millman, J. (1974) Criterion-referenced Measurement. Teoksessa W.J. Popham (toim.) Evaluation in Education: Current Applications. Berkeley:McCutchan.

Millman, J. (1978) Determinants of item difficulty: a preliminary investigation. Center for the Study of Evaluation, CSE Report No. 114.

Nedelsky, L. (1954) Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Novick, R., Lewis, C. & Jackson, H. (1973) The estimation of proportions in m groups. Psychometrika, 38, 19-46.

Novick, R., Lewis, C. (1974) Prescribing test length for criterion-referenced measurements. Teoksessa W. Harris, C. Alkin, and W. Popham (toim.), Problems in Criterion-Referenced Measurement. Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California.

Osburn, H.G. (1968) Item sampling for achievement testing. Educational and Psychological Measurement, 28, 95-104.

Popham, W.J. & Husek, T.R. (1969) Implications of criterion-referenced measurement. Journal of Educational Measurement, 6, 1-9.

Popham, W.J. (1972) Selecting objectives and generating test items for objectives based tests. Paper presented at Conference on Problems in Objectives Based Measurement. Center for the Study of Evaluation, UCLA.

Popham, W.J. (1974) Selecting objectives and generating test items for objectives based tests. Teoksessa Harris, C.W., Alkin, M.C., and Popham, W.J. (toim.) Problems in criterion referenced measurement,

CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation.

Popham, W.J. (1974) An approaching peril: Cloud-referenced tests. Phi Delta Kappan, 55, 9, 614-615.

Popham, W.J. (1975) Educational Evaluation. New Jersey: Prentice Hall. Englewood Cliffs.

Popham, W.J. (1976) Normative data for criterion-referenced tests? Phi Delta Kappan, 58, 593-594.

Popham, W.J. (1978) The standardized test flap flop. Phi Delta Kappan, 59, 7, 470-471.

Popham, W.J. (1978) Criterion-referenced rneasurement. Englewood Cliffs, New Jersey: Prentice Hall.

Rovinelli, R.J. & Hambleton, R.K. (1977) On the use of content specialists in the assesment of criterion-referenced test item validity. Dutch Journal for Educational Research, 2, 49-60.

Schutz, R.E. (1978) The design of measurement in instruction. Center for the Study of Evaluation, Evaluation Comment, Vol. 5, No. 4.

Scriven, M. (1967) Aspects of curriculum evaluation. Teoksessa Tyler, R. (toim.), Perspectives of Curriculum Evaluation. Chicago: Rand McNally.

Shavelson, R.J., Block, J.H. & Ravitch, M.M. (1972) Criterion-referenced testing: Comments on reliability. Journal of Educational Measurement, 9, 133-137.

Smith, R. & Tyler, R.W. (1942) Appraising and recording student progress. New York: Harper.

Smith, S. (1978) Decisions and dilemmas in constructing criterion-referenced tests: some questions and issues. Center for the Study of Evaluation, CSE Report No. 110.

Spinetti, J.A. (1973) A computer simulation study of tailored testing strategies for objective-based instructional programs. Unpublished doctoral dissertation, University of Massachusetss.

Swaminathan, H., Hambleton, R.K. & Algina, J.A. (1975) Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 12, 87-98.

Takala, S. (1976) Vieraiden kielten opetussuunnitelman, opettamisen ja oppimisen kysymyksiä. Jyväskylän yliopisto. Kasvatustieteiden tutkimuslaitoksen selosteita ja tiedotteita No. 68.

Takala, S. (1980) Vaatimustasojen määritteleminen opetussuunnitelmia laadittaessa. Jyväskylän yliopisto. Kasvatustieteiden tutkimuslaitoksen selosteita ja tiedotteita No. 145.

Tyler, R.W. (1973) Testing for accountability. Teoksessa A.C. Ornstein (toim.), Accountability for teachers and school administrators. Belmont, CA: Feardon Publishers.

Vahervuo, T. (1948) Oppilaitosten oppilasarvostelu. Kasvatusopillinen aikakauskirja, 85, 1, 5-27.

Vahervuo, T. (1951) Suhteellinen ja absoluuttinen arvostelusysteemi. Kasvatusopillinen aikakauskirja, 88, 5, 238-243.

Vahervuo, T. (1958) Arvosanojen antaminen. Helsinki: Otava.

Valette, R.M. (1971) Evaluation of learning in a second language. Teoksessa Bloom et al., 815-853.

Walker, C.B. (1978) Standards for evaluating criterion-referenced tests. Center for the Study of Evaluation, CSE Report No. 103.

Wedman, 1. (1973) Kriterierelaterade prov: Bakgrund, egenskaper och begränsningar. Pedagogiska rapporter, Umeå, nr 33.

Wedman, 1. (1973) Reliabilitets-, validitets- och diskriminationsmått för kriterierelaterade prov. Pedagogiska rapporter, Umeå, nr 34.

Wedman, 1. (1973) Mätproblem i norm- och kriterierelaterade prov. Pedagogiska rapporter, Umeå, nr 35.

Tiivistelmäkortti

Takala, S. (1980) Kriteerimittaamisen käsitteestä ja käytännön sovelluksista. Kasvatustieteiden tutkimuslaitos. Selosteita ja tiedotteita 146. Jyväskylän yliopisto. **ISBN** 951-678-337-6 ISSN 0357-122X

Raportti on osa tutkijan vieraiden kielten opetussuunnitelmaa käsittelevää tutkimusohjelmaa ja liittyy peruskoulun ensimmäiseen tilannekartoitukseen. Se täydentää myös tekijän aikaisempaa julkaisua, jonka teemana oli vaatimustasojen asettaminen opetussuunnitelmia laadittaessa. Raportissa tehdään selkoa kriteerimittaamisen (KRM) lähtökohdista, käsitteistä ja niiden kehittymisestä ja verrataan kriteerimittaamista perinteiseen normimittaamiseen (NRM). KRM:n suurimmaksi ansioksi todetaan se, että se antaa tarkan kuvauksen henkilön suoritustasosta tarkasti määritellyllä sisältöalueella. KRM edellyttää runsasta ja tiukkaa ennakkosuunnittelua, jossa sisällön asiantuntemus näyttelee suurta osaa. Empiiristä osioanalyysiä käytetäänkin vain puutteellisten osioiden paljastamiseen mutta ei osioiden karsinnan perusteena. Empiriaan perustuva osioiden karsinta heikentäisi sisällön (=kuvauksen) validiteettia, joka on tärkein validiteetin muoto KRM:ssa. KRM:ta voidaan käyttää hyväksi tarveanalyyseja suoritettaessa ja opetusta yksilöllistäessä. Erityisen lupaavalta näyttää sen käyttö opetussuunnitelmia ja koulujärjestelmän laadullista tuotosta arvioitaessa. KRM:n alueella on useita avoimia kysymyksiä, jotka vaativat osakseen tutkijoiden huomiota.

Hakusanat: evaluaatio, koetoiminta, mittaaminen, mittaustekniikka, metodologia, kriteerimittaaminen, normimittaaminen, moderni testiteoria

Abstract card

Takala, S. (1980) Kriteerimittaamisen käsitteestä ja käytännön sovelluksista. - On the concept and practical applications of criterion-referenced measurement. Institute for Educational Research. Bulletin 146. University of Jyväskylä, Finland. ISBN 951-678-337-6. ISSN 0357-122X

The report is part of the author's larger research programme dealing with some theoretical and practical problems related to the construction of FL curricula. It supplements an earlier report which dealt with the problem of setting standards in curriculum construction. The report deals with the concept and development of criterion-referenced measurement (CRM) and compares it with norm-referenced measurement (NRM). It is suggested that the greatest advantage of CRM is that it gives a good decription of a persons's status with respect to a well-defined domain of behaviors and content. CRM demands a lot of careful planning before the test is administered. Empirical item analyses are, accordingly, used only to suppiement content expertise in order to detect flawed items. It is not used to screen or select items, because this would jeopardize descriptive or content validity. CRM can be used for many purposes but especially promising is its use in curriculum evaluation and in the evaluation of the quality of the educational system. There are still ceveral problems in CRM, which need to be addressed by research. (In Finnish, English summary)

Descriptors: evaluation, test, measurement, meacurement technique, methodology

Takala, S. (1980) Kriteerimittaamisen käsitteestä ja käytännön sovelluksista. Kasvatustieteiden tutkimuslaitos. Selosteita ja tiedotteita 146. Jyväskylän yliopisto. ISBN 951-678-337-6 ISSN 0357-122X

Raportti on osa tutkijan vieraiden kielten opetussuunnitelmaa käsittelevää tutkimusohjelmaa ja liittyy peruskoulun ensimmäiseen tilannekartoitukseen. Se täydentää myös tekijän aikaisempaa julkaisua, jonka teemana oli vaatimustasojen asettaminen opetussuunnitelmia laadittaessa. Raportissa tehdään selkoa kriteerimittaamisen (KRM) lähtökohdista, käsitteistä ja niiden kehittymisestä ja verrataan kriteerimittaamista perinteiseen normimittaamiseen (NRM). KRM:n suurimmaksi ansioksi todetaan se, että se antaa tarkan kuvauksen henkilön suoritustasosta tarkasti määritellyllä sisältöalueella. KRM edellyttää runsasta ja tiukkaa ennakkosuunnittelua, jossa sisällön asiantuntemus näyttelee suurta osaa, Empiiristä osioanalyysiä käytetäänkin vain puutteellisten osioiden paljastamiseen mutta ei osioiden karsinnan perusteena. Empiriaan perustuva osioiden karsinta heikentäisi sisällön (=kuvauksen) validiteettia, joka on tärkein validiteetin muoto KRM:ssa. KRM:ta voidaan käyttää hyväksi tarveanalyyseja suoritettaessa ja opetusta yksilöllistäessä. Erityisen lupaavalta näyttää sen käyttö opetussuunnitelmia ja koulujärjestelmän laadullista tuotosta arvioitaessa. KRM:n alueella on useita avoimia kysymyksiä, jotka vaativat osakseen tutkijoiden huomiota.

Hakusanat: evaluaatio, koetoiminta, mittaaminen, mittaustekniikka, metodologia, kriteerimittaaminen, normimittaaminen, moderni testiteoria

Takala, S. (1980) Kriteerimittaamisen käsitteestä ja käytännön sovelluksista. - On the concept and practical applications of criterion-referenced measurement. Institute for Educational Research. Bulletin 146. University of Jyväskylä, Finland. ISBN 951-678-337-6. ISSN 0357-122X

The report is part of the author's larger research programme dealing with some theoretical and practical problems related to the construction of FL curricula. It supplements an earlier report which dealt with the problem of setting standards in curriculum construction. The report deals with the concept and development of criterion-referenced measurement (CRM) and compares it with norm-referenced measurement (NRM). It is suggested that the greatest advantage of CRM is that it gives a good decription of a persons's status with respect to a well-defined domain of behaviors and content. CRM demands a lot of careful planning before the test is administered. Empirical item analyses are, accordingly, used only to supplement content expertise in order to detect flawed items. It is not used to screen or select items, because this would jeopardize descriptive or content validity. CRM can be used for many purposes but especially promising is its use in curriculum evaluation and in the evaluation of the quality of the educational system. There are still several problems in CRM, which need to be addressed by research. (In Finnish, English summary)

Descriptors: evaluation, test, measurement, measurement technique, methodology

Kasvatustieteiden tutkimuslaitoksen **julkaisusarjoissa** ilmestyneet tämän raportin alaan liittyvät muut tutkimukset

Kasvatustieteiden tutkimuslaitoksen julkaisuja ISSN 0448-0953

Rapporter från Pedagogiska forskningsinstitutet

Reports from the Institute for Educational Research

15/1965 Pentti Pitkänen: Kielitaidon luokitusjärjestelmä ja sitä vastaava koetehtävätyypistö kielitaidon psykometrista mittaamista varten (40 s.) 6,50

36/1967 Valter Mäkelä - Pentti Pitkänen - Manu Renko: Eräiden muuttujien yhteydet ruotsin kielen alkeiden oppimiseen kansakoulun III luokalla. Väliseloste. - The relations between some variables and the learning of elementary Swedish in III grade of elementary school. Preliminary report (5 s.) (loppuunmyyty)

57/1970 Anthony May: Time for a change? A critical analysis of two broadcasts for schools: 'Time for English II'. - Muutoksen aika (+) (33 s.) 5,50

62/1970 oiva Ylinentalo: Lukutaito ja lukutaidos,a ilmenevät erot kansakoulun keski- ja yläasteella. - Reading ability and the differences in it in the middle and upper primary School (+) (44 s.) 7,-

70/1970 Raimo Konttinen: Opiskelijoiden englanninkielen taitojen ja niiden oppimisen yhteydet verbaaliseen lahjakkuuteen ja persoonallisuuden piirteisiin. - Relations of university students' previous attainment and learning of English to verbal aptitude and personality traits (38 s.) 6,50

86/1971 Jaakko Lehtonen: Fonologia ja kielenopetus. - Phonology and language teaching (51 s.) (n:o 168/1972 korvaa). 8,-

95/1971 Anthony May - David Wilson: English for the Upper Classes of the Finnish Secondary School: a New Approach. -(Suomen-kielinen lyhennelmä +) (68 s.) 10,-

115/1971 Sauli Takala: Kokeiluperuskoulujen yhteiset vieraiden kielten kokeet 1970-1971. - Achievement testing in foreign languages in 1970-1971 (24 s.) 4,50

119/1971 Sauli Takala: Tasokurssivalinnat, kurssien vaihtaminen ja tukiopetukseen osallistuminen oppilaalle vieraissa kielissä. - Course choices and changes and participation in remedial instruction in foreign languages (25 s.) 4,50

129/1972 Jorma Kuusinen: Luku- ja kirjoitushäiriöisten ja häiriöttömien psykolingvistiset kyvyt. - The psycholinguistic abilities of children suffering and not suffering from reading and writing disorders (39 s.) ISBN 951-677-018-5 6,50

145/1972 Glyn Hughes: Errors made by Finns in the translation of the six local cases: an analysis. - Suomalaisten kuuden paikallissijan kääntämisessä tekemät virheet: analyysi (39 s.) ISBN 951-677-043-6 6,50

156/1972 Jorma Kuusinen - Lea Blåfield: ITPA:n teoria, ominaisuudet ja käyttö. - The theory, characteristicc and use of ITPA (105 s.) ISBN 951-677-060-6 14,50

159/1972 Arvo Koponen: Vieraan kielen koulusaavutukset hajotetussa ja periodiopetuksessa 1. - Achievement in foreign languages in distributed and concentrated learning 1 (56 s.) ISBN 951-677-065-7 8,50

165/1972 Arvo Koponen: Asenteet ja asennemuutokset vieraita kieliä kohtaan hajotetussa ja periodiopetuksessa. - Attitudes and attitude changes towards foreign languages in distributed and concentrated teaching (25 s.) ISBN 951-677-076-2 4,50

168/1972 Jaakko Lehtonen: Kielenopetuksen fonoiogiaa. - Phonology of language teaching (+) (68 s.) ISBN 951-677-081-9 10,-

169/1972 Arvo Koponen: Vieraan kielen koulusaavutukset hajotetussa ja periodiopetuksessa II. - Achievement in foreign languages in distributed and concentrated teaching II (18 s.) ISBN 951-677-082-7 4,-

172/1973 Dieter Stellmacher: Grammatiktheorie und Sprachunterricht. - Grammatikteori och språkundervisning (47 s.) ISBN 951-677-086-X 7,50

179/1973 Matti Leivo: Kielitiede ja äidinkielen opetus. - Språkvetenskap och modersmålsundervisning. - Linguistics and the mother tongue teaching (29 s.) ISBN 951-677-098-3 5,-

194/1973 Sauli Takala: Ruotsin kielen koulusaavutukset peruskoulun ala-asteen ja yläasteen päättyessä. - Elevernas skolprestationer i svenska vid utgången av grundskolans låg- och högstadium. - Pupils' knowledge of Swedish at the end of the lower and upper level of the comprehensive School (+) (64 s.) ISBN 951-677-151-3 9,50

221/1974 Kimmo Leimu: Opetustoimen evaluaatiotyön hahmotusta ja muuan sovellutus. - En skissering av evalueringen inom undervisningsväsendet samt något om den praktiska tillämpningen. - A conception of educational evaluation and its application to Finnish conditions (+) (51 s.) ISBN 951-677-272-2 7,50

231/1974 Liisa Havola-Pitkänen: Englanninkielen koulusaavutukset. peruskoulun ala-asteen ja yläasteen päättyessä. - Skolprestationerna i engelska vid utgången av grundskolans låg- och högstadium (+). - Pupils' knowledge of English at the end of the lower and upper level of the comprehensive school (+) (97 s.) ISBN 951-677-327-3 12,50

236/1974 Pehr-Olof Rönnholm: Feltyper i rättskrivning. - Oikeinkirjoituksen virhetyypeistä (+).- Error types in spelling (+) (54 s.) ISBN 951-677-342-7 7,50

283/1977 Anneli Vähäpassi: Lukutaidon rakenteesta ja vaihtelusta peruskoulun kolmannella luokalla lukuvuonna 1973-74. - Läsfärdighetens struktur och variationer i läsfärdighet i grundskolans tredje årskurs läsåret 1973-74 (+). - On the structure and variability of reading skill in grade 3 of the comprehensive School in School year 1973-74 (+) (196 s.) ISBN 951-677-911-5 23,50

287/1978 Seppo Pietilä: Malliharjoitusmenetelmä oikeinkirjoituksen opetuksessa. - Skrivövningar efter modellord som en metod för undervisningen i rättskrivning (+). - Copying of written words as a method of teaching orthography (+) (192 s.) ISBN 951-677-944-1 23,-

290/1978 Kaija Kärkkäinen - Sauli Takala: Ylioppilastutkinnon kielikokeen rakennekoetyyppitutkimus. - Kartläggning av olika strukturprovstyper och deras tillämpbarhet i studentexamens språkprov (+). - A feasibility study of incorporating a structures test in the matriculation examination (+) (217 s.) ISBN 951-677-987-5 25,50

295/1979 Sauli Takala - Hannu Saari: Englannin kielen opetus ja koulusaavutukset Suomessa 1970-luvun alussa: IEA:n kansainväliseen yhteistoimintaan perustuva vertaileva ja kuvaileva tutkimus. - The teaching of English in Finland at the beginning of the 1970's: a comparative and descriptive study based on the IEA data (+) (182 s.) ISBN 951-678-119-9 21,50

Pedagogiska forskningsinstitutet. Notiser och rapporter

Institute for Educational Research. Bulletin

49/1975 Viking Brunell: Grundskolans centrala prov i läsförståelse i årskurserna 3-6 läsåret 1974-75. - Luetun ymmärtämisen arviointi vuosiluokilla 3-6 lukuvuonna 1974-75 (+). - Assessment of reading comprehension in grades 3-6 in 1974-75 (+) (116 s.) ISBN 951-677-533-0 14,50

62/1976 Liisa Havola-Pitkänen: Englannin kielen koulusaavutukset peruskoulun 6. luokan päättyessä. - Skolprestationerna i engeiska vid utgången av årskurserna 4, 5 och 6 i grundskolan (+). - Pupils' knowledge of English at the end of the 4th, 5th and 6th grades of the comprehensive school (+) (89 s.) ISBN 951-677-598-5 ..... 11.50

67/1976 Kaija Kärkkäinen: Peruskoulun yhteiset kokeet lukuvuonna 1974-75. Yläasteen ruotsin kielen kokeet ja koetuloksia. Ruotsin kieli toisena oppilaalle vieraana kielenä. - De centrala proven i Svenska på grundskolans högstadium läsåret 1974-75 (+). - Centrally administered achievement testc in the comprehensive school in academic year 1974-75. Tests and test resultc in Swedish at the upper level (+) (81 s.) ISBN 951-677-637-X 10,50

68/1976 Sauli Takala: Vieraiden kielten opetussuunnitelman, opettamisen ja oppimisen kysymyksiä. - Om läroplanen, undervisning och inlärning i främmande språk (+). - Some notes on foreign language syllabus, language teaching and learning (+) (154 s.) ISBN 951-677-641-8 18,50

73/1976 Viking Brunell: Grundskolornas centrala prov i modersmålet läsåret 1974-75. Provet i språkiakttagelser i årskurs 9. - Peruskoulun yhteiset äidinkielen kokeet (ruotsi) lukuvuonna 1974-75. IX luokan kielentuntemuksen kokeet (+). - Centrally administered achievement tests in mother tongue (Swedish) in 1974-75. The test for measuring the linquistic knowledge of the ninth grade pupils of comprehensive schools (+) (84 s.) ISBN 951-677-679-5 11,-

74/1976 Pentti Hakala - Kaija Kärkkäinen: Ruotsin kielen koulusaavutukset peruskoulun ala-asteen ja yläasteen päättyessä lukuvuonna 1972-73. - Skolprestationerna i Svenska vid utgången av grundskolans låg- och högstadium läsåret 1972-73 (+).- School achievements in Swedish at the end of the lower and the upper stage of the comprehensive School in 1972-73 (+). (260 s.) ISBN 951-677-680-9 .....

77/1976 Kimmo Leimu - Hannu Saari: Koulukoe- ja arviointitoiminnan kehittämisestä Suomessa. - Om utvecklandet av skolprovs- och evalueringsverksamheten i Finland (+). - On the development of School achievement testing and evaluation in Finland (+) (162 s.) ISBN 951-677-709-0 19,50

82/1977 Kaija Kärkkäinen: Keskiasteen koulujen yhteinen koe lukuvuonna 1973-74. Lukion kielen koe ja koetuloksia. - De centrala proven i Svenska i gymnasiet läsåret 1973-74 (+). - Centrally administered achievement tests in Swedish in the upper secondary School in academic year 1973-74 (+) (69 s.) ISBN 951-677-765-1 10,-

83/1977 Marja-Leena Husso - Eira Korpinen: Lukemisen ja kirjoittamisen valmiustestin kehittämisestä koulutulokkaille. Esikokeilun tulokset. - Development of a reading readiness test for School beginners. Results of pilot test (+) (74 s.)
ISBN 951-677-769-4 10,-

88/1977 Anneli Vähäpassi: Peruskoulun ja keskiasteen koulujen äidinkielen oppimistulosten arviointiprojekti. Luku- ja kirjoitustaidon tasosta peruskoulun kuudennella luokalla lukuvuonna 1974-75. - Utvärdering av elevernas läs- och skrivfärdigheter i grundskolans sjätte årskurs läsåret 1974-75. De centraia proven och provresultaten i modersmålet (finska) på grundskolans lågstadium (+). - The level of reading and writing in grade 6 of comprehensive School, 1974-75 (+) (133 s.)
ISBN 951-677-799-6 .....

92/1977 Kaija Kärkkäinen: Peruskoulujen yhteinen ruotsin kielen koe lukuvuonna 1973-74. Yläasteen päättövaiheen puheen ymmärtämisen koe ja koetuloksia. - De centrala proven i svenska (det för eleven andra främmande språket) läsåret 1973-74. Provet i hörförståelse vid utgången av grundskolans högstadium (+). - Results of the centrally administered test of Swedish listening comprehension in the final grade of the comrehensive school in 1973-74 (+) (138 s.)
ISBN 951-677-844-5 17,-

94/1977 Sauli Takala: Piirteitä vieraiden kielten opetuksesta. - Några aspekter på undervisningen i främmande språk (+). - Some aspects of FL teaching: (+) (127 s.)
ISBN 951-677-856-9 15,50

95/1977 Kaija Kärkkäinen - Sauli Takala: Vieraiden kielten rakenteiden hallintaa mittaavia koetyyppejä. - Kartläggning av strukturprovstyper. - Testing students' knowledge of grammaticai structures: an inventory of test types (105 s.)
ISBN 951-677-859-3 13,50

102/1977 Kaija Kärkkäinen: Keskiasteen koulujen yhteinen koe syksyllä 1976. Lukion ruotsin kielen lähtötasokoe ja koetuloksia. - De crntrala proven för mätning av utgångsnivån i svenska (det för eleven andra främmande språket) i gymnasiet hösten 1976 (+). - Centrally administered achievement, test, in secondary schools in autumn 1976. Pupils' knowledge of Swedish on entering the upper secondary School (+) (166 s.)
ISBN 951-677-899-2 20,-

109/1978 Pirjo Linnakylä: Taksonominen tavoitekuvaus äidinkielen oppimistulosten arvioinnin lähtökohtana. Peruskoulun yhteiset äidinkielen kokeet lukuvuonna 1972-73. - En taxonomisk målbeskrivning som utgångspunkt för utvärdering av inlärningsresultaten i modersmålet. Grundskolans centrala prov i modersmålet (finska) läsåret 1972-73 (+). - Taxonomic description of goals as the basis of achievement evaluation in mother tongue. Centrally administered achievement tests in mother tongue in the comprehensive School during School year 1972-73 (234 s.) ISBN 951-678-051-2 27,50

115/1978 Liisa Ahonen - Anja Korhonen - Terttu Marttila - Teuvo Piippo - Anneli Rusanen & Seija Värre: Peruskoulun kolmannen ja neljännen luokan englanninopetus sekä englannin- ja äidinkielen opetuksen integrointi. Opetussuunnitelman toteutusviitteitä ja -materiaalia. - Om engelskundervisningen i grundskolans tredje och fjärde årskurs och ett försök att integrera engelskundervisningen med undervisningen i modersmålet. Ett förslag till läroplansutveckling (+). - The teaching of English with special emphasis on integration with the teaching of mother tongue in grades 3 and 4 of the comprehensive School (+) (394 s.) ISBN 951-678-091-1 45,-

137/1980 Kaija Kärkkäinen - Pentti Määttä - Riitta Siimes - Sauli Takala: Erilaajuisia oppimääriä suorittaneiden menestyminen vuoden 1979 ylioppilaskirjoitusten kielikokeissa. - The relative success of former comprehensive vs. secondary School pupils in the FL tests included in the matriculation examination in 1979. (+) (173 s.)
ISBN 951-678-295-7 21,-

145/1980 Sauli Takala: Vaatimustasojen määritteleminen opetussuunnitelmia laadittaessa. - Standard setting in curriculum construction. (+) (55 s.) ISBN 951-678-336-8 8,-

146/1980 Sauli Takala: Kriteerimittaamisen käsitteestä ja käytännön sovelluksista. - On the concept and practical applications of criterion-referenced measurement (+) (128 s.)
ISBN 951-678-337-6 16,-

(+) merkityissä julkaisuissa on ko. kielinen pitempi tiivistelmä.
Julkaisuja voi tilata laitoksen osoitteella

Rapporter märkta med (+) är försedda med längre svenskspråkig sammanfattning.
Rapporterna kan beställas från Pedagogiska forskningsinstitutet

Reports marked with (+) have a longer English summary.
The reports can be ordered from the Institute for Educational Researcli

Tilaukset: Kasvatustieteiden tutkimuslaitos, Jyväskylän yliopisto
40100 Jyväskylä 10

Puh. 941-292378

w