Beoordelingssysteem

voor de

Kwaliteit van Tests

Deel 2

* Ingekort en aangepast voor de opdracht Diagnostiek en Testtheorie

INLEIDING

1. Inhoud van het beoordelingssyteem

Een test wordt beoordeeld op vijf categorieën. In dit vervolg worden de laatste twee besproken. Elke categorie bestaat uit meerdere vragen, waaronder één of meer basisvragen. Met behulp van de basisvragen wordt vastgesteld of aan bepaald minimum vereisten is voldaan, zonder welke verder beoordeling van de betreffende categorie overbodig wordt of niet mogelijk is.

Het tweede deel van de beoordeling van een test leidt tot een waardering op de volgende aspecten:

4. Betrouwbaarheid. Deze categorie bestaat uit drie vragen, waarvan één basisvraag. De tweede vraag is verdeeld in vier subvragen waarin de uitkomsten van verschillende typen betrouwbaarheidsonderzoek worden beoordeeld. Met behulp van de derde vraag wordt de kwaliteit van het uitgevoerde onderzoek bij het oordeel betrokken.

5a. Begripsvaliditeit. Deze categorie telt drie vragen, waarvan één basisvraag. Eerst worden de uitkomsten en vervolgens de kwaliteit van het uitgevoerde onderzoek naar de begripsvaliditeit beoordeeld.

5b. Criteriumvaliditeit. Deze categorie telt eveneens drie vragen, waarvan één basisvraag. Net als bij de begripsvaliditeit wordt eerst de hoogte van de uitkomsten beoordeeld en vervolgens worden deze geëvalueerd in het licht van de kwaliteit van de onderzoeksprocedure.

Het oordeel voor elk van deze categorieën kan zijn "onvoldoende", "voldoende" of "goed". De schaalpunten op de vragen zijn "-", "+/-" en "+". Een negatief oordeel op een *basisvraag* leidt direct tot het oordeel "onvoldoende" voor de betreffende categorie. De inhoud van de vragen, de toelichtingen en de wegingsvoorschriften worden gegeven in het volgende hoofdstuk. Bij de wegingsvoorschriften van sommige categorieën moeten somscores van items worden berekend. Hierbij wordt aan een negatief oordeel de score min één toegekend, aan een "+/-" oordeel de score nul en aan een positief oordeel de score plus één.

De functie van de toelichtingen is houvast te geven bij de beoordeling en niet om alle statistische of psychometrische achtergronden te verduidelijken. Bij onduidelijkheden dient men zelf de relevante literatuur op te zoeken en te bestuderen.

2. De betekenis van de beoordelingen

In het algemeen kan men stellen dat een "onvoldoende" voor een categorie op twee manieren tot stand kan komen: òf omdat de gevraagde informatie afwezig is, òf omdat de kwaliteit van de wèl aanwezige informatie negatief wordt beoordeeld. Zo kan bijvoorbeeld een "onvoldoende" voor de betrouwbaarheid van een test betekenen dat de betrouwbaarheid niet is onderzocht òf dat deze wel is onderzocht, maar dat dit onderzoek heeft aangetoond dat de test onvoldoende betrouwbaar is. Afwezigheid van onderzoeksgegevens wordt dus hetzelfde beoordeeld als wèl beschikbare onderzoeksgegevens die tot een negatief resultaat leiden, omdat de COTAN meent dat de auteur nu eenmaal verplicht is onderzoeksgegevens te verschaffen. In het boven gegeven voorbeeld betekent dat dat de test bij afwezigheid van gegevens als onvoldoende betrouwbaar wordt gezien tot het tegendeel is aangetoond.

Een tweede nuancering ten aanzien van de beoordeling "onvoldoende" is, dat één of meer "onvoldoendes" niet per se betekent dat een instrument onbruikbaar is. Zo kunnen één of meer schalen of subtests van een vragenlijst of test onvoldoende betrouwbaar zijn; dit hoeft echter niet te betekenen dat de andere schalen of sub-tests of de totaalscore onbruikbaar zijn. Bij tests die worden gebruikt voor belangrijke beslissingen op individueel niveau worden hoge eisen gesteld aan de betrouwbaarheid (zie de toelichting bij categorie 4). Zo wordt de betrouwbaarheid van een dergelijke test als "onvoldoende" beoordeeld indien deze lager is dan .80, bijvoorbeeld .75. Toch kan een dergelijke test nuttige informatie opleveren, bijvoorbeeld in combinatie met andere instrumenten. Ook is het binnen dit beoordelingssysteem mogelijk dat van deze zelfde test met een "onvoldoende" voor Betrouwbaarheid de Begrips- of de Criteriumvaliditeit als "voldoende" of zelfs als "goed" wordt beoordeeld, bijvoorbeeld omdat in selectiesituaties een validiteitscoefficient van .40 als hoog wordt beoordeeld. Zelfs een test met een lage voorspellende waarde kan in sommige gevallen nuttige informatie opleveren, afhankelijk van bijvoorbeeld toevalskans, selectieratio en kosten-baten verhouding.

Een derde nuancering betreft het feit dat in een beoordelingssysteem noodzakelijkerwijs grenswaarden worden genoemd waaraan tests moeten voldoen teneinde objectiviteit bij de beoordeling te garanderen. Zo wordt bij de categorie Betrouwbaarheid specifieke hoogtes van betrouwbaarheidscoëfficiënten genoemd waaraan moet worden voldaan voor een "voldoende" of "goed" beoordeling. Voor deze grenzen is geen sluitende

wetenschappelijke argumentatie te leveren, ze zijn gebaseerd op in het algemeen min of meer geaccepteerde adviezen van vooraanstaande deskundigen. Hiermee hangt samen dat in ieder geval van waarden die in de buurt van deze grenzen liggen nauwelijks is te beargumenteren waarom een bepaalde waarde net wel en een andere waarde net niet "voldoende" of "goed" is. Wèl wordt met het stellen van deze grenzen bewerkstelligd dat deze waarden voor alle tests in principe hetzelfde worden beoordeeld.

4. BETROUWBAARHEID

Basisvraag:

4.1. Worden gegevens over de betrouwbaarheid verstrekt	-		+

(Bij negatieve beoordeling van vraag 4.1 kan men direct doorgaan naar categorie 5).

4.2. Zijn de resultaten voldoende gelet op het beoogde type beslissingen dat met behulp van de test moet worden genomen?

a. paralleltestbetrouwbaarheid	-	+ / -	+	nvt
b. interne-consistentiebetrouwbaarheid	-	+ / -	+	nvt
c. test-hertestbetrouwbaarheid	-	+ / -	+	nvt
d. interbeoordelaarsbetrouwbaarheid	-	+ / -	+	nvt

4.3. Beoordeel hieronder de kwaliteit van het onderzoek:

a. Zijn de steekproeven op grond waarvan de betrouwbaarheidsgegevens zijn berekend in overeenstemming met het beoogde testgebruik?	-	+ / -	+
b. Maken de gegevens die worden verstrekt een gefundeerd oordeel over de betrouwbaarheid mogelijk?	-	+ / -	+

TOELICHTING BIJ CATEGORIE 4

Algemeen

Variantie in de scores van een test is opgebouwd uit betrouwbare en foutenvariantie. De bronnen van de foutenvariantie kunnen zeer verschillend zijn. De indices voor betrouwbaarheid met vermelding van de foutenbron maken het mogelijk over een voor een bepaald doel betrouwbare test te spreken. Met behulp van de traditionele betrouwbaarheidsmaten, zoals verwoord in vraag 4.2, wordt in feite de generaliseerbaarheid van scores over versies, items, tijdstippen en beoordelaars vastgesteld. Uit deze opsomming wordt duidelijk - maar om misverstand te vermijden wordt het nog eens gezegd - dat dè betrouwbaarheid van een test niet bestaat: we onderscheiden vormen van betrouwbaarheid naar de aard van de variantiebron, die in het betrouwbaarheidsonderzoek wordt geanalyseerd. Bovendien zullen de uitkomsten van het betrouwbaarheidsonderzoek variëren afhankelijk van de aard van de onderzochte groep (met name de homogeniteit van de groep met betrekking tot het gemeten begrip). Eventueel kan de betrouwbaarheid worden vastgesteld met behulp van een variantie-analytisch ontwerp, waarin meer facetten tegelijk worden geanalyseerd. Vaak zal men de betrouwbaarheid op klassieke wijze bepalen. Voor de beantwoording van de vragen maakt dit niets uit.
In het algemeen wordt slechts één beoordeling voor de betrouwbaarheid gegeven, ook wanneer een test verschillende scores oplevert. Dit is bijvoorbeeld het geval bij vragenlijsten die uit meerdere schalen bestaan. In dergelijke gevallen geeft de laagste coëfficiënt de doorslag in de beoordeling. Echter, wanneer het een duidelijke negatieve uitzondering betreft (bijvoorbeeld op een na alle subtests "goed" en een subtest "onvoldoende"), mag de hogere beoordeling worden aangehouden (in dit voorbeeld: "goed") en kan als voetnoot bij de beoordeling de uitzondering worden vermeld. Een andere situatie kan ontstaan wanneer de scores op de subtests worden gesommeerd tot één totaalscore, zoals bij sommige intelligentietests het geval is. Hierbij kunnen drie mogelijkheden worden onderscheiden:

- wanneer slechts de interpretatie van de totaalscore van belang is, hoeft uiteraard slechts de betrouwbaarheid van deze score te worden beoordeeld.
- wanneer door de testauteur wordt aangegeven dat de totaalscore weliswaar het belangrijkst is, maar dat ook interpretatie van subtestscores mogelijk is, worden deze laatste met de beoordelingscriteria die gelden voor één niveau lager dan die voor de totaalscore beoordeeld (zie ad. 4.2. Indien de totaalscore valt in de categorie "belangrijk", vallen sub-testscores in de categorie "minder belangrijk"). Hierbij zal het meestal voorkomen dat de scores op de subtests minder betrouwbaar zijn dan de totaalscore, maar kan de beoordeling voor beide gelijk zijn.

- wanneer door de testauteur geen onderscheid wordt gemaakt in het belang van totaalscore en subtestscores worden beide op dezelfde wijze (als even belangrijk) beoordeeld.

Wanneer één en ander leidt tot een verschillend oordeel over de betrouwbaarheid van subtests en totaal kan dit als voetnoot bij de beoordeling worden vermeld. Ook wanneer voor verschillende groepen de betrouwbaarheden worden gegeven èn deze betrouwbaarheden verschillen, wordt slechts één beoordeling gegeven. Mutatis mutandis geldt hiervoor dezelfde regel als hierboven gegeven: de laagst gevonden betrouwbaarheid geeft de doorslag, behalve wanneer het een duidelijke uitzondering betreft.

Voor de vaststelling van het eindoordeel van categorie 4 worden de volgende aanwijzingen gegeven:

- Bij negatieve beoordeling van de basisvraag wordt het eindoordeel "onvoldoende".
- Bij positieve beoordeling van de basisvraag levert vraag 4.2 met betrekking tot de hoogte van de betrouwbaarheidsmaat een voorlopig oordeel. Dit voorlopig oordeel kan naar beneden worden bijgesteld naar aanleiding van het antwoord op vraag 4.3 naar de kwaliteit van het uitgevoerde onderzoek. Bij het eindoordeel van de betrouwbaarheid dient men voorts díe coëfficiënt het zwaarst te wegen, die in overeenstemming is met het doel waarvoor de test wordt gebruikt. Wanneer men bijvoorbeeld over tijd wil voorspellen, dan is in eerste instantie een index van stabiliteit nodig en niet zozeer een consistentiemaat.

Aanwijzingen per vraag

Ad. 4.1.

● Worden betrouwbaarheidscoëfficiënten vermeld?
● Welke betrouwbaarheidscoefficienten worden vermeld? Geef van elke coefficiënt aan wat deze betekent.

Ad. 4.2.
Over de gewenste hoogte van een betrouwbaarheidscoëfficiënt kan geen algemene uitspraak worden gedaan, omdat het doel van het testgebruik hierop van invloed is. Nunnally & Bernstein (1994, p.265) geven aan dat een test die wordt gebruikt voor belangrijke beslissingen een betrouwbaarheid van minstens .90 moet bezitten. Met deze waarde als uitgangspunt zijn de volgende regels opgesteld:

- tests voor belangrijke* beslissingen op individueel niveau (bijvoorbeeld personeelsselectie, verwijzing naar speciaal onderwijs, opname/ontslag kliniek):
  onvoldoende: $r < .80$ voldoende: $.80 \leq r < .90$ goed $r \geq .90$
- idem, maar minder belangrijke* beslissingen (bijvoorbeeld voortgangscontrole, in het algemeen beschrijvend gebruik zoals bij beroepskeuzebegeleiding, therapie-indicatie):
  onvoldoende: $r < .70$ voldoende: $.70 \leq r < .80$ goed $r \geq .80$
- tests voor onderzoek op groepsniveau:
  onvoldoende: $r < .60$ voldoende: $.60 \leq r < .70$ goed $r \geq .70$.

* Met belangrijke beslissingen wordt bedoeld: beslissingen die op basis van de testscore worden genomen, die in principe, of op korte termijn, onomkeerbaar zijn, en die voor een belangrijk deel buiten de geteste om worden genomen.

Voor gebruik in het onderwijs noemt Frisbie (1988, p. 29) een ondergrens van .50 voor de betrouwbaarheid voor "teacher-made tests" als deze scores worden gecombineerd met andere informatie (toetsscores, observaties, cijfers voor opdrachten) tot een rapportcijfer. In feite is de test dan te beschouwen als een subtest, waarbij over het geheel van de testafnames op een of andere manier een totaalscore wordt berekend. Omdat met dit beoordelingssysteem slechts afzonderlijke tests worden beoordeeld, kan met een dergelijke wijze van gebruik geen rekening worden gehouden, tenzij dit door de testauteur expliciet als enige gebruikswijze wordt aanbevolen. Wèl wordt hiermee benadrukt dat ook tests met een als "onvoldoende" beoordeelde betrouwbaarheid, mits op een bepaalde wijze gebruikt, van nut kunnen zijn in het diagnostisch proces.

● Welke van de onder 4.2 (a t/m d) genoemde betrouwbaarheidscoefficienten is van toepassing op de pmtk en waarom?
● Over welke betrouwbaarheidscoefficienten genoemd onder 4.2 a t/m d die van toepassing zijn op de pmtk wordt in de handleiding ook daadwerkelijk informatie gegeven?
● Hoe beoordeel je de betrouwbaarheidscoefficienten (goed, voldoende, niet goed), en de informatie ter onderbouwing (goed, voldoende, niet goed) ? Betrek in je antwoord de suggesties genoemd onderd Ad 4.2 en ad 4.2.a t/m Ad 4.2.d.

Ad. 4.2.a.
De generaliseerbaarheid over testversies (parallelvormen bijvoorbeeld) kan worden bepaald door middel van een onderzoek, waarbij de twee tests met vergelijkbare iteminhoud en, in de ideale situatie, gelijke item-moeilijkheidsgraad, gemiddelde en variantie met elkaar worden gecorreleerd. De correlatie is een schatting van de betrouwbaarheid van beide tests.
De parallelvormbetrouwbaarheid kan van belang zijn bij pure "speed"-tests. De correlatie tussen testhelften, gevormd op basis van halve testtijd en/of halvering van het testmateriaal, kan als parallelvormbetrouwbaarheid worden opgevat. Vervolgens kan de correctie voor test-lengte worden toegepast.

Ad. 4.2.b.
De generaliseerbaarheid over items (of groepen items) binnen een test wordt gewoonlijk berekend door middel van Cronbach's coëfficiënt alfa. Indien coëfficiënt alfa is gebruikt, dient men er op bedacht te zijn dat deze wordt beinvloed door het aantal items, zodanig dat een groot aantal items bij matige inter-itemcorrelaties toch tot een hoge betrouwbaarheidscoëfficiënt kan leiden. De vroeger veel gebruikte split-halfcoëfficiënten worden

afgeraden onder andere omdat de uitkomsten afhankelijk zijn van de toevallige verdeling van items over test-helften.

Ad. 4.2.c.
De generaliseerbaarheid over tijd wordt bepaald door test-hertestcorrelaties. Een test wordt herhaald bij dezelfde onderzoeksgroep, waarbij het tijdsinterval en eventueel relevante gebeurtenissen in dat interval nauwkeurig dienen te worden vermeld. Als het tijdsinterval tussen de twee afnames lang genoeg is, kan dat een indicatie opleveren voor de stabiliteit van testscores.

Ad. 4.2.d.
Vooral bij observatie- en beoordelingsinstrumenten is het van belang of de scores over observatoren/beoordelaars kunnen worden gegeneraliseerd. Maten die hiervoor worden gebruikt zijn overeenstemmingsindexen zoals Cohen's kappa (Cohen, 1960, 1969), de coëfficiënt van Gower (1971), de identiteitscoëfficiënt (Zegers & Ten Berge, 1985) of andere maten die rekening houden met verschillen tussen zowel gemiddelden als varianties van beoordelaars (voor een overzicht zie Zegers, 1989). Ook variantie- en factoranalytisch onderzoek naar de strucur van het observator-/beoordelaarsgedrag kan hier relevant zijn.

Ad. 4.3.a.
Betrouwbaarheidscoëfficiënten moeten worden berekend voor de groepen waarvoor de test wordt gebruikt. Dit betekent dat deze per normgroep moeten worden berekend, aangezien de scores van cliënten met deze groepen worden vergeleken en het om de betrouwbaarheid van de meting binnen deze vergelijkingsgroep gaat. Het is daarom onjuist, en zelfs misleidend, om betrouwbaarheidscoëfficiënten te berekenen over het totaal van alle groepen, of, wat ook voorkomt, over een selectie van extreme groepen. De hoogte van de gevonden betrouwbaarheidscoëfficiënt is immers mede afhankelijk van de spreiding van de scores en deze zal in de totale groep bijna altijd en bij extreme groepen zeker hoger zijn dan bij de te gebruiken normgroepen.

- Zijn de steekproeven op grond waarvan de betrouwbaarheidsgegevens zijn berekend in overeenstemming met het beoogde testgebruik, maw zijn de betrouwbaarheidscoefficienten berekend voor de groepen waarvoor de test wordt gebruikt?

Ad. 4.3.b.
Enkele voorbeelden van informatie die beschikbaar moet zijn om de waarde van het betrouwbaarheidsonderzoek te kunnen beoordelen zijn:

- worden de standaarddeviaties van de scores bij test en hertestgroep vermeld?
- is/zijn de steekproe(f)(ven) waarover de betrouwbaarheidscoëfficiënt(en) is(zijn) berekend voldoende beschreven?

In het uiterste geval, wanneer geen enkele beschrijvende informatie bij de gerapporteerde betrouwbaarheidscoëfficiënten wordt gegeven, kan op grond van afwezigheid van deze gegevens een onvoldoende worden gegeven. Meestal zal echter wel enige informatie beschikbaar zijn, zodat de kwaliteit van het onderzoek is te beoordelen. Met name bij grensgevallen (onvoldoende/ voldoende of voldoende/goed) kan om redenen van gebrekkige informatie voor de lagere beoordeling worden gekozen.

5. VALIDITEIT

A. BEGRIPSVALIDITEIT

Basisvraag:

5.1. Worden gegevens over de begripsvaliditeit verstrekt? | - | | + |

(Bij negatieve beoordeling van vraag 5.1 kan men direct doorgaan naar categorie 5B).

5.2. Maken de resultaten voldoende aannemelijk dat het begrip zoals bedoeld wordt gemeten (of: maken de resultaten voldoende duidelijk wat wordt gemeten)? | - | + / - | +

5.3. Maken de gegevens die worden verstrekt een gefundeerd oordeel over de begripsvaliditeit mogelijk? | - | + / - | +

B. CRITERIUMVALIDITEIT

Basisvraag:

5.4. Worden er gegevens verstrekt over het verband test-criterium? | - | | +

(Bij negatieve beoordeling van vraag 5.3 kan men de rest van de vragen overslaan).

5.5. Zijn de resultaten voldoende gelet op het beoogde type beslissingen dat met de test moet worden genomen? | - | + / - | +

5.6. Zijn de resultaten voldoende gelet op het beoogde type beslissingen dat met de test moet worden genomen? | - | + / - | +

TOELICHTING BIJ CATEGORIE 5

Algemeen

Validiteit is de mate waarin een test aan zijn doel beantwoordt; kan men uit de testscores die conclusies trekken die men op het oog heeft? In de literatuur worden vele soorten validiteit onderscheiden, zo noemen Drenth en Sijtsma (1990) er al acht. De onderscheidingen hebben betrekking op het doel van het validiteitsonderzoek of op op het proces van validering door bepaalde data-analysetechnieken. In dit beoordelingssysteem wordt aangesloten bij de klassieke driedeling naar doel van het validiteitsonderzoek, zoals die onder andere in de Richtlijnen (Evers e.a., 1988) wordt gehanteerd: begripsvaliditeit, criteriumvaliditeit en inhoudsvaliditeit. Van deze drie is *inhouds*validiteit reeds aan de orde gekomen in categorie 1, omdat het als onderdeel van het testontwikkelingsproces wordt gezien.
Bij *begrips*validiteit gaat het erom te onderzoeken wat de test nu eigenlijk meet. Meet de test het bedoelde begrip of, gedeeltelijk of voornamelijk, iets anders? Vaak gebruikte methoden of technieken voor het aantonen van de begripsvaliditeit zijn: factoranalyse voor het aantonen van de ééndimensionaliteit, het vergelijken van de gemiddelde scores van groepen waarvan men mag verwachten dat ze verschillen zullen vertonen en het berekenen van correlaties met tests die hetzelfde zouden moeten meten (zogenaamde soortgenoten). Dit zijn in principe vrij eenvoudig uit te voeren onderzoeken die een eerste aanwijzing kunnen opleveren met betrekking tot de begripsvaliditeit, maar die elk op zich nog géén aanleiding geven tot een "voldoende" beoordeling. Slechts de opeenstapeling van dergelijke aanwijzingen, of meer uitgebreid structuur- of zgn. multi-trait-multi-method- onderzoek kan leiden tot de beoordeling "goed" of "voldoende".

Bij *criterium*validiteit onderzoekt men in hoeverre de testscore een goede voorspeller is van niet-testgedrag (retrospectief, gelijktijdig of predictief). Het is van belang dat verwachtingen worden gespecificeerd ten aanzien van het type criteria waarmee relaties worden verwacht. Dit is met name van belang wanneer een test uit verscheidene subtests of schalen bestaat. Overigens hoeft voor een "voldoende" of "goed" beoordeling niet de validiteit van *alle* subtests of schalen te worden aangetoond, omdat een enkele zeer valide schaal de test al tot een waardevol instrument kan maken. Gegevens met betrekking tot de criteriumvaliditeit kunnen ook betrokken worden bij het oordeel over de begripsvaliditeit (zij maken in feite ook deel uit van het proces van begripsvalidering, zie bijvoorbeeld Anastasi, 1986, en Messick, 1988), omdat deze gegevens tevens een bijdrage leveren aan de verheldering van hetgeen door de test wordt gemeten.

Voor de vaststelling van de eindoordelen van de categorieën 5A en 5B worden de volgende aanwijzingen gegeven:

- Bij negatieve beoordeling van de basisvragen 5.1 en/of 5.4 wordt het eindoordeel "onvoldoende" voor de betreffende categorie(ën).
- Bij positieve beoordeling van een basisvraag leveren de vragen 5.2 en/of 5.5 met betrekking tot de resultaten van het validiteitsonderzoek een voorlopig oordeel voor de betreffende categorie(ën). Dit voorlopige oordeel kan naar beneden worden bijgesteld naar aanleiding van de antwoorden op de vragen 5.3 en/of 5.6 naar de kwaliteit en volledigheid van de geleverde informatie.

Aanwijzingen per vraag

Ad. 5.1.
Het gaat hier om onderzoek van de interne of de externe structuur. De interne structuur kan worden onderzocht door associatiematen te bepalen tussen (groepen) items, items en test en tussen subtests. Ook kunnen procedures zoals proefpersonen hardop laten denken en inspectie van items worden gebruikt. De externe structuur wordt gewoonlijk onderzocht door relaties met andere tests te bepalen (convergente en discriminante validiteit).

- Wat verstaat men onder begripsvaliditeit?
- Welke gegevens gerapporteerd in de pmtk handleiding vallen onder "onderzoek van de interne of externe structuur", en waarom?

Ad. 5.2.
Zoals reeds eerder vermeld gaat het bij begripsvalidering vooral om de cumulatie van onderzoeksresultaten. Begripsvalidering is nooit af.

- Beargumenteer waarom je vind dat de gerapporteerde gegevens van onderzoek naar de interne en externe structuur van de PMTK getuigen van goede, voldoende, of onvoldoende begripsvaliditeit.

Ad. 5.3.
Om een gefundeerd oordeel te kunnen geven over de begripsvaliditeit dient bepaalde informatie beschikbaar te zijn, zoals de grootte van de steekproef waarover de resultaten zijn berekend, de gebruikte analysetechnieken, etc.

- Is er genoeg informatie in de pmtk handleiding om tot een gefundeerd oordeel te komen ten aanzien van begripsvaliditeit.
- Zo ja, welke essentiele informatie wordt inderdaad gerapporteerd? Zo nee, welke informatie wordt niet gerapporteerd die wel essentieel is om te komen tot een gefundeerd oordeel?

Ad. 5.4.
Gelet op de veelsoortigheid van dit type onderzoek kunnen hier nauwelijks algemene aanwijzingen worden gegeven. Een aandachtspunt: gewaarschuwd moet worden tegen de interpretatie van onderzoeksresultaten zonder specifieke verwachtingen vooraf. Dergelijk onderzoek krijgt al gauw het karakter van "vissen": post hoc zal men altijd wel een aantal interpreteerbare verbanden vinden wanneer men de test correleert met een groot aantal (toevallige beschikbaar zijnde) andere testscores, waarbij het aannemelijk is dat enkele van de significante correlaties op toeval berusten.

- Wat verstaat men onder criterium validiteit.
- Hoe wordt criteriumvaliditeit ook wel genoemd?
- Wat is het verschil tussen criterium validiteit en de naam die ook wel aan criteriumvaliditeit wordt gegeven?
- Worden er gegevens verstrekt over het verband test-criterium, zo ja, welke gegevens in de pmtk handleiding zijn dat?

Ad. 5.5.
Of één of meer validiteitscoëfficiënten voldoende zijn hangt van een aantal zaken af. Van belang zijn o.a. het doel van de test, de hoogte van de validiteitscoëfficiënten, de betrouwbaarheids-intervallen van de coëfficiënten, de winst die de test oplevert ten aanzien van al aanwezige informatie, de selectieratio en een kosten-baten-analyse. Voorts kan een test in verschillende situaties of bij verschillende groepen andere coëfficiënten opleveren of gedeelten van een criterium goed voorspellen. Zo wordt in selectiesituaties een validiteitscoëfficiënt van .40 als goed gezien, terwijl in

opleidingssituaties met gemak hogere coëfficiënten worden behaald. Naarmate de auteur meer expliciet is over het doel van de test, kan de beoordelaar beter uitmaken of de test daaraan een zinvolle bijdrage levert.

● Beargumenteer waarom je vind dat de gerapporteerde gegevens van onderzoek naar criterium validiteit van de PMTK getuigen van goede, voldoende, of onvoldoende criteriumvaliditeit.

Ad. 5.6.
Om een gefundeerd oordeel te kunnen geven over de criteriumvaliditeit dient bepaalde informatie beschikbaar te zijn, zoals de grootte van de steekproef waarover de resultaten zijn berekend, of het valideringsonderzoek onder dezelfde testcondities heeft plaatsgevonden als de condities waarin de test wordt gebruikt, informatie over de gebruikte analysetechnieken, etc.
Soms ligt de keuze van een criterium voor de hand en is het makkelijk beschikbaar (slagen-zakken, een rapportcijfer), in andere gevallen moeten criteriummaten apart worden geconstrueerd en verzameld. In beide gevallen dient het criterium zo volledig mogelijk te worden beschreven en dient te zijn aangegeven welke relevante gedragsaspecten wel en niet in de criteriummaat zijn opgenomen. Indien mogelijk dient de betrouwbaarheid van de criteriummaat te worden vermeld. Dit geldt met name voor samengestelde criteria. Wanneer de onderlinge relaties van de afzonderlijke elementen van het criterium laag zijn, kunnen beter afzonderlijke validiteitscoëfficiënten voor elk van de elementen worden gegeven.

● Is er genoeg informatie in de pmtk handleiding om tot een gefundeerd oordeel te komen ten aanzien van criteriumvaliditeit.
● Zo ja, welke essentiele informatie wordt inderdaad gerapporteerd? Zo nee, welke informatie wordt niet gerapporteerd die wel essentieel is om te komen tot een gefundeerd oordeel?