Samenvatting Hoorcolleges

VOCUS heeft deze samenvatting te danken aan Leonie Schiphorst.

Het gebruik van deze samenvatting is bedoeld als studeerhulp na het lezen van de verplichte literatuur. Gebruik van deze samenvatting is geheel voor eigen risico.

Soms wordt er verwezen naar bladzijden of tabellen in het originele boek.

Succes met studeren!

Hoorcollege 1 – 5 februari 2018.

8. Bij assessment gaat het vaak om personen. Bij evaluatie gaat het om producten, organisaties en beleid. De rekentoets is een selectieargument

9. er zijn aspecten waar elke hbo en wo opleiding moet werken, maar niet elke cursus.

12. Does= laat iets zien in de praktijk. Het moet aansluiten bij je doel. Je kunt sommige dingen niet direct toetsen in de authentieke situatie, studenten geneeskunde kunnen niet direct snijden in levende mensen. Bij 'shows' is er een soort simulatie.

13. Triangulatie: data van verschillende instrumenten ga je combineren om beter inzicht te krijgen in wat iemand gepresteerd heeft.

16. De titel van het boek is 'toetsen in het hoger onderwijs' maar het is ook toepasbaar buiten het hoger onderwijs.

18. Wat is je doel? Hoe meet je dat? Wat is je instrument? Hoe kom je tot een score?
Voor, tijdens en na het meten moet je een kwaliteitscontrole uitvoeren.

20. Psychometrische eis: vanuit de data kun je een gemiddelde berekenen.
edumetrische eis: de manier waarop data wordt verzameld en de gevolgen ervan.
betrouwbaarheid heeft te maken met omgevingsfactoren en verschillende beoordelaars.

23. De beslisboom gaat over de keuzes die je moet maken voor- en na afname van de vragenlijst. Inhoudsvaliditeit: welke constructen ga je meten? Begripsvaliditeit: welke items ga je toepassen om een construct te meten? Betrouwbaarheid: welke toets omstandigheden. Welke technieken passen bij welke data?

24.Facefadility=indruksvaliditeit, de onderzoeker heeft de indruk dat het valide is.

25. Hoge p-waarde: het tentamen is goed gemaakt. Rirwaarde: als iemand een item goed gemaakt heeft correleert die score met de overall score. Als iemand een moeilijk item goed heeft wil je dat hij/zij ook hoger scoort.

27. Testinstructie met verwachtingen, hoeveelheid items, cesuur en gevolgen.

30. De examencommissie checkt de procedures bijvoorbeeld of de cursus coördinator de toetsmatrijs maakt.

31. verticaal gaat over verschillende jaren. Horizontaal gaat over één jaar. Je moet dus ook kijken wat de opleiding als geheel toetst.

32. Toetsprogramma op opleidingsniveau is over alle cursussen heen. Je moet de kruisjes in het schema kunnen onderbouwen.

33. Vanuit assessmentoogpunt kijk je naar standaard 3 en 4. De opleiding moet op alle vier de standaarden een voldoende scoren.

34. Wat bij het ene niveau eruit komt voert weer door in het andere niveau. Als op beleidsniveau dingen niet goed gedaan zijn dan moet het facultaire beleid worden aangepast worden wat weer gevolgen heeft voor de opleidingen.
PDCA: Wat zijn de plannen, hoe zijn ze uitgevoerd, wat zijn de effecten daarvan, kunnen we daarvan leren.

Hoorcollege 2 – 12 februari 2018

1. Klassieke testtheorie geeft niet aan welk exacte waarde een betrouwbaarheidscoëfficiënt zou moeten hebben. Een lagere betrouwbaarheid verlaagt de kans om de correlatie tussen constructen goed te kunnen onderzoeken bij gebruik van de testsccore.

3.Latente variabelen: zijn niet direct observeerbaar, dit geef je weer met een rondje. Observeerbare variabelen geef je weer met een vierkant.

5. De score op X wordt als het goed is zoveel mogelijk bepaald door aanleg.

7. Operationalisaties zijn vaak test- en vragenlijsten.
Met een Likertschaal krijg je ordinale scores, een rangorde.

8. Met de itemscores construeer je een testscore, de itemscores krijgen dan een bepaald gewicht. De intervallen tussen de schaalpunten binnen een item moeten aan elkaar gelijk zijn maar ook over items.

9.Systematische meetfout: meet je wel wat je wilt meten?
Toevallige meetfout: fouten door situationele invloeden zoals vermoeidheid.

10. De systematische meetfout heeft invloed op de validiteit. De toevallige meetfout heeft invloed op de betrouwbaarheid.

12. Inhoudsvaliditeit kun je niet statistisch onderzoeken.

13. De begripsvaliditeit met de factoranalyse, componentenanalyse en de multitrek-mulimethode-aanpak. Begripsvaliditeit: welke en het aantal dimensies dat je meet. Criteriumvaliditeit meet je met een regressieanalyse.

14. Interne structuur: betrekking op de relaties tussen de latente variabelen die door de test worden gemeten. Over die relaties is een theorie die je meet met factoranalyse en componentenanalyse.
Externe structuur: er moet een hoge correlatie zijn tussen de test en andere tests. → convergente/divergente validiteit.

16.Ware score: persoonlijk gemiddelde van persoon A.
Testscore: som van ware score en toevallige meetfout (X = T + E)
Variantie binnen een persoon is altijd de variantie van toevallig meetfouten.

17. Toevallige meetfout= verschil tussen een van die testscores en de ware score. In de ideale situatie is het gelijk aan nul want dan is het betrouwbaar. Voor persoon A heb je dus een oneindige reeks van toevallige meetfouten. In de ideale situatie zijn er geen toevallig meetfouten.

19.Ware score: is iemands gemiddelde testscore over oneindig veel herhaalde afnames van dezelfde test bij gelijkblijvende structurele persoonskenmerken (= binnenpersoonsgemiddele). Variantie binnen een persoon is altijd de variantie van toevallige meetfouten.

21. Het is eigenlijk allemaal herhaling. De predictor is nu ware score. En de uitkomstvariabele is nu de testscore. Betrouwbaarheid .8= 80 procent van de testscorevariantie wordt verklaard door de ware score. De ware score heb je niet.

23. de betrouwbaarheden voor de twee parallell-tests zijn aan elkaar gelijk. Ook blijkt dat ze allebei gelijk zijn aan de correlatie van de twee testscores.

24. je hebt 2 testen die je afneemt bij dezelfde mensen. de tests bevatten niet dezelfde items maar ze zijn wel parallel. Rho is de gene die je wilt weten en R is de steekproefschatting. Variantie van de testscore in de populatie= variantie van de ware score en de variantie van de toevallige meetfout. Correlatie is een schatting van betrouwbaarheid.

28. Betrouwbaarheid S= betrouwbaarheid S'= correlatie tussen S en S'. pss'= betrouwbaarheid van een subtest.

30. Ware score= specifiek + gemeenschappelijk. De items meten hetzelfde en moeten dus een samenhang vertonen als schatting van betrouwbaarheid. S=systematisch en E=toevallig. S+E=uniek voor een item. Correlatie= schatting betrouwbaarheid subtest. Maar je wil de hele test.

31. Interne consistentie is gebaseerd op de correlaties tussen de items in dezelfde schaal. Je kijkt naar de overlap. Hoeveel variantie in de testscore kan worden verklaard door de predictoren. Items meten een gemeenschappelijke factor maar zijn niet identiek.

32. P2xc kun je schatten met een factoranalyse. Communaliteit is de ondergrens van betrouwbaarheid.

33. Een Cronbach's alfa geeft een schatting van de communaliteit gebaseerd op een één-factormodel. Het is een schatting van de ondergrens van de betrouwbaarheid. Guttmans labda2 is altijd hoger. Coëfficiënt omga geeft een betere schatting want deze kan ook bij 3 of meer factoren.

35. wanneer is een schattig van de betrouwbaarheid hoog of laag? --? Richtlijn NIP

36. Kleine steekproef→ veel spreiding. Als je de test vervolgens afneemt in een groep die homogener is dan is de betrouwbaarheid lager. Kleinere ware score variantie→ meer toevallige meetfouten → betrouwbaarheid lager. Je moet de test dus gebruiken bij een zelfde soort steekproef uit dezelfde populatie.

37. Itemanalyse: Je wil een aantal items weggooien (een zo klein mogelijke test) met een maximale betrouwbaarheid. Hoe belangrijker de beslissing hoe hoger de betrouwbaarheid moet zijn.

Hoorcollege 3– 19 februari 2018

1.Toetsen en testen als cultuurfenomeen

De politiek wil weg met doorgeslagen toetscultuur. Onderwijs moet niet alleen draaien om toetsen. Denk aan de cito-toets en rekentoets. De focus ligt op meetbare opbrengsten, als het tot goede scores leidt worden normeringen verscherpt.

Evidence-centered design

Voordat je een toets afneemt voer je een analyse over het domein. Je bepaalt welke kennis en vaardigheden de lerende moet opdoen. Deze variabelen kunnen niet direct worden geobserveerd, dus het gedrag en de prestaties moeten ook worden bepaald. De volgende stap is het bepalen van de soorten taken of situaties die dergelijke gedragingen of uitvoeringen zouden veroorzaken. Je zet dit op in een conceptual assessment framework (toetsmatrijs).

2. Het maken van een test

Als je iets meet, meet je iets onderliggends. Vroeger moest je weten wie je mondeling afnam omdat elke docent andere vragen stelde, men zei dat dit goed was voor de validiteit want in het echte leven gaat het ook zo.

Wat maakt een toets goed?

Een goede toets is niet altijd vastgesteld op verschillen. Dus hoe goed je het doet hangt niet af van hoe goed de rest het doet.

Transparantie: de duidelijkheid bij studenten over wat ze kunnen verwachten Over alle inhoud van de toetsen, criteria, normering en uitslagen van de toetsing.

Toetsmatrijs

Leerdoelen x beheersingsniveau (of: inhoudscategorieën x gedragscategorieën)
Hierin wordt aangegeven hoe de opgaven in een tentamen/toets zijn verdeeld over de leerstof, in relatie tot de vooropgestelde doelstellingen/onderwerpen.
Het helpt de docent bij het samenstellen van een toets die representatief is voor de cursus → validiteit.
Het helpt de student om zicht te krijgen op het belang van de verschillende onderdelen in de cursus → transparantie.

Ecologische validiteit is de mate waarin de onderzoeksresultaten uit een onderzoek overeenkomen met de alledaagse praktijk.

Toetsvormen

Gesloten vragen: ja/nee, juist/onjuist, meerkeuze en matchingtoetsen (‘geprecodeerd’)
Open vragen: essay, korte antwoord- en aanvultoetsen, mondelingtoetsen (‘vrije antwoorden’).
Casus: bestaat uit casus + vraag, de vraagvorm doet er minder toe dan de vraaginhoud.
De authenticiteit van de casusbeschrijving is essentieel.
Mondeling: heeft veel voordelen zoals herformuleren, doorvragen, hoog cognitief niveau, weinig tijd, voor dyslexie, feedback,

Performance assessment: een verzamelterm voor methoden waarmee gedrag kan worden uitgelokt bij studenten met als doel vast te stellen in hoeverre hogere-ordedoelstellingen zijn verworven. Er is sprake van performance assessment als de kandidaat onder zo natuurgetrouw mogelijke omstandigheden een zo realistisch mogelijke taak uitvoert.

Halo-effect: Matige prestaties van (in ogen van de beoordelaar) briljante of aardige studenten, worden dan hoger gewaardeerd dan exact dezelfde prestatie van middelmatige of minder vriendelijke studenten.
Signifisch effect: Strenge en soepele beoordelaars. De criteria die docenten hanteren bij de beoordeling kunnen verschillen.

<u>Bronnen van meetfouten bij vragenlijstonderzoek</u>
Satisficing : iemand heeft geen zin om de vragenlijst in te vullen en doet maar wat.
Sociaal wenselijk: mensen doen zich beter voor dan dat ze zijn.
Acquiescence: Neiging om met uitspraken in te stemmen
Geheugeneffecten: leidt tot een vertekend beeld
Contexteffecten: volgorde van vragen, als je een vraag op een bepaalde manier invult en er volgt daarna een vraag die er op lijkt dan ga je die op dezelfde manier invullen.
Intervieweffecten: creëert diegenen een veilige sfeer of juist niet.

3.Het maken van toetsitems
Iets wat de ene keer werkt kan de andere keer minder goed werken, je kan dus niet zomaar de test van iemand anders overnemen. Mensen veranderen, taal verandert, opleiding verandert ect.
Je kan nog zo overtuigd zijn dat de vraag correct is gesteld, het kan altijd zijn dat iemand het anders interpreteert of beredeneerd.
'Een test is niet beter dan de items waaruit hij bestaat. Een hele boel slechte items kunnen dus geen goede test maken.'

<u>Voordelen Meerkeuze-items</u>
-Makkelijk af te nemen en te verwerken, kost weinig tijd.
-Objectieve scoringsmethode
-Groter bereik meting kennisdomein????
-Je kunt bijna alles meten dus ook hogere denkprocessen.
- Je kunt veel vragen opnemen wat ten goede komt van de betrouwbaarheid.
- je kunt de psychometrische kwaliteit van toetsvragen makkelijk analyseren.

<u>Nadelen</u>
Gemakkelijk oppervlakkig
Gemakkelijke schijnobjectiviteit: de keuze voor de afleiders is subjectief.

<u>Na de toets: toets- en itemanalyse</u>
Item-restcorrelatie: de correlatie tussen de vraag (item) en eindscore (totaal score).
Een Ritwaarde van +1 betekent dat alle studenten die hoog op de toets scoorden, de betreffende vraag correct hebben beantwoord.
Itemdiscriminatie: elke vraag moet zo goed mogelijk onderscheid maken tussen studenten met een hoge en lage eindscore. Een goed discriminerende vraag heeft een positieve Rit-waarde.

Gecorrigeerde p-waarde (geldt alleen voor gesloten toetsvragen): een indicatie voor de hoeveelheid deelnemers die de vraag goed weten te beantwoorden zonder te gokken. De gecorrigeerde p-waarde valt altijd lager uit dan de ongecorrigeerde p-waarde.

Hoorcollege 4– 26 februari 2018

4. Per latente variabele wordt een subtest geconstrueerd. Subtests hebben geen gemeenschappelijke items. Waarom? Omdat de items specifiek voor één latente variabele is. Begripsvaliditeit: wordt het aantal latente variabelen ook gemeten?

5. Je hebt 7 latente variabelen. Er zijn geen correlaties tussen de unieke factoren onderling en ook niet tussen de unieke factoren en gemeenschappelijke factoren.
6. Dubbele pijl betekent correlaties, de gewone pijlen zijn regressies. F is item gemeenschappelijke factoren, U is item unieke factoren.

8. De predictor is gemeenschappelijk. Labda is de regressiecoefficient/slope.
Simple structure: de items van iedere subtest meten slechts één gemeenschappelijke factor. Iedere subtest meet dan maar één latente variabele. Elk item meet ook een unieke factor. Iedere unieke factor is onafhankelijk van alle item- gemeenschappelijke factoren.

10. Rijen horen bij de items, de kolommen bij de gemeenschappelijke factoren.
12. Voor ieder uniek element in de covariantiematrix van de itemscores is er een vergelijking.
14. Covariantie is een ongestandaardiseerde maat voor de samenhang tussen 2 variabelen.
15. De parameters zijn onbekenden. Het aantal onbekenden is groter dan het aantal vergelijkingen.

20. Niet-standaard confirmatief factormodel: je gaat alle schatten.
21. Restrictief: het legt restricties op op hoe de data mag zijn. Confirmatief: checken of dat wat je veronderstelt, terug te zien is de data.

22. Je ziet geen 1 dus er is gebruik gemaakt van 'Unit variance identification'
23. Het liefst wil je geen significant resultaat. Als de p waarde kleiner is het nominale significantieniveau dan H0 verwerpen (de H0 wil je hier niet verwerpen)

26. Het model past redelijk bij de data. Want beide zijn boven de .95. het is een toename ten opzichte van het onafhankelijkheidsmodel.
27. Je kunt kijken naar de proportie verklaarde variantie door de laatste formule eruit te halen en te delen door het totaal. Voor ieder item wordt een deel verklaard door item gemeenschappelijke factoren, dit geef je aan met omega.

28. communaliteit is een ondergrens van betrouwbaarheid bijvoorbeeld chronbach's alfa.
29. Hoe meer items je opneemt hoe hoger de communaliteit.

30. Je kunt de correlaties bepalen tussen subtestscore1 en subtestscore 2.
32. De factorscore is hoger dan de subtestscore en dus meer valide. Subtestscore= ongewogen. Bij factorscore: het wordt zo gewogen dat de corrrelatie maximaal is.

36. hoe bepaal je in een exploratieve analyse het aantal gemeenschappelijke factoren?
alle items hebben factorladingen op alle item gemeenschappelijke factoren.
38. Als het 1factormodel niet past dan probeer je een 2 factormodel → kruisende lijnen.

41. linksboven en rechtsonderin zijn de getallen laag → simpel structure. Aan de hand van de matrix ga je proberen de factoren in een goed passend model te interpreteren. Je hebt rotationele vrijheid.
44. Het aantal besloten factoren houdt je constant. In de onderste tabel staan de correlaties tussen

facotren. als de correlatie lager is dan .3 had je net zo goed varimax correlatie kunnen kiezen.

Hoorcollege 5– 5 maart 2018

2. Je wilt mensen van verschillende groepen met elkaar vergelijken. Meet invariantie: de test met hetzelfde op dezelfde manier in verschillende groepen. Toetsen of er gemiddelde groepsverschillen zijn op de item gemeenschappelijke factoren dus niet op de testscores. Dus de latente variabele.

3. Als de passing goed is dan is de begripsvaliditeit in orde. 2e vraag: hoe goed meten de items de itemgemeenschappelijke factoren? Het kan bijvoorbeeld zijn dat ze vooral unieke factoren meten. Communaliteit is de proportie variantie van die score verklaard door alle gemeenschappelijke factoren. Validiteit: proportie variantie van een subtest score die verklaard wordt door een specifieke factor.

4. q gemeenschappelijke factoren. Je veronderstelt een multipele regressie. Voor alle itemscores heb je een multipele regressie. Je hebt meerdere uitkomstvariabelen → multivariaat multipele lineaire regressie. Normaal zijn predictoren observeerbaar, nu niet. Unieke factoren zijn ongecorreleerd.

5. Itemscores zie je. Intercepten en regressie coëfficiënten schatten. 2 latente variabelen hebben een bepaalde samenhang→ correlatiematrix (schatbaar).

7. Op hoofddiagonaal staan de varianties.

8. Vector= matrix met maar 1 kolom of 1 rij. De vectoren en factorladingenmatrices kunnen verschillen tussen groepen. Doel van de test is vergelijken van mensen.

9. Meetinvariantie: de test moet hetzelfde meten voor verschillende groepen. Y-as: itemscore, x-as: itemgemeenschappelijke factor. De lijnen verschillen in intercept, dit is niet goed.

10. Nu verschillen niet alleen de intercepten maar ook de factorladingen. Alleen in het midden van de lijn is meetinvariantie. Intercept is beginpunt en factorlading de richtingscoëfficiënt.

11. Zelfde aantal item gemeenschappelijke factoren, items en zelfde factorladingen.

14. schattingen van vier parametermatrices. Intercepten, factorladingen, covarianties tussen itemgemeenschappelijke factoren en tussen item unieke factoren. De 1 gebruik je om te identificeren (unit loading identification). Hoe complexer het model hoe meer parameters ho minder precies.

15. De gemiddelden van de itemgemeenschappelijke factoren zijn op nul gezet.

16. In het model staan item gemeenschappelijke en unieke factoren door elkaar. Varianties op de diagonaal worden geschat.

17. Nu zijn de factorladingen matrices voor groepen helemaal gelijk.

18. Unit loading identification is standaard instelling. Restrictie voor de data: zelfde factorladingenschattingen voor de groepen.

20. schattingen van de varianties van item gemeenschappelijke factoren en unieke factoren zijn niet aan elkaar gelijk. Je zet de factorladingen matrix voor groepen gelijk.

21. je hebt pas wat aan het model bij sterke facotriele invariantie, hierbij zijn er ook gelijke vectoren van intercepten. Je hoeft maar voor eén groep de gemiddelden van de itemgemeenschappelijke factoren op nul te zetten, in de andere groepen kun je de gemiddelden schatten. Soort anova op latent niveau.

22. Strikte factoriele invariantie: nu veronderstel je dat ook de covariantiematrix van de item unieke factoren aan elkaar gelijk zijn.

26. Alleen de eerste groep is boven .95 dus er is geen meetinvariantie over groepen. Je meet wat anders bij mannen dan bij vrouwen.

30. Stel er is meetinvariantie dan moet je het nulhypothesemodel verwerpen. Er zijn verschillen op latent niveau.

32. Twee tests(methoden) en twee latente variabelen(trekken). Je moet laten zien dat je test is gerelateerd aan andere tests die dezelfde latente variabelen meten.

33. Multitrek-multimethode correlatiematrix. A1 en b1 zijn geconstrueerd voor dezelfde latente variabele.

35. Als ze allemaal gelijk zijn aan nul kun je de 2 latente variabelen onderscheiden van elkaar. De divergente validiteit moet laag zijn. De rode moeten lager zijn dan de zwarte. De zwarte moeten het liefst zo dicht mogelijk bij 1 liggen. Er is dan sprake van divergente validiteit.

36. In het ideale geval zijn ze aan elkaar gelijk want het zijn dezelfde schattingen van de correlaties tussen latente variabele 1 en latente variabele 2. Dit is de nulhypothese van geen methode-effecten.

37. Je hebt 4 subtestscores. Dubbele pijlen staan voor een correlatie(gestandaardiseerd) of een covariantie (niet gestandaardiseerd). Als de correlatie nul is, is er geen convergente validiteit.

39. Perfecte convergente validiteit is de correlatie gelijk aan 1. Dit komt nooit voor want er is altijd een toevallig meetfout. Als je geen perfecte correlatie hebt kan er nogsteeds perfecte convergente validiteit zijn.

40. als de latente variabelen totaal verschillend zijn is de correlatie gelijk aan nul.

41.aan de ene kant heb je een standaard confirmatief factormodel, ze meten dezelfde gemeenschappelijke factor. Je houdt rekening met methode-effecten. De subtestscores van test A hebben wat gemeenschappelijks.

43. Het model zonder bedoelde factoren is het model zonder convergente validiteit. Je toetst het ene model tegen het andere model. Als je het ene model verwerpt dan is er bewijs voor convergente validiteit.

44. Het model met perfect gecorreleerde bedoelde factoren heeft geen divergente validiteit. Je kan het toetsen tegen het algemeen model waarin je de correlatie schat.

45. Het model zonder methode-effecten. Als je het model verwerpt is er bewijs voor methode-effecten.

46. Convergente en divergente validiteit zijn wel wenselijk. Methode-effecten zijn niet wenselijk want dat meet je wat extra’s.

Hoorcollege 6 – 12 maart 2018

4. Meet je kennis vaardigheden of attitudes. En waar meet ik dit mee. Bijvoorbeeld performance assessment om te kijken of iemand een presentatie kan geven.

5. Je kunt alleen in de praktijk meten of dat wat de kunt ook daadwerkelijk doet. Als je het in de praktijk wil toepassen moet je rekening houden met de natuurgetrouwheid. Dan spreek je over competentietoetsing.
Competence based : What can you do?, Competences underlying behaviour , Task-independent.
Performance-based: Wat do you do?, Observable behaviour, Task-dependent.

6. Er zijn 4 methodieken die ontstaan door een combinatie van stimulus- en respons categorieën van assessmentvormen. Stimulus= de opdracht en verdere relevantie informatie zoals instructies
Proeve van bekwaamheid: Voordeel is dat er kan worden getoetst of de kandidaat het inzicht heeft om zelfstandig te beslissen of en hoe er moet worden gereageerd in bepaalde situaties.
Praktijktoets: Dit is efficiënter dan een proeve van bekwaamheid. Nadeel is dat je niet weet of de kandidaat de wil en het inzicht heeft om op eigen initiatief tot actie over te gaan.

Job performance appraisal: verschil tussen objectieve en subjectieve methoden. De assessor heeft de vrijheid om zelf te bepalen op welke observaties de beoordeling wordt gebaseerd. Er is geen directe observatie.
Mystery guest: expliciete opdracht en typical performance. Voordelen zijn een dichte benadering van de reële werksituatie en een hoge efficiënte door gerichte uitlokking van gedrag. Nadeel zijn hoge kosten en ethische vragen.

7. Untrosted professional acitivity: je heeft iemand het vertrouwen bepaalde handeling uit te voeren. Je moet bepaalde levels specificeren. Je denkt niet in rubrics die opbouwen in complexiteit.
High untrusted: je bent in staat een andere student te begeleiden bij het uitvoeren van een IPA.

8. Relatief normeren: cijfers worden bepaald door toets resultaten vergelijkenderwijs te waarderen.
Absoluut normeren: de behaalde toets resultaten worden afgezet tegen een vaste norm.

Holistische rubric: complexe vaardigheden worden zo veel mogelijk als een geheel beoordeeld. Het geeft een algemene impressie over de uitvoering van een taak doordat criteria worden samengenomen.
Voordeel: Ze leiden snel tot scoring, de betekenis van het geheel blijft behouden. Ze laten meer ruimte over voor interpretatie

Analytische rubric: er wordt juist gestreefd naar zo eenvoudig mogelijke criteria zodat alle relevante beoordelingsaspecten afzonderlijker van een kwaliteit worden voorzien.
Voordeel: de kans dat meer beoordelaars tot hetzelfde eindoordeel komen is groter→ grotere betrouwbaarheid. Het geeft de student meer inzicht in de complexiteit van de taak. Nadeel: tijdrovend. Je geeft de mensen meer richtlijnen over waar je naar gaat kijken bij de beoordeling. Je kan subvakjes maken → inhouds- en begripsvaliditeit.

10. Door rubrics wordt het concreet en transparant welke mate van beheersing bij welke beoordeling hoort. Je kunt hiermee het huidige ontwikkelniveau in kaart te brengen en mogelijkheden om verdere ontwikkelmogelijkheden benoemen.

De rubric in de afbeelding is analytisch want er zijn subvakjes. Verder staat er geen weging voor de subvakjes dus je mag compenseren → compensatorisch.

11. Standard setting methoden hebben als doel vaststellen van de onderscheidende aspecten voor het beoordelen van studenten. Standard setting is een procedure om te komen tot een beslisregel bij een opdracht om onderscheid te kunnen maken tussen twee of meerdere toetsscores.

13. De cesuurmethode van Angoff (1971) is ook een manier om een beredeneerde absolute cesuur te bepalen. Hierbij schatten experts hoeveel procent van de kandidaten die de stof nét voldoende beheersen, een vraag goed zal beantwoorden. Het gemiddelde van deze percentages bij alle vragen, geeft de cesuur (cut score.). In dit voorbeeld is de schatting gemaakt door 5 verschillende experts.

17. Meetfouten: (Inter)beoordelaarseffecten
Signifisch effect: Strenge en soepele beoordelaars. De criteria die docenten hanteren bij de beoordeling kunnen verschillen.
Halo-effect: Matige prestaties van (in ogen van de beoordelaar) briljante of aardige studenten, worden dan hoger gewaardeerd dan exact dezelfde prestatie van middelmatige of minder vriendelijke studenten.
Rater drift effect: refers to changes in rater behavior across different test administrations. Prior research has found evidence of drift.
Sequentie-effect: Voorafgaande beoordelingen zijn van invloed op de beoordeling.
→ Oplossing: onafhankelijke beoordeling of een tweede beoordelaar.

18. Een hoge p-waard betekent dat heel veel mensen de vraag goed hebben beantwoord. Item 5 moet eruit want lage gokkans. Rir-waarde is de correlatie tussen de itemscore en de score op de resterende items van dezelfde toets.

19. In dit geval had er moeten staan: 'wat is het allerbelangrijkste bij het maken van een goede toets volgens het boek?
20. Sommige items zijn positief geformuleerd en sommige zijn negatief geformuleerd. Je moet dus ompolen (hercoderen). Dit heeft geen invloed op de validiteit maar wel op de betrouwbaarheid.

26. De programmatische toetsing van van Vleuten draagt bij aan het leren van de student maar leidt ook tot een beoordeling van het presteren van de student. Betekenisvolle feedback en reflectie hierop zijn essentieel. Alle individuele toetsmomenten hebben hun focus op het geven van feedback. Dit wordt toegepast bij de studie diergeneeskunde. Een besluit voor promotie naar een volgend leerjaar of voor een diploma komt vervolgens tot stand door aggregatie van een veelvoud aan toetsmomenten.

27. Effectieve feedback geeft volgens Hattie en Timperley antwoord op 3 vragen:
Feedup: Wat is het einddoel van de student? (Wat zijn de beoordelingscriteria?)
Feedback: Hoe heeft de student de taak uitgevoerd? (Welke vooruitgang wordt geboekt?)
Feedforward: Welke aanpak is nodig om het doel te bereiken?

De feedback heeft betrekking op vier aspecten:
Task level- criteria en standaarden van de taak
Process level- proces van het probleem oplossen
Self-regulation level- niveau van het zelfsturingsproces
Self level- eigenschappen van de student

29. De manier van rapporteren en het taalgebruik door je rapport heen hangt sterk af van de doelgroep.