

De betrouwbaarheid van het mondeling examen nader bekeken

T. Klaassen, C.P.M. van der Vleuten, R.J. Rotteveel

Inleiding

Het mondeling examen is zowel nationaal als internationaal een veel gebruikte toetsvorm in het medisch onderwijs. De oorzaak van deze populariteit is wellicht dat docenten ervan overtuigd zijn dat het mondeling examen een valide beeld geeft van de competentie van een kandidaat. Ook het feit dat het examen een direct contact toelaat tussen kandidaat en examinerator draagt bij tot de populariteit. Doorgaans geven examinatoren aan dat, op grond van het 'klinische oordeel', snel duidelijk is wat voor vlees men in de kuip heeft.

In het medisch onderwijs wordt het mondeling examen vaak gebruikt in een klinische context. De kandidaat wordt ondervraagd aan de hand van een concreet patiëntenprobleem en beoordeeld op zijn of haar klinisch inzicht en begrip. In Nederland is deze vorm bekend als het 'patiëntexamen'. Ook in de Angelsaksische landen is het mondeling een van de populairste en belangrijkste examenvormen. Het daar als 'viva voce' aangeduide examen kent een korte en een lange variant, respectievelijk de 'short case' en de 'long case'.

Mondelinge examens kunnen ook verschillen in de mate waarin de afname en de scoring gestructureerd zijn. Het klassieke patiëntexamen kent weinig structuur. De examinatoren zijn autonoom in hun handelen en in het toekennen van waardering. In de meest gestructureerde vorm wordt gewerkt met protocollen zowel voor de handelwijze van de examinerator als voor de scoring van de antwoorden van de kandidaat. In dit zogenaamde Gestructureerde Mondelinge Examen (GME) speelt de examinerator zowel de rol van beoordelaar als die van gesimuleerde patiënt.¹

Het mondeling examen wordt echter ook verguisd, met name vanwege de lage betrouwbaarheid. Vanwege die betrouwbaarheidspro-

blemen is het mondeling reeds in de jaren zestig uit de Amerikaanse nationale examens verdwenen.² De overeenstemming tussen examinatoren is wisselend, maar over het algemeen laag.³⁻⁵ Naarmate het examen meer gestructureerd is, wordt de interbeoordelaarsbetrouwbaarheid wat hoger.^{6,7}

Vraagstelling

Bij de in de literatuur gerapporteerde betrouwbaarheden van het mondeling examen valt echter een kanttekening te plaatsen. Vrijwel alle studies hebben betrekking op de interbeoordelaarsbetrouwbaarheid: de mate waarin examinatoren het eens zijn wat betreft het oordeel over de kandidaat. Het is duidelijk dat onenigheid tussen examinatoren een belangrijke ruisfactor is in het mondeling examen. Er zijn echter méér factoren die in een toets ruis kunnen veroorzaken. Al deze factoren tezamen bepalen de betrouwbaarheid van een toetsvorm. Uit de literatuur over toetsing van medische competentie komt eenduidig naar voren dat de belangrijkste vorm van onbetrouwbaarheid wordt veroorzaakt door de variabiliteit in de kunde van kandidaten over onderdelen van de stof.⁸ De competentie van kandidaten is sterk afhankelijk van de inhoud van het bevroegde. Dit wordt ook wel als het probleem van de inhoudsspecificiteit aangeduid. Het betekent dat de toevallige steekproef uit de leerstof, die in een toets aan bod komt, van grote invloed kan zijn op de uiteindelijke waardering van de kandidaat. Zo kan in een mondeling examen de prestatie van een kandidaat op de ene casus weinig overeenkomst vertonen met die op een andere casus. Even goed als de onenigheid tussen examinatoren een ongewenst toevalsresultaat kan opleveren, kan ook de 'onenigheid' tussen casus over de be-

kwaamtheid van de kandidaat een ernstige foutenbron zijn. Om dit probleem te verhelpen dient het aantal getoetste casus in een toets groot te zijn. Doorgaans impliceert dat een toetsijfd van enkele uren, ongeacht de toetsvorm.⁹

Er zijn veel publikaties waarin resultaten worden gepresenteerd van onderzoek naar de invloed van het inhoudspecifiteitsprobleem bij verschillende toetsvormen. Het mondeling examen is daarop evenwel een uitzondering. Ons is slechts één publicatie bekend waarin op adequate wijze deze specifieke foutenvariante onderzocht is voor het mondeling examen. De gerapporteerde gegevens zijn slechts summier, omdat, vanwege de politieke gevoeligheid van de uitkomsten, de herkomst en samenstelling van de dataset niet mochten worden geopenbaard. Uit het onderzoek bleek dat de foutenvariante veroorzaakt door intercasusvariatie groot was en dat een groot aantal casus noodzakelijk bleek om betrouwbare gegevens te kunnen verkrijgen.

Adequaat betrouwbaarheidsonderzoek is belangrijk, omdat de mogelijkheid bestaat dat het mondeling examen op onjuiste gronden van onbetrouwbaarheid wordt beticht. De onbetrouwbaarheid zou vooral veroorzaakt kunnen worden doordat in het mondeling de leerstof inhoud onvoldoende wordt gedeekt. Daarmee zou het niet minder betrouwbaar of onbetrouwbaar zijn dan de meeste andere toetsvormen. Dat kan belangrijke consequenties hebben. Voor het verhogen van de betrouwbaarheid is het in dat geval zaak het aantal casus uit te breiden, dan wel de toets te verlengen. Dat schept ook goede mogelijkheden om de invloed van examinator-onenigheid te vermindern. Door verschillende examinatoren verschillende delen van de leerstofinhoud te laten beoordelen zullen milde en strenge oordelen elkaar per kandidaat neutraliseren. Zelfs bij matige interbeoordelaarsovereenstemming kan aldus een adequate betrouwbaarheid worden bereikt.^{10,11} Deze procedure zal de nodige

inspanningen en kosten vergen, maar ook dat is niet uniek voor het mondeling examen.

Het doel van deze studie is de betrouwbaarheid van een mondeling examen nader te onderzoeken door zowel interbeoordelaarsvariatie als intercasusvariatie te betrekken in een betrouwbaarheidschatting.

Methode

Onderwerp van studie was het mondeling examen psychiatrie dat wordt afgenomen aan het einde van een acht weken durend co-assistentie schap psychiatrie in het zesde studiejaar geneeskunde aan de Faculteit der Geneeskunde van de Rijksuniversiteit Limburg. Onderdeel van het stageprogramma is het schriftelijk uitwerken door de student van zes uit de praktijk voortkomende casus volgens een vast stramien (achtergrond en ziektegeschiedenis van de patient, diagnostiek en suggesties voor behandeling). Deze uitgewerkte casuïstiek dient een dag voor het mondeling examen te worden ingeleverd. De examinatoren selecteren twee casus waarover zij op het mondeling examen vragen zullen stellen. De student wordt eveneens niet op de hoogte gebracht van deze selectie.

De bij het co-assistentieschap betrokken vakgroepen leveren de examinatoren (psychiaters en psychologen). Bij de toewijzing van examinatoren aan examenkandidaten wordt ervoor gezorgd dat de examinatoren niet op de hoogte zijn van het functioneren van de student tijdens de stage.

De (ongetrainde, maar vaak ervaren) examinatoren beschikken over een examenformulier met onderwerpen, dat als richtlijn kan dienen. Ze zijn echter vrij niet al dan niet te gebruiken. Sommige examinatoren blijven dicht bij de casus, voor anderen is de casus aanleiding om ook vragen te stellen over andere gebieden van de psychiatrie. De toetsaankomst van het cijfer is een automatische afweging van de examinator, waarvoor geen nadere protocollering aanwezig is. Kortom, afgezien van de casusindeling en -selectie betreft het een ta-

melijk ongestructureerde vorm van het mondeling examen.

In elk mondeling examen wordt voor elke casus een half uur uitgetrokken. In 1994 heeft men de afnameprocedure gewijzigd. In de oude situatie waren per student twee examinatoren tegelijkertijd aanwezig die elk één casus afnam. Per casus gaf elke examinator onafhankelijk van de ander een heel cijfer. Het eindcijfer was het gemiddelde van de aldus verkregen vier cijfers. In de *nieuwere situatie* zijn de twee casus gehandhaafd, maar uit kostenoverwegingen nemen beide examinatoren apart en na elkaar elk één casus af. De onafhankelijk gegeven cijfers van beide examinatoren worden gemiddeld tot een eindcijfer.

Elke student dient een voldoende te halen voor het examen (6 of hoger). Wanneer een 5½ als eindcijfer wordt verkregen, volgt verhoging van een kwartier van het mondeling bij beide examinatoren. Wanneer een student gezakt is, dient het examen herhaald te worden op een later tijdstip bij een van dezelfde examinatoren.

Alleen de oorspronkelijke cijfers werden in de analyse opgenomen, dat wil zeggen met weglating van verlegingsresultaten en herkansingsgegevens. Complete gegevens werden verzameld van 141 studenten die in de oude situatie waren getaxamineerd en van 51 studenten die examen deden in de nieuwe situatie.

Naast beschrijvende statistische gegevens bespand de statistische analyse uit een schatting van de betrouwbaarheid gebruik makend van de generaliseerbaarheidstheorie.^{12,13} De twee afname-situaties verspreiden een verschillende aanpak, maar vulden elkaar goed aan. Voor de examens die afgenomen waren in de oude situatie werden variantiecomponenten geschat voor personen (studenten), casus (geenst binnen personen omdat de casuïstiek per student verschilde), beoordelaars (examinatoren, geenst binnen personen omdat verschillende studenten verschillende examinatoren toebedeeld kregen; gekruist met casus omdat

beide examinatoren beide casus afnamen) en hun interactie-effecten. Dit examen-design heeft als voordeel dat beoordelaarsvariatie van casusvariatie kan worden onderscheiden. Het heeft als nadeel dat de prestaties op de twee casus waarschijnlijk niet helemaal onafhankelijk van elkaar zijn, omdat dezelfde beoordelaars een oordeel geven over beide casus. Het voordeel over de ene casus kan doorwerken in het oordeel over de andere (halo-effect).

Voor de examens die afgenomen werden in de nieuwe situatie werden alleen variantiecomponenten geschat voor personen en casus. Omdat per casus slechts gegevens van één beoordelaar voorhanden waren, valt geen onderscheid te maken in verdere variantiebronnen. Beoordelaars-effecten en casuseffecten zullen de betrouwbaarheid beïnvloeden, maar onduidelijk zal zijn welk effect in welke mate bijdraagt. Het oordeel over de ene casus is echter in dit design wel onafhankelijk. Beide afname-situaties vallen elkaar daarmee wat voor- en nadelen betreft goed aan.

De variantiecomponenten werden gebruikt voor een schatting van generaliseerbaarheidsefficiënten (G) en standaardfouten (SEM) voor de bestaande, werkelijke situatie, maar ook geprojecteerd naar wat verkregen zou worden bij een ander aantal casus en beoordelaars. De G-coëfficiënten mogen worden opgevat als de verwachte correlatie die verkregen zou worden wanneer het mondeling uit andere casus en beoordelaars had bestaan. De SEM kan gebruikt worden om een betrouwbaarheidsinterval rondom het cijfer van een individuele uitslag te bepalen.

Resultaten

In tabel 1 is beschrijvende statistiek weergegeven van de eindcijfers die gegeven zijn in de oude en de nieuwe situatie. Opvallend is dat gemiddeld maar juist boven de zes wordt gescored. In de oude situatie zakte 18% van de deelnemers voor de eerste keer en in de nieuwe situatie 14%. Zoals gewoontelijk bij cijferwaar-

De betrouwbaarheid van het mondeling examen nader bekeken

T. Klaassen, C.P.M. van der Vleuten, R.J. Retzevel

Tabel 1. Beschrijvende statistieken betreffende eindcijfers van het mondeling examen psychiatrie

| Oude situatie | N | gemiddelde | | sd | minimum | maximum |
|-----------------|-----|------------|------|----|---------|---------|
| | | 1 | 2 | | | |
| Oude situatie | 141 | 6,31 | 1,30 | | 4 | 9 |
| Nieuwe situatie | 51 | 6,84 | 1,08 | | 5 | 9 |

Tabel 2. Schattingen van variantiecomponenten van het mondeling examen psychiatrie (P=personen, C=casus, B=beoordelaars, J=sson bakjes de standard error van deze schattingen)

| Variantiebron | Oude situatie | | Nieuwe situatie | |
|----------------|------------------------------|---------------------------------|------------------------------|---------------------------------|
| | Geschatte variantiecomponent | Percentage van totale variantie | Geschatte variantiecomponent | Percentage van totale variantie |
| P | 1,38 (1,77) | 78,86 | 0,61 (0,18) | 55,26 |
| C/P | 0,02 (0,03) | 1,14 | | |
| B/P | 0,04 (0,03) | 2,29 | | |
| Algemene error | 0,31 (0,04) | 17,71 | 0,51 (0,10) | 44,74 |

Tabel 3. Betrouwbaarheids-schattingen van het mondeling examen psychiatrie als functie van het aantal casus, beoordelaars en afnamestrategieën

| G-corr | Aantal casus | Toetsveld in uren | Oude situatie | | Nieuwe situatie | |
|--------|--------------|-------------------|---|------|---|------|
| | | | Aantal deze/alle examinatoren per casus | | Aantal verschillende examinatoren per casus | |
| | | | 1 | 2 | 1 | 2 |
| SEM | 1 | 0,5 | 0,79 | 0,88 | 0,79 | 0,88 |
| | 2 | 1,0 | 0,87 | 0,93 | 0,88 | 0,93 |
| | 3 | 1,5 | 0,90 | 0,95 | 0,92 | 0,96 |
| | 4 | 2,0 | 0,92 | 0,96 | 0,94 | 0,97 |
| SEM | 1 | 0,5 | 0,61 | 0,44 | 0,61 | 0,44 |
| | 2 | 1,0 | 0,45 | 0,33 | 0,43 | 0,31 |
| | 3 | 1,5 | 0,39 | 0,28 | 0,35 | 0,25 |
| | 4 | 2,0 | 0,35 | 0,25 | 0,30 | 0,22 |

deringen wordt geen gebruik gemaakt van het bereik van de gehele schaal. Opvallend is ook dat in de oude situatie gemiddeld lagere cijfers werden gegeven dan in de nieuwe situatie. Dit verschil is statistisch significant ($F(1, 190) =$

6,99, $p < 0,01$). Overige uitsplitsingen naar

sexe, volgorde van examinering (eerste, tweede of derde kandidaat die de examinator die dag beoordeelde) en score van eerste en tweede

examinator leverden geen statistisch significante verschillen op.

In tabel 2 worden de variantiecomponenten gegeven. Wat sterk opvalt is de grootte van de persoonscategorie. Blijkbaar is het mondeling in staat goed onderscheid te maken tussen studenten. In de oude situatie wordt zelfs drie kwart van de variantie veroorzaakt door verschillen tussen studenten; in de nieuwe situatie is dat nog altijd meer dan de helft. Opgemerkt moet echter worden dat de bijbehorende standaard errors eveneens groot zijn en er dus mogelijk sprake is van toevalsresultaten. Het verschil in P voor de oude en de nieuwe situatie zou veroorzaakt kunnen worden door deze grote standaard errors, maar ook door de eerder genoemde doorwerking van het oordeel over de ene casus op het oordeel over de andere. In de oude situatie wordt weinig variantie verklaard door C/P. Dit is een combinatie van het hoofd-effect C (moortilijkheidsgraadverschillen tussen casus) en het interactie-effect Personen x Casus (de inconsistentie van de kindte van studenten bij verschillende casus). Juist deze laatste term is doorgaans zeer groot in complexiteitsontingen. B/P zijn bevoordelaars-effecten (het hoofd-effect B, ofwel strengheidsverschillen tussen beoordelaars, en de interactie B x P, ofwel inconsistentie van beoordelaars over personen). Blijkbaar is er in de oude situatie sprake van beperkte examinatorvariancie.

De betrouwbaarheidsgegevens zijn samengevat in tabel 3 als functie van het aantal casus. Ter inschatting van de kosten en de vergelijkbaarheid met andere toetsvormen, is eveneens de toetsveld vermeld.

Door het meeruitgebreide design in de oude situatie kunnen effecten van verschillende afnamestrategieën worden nagegaan. In de feitelijke situatie bleven de examinatoren in beide casus dezelfde. De feitelijke betrouwbaarheid van het examen is dus 0,93. De SEM is 0,33. Een dubbele SEM (gemiddeld 1,96 x SEM) geeft het 95% betrouwbaarheidsinterval weer. Een 6 zou in de feitelijke situatie met 95% zekerheid ook (ongeveer) een 5,4 kunnen zijn of een

6,6. Geschied we zouden als minimumis hanteren dat een verschil van 1 punt met redelijke zekerheid wordt vastgesteld, dan geldt een minimale SEM van 0,51 ($1,96 \times 0,51 = 1$). Met dit criterium zou zelfs met één examinator kunnen worden volstaan of met één casus gebruik makend van twee examinatoren.

Het design laat het ook toe een projectie te maken naar een situatie waarin voor elke casus een andere examinator wordt ingezet. Opnieuw kan weer een uitsplitsing worden gemaakt voor één of twee examinatoren per casus (let wel: dat betekent dus dat bijvoorbeeld bij twee casus en één enkele examinator in totaal twee examinatoren voor het mondeling nodig zijn, bij dubbele examinatoren $2 \times 2 = 4$ examinatoren). Bij vergelijking van de vorige en deze afnamestrategie blijkt dat de laatste efficiënter is, met name wanneer meer casus worden gebruikt: het toevoegen van casus met verschillende examinatoren is voordeliger dan het toevoegen van casus bij dezelfde examinatoren of dan het toevoegen van meer examinatoren bij een gelijkblijvend aantal casus.

In de nieuwe situatie zijn de betrouwbaarheidsgegevens slechter dan in de oude situatie. Theoretisch zou de projectie vanuit de oude situatie naar één, per casus verschillende examinator met een mondeling bestaande uit twee casus ($G = 0,88$; $SEM = 0,43$) overeen moeten komen met twee casus in de nieuwe situatie ($G = 0,71$; $SEM = 0,50$). De verschillen tussen beide situaties in betrouwbaarheid zijn echter groot. Zoals eerder gezegd, kan dit verschil veroorzaakt worden door 'ruis' in de betrouwbaarheids-schattingen dan wel door een halo-effect in de oude situatie. Niettemin, ook in de nieuwe situatie blijft de betrouwbaarheid dusdanig dat verschillen van één cijfer serieuus mogen worden genomen.

Discussie

De gegevens in deze studie maken duidelijk dat het belangrijk is om bij een mondeling examen rekening te houden met verschillende

variantiebronnen. Ook kan op grond van deze gegevens beslist *niet* gesteld worden dat het mondelinge een onbetrouwbaar examen zou zijn. Zelfs wanneer gebruik gemaakt wordt van slechts één examinator kunnen betrouwbare resultaten worden bereikt door het examen te verleggen en de steekproef uit de leerstof te vergroten. Uit het gemiddelde verschil in cijferwaaierdijking tussen de oude en de nieuwe situatie lijkt het er wel op dat hij dubbele beoordelingen de examinatoren strenger worden. De conclusie is echter gerechtvaardigd dat een goede strategie in een mondelinge examen bestaat uit de inzet van meerdere examinatoren die verschillende inhoud onafhankelijk van elkaar beoordelen. Een omgekeerde strategie (kort mondelinge, dubbele examinatoren) valt af te raden. Wanneer de beschikbaarheid van examinatoren dat toehaalt, is, vooral bij korte mondelinge examens (dan wel examens die een gering deel van de stof dekken, ofwel weinig casus bevatten), een mengvorm bestaande uit paren examinatoren die verschillende inhoud onafhankelijk van elkaar beoordelen de optimale strategie. Bij langere mondelinge examens die de leerstof breder dekken staat de winst in betrouwbaarheid van dubbele examinatoren echter niet in verhouding tot de logistische kosten ervan.

Fantje relativerende opmerkingen zijn echter dringend op hun plaats. Het aantal waarnemingen, zowel met betrekking tot studenten, casuïstiek en examinatoren, is beperkt waardoor schattingsfouten in de betrouwbaarheidsberekeningen mogelijk zijn. Bovendien blijkt uit de verschillen tussen de oude en de nieuwe situatie dat halo-effecten in de oude situatie wellicht een rol hebben gespeeld. Ook ligt het voor de hand dat onderdeel van meerdere examinatoren die bij elkaar in één ruimte zitten nooit helemaal onafhankelijk zullen zijn en dat kan de betrouwbaarheidsgegevens verder geflatteerd hebben. Maar ook voor de nieuwe situatie, waarin beide effecten nauwelijks een rol kunnen spelen, is de geboekte betrouwbaarheid hoog te noemen en vergelijkbaar met, of

zelfs beter dan, toetsen bestaande uit schriftelijke gestloten vragen.⁹ Dat is eigenlijk een vreemde bevinding. De resultaten zijn ook behoudend beter dan die in de eerder genoemde studie over mondelinge examens, die gebaseerd waren op een veel groter aantal waarnemingen in Misschen dat een deel van de verklaring gezocht kan worden in de uniformiteit van de taak die de kandidaten in de stage moeten verrichten. Het blijft de vraag of dit in andere situaties, tot andere bevindingen zou leiden. Kortom, voorzichtigheid is geboden bij het trekken van conclusies op basis van deze ene studie, maar het lijkt ons van het grootste belang vergelijkbare betrouwbaarheidsstudies van mondelinge examens te gaan uitvoeren.

Literatuur

1. Maasch JL. Model for a criterion-referenced medical specialty test. Office of Medical Education Research and Development, Michigan State University in collaboration with the American Board of Emergency Medicine, 1980.
2. Van Han, I. Het mondeling examen. In: Meijer CM, Schepker AJA, Van der Vleuten CPM, eds. Medisch onderwijs in de praktijk. Assen: Van den Broek, 1995: 10-17.
3. Lantz ME, Stohli IA. The effect of rater severity on person ability measure: a test model analysis. *Am J Occup Ther* 1993; 47: 311-7.
4. McCormick WO. A practical oral rating scale: inter-rater reliability. *Can J Psychiatry* 1981; 26: 286-9.
5. Ruizin JW, Junberg L, Brownlie RC, et al. A study of the reliability of the grading process used in the American Board of Pediatrics oral examination. *J Med Educ* 1982; 57: 924-6.
6. Yang JC, Laude W. Improvement of reliability of an oral examination by a structured evaluation instrument. *J Med Educ* 1983; 58: 864-72.
7. Van Ham E, Gerritsma J. The assessment of clinical competence in general practice with their simulated recall. In: Bender W, Herstein RJ, Schepker AJA, Zwaersma RJ, eds. Teaching and assessing clinical competence. Groningen: Backwork, 1990: 306-9.
8. Van der Vleuten CPM, Van Linck SJ, Schuwirth LW. Toetsing op lesonaïschkijne in het medisch onderwijs. *Ned Tijdschr Geneesd* 1994; 138: 1388-92.

9. Van der Vleuten CPM, Newble DJ. Methods of assessment in certification. In: Newble D, Jolly B, Wakeford R, eds. Issues in the assessment of clinical competence. Cambridge: Cambridge University Press, 1994: 105-23.
10. Swanson DB. A measurement framework for performance-based tests. In: Hartik, Handen RM, eds. Further developments in assessing clinical competence. Montreal: Can-Heal, 1987: 13-45.
11. Van der Vleuten CPM, Norman GR, De Graaf E, Pitfalls in the pursuit of objectivity. *Med Educ* 1991; 25: 110-8.
12. Cronbach LJ, Gieser GC, Nanda H, Kaganman N. The dependability of behavioral measurements: generalizability for scores and profiles. New York: John Wiley and Sons, 1972.
13. Van der Vleuten CPM, Wijnen WHWE. Niet praktische dan een goede theorie: Generaliseerbaarheidstheorie als instrument voor betrouwbaarheidsstudies. *Bulletin Medisch Onderwijs* 1991; 10: 2-14.

DE AUTEURS

T. Klaassen is studente geneeskunde, Faculteit der Geneeskunde, Rijksuniversiteit Limburg.
C.P.M. van der Vleuten, psycholoog, is universitair hoogleraar bij de vakgroep Onderwijsontwikkeling en -evaluatie, Rijksuniversiteit Limburg.
R.J. Rotteveel, psycholoog, ten tijde van het onderzoek aangesteld als universitair docent bij de vakgroep Psychologie en Neuropsychologie, Faculteit der Geneeskunde, Rijksuniversiteit Limburg.

Correspondentie-adres:

C.P.M. van der Vleuten, Vakgroep Onderwijsontwikkeling & Onderwijsresearch, Rijksuniversiteit Limburg, Postbus 616, 6200 MD Maastricht.

Oproep

De tweede auteur van dit artikel roept personen die over soortgelijke gegevens van mondelinge examens beschikken op deze ter beschikking te stellen voor vergelijkbare analyse. In aanmerking kunnen datasets waarin meerdere examinatoren verschillende delen van de leerstof onafhankelijk van elkaar beoordelen (al of niet dubbel).