Betrouwbaarheid van beoordelingen van afstudeerwerk in het hbo; een literatuuronderzoek

Matthijs de Feber, november 2015

Inleiding

In 2011 kwam een hbo instelling negatief in het nieuws omdat studenten te makkelijk aan hun cijfers zouden zijn gekomen en daardoor konden afstuderen. Het gevolg daarvan was dat twee commissies onderzoek gingen doen naar het niveau en de betrouwbaarheid van de examens in het hogere beroepsonderwijs (hbo) in Nederland. De commissie-Dunnewijk (2011) concludeerde dat bij de door haar onderzochte opleidingen het te vaak voorkwam dat het niveau van de eindwerkstukken onvoldoende was, terwijl de opleidingen deze wel voldoende hadden beoordeeld. De commissie-Bruijn (2011) stelde dat: "de samenleving het zich eenvoudigweg niet kan veroorloven dat twijfels bestaan over de examinering binnen een voor Nederland zo belangrijke onderwijssector als het hbo."[1] De oplossing ligt volgens hem bij de objectivering van het afstudeerniveau van de student. Objectiviteit en betrouwbaarheid van beoordelen zijn de sleutelbegrippen in deze discussie. Vooruitlopend op een experimenteel onderzoek waarin geprobeerd wordt vast te stellen of examinatoren die niet betrokken zijn geweest bij het afstudeerproces van een kandidaat tot een betrouwbaarder oordeel komen dan de begeleiders van deze studenten, wordt in een literatuurstudie onderzocht wat er wetenschappelijk gepubliceerd is over dit onderwerp. Is het oordeel van een betrokken docent minder betrouwbaar dan de beoordeling van een niet betrokken onafhankelijke examinator? Op basis van het rapport van de commissie-Bruijn (2011) zou men mogen verwachten dat de betrokken docent inderdaad minder betrouwbaar is in zijn oordeel.

Selectie

De literatuurstudie heeft 14 relevante artikelen opgeleverd. Alvorens de verschillende publicaties inhoudelijk te bespreken is gekeken aan welke wetenschappelijk standaarden de artikelen voldoen. De artikelen zijn bijna allemaal peer-reviewed en gepubliceerd. De vraag is welke studies daadwerkelijk een effect aantonen en ingaan op causaliteit. Als basis hiervoor dient het artikel: Systematic Reviews in Education Research: When Do Effect Studies Provide Evidence? (Van Klaveren & De Wolf, 2013). In dit artikel wordt uiteen gezet dat vier inclusiecriteria relevant zijn:

I. De onderzoeksontwerpen (designs) zijn allemaal gerandomiseerde experimenten of quasi observationele studies.
II. De ontwerpen laten allemaal zien dat de geobserveerde kenmerken tussen de interventie- en controle groep gelijk verdeeld zijn.
III. Alle studies zijn gepubliceerd en peer-reviewed.
IV. Studies dienen uitsluitend te gaan over het onderwerp van onderzoek.

Voor deze literatuurstudie zijn criteria II en IV lichtelijk aangepast:

[1] *De Volkskrant* 14 mei 2012

(II.) *In de studie wordt in ieder geval ingegaan op een aantal geobserveerde verschillen en overeenkomsten tussen twee populaties.*
(IV.) *Er sprake van een verricht onderzoek (anders dan een literatuurstudie), dat gaat over het onderwerp van het onderzoek.*
In Tabel 1 zijn de criteria toegepast op de geselecteerde artikelen.

Tabel 1 *Inclusiecriteria*

Artikel	I.	II.	III.	IV.
Alexander & Praeger, 2009	X	X	√	√
Alvero, 2008	√	X	√	√
Baume et al., 2004	X	X	√	√
Bailey, 2014	X	X	√	X
Bloxham et al., 2015	X	X	√	√
Childs et al., 2009	X	X	√	√
Conway et al., 1995	X	X	√	√
Hand & Clewes, 2000	X	X	√	√
Hoyt, 2000	X	X	√	√
Klein et al., 2012	X	√	√	√
Knight, 2002	X	X	√	√
Streatmans, 2006	X	X	X	X
Vleuten & Schuwirth, 2005	X	X	√	√
Webster, 2000	X	X	√	√

Zoals uit Tabel 1 blijkt, zijn geen studies gevonden die aan alle inclusiecriteria voldoen. Een mogelijk verklaring daarvoor is dat, als er een goed onderzoeksdesign gebruik wordt, de uitkomstmaat in veel gevallen het cijfer is dat een student krijgt. In het primair en secundair onderwijs is het vaak nog mogelijk om te relateren aan CITO-scores of de cijfers van de centraal schriftelijke examens. Ook internationaal zijn er cijfers zoals PISA-scores en SAT waar een vergelijking mee kan worden gemaakt. In het hoger onderwijs lijkt dit niet te bestaan. Overigens laten Dee et al (2001) in een onderzoek zien dat er ook vraagtekens te zetten zijn bij de betrouwbaarheid van bepaalde gestandaardiseerde schoolexamens. Uiteraard zou het mogelijk zijn later succes op de arbeidsmarkt als uitkomst variabele op te nemen, maar dat vereist een studie op lange termijn terwijl er in de dynamiek van het hoger onderwijs zoveel verandert, dat het nauwelijks mogelijk is gegeven beoordelingen zinvol te koppelen aan dit succes. Daarom worden ook artikelen besproken die niet aan de bovenstaande criteria voldoen. Wellicht bieden ze toch inzichten en gedachten die verder onderzoek rechtvaardigen.

Beoordelen afstudeerwerk en betrouwbaarheid

Betrouwbaarheid wordt over het algemeen uitgedrukt in termen van reproduceerbaarheid van een beoordeling. Uiteraard zijn er veel bronnen van fouten of bias, die reproduceerbaarheid negatief beïnvloeden. Volgens Vleuten en Schuwirth (2005) is het van belang dat het beoordelen van competenties altijd afhankelijk is van context en inhoud en dat, om de reproduceerbaarheid te vergroten, er een grote steekproef (monster)

van deze inhoud en context genomen moet worden in de toets. Zij beweren in tegenstelling tot wat algemeen wordt aangenomen dat betrouwbaarheid niet afhankelijk is van objectiviteit en eventuele standaardisering, maar van vooral de grootte van de te beoordelen monster (steekproef). Dat leidt tot grotere reproduceerbaarheid en dat is vaak gerelateerd aan het aantal uren dat de toetsing duurt. Ze laten zien dat vrijwel iedere toetsvorm die korter dan vier uur duurt tot reproduceerbaarheid van minder dan 0.80 leidt, waarvan over het algemeen wordt aangenomen dat dat minder betrouwbaar is. Ervan uitgaande dat in de gangbare onderwijspraktijk toetsen minder lang dan vier uur duren en dat dat ook geldt voor afstudeerassessments is de vraag relevant of we in staat zijn om betrouwbaar te beoordelen. Onder de veelzeggende titel 'Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria' komen Bloxham et al. (2015) tot de conclusie dat er bijzonder weinig eenduidigheid is in het beoordelen van studentwerk. In het onderzoek werd studentwerk uit verschillende academische disciplines beoordeeld door zes willekeurig toegewezen beoordelaars. Op basis verschillende criteria werd door de beoordelaars bepaald welke volgorde van goed tot slecht ieder werkstuk gerangschikt zou moeten worden. De beoordeling zelf is hier dus geen uitkomstmaat; er is gekeken naar ordinale verschillen. Bij slechts één op de twintig beoordeelde werkstukken was overeenstemming te vinden.

Knight (2002) stelt eveneens dat summatieve (eind)assessments in het hoger onderwijs volstrekt onbetrouwbaar zijn. Het komt erop neer dat wat er getoetst moet worden veel te complex is om te toetsen en dat het vrijwel onmogelijk is dit kosteneffectief te doen. Vergelijk dit met de stelling over toetsduur en betrouwbaarheid van Vleuten en Schwuwirth, 2005. Knight ondersteunt zijn betoog echter nergens met cijfers en geeft niet aan hoe het wel zou kunnen als geld geen rol zou spelen. Toch ziet hij wel ruimte voor verbetering afgezien van een betere training van examinatoren stelt hij voor om twee examinatoren onafhankelijk van elkaar het werk te laten beoordelen en bij voorkeur zonder iets te weten over de achtergrond van de student zelf.

Baume et al. (2014) constateerde dat de betrouwbaarheid van beoordelingen in het hoger onderwijs niet al hoog wordt ingeschat. Zij proberen te achterhalen hoe examinatoren tot hun oordeel kwamen. Om te beginnen hebben ze een 140 portfolio's opnieuw laten beoordelen. Deze nieuwe beoordelingen vielen flink lager uit dan de oorspronkelijke beoordelingen. Aan de hand van interviews probeerden ze te achterhalen hoe die verschillen verklaard konden worden. Een van de conclusies die ze trokken was dat examinatoren ook 'door de beoordelingen heen' beoordelen. Dat wil zeggen dat ze een indruk hebben van het algemene niveau en met die indruk individuele beoordelingscriteria (her)beoordelen. Daardoor wordt gericht naar 'geslaagd' of 'gezakt' toe beoordeeld. Voor het verschil tussen de oorspronkelijke beoordelingen en de beoordelingen in dit experiment geven de auteurs niet een hele duidelijke verklaring: in het experiment moesten de examinatoren meer commentaar geven en men wist dat het niet 'echt' was. De auteurs merken op dat de originele examinatoren hun studenten kenden en vaak al van feedback hadden voorzien. Ze stellen het niet zo expliciet, maar het is goed mogelijk dat hieruit een belangrijk deel van de verschillen verklaard kan worden. Klein et al. (2012) lopen ook tegen dit probleem aan. In een kwantitatief onderzoek naar het afstudeerwerk van verschillende masteropleidingen van een

Nederlandse universiteit wordt de rol van de afstudeerbegeleider onderzocht. In dit artikel wordt geconstateerd dat deze begeleider traditioneel ook een van de twee examinatoren is en dat dat tot een belangenconflict kan leiden. Als begeleider heb je immers het leerproces beïnvloed en ben je dus medeverantwoordelijk voor het eindresultaat. Het onderzoek laat zien dat er een positieve relatie is tussen betrokkenheid van de begeleider en de uiteindelijke beoordeling. Het probleem van dit onderzoek (zoals gedeeltelijk ook erkend) is dat het subject van het onderzoek actief de uitkomstmaat van het onderzoek bepaalt. Een verklaring voor deze uitkomst zou kunnen zijn dat de meer betrokken begeleiders zich meer verantwoordelijk voelen voor een positief eindoordeel en daardoor hoger beoordelen. Het zou interessant zijn geweest te zien wat de verschillen in beoordeling tussen de twee beoordelaars zijn geweest.

Webster et al. (2000) wijzen er in hun artikel op dat in Groot Brittannië de politiek (net als in Nederland overigens) druk uitoefent om het niveau van academische opleidingen op peil te houden en te garanderen. Dit in tegenstelling tot de VS waar hoger onderwijs minder centraal georganiseerd is. In het kader daarvan zijn tachtig afstudeerwerkstukken onder de loep genomen en komen zij tot de conclusie dat beoordelingscriteria niet overal juist werden toegepast en dat het erop leek (conform de conclusie van Baume et al., 2004) dat er ook gebruik gemaakt werd van een breder impliciet beoordelingskader. Ook hier lijkt dus een bepaalde vooringenomenheid van een examinator een rol te spelen. Het kwalitatieve onderzoek van Hand (2000) komt tot een vergelijkbare conclusie: examinatoren in het hoger economische onderwijs in Groot Brittannië maken nauwelijks gebruik van het beschikbare beoordelingsmodel dat voorhanden is bij het beoordelen van afstudeerwerkstukken, maar hebben ook de neiging de scripties holistisch te benaderen. Straetmans (2006) zei daarover in zijn lectorale rede: “De verschuiving van een analytische naar een meer holistische beoordeling betekent dat er meer ruimte komt voor de subjectieve interpretaties van assessoren.” Childs et al. (2009) gaan in op de betrouwbaarheid van beoordelaars (betrouwbaar in de betekenis van accuratesse). Zij stellen vast dat er relatief weinig studies zijn naar deze accuratesse omdat er meestal geen externe criteria voorhanden zijn. In een experimentele setting bij een docentenopleiding (post-bachelor) in Canada hebben zij onderzocht of de variatie in beoordelingen te verklaren valt uit de algemene neiging van een individuele examinator te streng (horn effect) of te soepel (halo effect) te beoordelen. Zij concluderen dat er inderdaad bepaalde examinatoren structureel lagere en andere structureel hogere beoordelingen afgeven. De variatie tussen beoordelingen werd in dit onderzoek voor 15 procent hierdoor verklaard.

Betrouwbaarheid en rubrics

In veel opleidingen worden rubrics toegepast met als doel de betrouwbaarheid van de beoordeling van vaardigheden, beroepshandelingen of competenties te verhogen. Per vaardigheid (of competentie of handeling) zijn er een aantal criteria geformuleerd op basis waarvan de kandidaat beoordeeld wordt. Bij ieder criterium is een aantal niveaus beschreven waarop de kandidaat aantoont het criterium te beheersen. Normaal gesproken zijn er een vijftal niveaus opgenomen variërend van niet aangetoond tot een vergaande complexiteit. Iedere beschrijving vormt een ‘rubric’.

2. Theoretisch kader				
Onvoldoende: <5.5	Voldoende: 5.5 -6.5	Ruim voldoende: 6.5 - 7.5	Goed: 7.5 - 8.5	Zeer goed: 9-10
• Theoretisch kader ontbreekt • Geen relatie met de onderzoeksvraag • Onsamenhangend theoretisch kader • Geen duidelijke bronverwijzingen.	• Theoretisch kader is correct gekoppeld aan facetten van de hoofdvraag • De gebruikte bronnen zijn beperkt en voor de hand liggend (maar niet beperkt tot de voorgeschreven studieboeken). • Relevante variabelen, kernbegrippen, uit hoofd- en deelvragen worden uitgewerkt; de koppeling met de adviesvraag en deelvragen is duidelijk	• De gebruikte bronnen zijn voor een deel additioneel aan de voorgeschreven studieboeken (specifiek, internationaal, bredere context). • Bestudeerde literatuur wordt in een samenvatting aan elkaar verbonden. • Relevante variabelen, kernbegrippen uit adviesvraag en deelvragen worden systematisch uitgewerkt en hun relatie is beschreven	In aanvulling op "ruim voldoende" • De gebruikte bronnen zijn voor een belangrijk deel additioneel aan de voorgeschreven studieboeken (specifiek, internationaal, bredere context). • Uit de bronnen wordt een focus aangebracht getrokken die als kader dient voor het onderzoek.	In aanvulling op 'goed' • Er is sprake van een originele en kritische invalshoek. Er is een eigen conceptueel model ontwikkeld op basis van een kritische analyse van bestaande modellen en theorieën
Cijfer	Toelichting			

Dit is een praktische manier van beoordelen, maar ook een die tot discussie leidt. Cohen en Billsberry (2014) leggen in hun artikel wellicht de vinger op de zere plek. In een opiniërend artikel over het gebruik van rubrics voor managementopleidingen stellen zij dat er bij rubrics een identificeerbaar deugdelijk en meetbaar construct (hetgeen je wilt meten) nodig is. Dit ontbreekt meestal. Waarom is de ene manager wel succesvol en de andere niet? Zolang je dat niet weet kun je volgens hen geen zinvolle rubrics opstellen, Zonder goede objectieve uitkomstmaat is het lastig onderzoek te doen. De vraag is waarom zoveel opleidingen rubrics als beoordelingsmodel gebruiken als er geen evidence-based onderzoek is dat het gebruik ervan ondersteunt. De reden daarvoor ligt vermoedelijk in het sterke conceptuele karakter. Cohen en Billsberry (2014) geven aan dat er geen duidelijk meetbaar construct is dat getoetst moet worden. Dat is ten dele waar. Als dat er wel zou zijn zouden wellicht andere toetsvormen ook in aanmerking komen om verschillende vaardigheden te toetsen. Het voordeel van rubrics is dat je werkgevers juist goed kan betrekken in het benoemen van competenties die zijn graag bij hun werknemers zien (Bailey, 2014). Het is relatief eenvoudig gewenst en ongewenst gedrag te beschrijven en daar een beoordeling aan op te hangen. Daarnaast hebben rubrics ook het voordeel dat het voor studenten heel helder is wat ze moeten aantonen tijdens hun tentamen en dat de beoordeling van een toets zich eenvoudig laat vertalen naar concrete feedback (zie bijvoorbeeld Alexander & Praeger, 2009). In feite gaat deze discussie niet over de betrouwbaarheid van de rubrics, maar over de validiteit van het beoordelingsmodel. Meet je met een rubric wat je moet weten? Omdat het onderwerp van dit artikel meer over betrouwbaarheid gaat, laten we deze discussie verder links liggen.

Betrouwbaarheid binnen de psychologie

Een discussie over betrouwbaarheid en validiteit is niet een discussie die exclusief binnen de onderwijskunde gevoerd wordt. Binnen het domein van de psychologie worstel men ook met deze begrippen. Vragenlijsten en interviews moeten gevalideerd worden en de uitkomsten moeten betrouwbaar zijn. Conway et al. (1995) betogen dat psychologen vaak discussiëren over validiteit van interviews, maar dat betrouwbaarheid eerst bekeken moet worden, omdat dat vaak zuiverder gemeten kan worden dan validiteit. Het blijft immers arbitrair of de juiste vragen gesteld worden, de reproduceerbaarheid van de uitkomsten is statistisch uit te rekenen. Het bestuderen van betrouwbaarheid is de beste manier om de

kwaliteit van meetinstrumenten te isoleren. Door meta-analyses naar betrouwbaarheid uit te voeren kan men inzichten vergaren die helpen de validiteit van onderzoeken te verhogen. De meta-analyse over interviews betreffende de arbeidsmark liet zien dat training van interviewers (onderzoekers) tot een hogere betrouwbaarheid leidde en dat deze betrouwbaarheid verder vergroot wordt wanneer interviewers gevraagd werd de verschillende vragen separaat te beoordelen in plaats van met een algemeen holistisch oordeel te komen. Dat laatste - zo tonen verschillende hiervoor besproken artikelen aan - gebeurt in het onderwijs te weinig. En de stelling dat deze holistische benadering tot een lagere betrouwbaarheid leidt wordt dus door Conway et al (1995) onderschreven.

Hoyt (2000) houdt zich bezig met de zogenaamde 'raters bias'. Om te illustreren wat het probleem daarmee is, komt hij met een voorbeeld uit het hoger onderwijs. Namelijk de 'grade inflation': als beoordelaars een C als laagste beoordeling geven en dat als een onvoldoende beschouwen, maar deze C formeel voldoende (geslaagd) betekent, komt het erop neer dat alle leerlingen voldoende scoren en voldoende moet betekenen; gemiddeld of hoger. Statistisch kan natuurlijk niet iedereen gemiddeld of hoger scoren. Het probleem met deze raters bias is dus dat het gemiddelde verschuift; het heeft effect op de varantie en de co-variantie. Het grootste probleem blijft natuurlijk wanneer een deel van de waargenomen variantie veroorzaakt wordt door de variantie in bias tussen de verschillende beoordelaars en niet veroorzaakt wordt door de variantie tussen de te onderzoeken subjecten. Hoyt (2000) onderscheidt twee soorten van bias. De ene is de interpretatie van de uitkomstmaat. In onderwijstermen vertaald betekent het dat sommigen docenten een 7 een hoog cijfer vinden, terwijl anderen het zien als net iets beter dan voldoende. Het andere komt voort uit in hoeverre een examinator meer aspecten laat meewegen in zijn oordeel dan gevraagd. Dit kan schrijfstijl of aantrekkelijkheid zijn. Dit noemt hij dyade specifieke bias. Het probleem is dat deze verschillende typen bias elkaar zowel kunnen opheffen als versterken. Dit leidt tot onzuivere metingen en hij introduceert daarom ook een statistisch bi-variatie model om deze effecten zoveel mogelijk ongedaan te maken.

Betrouwbaarheid en grade inflation

Zoals hierboven beschreven laat Hoyt (2000) zien dat er een probleem met grade inflation is als de grens tussen onvoldoende en voldoende verschuift. Op betrouwbaarheid in de betekenis van reproduceerbaarheid heeft dit natuurlijk geen impact. De maatschappij wil er wel van uit kunnen gaan dat studenten met een diploma over een aantal competenties beschikken waarmee zij zich succesvol op de arbeidsmarkt kunnen voegen. Als de waarde van diploma's ter discussie komt te staan, kan dat ook maatschappelijke consequenties hebben. Dat hebben we in Nederland kunnen ondervinden naar aanleiding van de hbo-affaire. Alvero en al. (2010) hebben een interventie uitgevoerd om deze cijferinflatie tot staan te brengen. Allereerst wijzen zij erop dat niet alleen het verschuiven van de grens tussen voldoende en onvoldoende een probleem is. Er vindt ook een compressie aan de bovenkant plaats. Aangezien je in de VS nooit hoger dan een A kan halen proppen de beoordelingen tussen en C en A verder op. Er is dus steeds minder onderscheid mogelijk tussen gemiddelde, goede en excellente studenten. Op basis van diploma's kunnen werkgevers nu moeilijker onderscheid maken tussen afgestudeerden. Daarnaast is te verwachten dat goede studenten ontmoedigd

worden om hard te werken. Zij zullen immers sneller dezelfde beoordeling krijgen als minder getalenteerde collega-studenten zo betogen de auteurs. De eerste vraag die beantwoord moet worden is of er daadwerkelijk sprake is van cijferinflatie. Dat wordt door de universiteit die onderzocht is inderdaad bevestigd. De cijfers stijgen afgezet tegen de SAT scores significant (0.8p) tussen 1994-95 en 2000-01. Door dit probleem met een deel van de staf bespreekbaar te maken, beoordelingsmodellen onderling te bespreken en te kalibreren bleek deze cijferinflatie in ieder geval tijdelijk tot stilstand gebracht te kunnen worden. Het gedeelte van de staf dat niet meegenomen was in deze maatregel bleef als vanouds hoge cijfers geven.

Conclusie en discussie

Er is weinig evidence based onderzoek te vinden naar de betrouwbaarheid van beoordelingen in het hoger onderwijs omdat de uitkomstmaat die in het meeste onderwijskundig onderzoek wordt gehanteerd de beoordeling zelf is. Er zijn geen gestandaardiseerde toetsen zoals in het primair en secundair onderwijs het geval is. Zelfs bij gestandaardiseerde toetsen kan je vraagtekens zetten over de betrouwbaarheid van de uitkomsten. Langere termijn onderzoeken waarbij studenten na hun studie gevolgd worden met als doel te kijken of hun beoordelingen wel betrouwbaar waren zijn niet gevonden. Wel kan worden vastgesteld dat er redenen zijn om de betrouwbaarheid van beoordelingen (en de uitkomstmaat van veel onderzoeken) in twijfel te trekken. Zonder een causale conclusie te trekken is in enkele onderzoeken aangetoond dat er op het gebied van de betrouwbaarheid van de beoordeling van afstudeerwerk in de zin van reproduceerbaarheid veel voor verbetering vatbaar is. Datzelfde probleem bestaat ook in andere vakgebieden waaruit naar voren komt dat beoordelingen onderhevig zijn aan verschillende vormen van bias. Of rubrics zinvol zijn als instrument is de vraag, maar dat is een validiteitsdiscussie die buiten de vraagstelling van dit artikel valt. Wel zien we dat een fijnmazig beoordelingskader en gesprekken daarover tussen verschillende examinatoren de betrouwbaarheid van het oordeel ten goede komt. Rubrics lijken dus wel een basis te kunnen bieden voor een betrouwbaar oordeel.
Verschillende studies suggereren dat het kan uitmaken of een beoordelaar de beoordeelde kent. Op basis daarvan mag geconcludeerd worden dat het zonder meer zinvol is een onderzoek te doen naar de vraag in hoeverre de betrokkenheid van een examinator bij een afstudeerkandidaat van invloed is op de betrouwbaarheid van het oordeel van die examinator. Bij dat onderzoek moet dan wel een andere uitkomstmaat gezocht worden dan de beoordeling zelf.

Literatuurlijst

Alexander, C., & Praeger, S. (2009). *Smoke gets in your eyes: Using rubrics as a tool for building justice into assessment practices.* Conference paper

Alvaro Q. Barriga , Eric K. Cooper , Mary Ann Gawelek , Kristin Butela & Elizabeth Johnson (2008) Dialogue and Exchange of Information about Grade Inflation can Counteract its Effects, *College Teaching*, 56:4, 201-209

Bailey, J. R., (2014) Crossing the Rubric-con. In: *Journal of Management Education* 38(3) 313–318.

Baume, D., Yorke, M. & Coffey, M. (2004) What is happening when we assess, and how can we use our understanding of this to improve assessment? In: *Assessment & Evaluation in Higher Education*, 29:4, 451-477.

Bloxham, S., Den-Outer, B., Hudson, J. & Price, M. (2015): Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. In: *Assessment & Evaluation in Higher Education*.

Bruijn, J.A. (2011). *Vreemde ogen dwingen. Eindrapport Commissie externe validering examenkwaliteit hoger beroepsonderwijs.* Commissie-Bruijn, HBO-raad.

Childs, R. A., Ram, A. & Xu, Y. (2009). Combining Dual Scaling with Semi-Structured Interviews to Interpret Rating Differences. In: *Practical Assessment, Research & Evaluation* 14 (11).

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. In: *Journal of Applied Psychology, 80,* 565-579.

Cohen, M. & Billsberry, J. (2014). The Use of Marking Rubrics in Management Education: Issues of Deconstruction and Andragogy. In: *Journal of Management Education* Vol. 38(3) 352–358.

Dee, T. S., Dobbie William, J., Brian A., McCrary, J., & Rockoff, J. (2011), Rules and Discretion in the Evaluation of Students and Schools: The Case of the New York Regents Examinations. In: *Colombia Business School Research Paper.*

Dunnewijk, M. (2011). *Rapport van Bevindingen NVAO-Commissie Onderzoek Hogeschool Inholland.* Commissie-Dunnewijk, NVAO.

Hand, L. & Clewes, D. (2000). Marking the Difference: An investigation of the criteria used for assessing undergraduate dissertations in a business school. In: *Assessment & Evaluation in Higher Education*. 25:1, 5-21.

Hoyt, W., T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods* 5 (1) 64-86.

Klaveren, van C. & Wolf, de I. (2013). *Systematic reviews in education research: When do effect studies provide evidence*, working paper 46.

Kleijn, de R.,Mainhard, T., Meijer, P., Pilot, A. & Brekelmans, M. (2012). Master's thesis supervision: relations between perceptions of the supervisor-student relationship, final grade, percieved supervisor contribution to learning and student satisfaction. *Studies in Higher Education* 37 (8) 925-939.

Knight, P., (2002) Summative Assessment in Higher Education: practices in disarray. In: *Studies in Higher Education* 27 (3) 275-286.

Streatmans, G., (2006). *Bekwaam beoordelen en beslissen: beoordelen in competentiegerichte beroepsopleidingen.* Lectorale rede Saxion Hogescholen.

Vleuten, van der C. & Schuwirth, L. (2005). Assessing professional competence: from methods to programmes. *Medical Education*; 39: 309–317.

Webster, F., Pepper, D. & Jenkins, A. (2000). Assessing the undergraduate dissertation. *Assessment & Evaluation in Higher Education* 25 (1) 71-80.