

Alle kanten van de scriptie

De betrouwbaarheid van scriptiebeoordelingen voor de opleiding Toegepaste
Psychologie van Saxion te Deventer

Naam:	Demir Porovic
Studentnummer:	302971
Plaats:	Deventer
Datum:	16-10-2017
Opdrachtgever:	Academie Mens & Arbeid (AMA), opleiding Toegepaste Psychologie (TP), van de hogeschool Saxion te Deventer
1 ^e afstudeerbegeleider:	Jaap Veldhuijzen
2 ^e afstudeerbegeleider:	Rianne Marsman

Voorwoord

Voor u ligt de scriptie 'Alle kanten van de scriptie': De betrouwbaarheid van scriptiebeoordelingen aan de opleiding Toegepaste Psychologie (TP) van Saxion te Deventer. Het onderzoek voor deze scriptie heeft hoofdzakelijk op de Academie Mens & Arbeid (AMA) plaats gevonden. Deze scriptie is het product van mijn afstudeeropdracht voor de opleiding TP aan Saxion te Deventer in opdracht van de AMA.

Het maken van deze scriptie heeft iets langer geduurd dan oorspronkelijk gepland stond. Dit had te maken met persoonlijke omstandigheden waardoor het vaak niet lukte deadlines te halen. Ik wil bij deze mijn afstudeerbegeleiders, Jaap Veldhuijzen en Rianne Marsman, heel erg bedanken voor hun hulp, steun en geduld bij het tot stand komen van deze scriptie. Ik wil deze ruimte ook gebruiken om mijn excuses aan te bieden voor alle onregelmatigheden die deze situatie heeft veroorzaakt.

Binnen de scriptie zullen een aantal termen gebruikt worden die op meerdere manieren geïnterpreteerd kunnen worden en verwarrend kunnen zijn voor personen die niet betrokken zijn binnen het hoger onderwijs of dit onderzoek. Hieronder zal ik kort toelichten wat deze termen betekenen binnen dit onderzoek.

Vaak worden de termen internen en externen gebruikt. Deze termen slaan op het type beoordelaar. Met internen worden alle beoordelaars bedoeld die werkzaam zijn op Saxion en voor de opleiding Toegepaste psychologie (voortaan TP) scripties beoordelen. Met externen worden alle beoordelaars die werkzaam zijn binnen het werkveld van de psychologie bedoeld. De term beoordelingsprotocollen wordt gebruikt voor alle beoordelingsformulieren die worden gebruikt door de opleiding TP. Dit heeft betrekking op zowel het schriftelijke gedeelte als het mondelinge gedeelte. Deze twee gedeeltes slaan weer op de twee onderscheidenlijke onderdelen van de scriptie. Het schriftelijke gedeelte heeft betrekking tot alles wat met de rapportage van de scriptie te maken heeft. Het mondelinge gedeelte heeft betrekking tot alles wat te maken heeft met het eindgesprek. Daarnaast wordt er gesproken over kwalitatieve en kwantitatieve onderdelen. Het kwantitatieve onderdeel heeft betrekking tot al het onderzoek naar de beoordelingsprotocollen. Het kwalitatieve onderdeel heeft betrekking tot het onderzoek naar de opinie van beoordelaars. Het kwalitatieve gedeelte heeft ook kwantitatieve elementen maar wordt door eerdergenoemde reden onder dit onderdeel geschaard.

Het is daarnaast belangrijk om te vermelden dat met het woord scriptie zowel de beoordelingen van het schriftelijke gedeelte als ook het mondelinge gedeelte wordt bedoeld (ook wel de beoordelingsprotocollen genoemd). De reden hiervoor is dat bij beide beoordelingen dezelfde beoordelaars/adviseurs betrokken zijn. Hierdoor kan ook worden bepaald of er een eventuele discrepantie is tussen de interbeoordelaarsbetrouwbaarheid van het schriftelijke gedeelte en die van het mondelinge gedeelte. Als deze delen afzonderlijk vermeld worden dan heeft het alleen betrekking op het genoemde deel. Als enkel de term beoordelaars wordt gebruikt dan wordt hiermee zowel de externen als de internen mee bedoeld.

Nunspeet, Januari 2018

Samenvatting

Achtergrond en doelstelling: Door twijfels binnen het hoger onderwijs over de betrouwbaarheid van beoordelingen in het hoger onderwijs wil de Academie Mens & Arbeid, in deze de opleiding Toegepaste Psychologie, die betrouwbaarheid in kaart brengen. Daarbij gaat het specifiek om de interbeoordelaarsbetrouwbaarheid tussen internen en externen tijdens de (eind)examineringfase. Zodoende is de centrale onderzoeksvraag geformuleerd: wat is de interbeoordelaarsbetrouwbaarheid tussen internen en externen bij het schriftelijke – en mondelinge deel van de (eind)examinering? Door mistanden in het hoger onderwijs is externe validering volgens een rapport van de commissie Leers een steeds belangrijkere rol gaan spelen. Om deze mistanden in de toekomst te voorkomen en een kwalitatieve eindbeoordeling te krijgen is onderzoek nodig. Er zijn verschillende methoden vanuit de literatuur die duiden op factoren die invloed kunnen hebben op de interbeoordelaarsbetrouwbaarheid. Het gebruik van rubrics, trainingen en evaluatiemateriaal. Door het toepassen van verschillende intra klasse correlatiecoëfficiënten kan de interbeoordelaarsbetrouwbaarheid bepaald worden.

Methode: Er werd gebruik gemaakt van bestaande data vanuit de Academie Mens & Arbeid. Alle cijfers waren apart opgeslagen. Deze werden in SPSS ingevoerd en geanalyseerd. In totaal zijn 187 scripties meegenomen in het onderzoek. 107 hiervan waren bruikbaar voor de secundaire data analyse. Daarnaast werd een vragenlijst afgenomen met meerkeuze vragen en een viertal open vragen. Deze vragenlijst werd online afgenomen. De kwantitatieve data werd wederom in SPSS geanalyseerd. De kwalitatieve data is overgenomen in de bijlagen en inductief gecodeerd. De periode waarover het onderzoek gaat betreft het school jaar 2014/2015 en 2015/2016.

Resultaten: Voor bijna elke analyse werd er geen statistisch significante samenhang gevonden. Hoewel er samenhang gevonden is tussen internen is deze niet significant en berust dus op kans. Ook de onvoldoende samenhang tussen internen en externen berust op kans. Wel werd voor het aantal jaar dat een beoordelaar cijfers geeft en de hoogte van de cijfers een negatieve significante samenhang gevonden.

Conclusie: De beoordelingen van scripties door de in- en externe beoordelaars komen tussen de internen in hoge mate overeen. Tussen de internen en externen is er binnen het schriftelijke onderdeel van de scriptie vrijwel altijd onvoldoende correlatie. Binnen het mondelinge gedeelte is de correlatie altijd voldoende of goed. Hier dient bij opgemerkt te worden dat er ook een negatieve samenhang blijkt tussen het gegeven cijfer en de frequentie en duur van de gegeven beoordelingen omdat zowel interne als externe beoordelaars, naarmate ze langer en frequenter beoordelen, tot een lager cijfer neigen.

Aanbevelingen: De rol van de externe beoordelaar kan onder de loep genomen worden en op basis van de resultaten een meer wegende rol gegeven worden. Daarnaast moet vervolg onderzoek gedetailleerde kwaliteitsveranderingen die vanuit de beoordelaars komen achterhalen. Het trainen van beoordelaars zou kwaliteit kunnen waarborgen.

Inhoudsopgave

Hoofdstuk 1 – Inleiding	3
1.1 Aanleiding en opdrachtgever	3
1.2 Opdrachtgever	4
1.3 Onderzoeksvraag en deelvragen	4
1.4 Doelstelling	5
Hoofdstuk 2 – Theoretisch kader	6
2.1 Betrouwbaarheid	6
2.2 Interbeoordelaarsbetrouwbaarheid (IBB)	6
2.3 Interbeoordelaarsbetrouwbaarheid in het onderwijs	8
2.4 Conceptueel model	10
2.5 Aanzet onderzoek	10
Hoofdstuk 3 – Onderzoeksmethoden	11
3.1 Onderzoeksmethoden	11
3.2 Onderzoeksdoelgroep	11
3.3 Onderzoeksinstrumenten	11
3.4 Procedures	12
3.5 Analyses	13
3.6 Hypothesen	14
Hoofdstuk 4 – Onderzoeksresultaten	15
4.1 Uitvoering	15
4.2 Respons	15
4.3 Onderzoeksresultaten	16
4.3.1 Deelvragen over de interbeoordelaarsbetrouwbaarheid	16
4.3.2 Deelvragen over de vragenlijst	17
4.3.3 Meninge over opzet van de scriptie beoordelingen	17
Hoofdstuk 5 – Conclusies, discussie en aanbevelingen	21
5.1 Conclusies	21
5.1.1 Betrouwbaarheid afzonderlijke hoofdstukken van het schriftelijke gedeelte	21
5.1.2 Betrouwbaarheid eindcijfers van het schriftelijke gedeelte	21
5.1.3 Betrouwbaarheid eindcijfers van het mondelinge gedeelte	22
5.1.4 Tevredenheid onder de beoordelaars over de kwaliteit van de protocollen	22
5.1.5. Samenhang jaren ervaring en hoogte cijfers	22
5.1.6 De onderzoeksvraag	23

5.1.7. De meningen van beoordelaars	23
5.2 Discussie	24
5.3 Aanbevelingen	25
5.3.1 Plan van aanpak.....	27
Literatuurlijst	28
Bijlagen	I
Bijlage 1 Antwoorden op vragenlijst schriftelijke gedeelte.....	II
Bijlage 2 De vragenlijst	X
Bijlage 3 Het beoordelingsprotocol.....	XIII
Bijlage 4 Beoordelingsprotocol mondelinge gedeelte (het eindgesprek).....	XIV
Beoordelingsformulier eindgesprek TP	XIV

Hoofdstuk 1 – Inleiding

In dit hoofdstuk zullen de aanleiding en de opdrachtgever van het onderzoek aan bod komen. Vervolgens wordt ingegaan op de centrale onderzoeksvraag, deelvragen en doelstelling van het onderzoek.

1.1 Aanleiding en opdrachtgever

In 2010 ontstond er maatschappelijk en politiek commotie over de kwaliteit van diploma's aan de hogeschool INHolland voor de opleiding Media en Entertainment Management (Sahadat, 2010). Deze landelijke commotie heeft destijds gezorgd voor het aanstellen van een commissie die onderzoek doet naar de diplomering van langstudeerders bij de Hogeschool INHolland (Commissie Leers, 2010). Een aantal maanden later heeft de Inspectie van het onderwijs (2011), in haar rapport alternatieve afstudeertrajecten en de bewaking van het eindniveau in het hoger onderwijs, vermeld dat de Nederlands-Vlaamse Accreditatieorganisatie (NVAO) vaststelt dat een groot deel van de afstudeerwerkstukken niet van hbo-kwaliteit zijn. Grofweg een jaar later verscheen nog een rapport van de Inspectie van het onderwijs (2012) om de stappen die hogescholen hebben gezet richting het wegwerken van de tekortkomingen, die werden genoemd in het rapport uit 2011, in kaart te brengen.

Hoewel deze ontwikkelingen hoofdzakelijk over afstudeertrajecten die in bovengenoemde artikelen genoemd worden gaan, kan het belangrijk zijn om aan de hand hiervan te bepalen welke gevolgen deze ontwikkelingen voor het gehele hbo hebben gehad. Zo is in 2012 door een door de hbo-raad samengestelde commissie (beter bekend als de commissie Bruijn) het rapport 'vreemde ogen dwingen' uitgebracht waarin een van de oplossingen voor het ondervangen van een eventueel kwaliteitsgebrek van afstudeerwerkstukken gezocht wordt bij het, waar mogelijk, betrekken van externen binnen het afstudeertraject (Commissie externe validering examenkwaliteit hoger beroepsonderwijs, 2012). De adviezen die deze commissie geeft zijn hbo-breed en streven naar een instelling overstijgende toetsing van onder andere de afstudeerwerkstukken. Om dit te bewerkstelligen zal volgens commissie Bruijn (2012) een protocol opgesteld moeten worden. De aanbevelingen die de commissie Bruijn geeft hebben betrekking op meerdere aspecten binnen de toetsing en niet enkel de scriptie.

Binnen de opleiding Toegepaste Psychologie (TP) wordt gewerkt met drie beoordelaars. Twee beoordelaars komen van binnen de opleiding (internen). Zij gelden tevens als de afstudeerbegeleiders van de betreffende student. Daarnaast wordt er iemand van buiten de onderwijsinstelling (externen) ingeschakeld om scripties te beoordelen. Deze beoordelaar is werkzaam in het werkveld van TP. Binnen de opleiding gold voorheen dat de externen de rol van beoordelaar hadden en dus een bindend oordeel konden geven over het cijfer van de scriptie. Sinds het schooljaar 2016-2017 worden externen ingezet als adviseurs en hebben daardoor sindsdien logischerwijs een adviserende rol. In de huidige vorm beoordelen twee interne beoordelaars de scripties en is er een externe adviseur. Het is echter niet duidelijk in welke mate deze externen in hun beoordeling overeenkomen met de interne beoordelaars. Zoals in het rapport 'vreemde ogen dwingen' (2012) wordt aangegeven is het betrekken van externen mogelijk belangrijk voor het waarborgen van een objectieve beoordeling. Binnen dit onderzoek zal worden gezocht naar de mate van samenhang tussen beoordelingen van het schriftelijke gedeelte (alles wat te maken heeft met

het rapport) en het mondelinge gedeelte (alles wat te maken heeft met het eindgesprek). Voor het gemak worden beide onderdelen ondergebracht in de term “scriptie”.

Dit onderzoek zal hoofdzakelijk de mate van kwantitatieve overeenkomsten tussen internen enerzijds en tussen internen en externen anderzijds bij zowel het schriftelijke gedeelte als bij het mondelinge gedeelte vast moeten stellen. Het gaat hierbij eerst om de mate van overeenkomst tussen internen enerzijds en tussen internen en externen anderzijds bij beide genoemde gedeeltes.

1.2 Opdrachtgever

Dit onderzoek wordt in opdracht van het management van de Academie Mens & Arbeid (voortaan AMA), in de persoon van dhr. de Vries (teamleider van de opleiding Toegepaste Psychologie) uitgevoerd. Mede door enkele misstanden in het onderwijs wil de opdrachtgever weten hoe het gesteld is op de AMA voor de opleiding TP. Deze misstanden worden verderop in deze paragraaf uit een gezet.

De AMA wil naar aanleiding van het rapport ‘vreemde ogen dwingen’ (Commissie externe validering examenkwaliteit hoger beroepsonderwijs, 2012) weten hoe betrouwbaar de beoordelingen van scripties tussen de interne beoordelaars zijn en hoe deze beoordelingen zich verhouden tot de beoordelingen van de externen. De AMA is vooral geïnteresseerd in het kwantitatieve aspect van dit onderzoek. Hieronder valt de interbeoordelaarsbetrouwbaarheid van de gegeven beoordelingen. Dit onderzoek zal zich daarom ook voornamelijk richten op dit kwantitatieve aspect. Tevens wil de AMA graag input van de betrokken beoordelaars met betrekking tot hun mening over de beoordelingsprotocollen en procedures. Daartoe is, in overleg met de opdrachtgever, besloten om een vragenlijst op te stellen die ten doel heeft om de meningen van de beoordelaars ten aanzien van beoordelingsprotocollen en procedures te inventariseren. In hoofdstuk drie volgt een uitgebreidere uitleg over de vragenlijst en hoe deze toegepast zal worden.

1.3 Onderzoeksvraag en deelvragen

Op basis van bovenstaande kan de volgende onderzoeksvraag geformuleerd worden:

Wat is de interbeoordelaarsbetrouwbaarheid bij het beoordelen van scripties binnen de AMA-TP van Saxion te Deventer over de schooljaren 2014-2015 en 2015-2016?

De deelvragen die daaruit voortvloeien zijn als volgt:

- 1 Wat is de interbeoordelaarsbetrouwbaarheid tussen de drie beoordelaars over de cijfers van alle vijf afzonderlijke hoofdstukken en het totale cijfer van het schriftelijke gedeelte?
- 2 Wat is de interbeoordelaarsbetrouwbaarheid tussen de drie beoordelaars over de cijfers van alle drie afzonderlijke onderdelen en het totale cijfer van het mondelinge gedeelte?
- 3 In hoeverre zijn de beoordelaars tevreden over de kwaliteit van de beoordelingsprotocollen?
- 4 Wat vinden beoordelaars en adviseurs van de opzet van het schriftelijke gedeelte en het mondelinge gedeeltes wat zouden ze hier eventueel aan willen veranderen?
- 5 Wat is de samenhang tussen het aantal jaar dat de beoordelaars en adviseurs scripties beoordelen en het hoogte van het cijfer voor het schriftelijke gedeelte en het mondelinge gedeelte?

1.4 Doelstelling

Het voornaamste doel van dit onderzoek is om de AMA inzicht te geven in de mate van overeenstemming in beoordeling van de scripties tussen de interne en externe beoordelaars/adviseurs. De resultaten van dit onderzoek en de conclusies die daaruit worden getrokken, alsmede de aanbevelingen die worden gegeven, zullen bijdragen om dit inzicht te vergroten. Aan de hand van de interbeoordelaarsbetrouwbaarheid kan de AMA zien wat de betrouwbaarheid van de beoordelingen is tussen de interne beoordelaars en hoe de betrouwbaarheid van de beoordelingen van elk van de interne beoordelaars zich verhoudt tot de beoordelingen van de externe adviseurs. Dit kan invloed hebben op de rollen die de externe adviseurs toegekend kunnen krijgen vanuit de examinering van de opleiding TP of de AMA. Het gaat hier om een bindende (ook wel gelijkwaardige) beoordeling of een adviserende beoordeling. Dit kan van belang zijn als de externe een ander oordeel heeft dan de internen. Daarnaast kan de relevantie van het aantal beoordelaars onder de loep worden genomen. Met de uitkomsten van vragenlijst kan de AMA-TP eventueel aanpassingen maken aan de beoordelingsformulieren of algehele opzet van de beoordeling van scripties. De uitkomsten van het onderzoek kunnen leiden tot een aantal veranderingen binnen de procedure van het beoordeling van de scriptie alsmede de materialen die daarvoor gebruikt worden. De input van de beoordelaars is hierbij van groot belang zeker als er tussen de beoordelaars overeenkomende thema's worden genoemd. Uiteindelijk is het streven van de AMA om een kwalitatief betere beoordeling van de scriptie tot gevolg te hebben. Dit kan bewerkstelligd worden door het gebruik van statistische analyses. Hierdoor kan de kwaliteitsverbetering onderbouwd worden door zowel het advies van de betrokken beoordelaars als het gebruik van statistiek.

Hoofdstuk 2 – Theoretisch kader

In dit hoofdstuk zal een theoretisch kader gevormd worden door literatuur die relevant is voor het onderwerp van het onderzoek. Aan de hand van dit theoretische kader wordt een conceptueel model opgesteld.

2.1 Betrouwbaarheid

Binnen het onderzoek naar de interbeoordelaarsbetrouwbaarheid (voortaan afgekort als IBB) van de scriptie beoordelingen wordt gebruik gemaakt van betrouwbaarheid. Het is belangrijk om voor dit onderzoek een duidelijke definitie van betrouwbaarheid te hebben omdat betrouwbaarheid binnen verschillende vakgebieden net een andere betekenis kan hebben door de manier waarop het gebruikt wordt (Carmines & Zeller, 1979). Daarom is voor dit onderzoek gekozen voor de definitie van Trochim omdat deze gebruikt wordt binnen de context van sociale wetenschappen (2006): “de algehele consistentie van een waarde of score”. Trochim bedoelt hiermee betrouwbaarheid die binnen de psychometrie en statistiek wordt gebruikt. Een testscore bestaat uit een betrouwbare score en een foutmarge. Deze foutmarge kan ook een onbedoelde component inhouden waardoor niet wordt gemeten wat beoogd werd te meten (Evers, Lucassen, Meijer & Sijtsma, 2010). Trochim (2006) onderscheidt een viertal vormen van betrouwbaarheid. De IBB is een van de manieren om betrouwbaarheid te meten en meet dus de overeenstemming tussen twee of meerdere beoordelaars. Daarnaast heeft Trochim (2006) het over een interne consistentie betrouwbaarheid van een test die de betrouwbaarheid van items binnen eenzelfde test meten. Een derde vorm van betrouwbaarheid is de test-hertest betrouwbaarheid. Hierbij worden dezelfde methoden of beoordelaar gebruikt bij twee verschillende meetmomenten om de betrouwbaarheid vast te stellen. Hij besluit met de inter-methode betrouwbaarheid. Bij deze methode wordt gebruik gemaakt van verschillende tests en onderzoeksmethoden om te vast te stellen of de scores hetzelfde zijn. Voor het onderzoek is de IBB en de interne consistentie van belang omdat deze twee methoden om betrouwbaarheid te meten van toepassing zijn. In de volgende paragrafen wordt de IBB nader toegelicht.

2.2 Interbeoordelaarsbetrouwbaarheid (IBB)

De rode draad binnen dit onderzoek is de IBB die, zoals gesteld is, een vorm van betrouwbaarheid is waarmee kan worden bepaald in hoeverre twee beoordelaars overeenkomen in hun beoordeling en waarbij unieke analyses kunnen worden uitgevoerd om tot een betrouwbaarheidscoëfficiënt te komen (Kish, 1965). IBB kan op een aantal manieren worden vastgesteld.

Een aantal methoden die worden gebruikt zullen kort worden beschreven. Het is belangrijk om hier in het achterhoofd te houden dat bepaalde methodes het beste bij een bepaald meetniveau kunnen worden gebruikt. Deze wordt per methode toegelicht. Scott's pi wordt gebruikt voor het meten van betrouwbaarheid binnen menselijke communicatie en houdt rekening met toeval binnen de overeenkomst tussen beoordelaars. Het nadeel van de Scott's pi is dat het nominale data meet en vaak alleen tussen twee beoordelaars (Scott, 1955). Deze methode is derhalve niet geschikt voor het onderzoek.

Fleiss' kappa geeft voor dit onderzoek hetzelfde probleem omdat deze meting ook uitgaat van nominale data (Fleiss & Cohen, 1973). Het voordeel van Fleiss' kappa is echter dat deze een groter aantal beoordelaars in de meting kan meenemen (Shrout & Fleiss, 1979). Een andere methode is de Cohen's kappa (k). De Kappa is een robuuste manier van overeenkomst berekenen en neemt ook in overweging dat overeenstemming binnen beoordelingen als gevolg van toeval kunnen plaats vinden

(Hallgren, 2012). Er bestaat controverse omtrent de kappa omdat het volgens sommige onderzoekers makkelijker is om te kijken naar een eventueel verschil dan te kijken naar de samenhang (Pontius & Millones, 2011). De Kappa onder een waarde van 0.6 wordt gezien als onvoldoende terwijl een score van boven de 0.80 als goed wordt gezien (Evers, Lucassen, Meijer & Sijtsma, 2010). De kappa wordt gebruikt om de betrouwbaarheid van de een overeenstemming op nominaal meetniveau vast te stellen.

Als laatste methode voor het meten van IBB is het intraklassecorrelatiecoëfficiënt (Koch, 1982). Volgens Rhebergen et al. (2011) is het intraklassecorrelatiecoëfficiënt (voortaan afgekort als ICC) de meest gebruikte en geschikte statistische analyse om de IBB op het intervalmeetniveau te meten. Het grote voordeel van deze methode is dat er geen gepaarde observaties zijn maar dat ze als groepen gestructureerd worden. Iets wat voor dit onderzoek de ideale methode voor het meten van de IBB lijkt te zijn. Het meetniveau zal daarnaast moeten bepalen welke coëfficiënt het beste gebruikt kan worden om de interbeoordelaarsbetrouwbaarheid te meten. De coëfficiënten voor de kappa en ICC worden op dezelfde manier geïnterpreteerd zoals ook beschreven door Landis en Koch (1977) in tabel 1.

Tabel 1

Richtlijnen voor de kwalitatieve interpretatie van Cohen's kappa en de intraklassecorrelatiecoëfficiënt

Uitkomst correlatiecoëfficiënt	Kwalitatieve Interpretatie
(k)<0	Slecht
0-0,20	Gering
0,21 - 0,40	Matig
0,41 - 0,60	Redelijk
0,61 - 0,80	Goed
0,81 - 1,00	Bijna perfect

Noot. De interpretaties in de tabel zijn gebruikt zoals beschreven door Landis en Koch (1977).

Deze interpretatie is echter niet de enige omdat er geen eenduidige interpretatie van de bovengenoemde coëfficiënten is. Zo wordt binnen de commissie testaangelegenheden Nederland (COTAN) beoordelingssysteem voor de kwaliteit van tests (2010) een coëfficiënt onder de 0,60 geïnterpreteerd als onvoldoende, de score tussen 0,60 en 0,80 als voldoende en een score van 0,80 of hoger als goed. In tabel 2 is hier een weergave van te zien. Deze interpretatie heeft minder kwalitatieve interpretaties en daarmee trekt het ook een hardere lijn voor wat wel als een betrouwbare meting van overeenstemming is en wat niet.

Tabel 2

Richtlijnen voor de kwalitatieve interpretatie van Cohen's kappa en de intraklassecorrelatiecoëfficiënt volgens Evers, Lucassen, Meijer & Sijtsma (2010)

Uitkomst correlatiecoëfficiënt	Kwalitatieve Interpretatie
< 0,60	Onvoldoende
0,60 – 0.80	Voldoende
0,80 >	Goed

Een nog heel andere interpretatie voor beide coëfficiënten is die van Cicchetti (1994). Zijn kwalitatieve interpretaties worden weergegeven in tabel 3. De verschillende kwalitatieve interpretaties maken het moeilijk om een eenduidig antwoord te kunnen krijgen bij de uitkomsten van een betrouwbaarheidsonderzoek tussen beoordelaars.

Tabel 3

Richtlijnen voor de kwalitatieve interpretatie van Cohen's kappa en de intraklassecorrelatiecoëfficiënt volgens Cicchetti (1994)

Uitkomst correlatiecoëfficiënt	Kwalitatieve Interpretatie
< 0,4	Slecht
0,40 - 0,59	Redelijk
0,60 - 0,74	Goed
0,75 – 1,00	Uitstekend

De COTAN beoordelingssysteem voor de kwaliteit van tests (2010) stelt enkele voorwaarden aan het succesvol uitvoeren van een betrouwbaarheidsmeting met ICC. Een belangrijke voorwaarde is dat de kennis over de toepassing van de methode bij de beoordelaars bekend is zodat er geen verschillen ontstaan in de beoordelingen door verkeerde toepassing van de methode. Om een voldoende te kunnen geven moeten de door de methode gedefinieerde procedures en stappen gevolgd én constant blijven. Zonder deze voorwaarden mag geen voldoende als kwalitatieve interpretatie, zoals in tabel 2 beschreven, worden toegekend. Om uit de verschillende interpretaties een keus te kunnen maken is gekeken naar wat de meest relevante interpretatie voor dit onderzoek is. Doordat de COTAN beoordelingen binnen Nederland veel gebruikt worden en aansluiten op de beroepscode voor psychologen is dit de meest voor de hand liggende keuze (NIP, 2010).

2.3 Interbeoordelaarsbetrouwbaarheid in het onderwijs

Interbeoordelaarsbetrouwbaarheid wordt ook in het onderwijs gebruikt. Zo wordt deze gebruikt om de betrouwbaarheid van de beoordeling van open vragen in het staatsexamen te bepalen (de Jong, 1981). De normen uit het COTAN beoordelingssysteem worden vaak gebruikt binnen het hoger onderwijs om de IBB te interpreteren (tabel 2). Bij het kiezen van de beoordelaars beschrijven Evers et al (2010) dat er goed moet worden gekeken naar hoe de beoordelaars gekozen worden. Dat wil zeggen dat bij het gebruik van de IBB de beoordelaars op dezelfde manier de methode moeten toepassen volgens gestelde standaarden. Dit geldt voor zowel binnen als buiten het onderwijs. Binnen het onderwijs kunnen trainingen daarvoor gebruikt worden. De beoordelaars zullen dan getraind worden om de beoordelingsmethode op dezelfde manier te kunnen toepassen.

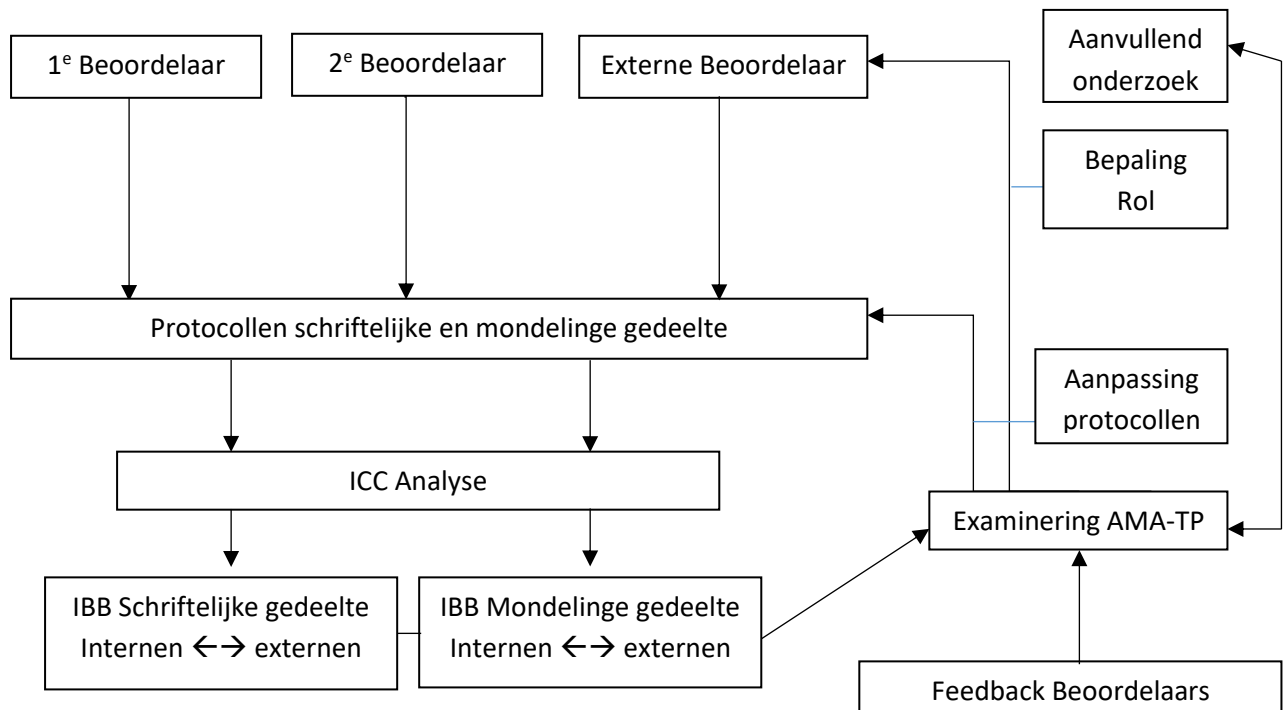
Om een eenduidige methode te kunnen gebruiken en de beoordelaars meer aanknopingspunten te geven om goed kennis te hebben van wat ze precies beoordelen kan gebruikt maken van rubrics (Andrade, 2000). De eenduidigheid van rubrics maakt het een zeer bruikbaar instrument (Jonsson, 2014). Rubrics is het gebruik van oplopende schalen met duidelijk beschreven criteria die aangeven waaraan voldaan moet worden (Brookhart & Chen, 2015). Deze criteria kunnen zowel holistisch als analytisch zijn. Holistisch houdt in dat het vaak algemener is en daarmee subjectiever. Vaak wordt dan vanaf een afstand gekeken naar de criteria. De analytische criteria zijn meer gericht op componenten van een prestatie en dus veel concreter (Jonsson & Svingby, 2007). Rubrics kunnen de kwaliteit van het toetsen binnen de scriptie verhogen en daarmee ook de kwaliteit van het onderwijs (Joosten- ten Brinke, 2011). De Nederlands-Vlaamse Accreditatieorganisatie (2014) stelt dat rubrics aan de kwaliteitseisen, transparantie, validiteit en betrouwbaarheid van toetsen kunnen bijdragen. Bij een duidelijke formulering van rubrics kunnen deze leiden tot een verbetering van studiesucces (Bruijns & Kok, 2013). Een kanttekening die bij het gebruik van rubrics binnen het onderwijs geplaatst

dient te worden is dat de IBB vaak fluctueert maar kan hier ook bevorderend voor werken (Jonsson & Svingby, 2007). Onderzoek naar een reden van eerdergenoemde fluctuaties is zeer schaars maar door de waargenomen tevredenheid van studenten over het gebruik van rubrics en de uit eerdergenoemde onderzoeken aantoonbare verbeteringen in studiesucces, kan gesteld worden dat rubrics wel degelijk een positieve bijdrage leveren binnen het onderwijs. Hierdoor kan het gebruik van rubrics wel degelijk een positieve invloed hebben op IBB (Westenberg, 2016). Het kan belangrijk zijn om ook de tevredenheid van de beoordelaars in het onderwijs te meten omdat dit construct een mogelijk kwaliteitsoordeel weerspiegelt over het beoordelingsmateriaal zoals dit ook in het onderzoek van Bensing (1991) naar voren kwam.

Uit het buitenland zijn enkele studies die voorbeelden tonen van de toepassing van IBB binnen (hoger) onderwijs. Deze bevinden zouden in Nederland voor scriptie beoordeling kunnen worden gebruikt. Onderzoek naar toepassing binnen Nederland is wel sterk aangeraden. In Australië is binnen het geneeskundig onderwijs voor een evaluatie van de studenten binnen dit onderwijs onderzoek geweest naar de interbeoordelaarsbetrouwbaarheid van evaluatie materiaal (genaamd SPEF-R) waaruit bleek dat een goed ontwikkeld evaluatie formulier goede resultaten gaf voor de IBB ongeacht de achtergrond of het aantal jaar aan ervaring van de beoordelaars (Rodger et al. 2014). In de Verenigde Staten van Amerika is ook een dergelijk onderzoek geweest binnen het hoger onderwijs om de consistentie van beoordelingen te bepalen die gebruikt worden voor accreditatie en om te kunnen bepalen of een training nodig is voor docenten om de interbeoordelaarsbetrouwbaarheid te kunnen verhogen. De training had weinig tot geen invloed op de IBB (Pufpaff, Clarke & Jones, 2015). Hieruit kan afgeleid worden dat een Interbeoordelaarsbetrouwbaarheidsonderzoek kan helpen bij het bepalen of een evaluatie-instrument wordt gebruikt of niet. De mogelijkheid om deze instrumenten aan te passen bestaat vanzelfsprekend ook. Daarnaast kan een dergelijk onderzoek ook leiden tot het veranderen van beleid ten opzichte van beoordelen. Zo kunnen beoordelaars getraind worden in het beoordelen. In het onderzoek van Pufpaff et al. (2015) bleek echter dat er geen significant verschil in verbetering van de IBB na de training was aangetroffen.

2.4 Conceptueel model

Op basis van bovenstaande literatuur en onderzoeksopzet kan het onderstaande conceptuele model gevormd worden.



Figuur 1. Het conceptuele model om de interbeoordelaarsbetrouwbaarheid van scripties te meten en het beloop van het gevolg van de uitkomsten daarvan.

2.5 Aanzet onderzoek

Dit onderzoek wordt uitgevoerd met de bedoeling om het beleid ten opzichte van de beoordelingen van scripties aan de Academie Mens & Arbeid te evalueren. Daarnaast zal het onderzoek een vervolgonderzoek kunnen inleiden wat leidt tot een kwalitatieve verbetering van het bovengenoemd beleid. In het rapport *Vreemde Ogen Dwingen* (2012) wordt gesuggereerd dat externe beoordelingen moeten helpen de beoordelingen betrouwbaar en transparant te houden. Daarbuiten wordt ook gekeken naar een mogelijke beleidsverandering op basis van het advies dat het rapport geeft. Een voorbeeld van dergelijke beleidsverandering zal de rolaanduiding van de externe beoordelaars kunnen omvatten. Dat wil in dit geval zeggen dat externe beoordelaars van het geven van een adviserend oordeel naar het geven van een bindend oordeel kunnen gaan. Vanuit dit onderzoek zal moeten blijken welke rol het beste bij de externen zou kunnen passen op basis van de data die beschikbaar is.

Hoofdstuk 3 – Onderzoeksmethoden

In dit hoofdstuk worden de gebruikte methoden van het onderzoek in een aantal delen uiteengezet. Deze delen bestaan uit de onderzoeksmethode, de onderzoeksdoelgroep, de gebruikte instrumenten, de procedures en de analyses.

3.1 Onderzoeksmethoden

Dit onderzoek bestaat uit een kwantitatief als een kwalitatief onderdeel. Het kwantitatieve onderdeel bestaat uit een statistische data analyse uitgevoerd over beoordelingen van alle 3 beoordelaars ten aanzien van de scripties van de opleiding TP binnen Saxion. Deze data is beschikbaar gesteld door de AMA-TP. De analyses worden eerst per hoofdstuk en vervolgens over het eindcijfer uitgevoerd. Dit wordt gedaan voor zowel het schriftelijke gedeelte als het mondelinge gedeelte. Vanwege de al beschikbare data kan hier worden gesproken over een secundaire kwantitatieve data analyse.

Het kwalitatieve onderdeel bestaat uit een vragenlijst (survey) die bij zowel interne als externe beoordelaars wordt afgenomen. Over het tot stand komen van deze vragenlijst wordt in paragraaf 3.3 dieper in gegaan. Hoewel de vragenlijst een mix is van kwalitatieve en kwantitatieve vragen zal, om duidelijkheid te houden binnen het onderzoek, de vragenlijst als kwalitatief onderdeel worden aangeduid omdat hierin kwalitatieve elementen aanwezig zijn. Deze vragenlijst bevat vragen waar op een Likertschaal maar ook op open-vragenantwoord wordt gegeven. Open vragen bieden meer vrijheid voor de respondenten om aan te geven waar ze eventuele verandering willen zien. Gesloten vragen richten zich meer op een keuze die de onderzoeker van de respondent verlangt (Jansen, Joostens & Kemper, 2004). Voor de Likertschaal is gekozen voor een schaal met vijf antwoordmogelijkheden omdat deze volgens Kim (2010) tezamen met een schaal met zeven antwoordmogelijkheden het meest betrouwbaar is. De gebruikte schaal gebruikt punten van helemaal oneens naar helemaal eens. In bijlage twee is de gehele vragenlijst te zien. Deze (gestructureerde) vragenlijst zal de meningen van internen en externen over de huidige beoordelingsprotocollen in kaart brengen. Het toevoegen van deze vragenlijst biedt gelegenheid voor beoordelaars/adviseurs om hun mening te geven over hun tevredenheid over de beoordelingsprotocollen.

3.2 Onderzoeksdoelgroep

Zoals eerder aangegeven staat binnen dit onderzoek de beoordelaar en adviseur centraal. De interne beoordelaars zijn docenten aan de hogeschool Saxion. De externen zijn werkzaam buiten Saxion en zijn enkel betrokken bij het beoordelen.

Alle interne en externe beoordelaars/adviseurs zullen gevraagd worden deel te nemen aan de enquête. In totaal zal de vragenlijst naar 58 personen verstuurd worden. Binnen deze groep zijn er 33 interne beoordelaars en 25 externe adviseurs.

3.3 Onderzoeksinstrumenten

Zoals eerder benoemd bevat dit onderzoek een kwalitatief en een kwantitatief onderdeel. Voor het kwantitatieve gedeelte (IBB) zijn de ingevulde beoordelingsprotocollen van de AMA-TP voor scriptiebeoordelingen en het programma Statistical Package for the Social Sciences ook wel SPSS 22 (IBM, 2013) nodig.

Voor het kwalitatieve onderdeel is, om de opinies van de beoordelaars en adviseurs over de beoordelingsprotocollen vast te leggen, een vragenlijst ontwikkeld. Deze vragenlijst is kort en gaat

over twee onderdelen van de scriptie. Deze twee onderdelen zijn het schriftelijke gedeelte en het mondelinge gedeelte. Voor beide onderdelen zijn bijna dezelfde vragen gesteld. De vragenlijst bevat vier demografische vragen: geslacht, leeftijd, type beoordelaar/adviseur en het aantal jaren dat beoordeeld wordt. De vragenlijst is per onderdeel opgebouwd uit vijf stellingen met twee open vragen om de tevredenheid van beoordelaars over zowel het schriftelijke als het mondelinge gedeelte te meten. Daarnaast wordt de respondenten gevraagd om per onderdeel (schriftelijk en mondeling) een cijfer te geven.

De vragen binnen de twee onderdelen (schriftelijk en mondeling) gaan over de opinies van de beoordelaars en adviseurs. Hen wordt gevraagd wat zij vinden van de kwaliteit, eenduidigheid en opzet van de beoordelingen. De tevredenheid over de kwaliteit van de beoordelingsprotocollen staat hierin centraal. Daarnaast zijn de twee open vragen uit de vragenlijst uit bijlage twee erop gericht om in kaart te brengen welke eventuele veranderingen de internen en externen zouden willen doorvoeren in de procedure van het beoordelen en de materialen die hiervoor gebruikt worden.

De vragen uit de bovengenoemde vragenlijst zijn geformuleerd op basis van de informatie die de opdrachtgever AMA-TP wilde verkrijgen. De geformuleerde vragen zijn vervolgens voorgelegd aan zowel de opdrachtgever als een vijftal experts van de AMA waaronder docenten. Dit om de vragen enige inhoudelijke validatie te geven. Na overleg volgde goedkeuring.

3.4 Procedures

Voor dit onderzoek wordt gebruik gemaakt van de cijfers die beoordelaars/adviseurs aan de studenten hebben gegeven. Deze cijfers zijn door de AMA digitaal opgeslagen. De beoordelingen van het schriftelijke gedeelte worden door de beoordelaars en adviseurs digitaal ingevuld in een Excel-bestand. Deze worden vervolgens per student in een digitale map bewaard. De beoordelingsformulieren voor het mondelinge gedeelte worden schriftelijk ingevuld en vervolgens gescand en in de map van de desbetreffende student geplaatst. De mappen van de studenten worden ingedeeld in de afstudeerperiode waarin ze afstuderen. Voor dit onderzoek is gebruik gemaakt van scripties die binnen de periode van november 2014 tot en met het afstudeermoment januari 2016 vallen.

Om alle data te kunnen analyseren moet het ingevoerd worden in versie 22 van SPSS (IBM, 2013). Hierin werden de beoordelingen gecodeerd zodat deze terug te vinden zijn in het geval dat dit nodig is. Om de ingevoerde cijfers terug te kunnen herleiden naar de beoordelingsformulieren is gebruik gemaakt van de studentnummers. Deze koppeling zal binnen dit rapport niet gebruikt worden om de privacy van studenten en beoordelaars/adviseurs te waarborgen. Daarna werden alle cijfers van de afzonderlijke hoofdstukken alsmede de eindcijfers van de scripties ingevoerd. Hetzelfde werd gedaan voor de cijfers van het mondelinge gedeelte. Door de cijfers van elk hoofdstuk te analyseren kan ook worden vastgesteld of er binnen de hoofdstukken een overeenstemming is in de IBB.

De 58 beoordelaars en adviseurs zullen allen een e-mail ontvangen waarin ze worden aangespoord om deel te nemen aan de vragenlijst. Hierin wordt ook de nadruk gelegd op het feit dat deelname betekent dat de respondent bijdraagt aan het verbeteren van het hbo-onderwijs in het bijzonder binnen de AMA. Daarnaast is de vragenlijst relatief kort en daardoor kan de drempel om de vragenlijst te maken ook lager zijn (Janssen, Joostens & Kemper, 2004). Door de vragenlijst digitaal af te nemen wordt het mogelijk om in korte tijd veel meer respondenten te ondervragen omdat de vragenlijst tegelijkertijd kan worden ingevuld door de respondenten. De vragenlijst blijft twee weken online staan. De vragenlijst is online uitgevoerd op de website: www.enquetesmaken.nl. Deze

website heeft een samenwerking met enkele hogescholen en universiteiten waardoor gratis gebruik kan worden gemaakt van veel diensten die deze website biedt. De respondenten krijgen halverwege de periode dat de vragenlijst online is een herinnering om de vragenlijst in te vullen. Om de beoordelaars en adviseurs verder te stimuleren is de mail verstuurd door de AMA. De data zal vanaf de website waar de vragenlijst op staat geëxporteerd worden naar Microsoft Excel. Vanuit de het Excel bestand zal de ingevulde tekst (de antwoorden) van de twee open vragen in bijlage één volledig te vinden zijn. Onder de tekst staan de codes die hieraan zijn gegeven. Deze data is inductief gecodeerd (Verhoeven, 2010). Vervolgens worden de thema's gekozen die het meest voorkomen. Dit is gedaan omdat thema's die vaak naar voren komen door meerdere beoordelaars van belang worden geacht. Dit belang is echter subjectief en wil niet zeggen dat veelvoorkomende thema's per se ook het belangrijkste zijn.

3.5 Analyses

Om de interbeoordelaarsbetrouwbaarheid van de beoordelingen te kunnen testen wordt gebruik gemaakt van de intraklassecorrelatiecoëfficiënt, voortaan ICC-A (Rhebergen et al. 2011). De reden hiervoor is dat het gaat om een ratio meetniveau (Verhoeven, 2010). Binnen deze analyse zal het *two-way random* model gebruikt worden omdat de beoordelaar telkens willekeurig is en niet vast staat (Shrout and Fleiss 1979). Om te bepalen wat de overeenkomst is tussen de beoordelaars en adviseurs wordt het type *absolute agreement* (ICC-A) gebruikt (de Vet, Terwee, Knol & Bouter, 2006). Eerst worden de analyses tussen de eerste en tweede beoordelaar uitgevoerd (intern) en daarna wordt de data van de twee beoordelaars afzonderlijk met de externe adviseur gecorreleerd. Dit gebeurt eerst per hoofdstuk en dan het totaal cijfer zowel voor het schriftelijke gedeelte als voor het mondelinge gedeelte.

De vragenlijst (survey) wordt op verschillende manieren geanalyseerd. Van de kwantitatieve beoordeling van de beoordelingsprotocollen zal, na het bepalen van de Cronbach's Alpha, een gemiddelde gemaakt worden voor de internen en externen. Hetzelfde wordt gedaan voor de analyse van de vragenlijst. Voor het construct 'tevredenheid over het testmateriaal' wordt de independent samples t-toets (Baarda, de Goede & van Dijkum, 2011) gebruikt om de twee groepen (internen en externen) te vergelijken. Dit construct is van het meetniveau interval. Dit meetniveau geldt overigens voor alle analyses met uitzondering van het kwalitatieve onderdeel. Voor zowel het gesprek als het schriftelijke gedeelte zal dezelfde analyse worden uitgevoerd. De primaire doelstelling is een mogelijke samenhang tussen het type beoordelaar en de uitkomst voor het construct tevredenheid te vinden. Daarnaast zal ook de Pearson's product-moment correlatiecoëfficiënt, ook wel Pearson correlatie genoemd (Howitt & Cramer, 2007), uitgevoerd worden bij het vaststellen van de samenhang naar het aantal jaar dat een beoordelaar scripties beoordeeld en de hoogte van het cijfer.

De kwalitatieve data zal inductief geanalyseerd worden (Verhoeven, 2010). Dit houdt in dat er vooraf niet wordt gezocht naar bepaalde thema's maar dat deze worden afgeleid aan de hand van de beschikbare data. Dit gedeelte van het onderzoek is exploratief en dient als uitgangspunt voor een vervolgonderzoek. De geanalyseerde kwalitatieve data zal gebruikt worden om de kwantitatieve data uit de vragenlijst aan te vullen zodat naast de bepaling van de IBB ook een platform ontstaat waar beoordelaars kunnen aangeven wat ze vinden van de beoordelingsprotocollen en het verloop van de beoordeling van zowel het schriftelijke als het mondelinge gedeelte.

3.6 Hypothesen

Op basis van bovenstaande literatuur van dit onderzoek kunnen de volgende hypothesen gevormd worden:

Door de aanbevelingen uit het rapport 'vreemde ogen dwingen' (2012) kan een voldoende overeenkomst tussen internen en een onvoldoende overeenkomst tussen internen en externen gehaald worden. In het rapport (2012) wordt beweerd dat internen te instemmend zijn met elkaar zouden zijn en dat "vreemde ogen" objectiviteit zouden brengen. Hierdoor kan er onvoldoende overeenkomst tussen internen en externen verwacht worden. De druk van vreemde ogen zou kunnen leiden tot meer objectiviteit. Op basis van bovenstaande kunnen de volgende hypothesen gevormd worden:

1. De samenhang tussen de interne beoordelaars over het schriftelijke gedeelte zal voldoende zijn maar tussen de internen en externen zal de samenhang onvoldoende zijn.

Deze hypothese wordt aangenomen als het correlatie coëfficiënt volgens de COTAN beoordeling tussen internen 0,60 of hoger is en tussen de internen en externen lager dan 0,60. De hypothese wordt verworpen bij elk ander resultaat.

2. De samenhang tussen de interne beoordelaars over het mondelinge gedeelte zal voldoende zijn maar tussen de internen en externen zal de samenhang onvoldoende zijn.

Deze hypothese wordt aangenomen als het correlatie coëfficiënt volgens de COTAN beoordeling tussen internen 0,60 of hoger is en tussen de internen en externen lager dan 0,60. De hypothese wordt verworpen bij elk ander resultaat.

Volgens Bensing (1991) is tevredenheid over het beoordelingsmateriaal een goede voorspeller van de kwaliteit daarvan. Aangenomen kan worden dat de tevredenheid van de beoordelaars, op een paar punten na, redelijk goed zit omdat er binnen de AMA geen klachten zijn binnen gekomen van beoordelaars die duiden op algehele ontevredenheid over de beoordelingsprotocollen. Daarnaast wordt verondersteld dat hoe meer ervaring een beoordelaar heeft, op basis van de literatuur, geen bepalende factor zal zijn voor de hoogte van het cijfer (Rodger et al. 2014). Op basis hiervan worden de volgende hypothesen gevormd:

3. De beoordelaars zijn over het algemeen tevreden over de beoordelingsprotocollen.

Deze Hypothese wordt aangenomen als er geen aangrijpende veranderingen door een grote groep (minimaal 10) wordt verlangd ongeacht het volume. De hypothese wordt verworpen als meer dan tien beoordelaars een aangrijpende verandering willen. Met aangrijpend wordt bedoeld dat een kern element wordt toegevoegd of verwijderd. Hoewel de kwantiteit geen bepalende rol speelt binnen het bepalen van de tevredenheid is toch gekozen voor een minimum van tien beoordelaars omdat dan verwacht kan worden dat het niet gaat om een uitschieter maar dat er meerdere beoordelaars zijn die niet tevreden zijn met de protocollen.

4. Er zal geen (statistisch significante) samenhang zijn tussen het aantal jaar (ervaring) dat een beoordelaar scripties beoordeeld en het hoogte van het cijfer.

Deze hypothese wordt aangenomen als de samenhang tussen het aantal jaar aan beoordelingen en het hoogte van het cijfer niet significant is. De hypothese wordt verworpen als de samenhang tussen de bovengenoemde variabelen significant is.

Hoofdstuk 4 – Onderzoeksresultaten

In dit hoofdstuk wordt de respons op de vragenlijst van het onderzoek beschreven alsmede eventuele bijzonderheden tijdens de uitvoering van het onderzoek. Daarna volgen de resultaten die per deelvraag worden beantwoord.

4.1 Uitvoering

De AMA heeft als opdrachtgever toegang verleend tot de digitaal opgeslagen beoordelingen. Deze beoordelingen zijn per student ingevoerd in een SPSS bestand. Voor elke student is drie keer een beoordelingsformulier ingevuld. De gegevens uit dit bestand zijn geanalyseerd door middel van de intraklassecorrelatiecoëfficiënt. Deze is uitgevoerd tussen de eerste en tweede beoordelaar, tussen de eerste beoordelaar en de adviseur en tussen de tweede beoordelaar en de adviseur. Deze analyses zijn uitgevoerd voor elk hoofdstuk afzonderlijk en voor de totale cijfers. Hetzelfde is gedaan het mondelinge gedeelte. In totaal zijn 187 scripties meegenomen in de analyse.

Om voldoende respons te krijgen voor de vragenlijst is een e-mail naar in totaal 58 beoordelaars verstuurd met een link naar de vragenlijst. De e-mail adressen zijn verkregen via de AMA. De vragenlijst is twee weken online geweest. In de e-mail werd het onderzoek kort toegelicht met daarbij een motivatie om zoveel mogelijk respondenten te krijgen. De ingevulde vragenlijsten werden ingevoerd in SPSS. Hier zijn verschillende analyses gedaan om de samenhang tussen geslacht of type beoordelaar (intern/extern) ten opzichte van de tevredenheid over de beoordelingen vast te stellen.

De antwoorden op de open vragen zijn allemaal terug te vinden in bijlage 1. Deze zijn per respondent geordend en inductief gecodeerd door algemene labels toe te kennen. Sommige antwoorden hebben meerdere labels omdat de er meerdere antwoorden zijn gegeven. Deze labels vormen de thema's die beoordelaars aankaarten ter verandering van het beoordelingsproces. De labels zijn geturfd en de meest prevalentie worden uitgelicht. In dit onderzoek wordt een thema gebruikt als het drie keer of vaker werd genoemd.

4.2 Respons

Voor het kwantitatieve deel van het onderzoek zijn in totaal 187 scripties beoordeeld. Hiervan waren er, vanwege missing values door bijvoorbeeld ziekte, 107 scripties voor het onderzoek bruikbaar.

Tabel 4

Respons per type beoordelaar.

Type Beoordelaar	n
Internen	23
Externen	8
Totaal	31

Voor het kwalitatieve deel van het onderzoek is aan alle beoordelaars en adviseurs die ten minste de afgelopen twee jaar een scriptie hebben beoordeeld een mail verstuurd. Hierin werd hen gevraagd om hun tevredenheid over de beoordelingsprotocollen te uiten. De respons voor dit kwalitatieve gedeelte is te zien in tabel 5. Na verstrekken van de vastgestelde periode van twee weken, die waren uitgetrokken voor het invullen van de vragenlijst, hebben in totaal 31 respondenten de vragenlijst ingevuld. 26 respondenten hebben de vragenlijst volledig ingevuld. In tabel 5 is te zien hoe de beoordelaars en adviseurs verdeeld zijn.

Tabel 5*Verdeling beoordelaar/adviseur en de respons op de vragenlijst.*

	(n)	Totale populatie	% Totale populatie
<i>Interne beoordelaars</i>	19	32	59
<i>(Externe) Adviseurs</i>	7	26	27
<i>Totaal</i>	26	58	45

4.3 Onderzoeksresultaten

De centrale vraag binnen dit onderzoek draait om de interbeoordelaarsbetrouwbaarheid van de scripties voor de opleiding Toegepaste Psychologie. Om deze vraag te kunnen beantwoorden zijn er deelvragen geformuleerd die gezamenlijk een antwoord kunnen vormen op de onderzoeksvraag. Om de interbeoordelaarsbetrouwbaarheid te kunnen vast stellen is de intraklassecorrelatiecoëfficiënt (ICC-Agreement voortaan ICC-A) gebruikt. De uitkomsten van deze analyses zijn te vinden in paragraaf 4.3.1.

4.3.1 Deelvragen over de interbeoordelaarsbetrouwbaarheid

De eerste deelvraag heeft betrekking op de ICC-A tussen de beoordelaars van de afzonderlijke hoofdstukken (tabel 6) en de ICCA-A van het gegeven eindcijfer van het schriftelijke gedeelte. Per hoofdstuk is de correlatie tussen de beoordelaars weergegeven en de interpretatie op basis van COTAN (2012) omdat deze interpretatie binnen het (psychologisch) onderwijs in Nederland vaak wordt aangehouden.

Tabel 6*Overzicht van de ICC-A voor de vijf afzonderlijke hoofdstukken van de scriptie tussen de verschillende beoordelaars.*

<i>Hoofdstuk</i>	<i>ICC-A</i>		
	<i>Internen</i>	<i>Eerste - Externe</i>	<i>Tweede - Externe</i>
<i>1</i>	0,51 (O)	0,25 (O)	0,16 (O)
<i>2</i>	0,55 (O)	0,27 (O)	0,21 (O)
<i>3</i>	0,57 (O)	0,23 (O)	0,19 (O)
<i>4</i>	0,59 (O)	0,17 (O)	0,26 (O)
<i>5</i>	0,62 (V)	0,38 (O)	0,30 (O)

Noot. Tussen haakjes staan de kwalitatieve interpretaties aan de hand van tabel 2. De afkortingen betekenen: G=Goed, V=Voldoende en O=Onvoldoende.

De ICC-A van de totale eindcijfers van het schriftelijke gedeelte tussen de interne beoordelaars is 0,65 (voldoende); tussen de eerste beoordelaar en adviseur 0,26 (onvoldoende); tussen de tweede beoordelaar en adviseur 0,29 (onvoldoende). Geen enkele correlatie was significant ($p > 0,147$ en $p < 0,156$).

De tweede deelvraag heeft betrekking op de ICC-A tussen de beoordelaars van de afzonderlijke onderdelen en het gegeven eindcijfer van het mondelinge gedeelte. In tabel 7 is te zien dat bij de beoordelingen van het mondelinge gedeelte kleine overeenstemmingen te zien zijn maar de mate waarin niet significant is.

Tabel 7

Een overzicht van de ICC-A voor het mondelinge gedeelte tussen de eerste en tweede beoordelaar (internen), de eerste beoordelaar en externe adviseur (Eerste – Externe) en de tweede beoordelaar en externe adviseur (Tweede – Externe).

Onderdeel	ICC-A		
	Internen	Eerste - Externe	Tweede - Externe
Presentatie	0,82 (G)	0,74 (V)	0,72 (V)
Inbreng	0,86 (G)	0,83 (G)	0,83 (G)
Stellingen	0,72 (V)	0,68 (V)	0,68 (V)

Noot. Tussen haakjes staan de kwalitatieve interpretaties aan de hand van tabel 2. De afkortingen betekenen: G=Goed, V=Voldoende en O=Onvoldoende.

De ICC-A van de totale eindcijfers van het mondelinge gedeelte tussen de interne beoordelaars is 0,9 (Goed); tussen de eerste beoordelaar en adviseur 0,86 (Goed); tussen de tweede beoordelaar en adviseur 0,85 (Goed). Geen enkele correlatie was significant ($p > 0,124$ en $p < 0,186$).

4.3.2 Deelvragen over de vragenlijst

De gebruikte vragenlijst heeft een hoge betrouwbaarheid zoals ook te zien is in tabel 8. De vragen over zowel de beoordelingsprotocollen als over het eindgesprek slaan op het construct tevredenheid.

Tabel 8

Betrouwbaarheidsanalyse Likertschaal voor de twee aspecten van de scriptie.

	Cronbach's Alpha
Vragen over de beoordelingsprotocollen	0,88
Vragen over het eindgesprek	0,92

De derde deelvraag heeft betrekking op hoe tevreden de beoordelaars en adviseurs zijn over de kwaliteit van de beoordelingsprotocollen. Daarvoor is een gemiddelde cijfer vanuit de vragenlijst nodig. Het gemiddeld cijfer van de interne adviseurs is 6,8 (sd: 0,9) en het gemiddelde cijfer van de adviseurs is 6,9 (sd: 0,9).

De vijfde deelvraag heeft betrekking op de samenhang tussen het aantal jaar dat de beoordelaars en adviseurs scripties beoordelen en het hoogte van het cijfer voor het schriftelijke gedeelte en het mondelinge gedeelte. Om de samenhang te berekenen tussen het aantal jaar dat een beoordelaar of adviseur scripties beoordeelt en de hoogte van het cijfer van het schriftelijke gedeelte, worden deze variabelen gecorreleerd met de Pearson correlatie omdat beide het meetniveau ratio hebben. Na de analyse blijkt dat er geen significante correlatie gevonden is voor het schriftelijke gedeelte ($r = 0,087$; $p = 0,674$). De cijfers voor de beoordelingsprotocollen zijn getoetst met de Pearson-correlatie. Beide zijn ratio meetniveau. Hier is echter wel een significante (negatieve) correlatie ($r = -0,447$) voor het aantal jaar dat de beoordelaar/adviseur heeft beoordeelt en de hoogte van het cijfer. Doordat het gaat om een negatieve correlatie betekent het in dit geval dat een beoordelaar die een kortere tijd scripties beoordeelt aan de AMA, hogere cijfers geeft ($r = -0,447$; $p = 0,022$). De correlatie is gemiddeld (Cohen, 1988).

4.3.3 Meningen over opzet van de scriptie beoordelingen

Alle meningen zijn in volledigheid terug te vinden in bijlage een. Hier worden enkel de meest voorkomende thema's kort toegelicht met enkele citaten om het thema te onderbouwen. Er zijn vier open vragen gesteld. De eerste twee vragen gaan over het schriftelijke gedeelte en de laatste twee vragen gaan over het mondelinge gedeelte. Per onderdeel hebben de vragen overlap daarom hebben

sommige vragen binnen een onderdeel hetzelfde thema. Telkens wordt de vraag getoond met een korte samenvatting van de thema's gevolgd door citaten. De deelvraag die bij de eerste twee vragen hoort is de volgende: *Wat vinden beoordelaars en adviseurs van de opzet van het schriftelijke gedeelte en wat zouden ze hieraan willen veranderen?*

Vraag 1: Wat zou volgens u beter kunnen bij het beoordelen van het schriftelijke gedeelte?

Voor deze vraag was een veel voorkomend thema er meer of minder onderdelen nodig zijn. Beoordelaars gaven aan dat een aantal criteria ontbreken waar zij graag de student op zouden willen scoren. Anderen zeiden juist dat er een teveel aan criteria zijn. Een overlappend thema is dat de criteria of onderdelen van beoordeling soms te specifiek zijn. De beoordelaars vonden dat dit algemener kon worden aangegeven zodat er genoeg ruimte is voor creativiteit van de student. Ook gaven beoordelaars aan dat de criteria in sommige gevallen op sommige onderdelen niet duidelijk genoeg zijn. Er mag daarnaast meer flexibiliteit komen voor het beoordelen waarbij meer ruimte is voor commentaar en toepasselijk kan worden beoordeeld op het type onderzoek. Daarnaast gaf een klein deel van de beoordelaars aan dat de vragenlijst prima is en er niets aan verandert hoeft te worden.

Enkele citaten uit de vragenlijst:

“De criteria zijn niet altijd even duidelijk geformuleerd, dit kan ervoor zorgen dat beoordelaars het heel anders interpreteren. Het probleem is alleen dat je het ook niet te specifiek wilt maken, want dan wordt het voor de student meer een invuloefening dan dat ze zelf kritisch na moeten denken. Het zou nuttig kunnen zijn om de beoordelingscriteria specifiek te maken op sommige onderdelen en deze alleen beschikbaar te maken voor docenten.”

“Minder beoordelingspunten per onderdeel (m.n. Minder per hoofdstukparagraaf), waardoor het nu erg op detail niveau blijft ipv op grotere geheel., 2. Ook andere aspecten meenemen, zoals samenvatting, schrijfstijl, soort bronnen, creativiteit), 3. Meer op inhoud ipv of het aspect aanwezig is (meer op kwaliteit dan op kwantiteit)”

“Ik zou graag veel minder criteria willen, dus minder punten wat ik moet beoordelen. Daarnaast zou ik graag cijfers geven per onderdeel. Ook zou ik graag het proces en functioneren van de student willen beoordelen.”

Vraag 2: Zou u de beoordelingen van het schriftelijke gedeelte anders aanpakken? Wat zou u dan anders doen?

Voor deze vraag was het meest voorkomende thema wederom dat er meer of minder criteria nodig zijn in de beoordeling van het schriftelijk gedeelte. Flexibiliteit is hierin ook belangrijk volgens de beoordelaars. Daarnaast vonden ze dat de beoordeling korter en duidelijker kon. Een klein deel vond de beoordeling in huidige vorm prima.

Enkele citaten uit de vragenlijst:

“Korter en bondig... maar het risico hiervan is dat de beoordeling minder transparant en objectief wordt. Voor zo iets belangrijks heb je veel informatie nodig om je beoordeling te rechtvaardigen. Denk bv aan de intelligentietest, deze test bevat niet alleen 10 items, of wel?”

“Kortere beoordelingsformulieren”

“ook beoordelen van, -samenvatting, -kwanti- en kwaliteit van de bronnenlijst”

“Nee, ben tevreden over huidige opzet.”

De deelvraag die hoort bij de laatste twee vragen gaat als volgt: *Wat vinden beoordelaars en adviseurs van de opzet van het mondelinge gedeelte en wat zouden ze hieraan willen veranderen?*

Vraag 3: Wat zou volgens u beter kunnen bij het beoordelen van het mondelinge gedeelte?

De beoordelaars geven voor deze vraag vaak diverse antwoorden waardoor er veel thema's zijn. Er komen een aantal thema's naar voren die enkele keren worden benoemd. Het gesprek volgt geen vast patroon en zorgt voor onduidelijkheid voor de beoordelaar en student. De weging van de beoordelingen staat niet vast en het kan volgens enkele beoordelaars betrouwbaarder. Ook wordt de berekening van de cijfers ervaren als tijdrovend en vervelend.

Enkele citaten uit de vragenlijst:

“Bij het eindgesprek is er meer een algemene beoordeling waar sprake van is. Ik persoonlijk gebruik de criteria meer als kapstok om het cijfer aan op te hangen, niet zozeer als maatstaf om het cijfer te bepalen.”

“en compleet ander formulier met andere criteria en andere weging. Nu is de weging 20-70-10, dat oogt al meteen als een bron van onbetrouwbaarheid.”

“Eindgesprek vind ik altijd een soort 'feestje', (tenzij de kandidaat natuurlijk erg gespannen is of slecht presteert). , Wel vind ik het gerommel met papieren bij de beoordeling nogal omslachtig. Iedereen moet z'n handtekening 6x zetten. Er moet veel worden berekend. Dat moet toch efficiënter kunnen.?”

“Het eindgesprek loopt niet volgens vaste richtlijnen. Het presentatiegedeelte is wel duidelijk. Hoe het daarna loopt, wie wat vraagt, of en zo ja, welke stellingen aan bod komen wisselt en wat we verwachten van studenten kan eenduidiger.”

Vraag 4: Zou u de beoordelingen van het eindgesprek anders aanpakken? Wat zou u dan anders doen?

De beoordelaars willen andere criteria waarop ze kunnen beoordelen. Sommigen vinden juist dat er te veel criteria zijn. Daarnaast is het invullen van de formulieren een knelpunt voor veel beoordelaars. Ze willen een digitaal formulier dat na eventueel overleg gezamenlijk wordt ingevuld om het gerommel met papier zo veel mogelijk te vermijden.

Enkele citaten uit de vragenlijst:

“misschien wat minder criteria en meer het onderbouwen van een cijfer ipv proberen met kleine criteria tot een cijfer te komen. bij een gesprek werkt dat toch anders dan een opdracht op papier”

“Automatische berekening van eindcijfer (bijv formulier laten invullen via tablet van 1e beoordelaar)”

“In een bestand werken die meteen met formules het eindcijfer berekend”

“Soms ligt het accent te sterk op het onderzoekstechnische deel. , Aandacht voor visie, eigen mening van de kandidaat waarin hij/zijn blijkt geeft van overstijgend kunnen denken, verbanden kunnen leggen, zaken in de context kunnen plaatsen, gericht op de beroepspraktijk.”

Hoofdstuk 5 – Conclusies, discussie en aanbevelingen

In dit hoofdstuk worden de onderzoeksvragen beantwoord en worden de mogelijke oorzaken daarvan besproken. Daarnaast staat het onderzoek ter discussie en worden mede daaruit aanbevelingen gedaan voor de AMA om vervolgstappen te ondernemen.

5.1 Conclusies

In dit onderzoek is getracht om de betrouwbaarheid tussen drie verschillende beoordelaars, waarvan één een adviserende rol heeft, vast te stellen. Om specifiek te zijn gaat het hier dus om interbeoordelaarsbetrouwbaarheid. Om de interbeoordelaarsbetrouwbaarheid vast te kunnen stellen binnen de AMA voor de opleiding Toegepaste Psychologie werd de onderzoeksvraag opgesplitst in enkele deelvragen. De uitkomsten van deze deelvragen geven een antwoord op de onderzoeksvraag. Daarom zullen eerst de conclusies die daaruit voortkomen worden besproken voordat er antwoord wordt gegeven op de hoofdvraag. Enkele deelvragen zijn samengevoegd in paragrafen om veelvuldige herhaling van tekst te voorkomen.

5.1.1 Betrouwbaarheid afzonderlijke hoofdstukken van het schriftelijke gedeelte

Om antwoord te kunnen geven op de deelvragen die betrekking hebben op de beoordeling van de vijf afzonderlijke hoofdstukken van het schriftelijke gedeelte zijn een drietal analyses uitgevoerd, te weten: tussen de eerste en tweede beoordelaar; tussen de eerste beoordelaar en de adviseur; tussen de tweede beoordelaar en adviseur. De correlatie tussen de interne beoordelaars is voor de afzonderlijke hoofdstukken onvoldoende op hoofdstuk vijf na. Tussen de eerste beoordelaar en de adviseur zien we een onvoldoende correlatie. Hetzelfde geldt voor de correlatie tussen de tweede beoordelaar en de adviseur. Hieruit kan geconcludeerd worden dat de samenhang tussen de internen groter is dan tussen de internen en externen. Ondanks dat de samenhang tussen internen onvoldoende is liggen de coëfficiënten gemiddeld hoger dan bij de samenhang tussen internen en externen. Daarnaast moet opgemerkt worden dat de waarden tussen de internen volgens COTAN criteria nagenoeg voldoende zijn. Een mogelijke verklaring voor de betere correlatie bij internen kan zijn dat de internen vaker scripties beoordelen dan externen en meer bezig zijn met beoordelen van schoolwerk. Door de uniformiteit van de beoordelingscriteria en de uitwisseling van informatie tussen docenten (internen) kunnen zij een meer uniforme kijk hebben gekregen op het beoordelen in tegenstelling tot de externe adviseur die minder bezig is met nakijken van schoolwerk en oordeelt vanuit een meer praktijkgerichte benadering.

5.1.2 Betrouwbaarheid eindcijfers van het schriftelijke gedeelte

Om antwoord te kunnen geven op de deelvragen die betrekking hebben op de beoordeling van de eindcijfers van het schriftelijke gedeelte zijn dezelfde drie soorten analyses, zoals beschreven in de vorige paragraaf, uitgevoerd als in paragraaf 5.1.1 met als enige verschil dat nu de eindcijfers werden gebruikt in plaats van de cijfers van de afzonderlijke hoofdstukken. De correlatie tussen internen is bij de eindcijfers goed. De analyse tussen de internen en externen laat weer een onvoldoende correlatie zien. De interne beoordelaars hebben meer overeenstemming onderling dan dat ze dat hebben met de externen. Hiervoor geldt eenzelfde mogelijke verklaring als die uit paragraaf 5.1.1. Tussen de internen is nu wel een voldoende correlatie te zien terwijl bij de afzonderlijke hoofdstukken slechts bij een hoofdstuk (vijf) te zien was. De waarden bij de afzonderlijke hoofdstukken lagen wel dicht tegen de voldoende aan. De voldoende op de eindcijfers kan betekenen dat de onvoldoende correlatie bij afzonderlijke hoofdstukken op toeval berust. Een andere verklaring kan zijn dat binnen de beoordelingsprotocollen onvoldoende verbinding staat

tussen de losse cijfers van de hoofdstukken en het eindcijfer. Op basis van de conclusies uit paragrafen 5.1.1 en 5.1.2 wordt de hypothese, die stelt dat, de samenhang tussen de interne beoordelaars over het schriftelijke gedeelte voldoende zal zijn maar tussen de internen en externen onvoldoende zal zijn verworpen. De harde lijn lag op een ICC-A van 0,60 en deze is bij de afzonderlijke hoofdstukken niet gehaald. Als men enkel uitgaat van het eindcijfer dan kan de hypothese worden aangenomen.

5.1.3 Betrouwbaarheid eindcijfers van het mondelinge gedeelte

Om de deelvragen die betrekking hebben op de beoordeling van het mondelinge gedeelte te kunnen beantwoorden zijn wederom dezelfde analyses uitgevoerd als in paragraaf 5.1.1 met de eindcijfers voor het mondelinge gedeelte. De uitkomsten uit paragraaf 4.3.1 laten een goede correlatie zien voor elke analyse. Dit betekent dat de beide interne beoordelaars en de adviseur een hoge overeenstemming hebben wat betreft het geven van het eindcijfer voor het mondelinge gedeelte. Dit staat in contrast met de uitkomsten uit de analyses over het schriftelijke gedeelte. Een mogelijke verklaring hiervoor kan zijn dat het mondelinge gedeelte meer praktisch gericht is en dat er veel interactie mogelijk is tussen de beoordelaars en afstudeerders. Hierdoor kan het oordeel genuanceerd worden wat leidt tot meer overeenstemming bij de beoordelaars. Op basis van de conclusie uit paragraaf 5.1.3 kan de hypothese, die stelt dat, de samenhang tussen de interne beoordelaars over het mondelinge gedeelte voldoende zal zijn maar tussen de internen en externen onvoldoende zal zijn verworpen worden.

5.1.4 Tevredenheid onder de beoordelaars over de kwaliteit van de protocollen

Om de vragen 'In hoeverre zijn de beoordelaars tevreden over de kwaliteit van de beoordelingsprotocollen?' en 'Wat vinden beoordelaars en adviseurs van de opzet van het schriftelijke gedeelte en het mondelinge gedeelte en wat zouden ze hieraan willen veranderen?' te kunnen beantwoorden is gebruik gemaakt van een tweetal open vragen per onderdeel. Hieruit blijkt dat een grote groep beoordelaars de huidige beoordelingsprotocollen op een of meerdere plaatsen zou willen aanpassen. Het merendeel betreft kleine aanpassingen waarbij door een enkeling ook om een ingrijpendere verandering wordt gevraagd. Op basis van deze gegevens kan gesteld worden dat een meerderheid van de beoordelaars over het algemeen tevreden zijn over de beoordelingsprotocollen ondanks dat ze kleine aanpassingen hieraan willen. De hypothese, die stelt dat de beoordelaars over het algemeen tevreden zullen zijn over de beoordelingsprotocollen daardoor worden aangenomen.

5.1.5. Samenhang jaren ervaring en hoogte cijfers

Om de vragen 'Wat is de samenhang tussen het aantal jaar dat de beoordelaars en adviseurs scripties beoordelen en het hoogte van het cijfer voor het schriftelijke gedeelte?' en 'Wat is de samenhang tussen het aantal jaar dat de beoordelaars en adviseurs scripties beoordelen en de hoogte van het cijfer voor het mondelinge gedeelte?' te beantwoorden is een Pearson correlatie uitgevoerd. Voor het schriftelijke gedeelte is geen statistisch significante samenhang gevonden. Voor de cijfers van het mondelinge gedeelte is wel een statistisch significante samenhang gevonden. Het betreft een negatieve samenhang wat betekent dat een beoordelaar/adviseur met een korter aantal jaren van beoordelen hogere cijfers geeft. Een mogelijke verklaring hiervoor kan zijn dat de meer ervaren beoordelaars kritischer of strenger beoordelen omdat zij bekender zijn met de protocollen en veel meer scripties beoordeeld hebben waardoor ze een strengere norm hanteren. Dat het enkel bij het mondelinge gedeelte significant is kan komen doordat het beoordelingsprotocol voor het mondelinge gedeelte minder gedetailleerde criteria bevat, waardoor er meer vrijheid is om kritischer

te zijn in tegenstelling tot het schriftelijke gedeelte waar de criteria in veel meer detail beschreven staan. De hypothese die op basis van Rodger et al. (2014) gesteld was wordt hiermee, vanwege het significante (negatieve) verschil dus verworpen.

5.1.6 De onderzoeksvraag

De ICC-A kent een zelfde interpretatie als het Kappa coëfficiënt (Landis & Koch, 1977). De meest gebruikte interpretatie is die van de COTAN (2012). Door deze interpretatie kan worden vastgesteld welke waarde de geanalyseerde gegevens hebben. Deze gegevens zullen essentieel zijn in het beantwoorden van de onderzoeksvraag: *Wat is de interbeoordelaarsbetrouwbaarheid voor het beoordelen van scripties binnen de academie mens en arbeid van de hogeschool Saxion te Deventer over de afgelopen twee jaar?*

Bij alle ICC-analyses behalve voor het mondelinge gedeelte werd vastgesteld dat in licht variërende mate de interne beoordelaars onderling redelijk tot goede overeenstemming hebben en volgens de COTAN (2012) een voldoende in het geven van cijfer voor zowel de afzonderlijke hoofdstukken als het eindcijfer voor het schriftelijke gedeelte. De samenhang tussen de interne beoordelaars en externe adviseurs was voor het schriftelijke gedeelte vrijwel constant onvoldoende. Bij het eindcijfer voor het gesprek viel op dat er een voldoende of goede correlatie was tussen alle drie beoordelaars en adviseur. Verder is alleen nog een statistisch significante correlatie gevonden tussen het aantal jaren van beoordelen en de hoogte van het cijfer dat gegeven wordt aan de studenten.

Op basis van bovenstaande luidt het antwoord op de eerder geformuleerde onderzoeksvraag als volgt: Over de interbeoordelaarsbetrouwbaarheid van de scripties van de opleiding Toegepaste Psychologie is niet een alomvattende uitspraak te doen. Dit komt door het gemeten verschil in het schriftelijke en mondelinge gedeelte. Duidelijk is dat de interne beoordelaars een over het algemeen voldoende overeenkomst hebben in zowel het schriftelijke gedeelte als het mondelinge gedeelte. De samenhang met de externe adviseurs is vrijwel constant onvoldoende met uitzondering van het mondelinge gedeelte.

Uit het geformuleerde antwoord kan geconcludeerd worden dat het belangrijk is om een externe beoordelaar te betrekken in het beoordelingsproces omdat deze een duidelijk verschillend oordeel laat zien tegenover de interne beoordelaars. Dit verschil komt in de resultaten van dit onderzoek duidelijk naar voren. Dit betekent dat een kritische blik van buitenaf bevorderend kan zijn voor het waarborgen van de kwaliteit en transparantie van scripties.

5.1.7. De meningen van beoordelaars

De meningen van de beoordelaars werd gevraagd door vier open vragen te stellen. De vragen zijn verdeeld in twee onderdelen: het schriftelijke deel en het mondelinge gedeelte. Per onderdeel werden twee vragen gesteld. Deze vragen hadden overlap waardoor in veel antwoorden ook overlap te zien is. Uit de antwoorden bleek dat de beoordelaars veel verschillende knelpunten en verbeteringen zien aan de beoordeling van hierboven genoemde onderdelen. Er zijn veel nuttige aanpassingen voortgekomen. De meest op het oog springende verbetering die voorgedragen wordt is dat er te veel criteria zijn waarop wordt beoordeeld. Deze veelvuldigheid aan criteria wordt vooral gemerkt bij de beoordeling van de afzonderlijke hoofdstukken van de scriptie. Zo vinden de beoordelaars dus dat er te specifiek op wordt beoordeeld en er dus een meer algemene beoordeling nodig is om de creativiteit van studenten te bevorderen. Er zijn echter ook beoordelaars die opperen dat de criteria aangevuld moeten worden met nieuwe onderdelen zoals de beoordeling van de samenvatting van de scriptie. Deze twee meningen lijken tegenstrijdig en daardoor kan er geen

eenduidig beleid gevoerd worden op basis van de antwoorden op dit punt. Daarnaast wordt er vaak verwezen naar de flexibiliteit van de beoordelingscriteria die vaak te strikt wordt genomen en waarbij er weinig ruimte is voor de creativiteit van de student. De criteria worden gezien als een afvinklijst voor studenten waardoor de aanwezigheid van onderdelen kan worden bepaald maar niet van welk niveau dit onderdeel is. Een klein deel van de beoordelaars vindt dat hier aandacht aan besteed moet worden. Een ander punt is dat de criteria en uitvoering van het mondelinge gedeelte vaak onduidelijk zijn. Hoewel voor zowel het schriftelijke gedeelte als het mondelinge gedeelte vaak dezelfde thema's komen bovendien wordt op de criteria van het mondelinge gedeelte vaker positief gereageerd. Veel beoordelaars vinden dat die criteria minder in aantal zijn en in sommige gevallen meer vrijheid bieden. Er wordt ook aangegeven dat er een vast patroon ontbreekt waardoor er onduidelijkheid heerst bij zowel beoordelaar als student.

5.2 Discussie

De kern van het onderzoek bestaat uit de interbeoordelaarsbetrouwbaarheid (kwantitatieve component) van de scripties echter is in het onderzoek ook een kwalitatieve component meegenomen. Hierdoor lijkt het of een deel van het onderzoek niet verwerkt is in de onderzoeksvraag. In zekere zin klopt dit. Voor de AMA ging het echter hoofdzakelijk om kwantitatieve component maar zij wilden hierbij ook kijken of er een verschil was in de tevredenheid over de beoordelingen. Er is opzettelijk door middel van vier open vragen ruimte gegeven aan de beoordelaars en adviseurs om hier hun eigen gedachten over de beoordelingen te uiten. Deze twee verschillende onderzoeksdelen kunnen daardoor voor verwarring zorgen over waar het onderzoek daadwerkelijk over gaat. De voorkeur vanuit de AMA gaat naar de kwantitatieve component en deze geniet ook de voorkeur als er conflicten ontstaan tussen de twee. In grote lijnen zullen de twee componenten aanvullend werken.

Het tweede punt gaat over de lengte van de vragenlijst die gebruikt is om de meningen van beoordelaars te peilen. Deze is met vier vragen heel erg kort. Dit is enigszins ook de bedoeling geweest maar hiermee wordt het risico gelopen dat er te weinig ruimte is om over een construct vragen te stellen. Analyses over de vragen die gebruik maken van de likert-schaal laten zien dat de vragenlijst betrouwbaar is. De interne consistentie is hoog dus zijn de gegevens die uit de vragenlijst komen zeer bruikbaar. Niet enkel omdat de vragenlijst meet wat het moet meten maar ook omdat het hergebruiken van deze vragen in een uitgebreidere vragenlijst voor vervolgonderzoek tijd bespaart. Dit is echter een logisch gevolg van de lengte van de vragenlijst (Jansen, Joostens & Kemper, 2004).

Een derde punt is dat de correlatiecoëfficiënten die gebruikt worden om de ICC-A te bepalen niet eenduidig zijn in hun interpretatie. Hierdoor kan het zijn dat deze niet universeel geaccepteerd worden en dus nadelige gevolgen kunnen hebben voor het onderzoek doordat in andere onderzoeken een andere interpretatie gebruikt wordt. Hierdoor kan moeilijk vergeleken worden welke methoden het beste gebruikt kunnen worden en welke buiten de vastgestelde norm vallen. Onenigheid over de interpretatie kan verschillende resultaten geven wat ervoor zorgt dat de resultaten van het onderzoek als niet eenduidig zijn (Gwet, 2010).

Een vierde punt zijn de vragen die gesteld zijn in de vragenlijst die gebruikt is om de meningen van beoordelaars te peilen. Deze vragen zouden veel overlappen waardoor het aantal effectief gereduceerd kan worden van vier naar twee. Vaak is een verwijzend antwoord te zien om aan te

tonen dat de vragen in betekenis weinig van elkaar verschillen. De vragen hadden beter als een vraag gesteld kunnen worden waardoor er ruimte zou zijn voor een andere soort vraag.

Een vijfde punt is de steekproefgrootte. De steekproefgrootte is minder dan de helft van populatie. Dit heeft enkel betrekking tot de afname van de vragenlijst en niet op de beoordelingsprotocollen. Bij de vragenlijsten kan een te kleine steekproef hebben deelgenomen om een betrouwbare uitspraak te kunnen doen voor een 95% betrouwbaarheidsinterval (Israel, 1992). Een te kleine steekproef kan echter tot gevolg hebben dat er geen significante samenhang kan worden gevonden (Sudman, 1976). Ook bij het gebruiken van subjectieve maten binnen het beoordelen van de scriptie kan een samenhang in de weg staan (Rui & Feldman, 2012). Het onderzoek lijkt hier echter niet onder te leiden omdat de samenhang tussen de internen wel te vinden is maar tussen de internen en externen minder aanwezig lijkt te zijn.

Een zesde punt is dat Bland en Altman (1986) beweren dat bij ICC een gewone correlatie ongeschikt is als er meer dan twee beoordelaars geanalyseerd worden. De vraag is in hoe verre is het combineren van de interne beoordelaars tot één groep en daarmee, feitelijk tellend, als één beoordelaar geschikt om betrouwbare analyses over uit te voeren. Bij een één op één analyse bestaat dit probleem niet. Kan een gemiddelde van de beoordelingen van internen gezien worden als een beoordelaar of twee. Het probleem hierbij is dat de voldoende correlatie tussen internen net voldoende was en dus kan berusten op kans. De vraag is wat dit doet met het gemiddelde en welke invloed dit uiteindelijk heeft op de correlatie tussen internen en externen.

Het laatste punt dat ter discussie staat is het aantal (externe) adviseurs dat heeft deelgenomen aan het onderzoek. De totale populatie is ongeveer in tweeën gesplitst qua verhoudingen. In de respons is deze verhouding niet terug te zien. Door deze geringe respons is het maar de vraag of de uitkomsten voor de groep adviseurs representatief genoeg is. In de toekomst zou deze respons gestimuleerd moeten worden zodat de verhoudingen met de totale populatie zo gelijk mogelijk zijn. Volgens de COTAN (2012) is een groep van minimaal 200 respondenten nodig om een voldoende te scoren op de grootte van de onderzoeksgroep. Binnen dit onderzoek ligt dat getal net onder de 200.

5.3 Aanbevelingen

Op basis van de resultaten van het kwantitatieve gedeelte (interbeoordelaarsbetrouwbaarheid) wordt aanbevolen om de externe adviseur meer te betrekken en wellicht een prominere rol te laten spelen binnen het beoordelen van scripties. Als vanouds zal de externe een bindende beoordeling of advies kunnen geven. De kijk van buitenaf kan voor een objectievere beoordeling van de scriptie zorgen of in ieder geval een andere kijk. De interne beoordelaars kunnen beïnvloed zijn door de beoordelingscultuur binnen de AMA of Saxion waardoor beoordelingen veel op elkaar kunnen lijken omdat op hetzelfde wordt gelet. Het vasthouden aan criteria hoeft niet per se een hogere kwaliteit te betekenen. Het kan creativiteit van de student wegnemen. Aan de andere kant kan ook gesteld worden dat de onervarenheid van de adviseur op het gebied van beoordelen kan zorgen dat de context van binnen de AMA niet helemaal begrepen wordt. Dit laatste draagt echter ook bij aan een objectievere kijk omdat de adviseur niet bezig is met de student of beoordelingscultuur maar puur objectief vanuit zijn of haar kennis de scriptie beoordeelt.

Bovenstaande toont aan dat er een onverkend gebied is waar veel mogelijkheden liggen voor een vervolgonderzoek. Dit onderzoek neigt naar de kwantitatieve component. De kwantitatieve component is binnen dit onderzoek uitvoerig besproken en geanalyseerd. Het misstaat natuurlijk niet om het onderzoek over twee jaar te herhalen om te zien of er verandering is opgetreden in de

interbeoordelaarsbetrouwbaarheid. De prioriteit ligt toch meer in het betrekken van de beoordelaars en adviseurs in het ontwikkelen van een kwalitatief goede beoordelingsprocedure. Hier kan een werkgroep voor worden opgericht die verder (kwalitatief) onderzoek zal doen naar wat beoordelaars en adviseurs vinden van bijvoorbeeld de beoordelingscriteria.

Het wordt zeer aanbevolen om het vervolgonderzoek zo eenduidig mogelijk te formuleren zodat er een duidelijke onderzoeksvraag is waar een goed onderbouwd antwoord op kan worden gegeven. Daarnaast zal goed over een eventuele vragenlijst moeten worden nagedacht. De constructen moeten duidelijk en begrijpelijk worden geformuleerd. Ook zal een onderzoek waarin de nadruk ligt op kwalitatief onderzoek een langere vragenlijst moeten voortbrengen met diepgaande vragen over de beoordelingsprotocollen om een gedetailleerde aanpassing te kunnen doen aan deze beoordelingsprotocollen.

Daarnaast kunnen uit de literatuur ook enkele aanbevelingen worden gedaan. Zo is iets wat vaak naar voren komt de training van beoordelaars. Hoewel deze zich vaak in een ander vakgebied bevinden buiten het onderwijs kan het toch nuttig zijn om te kijken naar wat deze onderzoeken of bevindingen kunnen betekenen voor het bevorderen van de interbeoordelaarsbetrouwbaarheid binnen het hoger onderwijs. Ook hier zal in eerste instantie onderzoek moeten worden gedaan binnen het hoger onderwijs om vast te stellen of dezelfde effecten terug gevonden kunnen worden. Zo is te zien in een onderzoek van Wang (2009) naar de IBB voor een cursus Engels als tweede taal dat een gecontroleerd protocol met gespecialiseerde beoordelaars een hogere kwaliteit en mate van overeenstemming tot gevolg heeft. In eerdere gevallen is aangetoond dat beoordelaars met behulp van zelfstudie van een bepaalde test voor creatief denken betrouwbaar kunnen beoordelen (Halpin & Halpin, 1974). De vraag hierbij blijft in hoeverre dit geldt voor docenten en experts uit het werkveld. Docenten zijn natuurlijk in hun werk vaak bezig met beoordelen van toetsen. Dit hoeft natuurlijk niet meteen te betekenen dat dezelfde vaardigheden of kennis nodig zijn om een scriptie te kunnen beoordelen. Getrainde beoordelaars kunnen ook beter bepalen waarop beoordeeld moet worden (Lievens & van Keer, 1999). Dit kan de objectiviteit en kwaliteit van de beoordelingen ten goede komen. Hier is dus echter onderzoek voor nodig.

Een andere aanbeveling vanuit de literatuur die iets concreter is over de beoordelingsprotocollen is het gebruik maken van rubrics binnen de beoordelingsprotocollen (Westenberg, 2016). Rubrics kunnen zorgen voor duidelijkere criteria waardoor de IBB hoger kan uitvallen en daarmee dus de kwaliteit van zowel de beoordelingsprotocollen als ook de gemaakte scripties kan verhogen. Studenten weten dan waar ze op moeten letten bij het schrijven en voorbereiden van een scriptie en de beoordelaars weten precies hoeveel punten kunnen worden toegekend op elk punt. Deze zullen dan heel concreet moeten worden gedefinieerd. Deze aanpassing zou ook een aantal van de punten uit de vragenlijst kunnen oplossen. Dit kan leiden tot meer tevredenheid bij beoordelaars. Natuurlijk wordt de creativiteit van de studenten hiermee niet bevordert maar deze wordt ook niet significant beperkt. Er kunnen dan ook minder criteria komen. In het algemeen lijkt deze aanpassing het meeste voordeel te bieden voor alle betrokken partijen. Ook hiernaar zal onderzoek moeten worden gedaan. Zowel een tevredenheidsonderzoek naar het gebruik en toepassing van de beoordelingsprocedure met Rubrics als ook de IBB hiervan.

5.3.1 Plan van aanpak

De examinering van de AMA zal een werkgroep kunnen formeren die de kwaliteit van de beoordelingen van de scripties zal moeten waarborgen en verbeteren. Deze werkgroep bestaat bij voorkeur uit docenten die zelf betrokken zijn bij het beoordelen van scripties binnen de AMA, Toegepaste Psychologie studenten die aan het afstuderen zijn en een of meerdere externe adviseurs. De werkgroep zal zich bezig houden met het doen van onderzoek naar en het evalueren van het beoordelingsproces. De evaluatie vindt bij voorkeur minimaal een keer per schooljaar plaats. De werkgroep hoeft geen vaste leden te hebben zolang alle vergaderingen en vorderingen goed bijgehouden worden.

In de eerste plaats zal de werkgroep onderzoek moeten doen naar wat de verschillende beoordelaars (zowel intern als extern) vinden van het gehele proces. Dit onderzoek zal hoofdzakelijk kwalitatief van aard zijn. De huidige vragenlijst kan gebruikt worden maar zal uitgebreid moeten worden. Er zal meer diepgang nodig zijn. De manier waarop deze vragenlijst wordt afgenomen kan verschillen van de huidige vorm. De voorkeur gaat uit naar semigestructureerde interviews die face-to-face worden afgenomen. De vragenlijst hoeft dan niet bij de gehele populatie te worden afgenomen wat veel tijd bespaart hoewel met een werkgroep de nodige mankracht om een grote steekproef uit te voeren beschikbaar is.

De evaluatie kan bestaan uit een steekproefsgewijze korte analyse in SPSS om te zien of er veranderingen in de interbeoordelaarsbetrouwbaarheid optreden. Ook zou er een platform moeten komen voor beoordelaars of adviseurs om suggesties te doen voor verbeteringen of aanpassingen. Deze kunnen door de werkgroep verder worden onderzocht. De werkgroep zal een adviserende rol hebben tegenover de AMA. Het is aan de AMA of ze de onderzochte toepassingen doorvoeren.

Het onderzoek naar trainingen zal moeten starten met een onderzoek naar trainingen voor beoordelaars binnen het hoger onderwijs of andere takken en niveaus van onderwijs binnen Nederland. Vervolgens kan in hetzelfde onderzoek een verkenning worden gedaan naar wat voor training of cursus eventueel nodig zou kunnen zijn. Een ander onderzoek kan dan de effectiviteit van deze training meten.

Daarnaast zou er ook onderzoek moeten worden gedaan naar de toepassing van rubrics binnen de beoordelingsprotocollen. Dit kan binnen de werkgroep maar kan ook als onafhankelijk onderzoek worden uitgevoerd. Bij voltooiing van het onderzoek en de toepassing van de rubrics is onderzoek nodig om te kijken naar de betrouwbaarheid hiervan. Ook de tevredenheid hierover is belangrijk om te meten. Kortom er zijn veel kanten die de AMA op kan maar er zal veel onderzoek nodig zijn om vast te kunnen stellen hoe effectief alle toepassingen en verandering daadwerkelijk zijn.

Literatuurlijst

- Andrade, H. G. (2000). Using Rubrics to Promote Thinking and Learning. *Educational Leadership*, 57(5), 13-18.
- Baarda, B., De Goede, M., Van Dijkum, C. (2007). *Basisboek statistiek met SPSS. Handleiding voor het verwerken en analyseren van en rapporteren over (onderzoeks) gegevens*. Noordhoff Uitgevers, Groningen.
- Bensing, J. M. (1991). Wie zegt dat di teen goed consult is? Kwaliteitsoordelen uit drie verschillende bronnen met elkaar vergeleken. *Huisarts en wetenschap*. Geraadpleegd op <http://www.nivel.nl/node/2430?database=ChoicePublicat&preref=90>
- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, pp. 307–310.
- Bruijns, V., & Kok, M. (2013). *Leidraad Toetsen en Beoordelen*. Amsterdam: Hogeschool van Amsterdam.
- Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, 67, 343-368. doi:10.1080/00131911.2014.929565
- Carmines, E. G., Zeller, R. A. (1979). Reliability and Validity Assessment. *Quantitative Applications in the Social Sciences*, Vol 17. SAGE Publications.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*. 6 (4): 284–290. doi:10.1037/1040-3590.6.4.284
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates, Hillsdale.
- Commissie externe validering examenkwaliteit hoger beroepsonderwijs (2012). *Vreemde ogen dwingen*. HBO-raad, vereniging van hogeschool, Den Haag.
- Commissie Leers (2011). *Veel ruimte, weinig rekenschap: Onderzoek naar het alternatieve afstudeertraject van de opleiding Media & Entertainment Management van INHolland te Haarlem*.
- Evers, A., Lucassen, W., Meijer, R. & Sijtsma, K. (2010). COTAN beoordelingssysteem voor de kwaliteit van tests. Nederlands Instituut van psychologen.
- Fleiss, J. L., Cohen, J. (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, Vol. 33 pp. 613–619
- Gwet, K. L. (2010) *Handbook of Inter-Rater Reliability*. Gaithersburg : Advanced Analytics, LLC
- Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol*.
- Halpin, G., Halpin, G. (1974), Can self-trained scorers reliably score the torrance tests of creative thinking? *Psychol. Schs.*, 11: 56–58. doi:10.1002/1520-6807(197401)11:1<56::AID-PITS2310110113>3.0.CO;2-A

- Howitt, D., Cramer, D. (2007). *Statistiek in de sociale wetenschappen*. Pearson Education Benelux, Amsterdam.
- IBM Corp. (2013). IBM SPSS Statistics for Windows, Version 22.0. IBM Corp, Armonk.
- Inspectie van het Onderwijs (2011). Alternatieve afstudeertrajecten en de bewaking van het eindniveau in het hoger onderwijs. Ministerie van onderwijs, cultuur en wetenschap, Utrecht.
- Inspectie van het Onderwijs (2012). Alternatieve afstudeertrajecten in het hoger onderwijs: Rapportage over de niveau-beoordelingen en de verbetertrajecten. Ministerie van onderwijs, cultuur en wetenschap, Utrecht.
- Israel, G. (1992). Determining sample size. Program Evaluation and Organizational Development, IFAS, University of Florida.
- Jansen, E. P. W. A., Joostens, T. H. & Kemper, D. R. (2004) Enquêteren, Het opstellen en gebruiken van vragenlijsten. Groningen: Noordhoff Uitgevers
- Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education*., 39, 840-852. doi:10.1080/02602938.2013.875117
- Jonsson, A., Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Education Research Review*, 2, 130-144. doi:10.1016/j.edurev.2007.05.002
- Joosten- ten Brinke, D. (2011). Eigentijds toetsen en beoordelen (lectorale rede). Geraadpleegd op <http://fontys.nl/web/file?uuid=9294b4fb-3152-47e3-8226-91b0f0603962&owner=458011ba-35e2-4ec8-bae3-ef97620e85c6&contentid=75208>
- de Jong, B. (1981). Correctievoorschriften voor de beoordeling van open vragen en interbeoordelaarsbetrouwbaarheid. *Tijdschrift voor onderwijsresearch*, nr. 5
- Kim, S. (2010). The Influence of Likert Scale Format on Response Result, Validity, and Reliability of Scale -Using Scales Measuring Economic Shopping Orientation. *Journal of the Korean Society of Clothing and Textiles*, Vol. 34, No. 6 p 913-927.
- Kish, L. (1968) *Survey sampling*. John Wiley & Sons Inc., New York.
- Koch, G. (1982). Intraclass correlation coefficient. *Encyclopedia of Statistical Sciences*. 4. New York: John Wiley & Sons. pp. 213–217.
- Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*.
- Lievens, F., Van Keer, E. (1999). Assessment centers in Vlaanderen : een wetenschappelijke evaluatie. Gent: Academia press.
- Nederlands instituut van Psychologen (2010). Algemene Standaard Testgebruik. Nederlands instituut van Psychologen. Geraadpleegd op <https://www.psynip.nl/wp-content/uploads/2016/09/Algemene-Standaard-Testgebruik.pdf>
- NVAO. (2014). Beoordelingskaders accreditatiestelsel hoger onderwijs Nederland. Geraadpleegd op https://www.nvao.net/system/files/procedures/Beoordelingskaders%20accreditatiestelsel%20NL%2019%20december%202014_0.pdf

- Pontius, R., Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*.
- Pufpaff, L. A., Clarke, L., Jones, R. E. (2015). The effect of Rater Training on Inter-Rater Agreement. *Mid-Western Educational Researcher*, vol. 27
- Rhebergen, D., Rouwenhorst, A., Comijs, H. C., Dols, A., Beekman, A. T. F., Terwee, C. B., Arts, D., Stek, M. L.(2011). De Interbeoordelaarsbetrouwbaarheid van de Nederlandse versie van de CORE; een studie bij opgenomen ouderen met depressie. *Tijdschrift voor de psychiatrie* 51.
- Rodger, S., Coleman, A., Caine, A., Chien, C., Copley, J., Turpin, M., Brown, T. (2014). Examining the inter-rater and test-retest reliability of the Student Practice Evaluation Form-Revised (SPEF-R) for occupational therapy students. *Australian Occupational Therapy Journal*.
- Rui, N., Feldman, J. (2012). Inter-rater reliability of a COP – a critical appraisal. *Us-China education review* B 3. David Publishing.
- Sahadat, I. (2010, 10 juli). Studenten kregen met truc diploma mee. *de Volkskrant*. Geraadpleegd op <http://www.volkskrant.nl/>
- Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3), 321-325
- Shrout, P. E., Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 86(2): 420-428.
- Sudman, S. (1976). *Applied Sampling*. New York: *Academic Press*.
- Trochim, W.M.K. (2006). *Research methods: the essential knowledge base*. 2nd edition. ISBN13: 978-1-305-47649-3
- Verhoeven, N. (2010). *Wat is onderzoek? Praktijkboek methoden en technieken voor het hoger onderwijs*. Boom onderwijs, Den Haag.
- de Vet, H. C., Terwee, C. B., Knol, D. L., Bouter, L. M.(2006). When to use agreement versus reliability measures. *J Clin Epidemiol* 59(10): 1033-1039.
- Wang, P. (2009). The Inter-Rater Reliability in Scoring Composition. *English Language Teaching*, 2(3), 39-43.
- Westenberg, C. (2016). Verantwoord Toetsen en Beoordelen met een Rubric: een Onderzoek naar het Mediërend Effect van Interbeoordelaarsbetrouwbaarheid op de Relatie tussen het Toepassen van een Rubric en Studenttevredenheid. Open Universiteit Nederland. Geraadpleegd op <http://dspace.ou.nl/bitstream/1820/7369/1/OWCWestenberg-Wiersum-10112016.pdf>

Bijlagen

Bijlage 1 Antwoorden op vragenlijst schriftelijke gedeelte

Respondent nummers zijn op volgorde van het export bestand van de vragenlijst. Onder de vragen staan de labels(L) om het antwoord onder te kunne brengen in een thema.

V1=Wat zou volgens u beter kunnen bij het beoordelen van het schriftelijke gedeelte?

V2= Zou u de beoordelingen van het schriftelijke gedeelte anders aanpakken? Wat zou u dan anders doen?

V3= Wat zou volgens u beter kunnen bij het beoordelen van het eindgesprek?

V4= Zou u de beoordelingen van het eindgesprek anders aanpakken? Wat zou u dan anders doen?

Respondent 1:

V1: Ik vind dit eigenlijk wel een prima formulier.

(L: Prima)

V2: nee, ik vind dit een prettig formulier om mee te werken.

(L: Prima)

V3:Bij het eindgesprek is er meer een algemene beoordeling waar sprake van is. Ik persoonlijk gebruik de criteria meer als kapstok om het cijfer aan op te hangen, niet zozeer als maatstaf om het cijfer te bepalen.

(L: Flexibiliteit)

V4: misschien wat minder criteria en meer het onderbouwen van een cijfer ipv proberen met kleine criteria tot een cijfer te komen. bij een gesprek werkt dat toch anders dan een opdracht op papier

(L: Minder criteria, opbouwend)

Respondent 2:

V1: Suggestieve vraag: hier impliceer je dat het formulier hoe dan ook verbeterd moet worden. , Verstandig is om neutraal te vragen wat men ervan vindt: wat goed is en wat niet goed is., Met je type vragen stuur (bias) je de mening van de respondenten en dat is niet juist. , Als onderzoeker dien je neutraal en objectief te zijn. , , Inhoudelijk opmerking: een beoordelingsformulier moet transparant en objectief zijn. Het huidige formulier doet een poging om dit mogelijk te maken.

(L: Transparantie en objectiviteit)

V2: Korter en bondig... maar het risico hiervan is dat de beoordeling minder transparant en objectief wordt. Voor zo iets belangrijks heb je veel informatie nodig om je beoordeling te rechtvaardigen. Denk bv aan de intelligentietest, deze test bevat niet alleen 10 items, of wel?

(L: Korter en bondiger)

V3: Weer twee suggestieve vragen. , Inhoudelijk: er moet meer aandacht besteed worden aan beroepsproducten.

(L: Meer aandacht beroepsproducten)

V4: geen idee.

Respondent 3:

V1: De criteria zijn niet altijd even duidelijk geformuleerd, dit kan ervoor zorgen dat beoordelaars het heel anders interpreteren. Het probleem is alleen dat je het ook niet te specifiek wilt maken, want dan wordt het voor de student meer een invuloefening dan dat ze zelf kritisch na moeten denken. Het zou nuttig kunnen zijn om de beoordelingscriteria specifiek te maken op sommige onderdelen en deze alleen beschikbaar te maken voor docenten.

(L: niet duidelijk(Balans tussen specifiek en algemeen)

V2: Het zou nuttig kunnen zijn om de beoordelingscriteria specifiek te maken op sommige onderdelen en deze alleen beschikbaar te maken voor docenten.

(L: Balans specifiek en algemeen)

V3: Ik heb nog geen eindgesprekken gevoerd, dat ga ik eind dit kwartiel voor het eerst doen. Ik kan hier dus niets zinnigs over zeggen, sorry!

V4: Ik heb nog geen eindgesprekken gevoerd, dat ga ik eind dit kwartiel voor het eerst doen. Ik kan hier dus niets zinnigs over zeggen, sorry!

Respondent 4:

V1: 1. Minder beoordelingspunten per onderdeel (m.n. Minder per hoofdstukparagraaf), waardoor het nu erg op detail niveau blijft ipv op grotere geheel., 2. Ook andere aspecten meenemen, zoals samenvatting, schrijfstijl, soort bronnen, creativiteit), 3. Meer op inhoud ipv of het aspect aanwezig is (meer op kwaliteit dan op kwantiteit)

(L: Te specifiek, meer criteria/onderdelen, kwaliteit ipv kwantiteit)

V2: Kortere beoordelingsformulieren

(L: korter)

V3: Niets

V4: Automatische berekening van eindcijfer (bijv formulier laten invullen via tablet van 1e beoordelaar)

(L: digitale formulieren)

Respondent 5:

V1: Andere wijze van cijferen., Onderscheid maken tussen kwalitatief - en kwantitatief onderzoek., Eenduidig erop wijzen dat het punten van aandacht zijn voor de hoofdstukinhoud maar dat deze aandachtspunten (zoals hypothesen bij hoofdstuk 2) ook in andere hoofdstukken (bv. in geval van hypothesen) in hoofdstuk 3 aan bod kunnen komen., Bij beschrijvende onderzoeken zijn m.i. conceptuele modellen overbodig.

(L: aanpassen op soort onderzoek, flexibiliteit)

V2: ook beoordelen van, -samenvatting, -kwanti- en kwaliteit van de bronnenlijst

(L: meer criteria/onderdelen)

V3: alles moet in zeer korte tijd beoordeeld worden - meer tijd

(L: meer tijd)

V4: zie hierboven

Respondent 6:

V1: We zijn al toe aan de tweede generatie onlineformulieren. De tekst van het item is helder, echter de antwoordcategorieën vormen een 4-puntsschaal. Een 5-puntsschaal waarbij het midden staat voor het rapportcijfer 5.5 is betrouwbaarder.

(L: schalen aanpassen)

V2: Het aantal items kan omlaag als iemand (JIJ?) met een factor -of homogeniteitsanalyse in de weer gaat. , , , Een 5-puntsschaal waarbij het midden staat voor het rapportcijfer 5.5 is betrouwbaarder

(L: schalen aanpassen, minder items)

V3: en compleet ander formulier met andere criteria en andere weging. Nu is de weging 20-70-10, dat oogt al meteen als een bron van onbetrouwbaarheid.

(L: andere weging)

V4: Criteria betreffende het niveau van argumenteren, parate kennis, oordeelsvermogen, out-of-the-box denken, verbanden leggen met maatschappelijke ontwikkelingen, verbanden leggen met ontwikkelingen in de psychologie, verbanden leggen met ontwikkelingen in de beroepspraktijk., , Verder moet de student van minimaal een aanbeveling een beroepsproduct kunnen aanprijzen met de criteria, kwaliteit, kosten besparing, effectiviteit, verwachte opbrengst, wijze van implementeren,

samenwerking met welke instanties en last-but-not-least, helpt dit product het leven van een of meer mensen op te vrolijken?
(L: meer criteria/onderdelen)

Respondent 7:

V1: Ik snap de tweede vraag hierboven niet., , Te detaillistisch, te arbeidsintensief en dwingt zowel de docent als de student om het creatieve product dat een onderzoek eigenlijk is een one size fits all keurslijf te wurmen. Ik zou graag een formulier hebben dat korter is, meer ruimte biedt voor afwijkingen van de standaard en waarin ik bovendien zaken als daadkracht, snelheid, creativiteit, zelfstandigheid en zelfredzaamheid kan belonen..

(L: te specifiek, meer criteria/onderdelen, korter)

V2: Niet zon keurslijf, minder details en ik zou niet afdwingen dat bij elk onderdeel commentaar ingevuld moet worden.

(L: te specifiek, flexibiliteit)

V3:Ik weet eigenlijk op dit moment niet meer zo heel precies wat er in dat formulier staat. Daarom alles op neutaal gezet. Ik vind ze meestal vrij lang.

(L: te lang)

V4:Ben ik wel tevreden over, vooral her feit dat de student snel duidelijkheid krijgt vind ik prettig. , Ik zou liever het hele fenomeen eindgesprek ter discussie willen stellen dan de manier waarop het beoordeeld wordt

(L: prima, eindgesprek in geheel ter discussie)

Respondent 8:

V1: Ik zou graag veel minder criteria willen, dus minder punten wat ik moet beoordelen. Daarnaast zou ik graag cijfers geven per onderdeel. Ook zou ik graag het proces en functioneren van de student willen beoordelen.

(L: minder criteria/onderdelen wel andere criteria)

V2: 7 cijfers geven per scriptie. 1 per hoofdstuk, plus 1 voor schrijfvaardigheid en 1 voor functioneren

(L: meer onderdelen beoordelen)

V3: weet ik niet

V4: geen idee

Respondent 9:

V1: Het probleem is dat bij een goede scriptie je eigenlijk heel snel weet dat het goed is. Echter, het is mogelijk dat omwille van de punten op het formulier, er dan een lagere cijfer uitkomt, juist omdat de student niet een copy van het beoordelingsformulier als basisstructuur voor de scriptie heeft gebruikt. Het omgekeerde kan ook, een matige scriptie voldoet aan alle punten. Het is een kwestie van vertrouwen, een goed eindproduct kan vanaf een redelijke onderzoeker (beoordelaar) prima worden herkend. Topproducten en slechte producten leiden altijd tot verwarring.

(L: gevoelsmatige beoordeling)

V2: Meer vertrouwen geven aan de docenten, en bij blind beoordelen met meer dan 1 punt verschil tussen de 3 beoordelaars altijd een externe board raadplegen.

(L: Vertrouwen op blinde beoordeling)

V3:Eigenlijk geldt hier hetzelfde als ik bij de scriptie heb geschreven. Echter, de gespreksvaardigheden van de beoordelaars vormen een onmiskenbaar onderdeel van het gesprek en dat maakt het wel subjectiever, dus in feite heb ik hier iets minder moeite met de handreikingen....

V4:Nog meer op hoofdlijnen handreikingen geven: precisie is schijnprecisie: beoordelen van een gesprek is geen exacte wetenschap

(L: te specifiek)

Respondent 10:

V1: Per subblok zou een commentaar gedeelte moeten komen, , Bovendien moet beter worden uitgelegd wat het doel van het verstrekken van deze gegevens is.

(L: meer commentaar mogelijk maken, uitleg)

V2:

V3: Er zou beter moeten worden nagedacht over de inhoud van het gesprek en de beoordeling. De eisen die aan de presentatie worden gesteld zijn niet voor iedereen duidelijk., , Omdat soms de presentatie anders is dan de scriptie komt, deze eigenlijk niet meer goed aan bod. De mondelinge verdediging zou toch het accent moeten hebben in het eind gesprek, , De behandeling van de stellingen komen niet aan bod

(L: geen duidelijke eisen)

V4: zie boven

(L: stellingen zijn belangrijkst

)

Respondent 11:

V1: Ik ben tevreden over de huidige opzet. Sommige criteria kunnen nog eens opnieuw onder de loep worden genomen, maar ik vind het voor 99% in orde.

(L: prima)

V2: Nee, ben tevreden over huidige opzet.

(L: prima)

V3: Gewicht toekennen aan afzonderlijke criteria. Je kunt nu zelf bepalen welk gewicht je toekent aan een criterium dat onder een bepaald onderdeel valt.

(L: geen vaststaande weging)

V4: zien hierboven

(L: vaststaande weging)

Respondent 12:

V1: De vragen lijken een absoluut antwoord te vragen. MAAR: het is vaak niet ja of nee aanwezig. Wat ik vooral mis is een niveau-aanduiding. Dus niet alleen ja of nee. , Voorbeeld:, 1. het probleem/de vraag is.....ervan zijn. kan weliswaar duidelijk en overtuigend zijn, maar is daarmee nog niet van voldoende niveau., 2. De mogelijkevakliteratuur. Is geen vraag bv naar volledigheid of meest relevante literatuur., , Kortom: dit beoordelingsformulier past meer bij het vak onderzoek/statistiek. Een scriptie vergt meer en overstijgende informatie en analyse.

(L: mist niveau aanduiding, meer onderdelen)

V2: Elk item in twee stappen beoordelen: a) of iets aanwezig is en b) zo ja: is het van voldoende niveau? Of Is het van de meest relevante kanten bezien?, Geef ook ruimte voor extra beoordelingsaspecten bv als gerapporteerd wordt over een gegeven presentatie op een congres.

(L: twee stappen beoordeling, meer onderdelen/criteria)

V3: nvt

V4: nvt

Respondent 13:

V1: Niet alle beoordelingscriteria zijn helder geformuleerd., Het werken met 4 keuze alternatieven werkt helemaal niet fijn. , Er kunnen alleen hele of halve cijfers voor een hoofdstuk gegeven worden, graag ook verder achter de komma

(L: niet duidelijk, meer schalen, alleen hele of halve cijfers)

V2: Ja, werken met helderder alternatieven en per criterium met cijfers werken.

(L: duidelijker en meer cijferen)

V3:de neutrale antwoorden van hierboven aanpassen

V4:In een bestand werken die meteen met formules het eindcijfer berekend

(L: digitaal formulier, makkelijke berekening)

Respondent 14:

V1: Ik vond de hoeveelheid cijfers die gegeven moet worden niet logisch. Vooral bij hoofdstuk 3 moeten er te veel deelcijfers geven worden. , , Bij H5 mis ik aandacht voor het probleem van de opdrachtgever en in welke mate er nu is bijgedragen een de oplossing van dit probleem., , Daarbij mis ik criteria over de samenhang tussen hoofdstukken. Sluit aan bij de opmerking hiervoor.

(L: te veel cijferen, samenhang ontbreekt, beroepsproduct)

V2: Zie hierboven

V3:-

V4:-

Respondent 15:

V1: De rol van een externe deskundige is nu niet geheel duidelijk. Is het adviserend of beoordelend? Meestal wordt het eindcijfer van de externe deskundige meegewogen in het gemiddelde. Maar niet altijd. Bijvoorbeeld, als de 2 docenten een voldoende geven en de externe deskundige een onvoldoende dan mag de student wel op gesprek komen en wordt het cijfer van de externe niet meegewogen. , , Voor een externe deskundige is het soms moeilijk een scriptie te beoordelen omdat de externe geen kennis heeft van het afstudeerproces en de eventuele verbeteringen die gemaakt zijn in de scriptie de afgelopen periode. Mijn inziens is het juist goed dat de externe alleen naar het eindresultaat kijkt. Dat is immers ook zijn of haar taak. Maar er ontstaat toch regelmatig wat onenigheid tussen docenten en de externe, met name als de externe de scriptie als enige met een onvoldoende beoordeeld. Uiteindelijk mag na overleg tussen de eerste beoordelaar en externe de student altijd op gesprek komen. Dit voelt niet altijd goed. Je hebt een onvoldoende gegeven en hier wordt in feite niks mee gedaan en zelfs niet eens meegewogen. , , Overigens ervaar ik het contact met de docenten als prettig en heb ik het gevoel dat de externe met respect wordt behandeld.

(L: rol externe beoordelaar duidelijker)

V2: Ik zou de rol van de externe geheel adviserend maken, wel geschreven feedback laten geven op de scriptie middels een standaard formulier. De externe geen eindcijfer laten geven. Dus alleen feedback zonder een eindcijfer. , , Of de externe tussentijds mee laten kijken. Misschien is het mogelijk dat een student tussentijds feedback ontvangt van een externe.

(L: externen alleen adviseren geen cijfer)

V3: Het beoordelingsformulier is prima. Echter het hanteren van het formulier verloopt niet altijd op dezelfde wijze. , , Soms wordt na afloop van het eindgesprek eerst overlegd en vullen we alle 3 een formulier in met precies dezelfde cijfers. , , Soms wordt na afloop van het eindgesprek door ons alle drie individueel een formulier ingevuld. Daarna gaan we overleggen wat we hebben ingevuld. En ontstaan er zelfs discussies.

(L: manier van invullen niet altijd gelijk)

V4: Ik zou ervoor kiezen om na afloop van het eindgesprek met de drie beoordelaars te overleggen (wat vinden we ervan) en dan gezamenlijk 1 formulier invullen.

(L: een ingevuld formulier voor alle beoordelaars na overleg)

Respondent 16:

V1: Er zit één zwak punt. Er staat: 23. Bij kwalitatief onderzoek is gebruik gemaakt van semigestructureerd of gestructureerde interviews. In geval van een literatuuronderzoek: de zoektermen en/of de zoekstrategieën zijn correct beschreven. , , Probleem is dat je hier een bolletje moet inkleuren, ook als het alleen een kwantitatief onderzoek betreft., Wat ik ook mis: een beoordeling van de mate van 'toegepaste psychologie' in een onderzoek. Ik heb net een onderzoek TP beoordeeld dat de mate van klanttevredenheid van cliënten over hun fysiotherapeut betrof. Dat

had ook het onderzoek van een marketingstudent kunnen zijn. Ik kan mijn frustratie hierover nergens in de beoordelingslijst kwijt.

(L: formulier flexibel maken, ruimte voor commentaar algemene zaken)

V2: Zie boven.

V3: Eindgesprek vind ik altijd een soort 'feestje', (tenzij de kandidaat natuurlijk erg gespannen is of slecht presteert). , Wel vind ik het gerommel met papieren bij de beoordeling nogal omslachtig. Iedereen moet z'n handtekening 6x zetten. Er moet veel worden berekend. Dat moet toch efficiënter kunnen.?

(L: digitaal formulier, makkelijker berekenen, te veel papier)

V4: zie boven

Respondent 17:

V1: De criteria zijn vaak onduidelijk hoe ze van toepassing zijn op de onderdelen. Het is wat mij betreft veel te gefragmenteerd met een kennelijk onderliggend idee dat 'meten is weten', maar dat komt de duidelijkheid van becijfering niet ten goede.

(L: te specifiek)

V2: Veel minder onderdelen en veel meer doelgericht op criteria. Nuances komen bij het gesprek wel aan bod., Daarnaast zijn een aantal criteria wat mij betreft zeer geforceerd, met name de eis om een 'conceptueel model' van de probleemstelling te formuleren. Dit levert bijna nooit de toegevoegde waarde die je ervan zou verlangen.

(L: minder criteria/onderdelen)

V3: Ik heb het beoordelingsformulier van het eindgesprek niet helemaal helder in mijn hoofd maar ik kan me wel herinneren dat het een stuk bevredigender werken is: sneller en dekkend voor de criteria.

(L: snel en efficiënt)

V4:-

Respondent 18:

V1: Het is mogelijk te beoordelen met --, -, + en ++. Welke cijfers daaraan gekoppeld zijn, is niet goed duidelijk. Soms geef ik bij drie keer een + een 7, soms een 8., , Niet alle beoordelingscriteria zijn even relevant.

(L: niet duidelijk)

V2: Boven genoemde punten aanpassen

(L: duidelijker maken)

V3: Het eindgesprek loopt niet volgens vaste richtlijnen. Het presentatiegedeelte is wel duidelijk. Hoe het daarna loopt, wie wat vraagt, of en zo ja, welke stellingen aan bod komen wisselt en wat we verwachten van studenten kan eenduidiger.

(L: geen vast patroon, onduidelijk)

V4: Bovenstaande punten aanpassen

(L: vast patroon, duidelijker maken)

Respondent 19:

V1: Bij de vragen over het Theoretisch kader kan wat mij betreft ook gekeken worden naar hoe de literatuur op onderwerp en inhoud gekozen is. > Aan de manier waarop de student met theorie en het verzamelen/herkennen van relevante info bezig is, kun je veel herleiden over 1. de mate van inzicht die student over het onderwerp heeft 2. de capaciteit om te redeneren, door te denken en een 'denklijn' te vinden/op te bouwen die aanleiding is voor scheppen van lijn én samenhang in de theoretische onderbouwing. Hierin kun je tevens de eigenheid vd student en zijn scheppend vermogen herkennen., , Dus theoretisch kader ook toetsen op: (denk)lijn, samenhang en het kunnen

herkennen van de vragen die het thema oproept/herkennen van relevante theorie die daar op aan sluit.

(L: meer criteria/onderdelen)

V2: Misschien zou in nog duidelijker een onderscheid willen maken tussen 1. het toetsen van abstractievermogen van de student & tot inzicht kunnen komen/ (nieuwe) beelden kunnen vormen 2. het toetsen over het vermogen om een juiste opzet te maken van onderzoek 3. het toetsen van de wiskundige vaardigheden die bij kiezen/toepassen van SPSS, Cronbach alfa, convergentie van onderzoek nodig zijn.

(L: meer criteria/onderdelen)

V3: Volgens mij blijft het enorm lastig om tot voldoende objectiviteit te komen, zeker voor de docenten die een student al langer kennen en meer of minder sympathie bij een student ervaren. Ik denk dat er misschien iets meer te objectiveren valt wanneer elke beoordelaar op een vast aantal punten vragen scores invult, of een soort Likert schaal?, Ik heb door de jaren diverse docenten meegemaakt die de leiding gaven bij het eindgesprek. Sommigen houden zich strak aan de structuur, anderen veel minder. Ik zou hier een meer eenduidige aanpak in willen zien.

(L: vast patroon ontbreekt)

V4: Ik zei hier al iets over in het vorige punt., Ik zou ook een wat objectievere manier zoeken om bij studenten die 'niet lekker in hun vel zitten' op dat moment en zo 'niet uit de verf komen' toch hun potentieel te kunnen herkennen én waarderen.

(L: vast patroon)

Respondent 20:

V1: Het huidige formulier legt de nadruk op de formele kanten van de samenstelling van een scriptie. Het kan zijn dat de student deze formele criteria aanvinkt in zijn scriptie en dan kun je als beoordelaar niet veel meer doen dan 'voldaan'. Een scriptie/onderzoeksverslag dient een logisch verhaal te zijn en dan komen lang niet altijd al die punten aan de orde.

(L: flexibiliteit)

V2: De beoordeling zou veel meer summier en minder geformaliseerd moeten zijn. Onderzoek is een creatief proces en dan heb je aan enkele hoofdpunten voldoende. Voor de beoordeling is het uitstekend om het huidige rapportageprotocol aan te bieden (maakt het ook gemakkelijker in de beoordeling) maar je moet er van kunnen afwijken, Goed onderzoek is mijn beoordeling op: 1:Wat is de vraag/het probleem van de opdrachtgever. 2:Hoe is dit in een onderzoeksvraag te vertalen (dan ook aangeven waar in de literatuur dergelijke vragen beantwoord worden, keuze van instrumenten, link met literatuur). 3:Hoe heb je dat uitgevoerd. 4:Wat waren de resultaten. 5:In hoeverre heb je de vraag beantwoord. 6:Wat is de aansluiting op bestaande kennis (literatuur).7: Hoe moet de opdrachtgever verder. Het wijkt niet veel af van wat er nu gebeurt maar door slechts de zeven kopjes te geven, geef je ruimte aan het creatieve proces dat onderzoek hoort te zijn. De beoordelaar wordt dan ook gedwongen tot lezen en beoordelen en niet tot turven

(L: flexibiliteit, te specifiek)

V3: Mijn ervaring tot nu toe is dat het eindgesprek een sterk halo effect kent. Een student doet het op een bepaalde manier en de cijfers richten zich op deze algemene beoordeling. Ik denk dat vooral 'overtuigingskracht' bepalend is in het eindgesprek. En dat is eigenlijk wel goed. Verder is mijn ervaring dat docenten ad hoc vragen stellen, er zit weinig strategie in. Ook is het voorgekomen dat een gezellige babbel ontstaat en dat is ook niet de bedoeling (lijkt mij). De entourage is echter weinig realistisch voor de beroepspraktijk.

(L: te vast patroon)

V4: je zou kunnen experimenteren met het splitsen van het eindgesprek in een openbaar deel, waarin de student voor een publiek van medestudenten zijn scriptie verdedigt (en ook vragen beantwoord), en een gesloten deel met de examinatoren (waar nog even doorgevraagd wordt). Dat is realistischer gezien de beroepspraktijk waar je ook voor een meute

fans/ongelovigen/onverschilligen/enthousiasten gezet wordt en vervolgens met de opdrachtgever nog even om tafel komt zitten.

(L: aanpassen concept eindgesprek)

Respondent 21:

V1: Ik vind dit eigenlijk wel een prima formulier.

(L: prima)

V2: Betrokkenheid van externe beoordelaar bij de opzet van het onderzoek is wenselijk.

(L: externen betrekken bij opzet)

V3:-

V4:-

Respondent 22:

V1: -

V2: -

V3: -het zou niet mogelijk moeten zijn dat een student met een blaadje in zijn handen staat en daar niet op afgerekend wordt in het formulier., De stellingen hebben een vreemde plek, en er hangt veel gewicht aan.

(L: stellingen wegen te zwaar)

V4: Soms ligt het accent te sterk op het onderzoekstechnische deel. , Aandacht voor visie, eigen mening van de kandidaat waarin hij/zijn blijkt geeft van overstijgend kunnen denken, verbanden kunnen leggen, zaken in de context kunnen plaatsen, gericht op de beroepspraktijk.

(L: meer criteria/onderdelen)

Bijlage 2 De vragenlijst

Kom
verder



Scriptiebeoordelingen Saxion

Pagina 1

Beste Beoordelaar,

De afgelopen jaren heeft u een of meerdere scripties beoordeeld van Toegepaste Psychologie studenten aan de Academie Mens & Arbeid van de hogeschool Saxion te Deventer. Op dit moment is een onderzoek gaande om de betrouwbaarheid tussen alle beoordelaars te bepalen. Daarnaast willen wij heel graag weten wat de beoordelaars van dit proces vinden en hoe ze dit proces hebben ervaren.

De vragenlijst bestaat uit een aantal meerkeuzevragen en enkele open vragen. De schalen zullen per vraag uitgelegd worden. Voor de open vraag wordt u verzocht om de vraag zo volledig en duidelijk mogelijk in te vullen. Daarnaast is de vragenlijst opgedeeld in twee delen. Het eerste deel zal gaan over de beoordeling van het schriftelijke gedeelte. Het tweede deel zal gaan over het eindgesprek. De vragenlijst is binnen 10 minuten in te vullen.

Pagina 2

Welk type beoordelaar bent u? *

- ☐ 1e/2e beoordelaar
- ☐ externe beoordelaar

Wat is uw geslacht? *

- ☐ Man
- ☐ Vrouw

Wat is uw leeftijd? *

Jaar

Hoe lang beoordeelt al u afstudeerprojecten voor de opleiding Toegepaste Psychologie binnen Saxion? ⓘ

Jaar

Pagina 3

De volgende vragen zullen gaan over de beoordeling van het schriftelijke gedeelte van de scriptie (het rapport).

Welk cijfer geeft u voor de kwaliteit van de beoordelingsformulieren voor het schriftelijke gedeelte van de scriptie? *

Geef uw cijfer

Geef op de schalen aan in hoeverre u het eens bent met onderstaande stellingen *

	Helemaal oneens	oneens	neutraal	eens	helemaal eens
De kwaliteit van de beoordelingscriteria van het schriftelijke gedeelte is goed	☐	☐	☐	☐	☐
De manier van cijferen van het schriftelijk gedeelte is duidelijk en logisch	☐	☐	☐	☐	☐
Ik ben tevreden over de opzet van de beoordeling van het schriftelijke gedeelte.	☐	☐	☐	☐	☐
De beoordelingsformulieren voor het schriftelijke gedeelte zijn eenduidig geformuleerd	☐	☐	☐	☐	☐
De beoordelingsformulieren voor het schriftelijke gedeelte zijn van zodanige kwaliteit dat het beoordelen makkelijker wordt	☐	☐	☐	☐	☐

Wat zou volgens u beter kunnen bij het beoordelen van het schriftelijke gedeelte?

Gelieve de vraag in volledige zinnen te beantwoorden.

Zou u de beoordelingen van het schriftelijke gedeelte anders aanpakken? Wat zou u dan anders doen? ⓘ

Pagina 4

De volgende vragen gaan over Het beoordelen van het eindgesprek.

Welk cijfer geeft u voor de kwaliteit van de beoordelingsformulieren van het eindgesprek? *

Geef uw cijfer

Geef op de schalen aan in hoeverre u het eens bent met onderstaande stellingen *

	Helemaal oneens	oneens	neutraal	eens	helemaal eens
De kwaliteit van de beoordelingscriteria van het eindgesprek zijn goed	☐	☐	☐	☐	☐
De manier van cijferen voor het eindgesprek is duidelijk en logisch	☐	☐	☐	☐	☐
Ik ben tevreden over de opzet van de beoordeling van het eindgesprek.	☐	☐	☐	☐	☐
De beoordelingsformulieren voor het eindgesprek zijn eenduidig geformuleerd	☐	☐	☐	☐	☐
De beoordelingsformulieren voor het eindgesprek zijn van zodanige kwaliteit dat het beoordelen makkelijker wordt	☐	☐	☐	☐	☐

Wat zou volgens u beter kunnen bij het beoordelen van het eindgesprek?

Gelieve de vraag in volledige zinnen te beantwoorden.

Zou u de beoordelingen van het eindgesprek anders aanpakken? Wat zou u dan anders doen? ⓘ

» [Redirection to final page of Enquêtes Maken \(wijzig\)](#)

Bijlage 3 Het beoordelingsprotocol

Bijlage 4 Beoordelingsprotocol mondelinge gedeelte (het eindgesprek)

Beoordelingsformulier eindgesprek TP¹

Naam student

Onderwerp scriptie

Studentnr.

Datum

Beoordelings-aspect	Indicatoren	Weging	Cijfer 1-10
Presentatie	- Het doel van de presentatie is duidelijk - De presentatie voegt iets aan de scriptie toe in de zin van inhoudelijke verbreding en/of verdieping van het afstudeeronderwerp - De opbouw van de presentatie heeft een logische en overzichtelijke ordening - De student maakt op overtuigende wijze contact met zijn toehoorders met adequaat gebruik van stemvolume en tempo. - De student geeft correcte antwoorden op vragen en gaat, daartoe uitgenodigd, de discussie aan o.b.v. correcte argumenten. - De gebruikte bronnen worden volgens de APA-regels op dia gepresenteerd en zijn van relevant niveau.	20%	
Inbreng student tijdens het gesprek	- De student staat duidelijk 'boven de stof' - De student geeft duidelijk en 'to the point' antwoord op de vragen - De student weet zijn antwoorden goed te onderbouwen - De student toont reflectieve vaardigheden t.a.v. de scriptie en noemt verbeterpunten. - De student weet zijn onderwerp duidelijk te linken aan relevante theorieën en maatschappelijke ontwikkelingen. - De student heeft een duidelijke visie op de TP-professies en op zijn eigen rol binnen het TP-werkveld.	70%	
De stellingen	- De stellingen zijn 'uitdagend', geen open deuren - De student weet zijn mening t.a.v. de stellingen goed onderbouwd te verdedigen	10%	
EINDCIJFER GESPREK			

Handtekening 1^e beoordelaar:

Handtekening 2^e beoordelaar:

¹ Eindgesprek moet minimaal een 5.5 zijn, wil de student slagen.