Universiteit
Leiden
The Netherlands

Design and evaluation of video portfolios : reliability, generalizability, and validity of an authentic performance assessment for teachers

Bakker, M.E.J.

Citation

Bakker, M. E. J. (2008, December 2). *Design and evaluation of video portfolios : reliability, generalizability, and validity of an authentic performance assessment for teachers*. *ICLON PhD Dissertation Series*. Leiden University Graduate School of Teaching (ICLON). Retrieved from https://hdl.handle.net/1887/13353

Version:	Corrected Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/13353

Note: To cite this publication please use the final published version (if applicable).

Nederlandse samenvatting

Hoofdstuk 1

In het eerste hoofdstuk van het proefschrift worden achtergrond, probleemstelling en onderzoeksvragen, context en relevantie van het onderzoek gepresenteerd. De ontwikkeling van instrumenten voor het beoordelen van docentcompetenties staat volop in de belangstelling. Uit onvrede met bestaande procedures, worden momenteel nieuwe beoordelingsprocedures ontwikkeld, ook wel 'authentieke performance assessments' genoemd. Een belangrijk kenmerk van deze beoordelingsprocedures is dat ze beogen recht te doen aan het complexe en contextgebonden karakter van lesgeven. In de nieuwe beoordelingsprocedures wordt veelal een mix van bewijsbronnen gebruikt die de verschillende componenten van het lesgeven bestrijken. Ook worden open taken ingezet die een beroep doen op het onmiddellijk en adequaat beslissen en handelen in de praktijk of in een context die vergelijkbaar is met de praktijk. Bij het ontwikkelen van nieuwe beoordelingsprocedures gaat ook de aandacht uit naar het waarborgen en evalueren van de kwaliteit van deze procedures. De nieuwe vormen van beoordelen brengen immers nieuwe bedreigingen van de betrouwbaarheid en validiteit met zich mee. Ten eerste spelen bij performance assessments beoordelaars een belangrijke rol; zij moeten de performance (het functioneren) van de docent interpreteren en beoordelen. Het blijkt dat het voor beoordelaars lastig is om objectief en betrouwbaar te scoren, omdat persoonlijke voorkeuren, vooroordelen en selectieve observatie moeilijk te vermijden zijn. Ten tweede wordt de validiteit van de performance assessments bedreigd door taakspecificiteit. De taken die in een assessment zijn opgenomen, blijken vaak aanzienlijk wisselende performances op te roepen bij respondenten, zelfs wanneer de taken uit eenzelfde domein komen. Ten derde is het moeilijk om een representatieve steekproef van assessmenttaken samen te stellen die alle relevante situaties en aspecten van lesgeven omvatten die docenten in de praktijk tegen kunnen komen.

Om die bedreigingen te reduceren, worden in de literatuur verschillende maatregelen aangedragen die in het design van de beoordelingsprocedure zouden kunnen worden opgenomen. Deze maatregelen bestaan uit het gebruiken van gepaste criteria en performanceniveaus, het gebruiken van een scoringsprocedure waarbij beoordelaars zorgvuldig een beoordeling opbouwen aan de hand van specifieke criteria en richtlijnen, het inzetten van meerdere beoordelaars, het gebruiken van een systematische en transparante scoringsprocedure, het trainen van beoordelaars in het

toepassen van de criteria en performanceniveaus en het standaardiseren van assessmenttaken. Hoewel gaandeweg een kennisbasis ontstaat over het ontwerpen van authentieke performance assessments, blijft het relatief ingewikkeld de ontwerpprincipes uit de literatuur om te zetten in een concrete beoordelingsprocedure. In dit proefschrift wordt op basis van ontwerpprincipes uit de literatuur een performance assessment ontwikkeld en geëvalueerd. Het proefschrift levert daarmee een bijdrage aan de kennisbasis met betrekking tot het realiseren van betrouwbare, generaliseerbare en valide performance assessments.

Het performance assessment dat in dit onderzoek is ontwikkeld, werd ingezet voor het beoordelen van de coachcompetentie van docenten in het MBO, met andere woorden, de competentie van MBO-docenten in het coachen van hun leerlingen die bezig zijn met een opdracht. De beoordelingsprocedure is speciaal voor deze docentcompetentie ontworpen, omdat dit een belangrijke competentie is geworden door de implementatie van zelfstandig en competentiegericht leren in het MBO. In de context van deze innovatie is in de regio Leiden en omstreken binnen de sector Techniek het MTS+ project gestart. Binnen het MTS+ project is een leeromgeving ontwikkeld die moet bijdragen aan zelfstandig en competentiegericht leren. In deze leeromgeving is het curriculum georganiseerd rond complexe en langlopende opdrachten die sterk gerelateerd zijn aan taken die mensen tegenkomen in de beroepspraktijk. Tijdens het uitvoeren van deze opdrachten worden de leerlingen gecoacht door hun docent.

In dit onderzoek is de coachcompetentie van docenten in het MBO beoordeeld op basis van een videodossier. Een videodossier bestaat uit een mix van bewijsbronnen die een compleet overzicht geven van de coachcompetentie van een docent. De belangrijkste bewijsbronnen in het dossier zijn de videofragmenten die verschillende kritische situaties tonen waarin een docent zijn of haar coachperformance laat zien. Verder zijn vier bronnen met contextinformatie toegevoegd: een samenvatting van wat er tijdens het videofragment te zien is en van wat er vooraf ging aan het videofragment, achtergrondinformatie over de leerlingen die tijdens het videofragment te zien zijn (leeftijd, vooropleiding, begeleidingsbehoefte, enz.), een beschrijving van het lesmateriaal dat tijdens het videofragment wordt gebruikt en een interview met de docent en met de leerling(en) waarin gereflecteerd wordt op de coachsituatie. Tot slot is er een training ontwikkeld voor beoordelaars waarin centraal staat hoe de criteria

voor competent coachen, de onderscheiden performanceniveaus en de scoringsregels toegepast dienen te worden tijdens het scoren van de videodossiers.

De centrale vraag van het onderzoek luidt: *in welke mate zijn beoordelingen op basis van een videodossier betrouwbaar, generaliseerbaar en valide?* Deze vraag is uitgewerkt in meer specifieke onderzoeksvragen die zijn onderzocht in drie deelstudies. De eerste deelstudie is een kleinschalig onderzoek waarin het ontwerp en de evaluatie van de beoordelingsprocedure centraal staan. In deze studie zijn de volgende onderzoeksvragen beantwoord:

1a In hoeverre komen beoordelaars tot dezelfde beoordelingen op basis van de ontworpen beoordelingsprocedure?

1b Welke aspecten van het videodossier stimuleren of belemmeren beoordelaars in het geven van valide interpretaties en beoordelingen?

In de tweede deelstudie is de betrouwbaarheid van de competentiebeoordelingen op basis van een videodossier nader onderzocht bij een grotere steekproef van beoordelaars. Daarnaast is in deze deelstudie ook de generaliseerbaarheid van competentiebeoordelingen nagegaan. De volgende onderzoeksvragen zijn beantwoord:

2a In hoeverre wordt de coachcompetentie van docenten in het MBO op basis van een videodossier betrouwbaar gescoord?

2b In hoeverre zijn de beoordelingen van afzonderlijke videofragmenten van docenten generaliseerbaar naar het beoogde universum van videofragmenten?

In de derde deelstudie zijn de betrouwbaarheid en validiteit van het scoren nader onderzocht. In deze deelstudie worden de volgende onderzoeksvragen beantwoord:

3a In hoeverre baseren verschillende beoordelaars hun beoordeling van de coachperformance in de videofragmenten op dezelfde bewijzen en argumenten?

3b Welk type bewijzen en argumenten rapporteren beoordelaars op de scoreformulieren?

3c In hoeverre rapporteren beoordelaars bewijzen en argumenten die corresponderen met het conceptuele kader dat is ontwikkeld voor het beoordelen van de coachcompetentie van docenten?

Voorafgaand aan de eerste deelstudie werd de beoordelingsprocedure ontworpen. De eerste stap in het ontwerp van deze procedure bestond uit een gedetailleerde analyse van de docentcompetentie coachen in de context van zelfstandig leren in het MBO. Op basis van deze domeinanalyse werden scoringsregels en een conceptueel kader gedefinieerd. De tweede stap in het ontwerp van de beoordelingsprocedure bestond uit de constructie van videodossiers met de hulp van een professionele filmploeg. Gedurende een periode van vier weken werden verschillende bronnen van bewijs verzameld rond een serie kritische coachsituaties in de praktijk. De derde stap in het ontwerp van de beoordelingsprocedure bestond uit het ontwerpen van een scoringsprocedure. Volgens deze procedure beoordeelden assessoren allereerst de coachperformance in afzonderlijke videofragmenten. Zij gebruikten hierbij specifieke criteria en beschrijvingen van competentieniveaus en werden aangespoord concrete bewijzen te zoeken waarop zij een beoordeling baseren. Vervolgens werd de beoordelaars gevraagd een overalloordeel te geven waarbij alle videofragmenten in beschouwing werden genomen. Ook hierbij gebruikten beoordelaars de beschrijvingen van de competentieniveaus en werden zij aangespoord hun beoordeling te onderbouwen met bewijzen en argumenten die betrekking hadden op de coachperformance in de afzonderlijke videofragmenten. Tot slot is er een training ontworpen waarin beoordelaars getraind werden in het toepassen van scoringsregels, het conceptuele kader en de scoringsprocedure. Nadat de beoordelingsprocedure was ontworpen, zijn de videodossiers beoordeeld door getrainde beoordelaars.

Hoofdstuk 2

In het tweede hoofdstuk wordt de eerste deelstudie beschreven. Het betreft een kleinschalig onderzoek waarin de interbeoordelaarsbetrouwbaarheid onderzocht werd evenals aspecten in het ontwerp van de beoordelingsprocedure die beoordelaars stimuleren of belemmeren in het maken van valide interpretaties en beoordelingen. Om een indicatie te krijgen van de overeenstemming in toegekende scores door beoordelaars, zijn scoreformulieren verzameld en analyses uitgevoerd. Op de scoreformulieren noteerden beoordelaars de scores die zij toekenden aan de getoonde performance in de videofragmenten. Voor de toegekende scores werd de Gower-coefficient bepaald als indicatie voor de interbeoordelaarsovereenstemming. Om inzicht te krijgen in aspecten van het ontwerp van de beoordelingsprocedure die beoordelaars belemmeren of stimuleren bij het komen tot valide interpretaties en beoordelingen, zijn alle beoordelaars geïnterviewd. Uit de resultaten van deze

deelstudie blijkt dat op basis van de ontworpen beoordelingsprocedure een acceptabel tot hoog niveau van interbeoordelaarsovereenstemming kon worden bereikt. Beoordelaars plaatsten hierbij wel de kanttekening dat het toepassen van de scoringsprocedure een aanzienlijke hoeveelheid tijd en energie kostte. Daarnaast waren de beoordelaars van mening dat de training een noodzakelijke conditie was voor het correct toepassen van deze procedure. Verschillende aspecten van de procedure bleken beoordelaars te stimuleren bij het komen tot interpretaties en beoordelingen:

- het conceptuele kader met beschrijvingen van leeractiviteiten en gerelateerde coachinterventies hielp de beoordelaars bij het beoordelen van de relevante aspecten van een coachperformance;
- de samenvatting van wat er gebeurt tijdens een kritische coachsituatie hielp de beoordelaars de relevante aspecten van de coachperformance te beoordelen;
- de contextinformatie, vooral het interview met de docent en de deelnemer(s), hielp de beoordelaars bij het begrijpen van het handelen van de docent en de gevolgen hiervan voor de deelnemers;
- beoordelaars gaven aan dat ze 'ongecompliceerde' coachsituaties gemakkelijker konden begrijpen en daarom gemakkelijker konden scoren. Onder 'ongecompliceerde coachsituaties' werden in het algemeen coachsituaties verstaan waarbij (a) de performance van de docent overeen komt met de toelichting op de performance van de docent tijdens het interview of (b) het coachen op een specifieke leeractiviteit duidelijke te onderscheiden was van het coachen op andere leeractiviteiten of (c) de deelnemers behoefte hadden aan coaching op een duidelijk te onderscheiden leeractiviteit.

Naast de aspecten die een positieve invloed hadden op het komen tot valide interpretaties en beoordelingen, werden ook aspecten gevonden die beoordelaars daarbij belemmerden:

- de coachperformance in afzonderlijke videofragmenten bleek moeilijker te beoordelen dan de coachperformance over verschillende videofragmenten heen, omdat de afzonderlijke videofragmenten maar kleine stukjes laten zien van wat er tussen docent en deelnemer(s) plaatsvindt;
- videofragmenten die langer duurden dan 15 minuten leken niet bij te dragen aan meer valide interpretaties en beoordelingen, omdat het moeilijk was voor beoordelaars om zich langer dan 15 minuten te concentreren op de coachperformance en omdat er volgens de beoordelaars geen nieuwe essentiële informatie werd toegevoegd na 15 minuten;

- het bleek soms moeilijk om het coachen op competentieniveau twee te onderscheiden van het coachen op competentieniveau drie, wat in de beoordelingsprocedure de kritieke scheiding is tussen 'onvoldoende' en 'voldoende';
- de mate waarin een docent 'praktijkgericht' coacht was niet te beoordelen op basis van de ontwikkelde videodossiers omdat de docenten nauwelijks gedrag vertoonden dat in overeenstemming was met dit criterium. Daardoor konden de beoordelaars hun beoordeling van het praktijkgerichte coachen alleen baseren op negatief bewijs in termen van gemiste kansen door de docent.

Hoofdstuk 3

In het derde hoofdstuk wordt de tweede deelstudie beschreven. Op basis van verschillende analyses werd getracht een indicatie te krijgen van de betrouwbaarheid van de beoordelingsprocedure. Er werd bepaald in welke mate scoringstendenties voorkwamen in het scoren door de beoordelaars, de interbeoordelaarsovereenstemming werd vastgesteld evenals de generaliseerbaarheid van toegekende scores over beoordelaars. Deze analyses werden uitgevoerd op een grotere steekproef dan in de eerste deelstudie. Daarnaast werden verschillende analyses uitgevoerd om een indicatie te krijgen van de generaliseerbaarheid van de scores naar een universumscore. Een universumscore verwijst in dit verband naar de score die een respondent behaald zou hebben, wanneer hij of zij alle mogelijke taken zou hebben uitgevoerd die er zijn om de competentie te meten. Er werd een rangorde bepaald van videofragmenten die zeer eenduidige scores uitlokten tot videofragmenten die zeer wisselende scores uitlokten bij de verschillende beoordelaars. De videofragmenten die zeer wisselende toegekende scores uitlokten bij de verschillende beoordelaars zijn een bedreiging voor de generaliseerbaarheid. Ook werd voor elk videofragment bepaald in welke mate de toegekende scores aan de coachperformance in het fragment overeenkwamen met scores die werden toegekend aan de coachperformance in andere fragmenten. Een belangrijke conclusie van deze deelstudie is dat de designprincipes van de beoordelingsprocedure lijken bij te dragen aan betrouwbaar scoren door beoordelaars. Het gebruiken van het ontwikkelde beoordelingskader, de competentieniveaus, de scoringsvoorschriften en de ontwikkelde dossiers door de beoordelaars tijdens het scoren en het volgen van de training gaan over het algemeen samen met betrouwbaar scoren door beoordelaars. Er werd een acceptabel niveau van interbeoordelaarsovereenstemming gevonden en ook

de generaliseerbaarheid van toegekende scores over de beoordelaars was hoog. Uit de resultaten bleek verder dat wanneer twee beoordelaars betrokken zijn bij de competentiebeoordeling op basis van een videodossier, een acceptabel niveau van interbeoordelaarsovereenstemming bereikt kan worden. Naast deze positieve resultaten, is uit deze deelstudie gebleken dat scoringstendenties voorkwamen. Beoordelaars waren niet in staat alle docenten even mild of streng te beoordelen. Bovendien bleken beoordelaars die een collega waren van de te beoordelen docent extreme beoordelingen te geven, zowel extreem positief als negatief.

Op basis van deze deelstudie kunnen ten aanzien van de generaliseerbaarheid over videofragmenten alleen tendensen worden beschreven. Definitieve conclusies over het minimale aantal videofragmenten dat nodig is om uitspraken te doen over het coachen van de docent kunnen dan ook niet getrokken worden. Het standaardiseren van de videofragmenten op basis van de definitie van een kritische situatie lijkt samen te gaan met positieve resultaten op het gebied van het generaliseren van toegekende scores over videofragmenten. De overeenstemming tussen toegekende scores aan een videofragment en de gemiddelde toegekende scores aan de rest van de videofragmenten is over het algemeen acceptabel tot goed, alleen de generaliseerbaarheid van toegekende scores aan de videofragmenten waarin de docent coacht op leerhouding is problematisch.

Hoofdstuk 4

In het vierde hoofdstuk wordt de derde deelstudie beschreven. In deze deelstudie is de validiteit van het scoren door beoordelaars onderzocht. Om de onderzoeksvragen van de deelstudie te kunnen beantwoorden, zijn verschillende kwantitatieve en kwalitatieve analyses uitgevoerd op de bewijzen en argumenten die beoordelaars rapporteerden op scoreformulieren om hun toegekende scores te rechtvaardigen. Op basis van deze analyses werd bepaald in welke mate constructirrelevante variantie en construct onderrepresentatie invloed hadden op het scoren door beoordelaars. Er werd een aanzienlijke variatie gevonden in bewijzen en argumenten die de verschillende beoordelaars aandroegen om een toegekende score te legitimeren. Ook wanneer eenzelfde score werd toegekend, bleken de bewijzen en argumenten uiteen te lopen. Er werd een grotere variatie gevonden in de argumenten dan in de verzamelde bewijzen. Op de scoreformulieren werd 58% tot 100% van de argumenten door maar een van de twaalf beoordelaars genoteerd. Verder bleek dat beoordelaars zowel concrete uitspraken deden over wat ze gezien hadden in het videodossier als, meer

abstracte interpretaties en beoordelingen gaven van wat ze gezien hadden in het videodossier. De concrete uitspraken werden voornamelijk gebruikt bij het aandragen van bewijzen en de abstracte uitspraken voornamelijk bij het aandragen van argumenten. De concrete bewijzen hadden betrekking op het gedrag van de docent: beoordelaars noteerden de vragen en feedback die de docenten inzetten tijdens het coachen. Deze bewijzen werden beschouwd als relevante bewijzen, omdat ze pasten binnen het conceptuele kader dat de beoordelaars zouden moeten gebruiken bij het beoordelen. Beoordelaars lijken dus redelijk in staat om relevante bewijzen te identificeren. Bij de cijfermatige beoordeling, schreven de beoordelaars een toelichting waarin ze interpretaties gaven van wat ze tijdens de videofragmenten hadden gezien en ook gaven ze een waardeoordeel hierover. Beoordelaars noteerden voornamelijk argumenten die betrekking hadden op het gedrag van de docent (18%) en de coachsituatie (14%). De waardeoordelen die beoordelaars noteerden in deze toelichting hadden betrekking op het gedrag van de docent (48%) en op de consequenties van het gedrag voor de deelnemers (19%). In het algemeen waren de bewijzen en argumenten consistent met het ontwikkelde conceptuele kader dat de beoordelaars verondersteld werden te gebruiken tijdens het beoordelen en werden er weinig construct-irrelevante bewijzen en argumenten aangedragen door beoordelaars. Het scoren door beoordelaars lijkt wel beïnvloed te worden door construct onderrepresentatie. Beoordelaars waren geneigd om in plaats van alle aspecten alleen een of twee aspecten van het conceptuele kader te gebruiken bij het beoordelen van de coachperformance.

Hoofdstuk 5

Op basis van de drie deelstudies worden in hoofdstuk 5 de algemene conclusies, beperkingen, suggesties voor vervolgonderzoek en praktische implicaties van het onderzoek besproken. Op basis van de drie deelstudies zijn tien algemene conclusies geformuleerd. Vijf van deze conclusies hebben betrekking op de mate waarin beoordelingen van een videodossier *betrouwbaar* zijn:

1. beoordelaars bereikten een acceptabel tot hoog niveau van overeenstemming voor het toekennen van (overall)scores wanneer zij de coachcompetentie van docenten uit het MBO beoordeelden;
2. beoordelaars bereikten een hoger niveau van overeenstemming voor het toekennen van overallscores dan voor het toekennen van scores aan de coachperformance in afzonderlijke videofragmenten;

3. twee beoordelaars waren nodig om een acceptabel niveau van interbeoordelaarsovereenstemming te verkrijgen;
4. het scoren door beoordelaars werd beïnvloed door scoringstendenties;
5. beoordelaars baseerden hun toegekende scores op verschillende bewijzen en argumenten, waarbij meer variatie werd gevonden in argumenten dan in bewijzen.

Twee conclusies hebben betrekking op de mate waarin de beoordelingen op basis van een videodossier *generaliseerbaar* zijn:

6. de beoordelingen die werden toegekend aan de videofragmenten van bepaalde docenten waren beter te generaliseren naar het beoogde universum van videofragmenten dan de beoordelingen van andere docenten;
7. de beoordelingen die werden toegekend aan videofragmenten waarin de docent coachte op bepaalde leeractiviteiten waren beter te generaliseren naar het beoogde universum van videofragmenten dan de beoordelingen van het coachen op andere leeractiviteiten.

Tot slot zijn drie conclusies getrokken die betrekking hebben op de mate waarin de beoordelingen op basis van een videodossier *valide* zijn:

8. beoordelaars ervoeren de contextinformatie die was toegevoegd aan het videodossier als noodzakelijke achtergrondinformatie voor een valide beoordeling van de coachcompetentie van de docenten;
9. beoordelaars waren in staat om tijdens het scoren bewijzen en argumenten te gebruiken die correspondeerden met het ontwikkelde conceptuele kader;
10. de validiteit van het beoordelen op basis van een videodossier werd wellicht bedreigd door het feit dat beoordelaars tijdens het scoren maar een of twee aspecten van het conceptuele kader gebruikten in plaats van alle aspecten.

Vervolgens worden in hoofdstuk 5 enkele beperkingen van het onderzoek beschreven. Ten eerste kon als gevolg van de gebruikte steekproef in deelstudie twee geen generaliseerbaarheidstudie worden uitgevoerd, maar werd op basis van itemrest correlaties en standaarddeviaties een indicatie gegeven van de generaliseerbaarheid over videofragmenten. Ten tweede is de studie in hoofdstuk vier gebaseerd op *gerapporteerde* bewijzen en argumenten door beoordelaars op scoreformulieren. Hierdoor werden bewijzen en argumenten die beoordelaars niet rapporteerden, maar die mogelijk wel een rol speelden in het beslisproces, buiten beschouwing gelaten. Deze inperking is aangebracht, omdat het bestuderen van expliciet gerapporteerde bewijzen en argumenten een logische eerste stap is bij het onderzoeken van constructirrelevante variantie en construct onderrepresentatie in het scoren door

beoordelaars. Ten derde is alleen bestudeerd *welke* bewijzen en argumenten beoordelaars aandroegen en niet *hoe* de bewijzen en argumenten gecombineerd werden tot een (eind)oordeel. Ook hier is de reden dat het bepalen van de gebruikte bewijzen en argumenten, een eerste logische stap is in het onderzoeken van het scoringsproces van beoordelaars. Pas na deze eerste stap kan onderzoek worden gedaan naar de wijze waarop beoordelaars bewijzen en argumenten combineren tot een (overall)oordeel.

Hoofdstuk 5 bevat tevens enkele suggesties voor vervolgonderzoek. De eerste lijn van vervolgonderzoek betreft het onderzoeken van de mate waarin de beoordelingen, verkregen op basis van een videodossier, kunnen worden geëxtrapoleerd naar prestaties buiten de assessmentcontext. De tweede lijn van vervolgonderzoek heeft betrekking op de mate waarin de ontwikkelde beoordelingsprocedure bijdraagt aan de professionele ontwikkeling van docenten die hebben deelgenomen aan het assessment. De derde lijn betreft vervolgonderzoek dat zich richt op het nader onderzoeken van kenmerken van videofragmenten.

Tot slot worden in hoofdstuk 5 enkele implicaties van het onderzoek beschreven voor de assessmentpraktijk. Ten eerste blijkt uit het onderzoek dat de voorgestelde designprincipes uit de literatuur die zijn toegepast bij de constructie van de videodossiers gebruikt kunnen worden voor het genereren van betrouwbare, generaliseerbare en valide beoordelingen. Over het algemeen werden positieve resultaten gevonden ten aanzien van het scoren door beoordelaars en de generaliseerbaarheid van beoordelingen wanneer de ontwikkelde beoordelingsprocedure werd ingezet tijdens het beoordelen van de coachcompetentie van docenten. Ten tweede blijkt uit het onderzoek dat in de praktijk volstaan kan worden met twee beoordelaars om tot betrouwbare beoordelingen te komen, mits het assessment wordt vormgegeven volgens de designprincipes die in dit onderzoek gebruikt zijn. Dit is een belangrijke implicatie, omdat het in de praktijk vaak niet mogelijk is om nog meer beoordelaars in te zetten bij assessments vanwege de hoge kosten die dit met zich mee zou brengen. Ten derde zijn er verschillende aanwijzingen uit het onderzoek naar voren gekomen voor de verbetering van trainingen voor beoordelaars. In dit soort trainingen zou bijvoorbeeld veel aandacht besteed moeten worden aan het onderscheid tussen performances die net wel en die net niet als voldoende kunnen worden aangemerkt. Beoordelaars blijken het moeilijk te vinden om performances op de grens van voldoende en onvoldoende te beoordelen.

Daarnaast blijkt dat beoordelaars tijdens de training gestimuleerd moeten worden om alle aspecten van het conceptuele kader te gebruiken, omdat ze anders geneigd zijn sommige aspecten buiten beschouwing te laten tijdens het beoordelen. Een aspect dat ook expliciet in de training tot uitdrukking zou moeten komen, is het creëren van een gedeeld conceptueel kader, zodat alle beoordelaars hetzelfde verstaan onder de competentie die ze moeten beoordelen en het conceptuele kader dat ze gebruiken tijdens het beoordelen. Deze maatregel zou de aanzienlijke variatie in aangedragen bewijzen en argumenten moeten verminderen. De laatste aanwijzing voor assessorentrainingen bestaat eruit dat er tijdens de training intensief geoefend moet worden in het bepalen van de consequenties van het handelen (coachen) van de docent voor de deelnemers. Beoordelaars waren erop gericht om concreet gedrag van de docent te beoordelen en waren minder geneigd om ook de consequenties van het gedrag voor de deelnemers mee te nemen in de beoordeling.