Beoordelingen door docenten: een analyse van de betrouwbaarheid van rapportcijfers[1]

R.M. van der Lans, W.J.C.M. van de Grift en K. van Veen

Samenvatting

In eerder onderzoek wordt gesteld dat docenten subjectieve beoordelaars zijn die zich bij het geven van cijfers niet beperken tot het becijferen van alleen de leerlingvaardigheid, maar cijfers geven voor een mengelmoes ('hodgepodge') van eigenschappen: de hodgepodgehypothese. Ook zouden docenten verschillen in mildheid; de mildheidshypothese. In dit onderzoek worden deze beide hypothesen onderzocht. Voor dit onderzoek zijn bij twee steekproeven proefwerkcijfers verzameld. De eerste steekproef telt 5988 proefwerkcijfers gegeven aan 192 leerlingen gedurende één schooljaar door 64 docenten. De tweede steekproef telt 29462 proefwerkcijfers gegeven aan 306 leerlingen gedurende drie opeenvolgende schooljaren door 52 docenten. Om de beoordelingsbias te onderzoeken werden een G-studie en D-studie uitgevoerd. De resultaten geven geen overtuigend bewijs voor de twee hypotheses. In het algemeen blijkt dat rapportcijfers een redelijk betrouwbaar onderscheid maken tussen minder en meer vaardige leerlingen ($E\rho^2 \geq .70$) en een betrouwbare beoordeling geven over de cesuur voldoende-onvoldoende ($\Phi_\lambda \approx .90$). Wanneer rapportcijfers op minder dan 8 proefwerken zijn gebaseerd dan is de betrouwbaarheid lager dan het criterium .70. Een aanzienlijk deel van de onbetrouwbaarheid in beoordeling kan worden verklaard door verschillen in de kwaliteit van de proefwerken en niet door mildheid of hodgepodgegedrag in de beoordeling van docenten.

Kernwoorden: Beoordeling, bias, cijfers, betrouwbaarheid, Generaliseerbaarheidstheorie

1 Inleiding

In deze studie wordt ingegaan op de beoordeling van leerlingen door docenten. Er is eerder op diverse manieren onderzoek verricht naar het becijferen door docenten (bijv., Bowers, 2009, 2010 2011; Brookhart, 1994, 2004; Marzano, 2002, Randall & Engelhard, 2010). In dit eerdere onderzoek wordt er vanuit gegaan dat docenten subjectieve beoordelaars zijn omdat docenten zich niet beperken tot het becijferen van alleen de leerlingvaardigheid, maar cijfers geven voor een mengelmoes ('hodgepodge') van eigenschappen waaronder de getoonde motivatie, de houding en het gedrag en de groei die de leerling heeft gemaakt (Brookhart, 1994; McMillan, Myran, & Workman, 2002). Ander onderzoek heeft zich gericht op verschillen in mildheid, zodat sommige docenten hun leerlingen becijferen ten aanzien van strengere eisen dan collega-docenten (bijv., Kuhlemeier & Kremers, 2013). Mede door deze veronderstelde subjectiviteit in cijfers is in veel landen een traditie ontstaan om leerlingvaardigheid ook te evalueren op basis van gestandaardiseerde toetsen en in sommige landen, met name Engeland, neemt deze traditie langzaam de evaluatie op basis van cijfers over (Standaert, 2014).

Toch wijst recent onderzoek uit dat juist de 'subjectieve' cijfers grotere voorspellende validiteit hebben voor het toekomstig schoolsucces dan gestandaardiseerde toetsen (Atkinson & Geiser, 2009; Bowers, 2009, 2010; Cliffordson, 2008; Thorsen & Cliffordson, 2008). Ook studies die construct validiteit van schoolcijfers bestuderen geven geen aanleiding om te veronderstellen dat schoolcijfers compleet onbetrouwbaar zijn (bijv., Brennan, Kim, Wenz-Gros, & Sieperstein, 2001; Südkamp, Kaiser, & Möller, 2012). Gerapporteerde correlaties tussen schoolcijfers en prestaties van dezelfde leerlingen op gestandaardiseerde toetsen zijn hoog en liggen doorgaans tussen 0.5 en de 0.6. Hieruit kan opgemaakt worden dat schoolcijfers niet een compleet subjectief oordeel zijn welke los staat van andere maten voor schoolsucces.

De literatuur geeft dus een gemengd beeld. Eerder onderzoek benadrukt zowel de onbetrouwbaarheid en subjectiviteit in becijfering, maar tegelijk geeft het indicaties dat schoolcijfers niet compleet subjectief kunnen zijn (i.e., het is niet mogelijk om meermaals hoge correlaties te vinden wanneer één van beide maten compleet onbetrouwbaar en subjectief zou zijn). Opvallend is echter, dat ondanks dat we uit voorgaande studies wel verwachtingen kunnen formuleren over de betrouwbaarheid van becijfering en rapportcijfers, geen van de genoemde studies de betrouwbaarheid van schoolcijfers heeft bestudeerd. Als gevolg hiervan weten we nog weinig van de exacte invloed van subjectiviteit op de betrouwbaarheid van gegeven rapportcijfers. Drany & Wilson (2008) merkten hierover vrij recent nog op:

"in contrast to the situation for trained raters… we know little about the *consistency* [onze cursivering] with which teachers apply the standards contained within a scoring guide to the work their students generate in the classroom" (p. 418).

Dit roept de vraag op hoe groot de invloed van beoordelingsbias is op de betrouwbaarheid van becijfering en de vraag hoe onderzoek naar betrouwbaarheid van schoolcijfers zou kunnen bijdragen aan de kennis over beoordelingsbias van docenten. In het vervolg van de theoretische achtergrond schetsen we een methodiek om beoordelingsbias in proefwerkcijfers te evalueren aan de hand van Generaliseerbaarheidtheorie (Cronbach, Gleser, Rajaratnam, & Nanda, 1972). De hoofdvraag van het artikel is: *In welke mate worden beoordelingen gemaakt op basis van door docenten gegeven proefwerkcijfers vertekend door beoordelingsbias?*

2 Theoretische achtergrond

2.1 Definiëring van beoordeling en beoordelingsbias

We zullen in de loop van de tekst veelvuldig terugkomen op enkele verwante begrippen: observatie, (proefwerk)cijfer, beoordeling, rapportcijfer en besluit. Om onderscheid te maken tussen deze begrippen maken we gebruik van het werk van Hofstee (1999). Hofstee (1999) spreekt van een beoordeling wanneer *mensen, op gezag van anderen, de kwaliteit(en) van iets van iemand vaststellen*. In dit onderzoek zijn het de docenten die op gezag van de school de vaardigheid van leerlingen in een schoolvak vaststellen. Hofstee (1999) beargumenteert dat beoordelingen zouden moeten zijn gebaseerd op meerdere observaties. Zulke observaties vinden in de scholen regelmatig plaats in de vorm van proefwerken en schooloverhoringen. In de tekst gebruiken we de woorden observatie en proefwerk daarom als synoniemen. De proefwerkcijfers die worden toegekend monden uit in een beoordeling in de vorm van een rapportcijfer. Daarom gebruiken we de woorden beoordeling en rapportcijfer ook als synoniemen. Een besluit is gebaseerd op een groter aantal beoordelingen; vaak over verschillende (niet gemakkelijk verenigbare) vaardigheden (Hofstee, 1999). In scholen leiden de rapportcijfers via een van tevoren afgesproken protocol tot besluiten over doubleren of versnellen.

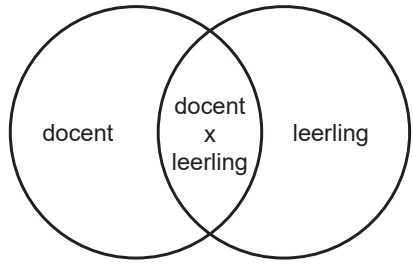
De definitie van Hofstee maakt tegelijk duidelijk dat er sprake is van beoordelingsbias wanneer een rapportcijfer van een leerling is gebaseerd op andere kwaliteiten dan de vaardigheid in een schoolvak. We zijn ons ervan bewust dat deze definitie een specifieke invulling geeft aan het doel van een rapportcijfer – namelijk het zichtbaar maken van de vaardigheid in een schoolvak– en dat dit kan worden gezien als specifieke invulling van het doel van onderwijs. We willen daarom benadrukken dat het niet onze bedoeling is om de impressie te wekken dat vaardigheid in het schoolvak – of in Biesta (2012) 's termen kwalificatie – het enige doel is van onderwijs. Ons uitgangspunt is dat cijfers slechts één eigenschap of doel tegelijk zouden moeten meten en niet een 'hodgepodge' van meerdere doelen. Wanneer onderwijs ook andere doelen, zoals de persoonlijke vorming van de leerling (of in Bieta's (2012) termen socialisatie en subjectivicatie), nastreeft en mee wil wegen in de besluitvorming over doubleren dan staat het de scholen vrij om

leerlingen daarin te becijferen. Door de tijd zijn er periodes geweest waarin scholen aparte cijfers gaven voor attitudes en/of motivatie – er zijn zelfs periodes geweest waarin veel waarde werd gehecht aan de cijfers voor attitude en motivatie (Standaert, 2014). De stelling dat cijfers één eigenschap tegelijk zouden moeten meten is overigens vaker ingenomen; onder andere door De Groot en Wijnen (1983), Brookhart (2004) en Randall en Engelhard (2010).

2.2 Beoordelingsbias: Enkele hypotheses en veronderstellingen

In het meeste onderzoek naar beoordelingsbias wordt gebruik gemaakt van zelfrapportagemethodes waarin docenten rapporteren over hun becijferingspraktijk (bijv., Brookhart, 1994, 2004; Cross & Frary, 1999; McMillan, Myran, & Workman, 2002; Randall & Engelhard, 2010). Dit onderzoek bevraagt docenten welke factoren zij meewegen bij het beoordelen van leerlingen. Uit dit onderzoek blijkt dat docenten hun beoordelingen ook van andere factoren laten afhangen dan van de vaardigheid. McMillan et al. (2002) rapporteren dat 39% van de docenten stelt dat zij in de beoordeling van hun leerlingen andere criteria dan de getoonde vaardigheid laten meewegen. In onderzoek van Cross & Frary (1999) stelt 37% van de docenten dat het gedrag en de houding van leerlingen – zoals interesse en nieuwsgierigheid – ook moet meewegen in de beoordeling. Recent zijn deze resultaten herbevestigd door Randall & Engelhard (2010). Op basis van deze resultaten wordt verondersteld dat in de beoordelingen van docenten een mengelmoes van factoren meewegen, waaronder gedrag en attitudes. Deze hodgepodge zou er vervolgens toe leiden dat docenten werk van dezelfde kwaliteit soms toch verschillend becijferen. De resultaten van dit zelfrapportage-onderzoek heeft ruim baan gegeven aan de veronderstelling dat cijfers voor een groot deel worden vertroebeld door docent-leerling-interacties.

De veronderstelling achter de hodgepodgehypothese kan worden weergegeven in een Venn diagram (Figuur 1). De cirkels in Figuur 1 noemen we facetten. Ieder facet geeft een variantie weer. Het facet docent geeft weer



Figuur 1.
Een Venndiagram van het gekruiste design. Wanneer de beoordelingsbias gering is dan zouden de facetten docent en docent × leerling klein zijn.

dat niet iedere docent tot dezelfde beoordelingen komt. Het facet leerling geeft weer dat niet iedere leerling dezelfde beoordeling krijgt. Het overlappende gedeelte geeft het interactie-effect weer tussen beide facetten. In een situatie waarbij de beoordeling vrij is van bias zou het facet leerling hoge variantie hebben, terwijl de facetten docent en docent × leerling (zeer) kleine variantie hebben.

In de hodgepodgehypothese wordt verondersteld dat een docent een hogere of lagere beoordeling geeft aan één leerling dan aan medeleerlingen, terwijl andere docenten deze ene leerling niet hoger of lager beoordelen. In dit geval zou er hoge variantie zijn in het facet docent-leerling-interactie en is er sprake van bias.

Zoals al opgemerkt, een beperking van het eerdere onderzoek naar de hodgepodgehypothese is dat het zich heeft beperkt tot zelfrapportage-onderzoek. Volgens Muijs (2006) geeft zelfrapportage-onderzoek een imperfecte indicator van het vertoonde gedrag van leraren. Het laat onvoldoende zien hoe de situatie werkelijk is. Het is daarom waardevol om te zoeken naar andere onderzoeksmethoden waarmee de hodgepodgehypothese kan worden onderzocht.

In een andere lijn van onderzoek naar beoordelingsbias wordt verondersteld dat docenten verschillen in strengheid. Dit kan getypeerd worden als de mildheidshypothese (bijv., Drany & Wilson, 2008). In deze hypothese wordt verondersteld dat de variatie in

het facet docent (zie Figuur 1) relatief hoog is. Deze hypothese is vooral onderzocht met quasi-experimenteel onderzoek (bijv., Drany & Wilson, 2008; Kuhlemeier & Kremers, 2013; Starch & Elliot, 1914), waarbij hetzelfde werk is beoordeeld door zowel de docent als een tweede beoordelaar. Ook in deze lijn van onderzoek wordt gesteld dat docenten subjectieve beoordelaars zijn, maar de resultaten zijn minder dramatisch. Drany en Wilson (2008) rapporteren bijvoorbeeld dat sommige docenten milder zijn dan andere docenten, maar ook dat er een redelijke mate van consistentie bestaat tussen docenten in hun beoordelingen.

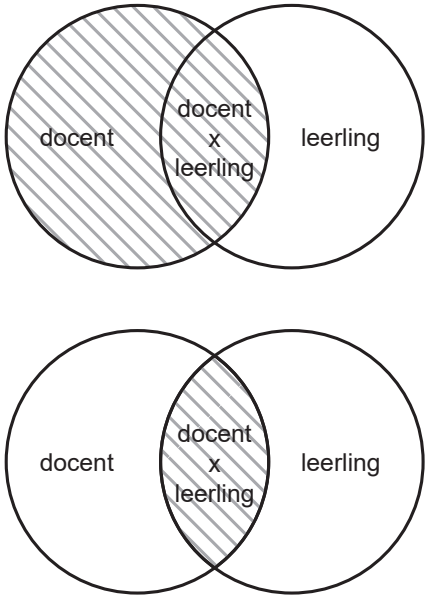
De conclusies uit dit eerdere quasi-experimentele onderzoek worden beperkt doordat de onderzoeksprocedure op een aantal punten afwijkt van de praktijksituatie waarin een docent cijfers toekent aan werk van leerlingen, in het bijzonder: (1) in het eerdere onderzoek wordt het werk van leerlingen becijferd door (tweede) beoordelaars die deze leerling niet kennen, terwijl een docent nooit het werk becijfert van leerlingen die onbekend zijn; (2) in het eerdere onderzoek worden cijfers gegeven die geen consequenties hebben voor de leerling, terwijl cijfers die worden gegeven in de school dit wel hebben; en (3) dit eerdere onderzoek is gebaseerd op de verschillen in becijfering van één proefwerk terwijl op scholen beoordelingen worden gegeven op basis van meerdere proefwerken. Deze beperkingen geven reden tot twijfel in hoeverre de resultaten in dit quasi-experimenteel onderzoek kunnen worden gegeneraliseerd naar de onderwijspraktijk. Op basis van dit eerdere quasi-experimenteel onderzoek kan gesteld worden dat docenten mogelijk verschillen in mildheid bij het eenmalig toekennen van cijfers aan proefwerken, maar het is niet duidelijk hoe docenten verschillen in mildheid in geval van het beoordelen van de leerling op basis van meerdere proefwerken.

Samenvattend kan gesteld worden dat in eerder onderzoek het theoretische model van Figuur 1 in beperkte mate is getoetst. Het empirische onderzoek heeft zich tot nog toe vooral gericht op de mildheidhypothese: de docent is milder naar alle leerlingen dan een collega-docent zou zijn. De hodgepodgehypothese is, voor zover ons bekend, alleen met zelfrapportage onderzocht. Ook worden de conclusies van eerder onderzoek naar de beoordelingsbias beperkt door onderzoeksprocedures die afwijken van de gangbare procedures op school.

2.3 Vormen van beoordeling; absolute en relatieve beoordeling

Er zijn twee analysemethoden die hierbij gebruikt kunnen worden die gebaseerd zijn op twee manieren om kwaliteiten te beoordelen: (1) op basis van een cesuur of (2) op basis van een rangorde (bijv., De Groot & Wijnen, 1983), ook wel beschreven als respectievelijk criteriumbeoordeling en normatieve beoordeling (bijv., Cliffordson, 2008). In navolging van Cronbach, et al. (1972) refereren we naar de eerste manier van beoordeling met de term absolute beoordeling en de tweede manier van beoordeling met de term relatieve beoordeling. Deze beide manieren van beoordelen leiden tot een andere analyse van de beoordelingsbias (Brennan, 2001; Cronbach, et al., 1972; Shavelson & Webb, 1991). Bij een absolute beoordeling is sprake van beoordelingsbias wanneer docenten geen overeenstemming hebben over het exacte rapportcijfer. Bij relatieve beoordelingen is sprake van beoordelingsbias wanneer docenten geen overeenstemming hebben over de rangorde van leerlingen: van minst naar meest vaardig. De Groot & Wijnen (1983) merken op dat de rapportcijfers in scholen zowel een relatieve beoordelingsfunctie hebben – de rapportcijfers dienen een rangorde weer te geven waarin ‘voldoende’ leerlingen worden onderscheiden van ‘goede’ leerlingen – en tegelijk door de cesuur ook absolute beoordelingsfunctie hebben – cijfers onder de 5.5 worden gezien als onvoldoende en cijfers gelijk of groter dan 5.5 als voldoende.

In Figuur 2 is gearceerd wat in een empirische analyse tot de beoordelingsbias wordt gerekend bij een absolute (boven) en een relatieve (onder) beoordeling. Om de beoordelingsbias vast te stellen van een absoluut besluit worden alle facetten die leiden tot afwijkingen in de beoordeling meegewogen. In het bovenste Venn diagramm worden daarom én het facet docent en het interactie



Figuur 2.
Een Venndiagram weergave waarin voor absolute (boven) en relatieve (onder) beoordelingen is gearceerd welke facetten worden gerekend tot beoordelingsbias.

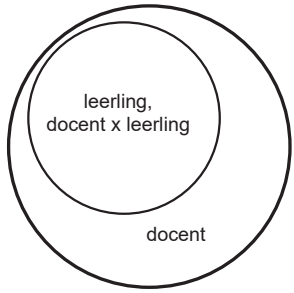
facet docent × leerling beschouwd als bron van bias. Om bias in relatieve besluiten vast te stellen worden alleen de facetten meegewogen die leiden tot wisselingen in de rangorde van leerlingen. In het onderste Venn diagramm wordt daarom alleen het interactie facet beschouwd als bias.

2.4 De schoolpraktijk; van theoretische analyse naar de praktische mogelijkheden

Het is moeilijk om een quasi-experimenteel onderzoek op te zetten waarin zowel de docent als een tweede (en derde) beoordelaar bekend zijn met de leerling en waarin deze docenten bij dezelfde leerlingen meerdere observaties uitvoeren (c.q. meerdere proefwerken geven en becijferen) om tot een beoordeling te komen. Toch zou een dergelijke onderzoeksopzet nodig zijn om de beoordelingsbias te meten. Een mogelijk alternatief zou gevonden kunnen worden in het gebruik van bestaande proefwerkcijfers. In scholen is veel data, in de vorm van proefwerkcijfers, aanwezig waarmee beoordelingsbias zou kunnen worden onderzocht.

Deze analyse met proefwerkcijfers is niet zonder problemen. Het belangrijkste obstakel is dat het op scholen gemeengoed is dat proefwerkcijfers worden toegekend aan één klas door één docent zonder dat andere collega's dit werk (steekproefsgewijs) voor een tweede maal becijferen. We weten daarom hoe één docent de proefwerken becijfert, maar missen informatie over hoe andere docenten diezelfde proefwerken zouden becijferen. Deze voor scholen gangbare procedure leidt tot een genest design (zie Figuur 3) in plaats van het meer volledige en in de theoretische analyse getoonde gekruiste design (Figuur 1).

Het belangrijkste kenmerk van het geneste design is dat de interactie tussen docent en leerling *confounded* is met het facet leerling. De *confound* is aangegeven in de Figuur 3 door beide facetten leerling en docent × leerling, te benoemen in dezelfde cirkel met daartussen een komma. Deze *confound* houdt in dat de variantie in cijfers tussen leerlingen wordt opgeteld bij de variantie in cijfers die ontstaat door docent × leerling interacties, zodat het facet leerling nu twee verklaringen heeft: (1) er is verschil in vaardigheid tussen leerlingen en (2) docenten beoordelen leerlingen van gelijke vaardigheid verschillend. Door deze



Figuur 3.
Een Venndiagram van het geneste design. Kenmerkend is dat het facet leerling binnen het facet docent zit. Als gevolg is het facet docent × leerling interactie confounded met het facet leerling. Dit is aangegeven door beide te noemen in dezelfde cirkel met daartussen een komma.

dubbele verklaring is het niet goed mogelijk om de hodgepodgehypothese te toetsen met het geneste design. Wel is het mogelijk om de mildheidshypothese te toetsen.

Er is een aantal mogelijkheden om met de cijfers beschikbaar op de scholen toch tot een gekruiste analyse te komen, maar deze mogelijkheden zijn gebaseerd op assumpties. De eerste mogelijkheid is om cijfers van dezelfde leerlingen te analyseren bij meerdere vakken. In dit geval hebben meerdere docenten de vaardigheid van de leerling beoordeeld en wanneer één docent tot een andere beoordeling komt dan de collega's zou dit duiden op bias. Lastig is dat de verschillen in beoordeling ook het gevolg kunnen zijn van verschillen in vaardigheid tussen de schoolvakken. Deze eerste mogelijkheid heeft dus de assumptie dat leerlingen niet noemenswaardig verschillen in vaardigheid tussen vakken: een leerling is goed in school of niet. Deze assumptie wordt niet ondersteund door eerder empirisch onderzoek (bijv., Bowers, 2011; Korobko, Glas, Bosker & Luyten, 2008; Thorsen & Cliffordson, 2008) waaruit blijkt dat verschillen tussen beoordelingen voor een belangrijk deel kunnen worden verklaard doordat leerlingen verschillen in hun vaardigheid tussen vakken.

Een andere mogelijkheid is om schoolcijfers van diverse schooljaren op te vragen. Leerlingen kunnen wisselen van docent bij opeenvolgende schooljaren. In deze methode wordt gebruik gemaakt van een specifieke eigenschap van ons becijferingsysteem, namelijk dat cijfers voor ieder proefwerk en ieder schooljaar opnieuw worden geijkt. De onderliggende assumptie is dat de vaardigheid van leerlingen in dezelfde mate toeneemt over de schooljaren. Het gevolg is dat wanneer een leerling een rapportcijfer 6.0 zou halen in het eerste schooljaar, dezelfde leerling wederom een rapportcijfer 6.0 zou halen in het tweede schooljaar. Natuurlijk is de leerling 'verbeterd', maar het werk in het tweede schooljaar is ook 'moeilijker' waardoor de 'verbetering' toch weer uitmondt in een 6.0. Ook deze assumptie is discutabel – er wordt gebruik gemaakt van een eigenschap die beschouwd kan worden als een zwakte van het systeem – maar toch is er vooralsnog geen empirisch bewijs dat aantoont dat deze assumptie niet klopt. In dit onderzoek zal daarom worden verkend of deze methode bruikbaar kan zijn om meer inzicht te krijgen in de hodgepodgehypothese.

2.5 Alternatieve interpretaties van de facetten

Tot nog toe is er gesproken over de facetten docent en docent × leerling interactie als indicators van beoordelingsbias. Dit is gangbaar in literatuur rondom beoordeling en becijfering. Toch zijn er alternatieve interpretaties denkbaar. We gaan hier kort in op de belangrijkste van deze alternatieve interpretaties en bespreken in de laatste alinea de gevolgen die deze alternatieve interpretaties hebben voor deze studie.

Het facet docent beschrijft alle verschillen in rapportcijfers tussen docenten. Behalve door beoordelingsbias kunnen zulke verschillen ook ontstaan omdat één docent een hogere kwaliteit van instructie heeft dan een andere docent. Een alternatieve interpretatie voor het facet docent is dus dat deze de verschillen in kwaliteit van instructie beschrijft (bijv., Hattie, 2009). Ook zouden we deze verschillen kunnen interpreteren als gevolg van een verschil in instroom. Wanneer klassen aan het begin van het schooljaar verschillen in niveau, dan zullen deze verschillen waarschijnlijk leiden tot lagere of hogere rapportcijfers aan het einde van het schooljaar (bijv., Wright, Horn, & Sanders, 1997).

Ook het docent × leerling interactie facet, dat betrekking heeft op dat één docent één leerling een rapportcijfer toekent dat een andere docent niet zou toekennen aan die ene leerling, kan zowel wijzen op bias als op een legitiem verschil. Zo kunnen legitieme verschillen ontstaan door succesvolle differentiatie in de instructie. Bij differentiatie in instructie besteedt een docent meer tijd aan een leerling die dat nodig heeft. Wanneer twee docenten verschillen in hun keuze welke leerling meer tijd nodig heeft, of in het geval dat één docent wel succes heeft met de differentiatie maar de andere docent niet, zal dit leiden tot een legitiem verschil in de rapportcijfers.

We stellen vast dat de facetten docent en

docent × leerling interactie dus niet alleen beoordelingsbias voorstellen, maar ook informatie kunnen bevatten over legitieme verschillen zoals verschil in kwaliteit van instructie, verschillen in instroom en verschil in differentiatie. Deze legitieme verschillen worden in deze studie meegewogen als bias. Het gevolg is dat in deze studie de mate van bias wordt overschat.

Op basis van bovenstaande richten we ons bij het beantwoorden van de hoofdvraag: *In welke mate worden beoordelingen op basis van door docenten gegeven proefwerkcijfers vertekend door beoordelingsbias?* op de volgende vier deelvragen:

- In welke mate verschillen docenten in mildheid van beoordelen?
- In welke mate wordt beoordeling door docenten vertekend door hodgepodge?
- Hoe betrouwbaar zijn rapportcijfers voor relatieve beoordeling?
- Hoe betrouwbaar zijn rapportcijfers voor absolute beoordeling onvoldoende - voldoende?

3 Methode

In dit onderzoek wordt ingegaan op de beoordelingsbias van docenten door een analyse van de betrouwbaarheid in schoolcijfers. Als methode gebruiken we Generaliseerbaarheidstheorie (Brennan, 2001; Shavelson & Webb, 1991). In deze methode wordt een G-studie uitgevoerd gevolgd door een D-studie. In een G-studie wordt de grootte van de facetten geanalyseerd. Deze grootte wordt uitgedrukt in een percentage variantie die door dit facet kan worden verklaard. De aandacht in de analyse gaat uit naar de grootte van de beoordelingsbias; de facetten docent en docent × leerling interactie. In de D-studie wordt geanalyseerd wat de gevolgen zijn van de beoordelingsbias op de betrouwbaarheid van de rapportcijfers.

Het artikel bestaat uit twee deelstudies. In de eerste deelstudie analyseren we de beoordelingsbias in het geneste design. In het geneste design kan alleen de mildheidshypothese worden getoetst en kan dus niet een compleet antwoord geven op onze hoofdvraag. Deze eerste studie heeft twee functies; (1) het verkennen van beoordelingsbias vanuit een herkenbaar startpunt en (2) de resultaten uit de eerste studie kunnen in de tweede studie worden gevalideerd wat de resultaten van beide deelstudies versterkt. In de tweede studie analyseren we beoordelingsbias in een gekruist design. In deze tweede studie wordt de hodgepodgehypothese getoetst.

3.1 Studie 1

3.1.1 Steekproef en procedure

Voor dit eerste onderzoek zijn 19461 cijfers verzameld. De steekproef bestaat uit alle cijfers die zijn gegeven aan 391 leerlingen in het voortgezet onderwijs (VO) gedurende één schooljaar. De tijdstippen waarop werd becijferd verschilde per klas. Deze 19461 cijfers zijn gevrijwaard van herkansingen. Verder zijn ook de kerst-, paas-, en zomerrapport gemiddeldes uit de cijferbestanden verwijderd.

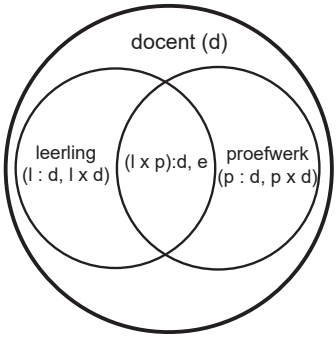
Het aantal cijfers per leraar schommelde van 29 cijfers in één schooljaar (vak Engels) tot 5 cijfers in één schooljaar (vak geschiedenis). Ook binnen vakken is er behoorlijke variatie in het aantal cijfers die docenten geven: collega-docenten Engels op andere scholen gaven bijvoorbeeld soms 11 cijfers. Dit toont de behoorlijke disbalans in de data en deze disbalans zou de resultaten – met name de grootte van de facetten – kunnen vertekenen (Brennan, 2001). Om dit te ondervangen is er gekozen om per docent acht cijfers willekeurig te selecteren. In het geval dat een leraar minder dan acht cijfers had gegeven gedurende het schooljaar werden alle cijfers geselecteerd. Na deze selectie bleven er 12190 cijfers over.

De cijfers zijn gegeven door 64 docenten die lesgaven aan 16 verschillende klassen verspreid over 7 scholen. Voor iedere klas zijn de cijfers voor geschiedenis, wiskunde, Nederlands en Engels verzameld. De leerlingen zaten in de onderbouw (11 – 15 jaar) van het voortgezet middelbaar beroepsonderwijs (VMBO), het hoger algemeen vormend onderwijs (HAVO) of het voorbereidend wetenschappelijk onderwijs (VWO).

De klassengrootte varieerde van 15 tot 29 met een gemiddelde van 25 waarbij er een sterk verband is tussen de klassengrootte en de onderwijssoort: in het VMBO telden alle klassen minder dan 25 leerlingen. Opnieuw is er sprake van een disbalans, ditmaal door verschil in klassengrootte. Gezien de vraag of er variatie is tussen docenten, en gezien de variatie tussen klassen niet goed kan worden onderscheiden van variatie tussen docenten, is besloten om deze disbalans te reduceren door van iedere klas 12 leerlingen willekeurig te selecteren. Van de oorspronkelijke 391 leerlingen werden er 192 geselecteerd. Het aantal cijfers daalde hierdoor van 12190 tot 5988.

3.1.2 Design

Het design heeft een geneste structuur (zie Figuur 4). De observaties bij een leerling zijn genest in docent (l : d). De ":" betekent 'genest in'. Naast dat leerlingen genest zijn in docenten heeft het design ook een facet proefwerk (p). Deze proefwerken zijn ingevoerd in de data als een *long form* (de Boeck, et al. 2011). Bij een *long form* worden alle proefwerkcijfers onder elkaar gezet in één kolom. Proefwerken worden beschouwd als gekruist met het facet leerling: alle leerlingen bij een docent hebben dezelfde proefwerken gemaakt. Ook worden proefwerken beschouwd als genest in docent: iedere docent geeft zijn of haar leerlingen andere proefwerken. Het interactie facet proefwerk × docent (p × d) is *confounded* met het facet proefwerk. Het interactie facet docent × leerling is *confounded* met het facet leerling. Het interactie facet leerling × proefwerk (l × p) is *confounded* met het residu (*e*).



Figuur 4.
Een Venndiagram weergave van het geneste design van de eerste studie (l × p): d.

3.1.3 Data analyse

In de eerste studie concentreren we ons op de mildheidshypothese. Analyses voor de G-studie zijn uitgevoerd in R met het package lme4 (Bates, Maechler, Bolker, & Walker, 2014). Om met lme4 een G-studie uit te voeren werd een random effects model gespecificeerd. De D-studie is gebaseerd op de output uit de G-studie. De betrouwbaarheidscoëfficiënten zijn berekend op basis van formules zoals in Brennan (2001) en Shavelson en Webb (1991). Voor de relatieve betrouwbaarheid geldt:

$$E(\rho^2) = \frac{\sigma^2_{(l:d)}}{\sigma^2_{(l:d)} + \frac{\sigma^2_{(l \times d):d,e}}{n_p}} \qquad (1)$$

We hebben de absolute betrouwbaarheid uitgerekend relatief aan de cesuur van 5.5. Deze keuze is gemaakt omdat absolute besluiten meestal het meer dichotome karakter hebben van onvoldoende of voldoende. Voor de absolute betrouwbaarheid met afkappunt (λ) hebben we de formule gehanteerd:

$$\Phi_\lambda = \frac{\sigma^2_{(l:d)} + ((\mu_{vak} - \lambda)^2 - (\sigma^2_T))}{(\sigma^2_{(l:d)} + ((\mu_{vak} - \lambda)^2 - (\sigma^2_T))) + \frac{\sigma^2_{(p:d)}}{n_p} + \frac{\sigma^2_{(l \times p):d,e}}{n_p}} \qquad (2)$$

Waarin λ = criterium voldoende-onvoldoende = 5.5; μ_{vak}= het gemiddelde van het schoolvak; $\sigma^2_{(l:d)}$ de variantie tussen leerlingen; $\sigma^2_{(p:d)}$ = de variantie tussen proefwerkcijfers; $\sigma^2_{(l \times p):d,\ e}$ = alle overige variantie; σ^2_T = de optelsom van de drie voorgenoemde facetten gedeeld door hun aantal levels. De subscripts duiden de nesting en kruising van de diverse facetten aan.

3.2 Studie 2

3.2.1 Methode

In het geneste design kan niet worden vastgesteld wat de mate van beoordelingsbias is die ontstaat doordat docenten gelijk werk van twee leerlingen toch ongelijk beoordelen,

bijvoorbeeld vanwege verschillen in de getoonde motivatie, persoonlijkheid of getoonde groei. Deze vormen van bias komen juist in zelfrapportage-onderzoek onder docenten prominent naar voren (bijv., Brookhart, 1994; McMillan, et al., 2002). In deze tweede studie verkennen we een onderzoeksopzet waarin het wel mogelijk is de invloed van zulk hodgepodgegedrag op de beoordeling te onderzoeken.

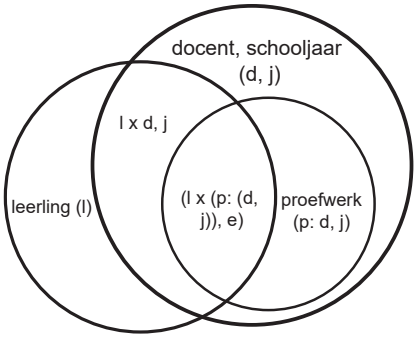
3.2.2 Steekproef en procedure

De steekproef voor de tweede studie bestond uit 56663 cijfers afkomstig van 2 scholen. Deze cijfers zijn gegeven aan proefwerken van alle proefwerksoorten, zoals luistertoetsen, schriftelijke overhoringen, presentaties, praktische opdrachten die werden gegeven aan 424 leerlingen gedurende drie opeenvolgende schooljaren in de onderbouw van de HAVO en VWO. Van deze 424 leerlingen waren van 306 leerlingen voor minimaal twee van de drie opeenvolgende jaren cijfers beschikbaar. De overige 118 (27.7%) leerlingen verhuisden naar een andere schoolsoort of doubleerden. De 118 verhuizers en doubleurs waren ongelijk verdeeld over de 3 opeenvolgende schooljaren; de meeste leerlingen verhuisden tussen het tweede en het derde schooljaar (18.7%). Hierdoor waren van deze leerlingen alleen cijfers beschikbaar in hun derde schooljaar. Enkele leerlingen verhuisden na het brugjaar (2.8%) waarna ze ofwel doubleerden ofwel nogmaals verhuisden, zodat van hen alleen cijfers beschikbaar waren in het tweede schooljaar. De overige 6.5 % verhuisde of doubleerde na het brugjaar. De 118 verhuizers waren ongelijk verdeeld over de klassen; waarbij de meeste klassen 1 tot 5 verhuizers of doubleurs telden. In het derde schooljaar werden, echter, soms bijna complete klassen geformeerd met verhuizers (20 leerlingen). Ook waren de rapportcijfers van deze 118 verhuizers of doubleurs meer homogeen (SD = .66) dan de rapportcijfers van de overige leerlingen (SD = .74). Uiteindelijk is besloten om de 118 verhuizers in de analyse te laten. We bespreken de gevolgen hiervan voor de interpretatie van de resultaten in de discussiesectie beperkingen.

Deze 56663 cijfers zijn gegeven door 52 docenten, waarvan 10 docenten geschiedenis, 12 docenten wiskunde, 16 docenten Nederlands en 15 docenten Engels. Ook in deze steekproef gaven de talen in het algemeen meer cijfers dan wiskunde en geschiedenis en om de disbalans te verkleinen werd besloten om willekeurig 8 cijfers per vak te selecteren. Dit resulteerde in een steekproef van 29462 cijfers. De klassengrootte varieerde van 22 tot 31.

3.2.3 Design

In het tweede design (zie Figuur 5) zijn proefwerkcijfers over meerdere schooljaren verzameld. Het verschil met het 'geneste' design in de eerste studie is dat dit design een apart facet heeft voor de docent × leerling ((d, j) × l) interactie. Omdat leerlingen wisselen van docent in opeenvolgende schooljaren kan dit design nagaan of één leerling door een docent hoger wordt beoordeeld dan dat deze leerling wordt beoordeeld door andere docenten in de vakgroep. Omdat iedere nieuwe docent altijd samengaat met een nieuw schooljaar is het facet docent in deze studie *confounded* met het facet schooljaar (j). Deze *confound* houdt in dat de variantie in cijfers tussen schooljaren wordt opgeteld bij de variantie in cijfers tussen docenten, zodat de variantie in het facet docent twee verklaringen heeft: (1) er is verschil tussen docenten en (2) er is verschil tussen de schooljaren. De assumptie is dat het facet docent vooral verschillen tussen docenten weergeeft, terwijl de schooljaren nauwelijks bijdragen aan de verschillen in het facet docent. Om na te gaan of deze assumptie verdedigbaar is werden de leerlingen geselecteerd die in opeenvolgende jaren dezelfde docenten hadden. In deze groep geeft het facet docent alleen verschillen weer tussen schooljaren en uitgaande dat de assumptie klopt zouden de correlaties de 1.00 moeten naderen. Omdat de leerlingen niet in beide jaren dezelfde proefwerken hebben gemaakt werd de correlatie berekend met de gemiddelde rapportcijfers. De Pearson correlaties tussen de rapportcijfers voor deze leerlingen waren: r = .70 (jaar 1 en jaar 2) r = .67 (jaar 2 en jaar 3) en r = .88 (jaar 1 en jaar 3). Deze correlaties naderen inderdaad de 1.00. Ook waren de correlaties beduidend hoger dan de Pearson correlaties voor de leerlingen die wel wisselden van docent in



Figuur 5.
Een Venndiagram weergave van het gekruiste design van de tweede studie: l × (p : (d, j)).

opeenvolgende jaren; r = .66, r = .37 en r = .55 respectievelijk. De verschillen in rapportcijfers tussen opeenvolgende schooljaren lijken dus inderdaad vooral te informeren over de wisselingen in docent. Net als in het geneste design zijn in dit design de andere interacties (docent × proefwerk (d × p) en leerling × proefwerk (l × p)) *confounded*.

3.2.4 Data analyse

De data analyse met het tweede design richt zich op hodgepodgehypothese. De analyse voor de G-studie is wederom gedaan met het lme4 package van R (Bates, et al., 2014). De D-studie is gebaseerd op de output van de G-studie. In de D-studie richten we ons op de relatieve betrouwbaarheid. De gebruikte formule is:

$$E(\rho^2) = \frac{\sigma^2_{(l)}}{\sigma^2_{(l)} + \frac{\sigma^2_{(d,j\,x\,l)}}{n_d} + \frac{\sigma^2_{l\,x\,(p:(d,j))}}{n_p n_d}} \qquad (3)$$

4 Resultaten

4.1 Studie 1

4.1.1 Beoordelingsbias door mildheid

In de Tabel 1 hieronder worden de resultaten van de G-studie met drie facetten (docent, leerling, proefwerk) gepresenteerd. Alhoewel Tabel 1 meer informatie geeft beperken we ons tot het bespreken in hoeverre de resultaten indicaties geven van beoordelingsbias.

Een indicatie van de mate waarin verschillen in mildheid tussen docenten de beoordeling vertroebelen wordt gegeven door de grootte van het facet docent (d). De resultaten geven weer dat verschillen in mildheid tussen docenten klein tot verwaarloosbaar is (d.w.z. variërend van 0 - 7%). Omdat het facet docent (d) ook een indicatie geeft van legitieme verschillen in de kwaliteit van lesgeven tussen docenten en de verschillen in instroom, kan geconstateerd worden dat de beoordelingsbias door docenten geen substantiële invloed heeft op de hoogte van de rapportcijfers van leerlingen.

4.1.2 Relatieve betrouwbaarheid van de jaarlijkse rapportcijfers

De resultaten geven ook een indicatie van de mate waarin cijfers worden verstoord door alle vormen van bias inclusief beoordelingsbias. Dit kan worden geanalyseerd met een D-studie. Een D-studie geeft een indicatie van de betrouwbaarheid van de rapportcijfers. De berekende betrouwbaarheid is een gemiddelde en deze kan fluctueren per steekproef (Brennan, 2001; Cronbach, et al., 1972). Daarom wordt in G-theorie gesproken over de verwachte betrouwbaarheid. We

Tabel 1
Resultaten G-studie met het geneste design (l × p): d. In de kolom onder '%' staan de percentages variantie voor ieder facet. In de kolom onder 'CI(%)' staat het 95% betrouwbaarheidsinterval.

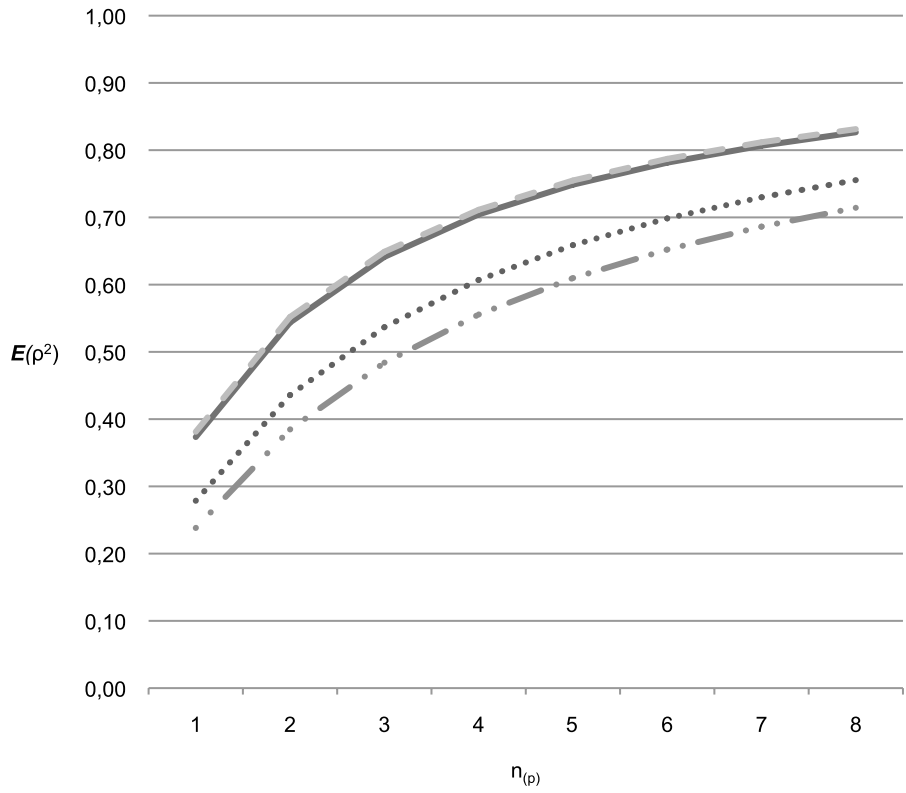
	Geschiedenis		wiskunde		Nederlands		Engels	
Facet	**%**	**CI(%)**	**%**	**CI(%)**	**%**	**CI(%)**	**%**	**CI(%)**
docent (d)	.07	.036 - .160	.02	.010 - .048	.00	.003 - .013	.00	.000 - .000
proefwerk (p)	.22	.169 - .283	.21	.169 - .277	.25	.182 - .299	.20	.183 - .274
leerling (l : d)	.27	.221 - .331	.29	.240 - .359	.19	.152 - .227	.22	.161 - .266
residu (l × p) : d, *e*	.45	.417 - .487	.47	.443 - .512	.63	.543 - .626	.57	.535 - .616

presenteren twee D-studies. De eerste voor de betrouwbaarheid van de relatieve beoordeling ($E\rho^2$) en de tweede voor de betrouwbaarheid van de absolute beoordeling voor de cesuur voldoende-onvoldoende (Φ_λ). In Figuur 6 wordt de verwachte betrouwbaarheid voor relatieve beoordelingen weergegeven voor ieder vak apart. Hierin is de verwachte betrouwbaarheid voor geschiedenis weergegeven met de ononderbroken lijn, voor wiskunde met de streepjeslijn, voor Nederlands met de stippel-streepjeslijn, en voor Engels met de stippellijn.

Uit Figuur 6 blijkt dat de bias in rapportcijfers afneemt naarmate er meer cijfers worden gegeven, maar ook dat de betrouwbaarheid van beoordelingen bij wiskunde en geschiedenis een hoger plafond heeft dan bij de talen. Dit zou het gevolg kunnen zijn van de grotere variatie in (deel)vaardigheden die in de talen worden getoetst: mondelinge taalvaardigheid, schrijfvaardigheid, leesvaardigheid, literatuur, grammatica en spelling.

4.1.3 Absolute betrouwbaarheid van de jaarlijkse rapportcijfers

Bij absolute besluiten gaat het erom hoe zeker we zijn dat een rapportcijfer lager of hoger is dan een gegeven criterium; in dit geval 5.5. We doen hiervoor opnieuw een D-studie. De resultaten (zie Figuur 7) laten zien dat docenten op een betrouwbare manier de excellente (> 9.0) en de 'onvoldoende' leerlingen (< 5.5) beoordelen. De laagste betrouwbaarheid is bij een rapportcijfer 7.0. Leerlingen die een rapportcijfer 7.0 krijgen toegekend hebben relatief veel proefwerkcijfers die hoger of lager dan een 7.0 waren. Daarom neemt



Figuur 6.

Grafiek van de D-studie gebaseerd op eerste geneste design (l × p): d. Op de y-as de verwachte relatieve betrouwbaarheid ($E\rho^2$). Op de x-as het aantal proefwerken. De ononderbroken lijn = geschiedenis, de streepjeslijn = wiskunde, de stippel-streepjeslijn = Nederlands, en de stippellijn = Engels.

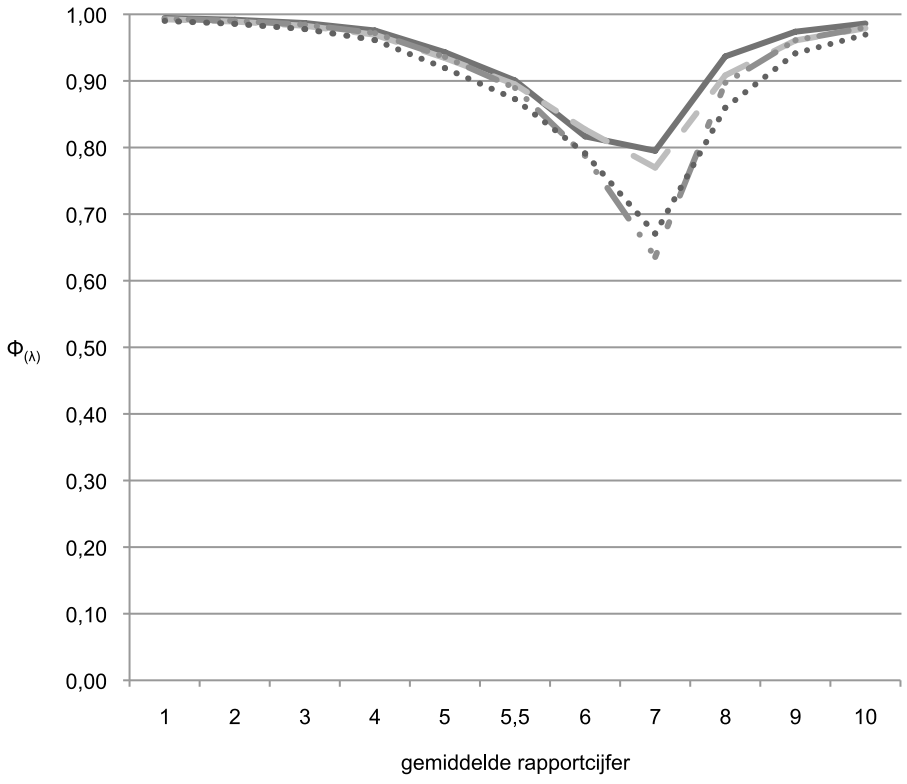
de onzekerheid toe of een leerling echt een rapportgemiddelde 7.0 zou moeten krijgen of eerder een 6.0 of een 8.0. Deze resultaten gaan er vanuit dat er acht proefwerken gedurende het jaar zijn gegeven.

Uit de Figuur 7 blijkt dus dat er consistentie is over de cesuur onvoldoende - voldoende. Wanneer er acht (of meer) proefwerken gedurende het schooljaar worden gegeven is de verwachte betrouwbaarheid van een beoordeling over de cesuur voldoende-onvoldoende voor geschiedenis: Φ_{λ} = .90, voor wiskunde: Φ_{λ} = .90, voor Nederlands: Φ_{λ} = .89 en voor Engels: Φ_{λ} = .87. De cijfers van de meeste leerlingen liggen dus consistent boven of consistent onder het criterium van 5.5, waardoor docenten voor de meeste leerlingen met behoorlijke zekerheid tot een beoordeling kunnen komen over doubleren.

4.2 Studie 2

4.2.1 Beoordelingsbias door 'hodgepodge' beoordeling

De resultaten in Tabel 2 geven een eerste inzicht in de grootte van de beoordelingsbias door de docent × leerling interactie in schoolcijfers. Deze eerste verkenning bij twee scholen en 52 docenten wijst erop dat deze interactie slechts een klein deel van de variantie in schoolcijfers kan verklaren (3 - 7%). Ervan uitgaand dat de variatie in het facet docent × leerling ook legitieme verschillen door differentiatie in instructie weergeeft, blijft er slechts ruimte voor een zeer klein percentage



Figuur 7.

De grafiek van de D-studie naar de betrouwbaarheid van absolute beoordelingen. Op de y-as de absolute betrouwbaarheid voor een cesuur besluit voldoende-onvoldoende ($\Phi_{(\lambda)}$). Op de x-as het gemiddelde rapportcijfer op basis van 8 cijfers. De ononderbroken lijn = geschiedenis, de streepjeslijn = wiskunde, de stippelstreepjeslijn = Nederlands, en de stippellijn = Engels.

Tabel 2

Resultaten G-studie van het tweede design: l × (p : (d, j)).In de kolom onder '%' staan de percentages variantie voor ieder facet. In de kolom onder 'CI(%)' staat het 95% betrouwbaarheidsinterval.

	geschiedenis		wiskunde		Nederlands		Engels	
Facet	**%**	**CI(%)**	**%**	**CI(%)**	**%**	**CI(%)**	**%**	**CI(%)**
docent(d, j)	.07	.037 - .096	.03	.020 - .053	.09	.059 - .153	.03	.021 - .055
proefwerk (p)|	.19	.162 - .227	.14	.119 - .165	.16	.136 - .189	.16	.138 - .192
leerling (l)	.23	.205 - .269	.30	.268 - .351	.17	.149 - .197	.26	.229 - .304
docent × leerling (d, j × l)	.07	.063 - .078	.07	.062 - .076	.04	.034 - .042	.03	.029 - .036
Residu (l × (p : d, j), e)	.45	.437 - .468	.45	.438 - .467	.55	.529 - .565	.51	.497 - .530

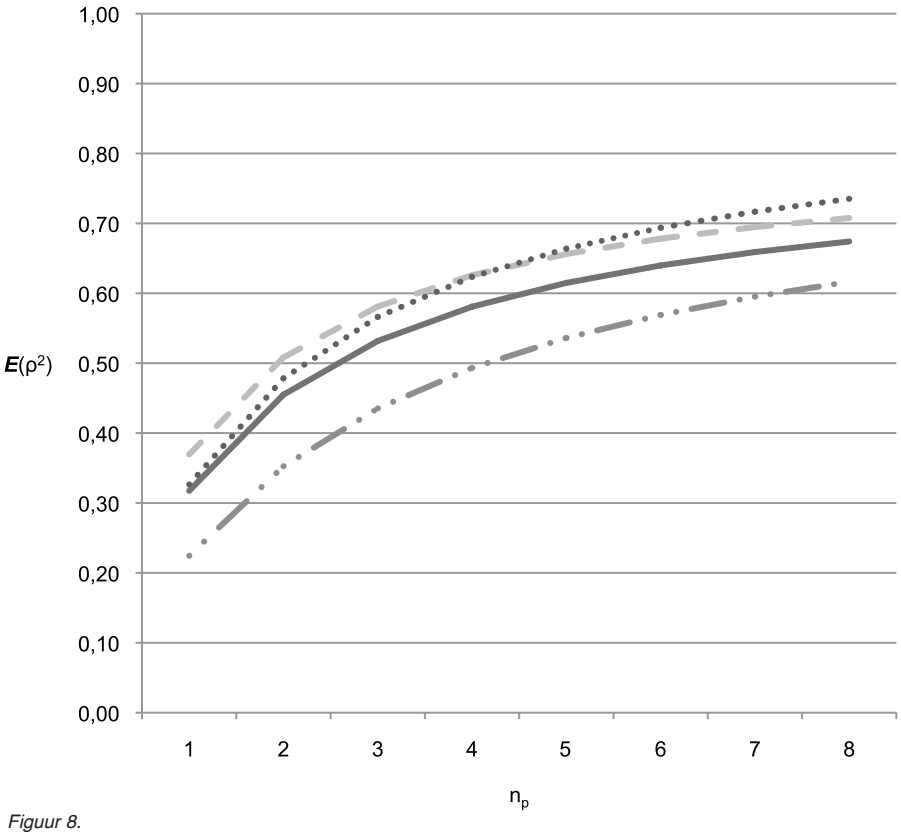
gevallen waarin docenten individuele leerlingen hoger beoordelen op basis van motivatie, persoonlijkheid of andere kenmerken anders dan de leerlingvaardigheid. Als er al sprake is van hodgepodge-beoordeling dan lijkt deze dus inconsistent over proefwerken.

Daarmee kan geconstateerd worden dat de mate van beoordelingsbias die ontstaat doordat leraren gelijkwaardig werk toch anders becijferen – een praktijk die naar voren komt uit zelfrapportage-onderzoek onder docenten – in deze steekproef niet zo sterk naar voren komt. Misschien dat één derde van de docenten dit doet, maar ze doen dit dan wel met maar enkele leerlingen en de grootte van deze bias is in de orde van tienden op het uiteindelijke rapportcijfer.

Dit design en deze steekproef bevestigen ook de eerdere resultaten over de mildheidshypothese. Ook hier zijn de percentages klein tot verwaarloosbaar (3 – 9%) en kan geconstateerd worden dat deze kleine verschillen in beoordeling tussen docenten niet alleen het gevolg zijn van bias, maar waarschijnlijk ook van legitieme verschillen. Er lijkt ook hier weinig reden om te argumenteren dat schoolcijfers in hoge mate bias vertonen.

Als laatste stellen we vast dat in beide designs de meeste variantie niet kan worden toegeschreven aan de meegewogen facetten. Het percentage residuele variantie schommelt in beide designs tussen de .45 en de .63. Hieruit kunnen we concluderen dat één enkel proefwerkcijfer maar zeer beperkt informeert over de vaardigheid van de leerling, maar eveneens beperkt informeert over beoordelingsbias van docenten of over de kwaliteit van het proefwerk. De voornaamste reden waarom docenten meerdere proefwerkcijfers dienen te verzamelen om tot betrouwbare beoordelingen te komen moet daarom gezocht worden in dit residu. Wanneer onderzoek wil nagaan hoe de becijfering efficiënter gemaakt zou kunnen worden – zodat minder cijfers nodig zijn om tot een betrouwbare beoordeling te komen – dan is het nodig om meer facetten mee te wegen dan alleen de kwaliteit van proefwerken en de beoordelingsbias van docenten.

Bovengenoemde resultaten suggereren dat de gemiddelde rapportcijfers dus redelijk vrij zijn van beoordelingsbias door de docent en docent × leerling interactie. Vervolgens is met een D-studie gepoogd een beter beeld te krijgen van de effecten op de betrouwbaarheid van beoordeling (Figuur 8). Hierin is de verwachte betrouwbaarheid voor geschiedenis weergegeven met de ononderbroken lijn, voor wiskunde met de streepjeslijn, voor Nederlands met de stippel-streepjeslijn, en voor Engels met de stippellijn. Het meewegen van de docent × leerling interactie zorgt voor een daling in de betrouwbaarheid. Deze is groter bij geschiedenis ($\Delta E\rho^2$ = .11) en wiskunde ($\Delta E\rho^2$ = .12), omdat de docent × leerling interactie hier een klein effect heeft (7%) dan bij Nederlands ($\Delta E\rho^2$ = .02) en Engels ($\Delta E\rho^2$ = .02), omdat bij de talen docent × leerling interacties verwaarloosbaar klein zijn (4 en 3%). De bovenstaande schattingen van de betrouwbaarheid van rapportcijfers geven weinig grond om te suggereren dat rapportcijfers, mits minimaal 8 proefwerken zijn gegeven, onbetrouwbaar zijn of in grote mate vertroebeld worden door bias.



Figuur 8.

Grafiek van de D-studie gebaseerd op het tweede gekruiste design (I × (p : (d, j)). Op de y-as de verwachte relatieve betrouwbaarheid ($E\rho^2$). Op de x-as het aantal proefwerken bij n_d = 1. De ononderbroken lijn = geschiedenis, de streepjeslijn = wiskunde, de stippel-streepjeslijn = Nederlands, en de stippellijn = Engels.

5 Conclusies en Discussie

In eerder zelfrapportage-onderzoek onder docenten werd gesteld dat docenten subjectieve beoordelaars zijn die niet tot een betrouwbare beoordeling van leerlingen kunnen komen. Docenten zouden kenmerken anders dan de getoonde vaardigheid, zoals motivatie, persoonlijkheid en de doorgemaakte ontwikkeling, meewegen in het uiteindelijke cijfer waardoor leerlingen met gelijke kwaliteiten toch anders beoordeeld worden (hodgepodgehypothese). Ook zouden docenten verschillen in mildheid van beoordelen (mildheidshypothese). In dit verkennende onderzoek is geen overtuigend bewijs gevonden voor deze beide hypotheses. De rapportcijfers van leerlingen lijken niet in hoge mate vertroebeld doordat leerlingen van gelijke kwaliteiten anders worden beoordeeld. Verder lijken de beoordelingen van leerlingen ook niet in grote mate te worden vertroebeld door verschillen in mildheid tussen docenten. We concluderen dat de rapportcijfers – mits minimaal 8 proefwerken worden gegeven in een schooljaar – redelijke betrouwbaarheid hebben voor relatieve beoordelingen ($E\rho^2 \geq .70$) en goede betrouwbaarheid hebben voor absolute beoordelingen over de cesuur van voldoende-onvoldoende ($\Phi_\lambda \approx .90$). Deze conclusie is in lijn met conclusies van andere recente werken waarin is geconcludeerd dat subjectiviteit in becijfering niet een zeer grote rol kan spelen gezien de correlaties tussen schoolcijfers en andere variabelen voor schoolsucces (bijv., Bowers, 2010; Südkamp, Kaiser, & Möller, 2012).

Dit betekent niet perse dat het percentage docenten dat rapporteert hodgepodgegedrag te vertonen incorrect is of dat docenten overdrijven. Het is nog steeds mogelijk dat 37-39% van de docenten soms hodgepodgegedrag vertonen. De resultaten suggereren eerder dat het aantal leerlingen waarbij deze 37-39% docenten ervoor kiezen om ook anderen facetten dan de vaardigheid te laten meewegen in hun beoordeling laag is en/of dat docenten niet telkens bij dezelfde leerlingen hodgepodgegedrag vertonen. Leerlingen worden vooral beoordeeld op basis van hun vaardigheid.

We sommen hieronder enkele andere belangrijke resultaten van deze studie op: (1) Voor rapportcijfers die zijn gebaseerd op minder dan 8 proefwerkcijfers is verwachte betrouwbaarheid voor de relatieve besluiten lager dan het criterium van .70. Er zijn dus minimaal 8 proefwerken nodig om tot een consistente rangorde te komen van minst tot meest vaardig in een schoolvak; (2) De subjectiviteit in beoordeling is het hoogst bij rapportcijfers van een 7.0. Voor een 7.0 is het dus het meest onzeker of dit ook echt een 7.0 is of eigenlijk een 6.0 of een 8.0; (3) Onbetrouwbaarheid in rapportcijfers kan vooral verklaard worden door verschillen in de kwaliteit van proefwerken. De verschillen in kwaliteit van proefwerken wegen zwaarder dan de subjectiviteit van beoordelaars; (4) Er zijn relevante verschillen tussen schoolvakken in de betrouwbaarheid van beoordeling waarbij de rapportcijfers voor de talen en vooral bij Nederlands, een lagere betrouwbaarheid hebben dan de beoordelingen voor geschiedenis en wiskunde. Toch geldt ook hier dat deze lagere betrouwbaarheid niet lijkt te komen door hogere subjectiviteit in beoordeling van de vakdocenten Nederlands of Engels.

5.1 Methodologische beperkingen

Bij de interpretatie van de resultaten van deze studie moet rekening worden gehouden met een aantal beperkingen. De belangrijkste beperking is de geringe steekproefgrootte van 64 en 52 docenten. Hierdoor is het aantal docenten per schoolvak gering. De resultaten voor de mildheidshypothese konden in de beide steekproeven worden geanalyseerd en deze kruis-validatie laat zien dat de resultaten plausibel zijn. Toch konden de resultaten voor de hodgepodgehypothese niet in beide steekproeven worden geanalyseerd en de grootte van de tweede steekproef – twee scholen en 58 docenten – beperkt de generaliseerbaarheid van de resultaten.

Een tweede beperking is dat bij deze methode opeenvolgende cijfers beschouwd worden als onafhankelijke waarnemingen. De assumptie van onafhankelijke waarnemingen wordt regelmatig geschonden, maar in het specifieke geval van schoolcijfers is onduidelijk hoe ernstig deze schending is.

De resultaten van de tweede deelstudie worden ook beperkt door de uitval (27.7%). De meeste uitvallers betroffen leerlingen die verhuisden naar een andere schoolsoort door buitengewoon hoge of lage prestaties. Uit het resultaat in Figuur 7 blijkt dat leerlingen aan de uitersten van de cijferschaal met hoge betrouwbaarheid worden beoordeeld. Met het wegvallen van deze leerlingen is de heterogeniteit in de steekproef toegenomen. Het meest logische gevolg van deze uitval is daarom dat we de beoordelingsbias hebben overschat.

Als laatste punt erkennen we dat het gebruikte criterium voor betrouwbaarheid: ($E\rho^2 \geq .70$) voor discussie vatbaar is. Psychometrici hebben geargumenteerd dat het criterium $\geq .70$ adequaat is voor explorerend en fundamenteel onderzoek, maar dat besluiten waarvan veel afhangt voor de personen in kwestie een hogere betrouwbaarheid vereisen (bijv., Nunnally, 1978). We merken op dat bij het criterium $\geq .70$ ook in het geval van een 'betrouwbaar' rapportcijfer nog steeds een aanzienlijk deel van het besluit wordt bepaald door facetten anders dan de vaardigheid.

5.2 Praktische en theoretische relevantie

Het is lang bekend dat een beoordeling op basis van één enkele proefwerk kwetsbaar is, omdat de leerling een minder moment kan hebben (Spearman, 1910). Om een goed beeld te krijgen van de kwaliteiten van een leerling zijn meerdere observaties nodig, maar schoolvakken verschillen in

hoge mate in de hoeveelheid proefwerken die ze geven aan de leerlingen. Uit dit onderzoek blijkt dat er minimaal 8 cijfers nodig zijn om tot een redelijk betrouwbare beoordeling te komen. Dit betekent dat er bij geschiedenis in de meeste gevallen meer cijfers nodig zijn. Voor de talen is het lastiger om op basis van dit onderzoek tot een advies te komen over het aantal proefwerken. We hebben gespeculeerd dat de lagere betrouwbaarheid bij de talen het gevolg zou kunnen zijn van de diversiteit in de (deel)vaardigheden die in dit vak worden beoordeeld. De veronderstelling is dat een leerling die bijvoorbeeld literair competent is, niet ook sterk hoeft te zijn in andere onderdelen zoals spreken of schrijven. Als deze veronderstelling klopt dan zijn er voor het vak Nederlands veel meer proefwerken nodig, omdat iedere aparte (deel) vaardigheid dan een aantal keren becijferd zou moeten worden. Wanneer deze veronderstelling niet klopt, dan kunnen de talen volstaan met 10-12 cijfers in een schooljaar.

Relevant voor de onderzoekspraktijk is dat de resultaten suggereren dat – mits er minimaal 8 cijfers zijn gegeven – de rapportcijfers een waardevolle bron van de informatie zijn over de kwaliteiten van een leerling. Dit suggereert dat de rapportcijfers bruikbaar zijn voor onderzoek naar het leren van leerlingen. Tegelijk suggereren de resultaten ook dat er weinig variatie is in rapportcijfers tussen docenten. Dit suggereert dat rapportcijfers onbruikbaar zijn voor onderzoek naar het functioneren van docenten. Andere vormen van data zijn nodig voor zulk onderzoek.

5.3 Mogelijkheden voor toekomstig onderzoek

Naast replicatie van de huidige resultaten in een grotere steekproef geven de resultaten van dit onderzoek nog andere suggesties voor vervolgonderzoek. Uit de resultaten blijkt dat de variatie tussen proefwerken groter is dan de variatie tussen beoordelaars. Vervolgonderzoek zou zich kunnen toeleggen op de vraag waarom klasgemiddeldes variëren van proefwerk op proefwerk. Er kunnen minimaal drie hypotheses worden getoetst: (1) de klasgemiddelde prestaties op proefwerken fluctueren omdat de proefwerken verschillen in moeilijkheid. Deze eerste hypothese zou erop duiden dat meerdere docenten die dezelfde proefwerken geven ook op dezelfde proefwerken lagere en hogere klasgemiddelde cijfers behalen; (2) de klasgemiddelde prestaties op proefwerken fluctueren omdat de docent in sommige onderwerpen beter lesgeeft dan in andere onderwerpen. Deze tweede hypothese zou erop duiden dat wanneer een docent klassen verschillende proefwerken geeft over hetzelfde onderwerp alle klassen voor hetzelfde lesonderwerp lagere of hogere cijfers behalen; (3) de klasgemiddelde prestaties op proefwerken fluctueren omdat de docent compenseert. Deze derde hypothese duidt op een vorm van beoordelingsbias die hier niet kon worden onderzocht. De hypothese voorspelt een hoge docent × proefwerk interactie.

Ten tweede suggereren de resultaten dat de betrouwbaarheid in beoordeling lager is bij de talen – en vooral bij Nederlands – dan bij geschiedenis en wiskunde. Vervolgonderzoek zou zich ook kunnen richten op het vakspecifieke in het betrouwbaar beoordelen van de vaardigheid bij Nederlands. In dit artikel hebben we gespeculeerd dat de lagere betrouwbaarheid in beoordeling bij het vak Nederlands deels zou kunnen liggen aan de uiteenlopende deelvaardigheden die in dit vak worden gedoceerd. Toekomstig onderzoek zou kunnen nagaan in hoeverre deze veronderstellingen gerechtvaardigd zijn en hoe hier beter mee omgegaan kan worden in de beoordeling.

Noot

[1] De eerste auteur wil Marjon Fokkens-Bruisma bedanken voor haar bereidheid om feedback te geven en mee te denken over de tekst.

Literatuur

Atkinson, R. C., & Geiser, S. (2009). Reflections on a Century of College Admissions Tests. *Educational Researcher, 38,* 665-667. DOI: 10.3102/0013189X09351981

Bates, D., Maechler, M., Bolker, B., Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using S4 Classes*. R package version 1.1-7. URL: http://CRAN.R-project.org/package=lme4

Biesta, G. J. J. (2012). *Goed onderwijs en de cultuur van het meten.* Den Haag: Boom Lemma.

Bowers, (2009). Reconsidering grades as data for decision making: More than just academic knowledge. *Journal of Educational Administration, 47*, 609-629. DOI: 10.1108/09578230910981107

Bowers, (2010). Grades and graduation: A longitudinal risk perspective to identify student dropouts. *Journal of Educational Research, 103,* 191-207. DOI: 10.1080/00220670903382970

Bowers, (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high-school. *Educational Research and Evaluation: An International Journal on Theory and Practice, 17,* 141-159. DOI: 10.1080/13803611.2011.597112

Brennan, R. L., (2001). *Generalizability Theory: Statistics for Social Science and Public Policy*. NY: Springer-Verlag, Inc

Brennan, R. T., Kim, J., Wenz-Gros, M., & Siperstein, G. M. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts comprehensive assessment system (MCAS). *Harvard Educational Review, 71,* 173-216.

Brookhart, S. M. (1994). Teachers' Grading: Practice and Theory. *Applied Measurement in Education, 7,* 279-301.

Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teacher College Record, 106,* 429-458.

Cliffordson, C. (2008). Differential prediction of study success across academic programs in the Swedish context: The validity of grades and tests as selection instruments for higher Education. *Educational Assessment, 13*, 56-75. DOI: 10.1080/10627190801968240

Cronbach, L. J., Gleser, C. G., Rajaratnam, N., & Nanda, H. (1972). *The dependability of behavioral measurements.* New York: Wiley.

Cross, L., H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by teachers and students alike. *Applied Measurement in Education, 12,* 53-72.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Abe, H., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. Journal of Statistical Software, 39, 1–25.

De Groot, A. D., & Wijnen, W. H. F. W. (1983). *Vijven en zessen. Cijfers en beslissingen: het selectieproces in ons onderwijs (10de druk).* Groningen, Nederland: Wolters Noordhoff

Drany, K., & Wilson M. (2008). An LLTM approach to the examination of teachers' ratings of classroom assessment tasks. *Psychology Science Quarterly, 50*, 417-432.

Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement.* New York, NY: Routledge

Hofstee, W., K., B. (1999). *Principes van beoordeling: Methodiek en ethiek van selectie, examinering en evaluatie.* Amsterdam, Nederland: Swets & Zeitlinger

Kuhlemeier, H., & Kremers, E. (2013). *De praktijk van de eerste en tweede correctie. Samenvatting van onderzoek naar het functioneren van het CSE.* Arnhem: Cito

Korobko, O., B., Glas, C. A. W., Bosker, R. J., & Luyten, J., W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement, 45,* 139-157.

Marzano (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education, 15,* 249-268. DOI: 10.1207/S15324818AME1503_2

McMillan , J. H., Myran, S., & Workman, D. (2002). Elementary Teachers' Classroom Assessment and Grading Practices, *The Journal of Educational Research, 95*, 203-213. DOI: 10.1080/00220670209596593

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation: An International Journal on Theory and Practice, 12,* 53–74.

Nunnally, J. C. (1978). *Psychometric Theory (2nd edition).* NY: McGraw-Hill.

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education, 26*, 1372–1380. DOI: 10.1016/j.tate.2010.03.008

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer.* Thousand Oaks, California: Sage Publications, Inc

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-295.

Standaert, R. (2014). *De becijferde school. Meetcultus en meetcultuur.* Leuven, België: Acco

Starch & Elliot, (1914). Reliability of grading work in mathematics. *The school review, 21,* 254-259.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgements of students' academic achievement: a meta-analysis. *Journal of Educational Psychology, 104,* 743-762.

Thorsen, C., & Cliffordson, C. (2008). The predictive validity of teacher-assigned criterion-referenced grades. *Educational Research and Evaluation: An International Journal on Theory and Practice,18,* 153-172. DOI: 10.1080/13803611.2012.659929

Wright, S., P., Horn, S., P., & Sanders, W., L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of personnel evaluation in education, 11,* 57-67.

Auteurs

Rikkert van der Lans is promovendus aan de lerarenopleiding van de Rijksuniversiteit Groningen. **Wim van de Grift** is hoogleraar onderwijskunde aan de Lerarenopleiding van de Rijkuniversiteit Groningen. **Klaas van Veen** is hoogleraar onderwijskunde en directeur van de Lerarenopleiding van de Rijkuniversiteit Groningen.

Correspondentieadres: Rikkert van der Lans, Grote Kruisstraat 2, 9712 TS Groningen. Email r.m.van.der.lans@rug.nl

Abstract

Teachers grading practices: an analysis of the reliability of teacher-assigned grade point average (GPA)

In previous research, teachers report that they use a hodgepodge of factors when grading students. This has led researchers to suspect that teacher-assigned grades are inflated by teacher-student interactions; the hodgepodge hypothesis. Teachers also are reported to differ in grading leniency; the leniency hypothesis. In this study these two hypotheses are investigated. Two samples of teachers-assigned grades were gathered. The first sample contained 5,988 grades awarded by 64 teacher to 192 students during one school year. The second sample contained 29,462 teacher-assigned grades awarded to 306 student by 52 teachers during three subsequent school years. Generalizability Theory is used to analyze bias. The results present little evidence to claim that school grades are considerably biased due to hodgepodge grading or teacher leniency. Unreliability in teacher-assigned grades is more due to the tests than due to teachers' hodgepodge or leniency.