Zijn we het eens? Interbeoordelaarsbetrouwbaarheid in de pedagogiek en het onderwijs

L. A. Van der Ark en D. ten Hove

Samenvatting

In een opvoedkundige context moet vaak beoordeeld worden: een onderwijzer beslist of het gedrag van een leerling bestraft wordt, een docent beoordeelt een scriptie of een raadsonderzoeker geeft een complexe assessment van een verdachte jongere. De interbeoordelaarsbetrouwbaarheid (IBB) geeft de mate weer waarin verschillende beoordelaars het met elkaar eens zijn. In dit paper behandelen we drie vragen over de IBB. De eerste vraag - Hoe bepaal je de IBB? - is niet eenduidig te beantwoorden omdat er veel coëfficiënten zijn om de IBB te schatten, die vaak verschillende resultaten opleveren. De keuze voor een IBB-coëfficiënt wordt bepaald door het doel en het design van het onderzoek, en door persoonlijke voorkeur. Het antwoord op de tweede vraag - Hoe kun je de IBB verhogen? - behelst het verhogen van de kwaliteit van de items, de rubrics en de beoordelingsprocedure, en de kunde van de beoordelaars. De derde vraag - Wanneer is de IBB hoog genoeg? - kan ook niet ondubbelzinnig worden beantwoord. We presenteren methoden om benchmarks voor één IBB-coëfficiënt in een andere te transformeren, maar stellen dat meer methodologisch onderzoek naar IBB vereist is om betere antwoorden op deze vraag te bieden.

Kernwoorden: **Cohen's Kappa, Interbeoordelaarsbetrouwbaarheid, Intra-klasse correlatiecoëfficiënt, Onderwijs, Pedagogiek**

1. Inleiding

Het Nutsseminarium voor Pedagogiek is van grote invloed geweest op het academisch denken over opvoeding in Nederland. Idenburg (1958) schreef dat vanwege het Nutsseminarium de pedagogiek voortaan betrokken zou worden "op het kind in zijn actuele situatie van gezin en school. Observeren en experimenteren zouden wezenlijke onderdelen van de studie gaan worden." Dit is 100 jaar na de oprichting van het Nutsseminarium goed te zien bij de wetenschappelijke pedagogiek in Nederland. Bij de Afdeling Pedagogische en Onderwijswetenschappen (POW) aan de Universiteit van Amsterdam, één van de nazaten van het Nutsseminarium, zijn experimenteren en observeren nog steeds de gangbare methoden bij het pedagogisch onderzoek. Dit artikel richt zich op observeren, en met name op het beoordelen van observaties. Een korte inventarisatie leert dat ongeveer 10% van alle wetenschappelijke artikelen die in 2018 door de afdeling POW gepubliceerd werd gebruik maakte van observatiemethoden. Deze artikelen betroffen vooral onderzoek bij hele kleine kinderen, beoordelingen van docenten, beoordelingen van geschreven werken door leerlingen en meta-analyses. Hieronder zijn ook publicaties in toptijdschriften zoals Child Development (bijv. Leijten et al., 2018), Journal of Abnormal Child Psychology (bijv. Moens, Weeland, Van der Giessen, Chhangur, & Overbeek, 2018) en Review of Educational Research (bijv. Egert, Fukkink, & Eckhardt, 2018). Studenten worden ook bij dit onderzoek betrokken, omdat zij in het kader van hun masterthesis of als onderzoeksassistent veelal gedragingen op video's beoordelen voor lopende onderzoeken.

De eerste auteur bekleedt als bijzonder hoogleraar de Kohnstammleerstoel bij de afdeling POW wegens de Vereniging ter Bevordering van de Studie der pedagogiek (VBSP) met als leeropdracht de 'kwantitatieve onderzoeksmethodologie ter bevordering van de academisering van het onder-

wijs'. De meningen verschillen of deze leeropdracht past bij een leerstoel die naar Kohnstamm genoemd is (zie bijvoorbeeld Levering, 2017). In dit paper demonstreren wij onderzoek dat voorkomt uit deze leeropdracht binnen de Kohnstammleerstoel en laten wij zien hoe dit onderzoek meerwaarde heeft voor de pedagogiek en onderwijskunde. Hierbij gaan wij in op zowel de mogelijkheden als beperkingen van kwantitatieve methoden bij observeren en beoordelen.

Het methodologische onderzoek naar observeren en beoordelen vormt een belangrijke steunpilaar voor het pedagogisch onderzoek, want als de observaties en beoordelingen niet deugen, dan kunnen ook de resultaten van het pedagogisch onderzoek niet worden vertrouwd. Als meerdere beoordelaars hetzelfde gedrag, gesprek of werkstuk observeren, is het wenselijk dat hun beoordelingen overeenkomen. Als dat niet het geval is, dan heeft een beoordelaar mogelijk iets over het hoofd gezien, is de beoordelingsprocedure mogelijk onduidelijk, zijn de beoordelaars mogelijk niet goed getraind, of is datgene dat beoordeeld moet worden multi-interpretabel. Voorbeelden van het laatste zijn het beoordelen van een kind op een onduidelijk item, of het beoordelen van kunst, waarbij de visie van de beoordelaar van invloed kan zijn (Van de Kamp, 2012).

De mate waarin beoordelingen tussen beoordelaars overeenkomen wordt de *interbeoordelaarsbetrouwbaarheid* (IBB) genoemd. Als de IBB gelijk is aan 1 dan komen de beoordelingen van verschillende beoordelaars perfect overeen. Als de IBB gelijk is aan 0 dan zijn de beoordelingen volkomen willekeurig. Voor zowel de wetenschap als de praktijk is een hoge IBB om twee redenen van belang. Ten eerste betekent een lage IBB dat minder precies gemeten wordt. Stel dat twee groepen van 10 beoordelaars de gedragingen van een kind observeren, en het gedrag beoordelen met een rapportcijfer (1 tot 10) om het niveau van zelfredzaamheid van het kind aan te geven. In Groep A wordt vijf keer een 5 en vijf keer 6 gegeven; in Groep B worden alle cijfers van 1 tot en 10 gegeven. In beide gevallen is het geschatte het niveau van zelfredzaamheid gelijk aan M = 5.5, maar in Groep A - waar de beoordelingen redelijk overeen komen - is het niveau van zelfredzaamheid bijna zes keer zo precies vastgesteld ($SD \approx 0.53$) als in Groep B ($SD \approx 3.03$) - waar de beoordelingen niet overeenkomen. Hierdoor bestaat er veel meer vertrouwen in de beoordelingen van Groep A.

Ten tweede betekent een lage IBB dat voorspellingen die gedaan worden met de beoordelingen, bijvoorbeeld in een regressie- of correlatieanalyse, vertekende resultaten opleveren. Uit de literatuur van de correctie attenuatie (Spearman, 1904) is bekend dat effecten kleiner worden als de betrouwbaarheid (en dus ook de IBB) afneemt. Stel dat de zelfredzaamheid van een groep kinderen perfect betrouwbaar beoordeeld is (IBB = 1), dat de zelfredzaamheidsbeoordelingen (X) gebruikt zijn om schoolsucces op de middelbare school (Y) te voorspellen, en dat de correlatie tussen die twee variabelen gelijk is aan $r_{XY} = 0.80$. Als de IBB van de zelfredzaamheidsbeoordelingen niet perfect betrouwbaar gemeten zou zijn (IBB = 0.50), dan zou de correlatie met het schoolsucces naar verwachting gereduceerd zijn tot $0.80 \times \sqrt{0.50} \approx 0.57$: een aanzienlijke afname.

Voor de pedagogische praktijk kan een lage IBB ook onwenselijke gevolgen hebben. Vaak wordt in de praktijk een beoordeling maar door één beoordelaar gedaan, zoals door een docent die cijfers moet geven, of door een hulpverlener die het gedrag van een cliënt moet inschatten. De IBB is dan niet te bepalen omdat daar meerdere onafhankelijke beoordelingen voor nodig zijn. Het niet kunnen bepalen van de IBB is met name een probleem als er voor de beoordeelde veel op het spel staat. Een voorbeeld daarvan zijn beoordelingen met het Landelijk Instrumentarium Jeugdstrafrechtketen (LIJ; zie Van der Put et al., 2011, voor details), een observatie-instrument dat wordt gebruikt om het dynamisch risicoprofiel te bepalen van jongeren die met justitie in aanmerking zijn gekomen. Omdat het LIJ gebruikt wordt voor het bepalen van zowel de behandeling als de strafeis van de jongeren, staat er veel op het spel. Voor zowel de jongere, de officier van justitie, als de behandelaar is het van belang dat de beoordeling het gedrag van de betreffende jongere reflecteert, en zo min mogelijk beïnvloed

wordt door de raadsonderzoeker die het LIJ toevallig heeft ingevuld. Bij dergelijke belangrijke beslissingen is het daarom verstandig de IBB te laten onderzoeken (Van der Ark, Van Leeuwen, & Jorgensen, 2018).

IBB is dus belangrijk voor veel pedagogisch onderzoek. Het is echter niet makkelijk om wijs te worden uit de tal van coëfficiënten die beschikbaar zijn om de IBB te bepalen, om de numerieke resultaten te interpreteren, en om onderzoeken zo op te zetten dat de IBB gemaximaliseerd wordt. In het vervolg van dit artikel gaan we hier verder op in: Eerst behandelen we de definitie van IBB. Vervolgens gaan we in op drie vragen: Hoe bepaal je de IBB?, Hoe kun je de IBB verhogen? en Wanneer is de IBB hoog genoeg? Ten slotte beschrijven we kort lopend methodologisch onderzoek bij POW dat als doel heeft de drie bovengenoemde vragen beter te beantwoorden.

2. Hoe bepaal je de interbeoordelaarsbetrouwbaarheid?

Een algemene beschrijving van IBB is de mate van overeenstemming of gedeelde variantie tussen twee of meer beoordelaars die hetzelfde object of dezelfde persoon beoordelen. Een eenduidige meer precieze beschrijving valt niet te geven, want er bestaan verschillende ideeën over wat IBB precies is. Voor een overzicht verwijzen we naar Gwet (2014) of Hallgren (2012). De meeste onderzoekers associëren IBB met de *Cohens kappa* (Cohen, 1960), maar er is een enorme hoeveelheid aan coëfficiënten die pretenderen de mate van IBB aan te geven. Popping (1988) identificeerde alleen al voor nominale data 38 coëfficiënten. Het computerprogramma irr (Gamer, Lemon, Fellows, & Singh, 2012) kan 17 verschillende coëfficiënten voor de IBB berekenen, en sommige coëfficiënten hebben ook nog verschillende versies. Al deze coëfficiënten schatten de IBB op een andere manier. Zhao, Liu en Deng (2013) typeerden vier verschillende typen IBB-coëfficiënten[1], terwijl IBB-coëfficiënten gebaseerd op generaliseerbaarheidstheorie (bijv. Shavelson, Webb, & Rowley, 1989) een belangrijke vijfde type is, en er mogelijk nog meer typen zijn.

De keuze voor een IBB-coëfficiënt is niet helemaal vrij omdat sommige coëfficiënten niet geschikt zijn voor een bepaald type meetniveau, type beoordeling, of het onderzoeksdesign. De intra-klasse correlatiecoëfficiënt (ICC; Shrout & Fleiss,1979) is bijvoorbeeld niet geschikt voor beoordelingen op nominaal en ordinaal meetniveau, Cohens kappa is alleen geschikt voor dichotome beoordelingen, zoals ja/nee of goed/fout, en de ICC heeft verschillende varianten voor verschillende onderzoeksdesigns (voor een determinatiesleutel, zie bijvoorbeeld Koo & Li, 2016). Echter, het aantal coëfficiënten waaruit gekozen kan worden blijft groot.

Net zoals er verschillende pedagogische stelsels naast elkaar kunnen bestaan (Kohnstamm, 1935, p. 5) waaruit een pedagoog op basis van overtuiging en geloof een keuze maakt, bestaan er ook verschillende vormen van IBB naast elkaar waaruit de onderzoeker op basis van overtuiging een keuze maakt. Verschillende onderzoekers kunnen ook tot verschillende keuzes komen. Deze keuze zou gemaakt moeten worden op basis van de achterliggende statistische theorieën over wat IBB is, hoewel in de praktijk mogelijk ad-hoc overwegingen een rol spelen: Wat is bekend? Wat is beschikbaar in de software? Over de achterliggende theorieën wordt ook gedebatteerd. Krippendorff (2004) en Zhao et al. (2013) beargumenteerden bijvoorbeeld waarom sommige coëfficiënten, waaronder de bekende coëfficiënt Cohens kappa, ongeschikt zouden zijn om inzicht te krijgen in de IBB. Het belangrijkste argument tegen de kappa coëfficiënt is dat deze gevoelig is voor de marginale verdeling van scores.

Onderzoek van Ten Hove, Jorgensen en Van der Ark (2018) maakt duidelijk dat de keuze van de IBB-coëfficiënt een enorme impact kan hebben op de resultaten van een onderzoek. Zij berekenden op vier verschillende datasets 20 verschillende IBB coëfficiënten. Over het algemeen waren de verschillen zeer groot. Voor één dataset lagen de waarden van de IBB coëfficiënten zelfs tussen 0.10 (Krippendorffs alpha) en 0.92 (Finn's twee-weg coëfficiënt). Dus de ene coëfficiënt

gaf aan dat de IBB zeer laag was, terwijl een andere aangaf dat de IBB zeer hoog was.

Wij prefereren de ICC, omdat deze ons inziens theoretisch het best verankerd is, in de generaliseerbaarheidstheorie. Binnen de generaliseerbaarheidstheorie is een beoordeling (X) op te splitsen in een systematisch deel T (van het Engelse *true score*), een beoordelaarseffect R (van het Engelse *rater effect*) en meetfout E (van het Engelse *error*):

$$X = T + R + E. \quad (1)$$

Het systematisch deel (T) is de gemiddelde beoordeling van een persoon, als deze persoon onder verder identieke omstandigheden door oneindig veel verschillende beoordelaars wordt beoordeeld. Het beoordelaarseffect (R) is de systematische neiging van de beoordelaar. Waar de ene beoordelaar bijvoorbeeld structureel neigt om gedrag positief te beoordelen, neigt een andere beoordelaar om gedrag negatief te beoordelen, waardoor de beoordelaar de beoordeling ongewenst beïnvloedt. De meetfout (E) is de niet-systematische ruis die de beoordeling vertroebelt. Idealiter zijn het beoordelaarseffect (R) en de meetfout (E) gelijk aan 0. zodat de beoordeling precies is wat je zou verwachten. De ICC is gedefinieerd als het gedeelte van de variantie in geobserveerde beoordelingen dat toe te schrijven is aan de variantie in verwachte beoordelingen. Als σ^2_X de variantie is van de geobserveerde beoordelingen en σ^2_T de variantie van de verwachte beoordelingen, dan geldt[2]

$$\text{ICC} = \sigma^2_T / \sigma^2_X = \sigma^2_T / (\sigma^2_T + \sigma^2_R + \sigma^2_E). \quad (2)$$

Als alle beoordelaarseffecten (R) en meetfout (E) gelijk zijn aan 0 dan zijn er ook geen verschillen in beoordelaarseffecten en meetfout, dat wil zeggen dat $\sigma^2_R = \sigma^2_E = 0$ en de ICC = 1. Als een beoordeling alleen maar uit meetfout en beoordelaarseffecten bestaat zijn alle verwachte beoordelingen (T) gelijk aan elkaar, en is $\sigma^2_T = 0$ is de ICC gelijk aan 0. Naast een sterke theoretische inbedding heeft de ICC meer voordelen. Door de ICC Bayesiaans te schatten (nog niet mogelijk in bestaande gebruikersvriendelijke software) kan ook goed met ontbrekende gegevens worden omgegaan (Van der Ark et al., 2018). Een nadeel is dat de ICC in principe alleen geschikt is voor continue beoordelingen, maar via een logit transformatie kan de ICC geschikt gemaakt worden voor beoordelingen met geordende categorieën. Alleen voor nominale beoordelingen is de ICC echt niet geschikt en dan ligt onze voorkeur bij Krippendorffs alfa (Krippendorff, 2011), die breed toepasbaar is, goed met ontbrekende gegevens kan omgaan, en niet de eerder genoemde nadelen van Cohens kappa heeft.

3. Hoe kun je de interbeoordelaarsbetrouwbaarheid verhogen?

Ongeacht de coëfficiënt die gekozen wordt om de IBB te schatten zijn er tal van maatregelen die de IBB verhogen. Wij onderscheiden twee verschillende typen maatregelen: Maatregelen met betrekking tot items en rubrics, en maatregelen met betrekking tot procedure en beoordelaars.

3.1 Items en rubrics

Items zijn de stimuli waarop een subject door een beoordelaar beoordeeld wordt, inclusief de mogelijke beoordeelopties. Een beoordeling kan uit één item bestaan (bijv. het aantal seconden die een leerling nodig had om een bepaalde opgave te maken) of uit meerdere items (bijv. het LIJ, waar raadsonderzoekers de jongeren beoordelen op 259 items en subitems onderverdeeld in 10 domeinen). Naar verwachting wordt de variantie van de meetfout (σ^2_E) kleiner als het aantal kwalitatief gelijkwaardige of betere items toeneemt (vergelijk Lord & Novick, 1968, pp. 112-114). Uit Formule 2 blijkt dat wanneer σ^2_E daalt en de overige variantiecomponenten gelijk blijven de IBB toeneemt. Dus meer (goede) items heeft een positief effect op de IBB.

De items dienen zo geformuleerd te worden dat het voor beoordelaars duidelijk is wat er wordt bedoeld. Duidelijk formuleren reduceert ruis, waardoor σ^2_E (Formule 2) daalt en de IBB toeneemt. Hofstee (1991; voor een meer uitgebreide behandeling zie Case & Swanson, 1996)[3] schreef enkele richtlijnen voor het schrijven van items. Tabel 1 geeft twee items uit het LIJ weer met een relatief hoge IBB (item 2.8: ICC = 0.93 ; item 2.10: ICC = 0.83; Van der Ark et al., 2018) en tabel

Tabel 1
Twee vragen uit het LIJ met een relatief hoge IBB

2.8 Schoolprestaties van de jeugdige gedurende de afgelopen zes maanden	O Presteert goed (gemiddeld 7 of hoger) O Presteert voldoende (gemiddeld 6) O Presteert zwak (gemiddeld 5) O Presteert slecht (gemiddeld 4 of lager) O Onbekend
2.10 Spijbelt de jeugdige? Het gaat hier om de afgelopen zes maanden. 'Soms' betekent wel eens gespijbeld, maar geen leerplichtmelding. 'Vaak' betekent: melding bij leerplicht en/of minimaal drie dagen onafgebroken ongeoorloofde afwezigheid of afwezigheid gedurende meer dan een achtste deel van de onderwijstijd in een periode van vier opeenvolgende weken. (NB: Spijbelen aansluitend aan een vakantie telt niet mee)	O Niet of nauwelijks, jeugdige is altijd aanwezig O Soms O vaak O Onbekend

Tabel 2
Richtlijnen voor het Formuleren van Items (Hofstee, 1991) met Voorbeelden

Richtlijn	Item	Geprefereerde formulering	Minder geschikte formulering
Gebruik derde persoon enkelvoud	2.8	Schoolprestaties van de jeugdige …	Jouw schoolprestaties …
Gebruik observeerbare termen	2.8	O Presteert slecht (gemiddeld 4 of lager)	O Presteert slecht
Gebruik standaard Nederlands	2.8	… gedurende de afgelopen zes maanden	… de laatste zes maanden
Vermijd zogenaamde modifiers	2.8	Het gaat hier om de afgelopen zes maanden.	Het gaat hier om de afgelopen zes maanden, hoewel de meningen over de tijdsperiode waarin verzuim geregistreerd moet worden uiteenlopen.
Vermijd suggestief taalgebruik	2.10	Spijbelt de jeugdige?	De jeugdige spijbelt zeker ook?
Vermijd moeilijke woorden	2.10	… onafgebroken ongeoorloofde afwezigheid…	… onafgebroken wederrechtelijke afwezigheid…
Vermijd ontkenningen	2.10	'Soms' betekent wel eens gespijbeld, maar geen leerplichtmelding. …	'Soms' betekent niet dat er niet gespijbeld wordt, maar geen leerplichtmelding. …
Vermijd jargon en idiomatisch taalgebruik	2.10	(NB: Spijbelen aansluitend aan een vakantie telt niet mee)	(NB: Luxe-spijbelen telt niet mee)
Vermijd racistisch, seksistisch, etnocentrisch en androcentrisch taalgebruik	2.8	O Presteert goed (gemiddeld 7 of hoger)	O Hij presteert goed (gemiddeld 7 of hoger)

2 geeft Hofstee's richtlijnen met voor elke richtlijn een voorbeeld van een goede en een minder goede formulering, gebaseerd op de twee items in tabel 1. Als van deze richtlijnen wordt afgeweken is het minder duidelijk welk gedrag geobserveerd moet worden of hoe het gedrag beoordeeld moet worden en moet de beoordelaar intuïtief aanvullen wat de bedoeling. Hierdoor ontstaat naar verwachting meer ruis in de beoordelingen, en neemt σ^2_E naar verwachting toe, en de IBB af. Als onduidelijk is welk gedrag geobserveerd moet worden is het ook mogelijk dat beoordelaarseffecten een grotere rol gaan spelen, en kan σ^2_R toenemen, wat ook een negatief effect heeft op de IBB. Naast de items moeten ook de beoordelingscategorieën duidelijk beschreven zijn in zogenaamde rubrics. Rubrics zijn criteria waarop gedrag beoordeeld wordt, plus de verschillende niveaus waarop deze criteria

zijn beschreven. Hier gelden ook min of meer de richtlijnen uit tabel 2.

Als nog onbekend is hoe de items in de praktijk werken, verdient het de aanbeveling om de items te testen in pilotstudies, waarin zowel beoordelaars als experts een rol spelen. Dit heeft vooral als doel om voor alle beoordelaars onduidelijkheden weg te nemen (reductie van σ^2_E) en om mogelijke beoordelaarseffecten tegen te gaan (reductie van σ^2_R). Experts op het gebied van het te beoordelen construct kunnen vanuit hun expertise mogelijk slechte items eruit halen (reductie van σ^2_E). Met behulp van beoordelaars die proefbeoordelingen doen (bijvoorbeeld van een video met gedragingen van dezelfde persoon) kan men via hardop-denkprotocollen, focusgroepen of interviews erachter komen welke items grote variatie in beoordelingen veroorzaken en waarom dat gebeurt. Op basis van deze pilots kunnen de items of de rubrics aangepast worden. Als er wel al data beschikbaar zijn is het ook mogelijk om voor elk item afzonderlijk de IBB te berekenen, en items met een lage IBB nader te inspecteren, en eventueel aan te passen, de rubrics aan te passen of de items geheel te verwijderen. Voor uitgebreidere mogelijkheden rond het schalen van vragenlijsten waar beoordelaars een rol spelen wordt verwezen naar Snijders (2001, zie ook Koopman, Zijlstra, de Rooij, & Van der Ark, 2019).

3.2 Procedure van de beoordeling en kennis van de beoordelaars

Naast rubrics is ook de beoordelingsprocedure van belang. Een duidelijk beschreven procedure is belangrijk voor de standaardisatie van de beoordeling. Naarmate de procedure beter gestandaardiseerd is, weten observatoren/beoordelaars beter wat van hen verwacht wordt, en nemen naar verwachting zowel de ruis als de beoordelaarseffecten af (σ^2_E en σ^2_R dalen), waardoor de IBB naar verwachting stijgt. Tot de procedure behoren zaken als de setting van de observatie en beoordeling, de introductie van de beoordelaar aan degenen die geobserveerd worden, de mate waarin de observator rapport dient te creëren, de mate van terugkoppeling aan degenen die geobserveerd worden en de maximale tijd tussen de observatie en de beoordeling. Ook is het van belang dat observatoren voldoende tijd en ruimte hebben voor hun beoordelingen en bijvoorbeeld niet tijdens een beoordelingsprocedure weggeroepen worden voor andere zaken. Een bijkomend voordeel van een goed beschreven procedure is dat andere studies er ook gebruik van kunnen maken, wat het mogelijk maakt resultaten over studies te vergelijken (Thomassin, Raftery-Helmer, & Hersh, 2018).

Ten slotte dient de observator/beoordelaar zelf op de hoogte te zijn van de psychologische processen die een rol spelen bij het observeren en beoordelen. Het gaat hier om zogenaamde cognitieve vertekeningen (Tversky & Kahnemann, 1974; zie ook Kahnemann, 2011). De bekendste is waarschijnlijk de *halo bias*, waarbij een globale positieve of negatieve indruk tijdens een observatie de beoordelingen op meer specifieke aspecten beïnvloedt (Cooper, 1981). Door de jaren heen zijn meer dan 100 verschillende vertekeningen bedacht en onderzocht (zie 'List of cognitive biases', 2018, voor een overzicht). De aanwezigheid van cognitieve vertekeningen zal de spreiding in de beoordelaarseffecten (σ^2_R) vergroten, en daarmee de IBB verlagen (formule 2). Scholing en intervisie waarbij de beoordelingsprocedures goed uitgelegd worden en waarbij de beoordelaars bewust worden van de mogelijke cognitieve vertekeningen, kunnen, samen met duidelijke items en rubrics, bijdragen aan het verbeteren van de kwaliteit van de beoordeling en het vergroten van de IBB.

4. Wanneer is de interbeoordelaarsbetrouwbaarheid hoog genoeg?

Landis en Koch (1977) stelden richtlijnen op voor de interpretatie van Cohens kappa (κ): $\kappa \leq 0.20$ is *onvoldoende* (slight), $0.20 < \kappa \leq 0.40$ is *matig* (fair), is *voldoende* (moderate), $0.60 < \kappa \leq 0.80$ is *goed* (substantial) en $0.80 < \kappa \leq 1$ is *uitstekend* (excellent). Deze richtlijnen zijn zeer populair, en andere richtlijnen worden zelden gebruikt (Gwet, 2014). Deze richtlijnen hebben twee problemen. Ten

eerste gelden de richtlijnen niet voor andere IBB-coëfficiënten[4]. Immers Ten Hove et al. (2018) hebben laten zien dat bij dezelfde set beoordelingen verschillende coëfficiënten diverse waarden aannemen. Ten tweede, is het volgens ons onmogelijk om een enkele richtlijn te geven voor een coëfficiënt. De eis die men stelt aan de IBB dient afhankelijk te zijn van het doel van de beoordeling. Bij een beoordeling die bedoeld is voor individuele diagnostiek met belangrijke consequenties (bijv. of een asielzoeker in Nederland mag blijven) gelden andere maatstaven dan bij beoordelingen die alleen voor onderzoek op groepsniveau gebruikt worden (bijv. het coderen van gedrag bij studenten om factoren van sociale mimicry te onderzoeken; Salazar Kämpf, et al., 2018).

Om de ICC en Krippendorffs alfa te kunnen interpreteren, transformeerden Van der Ark et al. (2018) de door Landis en Koch (1977) voorgestelde richtlijnen voor Cohens kappa naar de ICC en Krippendorffs alfa. Eerst werden verschillende datasets gesimuleerd, waarbij elke dataset bestond uit één beoordeling door 100.000 beoordelaars. Vervolgens werd voor elke dataset Cohens kappa, ICC(2,1) en Krippendorffs alfa berekend, en met een tweedegraads polynoom regressiemodel de relaties berekend tussen Cohens kappa en de ICC en tussen Cohens kappa en Krippendorffs alfa. Ten slotte werden met de parameter schattingen uit dit polynome regressiemodel de richtlijnen van Landis en Koch getransformeerd (Tabel 3). Hierdoor zijn de richtlijnen van Landis en Koch ook beschikbaar voor andere coëfficiënten, maar wat opvalt is dat ook het aantal categorieën waarop beoordeeld wordt een grote rol speelt, en waarschijnlijk ook andere factoren waar in dit onderzoek geen rekening mee is gehouden.

Het tweede probleem, dat de betekenis van een IBB coëfficiënt afhankelijk is van de context waarin hij wordt gebruikt, is moeilijker oplosbaar. Eerst moet worden onderzocht wat het effect van IBB op het beoogde resultaat van het onderzoek is. In het geval dat een observatie en beoordeling gebruikt worden bij een voorspelling (bijv. in een regressieanalyse), zou met een simulatiestudie het effect van de IBB op de statistische power van de beslissing onderzocht moeten worden. In het geval dat een observatie en beoordeling moet leiden tot de kwantificering van de beoordeelde op een construct (denk aan docenten die middels de beoordeling van een open-vragen tentamen willen weten wat het niveau van Engelse taalvaardigheid is van hun studenten), zou met een simulatiestudie het effect van de IBB op de meetprecisie onderzocht moeten worden. Dergelijke methoden zijn momenteel nog nauwelijks beschikbaar.

Tabel 3
Transformatie van de richtlijnen voor Cohens kappa naar ICC en Krippendorffs alfa

Richtlijn	*C*	Cohens kappa	ICC[a]	Krippendorffs alfa
Onvoldoende	2	≤ 0.20	≤ 0.24	≤ 0.20
	3	≤ 0.20	≤ 0.40	≤ 0.35
	5	≤ 0.20	≤ 0.55	≤ 0.50
Matig	2	0.21 - 0.40	0.24 - 0.41	0.21 - 0.40
	3	0.21 - 0.40	0.41 - 0.60	0.36 - 0.55
	5	0.21 - 0.40	0.56 - 0.75	0.51 - 0.72
Voldoende	2	0.41 - 0.60	0.42 - 0.60	0.41 - 0.60
	3	0.41 - 0.60	0.61 - 0.75	0.56 - 0.73
	5	0.41 - 0.60	0.76 - 0.85	0.73 - 0.85
Goed	2	0.61 - 0.80	0.61 - 0.80	0.61 - 0.80
	3	0.61 - 0.80	0.76 - 0.88	0.74 - 0.86
	5	0.61 - 0.80	0.86 - 0.91	0.86 - 0.90
Uitstekend	2	0.81 - 1.00	0.80 - 1.00	0.81 - 1.00
	3	0.81 - 1.00	0.89 - 1.00	0.87 - 1.00
	5	0.81 - 1.00	0.92 - 1.00	0.91 - 1.00

Noot. *C* = aantal categorieën dat gebruikt wordt bij de beoordeling. a = Het betreft hier de ICC voor tweeweg random effecten, absolute overeenstemming en één beoordelaar.

5. Discussie

Ook 100 jaar na de oprichting van het Nutseminarium zijn observaties en beoordelingen nog zeer belangrijk in pedagogische en onderwijskundig onderzoek. De vraag 'Zijn we het eens?' is bij beoordelingen veel lastiger te beantwoorden. De laatste 60 jaar is veel onderzoek gedaan naar vertekeningen bij observeren en beoordelen, dat handvatten biedt om de IBB zo hoog mogelijk te krijgen. Echter, de verschillende IBB-coëfficiënten die doorgaans gebruikt worden hebben vaak verschillende interpretaties, leveren verschillende waarden op als ze op dezelfde set beoordelingen worden toegepast, en zijn alleen vergelijkbaar na vrij ingewikkelde omrekenprocedures.

Twee grotere problemen zijn nog helemaal niet opgelost: Het eerste probleem is dat algemene richtlijnen, zoals die opgesteld door Landis en Koch (1977), niet volstaan om te bepalen of de IBB hoog genoeg is. Effecten van de IBB op de power en meetprecisie zijn veel belangrijker en nuttiger om te weten dan een algemene kwalificatie. De uit de generaliseerbaarheidstheorie voortkomende ICC is verwant aan de testscorebetrouwbaarheid zoals gedefinieerd in de klassieke testtheorie (zie Cronbach, Rajaratnam, & Gleser, 1963). De effecten van testscorebetrouwbaarheid op power en meetprecisie zijn bekend (bijv. Lord & Novick, 1968). Nieuw onderzoek moet uitwijzen wat de effecten zijn van de verschillende ICC varianten op power en meetprecisie. Als die effecten bekend zijn, kan de vraag hoe hoog de IBB moet zijn, beoordeeld worden aan de hand van de mate waarin de IBB de power of meetprecisie beïnvloedt.

Het tweede probleem zijn beoordelingen waar afhankelijkheden in zitten. Voor dit type beoordelingen kan de IBB nog niet bepaald worden. Een eerste voorbeeld zijn beoordelingen met zogenaamde *interdependenties*. Een voorbeeld hiervan zijn twee kinderen die samen spelen en waarvan de gedragingen van de kinderen zowel apart als in interactie met elkaar beoordeeld moeten worden. Omdat in de pedagogiek beoordelingen bijna altijd binnen een context plaatsvinden, wordt dit node gemist. Een tweede voorbeeld zijn beoordelingen waarbij de beoordeelden genest zijn in verschillende groepen. Denk hierbij bijvoorbeeld aan het beoordelen van kinderen waarvan sommigen in dezelfde klas zitten, en op die manier afhankelijk van elkaar zijn. Het moet onderzocht worden of het negeren van deze afhankelijkheid de schatting van de IBB vertekent, en indien dit het geval is, moeten alternatieve schattingsmethoden ontwikkeld worden. Daarnaast zijn methoden nodig die de IBB kunnen schatten wanneer een onderzoeker geïnteresseerd is in verschillende subcomponenten van de geobserveerde, geneste, data (zie bijvoorbeeld Salazar Kämpf et al., 2017). Er is ons inziens nog veel werk te doen voordat we de vraag "in welke mate zijn we het eens" wetenschappelijke kunnen beantwoorden. Binnen het zojuist gestarte promotieonderzoek van de tweede auteur, zijn we nu, binnen het door Kohnstamm 100 jaar geleden opgerichte Nutsseminarie, bezig om deze vragen te beantwoorden.

Noten

[1] Type 1: IBB gebaseerd op overeenstemming waarbij niet voor kans wordt gecorrigeerd. Type 2: IBB gebaseerd op overeenstemming waarbij de kanscorrectie gebeurt op basis van de beoordelingscategorieën. Type 3: IBB gebaseerd op overeenstemming waarbij de kanscorrectie gebeurt op basis van de scoreverdeling. Type 4: IBB gebaseerd op overeenstemming waarbij de kanscorrectie gebeurt op basis van zowel de scoreverdeling als de beoordelingscategorieën.

[2] Er zijn verschillende typen ICC voor verschillende onderzoeksdesigns. Formule 2 betreft een ICC is voor de mate overeenstemming wanneer er in de praktijk slechts één beoordelaar is; voor een overzicht zie Koo en Li (2016). Deze formule kan het best gebruikt worden om het effect van maatregelen ter verhoging van de IBB op de hoogte van de IBB te illustreren.

[3] Hoewel bedoeld voor vragenlijsten, gelden de opgenomen richtlijnen ook voor items die bij beoordelingen gebruikt worden.

[4] Zonder verder op details in te gaan is het zelfs twijfelachtig of de richtlijnen geschikt zijn voor Cohens kappa zelf. De numerieke waarde van

kappa hangt onder andere af van het aantal beoordelaars en de verdeling van scores. Hierdoor lijkt een generalisering moeilijk.

Literatuur

Case, S. M., & Swanson, D. B. (1998). *Constructing written test questions for the basic and clinical sciences* (2nd ed.). Philadelphia, PA: National Board of Medical Examiners.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20. 37-46. doi:10.1177/001316446002000104

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90,* 218-244. doi: 10.1037/0033-2909.90.2.218

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*, 137-163. doi: 10.1111/j.2044-8317.1963.tb00206.x

Egert, F., Fukkink, R. G., & Eckhardt, A. G. (2018). Impact of in-service professional development programs for early childhood teachers on quality ratings and child outcomes: A meta-analysis. *Review of Educational Research, 88*, 401-433. doi: 10.3102/0034654317751918

Gamer, M., Lemon, J., Fellows I., & Singh, P. (2012). irr: Various coefficients of interrater reliability and agreement [Computer software]. Geraadpleegd via https://CRAN.R-project.org/package=irr

Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Gaithersburg, MD: Advanced Analytics LLC.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*, 23-34. Geraadpleegd via http://www.tqmp.org/RegularArticles/vol08-1/p023/p023.pdf

Hofstee, W. (1991). Richtlijnen voor het schrijven van vragenlijstitems. Ongepubliceerd rapport. Geraadpleegd via https://www.researchgate.net/publication/247409815_Richtlijnen_voor_het_schrijven_van_vragenlijstitems_Guidelines_for_writing_inventory_items

Idenburg, Ph. J. (1958). Het Nutsseminarium als centrum van opleiding en onderzoek. *Pedagogische Studien, 35*, 258-267.

Kahnemann, D. (2011). *Thinking fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kohnstamm, Ph. (1935). *Paedagogiek als wetenschap* (Mededeling No. 28 van het Nutsseminarium voor de Paedagogiek). Amsterdam: Universiteit van Amsterdam.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155-163. doi:10.1016/j.jcm.2016.02.012

Koopman, L., Zijlstra, B. J. H., De Rooij, M. de, & Van der Ark, L. A. (in press). Bias of two-level scalability coefficients and their standard errors. *Applied Psychological Measurement*.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research, 30,* 411-433.

Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability. Rapport Geraadpleegd op 29 januari 2019 via https://repository.upenn.edu/asc_papers/43/

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174. doi:10.2307/2529310

Leijten, P., Melendez-Torres, G. J., Gardner, F., Van Aar, J., Schulz, S., & Overbeek, G. (2018). Are relationship enhancement and behavior management "the golden couple" for disruptive child behavior? Two meta-analyses. *Child development*, advanced online publicatie. doi: 10.1111/cdev.13051

Levering, B. (2017). Ik zou de pianostemmer van de pedagogiek willen zijn. Interview met Andries van der Ark. *Pedagogiek in de praktijk, 95*, 6-13.List of Cognitive Biases (n. d). In *Wikipedia*. Geraadpleegd op 29 januari 2019 via https://en.wikipedia.org/wiki/List_of_cognitive_biases

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Popping, R. (1988). On agreement indices for nominal data. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research* (pp. 90-105). Londen, Engeland: Palgrave Macmillan. doi:10.1007/978-1-349-19051-5_6

Salazar Kämpf, M., Liebermann, H., Kerschreiter,

R., Krause, S., Nestler, S. & Schmukle, S. C. (2018). Disentangling the sources of mimicry: Social relations analyses of the link between mimicry and liking. *Psychological Science, 29,* 131-138. doi: 10.1177/0956797617727121.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*, 922-932. Geraadpleegd via https://web.stanford.edu

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428. doi:10.1037/0033-2909.86.2.420

Snijders, T. A. B. (2001). Two-level nonparametric scaling for dichotomous data, in A. Boomsma, M. A. J. van Duijn, & T.A.B. Snijders (Red.), *Essays on item response theory* (pp. 319-338). New York, NY: Springer. doi: 10.1007/978-1-4613-0169-1_17

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*, 72-101. doi: 10.2307/1412159

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2018). On the usefulness of interrater reliability coefficients. In M. Wiberg, S. A. Culpepper, R. Janssen, J. González, & D. Molenaar (Red.), *Quantitative Psychology*; *The 82nd Annual Meeting of the Psychometric Society, Zurich, Switzerland, 2017* (pp. 67-76). New York, NY: Springer. doi: 10.1007/978-3-319-77249-3_6

Thomassin, K., Raftery-Helmer, J., & Hersh, J. (2018). A Review of Behavioral Observation Coding Approaches for the Trier Social Stress Test for Children. *Frontiers in Psychology: Quantitative Psychology and Measurement, 9*, 2610. doi: 10.3389/fpsyg.2018.02610

Tversky, A., & Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131. doi: 10.1126/science.185.4157.1124

Van der Ark, L. A., Van Leeuwen, J. L., & Jorgensen, T. D. (2018). Interbeoordelaars-betrouwbaarheid LIJ; Onderzoek naar de interbeoordelaarsbetrouwbaarheidheid van het Landelijk Instrument Jeugdstrafrechtketen (Rapport 2829). Den Haag: WODC. Geraadpleegd via https://www.wodc.nl/binaries/2694_Volledige_Tekst_tcm28-357630.pdf

Van der Kamp, M.-T. (2012). Beoordelen van creatieve beeldende producten en processen van leerlingen in het voortgezet onderwijs: Een literatuuronderzoek naar criteria voor beeldende producten en processen in een hedendaagse context van kunst en kunsteducatie. Intern rapport Expertisecentrum Vakdidactiek Kunsttheorie. Geraadpleegd via http://expertisecentrum-kunsttheorie.nl/cms_data/litozbpr.pdf

Van der Put, C. E., Spanjaard, H. J. M., Van Domburgh, L., Doreleijers, T. A. H., Lodewijks, H. P. B., Ferwerda, H. B., Bolt, R. B., & Stams, G. J. J. M. (2011). Ontwikkeling van het Landelijke Instrumentarium Jeugdstrafrechtketen (LIJ). *Kind & Adolescent Praktijk, 10.* 76-83.

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association, 36*, 419-480. doi:10.1080/23808985.2013.11679142

Auteurs

L. Andries Van der Ark, Bijzonder hoogleraar en universitair hoofddocent, Pedagogische en onderwijswetenschappen, Universiteit van Amsterdam. **Debby Ten Hove**, Promovendus en docent, Pedagogische en onderwijswetenschappen, Universiteit van Amsterdam.

Correspondentie: L.A.vanderArk@uva.nl, D.tenHove@uva.nl

Abstract

Do we agree? Interrater reliability in education

Ratings and assessments are daily practices in education: Teachers decide whether a student's behaviour should be punished, teachers grade students' theses, and child-protection officers give complex assessments of juvenile delinquents. The interrater reliability (IRR) expresses the level of agreement among raters. We discuss three questions. The first question– How to determine the IRR? –cannot be answered unequivocally. Many coefficients are available to estimate the IRR, but often produce different results. The choice among coefficients depends on the research goal, research design, and personal preference. The answer to the second

question–How to increase the IRR?–entails increasing the expertise of the raters', and the quality of the items, rubrics and procedure. To answer the third question–When is the IRR high enough?–we present a method to transform benchmarks for one IRR coefficient to another, but methodological research is required to provide a better answer to this question.

Keywords: Cohen's Kappa, education, interrater reliability, intra class correlation coefficient, pedagogics