Leerlinggedrag volgsysteem

De betrouwbaarheid van een gedragsobservatie instrument en de samenhang tussen leergedrag en leerprestatie bij jongeren in de tweede klas van het voortgezet onderwijs.

T.J.M. van Dijk - 0854875

Leiden Universiteit - Masterproject Leerproblemen

1ste lezer: Mevr. Dr. M.J.A.J. Verhallen

2de lezer: Mevr. Dr. C.A. Espin

Juli 2012

Samenvatting

Hoewel bekend is dat het cognitieve vermogen van een leerling een grote invloed heeft op de leerprestaties van die leerling, is minder algemeen bekend dat ook het leergedrag van een leerling een grote invloed op de leerprestaties heeft. Hoewel er veel onderzoek is gedaan naar adequaat instrumentarium om de cognitie te onderzoeken, is dit niet het geval voor het leergedrag. In dit onderzoek wordt de betrouwbaarheid van een zelf ontworpen observatie instrument om het leergedrag in kaart te brengen onderzocht en met dit instrument wordt tevens de relatie tussen leergedrag en leerprestatie onderzocht. Hierbij is gebruik gemaakt van systematische directe observatie en momentary time sampling. De respondenten waren 25 meisjes en 13 jongens uit de tweede klas van het praktijkonderwijs en het onderwijs voor nieuwkomers. Zij waren afkomstig uit twee scholen. Met behulp van correlatietoetsen zijn de interbeoordelaarsbetrouwbaarheid, test-hertestbetrouwbaarheid en de samenhang tussen leergedrag en leerprestatie berekend. Het is gebleken dat er een goede interbeoordelaars- en test-hertestbetrouwbaarheid bestaan voor dit instrument als geheel. Het passieve niet-aan-taak gedrag kon echter niet betrouwbaar worden gemeten. Het is opvallend dat er geen samenhang is gevonden tussen leergedrag en leerprestatie. Dit is mogelijk te verklaren door de specifieke groep leerlingen die aan het onderzoek deelnam. Het is mogelijk dat deze groep leerlingen een afwijkend beeld laat zien van andere leerlingen uit de tweede klas van het voortgezet onderwijs. Beperkingen en aanbevelingen worden besproken.

Introductie

Er bestaat een relatie tussen gedrag en schools presteren, zoals blijkt uit verschillende onderzoeken (Arnold et al., 2005; Reid, Gonzalez, Nordness, Trout, & Epstein, 2004; Trout, Nordness, Pierce, & Epstein, 2003). Zo blijkt uit longitudinale studies dat leesproblemen bij jongens en meisjes op jonge leeftijd leiden tot gedragsproblemen op latere leeftijd (Maughan, Pickles, Hagell, Rutter, & Yule, 1996; Williams & McGee, 1994). De benedengemiddelde leerprestaties van leerlingen zouden hun gedrag op een negatieve wijze kunnen beïnvloeden. Het is mogelijk dat leerlingen door hun benedengemiddelde schoolresultaten een negatief zelfbeeld ontwikkelen, wat het gedrag beïnvloedt. Door het ervaren van meerdere faalmomenten tijdens het leren ontstaan niet alleen cognitieve achterstanden, maar ook problemen in motivatie en uiteindelijk in het gedrag (Morgan, Farkas, Tufis, & Sperling, 2008). Het negatief beïnvloede gedrag kan zijn uiting hebben in minder betrokkenheid van de leerling tot de les.

Morgan et al. (2008) hebben echter gevonden dat deze samenhang van bidirectionele aard is. Dit betekent dat de schoolprestatie een risicofactor is voor het gedrag, maar dat het gedrag ook een risicofactor is voor een lagere schoolprestatie. Doordat leerlingen minder gemotiveerd zijn, zullen zij minder opletten tijdens de instructie en minder aandacht besteden aan de leerstof, waardoor zij slechter zullen presteren op school (Roeser, van der Wolf, & Strobel, 2001). Op deze manier kan een negatieve causale spiraal ontstaan, waarbij een tekort aan oefening en (lees)vaardigheid kan leiden tot frustratie, wat demotiveert om in het vervolg een poging te doen om vaardigheden onder de knie te krijgen (Stanovich, 1986). Deze negatieve spiraal kan worden doorbroken door de leerprestatie direct te verbeteren of door deze indirect te verbeteren via het leergedrag. Het is mogelijk dat wanneer het leergedrag verbetert, dit een positieve invloed heeft op de leerprestaties van leerlingen. De betere leerprestaties kunnen op de lange termijn het toekomstperspectief voor de individuele leerling verbeteren, aangezien de leerling meer kansen krijgt om zich te ontwikkelen door het volgen van onderwijs op een adequaat niveau voor die leerling. Uiteindelijk kan dit ertoe leiden dat elke leerling gebruik kan maken van zijn of haar capaciteiten.

Hieruit blijkt dat het effectief kan zijn om zowel het leergedrag als de leerprestaties te verbeteren, aangezien zij elkaar versterken. Echter, in scholen wordt over het algemeen meer de nadruk gelegd op het meten en ontwikkelen van de cognities ter verbetering van de leerprestaties, dan op het vergroten van het leergedrag. Dit is het gevolg van het feit dat het fenomeen leerprestatie voornamelijk vanuit het normatief ontwikkelingsperspectief wordt benaderd vanuit scholen. Wanneer dit perspectief wordt aangenomen, kunnen de leerresultaten van een leerling vergeleken worden met die van leeftijdsgenoten om de leerprestatie vast te stellen (Sattler, 2008). De leerresultaten van leeftijdsgenoten vormen dan de norm. Dat scholen over het algemeen vanuit dit perspectief werken, blijkt uit het feit dat het cijfer voor een toets afhangt van de resultaten van de medeleerlingen en bijvoorbeeld ook de Wechsler Intelligence Scale for Children (WISC) werkt met normen (Kievit, Tak, & Bosch, 2009). Een perspectief wat echter minder vaak binnen het onderwijs wordt aangenomen, maar wat toch interessant is, is het cognitieve-gedragsmatige perspectief (Sattler, 2008). Dit perspectief neemt aan dat zowel cognitieve mogelijkheden als het leergedrag de leerprestatie kunnen beïnvloeden. Dit wordt ondersteund door een synthese van Hattie (2009), waarin meer dan 800 meta-analyses

zijn opgenomen. De 800 meta-analyses hadden alle betrekking op voorspellers van de leerprestatie van leerlingen. Het leergedrag in de klas (waar concentratie en het richten van de aandacht op de taak onder werd verstaan) nam een zesde plaats in op de lijst van meest invloedrijke voorspellers van leerprestatie. Er werd een effectgrootte van d = .80 gevonden, wat een groot effect is (Hattie, 2009). Vanuit dit perspectief wordt het leergedrag gezien als een factor die de mate waarin de cognitie van een kind tot uiting kan komen in de weg kan zitten. Wanneer een kind hierdoor niet volledig zijn of haar cognitieve mogelijkheden kan exploreren, dan zal dit ook een negatief effect hebben op de leerprestatie. Om gewenste leeruitkomsten te verkrijgen is het vanuit dit perspectief daardoor van belang dat wordt ingezet op het verbeteren van zowel cognities als het leergedrag, om indirect de leerprestatie te verbeteren. Dit betekent dat het niet voldoende is om het kind te ondersteunen in de ontwikkeling van zijn of haar cognitieve mogelijkheden, maar dat het kind ook hulp moet krijgen in het ontwikkelen van zijn of haar leergedrag. Het meten en het ontwikkelen van meetinstrumenten voor het leergedrag zou daarom een belangrijke en betekenisvolle focus van onderzoek en interventie moeten zijn (Leung, Lo, & Leung, 2012). Toch blijkt de ontwikkeling van meetinstrumenten voor het meten van leergedrag een achterstand te hebben opgelopen ten opzichte van die van de leerprestaties. Hoewel er een grote onderzoeksbasis is met betrekking tot meetinstrumenten voor de voortgang van de leerprestatie is er veel minder kennis over instrumenten voor het meten en het monitoren van leergedrag (Chafouleas et al., 2010a). De behoefte hieraan is echter groot (Chafouleas et al., 2010a; Chafouleas, Volpe, Gresham, & Cook, 2010b; Hintze & Matthews, 2004; Riley-Tillman, Methe, & Weegar, 2007).

Leergedrag

Voor het kunnen ontwikkelen van een instrument is het van belang dat men goed weet wat er met het instrument gemeten moet worden. Het begrip leergedrag moet gedefinieerd en geoperationaliseerd worden, voordat men over kan gaan tot het ontwikkelen van een instrument om het leergedrag te meten. Morgan et al. (2008) onderscheiden in hun studie naar de samenhang tussen gedrag en leerprestatie verschillende vormen van gedrag, namelijk *benadering tot leren*, *problemen in zelfcontrole, problemen in sociale vaardigheden, internaliserend probleemgedrag* en *externaliserend probleemgedrag*. De categorie *benadering tot leren* uit dit onderzoek

staat voor de mate waarin een kind voordeel haalt uit zijn of haar klassenomgeving. Dit betekent dat wordt geobserveerd of een leerling aandacht op de taak richt en deze kan vasthouden, graag wil leren, onafhankelijk kan leren, zijn of haar aandacht kan verleggen tussen taken en het leren kan organiseren en/of structureren (Morgan et al., 2008). De gedragsvorm *problemen in zelfcontrole* geeft inzicht in hoeverre een kind zijn of haar gedrag kan reguleren. De gedragscategorie *problemen in sociale vaardigheden* is een weergave van de mogelijkheden van een kind om vriendschappen te initiëren en te behouden. *Externaliserend gedrag* beslaat gedrag dat naar buiten is gericht (zoals ruziemaken of impulsief gedrag), terwijl *internaliserend gedrag* naar binnen is gericht (zoals angst, eenzaamheid en verdriet).

Uit het onderzoek van Morgan et al. (2008) bleek dat leesproblemen de grootste invloed hebben op de categorie *benadering tot leren* en dat zij een minder grote invloed hebben op de andere vier categorieën. Verder blijkt uit dit onderzoek dat de categorie *benadering tot leren* de enige categorie van gedragsproblemen is die invloed uitoefent op leesproblemen. Van de vijf verschillende categorieën van gedrag die in het onderzoek zijn meegenomen heeft *benadering tot leren* (het leergedrag) daardoor de grootste samenhang met leerprestaties (Morgan et al., 2008). Het is daardoor de beste indicator van leergedrag. De andere vier vormen zullen zonder twijfel ook in de klassensituatie hun weerslag hebben, maar de definities uit het onderzoek van Morgan et al. (2008) weerspiegelden een andere focus bij deze begrippen. Opvallend aan het onderzoek van Morgan et al. (2008) is dat alleen de samenhang met de leesprestatie is onderzocht, terwijl de samenhang op andere leergebieden (zoals wiskunde) mogelijk anders kan zijn. Onderzoek van Gruber, DuPaul, Jitendra, Volpe en Lorah (2004) toont namelijk aan dat leerlingen meer actieve betrokkenheid tonen bij wiskunde dan bij taal. Echter, bij taal werd meer passieve betrokkenheid bij het leren getoond. De onderzoekers hebben hierbij echter geen rekening gehouden met de betrouwbaarheid waarmee het gedrag kan worden gescoord bij deze afzonderlijke vakken.

De term *benadering tot leren* volstaat echter niet. Deze term legt namelijk voornamelijk de nadruk op het passieve karakter van het leergedrag. Een term die vergelijkbaar is met *benadering tot leren*, maar juist meer gericht is op de actieve vorm van leergedrag, is het begrip *engagement*, oftewel betrokkenheid. Betrokkenheid is de uiting van het leergedrag van een leerling (Fredricks, Blumenfeld, & Paris, 2004; Johnson, McGue, & Iacono, 2005; Miles & Stipek, 2006;

Roeser et al., 2001). Fredricks et al. (2004) definiëren betrokkenheid als het jezelf involveren en bezighouden met iets en erdoor aangetrokken worden. Het vergroten van de betrokkenheid tot het leren zorgt voor minder verveling bij de leerling, minder dropout en verbetering van de leerprestaties (National Research Council & Institute of Medicine, 2004). Hierbij kan onderscheid worden gemaakt in gedragsmatige, emotionele en cognitieve betrokkenheid. Gedragsmatige betrokkenheid is belangrijk voor de preventie van drop-out en het behalen van positieve schoolresultaten. Deze vorm van betrokkenheid heeft betrekking op participatie. Het betreft de betrokkenheid bij schoolse, sociale en buitenschoolse activiteiten. Emotionele betrokkenheid heeft betrekking op de reacties van leerlingen op docenten, medeleerlingen en op de school als geheel. De laatste vorm van betrokkenheid, cognitieve betrokkenheid, gaat over de wil en het idee om te investeren om complexe ideeën en vaardigheden onder de knie te krijgen. Deze termen hebben in grote mate overlap, aangezien zij alle drie betrekking hebben op de interesses, motivaties en inzet van een leerling (Fredricks et al. 2004). Op basis van de voorgaande definities kan gedragsmatige betrokkenheid worden gezien als het leergedrag wat een leerling in de klas laat zien. De twee termen benadering tot leren en gedragsmatige betrokkenheid geven samen een complete definitie van het leergedrag van leerlingen, waarbij de benadering tot leren meer de nadruk legt op de aandacht die wordt gericht op de taak en de gedragsmatige betrokkenheid de meer actieve participatie benadrukt. Echter, ook wanneer men deze begrippen gebruikt om leergedrag te definiëren blijft het een abstract concept. Iedereen kan namelijk zijn eigen invulling geven aan de termen ‘betrokkenheid’ en ‘participatie’. De subjectieve invulling door afzonderlijke onderzoekers leidt tot onbetrouwbare metingen en het is om deze reden van groot belang dat het begrip leergedrag niet alleen wordt gedefinieerd, maar dat de gedragingen die kenmerkend zijn voor verschillende vormen van leergedrag ook worden geoperationaliseerd (Salvia & Ysseldyke, 2004). Dit heeft tot gevolg dat een observator het gedrag beter kan herkennen en dat de meting objectiever kan worden uitgevoerd (Hintze, 2005). Om deze reden moeten een aantal indicatoren van het leergedrag en de betrokkenheid tot het leren worden onderscheiden.

Indicatoren leergedrag

Een bestaand instrument voor het meten van leergedrag is de *Behavioral Observation of Students in Schools (BOSS*; Shapiro, 2004). In een studie van Hintze en Matthews

uit 2004 werd een schaal met drie indicatoren ontwikkeld die afgeleid waren van de BOSS. Deze indicatoren waren *Actief betrokken gedrag*, *Passief betrokken gedrag* en *Niet aan taak*. Wanneer een leerling actief gericht is op een taak valt dit gedrag binnen de categorie *Actief betrokken gedrag*. Binnen deze categorie vallen gedragingen als schrijven, het opsteken van de hand en hardop lezen. Als de betrokkenheid van de leerling van een passief karakter is, zoals het luisteren naar instructie of het kijken naar lesmateriaal, is sprake van *Passief betrokken gedrag*. Al het gedrag wat niet binnen deze twee eerste categorieën viel werd behandeld als *Niet aan taak* (Hintze & Matthews, 2004). Met behulp van deze operationalisatie werden leerlingen twee keer per dag, gedurende tien dagen, geobserveerd. Het bleek echter dat met dit aantal observaties geen voldoende betrouwbaarheid en validiteit konden worden bereikt. Een kwart van de variantie in de data was te verklaren door fouten in de observatie waaruit blijkt dat, zoals Hintze en Matthews (2004) zelf opmerken, "makkelijker niet altijd beter is als het gaat om systematische directe observatie" (p. 268). Zij concluderen dat leergedrag mogelijk een multidimensionaal karakter heeft, en dat meer dan drie categorieën nodig zijn om het goed in kaart te brengen. Uit dit onderzoek bleek dat pas bij 8 tot 40 metingen betrouwbare data kon worden verkregen, afhankelijk van de variabiliteit in het gedragspatroon van de specifieke leerling (Hintze & Matthews, 2004). Dit is niet wenselijk, aangezien leerkrachten op basis van het leergedrag snel en accuraat moeten kunnen beslissen of een interventie nodig is bij een leerling.

Vanuit het denkkader van Hintze en Matthews (2004) dat leergedrag een multidimensionaal karakter heeft dat niet te vangen is binnen drie categorieën lijkt het verstandig om meer categorieën te gebruiken bij het in kaart brengen van het leergedrag. Het is mogelijk dat de betrouwbaarheid van de operationalisatie van leergedrag vergroot wordt als ook binnen de categorie *niet aan taak* een verdere opdeling wordt gemaakt, zoals dat ook het geval was voor de categorie *aan taak*. Deze denkwijze past binnen het kader dat Suldo en Shaffer (2008) schetsten in hun onderzoek waarin een zogenaamd Dual-Factor Model of Mental Health werd gehanteerd. In dit model werd op zowel de positieve als de negatieve kanten van de mentale gezondheid van leerlingen gefocust en het bleek dat door deze werkwijze een beter beeld kon worden verkregen van de mentale gezondheid van leerlingen. Voor het concept leergedrag betekent dit dat ook binnen de categorie *niet aan taak* een onderscheid in een passieve en een actieve vorm relevant kan zijn. Dit heeft tot gevolg

dat er vier categorieën ontstaan waarbij niet alleen het aan-taak gedrag een passieve en actieve vorm heeft, maar ook het niet-aan-taak gedrag een passieve en actieve vorm heeft. Het passieve niet-aan-taak gedrag is in deze indeling het gedrag dat niet op de taak is gericht, zoals uit het raam staren. Bij actief niet-aan-taak gedrag, oftewel storend gedrag, is het kind niet alleen zelf van zijn taak, maar houdt hij of zij ook anderen van hun taak. Dit kan worden vorm gegeven in bijvoorbeeld kletsen of anderen fysiek aanraken.

Het is opvallend dat Hintze & Matthews (2004) storend gedrag niet als categorie hebben opgenomen, terwijl het vanuit de literatuur een belangrijke factor blijkt te zijn in de ontwikkeling van de leerprestatie (Johnson, McGue, & Iacano, 2005; McCall, Evahn, & Kratzer, 1992; Nelson, Benner, Lane, & Smith, 2004). Het is mogelijk dat zij niet voor deze categorie hebben gekozen, aangezien het verband tussen storend gedrag en leerprestatie op de basisschoolleeftijd nog indirect is en concentratieproblemen een belangrijke derde variabele zijn (Hinshaw, 1992). Echter wanneer leerlingen naar de middelbare school gaan en in de adolescentie terecht komen, is de samenhang tussen het storend gedrag en achterstanden in de leerprestatie meer direct en staat het antisociale gedrag in de klas zelf op de voorgrond (Johnson, McGue, & Iacano, 2005). Het is daardoor vooral bij leerlingen in het voortgezet onderwijs betekenisvol om het storende gedrag van leerlingen te onderzoeken. De hier genoemde onderzoeksresultaten zijn echter voornamelijk afkomstig van onderzoek bij leerlingen in het basisonderwijs. De samenhang tussen leergedrag en leerprestatie bij jongeren in het voortgezet onderwijs is in een veel mindere mate onderzocht. Het is daarom niet met zekerheid te zeggen dat de resultaten ook op die groep van toepassing zijn. Het is mogelijk dat een meer uitgebreide operationalisatie van het concept leergedrag, waarbij ook een opdeling in de categorie niet-aan-taak gedrag wordt gemaakt eenzelfde resultaat tot stand kan brengen als in het onderzoek van Suldo en Shaffer (2008). De data worden op deze manier niet op een grote hoop gegooid, maar worden onderscheiden in verschillende categorieën.

Het is echter ook mogelijk dat het leergedrag betrouwbaarder kan worden gemeten, wanneer juist gebruik wordt gemaakt van dichotome schalen. Dit betekent dat slechts twee categorieën worden gehanteerd en dit zijn er dus juist minder in plaats van meer dan in het onderzoek van Hintze en Matthews (2004). Wanneer gebruik wordt gemaakt van een schaal met twee categorieën, zullen er namelijk per categorie meer scores zijn dan wanneer gebruik wordt gemaakt van drie of vier

categorieën, aangezien de scores over minder categorieën verdeeld worden. Dit heeft tot gevolg dat robuustere categorieën ontstaan, aangezien er meer scores in iedere categorie aanwezig zijn. Robuustere categorieën leiden tot meer betrouwbaarheid. Al in 1985 werd namelijk door Thorndike vastgesteld dat wanneer de steekproef groter is, men kan verwachten dat de betrouwbaarheid hoger zal uitvallen. Het aantal scores dat binnen een bepaalde categorie valt kan worden gezien als een steekproef van het gedrag dat binnen die categorie valt. Als er dus meer scores binnen een categorie vallen, dan leidt dat tot een grotere steekproef van dat gedrag en deze grotere steekproef is betrouwbaarder te meten dan kleinere steekproeven die men zou aantreffen wanneer meer categorieën worden gebruikt (Thorndike, 1985).

In onderzoek van Chafouleas et al. (2010a) is gebruik gemaakt van een schaal met twee categorieën, namelijk schoolse betrokkenheid en storend gedrag. Hoewel in dit onderzoek de categorie schoolse betrokkenheid zowel uit actief als passief aan-taak gedrag bestaat, werd onder storend gedrag alleen het actieve niet-aan-taak gedrag verstaan en passief niet-aan-taak gedrag kreeg daardoor geen rol in de observatieschaal van het onderzoek van Chafouleas et al. (2010a). Toch is in dit onderzoek gebleken dat met een dichotome schaal betrouwbare data kunnen worden verkregen.

Riley-Tillman, Christ, Chafouleas, Boice-Mallach en Briesch (2011) hebben in hun onderzoek naar het belang van de duur van observatie tevens gebruik gemaakt van een dichotome schaal. Dit was precies dezelfde schaal als was gebruikt in het onderzoek van Chafouleas et al. (2010a). Uit deze studie bleek dat hoe langer de observatie duurt, hoe meer storend gedrag wordt overschat door observatoren. In deze studie werd onderscheid gemaakt in observaties van 5, 10 en 20 minuten. De duur van de observatie bleek echter geen invloed te hebben op de accuratesse van de schatting van actief leergedrag.

De vier categorieën actief leergedrag, passief leergedrag, off-task gedrag en storend gedrag kunnen op twee manieren in dichotome schalen worden ingedeeld. De eerste manier is actief tegenover passief gedrag. Hierbij worden gedragingen die bij een indeling in vier categorieën binnen actief leergedrag of storend gedrag zouden vallen, samengevoegd tot één categorie, namelijk actief gedrag. In het geval van passief gedrag betekent dit dat de categorieën passief aan-taak gedrag en passief niet-aan-taak gedrag worden samengevoegd. Een tweede manier waarop de vier categorieën kunnen worden samengevoegd tot twee categorieën is de dichotome

schaal positief gedrag tegenover negatief gedrag. In het geval van positief gedrag worden de categorieën actief leergedrag en passief leergedrag samengevoegd, terwijl de combinatie van off-task gedrag en storend gedrag de categorie negatief gedrag oplevert. Deze indeling lijkt in grote mate op de indeling die door Chafouleas et al. (2010a) is gebruikt, behalve dat Chafouleas et al. (2010a) binnen de negatieve schaal geen ruimte lieten voor passief niet-aan-taak gedrag. Uit onderzoek van Butler (1990) is gebleken dat 80% van het leergedrag positief gedrag is, tegenover slechts 20% negatief gedrag.

Het is van belang dat de categorieën die worden gebruikt zo duidelijk mogelijk worden gedefinieerd en geoperationaliseerd. Wanneer categorieën niet goed zijn gedefinieerd en geoperationaliseerd, kan dit een belangrijke bron van variantie zijn die de betrouwbaarheid van het scoren kan verkleinen (Brown-Chidsey, 2005; Kobak et al., 2009).

Culturele- en sekseverschillen

Onderzoek naar de samenhang tussen leergedrag en –prestatie is voornamelijk gedaan bij autochtone kinderen, waarbij de factoren cultuur of etniciteit in zijn geheel niet zijn meegenomen in de onderzoeken. Uit onderzoek naar de leerprestatie is echter gebleken dat allochtone jongeren over het algemeen minder goede leerprestaties laten zien, wat het gevolg kan zijn van beperkingen in taalkennis en -begrip (Xiong, Eliason, Detzner, & Cleveland, 2005), het gevoel van culturele afstand door de familieachtergrond en de sociaal-economische status (Kiang, Supple, Stein, & Gonzalez, 2012) of door discriminatie en stereotypering (Lee & Stacey, 2001). In meerdere studies is aangetoond dat meisjes over het algemeen beter presteren in wiskunde en taal dan jongens (Landgren, Kjellman, & Gillberg, 2003; McDermott, Goldberg, Watkins, Stanley, & Glutting, 2006) en dat dit ook bij immigrantenjongeren het geval is (Brandon, 1991; Rong & Brown, 2001; Suárez-Orozco et al., 2010).

Uit onderzoek naar het leergedrag van immigrantenjongeren blijkt ook dat er juist bij deze groep sprake is van minder betrokkenheid tot het leren, wanneer zij het gevoel hebben niet te worden gewaardeerd in de klas (Marchant, Paulson, & Rothlisberg., 2001). Ook is bekend dat jongens gemiddeld meer storend gedrag en concentratieproblemen laten zien dan meisjes (Butts et al., 1995; Johnson, McGue, & Iacano, 2005; Rhee, Waldman, Hay, & Levy, 2001). Hieruit blijkt dat zowel de

leerprestatie als het leergedrag bij allochtone jongeren negatief afwijken van de norm van autochtone leerlingen en dat de jongens op beide gebieden de grootste afwijking vertonen. Dit maakt het interessant om ook voor de allochtone groep leerlingen te onderzoeken of de samenhang tussen leergedrag en leerprestatie aanwezig is en daarnaast of de samenhang verschilt tussen jongens en meisjes.

Hoewel informatie over de leerprestatie over het algemeen wordt samengevat in een cijfer, is dit voor leergedrag zelden het geval. Over het algemeen wordt de informatie over het leergedrag verkregen door middel van observatie. Aangezien docenten tijdens hun lessen ook instructie moeten geven en aandacht aan andere leerlingen moeten geven hebben zij niet veel tijd over tijdens hun lessen om observaties uit te voeren. Bij het meten van leergedrag is het daardoor van belang dat er niet alleen betrouwbare en valide data worden verzameld, maar dat deze ook snel en effectief te verkrijgen zijn (Espin et al., 2000; Evans & Owens, 2010; Riley-Tillman, Chafouleas, & Briesch, 2007). Er zijn meerdere methoden ontwikkeld om aan deze eisen te voldoen.

Leergedrag in kaart

Een methode die vaak wordt gebruikt om (leer)gedrag te beoordelen is de beoordelingsschaal. Hierbij wordt de mate waarin bepaald gedrag aanwezig is gescoord door de leerkracht door dit op een schaal aan te geven. Deze beoordeling is echter gebaseerd op de eerdere ervaringen die een leerkracht met de leerling heeft gehad en worden niet gedaan op het moment dat het gedrag zich voordoet. Dit maakt dat dit instrument afhankelijk is van de herinnering van een leerkracht, wat ten koste gaat van de objectiviteit van het instrument (Shapiro & Clemens, 2005). Een ander nadeel is dat de beoordelingsschalen vaak veel items kennen en daardoor moeilijk op herhaalbare basis te gebruiken zijn (Riley-Tillman et al., 2007).

Een tweede veel gebruikt methode voor het formatief meten van leergedrag is systematische directe observatie. Het is een objectieve en accurate methode met een goede sensitiviteit en specificiteit (Riley-Tillman, Chafouleas, Briesch, & Eckert, 2008; Wilson & Reschly, 1996). Met de sensitiviteit van systematische directe observatie wordt bedoeld dat het gedrag van de leerlingen dat binnen een bepaalde categorie van leergedrag valt ook in deze categorie wordt gescoord, terwijl de specificiteit betrekking heeft op de mate waarin het gedrag van leerlingen dat niet binnen een bepaalde categorie valt ook niet in die categorie wordt gescoord.

Systematische directe observatie is een vorm van observatie waarbij specifieke gedragingen worden geobserveerd. Deze gedragingen zijn van tevoren geoperationaliseerd en de observatie vindt plaats met gestandaardiseerde procedures, waardoor observaties op een objectieve wijze kunnen worden uitgevoerd. Verder worden de tijd en plaats van observatie bewust en met aandacht geselecteerd en gespecificeerd. Als laatste is de manier van scoren en samenvatten van de gedragingen voor alle observatoren precies hetzelfde (Salvia & Ysseldyke, 2004). De data kunnen op deze manier meerdere functies krijgen, zoals het identificeren van leerlingen voor interventie, het voorzien in een baseline en het maken van doelen (Shapiro & Clemens, 2005). In tegenstelling tot de beoordelingsschaal vindt de scoring bij systematische directe observatie plaats op het moment dat het gedrag voorkomt, wat de meest objectieve scoring als resultaat heeft (Christ, Riley-Tillman, & Chafouleas, 2009; Cone, 1978). Daarbij komt dat systematische directe observatie kan worden gebruikt in verschillende klassensituaties en voor verschillende doeleinden (Hintze, Volpe, & Shapiro, 2002).

Systematische directe observatie is bovendien een systematische en kwantitatieve methode. Dit betekent dat gestandaardiseerde procedures worden gevolgd. Deze methodes zijn beter repliceerbaar dan kwalitatieve methodes en het is gemakkelijker om te beoordelen of er daadwerkelijk veranderingen in gedrag optreden op de langere termijn (Shapiro & Clemens, 2005). Ook kunnen op basis van de data die voortkomen uit deze methode specifieke doelen worden opgesteld met betrekking tot het leergedrag. Wanneer een kwalitatieve methode wordt gebruikt, dan wordt een narratieve beschrijving van het kind opgesteld. Deze methodes zijn echter niet goed repliceerbaar, terwijl dit juist wel van groot belang is op de langere termijn wanneer men het gedrag wil volgen. Ook geven deze methodes meer mogelijkheid tot subjectieve interpretaties van leergedrag door een docent of onderzoeker. Het is om deze redenen moeilijker om doelen te stellen op de langere termijn en om te beoordelen of verandering in gedrag is voorgekomen (Shapiro & Clemens, 2005).

Zwaktes van systematische directe observatie die meerdere malen in de literatuur wordt genoemd zijn de grote kosten in tijd en geld die deze meetmethode met zich meebrengt (Chafouleas et al., 2010a; Hintze & Matthews, 2004; Pelham, Fabiano & Massetti, 2005; Riley-Tillman et al., 2007; Riley-Tillman et al., 2008). Een beoordelingsschaal is kosteneffectiever, maar is voornamelijk van een evaluatief karakter (Chafouleas et al., 2010a). De beoordelingsschaal is daardoor geen effectief

instrument wanneer men het gedrag niet alleen wil beoordelen, maar ook wil volgen (Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007).

Een laatste methode die wordt gebruikt in het onderzoek naar leergedrag is het gebruik van *Daily Behavior Report Cards* (DBRC). DBRC is een methode waarbij na een observatie het gedrag op een schaal wordt weergegeven. Het is hiermee een mengvorm van een beoordelingsschaal (aangezien de mate van gedrag moet worden aangegeven op een schaal) en systematische directe observatie (aangezien dit direct gebeurt na de observatie). Uit onderzoek van Riley-Tillman et al. (2008) bleek dat hoewel DBRC en systematische directe observatie beiden net zo acceptabel werden gevonden en vergelijkbare resultaten gaven, observatie de voorkeur verdient boven DBRC aangezien observatie een instrument is dat al veel meer wordt gebruikt en dat personeel hier in veel gevallen ook al beter in getraind is. Het lijkt daarom verstandig om observatie te gebruiken in plaats van DBRC, zodat de leerkrachten zo min mogelijk met nieuw en onbekend materiaal te maken krijgen als zij met een nieuw instrument te maken krijgen. DBRC heeft de laatste jaren aan populariteit gewonnen onder een andere naam, namelijk Daily Behavior Rating (DBR). Hoewel de naam verschilt is de vorm van dit instrument identiek aan DBRC (Chafouleas et al., 2007). Onderzoek met dit instrument laat wisselende resultaten zien met betrekking tot betrouwbaarheid en validiteit. Zo laten Chafouleas et al. (2010a) zien dat de DBR pas valide resultaten geeft bij 60 metingen. Daarnaast blijkt de betrouwbaarheid in grote mate af te hangen van de observator en wordt aangeraden om enkel data van de DBR te gebruiken die van dezelfde observator afkomstig zijn (Chafouleas et al., 2010a).

Momentary Time Sampling

Systematische directe observatie kan op meerdere manieren worden vorm gegeven (Shapiro & Clemens, 2005). Zo kan de vorm van het gedrag worden genoteerd (topografie), het aantal keer of de snelheid dat het gedrag voorkomt, de duur ervan, de intensiteit, of de snelheid dat het voorkomt na een bepaalde stimulus (Martin & Pear, 2011). Strategieën die hierbij kunnen worden gebruikt kunnen ook variëren. Echter, niet elke strategie is te gebruiken om elk gedrag te observeren. Verschillende strategieën om gedrag te observeren zijn continu observeren, interval observatie of time-sampling observatie (Martin & Pear, 2011). Continu observeren houdt in dat elke keer dat een respondent bepaald gedrag laat zien tijdens een vooraf vastgestelde tijdsperiode dit wordt genoteerd. Bij interval observatie geeft de observator na elk kort

interval aan of het gedrag wel of niet is voorgekomen. De laatste strategie die kan worden gehanteerd is time sampling. Bij deze methode wordt een observatiesessie opgedeeld in gelijke intervallen en wordt het gedrag in deze intervallen gescoord.

Time sampling methodes hebben het voordeel dat niet elke gedraging van de leerling hoeft te worden opgemerkt en te worden genoteerd, wat bij veel andere observatiemethodes wel het geval is (Shapiro & Clemens, 2005). Dit is van groot belang wanneer het uiteindelijke doel is om ook leerkrachten zelf dit instrument te laten gebruiken tijdens hun lessen. Het is namelijk belangrijk dat de leerkracht zijn of haar les zo vloeiend mogelijk kan voortzetten tijdens observatiemomenten. Een specifieke vorm van time sampling die de voorkeur verdient bij het onderzoeken van leergedrag is Momentary Time Sampling (MTS). MTS is een manier van meten waarbij het gedrag van een leerling wordt gescoord op het precieze moment dat het interval begint (Shapiro & Clemens, 2005). MTS verdient de voorkeur boven een aantal andere observatiemethoden (Watson & Steege, 2003). Er kan bijvoorbeeld geen gebruik worden gemaakt van duration recording aangezien het observeren van meerdere vormen van gedrag bij deze methode niet mogelijk is, terwijl leergedrag wel in categorieën wordt ingedeeld. MTS verdient bovendien de voorkeur boven andere vormen van time sampling, aangezien het de kleinste onder- en overschatting van het gedrag oplevert ten opzichte van deze andere vormen (Lentz, 1988). Hoewel MTS als strategie de voorkeur krijgt binnen de systematische directe observatie, moet dit wel op een consistente wijze worden toegepast.

Het monitoren van leergedrag

Hoewel het van belang is dat het leergedrag betrouwbaar kan worden gemeten op bepaalde momenten in de tijd, is het tevens belangrijk dat het leergedrag gevolgd kan worden zodat hier interventies op gebaseerd kunnen worden. Het probleemoplossingsmodel kan worden gebruikt om een probleem te identificeren en een interventie te selecteren. De observaties leiden tot een indicatie van het leergedrag voor iedere leerling. Het probleemoplossingsmodel is een model dat wordt gebruikt om een discrepantie tussen huidig en gewenst of verwacht functioneren te verkleinen of in zijn geheel te doen laten verdwijnen (Brown-Chidsey, 2005). Hoewel in het onderwijs probleemoplossing vaak is gericht op de educatieve ontwikkeling kan het ook worden gebruikt om problemen in het gedrag aan te pakken (Deno, 2005). Deno (2005) beschrijft een ontwikkeling in de manier waarop het probleemoplossingsmodel

wordt gebruikt. In klassiekere modellen werd een probleem of stoornis voornamelijk geïsoleerd benaderd, waarbij men zich richtte op het kind zelf. Tegenwoordig worden de problemen van kinderen meer gezien in een context van invloeden. Een voorbeeld van deze contextgerichte probleemoplossing is Response to Intervention (RtI).

RtI is een proces waarbij leerlingen die een risico lopen voor zwak schools presteren systematisch kunnen worden geïdentificeerd en geholpen met hun problemen (Deno et al., 2009). Bij RtI wordt de omgeving aangepast aan de mogelijkheden van het kind. RtI is een proces dat bestaat uit verschillende lagen. Alle leerlingen bevinden zich in eerste instantie in de eerste laag (Tier I). Zij krijgen allen instructie die effectief genoeg is voor ongeveer 80% van de leerlingen (Tilly, 2008). Wanneer een leerling zich niet genoeg ontwikkelt op basis van de algemene instructie krijgt hij of zij extra instructie of een interventie, waardoor deze leerling in de tweede laag van RtI belandt (Tier II). De leerlingen die ook met deze extra ondersteuning niet genoeg kunnen ontwikkelen komen in een groep terecht waarin nog intensievere hulp wordt gegeven. Dit is de derde laag van RtI (Tier III) en dit kan betekenen dat een leerling naar het speciaal onderwijs wordt verwezen (Fuchs, Mock, Morgan, & Young, 2003). Hoewel deze vorm van identificeren voornamelijk wordt gebruikt voor de leerprestatie van leerlingen, geven Fuchs et al. (2003) aan dat het belangrijkste ingrediënt van RtI is dat er een technisch systeem wordt geboden voor het identificeren van problemen en het bijhouden van groei met betrekking tot deze problemen. Dit essentiële onderdeel van RtI is ook toepasbaar op het leergedrag van leerlingen, in plaats van de leerprestatie (Deno et al., 2009).

RtI is een vorm van meten waarbij gebruik wordt gemaakt van Curriculum-based Measurement (CBM). CBM is ontwikkeld om op een simpele en snelle manier de groei in vaardigheden van leerlingen te volgen. De resultaten hiervan kunnen gebruikt worden om de instructie aan te passen. De gedachte hierachter is dat wanneer de instructie wordt aangepast op de leerling, hij of zij beter zou kunnen presteren (Stecker, Fuchs, & Fuchs, 2005). Hierbij wordt het belang van de omgeving dat ook binnen RtI een grote rol speelt benadrukt. Het belangrijkste doel van CBM is dat met behulp van formatieve informatie de leerkracht zijn instructiemethoden kan evalueren (Deno, 2005).

Het huidige onderzoek

In voorliggend onderzoek wordt een instrument voor het meten en volgen van leergedrag ontworpen en getest op zijn betrouwbaarheid. Het kan mogelijk een invulling geven aan de behoefte aan een betrouwbaar instrument voor het meten en volgen van leergedrag (Chafouleas et al., 2010a; 2010b). Vanuit de literatuur lijkt het hierbij van belang om gebruik te maken van systematische directe observatie en Momentary Time Sampling. Het instrument dat voor dit onderzoek is ontworpen maakt hier gebruik van. Er wordt hierbij gebruik gemaakt van observaties van acht minuten. Aangezien dit tussen de grenzen van vijf en tien minuten van het onderzoek van Riley-Tillman et al. (2011) valt, zal de overschatting van storend gedrag relatief klein zijn. Er is bewust niet voor een observatieduur van vijf minuten gekozen aangezien er in het onderzoek van Riley-Tillman et al. (2011) in deze korte duur te weinig actief leergedrag kon worden geobserveerd. Toch is de observatieduur in dit onderzoek onder de tien minuten gehouden om tevens rekening te houden met de resultaten van het onderzoek van Riley-Tillman et al. (2011) waarin een observatieduur van ongeveer vijf minuten werd aangeraden aangezien dat de minste overschatting van storend gedrag geeft. Er is om deze redenen gekozen voor een observatieduur van acht minuten om zowel het actieve leergedrag als het storende gedrag adequaat in kaart te kunnen brengen en daarmee een volledig beeld te krijgen van het leergedrag.

Verder bleek uit het onderzoek van Hintze en Matthews (2004) dat het verstandig lijkt om ook de categorie *niet aan taak* verder op te delen in deelcategorieën om de betrouwbaarheid van het instrument te vergroten. In dit onderzoek wordt daarom gekozen om ook binnen de categorie *niet aan taak* een indeling te maken in actief of passief gedrag. Dit betekent dat in dit onderzoek de categorieën *Actief leergedrag, Passief leergedrag, Off-task* en *Storend gedrag* zijn gebruikt. De vier categorieën kunnen worden gezien als een schaal waarop het gedrag wordt beoordeeld. Het meest positieve gedrag is actief leergedrag, terwijl het meest negatieve uiterste wordt weerspiegeld door storend gedrag.

Naast het beoordelen van de betrouwbaarheid van het instrument als geheel zal tevens worden onderzocht hoe betrouwbaar de verschillende categorieën worden gescoord. Er wordt verwacht dat passief gedrag vaker wordt geobserveerd dan actief gedrag. Om deze reden wordt verwacht dat het actieve gedrag een lagere betrouwbaarheid zal hebben dan het passieve gedrag, aangezien actief gedrag minder

wordt geobserveerd. De steekproef van passief gedrag wordt verwacht groter te zijn en zal om die reden een hogere betrouwbaarheid hebben (Thorndike, 1985).

Hoewel er enerzijds kan worden aangenomen dat met vier categorieën betrouwbaar kan worden geobserveerd, kan anderzijds worden verwacht dat met het instrument betrouwbaarder kan worden geobserveerd wanneer de categorieën worden samengevoegd tot een dichotome indeling, aangezien dit tot robuustere categorieën leidt. Er zal daarom in dit onderzoek een vergelijking worden gemaakt tussen de betrouwbaarheid van het instrument wanneer er vier categorieën worden gebruikt en wanneer er twee categorieën worden gebruikt. Er worden twee verschillende dichotome schalen gevormd, namelijk passief tegenover actief gedrag en positief tegenover negatief gedrag. Er wordt verwacht dat het actieve gedrag minder vaak zal worden geobserveerd dan het passieve gedrag en om die reden wordt verwacht dat passief gedrag betrouwbaarder kan worden gemeten (Thorndike, 1985). Ook wordt verwacht dat positief gedrag betrouwbaarder kan worden gescoord dan negatief gedrag wanneer gebruik wordt gemaakt van deze dichotome verdeling. Net als bij de schaal met actief en passief gedrag wordt verwacht dat positief gedrag betrouwbaarder wordt gescoord aangezien het meer zal voorkomen (Butler, 1990; Thorndike, 1985).

Samenhang leergedrag en leerprestatie

De tweede onderzoeksvraag heeft betrekking op de samenhang tussen leergedrag en leerprestatie. Hierbij wordt onderzocht of de samenhang verschilt bij verschillende vakken. Er wordt verwacht dat leerlingen tijdens de wiskundeles meer actief leergedrag vertonen, terwijl zij tijdens de taalles meer passief leergedrag laten zien (Gruber et al., 2004). Verder kan worden verwacht dat het leergedrag van de leerling betrouwbaarder kan worden beoordeeld tijdens Nederlands, dan bij rekenen. De laatste hypothese volgt uit de verwachting dat actief gedrag betrouwbaarder wordt gescoord dan passief gedrag. Aangezien uit het onderzoek van Gruber et al. (2004) bleek dat tijdens wiskunde van meer actief leergedrag sprake is wordt verwacht dat het gedrag tijdens dit vak betrouwbaarder kan worden gescoord. Als er sprake is van een samenhang tussen het leergedrag en leerprestatie dan zal deze groter zijn bij wiskunde dan bij taal, aangezien het leergedrag bij wiskunde betrouwbaarder kan worden gescoord. De laatste vraag die wordt beantwoord in dit onderzoek is de vraag of er een grotere samenhang is tussen leergedrag en leerprestatie bij jongens of bij meisjes. Dit is nog niet eerder onderzocht.

Hoewel de samenhang tussen leergedrag en leerprestatie al uitvoerig is bestudeerd door Morgan et al. (2008), heeft dit onderzoek een andere doelgroep om de samenhang tussen deze twee concepten voor deze doelgroep te bevestigen. In deze studie wordt namelijk onderzocht of de samenhang tussen leergedrag en leerprestaties ook bestaat voor een zeer specifieke groep leerlingen. Deze groep leerlingen volgt praktijkonderwijs en onderwijs voor nieuwkomers. Hoewel het onderzoek van Morgan et al. (2008) zich richtte op een groep kinderen, waarbij grote variatie bestond in onder andere leeftijd, schoolniveau, sociaal-economische status en ras, is het niet zeker of de gevonden samenhang ook bij deze specifieke groep bestaat. Tevens is het zo dat het onderzoek van Morgan et al. (2008) zich richtte op kinderen in de kleuterklas tot aan groep 5. Dit onderzoek richt zich op jongeren in de tweede klas van het voortgezet onderwijs en er kan daardoor met dit onderzoek worden onderzocht of de gevonden samenhang in de kleuterklas en groep 5 nog steeds van kracht is in de tweede klas van het voortgezet onderwijs. Een laatste beperking van het onderzoek van Morgan et al. (2008) is dat het zich alleen heeft gericht op de samenhang tussen leesproblemen en leergedrag. In deze studie zal naast de prestatie voor het vak Nederlands ook de prestatie voor het vak Rekenen worden onderzocht op zijn samenhang met leergedrag.

Onderzoeksvragen

1. In hoeverre is het voor dit onderzoek ontworpen observatie instrument betrouwbaar in het meten van leergedrag van leerlingen uit de tweede klas van het middelbaar onderwijs?
    - Heeft het instrument een voldoende interbeoordelaarsbetrouwbaarheid en test-hertestbetrouwbaarheid?
    - Zijn de beoordelingen van het gedrag betrouwbaarder voor een of meerdere van de vier categorieën dan voor de rest van de categorieën?
    - Wordt met een dichotome schaal bestaande uit de categorieën actief en passief gedrag betrouwbaarder geobserveerd dan met een schaal bestaande uit vier categorieën?
    - Wordt met een dichotome schaal bestaande uit de categorieën positief en negatief gedrag betrouwbaarder geobserveerd dan met een schaal bestaande uit vier categorieën?

2. Bestaat er een samenhang tussen het leergedrag van leerlingen en hun leerprestatie?
    - Is de samenhang tussen leergedrag en leerprestatie verschillend bij Nederlands en rekenen?
    - Is de samenhang tussen leergedrag en leerprestatie verschillend bij jongens en meisjes?

Methode

Participanten

Dit onderzoek is ondernomen in het kader van een project van een scholengroep en valt daarmee binnen het SLOA-project van de VO-raad. Het onderzoek is uitgevoerd bij twee scholen die onderdeel zijn van deze scholengroep en hiervoor was geen verdere selectie nodig. De scholen bevonden zich in de binnenstad van Den Haag, een middelgrote stad in de provincie Zuid-Holland. Het onderwijs dat werd aangeboden op de scholen was praktijkonderwijs op de ene school en onderwijs aan nieuwkomers op de andere school. Het onderzoek is uitgevoerd met hele klassen uit deze scholen. Aan dit onderzoek hebben 38 kinderen uit de tweede klas van het middelbare onderwijs meegedaan, uit vijf verschillende klassen. De groep leerlingen bestond uit 25 (65,8%) meisjes en 13 (34,2%) jongens. De leeftijd van de leerlingen in deze groep varieerde tussen de 161 maanden (13,4 jaar) en de 235 maanden (19,6 jaar). Zij waren gemiddeld 185 maanden (15,4 jaar) oud (SD = 19,14). Van de groep leerlingen volgden er 8 (21,1%) praktijkonderwijs en 30 (78,9%) onderwijs voor nieuwkomers. De leerlingen en hun ouders of verzorgers hebben passief toestemming gegeven voor het onderzoek. Dit betekent dat de ouders of verzorgers werd gevraagd zelf de onderzoekers ervan op de hoogte te brengen als zij wilden dat hun kind niet deelnam aan het onderzoek.

Design

In dit onderzoek werd een 2x2x2-design gehanteerd. De leerlingen konden op drie dimensies worden ingedeeld. Deze dimensies waren het vak (rekenen of taal), de week (eerste of tweede week) en de sekse (jongen of meisje).

Observatieschaal

Leerlingen zijn geobserveerd met een leerlinggedragbeoordelingslijst voor leerkrachten, gebaseerd op de Pupil Observation Procedure (Espin & Yell, 1994). De voornaamste aanpassingen op dit instrument waren een meer uitgebreide operationalisatie van de vier types leergedrag en het niet observeren van het gedrag van de leerkracht. Het gedrag werd in vier categorieën ingedeeld, namelijk *actief leergedrag, passief leergedrag, off-task* en *storend gedrag*. De definities voor deze vier vormen van leergedrag zijn vertalingen van de definities die zijn gebruikt door Espin en Yell (1994). Naast een definitie zijn voor elke categorie een aantal voorbeeldgedragingen gedefinieerd om de categorieën te operationaliseren. Hiervoor is gebruik gemaakt van voorbeelden uit onderzoeken van Espin en Yell (1994), Shapiro (2004; BOSS) en Hintze en Matthews (2004). Verder zijn voorbeelden toegevoegd waarvan de beoordelaars in overleg tot consensus zijn gekomen.

De definitie van actief leergedrag die in dit onderzoek is gebruikt was: 'De leerling is mondeling, schriftelijk of motorisch aan het reageren op vragen van de leraar of schriftelijk materiaal.' Voorbeelden van actief leergedrag zijn vragen stellen, hardop voorlezen, reageren op de instructie van de leerkracht, etc. De definitie van passief leergedrag was: 'De aandacht van de leerling is gericht op de taak zoals deze gedefinieerd is door de leraar. De ogen van de leerling zijn gericht op de huidige taak (dat wil zeggen: de ogen van de leerling zijn gericht op de leraar als deze instructie aan het geven is, en naar het leermateriaal als de leerling zelfstandig werkt).' Voorbeelden voor passief leergedrag zijn stillezen, luisteren naar de leerkracht of medeleerling, kijken naar de leermaterialen, etc.

Voor off-task gedrag is de volgende definitie gehanteerd: 'De aandacht van de leerling is niet gericht op de taak zoals deze gedefinieerd is door de leraar. De ogen van de leerling zijn niet gericht op de huidige taak.' Bij off-task gedrag kan worden gedacht aan gedragingen als staren, bladeren zonder te lezen, lezen van irrelevante informatie, etc. Tenslotte is voor de categorie storend gedrag de volgende definitie gebruikt: 'Elk gedrag, door de leerling veroorzaakt, dat inbreuk maakt op de leeromgeving van zichzelf of anderen.' Dit is gedrag als schreeuwen, voor de beurt praten, niet op de plaats zitten, etc.

Het doel van dit instrument was om snel en betrouwbaar deze vier categorieën leergedrag in kaart te brengen. In dit onderzoek zijn de vier categorieën tevens samengevoegd om twee dichotome schalen te vormen om te bestuderen of

betrouwbaardere metingen kunnen worden verkregen met deze schalen. Een doel van het gebruikte instrument wat in dit onderzoek niet aan de orde komt was het volgen van de ontwikkeling van het leergedrag. Het instrument maakt hierbij gebruik van het probleemoplossingsmodel, response to intervention en curriculum-based measurement om de ontwikkeling van het leergedrag van de leerling in kaart te brengen en te kunnen blijven volgen.

Het instrument maakte gebruik van momentary time sampling, waarbij de observatiesessie werd opgedeeld in gelijke intervallen van 10 seconden en het gedrag aan het begin van ieder interval werd gescoord. Gedurende één minuut is elk interval het gedrag van een individuele leerling gescoord in één van de vier categorieën. Wanneer het gedrag van een leerling binnen een bepaalde categorie viel, werd een score van 1 toegekend in deze categorie en een score van 0 in de drie overige categorieën. De beoordelaar had vooraf de volgorde waarin verschillende kinderen werden beoordeeld genoteerd.

Het leergedrag was een variabele op ordinaal niveau. Elke leerling haalde een score op elk van de vier categorieën. Deze categorieën waren op zichzelf variabelen op ratio niveau. De minimale score was hierbij 0, terwijl de maximale score per categorie het aantal minuten was dat een leerling was beoordeeld maal zes, aangezien er elke minuut zes metingen werden gedaan. Dit betekent dat wanneer een leerling in een minuut observatie driemaal actief leergedrag vertoonde, eenmaal storend gedrag vertoonde en tweemaal passief leergedrag vertoonde, hij of zij een 3 scoorde voor actief leergedrag, een 1 voor storend gedrag, een 2 voor passief leergedrag en een 0 voor off-task gedrag. Er is gerekend met optellingen van deze ruwe scores voor zowel de berekening van de interbeoordelaars- en test-hertestbetrouwbaarheid, als de samenhang tussen leergedrag en leerprestatie.

Betrouwbaarheid

Wanneer een instrument het leergedrag probeert te meten met systematische directe observatie zijn meerdere methoden mogelijk om de betrouwbaarheid van dat instrument te berekenen. Ten eerste wordt over het algemeen de interbeoordelaarsbetrouwbaarheid berekend voor het instrument (Hintze & Matthews, 2004; Johnston & Pennypacker, 1993). Voor een goede interbeoordelaarsbetrouwbaarheid is het van belang dat er een grote mate van overeenstemming is tussen twee observatoren die tegelijkertijd maar onafhankelijk van elkaar het gedrag beoordelen (Kazdin, 1982). De

observaties worden uitgevoerd gedurende twee weken waarin dezelfde leerlingen in iedere week tijdens een willekeurige taalles en een willekeurige rekenles worden geobserveerd. Aangezien wordt verwacht dat de omgevingsfactoren in de twee lessen in grote mate gelijkenissen vertonen (qua vak, klassenopstelling, positie in de klas etc.) wordt ook de test-hertestbetrouwbaarheid van dit instrument onderzocht. Dit is de consistentie van een respons van een deelnemer over de tijd heen. Wanneer een hoge correlatie wordt gevonden tussen de twee meetmomenten, dan is sprake van een hoge test-hertestbetrouwbaarheid (Leary, 2008).

Interbeoordelaarsbetrouwbaarheid

In de wetenschappelijke literatuur kunnen vele voorbeelden worden gevonden van grenzen van correlatiecoëfficiënten om een instrument als zijnde betrouwbaar te kunnen interpreteren. Zo geeft Leary (2008) aan dat een correlatiecoëfficiënt van .70 of hoger voldoende is om de interbeoordelaarsbetrouwbaarheid van een instrument te garanderen. Salvia en Ysseldyke (2004) vinden echter dat een coëfficiënt van .70 slechts voldoende is wanneer een instrument wordt gebruikt voor screening van leerlingen. Zij stellen dat een instrument pas voldoende betrouwbaar is om belangrijke programma- en instructieveranderingen te ondersteunen wanneer een coëfficiënt van .90 of hoger wordt bereikt. Het doel van deze studie was een screening van leerlingen op hun leergedrag en om die reden kan ook op basis van de richtlijnen van Salvia en Ysseldyke (2004) de coëfficiënt van .70 worden aangehouden. Ook in dit onderzoek wordt een benedengrens van .70 aangehouden voor een voldoende interbeoordelaarsbetrouwbaarheid.

Verder zal gebruik worden gemaakt van de grenzen die Cohen (1988) heeft opgesteld. Hij beschrijft dat een voldoende correlatie wordt gevonden bij een correlatie tussen de r = .30 en de r = .50. Alle waarden die lager zijn worden gezien als zijnde een kleine of zelfs niet substantiële correlatie. Een correlatie van r = .50 tot r = .70 wordt beschouwd als een hoge correlatie. Een zeer hoge correlatie wordt bereikt als deze tussen de r = .70 en r = .90 valt. Een correlatie is volgens Cohen (1988) bijna perfect als het r = .90 of hoger is. Deze interpretatie van Cohen (1988) is echter van een algemene aard en is niet toegespitst op het berekenen van de betrouwbaarheid van een instrument zoals de eerder genoemde onderzoeken dat wel waren. Daarom is de grens van een correlatie van .70 voor een voldoende betrouwbaarheid aangenomen, aangezien deze grens uit de eerder genoemde

onderzoeken naar voren kwam. Wel worden de grenzen van Cohen (1988) aangehouden voor een meer specifieke interpretatie van de correlatiewaarden. Dit betekent dat een correlatie van r = .90 of hoger als bijna perfect gezien wordt, terwijl een correlatie van r = .30 of lager als niet substantieel beschouwd wordt. Alle correlaties tussen de r = .30 en r = .70 worden beschouwd als zijnde matig, maar niet voldoende om de betrouwbaarheid van het instrument aan te tonen.

Voor het percentage van overeenkomst tussen observatoren zijn geen duidelijke criteria voor handen die uit onderzoek naar voren zijn gekomen (Topf, 1986). Wel bestaat er consensus binnen de gedragswetenschappen over welke percentages als grens dienen om een voldoende interbeoordelaarsbetrouwbaarheid te kenmerken. Hierbij wordt een percentage overeenkomst van minstens 70% als zijnde noodzakelijk beschreven en wanneer 90% overeenkomst wordt bereikt tussen observatoren dan is sprake van een goede betrouwbaarheid (House, House, & Campbell, 1981). De grens van een voldoende betrouwbaarheid zal op basis van deze gegevens een percentage van 70% overeenkomst zijn. Tijdens het beoordelen van de trainingsvideo behaalden de observators een interbeoordelaarsbetrouwbaarheid van gemiddeld 57%, wat niet voldoende is. De distinctie tussen positief en negatief gedrag werd wel voldoende gemaakt met een percentage van 72%. De hoogste interbeoordelaarsbetrouwbaarheid (74%) werd gevonden wanneer actief gedrag tegenover passief gedrag moest worden beoordeeld. Hieruit blijkt dat tijdens de training de dichotome schalen een voldoende betrouwbaarheid werd behaald, maar dat dit voor de indeling in vier categorieën niet het geval was.

Hoewel het instrument is ontworpen voor groepsdocenten, zal in dit onderzoek de observatie worden uitgevoerd door studentonderzoekers. Uit onderzoek van Chafouleas et al. (2010a) blijkt echter dat de betrouwbaarheid van de beoordeling van het gedrag door leerkrachten even goed is als de betrouwbaarheid van de beoordeling door onderzoekers. Een vergelijkbaar resultaat werd gevonden in onderzoek van Chafouleas, Mc Dougal, Riley-Tillman, Panahon en Hilt (2005), waar een grote mate van gelijkheid in het beoordelen van leergedrag tussen leerkrachten en externe observatoren werd gevonden. Hoewel in dit onderzoek niet de groepsdocent van de leerlingen het gedrag beoordeelt, maar een externe observator, kan toch worden verwacht dat de resultaten van dit onderzoek ook gelden voor de groepsdocent.

Test-hertestbetrouwbaarheid

De grenzen voor correlatiewaarden die bij de interbeoordelaarsbetrouwbaarheid zijn genoemd, waren bij de meeste studies vastgesteld voor de betrouwbaarheid in het algemeen. Alleen Leary (2008) heeft specifiek voor verschillende vormen van betrouwbaarheid aparte grenzen opgesteld. Voor de interbeoordelaars- en test-hertestbetrouwbaarheid zijn de grenzen hetzelfde. Leary (2008) beschrijft namelijk ook voor de test-hertestbetrouwbaarheid een correlatie van .70 of hoger als voldoende. Deze grens werd daarom ook voor de test-hertestbetrouwbaarheid gebruikt in dit onderzoek. Voor een meer gedetailleerde interpretatie van de betrouwbaarheid werden de grenzen van Cohen (1988) gebruikt, zoals beschreven bij de interbeoordelaarsbetrouwbaarheid.

Leerprestaties

Naast het gedrag is ook de leerprestatie van elke leerling bepaald. Hiervoor zijn cijfers gebruikt van het vak waarbij de leerling is geobserveerd. De vakken waarvan de cijfers zijn opgevraagd zijn Nederlands en rekenen/wiskunde. Door het leergedrag en de leerprestatie met elkaar te vergelijken, is onderzocht of er een relatie bestaat tussen deze variabelen. Hierbij is tevens het mentorcijfer van iedere leerling gebruikt. Het mentorcijfer is een cijfer dat de leerkracht aan de leerling geeft op basis van zijn of haar leergedrag en –prestatie. Om de subvragen binnen de tweede onderzoeksvraag te kunnen beantwoorden zijn naast de variabelen leergedrag en leerprestatie ook de variabelen sekse (jongen/meisje) en vak (Nederlands/rekenen) meegenomen in de berekeningen.

Leerprestatie was een variabele op interval niveau en bestond uit een enkel cijfer wat het gemiddelde is van alle cijfers die een leerling heeft behaald voor het specifieke vak waarbij hij is beoordeeld op zijn leergedrag. Aangezien er op de scholen werd gewerkt met cijfers op een schaal van tien, was de hoogst mogelijke waarde op deze variabele 10 en de laagst mogelijke waarde was 1. Het mentorcijfer was ook een variabele op interval niveau, maar dit cijfer varieerde tussen de 1 en de 4.

Procedure

De onderzoekers die het gedrag van de leerling voor dit onderzoek beoordeelden zijn vooraf getraind met behulp van een trainingsvideo. Voordat de onderzoekers in de klassen de leerlingen gingen observeren hebben zij eerst kennis gemaakt met de

docenten. De docenten waren in alle gevallen bekend gemaakt met de komst van de onderzoekers en het doel van het onderzoek en van de observatie. Voordat de les begon werden de beoordelaars aan de leerlingen voorgesteld en werd kort het doel van hun aanwezigheid in de klas uitgelegd. De leerlingen werd uitgelegd dat zij zo goed als zij konden de onderzoekers moesten negeren en de les moesten beschouwen als een normale les.

De beoordeling van het leergedrag werd gedaan door drie onderzoekers. De observatoren zaten tijdens de observatie voorin de klas op een plek waar zij de leerlingen zo goed mogelijk konden zien. De observatie van verschillende leerlingen werd uitgevoerd gedurende het hele lesuur. Wanneer een meting niet kon worden volbracht (bijvoorbeeld wanneer een leerling ziek was of onder de observatie de klas verliet) werd de observatie direct vervolgd bij de volgende leerling. Wanneer de eerdere observatie al was gestart werd deze zo snel mogelijk voortgezet, nadat de onderzoeker klaar was met een andere leerling. De verloren tijd werd op deze manier zo snel mogelijk ingehaald. Incomplete data (minuten waarvan niet elke tien seconden kon worden geobserveerd) werden niet meegenomen in het onderzoek. De observaties werden uitgevoerd in taallessen (Nederlands) en rekenlessen (Wiskunde).

Statistische analyse

Om de betrouwbaarheid van het gebruikte instrument te berekenen zijn twee correlaties berekend met behulp van de Pearson's correlatietoets. Dit zijn de interbeoordelaarsbetrouwbaarheid en de test-hertestbetrouwbaarheid. Met betrekking tot de interbeoordelaarsbetrouwbaarheid is tevens het percentage overeenkomst tussen de observatoren berekend. De interbeoordelaarsbetrouwbaarheid is per duo observatoren berekend. Aangezien er drie observatoren waren, levert dit drie duo's op waartussen de overeenkomst in beoordeling kan worden berekend.

De relatie tussen het leergedrag en de leerprestatie is berekend met behulp van een Pearson's correlatietoets. Er is voor deze toets gekozen, aangezien er geen sprake was van een afhankelijke en onafhankelijke variabele. De samenhang kon daarom het beste worden berekend met behulp van correlatietoetsen per categorie. Hierbij werd tevens onderscheid gemaakt in de groepen in sekse en in het vak dat is gevolgd. Met behulp van dit onderscheid konden de subvragen die hier betrekking op hadden worden beantwoord.

Resultaten

Descriptieve analyse

Uit de data die in twee weken is verzameld is gebleken dat niet alle categorieën van leergedrag in dezelfde frequentie voorkomen. Er is gebruik gemaakt van 608 valide minuten aan observatie in dit onderzoek. Aangezien er zes observaties per minuut werden gedaan, betekent dit dat er 3648 observaties hebben plaatsgevonden. Bij 868 observaties (23,8%) hiervan werd actief leergedrag geobserveerd. In 1781 (48,8%) van de observaties werd passief leergedrag geobserveerd. Er werd in totaal 823 keer (22,6%) off-task gedrag geconstateerd. Slechts 176 (4,8%) keer werd storend gedrag geobserveerd. Dit betekent dat de leerlingen het meest passief leergedrag vertoonden, terwijl off-task gedrag en actief leergedrag beiden ongeveer in een kwart van de gevallen voorkwamen. De leerlingen lieten minder storend gedrag zien dan de andere vormen van gedrag. Wanneer afzonderlijk naar de twee weken wordt gekeken valt op dat in de eerste week er relatief meer positief leergedrag is vertoond (actief en passief leergedrag), terwijl er relatief minder negatief gedrag is geobserveerd (off-task en storend gedrag) in deze week. In tabel 1 staan deze getallen samengevat.

Tabel 1.

Verdeling van observaties over de categorieën

		Actief	Passief	Off-task	Storend gedrag
Week 1	Absoluut	448	967	343	66
	Percentueel	24,6	53,0	18,8	3,6
Week 2	Absoluut	420	814	480	110
	Percentueel	23,0	44,6	26,3	6,0
Totaal	Absoluut	868	1781	823	176
	Percentueel	23,8	48,8	22,6	4,8

Voor iedere categorie is onderzocht of er sprake is van een normale verdeling. Hierbij is onderscheid gemaakt in de week waarin de data is verkregen, aangezien deze data ook gebruikt is voor het berekenen van de test-hertestbetrouwbaarheid. Met behulp van een Kolmogorov-Smirnov toets voor de normale verdeling is gebleken dat niet alle categorieën normaal zijn verdeeld volgens deze toets. Van de data die in de eerste week zijn verzameld waren de categorieën actief leergedrag (AL; p=.200), passief

leergedrag (PL; p=.200) en off-task gedrag (OT; p=.085) normaal verdeeld volgens deze toets. De categorie storend gedrag (SG; p<.001) is niet normaal verdeeld volgens deze toets. De categorieën van de tweede week laten eenzelfde uitkomst zien. De categorieën actief leergedrag (p=.200), passief leergedrag (p=.200) en off-task gedrag (p=.200) zijn in deze week normaal verdeeld, terwijl de categorie storend gedrag (p=.008) in de tweede week niet normaal was verdeeld volgens de Kolmogorov-Smirnov toets. De variabelen actief (AG) en passief gedrag (PG) en positief (POS) en negatief (NEG) gedrag van de dichotome schalen waren in beide weken normaal verdeeld (p=.200 in alle gevallen).

Ook de leerprestaties zijn numerieke variabelen en voor de toetsen die worden gebruikt bij de tweede hoofdvraag moeten deze ook normaal verdeeld zijn. Niet van alle 38 leerlingen is zowel data voor het leergedrag en de leerprestatie aanwezig, wat leidt tot een kleiner aantal leerlingen. De verdelingen van de leerprestatie voor het vak Nederlands en het vak Rekenen waren beiden normaal verdeeld volgens de Kolmogorov-Smirnov toets. De verdeling van het mentorcijfer dat de leerlingen hebben gekregen was echter niet normaal verdeeld volgens deze toets.

Aangezien de Kolmogorov-Smirnov in sommige gevallen een te strenge toets is voor normaliteit is tevens de gestandaardiseerde scheefheid (skewness) en gepiektheid (kurtosis) berekend. Om deze maat te berekenen wordt de scheefheid/gepiektheid door de bijbehorende standaardmeetfout gedeeld. Er kan worden aangenomen dat een variabele normaal verdeeld is als deze maat tussen de -3 en de 3 valt. Eerst is een logaritmische transformatie uitgevoerd op de variabelen storend gedrag van week 1 en week 2, waardoor de verdeling meer een normale verdeling benaderde. Hoewel deze getransformeerde variabelen volgens de Kolmogorov-Smirnov toets nog niet als normaal werden geïdentificeerd, vielen de gestandaardiseerde scheefheid en gepiektheid binnen de grenzen van -3 en 3. Voor het mentorcijfer hoefde geen logaritmische transformatie te worden uitgevoerd. De originele verdeling had een goede gestandaardiseerde scheefheid en gepiektheid. In tabel 2 zijn de descriptieve waarden van alle categorieën samengevat.

Bij het opstellen van het databestand waarmee is gewerkt is zorgvuldig te werk gegaan, om ervoor te zorgen dat er geen sprake is van missende data. Het uiteindelijke databestand bevatte daarom geen enkele missende waarde. Na een analyse van uitbijters en extreme waarden met behulp van boxplots konden slechts enkele uitbijters worden gevonden en nooit meer dan één per onderzochte variabele.

Een waarde werd hierbij als uitbijter geïnterpreteerd, wanneer hij anderhalve box van de centrale box af lag. Wanneer de waarde drie boxen van de centrale box aflag werd deze als een extreme waarde gezien. De uitbijters waren niet verwijderd voor de verschillende berekeningen, aangezien de data normaal verdeeld waren.

Tabel 2.

Beschrijvende Gegevens van de Categorieën en leerprestaties

	N	*M*	S_d	Gest. Scheefheid*	Gest. Gepiektheid**	*p* (K-S)***
Actief leergedrag week 1	38	11,79	6,01	1,49	0,16	.200
Passief leergedrag week 1	38	25,45	7,54	1,03	-0,73	.200
Off-task gedrag week 1	38	9,03	5,38	2,85	1,35	.085
Storend gedrag week 1	38	0,33	0,31	0,86	-1,74	<.001
Actief leergedrag week 2	38	11,05	4,86	0,42	-0,40	.200
Passief leergedrag week 2	38	21,42	8,68	-0,01	-0,04	.200
Off-task gedrag week 2	38	12,63	8,13	2,29	1,40	.200
Storend gedrag week 2	38	0,45	0,37	0,43	-1,74	.008
Actief gedrag week 1	38	13,53	6,29	0,97	-0,89	.200
Passief gedrag week 1	38	34,47	6,29	-0,97	-0,89	.200
Positief gedrag week 1	38	37,24	6,34	-2,23	2,01	.200
Negatief gedrag week 1	38	10,76	6,34	2,23	2,01	.200
Actief gedrag week 2	38	13,95	5,80	-0,48	-1,05	.200
Passief gedrag week 2	38	34,05	5,80	0,48	-1,05	.200
Positief gedrag week 2	38	32,47	8,83	-0,78	-0,42	.200
Negatief gedrag week 2	38	15,53	8,83	0,78	-0,42	.200
Cijfer Nederlands	29	6,87	0,86	1,18	0,03	.200
Cijfer Rekenen	18	7,04	1,00	2,30	0,99	.107
Mentorcijfer	36	2,89	0,78	-0,44	-0,64	<.001

* Gest. Scheefheid = scheefheid/standaardmeetfout na eventuele logtransformatie.

** Gest. Gepiektheid = gepiektheid/standaardmeetfout na eventuele logtransformatie.

*** *p* (K-S) = p-waarde op de Kolmogorov-Smirnov toets voor normaliteit.

Betrouwbaarheid van de losse categorieën

De eerste hoofdvraag die in dit onderzoek werd gesteld was de volgende: *In hoeverre is het voor dit onderzoek ontworpen observatie instrument betrouwbaar in het meten van leergedrag van leerlingen uit de tweede klas van het middelbaar onderwijs?* Om deze vraag te beantwoorden zijn de test-hertestbetrouwbaarheid en de interbeoordelaarsbetrouwbaarheid van het instrument berekend. Op deze wijze is bij beantwoording van de eerste hoofdvraag tegelijkertijd een antwoord gegeven op de eerste twee subvragen die hierbij zijn gesteld. Deze twee vragen waren namelijk de volgende: *Heeft het instrument een goede interbeoordelaarsbetrouwbaarheid en test-hertestbetrouwbaarheid?* en *Zijn de beoordelingen van het gedrag betrouwbaarder voor een of meerdere van de vier categorieën dan voor de rest van de categorieën?*

Test-hertestbetrouwbaarheid

Met behulp van de Pearson's correlatietoets is de test-hertestbetrouwbaarheid berekend. Hiervoor zijn de data die zijn verzameld in de eerste week, voor iedere categorie vergeleken met de data die zijn verzameld in de tweede week met behulp van de Pearson's correlatietoets. Zoals in tabel 3 te zien is, bestaat er geen significante samenhang voor de categorieën Actief Leergedrag (AL) en Off-task gedrag (OT), maar is er wel sprake van een significante samenhang voor de categorieën Passief Leergedrag (PL) en Storend Gedrag (SG). Een correlatie van .70 wordt voor geen van de categorieën bereikt.

Tabel 3

Test-hertestbetrouwbaarheid van de categorieën.

	r	*p*
AL	.064	.702
PL	.381	.018
OT	.258	.118
SG	.351	.031
AG	.184	.268
PG	.184	.268
POS	.298	.069
NEG	.298	.069

Interbeoordelaarsbetrouwbaarheid

Om de interbeoordelaarsbetrouwbaarheid te onderzoeken zijn de Pearson's correlatietoets en het percentage van overeenkomst tussen de duo's observatoren berekend. In een aantal lessen is het gedrag van dezelfde leerlingen op hetzelfde moment beoordeeld door twee onafhankelijke onderzoekers. De observatoren die dit duo bevatte, wisselden echter en om die reden zijn drie duo's ontstaan met ieder hun eigen correlatie en percentage overeenkomst, zoals in tabel 4 is weergegeven. Het gemiddelde percentage overeenkomst was 84,6%. Wanneer de percentages overeenkomst per categorie werden bestudeerd, werd duidelijk dat off-task gedrag het laagste percentage overeenkomst heeft (64,5%), terwijl de categorieën actief leergedrag, passief leergedrag en storend gedrag allemaal een overeenkomst in beoordeling van op of rond de 75% behaalden (zie tabel 4).

Tevens is de Pearson's correlatietoets uitgevoerd om de interbeoordelaarsbetrouwbaarheid te berekenen. Per categorie is berekend in hoeverre de beoordelaars overeenstemden. Aangezien er verschillende leerlingen door de afzonderlijke duo's zijn beoordeeld en de data om deze reden niet vergelijkbaar was, is geen correlatie voor het totaal berekend. Opvallend is de lage correlatie van r = .432 (p = .468) binnen de categorie off-task gedrag (OT). Binnen de categorieën actief leergedrag (AL) is eenmaal een correlatie van r = .731 (p = .039) gevonden. De categorie passief leergedrag (PL) had binnen één duo een correlatie van r = .761 (p = .135). Binnen de categorie storend gedrag (SG) werd voor één duo een correlatie van r = .745 (p = .034) gevonden. Verder zijn in alle categorieën alleen correlaties van r = .945 (p = .015) of hoger gevonden (zie tabel 4).

Tabel 4

Interbeoordelaarsbetrouwbaarheid totaal en per categorie.

		obs. 1 – obs. 2	obs. 2 – obs. 3	obs. 1 – obs. 3	totaal
Totaal %		88,7%	82,5%	80,8%	84,6%
AL	%	76,7%	75,0%	64,3%	74,2%
	r	.731	.945	.958	
	p	.039	.015	.010	
PL	%	82,2%	75,3%	70,4%	76,7%

	r	.957	.761	.955	
	p	<.001	.135	.011	
OT	%	77,1%	56,3%	60,0%	64,5%
	r	.968	.432	.981	
	p	<.001	.468	.003	
SG	%	87,5%	75,0%	70,0%	75%
	r	.745	.968	.986	
	p	.034	.007	.002	
AG	%	82,0%	71,4%	78,1%	77,8%
	r	.883	.912	.408	
	p	.004	.031	.495	
PG	%	92,9%	89,4%	92,6%	91,7%
	r	.883	.912	.408	
	p	.004	.031	.495	
POS	*%*	92,6%	86,7%	74,4%	86,0%
	r	.949	.995	.626	
	p	<.001	<.001	.295	
NEG	*%*	76,7%	62,9%	62,7%	59,9%
	r	.949	.995	.626	
	p	<.001	<.001	.295	

Betrouwbaarheid van de dichotome schalen

Actief en passief gedrag

Naast de vraag of er verschillen in betrouwbaarheid bestaan in het scoren van de vier afzonderlijke categorieën, is ook de vraag gesteld of er een verschil bestaat in de betrouwbaarheid van het beoordelen van actief of passief gedrag. Deze subvraag is de volgende: *Is er een verschil in betrouwbaarheid van het beoordelen van actief gedrag (actief leergedrag en storend gedrag) en passief gedrag (passief leergedrag en off-task)?* Ook voor deze subvraag is de test-hertestbetrouwbaarheid en de interbeoordelaarsbetrouwbaarheid beoordeeld. In het geval van de test-hertestbetrouwbaarheid is een Pearson's correlatietoets uitgevoerd. Voor de interbeoordelaarsbetrouwbaarheid is ook een percentage van overeenkomst tussen de observatoren berekend. De correlatie

van het actieve gedrag (AG) en het passieve gedrag (PG) is logischerwijs met elkaar verbonden. Elk gedrag is gescoord en de beoordeling valt altijd in één van de twee categorieën. Als er daarom een verschil in score bestaat in Actief Gedrag, dan moet ditzelfde verschil ook in Passief Gedrag bestaan, aangezien elk gedrag wordt gescoord. Om deze reden zal de correlatie van scores voor Actief Gedrag altijd even hoog zijn als de correlatie voor Passief Gedrag. Het is daarom in dit geval betekenisvoller om te kijken naar de percentages van overeenkomst om het verschil in betrouwbaarheid van scoren te onderzoeken. Uit de percentages blijkt dat voor de categorie Actief Gedrag gemiddeld 77,8% overeenstemming tussen de observatoren bestond, terwijl dit voor de categorie Passief Gedrag gemiddeld 91,7% was.

Voor de berekening van de test-hertestbetrouwbaarheid is tevens een Pearson's correlatietoets gebruikt. Ook hier is het vanuit theoretisch perspectief logisch dat de correlatie tussen de verschillende meetmomenten even hoog is. Zoals in tabel 3 is te zien is er sprake van een correlatie van $r = .184$ ($p = .268$) voor zowel de categorie Actief Gedrag als Passief Gedrag.

Positief en negatief gedrag

De laatste subvraag van de eerste hoofdvraag had betrekking op een andere dichotome indeling van de schaal die kan worden gemaakt. Deze indeling bestaat uit positief gedrag (POS) enerzijds en negatief gedrag (NEG) anderzijds. Zoals is weergegeven in tabel 4 is het percentage overeenstemming voor positief gedrag gemiddeld 86,0%. Voor de categorie negatief gedrag is dit 59,9%. Aangezien hier wederom sprake is van een dichotome verdeling, zijn de correlatiewaarden van de interbeoordelaarsbetrouwbaarheid voor beide categorieën hetzelfde. Ditzelfde geldt voor de correlatiewaarden van de test-hertestbetrouwbaarheid. In tabel 3 is te zien dat deze voor beide categorieën een waarde van $r = .298$ met een bijbehorende significantiewaarde van $p = .069$ hebben.

Samenhang leergedrag en leerprestatie

De tweede hoofdvraag van dit onderzoek heeft betrekking op de relatie tussen leergedrag en leerprestatie. De specifieke vraag die hierbij werd gesteld was: *Bestaat er een samenhang tussen het leergedrag van leerlingen en hun leerprestatie?* Hierbij werd onderscheid gemaakt in het vak dat is gevolgd tijdens de observatie. Observaties hebben plaatsgevonden tijdens de lessen Nederlands en rekenen. Daaruit volgt de

volgende subvraag: *Is de samenhang tussen leergedrag en leerprestatie verschillend bij Nederlands en rekenen?* Om deze vraag te kunnen beantwoorden is een Pearson's correlatietoets uitgevoerd om de correlatie te berekenen tussen het leergedrag en het rapportcijfer voor rekenen enerzijds en het leergedrag en het rapportcijfer voor Nederlands anderzijds. Zoals te zien in tabel 5 heeft het cijfer dat leerlingen behalen voor het vak Nederlands geen significante samenhang met alle vier de categorieën. Ook binnen de dichotome schalen was er geen sprake van een significante samenhang met het cijfer voor het vak Nederlands. Vergelijkbare resultaten werden gevonden tussen het cijfer voor het vak rekenen en de vier categorieën. Ook binnen de dichotome schalen was er geen sprake van significante relaties met het cijfer voor het vak rekenen.

Tabel 5.

Correlaties leergedrag (4 categorieën) en leerprestatie

		AL	PL	OT	SG
Cijfer rekenen	Totaal *r* (*p*)	-.129 (.588)	.220 (.351)	.031 (.897)	-.326 (.161)
	Jongens *r* (*p*)	-.315 (.605)	.314 (.607)	-.072 (.908)	-.370 (.540)
	Meisjes *r* (*p*)	-.128 (.650)	.273 (.324)	.034 (.904)	-.403 (.136)
Cijfer Ned.	Totaal *r* (*p*)	-.022 (.909)	.135 (.477)	-.195 (.301)	.239 (.204)
	Jongens *r* (*p*)	.002 (.995)	.244 (.527)	-.382 (.311)	.510 (.160)
	Meisjes *r* (*p*)	-.173 (.452)	-.019 (.936)	.114 (.622)	.205 (.372)

Tabel 6.

Correlaties leergedrag (2 categorieën) en leerprestatie

		AG	PG	POS	NEG
Cijfer rekenen	Totaal *r* (*p*)	-.266 (.258)	.266 (.258)	.126 (.598)	-.126 (.598)
	Jongens *r* (*p*)	-.348 (.565)	.348 (.565)	.122 (.846)	-.122 (.846)
	Meisjes *r* (*p*)	-.319 (.246)	.319 (.246)	.160 (.568)	-.160 (.568)
Cijfer Ned.	Totaal *r* (*p*)	.069 (.718)	-.069 (.718)	.122 (.522)	-.122 (.522)
	Jongens *r* (*p*)	.202 (.603)	-.202 (.603)	.240 (.535)	-.240 (.535)
	Meisjes *r* (*p*)	-.088 (.704)	.088 (.704)	-.172 (.456)	.172 (.456)

Onderscheid in sekse

De laatste subvraag bij de tweede hoofdvraag is als volgt: *Is de samenhang tussen leergedrag en leerprestatie verschillend bij jongens en meisjes?* Aangezien er sprake is van een samenhang en geen causaal verband zijn correlatietoetsen uitgevoerd. De Pearson's correlatietoets die voor de groep als geheel is uitgevoerd is wederom uitgevoerd, maar nu eenmaal voor de groep jongens en eenmaal voor de groep meisjes. In tabel 5 en 6 zijn de uitkomsten van deze toetsen onder de totaalscore weergegeven. De correlaties tussen het leergedrag en de leerprestatie van zowel de jongens als de meisjes bij het vak rekenen waren niet significant voor alle vier de categorieën. Dezelfde toets is gebruikt om de correlatie tussen de categorieën van het leergedrag en de leerprestatie voor jongens en meisje apart te berekenen voor het vak Nederlands. Voor zowel de meisjes als de jongens kon voor geen van de vier categorieën een significant verband tussen het leergedrag en de leerprestatie worden gevonden.

Ook zijn deze berekeningen gedaan voor de samengevoegde categorieën om te onderzoeken of de samenhang tussen leerprestatie en leergedrag andere waarden aannam wanneer er minder variatie in de data van het leergedrag zat. Dit betekent dat dezelfde berekeningen zijn uitgevoerd voor de categorieën actief gedrag (AG) en passief gedrag (PG) enerzijds en positief gedrag (POS) en negatief gedrag (NEG) anderzijds. In tabel 6 zijn deze waarden overzichtelijk samengevat. Net als bij de correlaties van de samenhang voor de gehele groep leerlingen, waren ook bij de

indeling in jongens en meisjes de correlatiewaarden van de tegenovergestelde categorieën even hoog, aangezien gebruik wordt gemaakt van een dichotome schaal. Dit betekent eveneens dat wanneer de ene waarde positief was, de tegenovergestelde waarde negatief was. Als leerlingen bijvoorbeeld meer actief gedrag lieten zien als zij een hoger cijfer haalden, dan betekent dit impliciet dat zij ook minder passief gedrag vertoonden als zij een hoger cijfer haalden. Voor zowel de jongens als de meisjes werden geen significante relaties tussen de leerprestatie voor het vak rekenen en de dichotome schalen gevonden. Eenzelfde beeld werd gevonden wanneer de samenhang tussen de leerprestatie voor Nederlands en het leergedrag werd onderzocht. Ook voor dit vak werden geen relaties gevonden tussen het leergedrag en de leerprestatie, wanneer gebruik werd gemaakt van de dichotome schalen.

Discussie

Samenvatting doel van de studie

Dit onderzoek had een tweeledig doel. Ten eerste is de betrouwbaarheid van het gebruikte instrument onderzocht. Dit instrument is gebruikt om het leergedrag van leerlingen mee in kaart te brengen. Aangezien de resultaten die met het instrument worden verkregen, worden gebruikt om het onderwijs van leerlingen aan te passen, is het van groot belang dat betrouwbare observaties kunnen worden gedaan en adequate conclusies worden getrokken uit data die met dit instrument zijn verkregen. De betrouwbaarheid van het instrument geeft een inzicht in de mogelijkheden van het instrument om consistente informatie over het leergedrag van leerlingen te kunnen geven. Pas wanneer een instrument betrouwbaar is, laten verschillen in de verkregen data ook daadwerkelijk verschillen in het leergedrag van de leerling zien (Leary, 2008). Dit betekent dat alleen inhoudelijke beslissingen kunnen worden gemaakt voor de ondersteuning van leerlingen in hun leergedrag, wanneer het instrument dat is gebruikt om dat leergedrag te meten ook betrouwbaar is. Het eerste doel van dit onderzoek was daarom om vast te stellen of het gebruikte instrument voldoende betrouwbaar was. Voor de beantwoording van de tweede onderzoeksvraag is van dit instrument gebruik gemaakt. Dit tweede doel was te onderzoeken of er sprake is van een samenhang tussen het leergedrag en de leerprestatie van leerlingen. Mocht dit het geval zijn, dan lijkt het verstandig om naast het ontwikkelen van de cognitieve

vermogens van kinderen (wat al veel gebeurd in scholen) ook in te zetten op de ontwikkeling van het leergedrag om indirect de leerprestatie te verbeteren.

Interpretatie en verklaring van de resultaten

Opvallend is dat slechts 72,6% van de observaties actief leergedrag of passief leergedrag betreft. Dit is opvallend, aangezien uit onderzoek van Butler (1990) is gebleken dat deze vormen van leergedrag in 80% van de observaties voorkomt. Het beperkte positieve leergedrag van de leerlingen is wellicht te verklaren door het soort onderwijs dat deze leerlingen volgen. Jongeren die twee jaar of minder in Nederland wonen, komen in aanmerking voor het onderwijs voor nieuwkomers. Het is mogelijk dat een gebrek in kennis van de Nederlandse taal de leerlingen demotiveert en tot gevolg heeft dat leerlingen minder actief en/of passief leergedrag vertonen. Leerlingen die niet of slechts beperkt bekend zijn met de Nederlandse taal laten minder motivatie zien om aan-taak gedrag te vertonen, wat het gevolg is van meerdere faalervaringen (Bender & Wal, 1994). Van jongeren in het praktijkonderwijs is het bekend dat zij over het algemeen meer ondersteuning nodig hebben om zich te concentreren op hun schooltaken (Schafrat, 2007). Het is om die reden te verklaren dat zij minder lang hun concentratie bij hun taak kunnen houden. Er is geen norm beschikbaar voor het percentage storend gedrag dat leerlingen laten zien in de klas. Er kwam echter in deze studie niet opvallend veel storend gedrag voor bij de kinderen.

Interbeoordelaarsbetrouwbaarheid

Het blijkt dat voor de vier categorieën (met ieder drie duo's observatoren) er in acht van de twaalf gevallen een bijna perfecte correlatie (r = .90 of hoger) wordt gevonden. Van de overige vier correlaties zijn er 3 voldoende (r = .70 tot r = .90) en er komt slechts bij één duo observatoren een matige correlatie (r = .30 tot r = .70) voor. Uiteindelijk waren twee van de twaalf correlaties niet significant, zo bleek uit de toets. Echter, tien van de twaalf correlaties waren dit wel. Dit betekent dat de interbeoordelaarsbetrouwbaarheid van het instrument voldoende kan worden bevonden op basis van de correlatiewaarden wanneer de beoordeling in 4 categorieën kan worden ingedeeld.

Hoewel werd verwacht dat de correlaties hoger zouden uitvallen wanneer met minder categorieën werd gewerkt, kwam dit uit de resultaten niet duidelijk naar voren. Hoewel bij een indeling in actief en passief gedrag twee van de drie duo's een

voldoende correlatie behaalden (waarvan één bijna perfect), was de correlatie van het derde duo matig. Ook in een indeling in positief en negatief gedrag viel ditzelfde duo uit en behaalde zij slechts een matige correlatie, waar de andere twee duo's beiden een bijna perfecte correlatie behaalden.

Het feit dat dichotome schalen geen hogere interbeoordelaarsbetrouwbaarheid laten zien kan op twee manieren verklaard worden. Ten eerste lijken de lage correlaties voornamelijk bij één van de drie duo's voor te komen. Het is mogelijk dat de beoordelaars van dit duo het gedrag op een verschillende manier beoordeelden, terwijl de andere duo's het gedrag op eenzelfde manier beoordeelden. Het lijkt er op dat het afwijkende duo de correlatie op een negatieve manier beïnvloedt. Wanneer de correlaties van dit duo niet in ogenschouw worden genomen vallen vier bijna perfecte correlaties weg bij de indeling in vier categorieën, terwijl vier matige correlaties wegvallen in de dichotome schalen. In de correlaties die overblijven ontstaat dan alsnog een beeld waarin de dichotome schalen een meer betrouwbare meting geven van het leergedrag. Een tweede verklaring is het feit dat bij een indeling in vier categorieën sprake lijkt te zijn van een plafondeffect. De correlaties en daarmee ook de betrouwbaarheid zijn in deze categorieën al zeer hoog en laten weinig verbetering toe door ze samen te voegen in dichotome schalen.

Een subdoel bij het onderzoeken van de betrouwbaarheid was om te onderzoeken of één of meerdere categorieën betrouwbaarder of juist minder betrouwbaar was te beoordelen dan de andere. Wanneer naar de correlaties wordt gekeken, blijken slechts twee van de twaalf correlaties niet significant te zijn. Deze twee correlaties vallen in de categorieën passief leergedrag en off-task gedrag. Dit betekent dat in beide schalen maar één keer een niet-significante correlatie voorkomt, terwijl de correlaties van de andere twee duo's wel significant zijn. Het lijkt er daardoor op dat de niet-significante correlatie niet consistent voorkomt. Een andere meetmethode kan wellicht een beter inzicht geven in de vraag of één of meerdere categorieën betrouwbaarder kan worden gescoord, namelijk het percentage overeenkomst tussen observatoren. Wanneer hier naar wordt gekeken komen de observatoren voor drie van de vier categorieën gemiddeld voor ongeveer 75% overeen in hun beoordeling. Alleen de categorie off-task gedrag lijkt hier een uitval te laten zien, aangezien de duo's voor deze categorie gemiddeld in slechts 64,5% van de gevallen overeenkomen in hun scores. Dit betekent dat het instrument als geheel betrouwbaar is, maar dat de categorie off-task gedrag met dit instrument niet

betrouwbaar kan worden gemeten. Hoewel is geprobeerd om een uitgebreide definitie en operationalisatie van de categorie op te stellen in dit onderzoek, is gebleken dat de beoordelaars in dit onderzoek toch te vaak het gedrag dat binnen deze categorie viel verschillend beoordeelden. Het is opvallend dat juist dit de categorie is die in eerdere onderzoeken, zoals die van Hintze & Matthews (2004) en die van Chafouleas et al. (2010a), niet is opgenomen. Hoewel er geen verklaring wordt gegeven voor het feit dat zij deze categorie niet hebben opgenomen, is het mogelijk dat het geen duidelijk afgebakende categorie is. Dit kan een verklaring zijn voor het feit dat de categorie niet betrouwbaar wordt gemeten met dit instrument.

Wanneer eenzelfde vergelijking wordt gemaakt binnen de dichotome schalen is er sprake van een groter verschil in percentage overeenkomst tussen de categorieën. Wanneer de indeling actief tegenover passief gedrag wordt gemaakt vallen beide percentages van overeenkomst boven de 70%, wat betekent dat beide categorieën betrouwbaar kunnen worden gemeten. Aangezien de categorie passief gedrag een percentage van hoger dan 90% overeenkomst tussen de observatoren heeft, kan tevens worden gezegd dat er voor deze categorie sprake is van een goede interbeoordelaars-betrouwbaarheid. Dit betekent dat de categorie passief gedrag betrouwbaarder kan worden gemeten dan actief gedrag. Dit was verwacht vanuit de literatuur, aangezien werd verwacht dat positief gedrag meer zou voorkomen en daardoor betrouwbaarder te meten zou zijn. (Thorndike, 1985).

Wanneer de dichotome schaal waarin positief tegenover negatief gedrag wordt gesteld wordt bestudeerd, valt op dat het positieve gedrag betrouwbaar kan worden gemeten aangezien een percentage van boven de 70% wordt bereikt. Voor de categorie negatief gedrag werd deze grens niet bereikt en deze categorie kan daardoor niet betrouwbaar worden gemeten. Dit betekent dat het positieve gedrag betrouwbaarder kan worden gemeten dan het negatieve gedrag. Dit resultaat was verwacht, aangezien werd verwacht dat positief gedrag meer zou voorkomen en daardoor betrouwbaarder te meten zou zijn (Butler, 1990; Thorndike, 1985). Dat er geen betrouwbare meting van negatief gedrag kan worden verkregen is te verklaren vanuit het feit dat deze categorie een samenvoeging is van het off-task gedrag en het storende gedrag. Daardoor is het mogelijk dat dit lage percentage voortkomt uit het lage percentage dat werd gevonden voor de categorie off-task gedrag.

Test-hertestbetrouwbaarheid

De test-hertestbetrouwbaarheid van het instrument is laag gebleken. Hoewel voor de categorieën passief leergedrag en storend gedrag nog een statistisch significante correlatie werd gevonden, waren de correlaties matig en niet voldoende. De correlaties van de test-hertestbetrouwbaarheid van de categorieën actief leergedrag en off-task gedrag waren statistisch niet significant. De correlaties van deze twee categorieën waren erg klein (off-task gedrag) of niet substantieel (actief leergedrag).

Een mogelijke verklaring hiervoor, die ook door Johnston en Pennypacker (1993) en Hintze en Matthews (2004) wordt genoemd, is dat de situatie waarin de leerling leergedrag laat zien verschillend is tussen de observatiemomenten, wat tot gevolg heeft dat ook het leergedrag van de leerling anders is. Om de test-hertestbetrouwbaarheid te berekenen zijn in dit onderzoek de data van een willekeurige rekenles en een willekeurige taalles voor iedere week (vier lessen in totaal) gebruikt. De data zijn daardoor niet afkomstig van lessen op precies hetzelfde moment op de dag of zelfs van lessen op verschillende dagen in de week onderwezen door verschillende docenten. Het gevolg zou daardoor kunnen zijn dat de situatie anders is voor de leerlingen. Dit heeft de eerder beschreven nadelige uitwerking op de test-hertestbetrouwbaarheid, aangezien het de verwachting dat het leergedrag over de weken hetzelfde blijft minder aannemelijk maakt. Om deze verklaring te onderzoeken is de toets opnieuw gedaan, maar nu alleen met de leerlingen van wie de observatie wel in iedere week op hetzelfde moment heeft plaatsgevonden. Opvallend was dat de correlatie niet verbeterde en zelfs slechter werd wanneer de toets alleen met deze leerlingen werd uitgevoerd. De correlaties werden voor alle categorieën, behalve het storend gedrag, lager. Ook de correlaties van de categorieën actief gedrag en passief gedrag waren voor deze groep kinderen juist lager. De lage test-hertestbetrouwbaarheid lijkt daardoor niet te verklaren door verschillen in de situatie waarin de leerlingen werden geobserveerd.

Een tweede mogelijke verklaring is het feit dat een aantal leerlingen met extreme verschillen in hun gedrag, het gehele beeld in zo'n grote mate beïnvloeden dat significante correlaties niet langer significant zijn. Om deze reden is dezelfde correlatietoets nogmaals uitgevoerd voor de groep leerlingen die minder dan 15 scores verschil hadden over de weken in een willekeurige categorie. Er is voor een verschil van 15 gekozen, aangezien dit er toe leidde dat de meest extreme verschillen niet werden meegenomen terwijl er wel een grote groep leerlingen overbleef om de

berekeningen mee uit te voeren. Dit leidde er namelijk toe dat slechts tussen de drie en zes leerlingen die varieerden tussen de categorieën niet werden meegenomen in de toetsing. Echter, in de categorie storend gedrag kon bij geen enkele leerling een verschil in scores van 15 of meer worden gevonden en kon dit niet als verklaring worden aangedragen. Toch waren de resultaten zonder deze groep leerlingen in grote mate verschillend van de resultaten met deze groep leerlingen bij alle andere categorieën. Zo blijkt dat met deze meer selecte groep voor elke categorie een significante correlatie voor de test-hertestbetrouwbaarheid wordt bereikt. Het is daarom zeer goed mogelijk dat de leerlingen met zeer extreme verschillen in hun scores de data in een dermate grote mate beïnvloeden dat het gehele beeld van de test-hertestbetrouwbaarheid is beïnvloed.

Er is geprobeerd om de extreme verschillen die zijn gevonden te verklaren. Ten eerste is onderzocht of er sprake kan zijn geweest van een leereffect. Het is mogelijk dat de leerlingen in de tweede week beter wisten wat de onderzoekers kwamen doen en dat zij daarom in de tweede week meer positief gedrag vertoonden en minder negatief gedrag (Rousson, Gasser, & Seifert, 2002). Het bleek echter dat het beeld totaal andersom was. Deze leerlingen lieten namelijk in de tweede week minder postief gedrag zien en meer negatief gedrag. Dit geeft aanleiding voor een andere mogelijke verklaring. Het is namelijk zeer goed mogelijk dat een bepaalde vorm van angst voor en bewustzijn van de observatoren ervoor zorgt dat leerlingen in de eerste week sociaal gewenst gedrag laten zien. Dit wordt reactiviteit genoemd (Davidshofer, Murphy, & Charles, 2005). De leerlingen zouden dan in de tweede week meer gewend zijn aan de observatoren in de klas en meer natuurlijk gedrag laten zien dan in de eerste week. Het is mogelijk dat de leerlingen met de meest extreme verschillen meer beïnvloedbaar zijn door hun omgeving, dan de andere leerlingen en daardoor grotere verschillen laten zien in hun gedrag over de twee weken. Wanneer een leerling meer dan gemiddeld afhankelijk is van zijn of haar omgeving in zijn leerproces dan wordt dit differentiële ontvankelijkheid genoemd (Kegel, Bus, & van IJzendoorn, 2011).

Samenhang leergedrag en leerprestatie

Wanneer de correlaties tussen het leergedrag en de leerprestatie worden geanalyseerd wordt er geen enkele statistisch significante correlatie gevonden. Wanneer een grens van $r = .70$ wordt aangehouden voor een voldoende grote correlatie, wordt tevens

geen voldoende correlatie gevonden voor één van de vier categorieën. Ditzelfde geldt wanneer de vier categorieën worden samengevoegd tot de dichotome schalen. Voor geen van beide schalen kan een significante of voldoende correlatie worden gevonden. Ook wanneer onderscheid wordt gemaakt in sekse is er voor geen van de vier categorieën en voor geen van de dichotome schalen sprake van een significante of voldoende correlatie. Deze resultaten waren niet verwacht, aangezien de samenhang tussen leergedrag en leerprestatie al meerdere malen is aangetoond in onderzoek (Arnold et al., 2005; Reid et al., 2004; Trout et al., 2003).

Hoewel er geen statistisch significante samenhang is gevonden, is er wel sprake van een trend in de data. Ook hier kunnen betekenisvolle resultaten uit worden gedestilleerd, hoewel deze niet significant zijn. Zo bestaat er een consistent beeld waarin meer actief leergedrag met een lager cijfer voor zowel rekenen als Nederlands samenhangt. Dit is opvallend, aangezien men zou verwachten dat wanneer een leerling meer actief leergedrag vertoont, deze ook een betere prestatie levert, zoals dit ook is gebleken uit eerder onderzoek (Arnold et al., 2005; Morgan et al., 2008; Reid et al., 2004; Trout et al., 2003). Toch is deze uitkomst te verklaren vanuit het feit dat het vertonen van actief leergedrag niet hoeft te betekenen dat een leerling de leerstof ook daadwerkelijk begrijpt. Zo vallen activiteiten als het stellen van vragen en het beantwoorden van vragen ook binnen de categorie actief leergedrag. Hierbij wordt echter niet beoordeeld of de leerling de leerstof na het stellen van zijn vraag wel begrijpt, of dat hij zelf een adequaat antwoord geeft op de vraag van de leerkracht. Dat een leerling actief leergedrag laat zien hoeft daardoor niet te betekenen dat een leerling de leerstof begrijpt en kan in het geval van het stellen van vragen zelfs een actief teken zijn dat de leerstof juist niet wordt begrepen. Bij deze specifieke doelgroep is het ook mogelijk dat de beperkte taalbeheersing ervoor zorgt dat leerlingen erg lang bezig zijn met een opdracht. Zij kunnen dan weldegelijk veel actief leergedrag vertonen (zoals het maken van opdrachten), maar alsnog weinig van de stof begrijpen omdat zij erg lang over een opgave doen en hem door een beperkt taalbegrip niet begrijpen. In dat geval is het vertonen van het actieve leergedrag niet voldoende om de informatie op te kunnen pakken, wat een negatieve invloed heeft op de leerprestatie.

Met betrekking tot passief leergedrag wordt wel een stabiele (hoewel niet-significante) positieve trend gevonden met de leerprestatie voor zowel Rekenen als Nederlands. Het valt binnen de verwachting dat wanneer een leerling luistert naar een

uitleg of tekst met betrekking tot het onderwerp leest, hij kennis opdoet die hij kan gebruiken voor latere toetsen en daardoor een hoger cijfer kan halen voor die toets of dat vak. De categorie off-task gedrag laat echter geen consistent beeld zien, zoals actief leergedrag en passief leergedrag dat wel deden. De correlaties zijn afwisselend positief en negatief wanneer wordt gekeken naar het vak en de sekse apart. Er lijkt daardoor geen enkele trend in de resultaten voor deze categorie te zijn. Dit is mogelijk te verklaren door het feit dat deze categorie niet betrouwbaar geobserveerd kon worden.

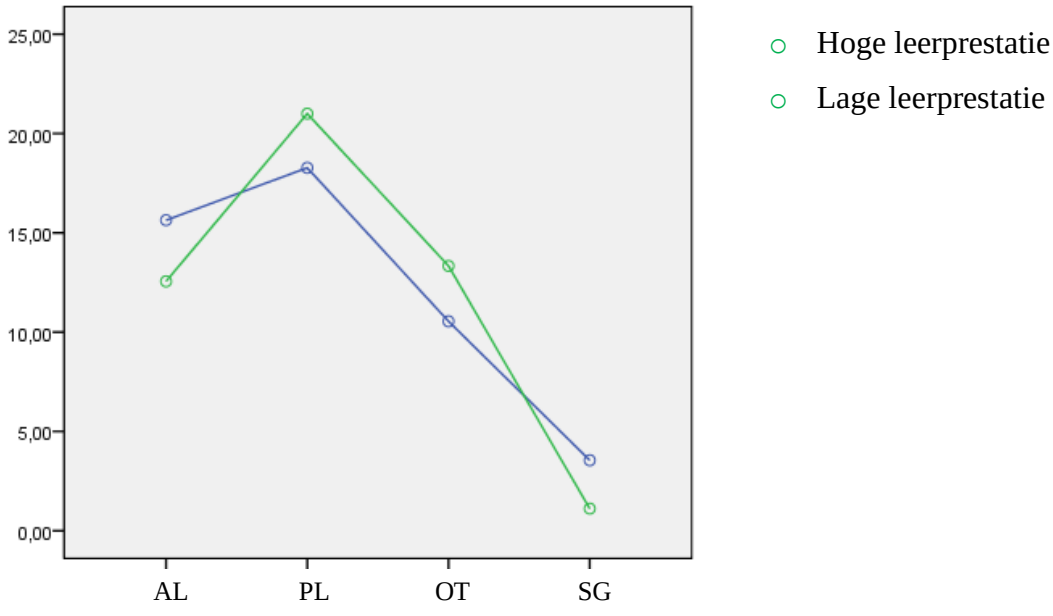
De categorie storend gedrag laat wel een trend zien. Zowel jongens als meisjes halen een lager cijfer voor Rekenen bij het vertonen van meer storend gedrag. Dit is te verwachten, aangezien een leerling bij het vertonen van storend gedrag geen notie kan nemen van de leerstof. Opvallend is echter dat de trend in de data voor het vak Nederlands volledig andersom is. Dit betekent dat meer storend gedrag bij zowel jongens als meisjes samenhangt met een hoger cijfer. Hoewel de samenhang niet significant is, is dit toch een opvallende trend.

Wanneer naar trends wordt gezocht in de dichotome schalen blijkt dat voor het vak Rekenen een hoger cijfer samenhangt met minder actief gedrag en meer passief gedrag. Voor het vak Nederlands is geen duidelijke trend te ontdekken. Ook blijkt, zoals verwacht, meer positief gedrag samen te hangen met een hoger cijfer voor Rekenen terwijl meer negatief hier negatief mee correleert. Ook voor deze dichotome schaal is het beeld minder duidelijk voor het vak Nederlands.

Aangezien de onderzoekers het opvallend vonden dat er geen samenhang kon worden gevonden tussen het leergedrag en de leerprestatie is onderzocht of de samenhang bij deze leerlingen wellicht indirect aanwezig was. Op de scholen wordt door de mentor naast normale cijfers voor de leerprestatie ook een mentorcijfer gegeven. Dit cijfer weerspiegelt de impressie die de mentor van de leerling heeft. Hierbij worden het leergedrag en de leerprestatie van de leerling beoordeeld. Het is mogelijk dat het mentorcijfer een mediërende factor is tussen het leergedrag en de leerprestatie. Als dit het geval is, dan zou het mentorcijfer met zowel de leerprestatie als het leergedrag voor beide vakken moeten samenhangen. Om deze hypothese te toetsen is een Pearson’s correlatietoets gedaan met daarin het mentorcijfer opgenomen. Uit de toets bleek dat er echter ook maar weinig samenhang bestond tussen het mentorcijfer en de leerprestatie (taal én rekenen) en/of het leergedrag van de leerlingen. Hoewel een significante correlatie met het rekencijfer en het aan-taak

gedrag van de leerlingen kon worden gevonden, waren de andere correlaties allemaal klein of niet substantieel. Hiermee wordt de hypothese dat het mentorcijfer een mediërende factor is voor de samenhang tussen leergedrag en leerprestatie verworpen.

Een mogelijke verklaring voor de lage samenhang is dat de grote variatie in het cijfer het algehele beeld vertroebelt. Om deze reden is het gemiddelde van de cijfers voor ieder vak berekend en is de schaal in twee delen opgebroken op het punt van het gemiddelde. Op deze manier ontstond een dichotome schaal. Er is met behulp van een repeated measures ANOVA onderzocht of er met deze dichotome schaal wel een samenhang tussen leergedrag en leerprestatie bestaat. Met deze toets wordt onderzocht of de leerprestatie verschilt tussen de groepen van het leergedrag. Mauchly's test voor sfericiteit geeft aan dat de assumptie van sfericiteit geschonden is ($\chi^2(2) = 15{,}936$, $p = .007$). Dit betekent dat de variantie van de verschillende waarden binnen de verschillende categorieën niet gelijk is. Om deze reden wordt een Greenhouse-Geisser correctie toegepast, wat betekent dat er een aanpassing wordt gemaakt in het aantal vrijheidsgraden, waardoor alsnog aan de aanname van sfericiteit kan worden voldaan. Uit de repeated measures ANOVA blijkt dat de gemiddelde leerprestatie niet significant verschilt voor de verschillende categorieën ($F(2.161, 38.892) = 0.775$, $p < .477$). Dit betekent dat er ook voor deze dichotome schaal voor de leerprestatie geen samenhang bestaat tussen het leergedrag en de leerprestatie van de leerlingen. Figuur 1 laat zien dat leerlingen met zowel een hoge als een lage leerprestatie eenzelfde patroon laten zien in hun leergedrag.



Figuur 1. Samenhang leerprestatie en leergedrag.

Een andere mogelijke verklaring is dat er geen samenhang tussen het leergedrag en de leerprestatie bestaat, omdat de leerlingen van slechts twee specifieke schooltypen afkomstig zijn. Voor het onderwijs voor nieuwkomers is dit te verklaren vanuit de mogelijkheid dat voor deze jongeren nog een beeld moet worden gevormd over wat hun cognitieve mogelijkheden zijn. Het is mogelijk dat deze leerlingen tot meer in staat zijn dan het onderwijs waar zij vooralsnog in zijn geplaatst toelaat. Dit zou kunnen leiden tot verveling, waardoor een leerling weinig positief gedrag laat zien maar toch hoge cijfers haalt, een fenomeen wat ook wordt gezien bij hoogbegaafde kinderen (Gur, 2011; Porter, 1999). Voor de kinderen uit het praktijkonderwijs is bovendien een lage intelligentie of een leerachterstand niet de enige mogelijke indicatie voor deze vorm van onderwijs. Een andere mogelijke indicatie is sociaal-emotionele problematiek (RVC-VO, 2012). Dit betekent dat de focus van de problematiek niet bij alle kinderen cognitief van aard is. Het is daarom mogelijk dat deze kinderen wel goed presteren, maar vanuit hun sociaal-emotionele problematiek moeite hebben om hun leergedrag te controleren.

Conclusie

Met betrekking tot de eerste onderzoeksvraag is gebleken dat het ontworpen instrument een voldoende interbeoordelaarsbetrouwbaarheid heeft als geheel, maar dat de categorie off-task gedrag niet betrouwbaar wordt geobserveerd met dit instrument. Tevens is er sprake van een hoge test-hertestbetrouwbaarheid, wanneer de leerlingen met de meest extreme verschillen in score niet worden meegenomen in de toets. Ook is gebleken dat passief gedrag en positief gedrag relatief betrouwbaarder kunnen worden gescoord dan actief gedrag en negatief gedrag.

In dit onderzoek is geen samenhang tussen het leergedrag en de leerprestatie geconstateerd. Bij verschillende leerprestaties laten leerlingen eenzelfde patroon in hun leergedrag zien, waarbij passief leergedrag het meeste voorkomt en storend gedrag het minste. Wanneer naar trends in de data wordt gekeken valt voornamelijk op dat actief gedrag (niet significant) negatief samenhangt met de leerprestatie en dat storend gedrag bij rekenlessen (niet significant) negatief samenhangt met de leerprestatie, terwijl dit bij het vak Nederlands een (niet significante) positieve samenhang is.

Beperkingen onderzoek

Zoals elk wetenschappelijk onderzoek ging dit onderzoek gepaard met beperkingen. Een eerste beperking is het feit dat er gebruik is gemaakt van correlationeel onderzoek. Dit betekent namelijk dat de resultaten niet kunnen worden besproken in termen van causale relaties. Er kan alleen worden geconcludeerd dat er sprake is van een samenhang, maar niet welke richting deze op is of dat hij wellicht bidirectioneel van aard is. Aangezien onderzoek naar dit verband nog erg weinig aandacht heeft gekregen, is dit onderzoek toch waardevol aangezien het een exploratie is van de eventuele samenhang tussen leergedrag en leerprestatie.

Een tweede beperking bestaat erin dat het ontwikkelen van een nieuw instrument veel onzekerheden met zich meebrengt, die later in het onderzoek pas daadwerkelijk hun uitwerking hebben. Zo is in de training geen aandacht besteed aan het beoordelen van het lestype, waarbij onderscheid gemaakt werd in instructiemomenten en in zelfstandig werken. Het lijkt aannemelijk dat dit invloed heeft op het leergedrag. Shapiro (2004) veronderstelt namelijk dat het mogelijk is dat een verschil in de hoeveelheid teacher-directed instruction (TDI) variatie in leergedrag kan verklaren. Wanneer er sprake is van veel TDI in een lestype is de leerkracht meer actief betrokken bij het lesgeven. Dit is bijvoorbeeld het geval wanneer een leerkracht instructie geeft. Wanneer een leerling echter zelfstandig werkt kan worden verwacht dat hij of zij minder actieve betrokkenheid van de leerkracht krijgt, tenzij hierom wordt gevraagd. Er kan worden verwacht dat de leerling meer actief leergedrag vertoont tijdens instructie, aangezien de aandacht van de leerling in deze situatie sterk wordt gestuurd door de leerkracht en er in grote mate controle is op het leergedrag van de leerling (Shapiro, 2004). De trainingsvideo was echter niet geschikt om het lestype te beoordelen, waardoor het gevolg was dat de beoordelaars niet geoefend waren in het beoordelen van het lestype. Het gevolg is geweest dat het lestype inconsequent is gecodeerd en om die reden is deze variabele uit dit onderzoek gelaten. Een tweede voorbeeld van de onzekerheden die men tegenkomt bij het ontwikkelen van een nieuw instrument is dat vooraf geen rekening is gehouden met het coderen van momenten dat leerlingen wachten. Ook dit kwam niet in de trainingsvideo voor. Na de eerste observaties werd echter geconstateerd dat het wachten van leerlingen niet als zijnde leergedrag kan worden beoordeeld. Er is op dat moment geconcludeerd dat de observator het beste een ‘w’ (van ‘wachten’) noteert op het observatieformulier. Later bleek echter dat door deze notatie de gehele minuut aan

observatie ongeldig werd en niet voor het onderzoek kon worden gebruikt. Het lijkt bij nader inzien effectiever om niks te noteren op het observatieformulier wanneer een leerling moet wachten en het observeren te hervatten als de leerling klaar is met wachten en weer leergedrag vertoont. Op deze manier kunnen zo veel mogelijk bruikbare data worden verzameld.

Verder is ook sprake van een relatief kleine groep leerlingen waarmee is getoetst. Hoewel voor de toetsing van de test-hertestbetrouwbaarheid een groep van 38 leerlingen is gebruikt, is dit voor het berekenen van de samenhang tussen leergedrag en leerprestatie een groep van slechts 18 leerlingen. De kleine steekproef heeft tot gevolg dat het onderzoek minder zogenaamde *power* (kracht) heeft. De resultaten zijn een afspiegeling van het leergedrag van slechts een kleine groep leerlingen en daardoor hebben eventuele afwijkende gedragingen meer invloed op het gehele beeld en de uiteindelijke resultaten. Een tweede gevolg van een kleine steekproef is het feit dat significante verschillen/effecten moeilijker kunnen worden opgespoord en dat resultaten daarom vaker niet-significant zijn.

Hieraan verwant is de beperking dat de steekproef niet geheel aselect is getrokken, aangezien de leerlingen uit slechts vijf verschillende klassen kwamen. Deze vijf klassen zijn ook afkomstig uit slechts twee scholen met specifieke vormen van onderwijs (praktijkonderwijs en onderwijs voor nieuwkomers). Dit heeft tot gevolg dat voorzichtigheid is geboden wanneer men de resultaten wil generaliseren naar andere groepen leerlingen die andere vormen van onderwijs volgen. Er is namelijk geen zekerheid of de resultaten ook op deze andere groepen leerlingen van toepassing zijn.

Theoretische implicaties en aanbevelingen voor vervolgonderzoek

Met dit onderzoek is een instrument geboden waarmee betrouwbare data van het leergedrag van leerlingen kan worden verkregen, voor het instrument als geheel en voor drie van de vier categorieën. Echter, uit het onderzoek is gebleken dat het off-task gedrag men met behulp van dit instrument niet betrouwbaar kan worden gemeten. Wanneer een indeling in dichotome schalen wordt gemaakt, kunnen actief en passief gedrag betrouwbaar worden gemeten, net zoals positief gedrag. Negatief gedrag kan echter niet betrouwbaar worden gemeten met dit instrument. Het instrument kan voor verder onderzoek binnen dit thema worden gebruikt. Het wordt aanbevolen om observatoren te trainen, voordat zij in de praktijk met dit instrument gaan werken.

Verder biedt dit onderzoek het inzicht dat niet in alle gevallen mag worden aangenomen dat het leergedrag en de leerprestatie een samenhang vertonen. Wellicht is de samenhang tussen leerprestaties en leergedrag voor leerlingen uit het praktijkonderwijs en onderwijs voor nieuwkomers niet of in mindere mate aanwezig. In dit onderzoek is gevonden dat leerlingen uit de tweede klas van het middelbaar onderwijs die deze vormen van onderwijs volgen geen duidelijke samenhang vertonen in hun leergedrag en hun cijfers. Het is daarom van groot belang dat de samenhang die in eerdere studies is gevonden wordt getoetst voor verschillende groepen leerlingen van verschillende leerniveaus en leeftijden om de generaliseerbaarheid van deze resultaten te bewijzen.

Een theoretische aanbeveling voor vervolgonderzoek is tevens dat bij het gebruik van dit instrument ook consistent het lestype wordt bijgehouden tijdens de observatie. Het is zeer goed mogelijk dat er een samenhang is tussen het soort les (instructie, zelfstandig werken of computerles) en het gedrag dat leerlingen laten zien. In voorliggend onderzoek is het helaas niet mogelijk geweest om deze hypothese te toetsen en dit laat de mogelijkheid open voor vervolgonderzoek om zich hier op te richten. Op dit moment levert dit een gat in het kennisbestand op, terwijl de variabele lestype waardevolle informatie kan geven over het leergedrag. Het is hierbij aanbevolen om ook tijdens de training van de observatoren al aandacht hieraan te schenken, waardoor het voor de observatoren een automatisme wordt om naast het leergedrag het lestype te beoordelen.

Hoewel er in dit onderzoek tijdens twee verschillende vakken is geobserveerd (Nederlands en Rekenen) is het mogelijk dat de samenhang met het leergedrag wel bij andere vakken bestaat zoals zaakvakken (als aardrijkskunde of geschiedenis) of creatieve vakken (als muziek of drama). Dit blijft vooralsnog een open vraag en het is een aanbeveling voor vervolgonderzoek om juist ook voor andere soorten vakken de samenhang tussen leergedrag en leerprestatie te exploreren.

Een laatste aanbeveling betreft het onderzoeken van cognities en leergedrag samen. In dit onderzoek is alleen de samenhang tussen het leergedrag en de leerprestatie onderzocht, maar het is interessant om in vervolgonderzoek ook de cognitie te onderzoeken. Het wordt namelijk in het cognitief-gedragsmatige perspectief verondersteld dat zowel de cognitieve mogelijkheden als het leergedrag de leerprestatie beïnvloeden. Vanuit dit perspectief is het daarom interessant om ook de

samenhang tussen leergedrag en de cognitieve mogelijkheden enerzijds en de cognities en de leerprestatie anderzijds te onderzoeken.

Praktische implicaties

De achterliggende motivatie voor het ontwikkelen van een nieuw instrument om het leergedrag te beoordelen, was om scholen hiermee te helpen. Het ontworpen instrument is makkelijk en snel te gebruiken en in te vullen en daardoor zeer gebruiksvriendelijk voor de praktijk. Ook hier dient bij opgemerkt te worden dat het wordt aangeraden om leerkrachten eerst te trainen in het gebruik van dit instrument, zodat de observatie vloeiend verloopt waarbij de leerling niet wordt gestoord in zijn of haar leergedrag.

In dit onderzoek is tevens uitgebreid stilgestaan bij het aantal categorieën dat nodig is om betrouwbaar het leergedrag te scoren. Hierbij is in eerste instantie gebruik gemaakt van vier categorieën, namelijk actief leergedrag, passief leergedrag, off-task gedrag en storend gedrag. Dezelfde betrouwbaarheidstoetsen zijn echter ook uitgevoerd op dichotome schalen, met maar twee categorieën. Deze schalen waren actief tegenover passief gedrag en positief tegenover negatief gedrag. In dit onderzoek lijkt een indeling in een dichotome schaal de meest betrouwbare resultaten op te leveren, aangezien dit meer robuuste schalen oplevert. Het is echter de vraag of dit in de praktijk ook inhoudelijk waardevolle informatie oplevert. Wanneer een leerling bijvoorbeeld veel negatief gedrag laat zien kan het erg waardevol zijn bij het opstellen van een interventie om informatie te hebben over de vraag of een leerling meer dan normaal passief gedrag of storend gedrag laat zien. In het eerste geval zou de problematiek in de concentratie kunnen worden gezocht, terwijl in het tweede geval de problematiek meer sociaal-emotioneel van aard zou kunnen zijn. Wanneer deze informatie niet aanwezig is, kan een meer dan normale hoeveelheid negatief gedrag tot veel verschillende interventies leiden, terwijl een meer gedifferentieerd beeld in vier categorieën meer handreikingen voor de interventie kan geven.

Literatuurlijst

Arnold, E. M., Goldston, D. B., Walsh, A. K., Reboussin, B. A., Daniel S. S., Hickman, E., & Wood, F. B. (2005). Severity of emotional and behavioural problems among poor and typical readers. *Journal of Abnormal Child Psychology, 33,* 205-217.

Bender, W. N., & Wall, M. E. (1994). Social-emotional development of students with learning disabilities: A meta-analysis. *Journal of Learning Disabilities, 23,* 298-305.

Brandon, P. R. (1991). Gender differences in young Asian Americans' educational attainments. *Sex Roles, 25*, 45–61.

Brown-Chidsey, R. (2005). Introduction to problem-solving assessment. In R. Brown-Chidsey (Ed.), *Assessment for Intervention: A problem-solving approach* (pp. 3-9). New York: Guilford Press.

Butler, A. (1990). *Validation of a classroom observation code for behaviour disordered and learning disabled students.* Unpublished master's thesis, University of Utah, Salt Lake City.

Butts, J. A., Snyder, H. N., Finnegan, T. A., Aughenbaugh, A. L., Tierney, N. J., Sullivan, D. P., & Poole, R. S. (1995). *Juvenile court statistics: 1992*. Washington, DC: Office of Juvenile Justice and Delinquency Prevention.

Chafouleas, S. M., Briesch, A. M., Riley-Tilman, T. C., Christ, T. J., Black, A. C., & Kilgus, S. P. (2010a). An investigation of the generzalizability and dependability of direct behavior rating single item scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students. *Journal of School Psychology, 48,* 219-246.

Chafouleas, S. M., Christ, T. J., Riley-Tillman, T. C., Briesch, A. M., & Chanese, J. A.M. (2007). Generalizability and dependability of direct behavior ratings to assess social behavior of preschoolers. *School Psychology Review, 36*, 63−79.

Chafouleas, S. M., McDougal, J. L., Riley-Tillman, T. C., Panahon, C. J., & Hilt, A. M. (2005). What do daily behavior report cards (DBRCs) measure? An initial comparison of DBRCs with direct observation for off-task behavior. *Psychology in the School, 42,* 669-676.

Chafouleas, S. M., Volpe, R. J., Gresham, F. M., & Cook, C. R. (2010b). School-based behavioral assessment within problem-solving models: current status and future directions. *School Psychology Review, 39*, 343-349.

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development and use of Direct Behavior Rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention, 34*, 201−213.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* New Jersey: Lawrence Erlbaum.

Cone, J. D. (1978). The behavioral assessment grid (BAG): A conceptual framework and a taxonomy. *Behavior Therapy, 9*, 882−888.

Cone, J. D. (1988). Psychometric considerations and the multiple models of behavioural assessment. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical approach* (pp. 42-66). New York: Pergamon.

Davidshofer, K. R., Murphy, Charles O. (2005). *Psychological testing : principles and applications*. Upper Saddle River, N.J.: Pearson/Prentice Hall

Deno, S. L. (2005). Problem-solving assessment. In R. Brown-Chidsey (Ed.), *Assessment for Intervention: A problem-solving approach* (pp. 10-42). New York: Guilford Press.

Deno, S. L., Reschly, A. L., Lembke, E. S., Magnusson, D., Callender, S. A., Windram, H., & Stachel. N. (2009). Developing a school-wide progress-monitoring system. *Psychology in the schools, 46,* 44-55.

Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *The Journal of Special Education, 34,* 140-153.

Evans, S. W., & Owens, J. S. (2010). Commentary. Behavioral assessment within problem-solving models: finding relevance and expanding feasibility. *School Psychology Review, 39,* 427-430.

Fredricks, J. A., Blumenfeld, P. C. & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*, 59-109.

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-Intervention: definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice, 18,* 157-171.

Gur, C. (2011). Do gifted children have similar characteristics? An observation of three gifted children. *Procedia Social and Behavioral Sciences, 12,* 493-500.

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement.* Oxford: Taylor & Francis.

Hinshaw, S. P. (1992b). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin, 111,* 127–155.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34,* 507-519.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review, 33,* 258-270.

Hintze, J. M., Volpe, R. J., & Shapiro, E. S. (2002). Best practices in the systematic direct observation of student behavior. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 993−1006). Washington, DC, US: National Association of School Psychologists.

House, A. E., House, B. J., & Campbell, M. B. (1981). Measures of interobserver agreement: calculation formulas and distribution effects. *Journal of Behavioral Assessment, 3,* 37-57.

Johnson, W., McGue, M., & Iacono, W. G. (2005). Disruptive behavior and school grades: genetic and environmental relations in 11-Year-Olds. *Journal of Educational Psychology, 97,* 391-405.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

Kegel, C. A. T., Bus, A. G., & van IJzendoorn, M. H. (2011). Differential susceptibility in early literacy instruction through computer games: the role of the dopamine D4 receptor gene (DRD4). *Mind, Brain and Education, 5,* 71-78.

Kievit, T., Tak, J. A., & Bosch, J. D. (2009). *Handboek psychodiagnostiek voor de hulpverlening aan kinderen.* Utrecht: De Tijdstroom.

Kobak, K. A., Brown, B., Sharp, I., Levy-Mack, H., Wells, K., Ockun, F., & Williams, J. B. W. (2009). Sources of unreliability in depression ratings. *Journal of Clinical Psychopharmacology, 29,* 82-85.

Landgren, M., Kjellman, B., & Gillberg, C. (2003). A school for all kinds of minds: The impact of neuropsychiatric disorders, gender and ethnicity on school-related tasks administered to 9-10 year old children. *European Child and Adolescent Psychiatry*, *12*, 162-171.

Leary, M. R. (2008). *Introduction to behavioural research methods.* Boston: Pearson Education.

Lee, J. C., & Stacey, J. (2001). More than 'model minorities' or 'delinquents': A look at Hmong American high school students. *Harvard Educational Review, 71*, 505–528.

Lentz, F. E. (1988). Direct observation and measurement of academic skills: A conceptual review. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Conceptual foundations and practical applications* (pp. 76-120). New York: Guilford Press.

Leung, C., Lo, S. K., & Leung, S. S. L. (2012). Validation of a questionnaire on behaviour academic competence among Chinese preschool children. *Research in Developmental Disabilities, 33*, 1581-1593.

Marchant, G. J., Paulson, S. E., & Rothlisberg, B. A. (2001). Relations of middle school students' perceptions of family and school contexts with academic achievement. *Psychology in the Schools, 38*, 505–519.

Martin, G., & Pear, J. (2011). *Behavior modification: What it is and how to do it.* Boston: Pearson Education.

Maughan, B., Pickles, A., Hagell, A., Rutter, M., & Yule, W. (1996). Reading problems and antisocial behaviour: Developmental trends in comorbidity. *Journal of Child Psychology and Psychiatry, 37*, 405–518.

McCall, R. B., Evahn, C., & Kratzer, L. (1992). *High school underachievers: What do they achieve as adults*. Newbury Park: Sage.

McDermott, P. A., Goldberg, M. M., Watkins, M. W., Stanley, J. L., & Glutting, J. J. (2006). A nationwide epidemiologic modeling study of LD: Risk, protection, and unintended impact. *Journal of Learning Disabilities*, *39*, 230-251.

Miles, S. B., & Stipek, D. (2006). Contemporaneous and longitudinal associations between social behavior and literacy achievement in a sample of low-income elementary school children. *Child Development, 77*, 103-117.

Morgan, P. L., Farkas, G., Tufis, P. A., & Sperling, R. A. (2008). Are reading and behaviour problems risk factors for each other? *Journal of Learning Disabilities, 41*, 417-436.

National Research Council & Institute of Medicine. (2004). *Engaging schools: Fostering high school students' motivation to learn.* Washington, DC: National Academy Press.

Nelson, J. R., Benner, G. J., Lane, K., & Smith, B. W. (2004). Academic achievement of K–12 students with emotional and behavioral disorders. *Exceptional Children, 71,* 59–73.

Pelham, W. E., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*, 449−476.

Porter, L. (1999). *Gifted young children.* Buckingham: Open University Press.

Reid, R., Gonzalez, J. E., Nordness, P. D., Trout, A. & Epstein, M. H. (2004). A meta-analysis of the academic status of students with emotional/behavioural disturbance. *The Journal of Special Education, 38*, 130-143.

Rhee, S. H., Waldman, I. D., Hay, D. A., & Levy, F. (2001). Aetiology of the sex difference in the prevalence of *DSM–III–R* ADHD: A comparison of two models. In F. Levy & D. A. Hay (Eds.), *Attention, genes, and attention deficit hyperactivity disorder* (pp. 139–156). Philadelphia: Psychology Press.

Riley-Tillman, T. C., Chafouleas, S. M., & Briesch, A. M. (2007). A school practitioner's guide to using daily behavior report cards to monitor student behavior. *Psychology in the Schools, 44*, 77-89.

Riley-Tillman, T. C., Chafouleas, S. M., Briesch, A. M., & Eckert, T. L. (2008). Daily behaviour report cards and systematic direct observation: An investigation of the acceptability, reported training and use, and decision reliability among school psychologists. *Journal of Behavioral Education, 17,* 313-327.

Riley-tillman, T. C., Christ, T. J., Chafouleas, S. M., Boice-Mallach, C. H., & Briesch, A. (2011). The impact of observation duration on the accuracy of data obtained from Direct Behavior Rating (DBR). *Journal of Positive Behaviour Interventions, 13,* 119-128.

Riley-Tillman, T. C., Methe, S. A., & Weegar, K. (2009). Examining the use of direct behavior rating on formative assessment of class-wide engagement: a case study. *Assessment for effective intervention, 34*, 224-230.

Roeser, R. W., van der Wolf, K., & Strobel, K. R. (2001). On the relation between social–emotional and school functioning during early adolescence preliminary findings from Dutch and American samples. *Journal of School Psychology, 39*, 111-139.

Rong, X. L., & Brown, F. (2001). The effects of immigrant generation and ethnicity on educational attainment among young African and Caribbean Blacks in the United States. *Harvard Educational Review 70*, 537–565.

Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater and test–retest reliability of continuous measurements. *Statistics in Medicine, 21,* 3431-3446.

Salvia, J., & Ysseldyke, J. E. (2004). *Assessment (9th ed.).* Boston: Houghton Mifflin.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations.* San Diego: Jerome M Sattler Publisher Inc.

Shafrat, W. (2007). Arbeidskundig onderzoek. *Werknemer in opleiding.* ‘s Hertogenbosch.

Shapiro, E. S. (2004). *Academic skills problems: Direct assessment and intervention (3rd ed.).* NY: The Guilford Press.

Shapiro, E. S. & Clemens, N. H. (2005). Conducting systematic direct classroom observations to define school-related problems. In R. Brown-Chidsey (Ed.), *Assessment for Intervention: A problem-solving approach* (pp. 175-199). New York: Guilford Press.

Stanovich, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*, 360-407.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: *Review of research. Psychology in the Schools, 42,* 795 – 819.

Suárez-Orozco, C., Gaytán, F. X., Bang, H., Pakes, J., O’Connor, E., & Rhodes, J. (2010). Academic trajectories of newcomer immigrant youth. *Developmental Psychology, 46,* 602–618.

Tilly, W. D. (2008). The evolution of school psychology to science- based practice: Problem solving and the three-tiered model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 17–36). Bethesda, MD: National Association of School Psychologists.

Thorndike, R. (1985). Testing the test: Reliability. *Journal of Counseling and Development, 63,* 528-530.

Topf, M. (1986). Three estimates of interrater reliability for nominal data. *Nursing Research, 35,* 253-255

Trout, A. L., Nordness, P. D., Pierce, C. D., & Epstein, M. H. (2003). Research on the academic status of children with emotional and behavioural disorders: A review of the literature from 1961 to 2000. *Journal of Emotional and Behavioural Disorders, 11,* 198-210.

Watson, T. S., & Steege, M. W. (2003). *Conducting school-based functional behavioral assessments: A practitioner's guide.* New York: Guilford Press.

Williams, S., & McGee, R. (1994). Reading attainment and juvenile delinquency. *Journal of Child Psychology and Psychiatry, 35,* 441–459.

Wilson, M. S., & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review, 25,* 9–23.

Xiong, Z., Eliason, P. A., Detzner, D. F., & Cleveland, M. J. (2005). Southeast Asian immigrants' perceptions of good adolescents and good parents. *Journal of Psychology: Interdisciplinary and Applied, 139,* 159–175.

Bijlage 1:

Operationalisatie categorieën (BOSS / Shapiro (2004) en Hintze & Matthews (2004))

Actief leergedrag

Definitie: De leerling is mondeling, schriftelijk of motorisch aan het reageren op vragen van de leraar of schriftelijk materiaal

- beantwoorden vraag van leerkracht of leermateriaal
- reageren op instructie leerkrachtschrijven/rekenen/wijzen
- vinger/hand opsteken
- vraag stellen
- hardop (voor)lezen
- vraag beantwoorden
- praten met anderen (medeleerlingen of leerkracht) over leerstof
- bladeren door een boek of schrift om iets te zoeken wat gerelateerd is aan de leerstof

Aandachtspunten:

- Het actieve gedeelte zit hem in de meeste gevallen in het feit dat de leerling de aandacht op zichzelf vestigt door verbale of non-verbale taal.

Passief leergedrag

Definitie: De aandacht van de leerling is gericht op de taak zoals deze gedefinieerd is door de leraar. De ogen van de leerling zijn gericht op de huidige taak (dat wil zeggen: de ogen van de leerling zijn gericht op de leraar als deze instructie aan het geven is, en naar het leermateriaal als de leerling zelfstandig werkt)

- stil lezen
- luisteren naar de leerkracht of medeleerling (die antwoord geeft op een vraag)
- kijken naar bord of leerkracht tijdens instructie
- kijken naar leermaterialen
- luisteren naar een gesprek van medeleerlingen over de leerstof

Aandachtspunten:

- Het grote verschil met ‘off-task’ is dat de leerling bij dit gedrag wel een geïnteresseerde uitdrukking toont voor de lesstof.
- Het kijken naar de leerkracht of leermaterialen kan passief leergedrag zijn, maar het is ook mogelijk dat de leerling hierbij een dromerige blik heeft. In dit laatste geval wordt de categorie ‘off-task’ gescoord.

Off-task

Definitie: De aandacht van de leerling is niet gericht op de taak zoals deze gedefineerd is door de leraar. De ogen van de leerling zijn niet gericht op de huidige taak.

- niet luisteren naar leerkracht
- geen reactie op instructie van leerkracht
- naar buiten staren / voor zich uit staren
- met ogen dicht zitten
- doelloos bladeren door een boek of schrift
- lezen van irrelevante informatie

- luisteren naar een gesprek van peers dat niet over de leerstof gaat
- plukken aan kleding (of ander materiaal)
- niet gericht op het leermateriaal
- tekenen of schrijven ongerelateerd aan de leerstof

Aandachtspunten:

- Het moet duidelijk zijn dat de leerling niet bezig is met datgene wat de leerkracht hem/haar heeft opgedragen. De leerling stoort hier echter niemand mee.
- Let bij het luisteren naar en spreken met medeleerlingen op of dit gaat over de leerstof. Als het niet verstaanbaar is, dient dit opgemaakt te worden uit de non-verbale gedragingen (wijzen naar de boeken / in de boeken gedoken vs. lachen / achterstevoren zitten bijvoorbeeld).

Storend gedrag

Definitie: Elk gedrag, door de leerling veroorzaakt, dat inbreuk maakt op de leeromgeving van zichzelf of anderen.

- door de klas schreeuwen
- herrie maken wanneer dit niet van toepassing is (vocaal of met instrumenten/objecten)
- praten voor de beurt of wanneer niet door leerkracht vereist
- niet op zijn plaats (wanneer wel vereist door leerkracht)
- lichamelijk contact met de leerkracht, andere leerlingen of hun bezit (Slaan/porren/duwen/trekken/schoppen/breken/afpakken/scheuren/beetpakken)
- praten met anderen over dingen die niet met de leerstof te maken hebben
- laten zien van dingen die niet met de leerstof te maken hebben aan anderen
- verbaal agressief gedrag
- spelen met de pen (of ander materiaal)
- knippen met vingers
- slaan op tafel
- talking back
- arguing
- task refusal
- social rudeness

Aandachtspunten:

- Het storende gedrag dient niet binnen de klassensituatie vereist te zijn. Als de leerkracht bijvoorbeeld samen met de leerlingen muziek maakt of hen vraagt hard te schreeuwen, dan is dit uiteraard geen storend gedrag. Het gedrag dient daadwerkelijk storend te zijn voor anderen (leerkracht of medeleerlingen).