# De bruikbaarheid van observatie-instrumenten als toetsingsmethode in het kleuteronderwijs


| | |
|---|---|
| Naam: | Leoni Janssen |
| Studentnummer: | S4900987 |
| Datum: | 16 Juni 2023 |
| Cursus: | Masterthese Orthopedagogiek PAMA5166 |
| Faculteit: | Gedrags- en Maatschappijwetenschappen, Rijksuniversiteit Groningen |
| Eerste beoordelaar: | Dr. N. Frans |
| Tweede beoordelaar: | Prof. Dr. T. Kretschmer |
| Aantal woorden: | 7979 |

## Samenvatting

Op politiek en wetenschappelijk niveau is er groeiende kritiek ontstaan op het gebruik van gestandaardiseerde kleutertoetsen. In Nederland moeten scholen daarom vanaf 2022 observatie-instrumenten gebruiken bij het volgen van de ontwikkeling van kleuters. Hoewel er weinig bekend is over de percepties van leerkrachten op deze nieuwe vorm van toetsing, bepalen die percepties wel hoe toetsen en toetsresultaten gebruikt worden. Het doel van dit onderzoek was om de percepties van leerkrachten over de bruikbaarheid van het observatie-instrument 'Kleuter in beeld' van Cito te onderzoeken. Dit nieuwe toetsingsinstrument werd vergeleken met de oude gestandaardiseerde Kleutertoetsen van Cito. Brown's CoA-III-A vragenlijst is gebruikt om de percepties te meten. Daarnaast is onderzocht hoe leerkrachten gebruik maken van de kindroute uit 'Kleuter in beeld'. Het onderzoek toonde aan dat leerkrachten 'Kleuter in beeld' bruikbaar vonden voor dezelfde formatieve en summatieve doelen als de oude Kleutertoetsen. Analyse van de CoA-III-A toonde aan dat leerkrachten beide toetsen bruikbaar vonden om het onderwijs te verbeteren en het niveau en de vaardigheden van leerlingen te beoordelen. Zij zagen de toetsen niet als middel om de effectiviteit en kwaliteit van scholen te beoordelen. De correlatie tussen deze doelen liet zien dat leerkrachten de toetsen voor alle doelen, of voor geen van de doelen bruikbaar vonden. Leerkrachten vonden het observatie-instrument daarnaast iets eerlijker en nauwkeuriger. De kindroute werd afwisselend ingezet. De meeste leerkrachten gebruikten de kindroute echter bij minder dan 10% van de leerlingen. Vervolgonderzoek is nodig om een beter begrip te krijgen over de gevonden verschillen.

## Abstract

The use of standardised assessments for kindergarten students is under ever-growing criticism at both political and scientific level. Consequently, in the Netherlands, schools are required to use observation instruments to monitor the development of young children starting from 2022. However, there is limited knowledge about teachers' perceptions regarding this new form of assessment, despite the fact that these perceptions significantly influence the use of assessments and assessment results. This research aimed to investigate teachers' perceptions of the usability of the observation instrument called 'Kleuter in beeld' developed by Cito, and compare it with the previous standardised Kindergarten assessment, also provided by Cito. The study also examined how teachers utilized the 'child route' of 'Kleuter in beeld'. Brown's CoA-III-A questionnaire was used to measure teachers' perceptions. The research revealed that teachers found 'Kleuter in beeld' useful for the same formative and summative purposes as the old Kindergarten assessments. Analysis of the CoA-III-A indicated that teachers considered both assessments valuable for improving education and student accountability, while not viewing them as means for school accountability. The correlation between these objectives demonstrated that teachers either found the assessments useful for all objectives or not useful for any. Additionally, teachers perceived the observation instrument to be slightly fairer and more accurate. The child route was utilized to varying extents, but the majority of teachers used it for fewer than 10% of their students. Further research is required to gain a better understanding of the observed differences.

## Dankwoord

Deze thesis bewijst dat je met een beetje eigenwijsheid een heel eind kan komen. Het afronden ervan heb ik mede te danken aan de steun en toewijding van de mensen om mij heen. Ik zou niet half zijn wie ik nu ben zonder hen.

Allereerst wil ik mijn thesisbegeleider Niek Frans bedanken voor zijn buitengewone begeleiding en mentorschap. Ik heb het als bijzonder waardevol ervaren dat het proces zo soepel en in goede samenwerking verliep. De aansluiting die jij vond op mijn begeleidingsbehoeftes, het vertrouwen, de ruimte voor frustraties en de gezonde dosis humor hebben me enorm geholpen. Daarnaast heb je me tijdens het onderzoek op verschillende manieren enthousiast gemaakt en uitgedaagd. Dat maakte dat ik plezier heb gehad in het schrijven van de thesis. Je wist op de juiste momenten wat ik nodig had en ik ben dankbaar voor je betrokkenheid.

Isabel, jij kent mij beter dan ik mezelf ken. Je energie en je overweldigende overtuiging van mijn kunnen hebben mij, niet alleen dit jaar, maar in mijn hele studiecarrière door de moeilijke momenten getrokken. Je hebt me geholpen te leren vertrouwen op mezelf. Je vriendschap is van onschatbare waarde voor me. Je bent een inspiratiebron en onze vele gesprekken helpen me om groots te mogen denken en ambitieus te zijn.

Mama en Josien, ik heb altijd het gevoel gehad dat jullie tweehonderd procent achter mij stonden. De onvoorwaardelijke steun en het idee dat ik op elk moment bij jullie terecht kan voor een bord eten, statistische vragen of een knuffel betekenen veel voor me. Mama, jij liet me mijn eigen pad kiezen, meer had ik niet kunnen vragen. Ik denk dat papa trots zou zijn.

Lieve Brunchies (Harlynn, Robin, Tanja, Joost), bedankt voor alle adviezen, dat ik een uitlaatklep had bij jullie en dat we groots kunnen dromen. Aan mijn lieve schoonouders: bedankt dat ik welkom was in jullie huis voor het afschrijven van mijn thesis.

Tenslotte wil ik Johannes, mijn lief bedanken. Je bent alles voor mij. Jij bent de reden dat ik elke keer door kon gaan als het echt niet meer lukte. Je brengt me terug op de grond als alles chaos is. Je oneindige geduld, je nuchterheid en zachtheid is alles wat ik nodig heb. Je bent mijn grootste supporter en ik hoop dat je weet dat ik ook die van jou ben.

En aan alle mensen die ooit tegen me gezegd hebben dat ik iets niet kon. Ik heb jullie het tegendeel bewezen.

## Inleiding

In het regeerakkoord 2017-2021 heeft het kabinet besloten dat er gedurende de kleuterperiode geen gestandaardiseerde pen- en papiertoetsen meer mogen worden afgenomen. Volgens het kabinet is het niet passend bij de ontwikkeling van kleuters om schoolse toetsen te gebruiken. Met name omdat kleuters zich sprongsgewijs en spelenderwijs ontwikkelen, aldus het kabinet. Op een plat vlak toetsen maken doet volgens het kabinet geen recht aan die ontwikkeling. Het volgen van leerlingen aan de hand van een leerlingvolgsysteem (LVS) blijft, ook in de kleuterklassen, echter wel verplicht. Scholen moeten sinds augustus 2022 daarom aan de hand van observatie-instrumenten kleuters volgen (Van Engelshoven, 2018).

Ondanks de groeiende kritiek vanuit de politiek kan toetsing niet zomaar worden geschrapt uit het kleuteronderwijs. Een kerndoel van toetsing is namelijk onderwijssystemen effectiever en efficiënter maken (Tóth & Csapó, 2022). Toetsing speelt om die reden een cruciale rol. Volgens Brown en Harris (2009) heeft toetsing drie belangrijke doelen binnen het onderwijs. In de eerste plaats beschrijven zij dat toetsing het leerproces van leerlingen en de kwaliteit van de lessen kan verbeteren (Improvement). Toetsing helpt het leerproces wanneer het informatie geeft die als feedback gebruikt kan worden om lessen en leeractiviteiten aan te passen. Het tweede doel dat Brown en Harris benoemen is dat toetsing leerlingen individueel verantwoordelijk houdt voor hun leerproces. Dit houdt in dat leerlingen bijvoorbeeld een score of cijfer krijgen op hun werk, dat het niveau van een vaardigheid wordt beoordeeld aan de hand van criteria of dat leerlingen bepaalde diploma's of certificaten krijgen (Brown & Harris, 2009). Dit zien we bijvoorbeeld bij het halen van een zwemdiploma. De niveaus 'A', 'B', en 'C' laten zien in hoeverre een kind vaardig is in het water. Toetsing zegt in dit geval dus iets over het niveau en de vaardigheden van een leerling (Student Accounting). Het laatste doel wat ze beschrijven is dat toetsing gebruikt kan worden om te evalueren hoe goed of slecht leerkrachten, scholen of systemen het doen. Toetsing toont dan de effectiviteit en kwaliteit van leerkrachten en scholen aan. Leerkrachten en scholen leggen door middel van toetsing als het ware verantwoording af voor de kwaliteit van het onderwijs dat zij geven (School Accounting).

De doelen van toetsing die Brown en Harris beschrijven kunnen als summatieve of formatieve doelen worden beschouwd (Newton, 2007). Volgens Taras (2005) leidt elk proces van toetsing uiteindelijk tot een summatief doel: een soort beoordeling, danwel van een school, een leerkracht of een leerling. Toetsing kan echter, in het proces tot die beoordeling, ook een formatief doel dienen. Toetsing wordt formatief op het moment dat informatie uit toetsing wordt gebruikt om lessen aan te passen op de behoeftes van leerlingen (Black et al., 2003). Taras (2005) beschrijft dat toetsing formatief is wanneer het feedback geeft over het 'gat' tussen het niveau van de getoetste vaardigheid en het benodigde niveau. Die feedback moet een indicatie geven over hoe er voor verbetering gezorgd kan worden, zodat het benodigde niveau gehaald kan worden. Bij het afleggen van verantwoording door leerkrachten, scholen of leerlingen hoort een bepaalde beoordeling, waardoor dit doel een summatief karakter laat zien. Het verbeteren van het leerproces en de kwaliteit van lessen laat juist een informerend gebruik van toetsing en toetsresultaten zien, en is daarom formatief (Newton, 2007).

De rol die toetsing exact vervult in het onderwijs, hangt af van de opvattingen van leerkrachten over de doelen van toetsing (Izci, 2016; Vandeyar & Killen, 2007). Die opvattingen hebben invloed op hoe zij toetsing gebruiken, hoe toetsresultaten worden benut,

welke beslissingen eraan worden gekoppeld en hoe het onderwijs wordt vormgegeven (Barnes et al., 2015; Calderhead, 2011; Frans et al., 2020). Wanneer leerkrachten toetsinstrumenten vooral zien als verantwoordingsmiddel, kan dit leiden tot verhoogde prestatiedruk. Deze druk belemmert leerkrachten bij het effectief toepassen van onderwijsstrategieën en heeft daardoor een negatieve invloed op de onderwijspraktijk (Certo, 2006; Condliffe & Plank, 2013; Saeki et al., 2018; Tóth & Csapó, 2022). Anderzijds zullen leerkrachten toetsresultaten eerder gebruiken als informatiebron om hun lessen aan te passen wanneer zij toetsing beschouwen als middel om het onderwijs te verbeteren. De opvattingen van leerkrachten over de doelen van toetsing kunnen dus dienen als leidraad voor hun handelen in de onderwijspraktijk (Barnes et al., 2017; Fives & Buehl, 2012).

Deze percepties van leerkrachten op de doelen van toetsing zijn mogelijk van meerdere factoren afhankelijk. Zo kunnen de opleiding van leerkrachten, persoonlijke ervaringen, schoolcontext, schoolbeleid, toetsingssoorten en verwachtingen van onderwijsautoriteiten allemaal die percepties beïnvloeden (Barnes et al., 2015, 2017; Bonner, 2016; Brown, 2004; Brown & Harris, 2009; Daniels et al., 2014; Fives & Buehl, 2012; Monteiro et al., 2021). Deze invloeden zijn veranderlijk en naarmate een leerkracht bijvoorbeeld nieuwe ervaringen opdoet, of het schoolbeleid verandert, kunnen ook bestaande percepties uitgedaagd worden (Bonner, 2016; Brown, 2004; Izci, 2016).

Hoewel er veel onderzoek is gedaan naar de percepties van leerkrachten ten opzichte van gestandaardiseerde toetsen, is er nog weinig bekend over hun percepties op informele toetsingsmethoden zoals observatie (Barnes et al., 2017; Brown, 2004; Brown et al., 2011; Brown & Harris, 2009; Frans et al., 2017, 2020; Hanes, 2010; Remesal, 2007). Dit is opmerkelijk omdat die percepties van leerkrachten juist afhankelijk zijn van de context (Brown et al., 2011; Daniels et al., 2014). Het gebruik van observatie als toetsingsmethode wordt daarnaast steeds meer toegepast in het kleuteronderwijs en krijgt een belangrijkere rol (Brodie, 2013; Pellegrini, 2001; Thornton-Lang, 2014). De aankondiging van het kabinet dat leerkrachten in het kleuteronderwijs observatie-instrumenten moeten gebruiken om leerlingen te volgen, is hier een illustratie van.

De toegenomen focus op observatie als toetsingsmethode, komt onder andere voort uit de groeiende kritiek op het gebruik van gestandaardiseerde toetsen bij kleuters. Van Dijk en Van Geert (2007) laten in hun onderzoek bijvoorbeeld zien dat de snelheid waarmee de vaardigheid van een kleuter zich ontwikkelt, verschilt. Daarnaast verschilt ook het tijdstip waarop de ontwikkeling van een vaardigheid begint per kind. Dit betekent dat het meten van vaardigheden op een vast moment, zoals bij gestandaardiseerde toetsen gangbaar is, weinig zegt over de individuele ontwikkeling van een kind. In plaats van toetsen met momentopnames, wordt regelmatige monitoring van de vooruitgang als wenselijker beschouwd door (Snowling en collega's (2012) en Van Dijk en Van Geert (2007).

Observatie-instrumenten bieden een kans om flexibeler ingezet te worden dan gestandaardiseerde toetsen. Gestandaardiseerde toetsen zijn vaak gebonden aan vaste meetmomenten, terwijl een observatie-instrument gebruik kan maken van informatie die over een langere periode is vergaard. Het invullen van zo'n observatie-instrument is dan minder snel gebonden aan één vast moment, of kan bijvoorbeeld worden opgedeeld. Die mogelijke flexibiliteit kan volgens Nolan en Highhouse (2014) de attitudes van leerkrachten over toetsen beïnvloeden. Hun onderzoek laat zien dat de mate van autonomie samenhangt met intenties om

bepaalde toetsingsvormen te gebruiken. Wanneer beoordelaars invloed hebben op de uitvoering van toetsen en het gebruik van resultaten, stijgt de intentie om ze actief in te zetten. De zelfdeterminatietheorie onderbouwd dit effect. Het veronderstelt dat de intrinsieke motivatie van mensen onder andere afhankelijk is van de mate van vrijheid die zei ervaren in hun handelen en het maken van keuzes (Deci & Ryan, 2012).

Ondanks dat observatie tot positievere attitudes bij leerkrachten zou kunnen leiden, zijn er ook enkele nadelen verbonden aan observatie als toetsingsvorm. Observatie is, mede door die flexibiliteit, bijvoorbeeld gevoelig voor waarnemersfouten (observer bias). De observator (de leerkracht) focust en hecht hierbij, bewust of onbewust, meer waarde aan informatie die haar overtuigingen bevestigd. Hierdoor ontstaan systematische afwijkingen. Verwachtingen, interpretaties en vooroordelen beïnvloeden dan de toetsresultaten en leiden mogelijk tot vertekening (Sleegers et al., 2019). Zelfs wanneer mensen worden geconfronteerd met informatie die hun overtuigingen weerlegt, kunnen die overtuigingen blijven bestaan (Nickerson, 1998).

Observatie onderscheidt zich van andere toetsingsmethoden doordat de beoordeling onder andere plaatsvindt in de natuurlijke context (bijvoorbeeld klassenactiviteiten), non-verbale signalen meegenomen kunnen worden en het vaak minder tijdsgebonden en flexibeler is (Brodie, 2013; Pellegrini, 2001; Thornton-Lang, 2014). De invloed van deze observatie-specifieke factoren op de percepties van leerkrachten over toetsing zijn tot op heden nauwelijks onderzocht. Tegelijkertijd weten we dat, zoals eerder beschreven, contextuele factoren invloed hebben op die percepties. De percepties van leerkrachten over toetsing en waarom getoetst wordt, beïnvloeden vervolgens hun lessen. Als leerkrachten toetsing vooral zien als middel om individuele behoeftes van leerlingen te herkennen, dan zullen zij eerder differentiëren, gerichte interventies inzetten of instructies aanpassen. Als zij toetsing vooral zien als een middel om de prestaties van scholen en leerkrachten te controleren, dan zullen zij hun lessen juist meer richten op het voorbereiden van leerlingen op de getoetste onderwerpen (Barnes et al., 2015; Frans et al., 2020; Gollub et al., 2002; Prawat, 1992). De ontwikkelingen van het toetsingsbeleid in Nederlandse kleuterklassen, vragen om een beter begrip over óf, en hoe toetsingsmethodes de percepties van leerkrachten beïnvloeden.

**Onderzoek context**

Om te onderzoeken of de percepties van leerkrachten veranderen als observatie wordt ingezet als toetsingsinstrument, worden deze percepties onderzocht bij leerkrachten die het observatie-instrument ‘Kleuter in beeld gebruiken’.

‘Kleuter in beeld’ is een nieuw volgsysteem waarbij observaties van de leerkracht centraal staan (Schouwstra & Vloedgraven, 2020). Het biedt een nieuwe manier van het volgen van de ontwikkeling van kleuters. De observaties worden onderscheiden in indirecte en directe observaties. Bij indirecte observaties worden gevolgen en resultaten van het gedrag, en niet het gedrag zelf, door de leerkracht geobserveerd. Bij directe observaties observeert de leerkracht het gedrag op het moment dat het zich voordoet. Het observatie-instrument vertaalt indirecte observatie naar de leerkrachtroute. Daarin beschrijft de leerkracht, op basis van gestructureerde observaties die ze in de klas heeft gedaan, per vaardigheid in welke mate een kind de vaardigheid beheerst. Dit kan aangegeven worden met behulp van vijf niveaus (<E1, E1, M2, E2, >E2). Er wordt per vaardigheid uitgelegd hoe de niveaus geïnterpreteerd kunnen worden.

De directe observatie wordt vertaald naar de kindroute, die bestaat uit extra opdrachten en activiteiten die de leerkracht samen met de leerling kan uitvoeren. Als de leerkracht twijfels heeft over de vaardigheid van een leerling, of als ze meer informatie wil, kan ze kiezen voor de kindroute.

'Kleuter in beeld' is in samenwerking met het werkveld ontwikkeld en biedt de leerkracht meer regie en flexibiliteit, in vergelijking met de oude kleutertoetsen taal en rekenen (Lansink & Hemker, 2012; Schouwstra & Vloedgraven, 2020). De leerkracht plant bijvoorbeeld zelf de inzet van het instrument op een moment waarop zij de informatie nodig heeft. Hoe vaak de leerkrachtroute wordt ingevuld, bij welke leerlingen en hoe vaak de kindroute wordt ingezet, bepaald de leerkracht tevens zelf. De oude kleutertoetsen werden meestal twee keer per jaar, individueel op de computer of in de groep op papier afgenomen (Lansink & Hemker, 2012; Op den Kamp & Keuning, 2012). Voordat leerkrachten observatie-instrumenten zoals 'Kleuter in beeld' moesten inzetten, werden voornamelijk deze (gestandaardiseerde) kleutertoetsen ingezet om te ontwikkeling te volgen (Lansink & Hemker, 2012; Op den Kamp & Keuning, 2012). De percepties van leerkrachten over de doelen van toetsing zijn in eerder onderzoek belicht aan de hand van deze kleutertoetsen (Frans et al., 2020). Uit dit onderzoek kwam naar voren dat leerkrachten de kleutertoetsen niet uitsluitend geschikt vonden voor het afleggen van verantwoording (door scholen, leerkrachten of leerlingen) of uitsluitend voor de verbetering van het onderwijs. Sommige leerkrachten zagen de toetsen als een positieve bevestiging van hun eigen observaties. Anderen zagen het juist als een negatieve tegenstelling van hun eigen observaties. De opvattingen van leerkrachten over de kleutertoetsen werd beïnvloed door de samenstelling van de klas, het management en de toegeschreven doelen van de toets (Frans et al., 2020). De bevindingen uit dit onderzoek vormen een kader waaruit de percepties van leerkrachten die observatie als toetsing gebruiken worden belicht.

Om inzicht te krijgen op de vraag of de percepties van leerkrachten op het nieuwe observatie-instrument anders is in vergelijking met gestandaardiseerde kleutertoetsen, is de volgende onderzoeksvraag opgesteld:

'*In hoeverre zien leerkrachten het observatie-instrument 'kleuter in beeld' als een bruikbaar instrument om de ontwikkeling van kleuters te in beeld te brengen?'*.

Antwoord op de onderzoeksvraag zal gegeven worden door het beantwoorden van de volgende deelvragen:

(1) In hoeverre zien leerkrachten het observatie-instrument 'Kleuter in beeld' als bruikbaar instrument om informatie krijgen voor de verbetering van onderwijs, in vergelijking met de gestandaardiseerde toetsen voor kleuters?

(2) In welke mate gebruiken leerkrachten de kind-route om meer inzicht te krijgen in de taalontwikkeling van leerlingen?

(3) In hoeverre is er een relatie tussen jaren werkervaring en het gebruik van de kindroute?

# Methode

## Onderzoeksdesign

Er is een kwantitatief correlationeel onderzoek uitgevoerd met een cross-sectioneel design. Er werd onderzocht wat de attitudes van leerkrachten waren over het nieuwe observatie-instrument 'Kleuter in beeld' (Schouwstra & Vloedgraven, 2020) in vergelijking met de oude gestandaardiseerde kleutertoetsen van Cito (Lansink & Hemker, 2012). De data over deze oude gestandaardiseerde kleutertoetsen werd in eerder onderzoek van Frans en collega's (2020) verzameld. Er is gekozen om deze bestaande data te gebruiken, omdat de gestandaardiseerde kleutertoetsen niet meer gebruikt mogen worden in het kleuteronderwijs. Daardoor was het niet mogelijk nieuwe data over de gestandaardiseerde toetsen te verzamelen.

## Onderzoekspopulatie en Steekproef

De doelpopulatie van het onderzoek bestond uit basisschoolleerkrachten die lesgeven in groep 1 en 2 van het primair onderwijs en gebruik maakten toetsingsinstrumenten van Cito. De totale steekproef werd opgedeeld in twee groepen. De eerste groep bestond uit basisschoolleerkrachten die het observatie-instrument 'Kleuter in beeld' gebruikten. Hierbij werd nieuwe data verzameld door middel van een gelegenheidssteekproef. De tweede groep bestond uit basisschoolleerkrachten die gebruik maakten van de oude gestandaardiseerde kleutertoetsen voor taal en rekenen. De data van deze steekproefgroep kwam uit eerder onderzoek (Frans et al., 2020). De steekproef uit dit onderzoek bestond uit 97 participanten, waarvan 63% leerkracht in het kleuteronderwijs waren, 30% intern begeleider en 3% combineerde deze functies. In het kader van het huidige onderzoek werd de data van intern begeleiders uitgesloten.

(Cohen (1988) beschrijft voor een ongepaarde *t*-toets om gemiddelden te vergelijken bij een $\alpha = .05$ en power van 0.8 dat voor een medium effect de steekproef minimaal uit 64 participanten per subgroep moest bestaan. Bij een groot effect uit 28 participanten. De steekproefgroep van leerkrachten die de kleutertoetsen gebruikten bestond uit 63 participanten. Omdat dit onderzoek gebonden was aan de hoeveelheid respondenten uit de bestaande data (Frans et al., 2020), is er gekozen om in ieder geval een vergelijkbaar aantal reacties voor de nieuwe data te verzamelen, wat voldoende power zou moeten geven om een medium verschil ($d = .50$) aan te tonen. Er bestond echter wel onzekerheid op het vinden ten minste een medium effect.

## Procedure

Voorafgaand aan de start van het onderzoek, werd het onderzoeksvoorstel voorgelegd aan de Ethische commissie van Pedagogische- en Onderwijswetenschappen aan de Rijksuniversiteit Groningen. Na de goedkeuring van het onderzoeksvoorstel werd gestart met de werving.

De onderzoeker heeft leerkrachten in het kleuteronderwijs gecontacteerd via haar netwerk. Via een sneeuwbalmethode werd gevraagd of participanten andere leerkrachten kenden voor het onderzoek. Op sociale mediawebsites heeft de onderzoeker tevens kleuterleerkrachten benaderd voor deelname aan het onderzoek. Verder heeft de onderzoeker open data via de instellingsinformatie van Dienst Uitvoering Onderwijs (DUO) gebruikt om

basisscholen te contacteren. Basisgegevens zoals adressen, emailadressen en telefoonnummers zijn hierin opgenomen. Wanneer participanten werden benaderd, werd vermeld dat er voor het onderzoek gezocht werd naar kleuterleerkrachten die het observatie-instrument 'Kleuter in beeld' gebruiken. Leerkrachten konden deelnemen aan het onderzoek door een online maximaal tien minuten durende Qualtrics vragenlijst in te vullen die samen met een informatiebrief werd verstuurd via e-mail.

In de uitnodiging en voorafgaand aan de vragenlijst werden participanten geïnformeerd over het doel en de aard van het onderzoek en de voorwaarden voor deelname. Participanten werden geïnformeerd dat het onderzoek maximaal tien minuten duurde. Daarnaast werden zij geïnformeerd over hoe privacy wordt gewaarborgd en dat deelname anoniem was. Gegevens die herleidbaar zijn, zoals namen en IP-adressen, werden niet opgeslagen. De geanonimiseerde gegevens worden na afronding van het onderzoek voor een duur van maximaal vijf jaar opgeslagen op een beveiligde server van de Rijksuniversiteit Groningen en is voor die duur inzichtelijk voor de onderzoeker, de thesisbegeleider en de directeur onderzoek. De participanten werden geïnformeerd over de vrijwilligheid van deelname en de vrijheid om op ieder moment te stoppen met het invullen van de vragenlijst. Tenslotte werd de participant gevraagd om schriftelijk toestemming te geven voor het gebruik van de data binnen dit onderzoek door het aanvinken van een verklaring.

**Onderzoeksinstrumenten**

De Conceptions of Assessment Abridged Version (CoA- III-A; Brown, 2013) vragenlijst is gebruikt om de attitudes van de participant over het observatie-instrument 'Kleuter in beeld' te meten (Brown, 2013). Deze vragenlijst is tevens gebruikt om data te verzamelen over de oude kleutertoetsen (Frans et al., 2020). Er werd een naar het Nederlands vertaalde versie gebruikt. Het instrument meet de percepties van leerkrachten over vier verschillende doelen van toetsing met vier verschillende subschalen: toetsing verbetert het onderwijs ('Improvement', 12 items), toetsing laat scholen verantwoording afleggen ('School Accountability', 3 items), toetsing houdt leerlingen verantwoordelijk voor hun leerproces ('Student Accountability', 3 items) en toetsing is irrelevant ('Irrelevance', 9 items). Leerkrachten gaven aan de hand van een zespuntsschaal aan in hoeverre ze het eens waren met stellingen over deze doelen (Brown, 2006; Calderhead, 2011). Hoe hoger de score, hoe meer participanten het eens waren met de stelling. Het subonderdeel 'toetsing verbeterd het lerend vermogen' werd niet meegenomen. Uit het onderzoek naar de oude kleutertoetsen blijkt dat deze stellingen voornamelijk werden beantwoordt met 'niet van toepassing' (Frans et al., 2020). Ze lijken niet passend bij de kleuterleeftijd, omdat kleuters bijvoorbeeld nog niet in staat zijn feedback over hun leren te gebruiken. De CoA-III-A gaf antwoord op de eerste deelvraag.

Brown onderzocht met een confirmatieve factoranalyse de psychometrische eigenschappen van de CoA-III-A bij basisschoolleerkrachten in Queenstown ($N$= 692) en Nieuw-Zeeland ($N$= 525). De verworven data uit de CoA-III-A vormden een goede 'fit' met het (multi-)factormodel. Dit factormodel is ontwikkeld door middel van drie modelvormende studies en bestaat uit subschalen die percepties van leerkrachten op toetsing meten (Brown, 2006). De fit van de verworven data en het factormodel liet zien dat het model de structuur en de variabiliteit van de data goed weer kon geven. Met andere woorden: de verworven data uit het onderzoek spreekt de structuur die Brown verondersteld niet tegen en levert bewijsvoering

voor de constructvaliditeit (Brown, 2006). Frans en collega's (2020) voerden daarnaast een exploratieve mokkenschaalanalyse uit waaruit bleek dat de subschaal 'toetsing is irrelevant' een relatief zwakke schaal vormde. 'Toetsing verbeterd onderwijs' en 'toetsing laat scholen verantwoording afleggen' vormden sterkere, maar gecorreleerde, schalen. Bij het beantwoorden van de CoA-III-A is voorafgaand aan participanten gevraagd om expliciet het observatie-instrument 'Kleuter in beeld' in gedachten te nemen.

Voor het beantwoorden van deelvraag twee en drie is uitgevraagd hoe vaak leerkrachten de kindroute het afgelopen kalenderjaar hebben gebruikt en hoeveel leerlingen er toen in de klas zaten. Aan het begin van de vragenlijst werden tenslotte twee vragen geïmplementeerd die controleerden dat de juiste participanten de vragenlijst invulden. Eerst werd gevraagd of de participant leerkracht was van groep 1 of 2, daarna of zij 'Kleuter in beeld' gebruikte. Vervolgens werd algemene informatie over geslacht, leeftijd en jaren werkervaring gevraagd. De antwoordmogelijkheden voor leeftijd waren opgedeeld in de categorieën 'jonger dan 25 jaar', '25 tot 35 jaar', '36 tot 45 jaar', '46 tot 55 jaar', '56 tot 60 jaar' en '61 jaar of ouder'. Hiervoor is gekozen om de anonimiteit van de respondenten te waarborgen.

**Data-Analyse**

Verkregen data zijn in de statistische softwareprogramma's Statistical Product and Service Solutions (SPSS) versie 28 (IBM Corporation, 2020) en R 4.2.2 (R Core Team, 2021) geanalyseerd. Er is antwoord gegeven op de eerste onderzoeksvraag door eerst de itemscores van de CoA-III-A te onderzoeken. De itemtotaalcorrelaties, itemrestcorrelaties en de Cronbach's Alpha zijn berekend in R met het package 'psych' (Revelle, 2021). Hierna is er samen met de thesisbegeleider een confirmatieve Mokkenschaalanalyse uitgevoerd met het R package 'mokken' om de schaalbaarheid van de items te onderzoeken (Van der Ark & Luijten, 2021). Een mokkenschaalanalyse is een statistische techniek die de schaalbaarheid en consistentie van items beoordeelt (Baghaei, 2021; Sijtsma & Molenaar, 2002). De assumptie van monotoniciteit is gecontroleerd met itemresponscurves. Vervolgens zijn de gemiddelde somscores van de CoA-III-A schalen en de scores op individuele items bekeken. Hiervoor zijn telkens de resultaten van de beide steekproeven vergeleken. Belangrijke overeenkomsten en verschillen, zowel op schaal- als itemniveau zijn beschreven. Daarnaast zijn er ongepaarde *t*-toetsen uitgevoerd om te onderzoeken of gevonden verschillen tussen de steekproeven bij de schalen significant waren. Normaliteit is hierbij getoetst met histogrammen. De assumptie van gelijke variantie is gecontroleerd door te bepalen of de standaarddeviatie van de ene groep maximaal twee keer zo groot was als de standaarddeviatie van de andere groep. Er is uitgegaan van een significantiegrens van $\alpha = .05$.

Deelvraag twee is beantwoord met beschrijvende statistiek. Het gebruik van de kindroute is in percentages berekend door het aantal keer dat de kindroute is gebruikt te delen door de klassengrootte en te vermenigvuldigen met 100. Er is gekozen om met percentages te rekenen om te controleren op verschillen in klassengroottes. Deelvraag drie is beantwoord door de Spearman correlatie te berekenen.

Er is een Fisher exact test uitgevoerd om te onderzoeken of er een verband was tussen de gevolgde opleiding en de steekproef. Tenslotte is er een Chi-kwadraat toets uitgevoerd om te onderzoeken of er een significant verschil was tussen de steekproef en leeftijd en een ongepaarde t-toets voor de steekproef en het aantal jaren werkervaring.

# Resultaten

## Problemen in de dataset

Er zijn 156 participanten begonnen met het invullen van de vragenlijst. Er zijn 19 participanten uit de vragenlijst geleid omdat zij aangaven geen leerkracht te zijn van groep 1 en/of 2 in het kleuteronderwijs en 69 omdat zij geen gebruik maakten van 'Kleuter in beeld'. Er zijn 9 participanten uit de dataset verwijderd omdat ze na de algemene vragen niets meer hebben ingevuld. De uiteindelijke dataset over 'Kleuter in beeld' bestaat uit reacties van 62 participanten. In totaal hebben 8 participanten de CoA-III-A (24 vragen) niet volledig ingevuld. Hiervan hebben vier participanten alleen de eerste 6 vragen ingevuld, twee alleen de eerste 12 vragen, één heeft de eerste 14 vragen ingevuld en één participant heeft 22 vragen ingevuld. Deze reacties zijn meegenomen in de verdere analyses. Er waren twee participanten die alleen de vragen over de kindroute en de klassengrootte niet hebben ingevuld.

Vier participanten gaven een opleiding aan die afweek van de antwoordopties. Eén participant gaf aan de Pabo en hbo-pedagogiek te hebben gedaan en één de verkorte PABO. Deze participanten zijn onderverdeeld in de groep PABO, omdat die opleiding het meest relevant was of inhoudelijk hetzelfde. Eén participant gaf aan de PA en KLOS te hebben gedaan. Deze participant is meegenomen in de groep PA, omdat dit de meeste recent gevolgde opleiding is. Eén participant gaf aan de Master jonge kind en gedrag te hebben gedaan en is onderverdeeld in de groep WO, omdat het om een universitaire opleiding gaat.

## Steekproefbeschrijving

De steekproef bestaat uit 125 leerkrachten waarvan 63 leerkrachten de oude kleutertoetsen (KT) gebruikten en 62 leerkrachten 'Kleuter in beeld' (KiB). In Tabel 1 is de leeftijd van leerkrachten in beide steekproeven te zien. De grootste groep leerkrachten was tussen de 45 en 55 jaar oud. De ruime meerderheid van de leerkrachten was tussen de 25 en 55 jaar oud. Er was één leerkracht die deze vraag heeft overgeslagen. De leerkrachten uit de KT steekproef waren iets ouder dan leerkrachten uit de KiB steekproef. Zo was 59% ouder dan 45 ten opzichte van 42% van de leerkrachten die het observatie-instrument gebruiken. Het verschil in leeftijd was niet significant ($\chi^2(5) = 7{,}38$, $p = .194$).

In Tabel 1 is tevens de verdeling van de opleiding per steekproef te zien. In totaal hebben 123 leerkrachten hun opleiding opgegeven en twee leerkrachten hebben niets ingevuld. De meerderheid van de steekproef heeft de PABO gevolgd. Daarbij hadden leerkrachten die het observatie-instrument gebruikten, vaker de PABO als vooropleiding dan leerkrachten die de kleutertoetsen gebruikten. Ongeveer driekwart van de leerkrachten uit de KiB steekproef volgde de PABO (74%) tegenover ongeveer de helft van de KT steekproef (49%). De groep leerkrachten die de KLOS heeft gevolgd was groter in de KT steekproef dan in de KiB steekproef (resp. 46% en 18%). Een kleine groep leerkrachten volgden een andere opleiding zoals de PA of een universitaire opleiding. Uit een Fisher's exact test komt naar voren dat er een significant verschil was in opleiding tussen de steekproeven ($p = .003$).

**Tabel 1**

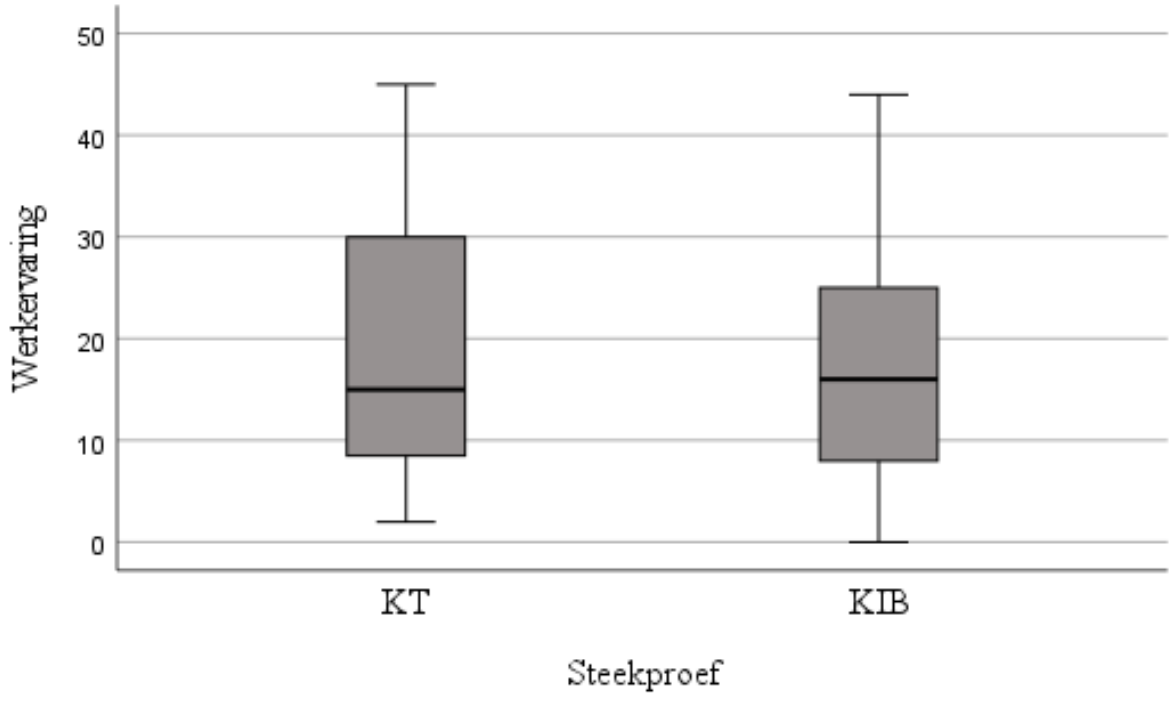
*Verdeling leeftijd en opleiding per steekproef*

| | | | Steekproef | | |
|---|---|---|---|---|---|
| | | | KT* | KIB* | Totaal |
| Leeftijd | < 25 | | 1,9% | 8,1% | 5,2% |
| | 25 - 35 | | 18,9% | 27,4% | 23,5% |
| | 36 - 45 | | 20,8% | 22,6% | 21,7% |
| | 46 - 55 | | 30,2% | 22,6% | 26,1% |
| | 56 - 60 | | 20,8% | 8,1% | 13,9% |
| | > 61 | | 7,5% | 11,3% | 9,6% |
| | | Totaal | 100,0% | 100,0% | 100,0% |
| Opleiding | PABO | | 49,2% | 74,2% | 61,8% |
| | KLOS | | 45,9% | 17,7% | 31,7% |
| | PA | | 1,6% | 4,8% | 3,3% |
| | WO | | 3,3% | 3,2% | 3,3% |
| | | Totaal | 100,0% | 100,0% | 100,0% |

In Figuur 1 is de verdeling van jaren werkervaring per steekproef weergeven in boxplots. De gemiddelde jaren werkervaring van leerkrachten uit KT steekproef was 19,38 jaar (*SD*= 12,3) en uit de KiB steekproef 17,55 jaar (*SD* = 12,1). De mediaan van KT steekproef was 15 en van KiB 16. Er zijn geen grote verschillen in het aantal jaren werkervaring in beide groepen. De verschillen waren daarnaast niet significant (*t* (123) = .841, *p*= .402, tweezijdig, 95% CI [-2,479, 6,144].

**Figuur 1**

*Spreiding aantal jaren werkervaring per steekproef*



Uit de analyse van het gebruik van de kindroute komt naar voren dat de meeste leerkrachten (52%) de observatie-instrumenten voor taal, rekenen, motoriek en sociaal emotioneel gebruikten. Afgerond 37% van de leerkrachten gebruikten alleen de instrumenten voor taal en rekenen. Er waren twee leerkrachten die alleen het instrument taal of rekenen gebruikten en vijf leerkrachten die een combinatie van drie instrumenten gebruikten.

**Percepties op toetsing**

In Tabel 2 zijn de itemtotaalcorrelatie, de itemrestcorrelatie, de schaalbaarheidscoëfficiënt en de scores per item weergegeven. Daarnaast zijn de *H*-coëfficiënten en de Cronbach's alpha per schaal weergegeven. De itemtotaalcorrelaties in de 'Improvement' schaal en de 'School Accountability' schaal waren allemaal gemiddeld tot sterk ($r_{it} > .50$). Ook de itemrestcorrelaties waren voldoende ($r_{ir} > .30$). In de 'Student Accountability' schaal waren tevens de itemtotaalcorrelaties sterk, maar was de itemrestcorrelatie van één item aan de lage kant ($r_{ir}$=.30). In de 'Irrelevance' schaal waren alle itemtotaalcorrelaties gemiddeld tot sterk, behalve van item 29, die een zwakke itemtotaalcorrelatie heeft ($r_{it}$ = .11) en een negatieve itemrestcorrelatie ($r_{ir}$ = *-.065*). Dit item correleerde negatief met vijf van de acht andere items in de schaal en de correlatie met de overige drie items was lager dan .09. De negatieve correlaties kunnen mogelijk verklaard worden doordat leerkrachten bij '*Toetsen heeft weinig impact op het lesgeven*' zowel aan een negatieve als positieve impact kunnen denken. Hun antwoord wordt in dat geval beïnvloed door impact als positief of negatief te beschouwen. Voor het uitvoeren de verdere analyses is item 29 verwijderd uit de schaal.

Uit de mokkenschaalanalyse komt naar voren dat 'School Accountability' een sterke schaal vormt (*H*= .67). De 'Student Accountability' schaal vertoonde een zwakke schaal (*H*= .34), terwijl de 'Improvement' en de 'Irrelevance' schaal weer sterker waren (*H*= .40 en *H*= .44). In de 'Improvement' schaal had item 23 een lage schaalbaarheidscoëfficiënt ($H_i$ = .25). Deze lage waarde geeft aan dat dit item mogelijk niet goed past in de hiërarchische structuur van de schaal. Tenslotte was de schaalbaarheidscoëfficiënt van item 2 in de 'Student Accounting' schaal ook aan de lage kant ($H_i$ = .30). Cronbach's alpha laat zien dat de interne consistentie van de 'Improvement', 'School Accountability' en 'Irrelevance' schalen goed waren (α> .81). Cronbach's alpha van de 'Student Accountability', was aan de lage kant (α= .55). De items in de schaal lijken mogelijk niet consistent te meten wat ze veronderstellen te meten.

De 'Improvement' schaal was positief gecorreleerd aan de 'School Accountability' schaal (*r*=.61*)* en de 'Student Accountability' schaal (*r*=.50*)*. De beide 'Accountability' schalen waren tevens positief aan elkaar gecorreleerd (*r*=.42). De 'Student Accountability' en de 'Irrelevance' schaal waren vrijwel niet aan elkaar gecorreleerd (*r* = -.13). De 'School Accountability' en de 'Improvement' schalen waren beide negatief gecorreleerd aan de "Irrelevance' schaal (*r*= -.35 en *r*= –.59). De items in de schalen voldeden volgens de itemresponscurves aan de assumptie van monotoniciteit.

**Tabel 2**

*Resultaten mokkenschaalanalyse en gemiddelde scores*

| Item | Rit | Itemrest-correlatie | *Hi* | *H* | Cronbach's Alpha | Gemiddelde mate van instemming per item | |
|---|---|---|---|---|---|---|---|
| | | | | | | Kleutertoetsen | Kleuter in beeld |
| **Improvement** | | | | .40 | .84 | | |
| 3 | .66 | .56 | .40 | | | 3,62 | 3,71 |
| 5 | .72 | .60 | .43 | | | 3,68 | 3,69 |
| 6 | .72 | .64 | .45 | | | 3,19 | 3,42 |
| 13 | .71 | .61 | .43 | | | 3,24 | 3,64 |
| 15 | .69 | .59 | .43 | | | 4,05 | 4,12 |
| 16 | .70 | .61 | .43 | | | 2,83 | 2,98 |
| 23 | .50 | .34 | .25 | | | 3,83 | 2,93 |
| 25 | .57 | .43 | .32 | | | 2,89 | 3,05 |
| 26 | .77 | .68 | .48 | | . | 3,13 | 3,45 |
| **School Accountability** | | | | .67 | .84 | | |
| 1 | .86 | .66 | .64 | | | 2,84 | 3,00 |
| 11 | .89 | .75 | .71 | | | 2,65 | 2,62 |
| 21 | .86 | .68 | .65 | | | 2,57 | 2,85 |
| **Student Accountability** | | | | .34 | .55 | | |
| 2 | .68 | .30 | .30 | | | 4,21 | 3,52 |
| 12 | .78 | .44 | .38 | | | 3,48 | 3,38 |
| 22 | .71 | .35 | .32 | | | 3,24 | 3,31 |
| **Irrelevance** | | | | .44 | .84 | | |
| 8 | .78 | .64 | .48 | | | 3,51 | 3,06 |
| 9 | .68 | .59 | .45 | | | 2,70 | 2,64 |
| 10 | .68 | .55 | .44 | | | 4,24 | 4,07 |
| 18 | .76 | .66 | .49 | | | 3,03 | 2,55 |
| 19 | .68 | .58 | .46 | | | 2,23 | 2,40 |
| 20 | .55 | .41 | .32 | | | 3,54 | 4,02 |
| 28 | .67 | .53 | .40 | | | 3,16 | 3,02 |
| *29* | *.11* | *-.065* | NA | | | 2,81 | 3,04 |
| 30 | .73 | .64 | .48 | | | 3,24 | 2,69 |

Op itemniveau waren er een aantal opvallende verschillen te zien tussen de scores per steekproef (zie Tabel 2). Leerkrachten die de kleutertoetsen gebruikten vonden dat die toetsen hogere denkvaardigheden van leerlingen meten (*M*= 3.83*).* Leerkrachten die 'Kleuter in beeld' gebruiken, vonden juist dat het observatie-instrument niet erg geschikt was om dit te meten (*M*=2.93). Hoewel leerkrachten die 'Kleuter in beeld' gebruikte het er enigszins mee eens waren dat toetsing leerlingen in categorieën plaatst (*M*=3.52), waren leerkrachten die de kleutertoetsen gebruikten het hier veel sterker mee eens (*M*=4.21). Toetsing doormiddel van de kleutertoetsen werden daarnaast meer als een onnauwkeurig proces gezien (*M*=3,24) dan toetsing met 'Kleuter in beeld' (*M*= 2.69). Ook zagen we dat leerkrachten die 'Kleuter in beeld'

gebruikten, in plaats van de kleutertoetsen, minder van mening waren dat toetsen leerkrachten dwingt om op een manier les te geven die tegen hun overtuiging in gaat (resp. *M*=3.06 en *M*=3.51). Verder vonden leerkrachten die 'Kleuter in beeld' gebruikten de toetsresultaten uit deze toets betrouwbaarder (resp. *M*=3.45 en *M*=3.12). Bovendien waren zij het er meer mee eens dat toetsen vaststelt wat leerlingen geleerd hebben (resp. *M*=3.64 en *M*=3.24). Tenslotte werd 'rekening houden met fouten en onnauwkeurigheden in het meten' belangrijker geacht door leerkrachten die 'Kleuter in beeld' gebruikten (resp. *M*=4.02 en *M*=3.54).

In Tabel 3 zijn de gemiddelde scores op de schalen per steekproef te zien. Als we de gemiddelden per schaal vergelijken tussen de steekproeven KT en KiB, dan zien we dat de scores veelal met elkaar overeenkomen. De verschillen tussen de gemiddelde scores op de schalen per steekproef is minimaal. Het meeste verschil is te vinden in de gemiddelde scores op de 'Student Accounting' schaal. De standaarddeviatie van elke schaal laat zien dat er voldoende variante zit in de antwoorden van participanten. De spreiding van scores op de schalen was bij de steekproef observatie-instrument telkens iets hoger. Op de stellingen in de 'School accounting' schaal lijken beide steekproeven het enigszins oneens te zijn. Op de overige schalen scoren beide steekproeven telkens 'een beetje eens'.

**Tabel 3**
*Gemiddelde schaalscores per steekproef*

| | Steekproef | | | |
|---|---|---|---|---|
| | KT | | KIB | |
| | Mean | *SD* | Mean | *SD* |
| Improvement | 3,38 | .725 | 3,45 | .939 |
| School Accounting | 2,70 | .968 | 2,84 | 1,163 |
| Student Accounting | 3,65 | .843 | 3,39 | 1,093 |
| Irrelevance | 3,23 | .834 | 3,06 | .935 |

*Noot.* Bij steekproef 'KT' was *N*=62 en bij 'KiB' *N*=55

De gemiddelde score op de 'Improvement' schaal, te vinden in Tabel 3, lag bij de KiB steekproef iets hoger. Leerkrachten die 'Kleuter in beeld' gebruikten waren het er gemiddeld genomen iets sterker mee eens dat toetsing het lerend vermogen van leerlingen en de kwaliteit van lessen verbeterd. De Cohens d score van 0.09 geeft een zeer klein effect aan. Het verschil tussen de gemiddelde scores, zoals weergeven in Tabel 3, was volgens de ongepaarde t-toets niet significant ($t$ (115) = -.456, $p$= .649, tweezijdig, 95% CI [-.376, .235]). Leerkrachten zagen toetsing daarnaast, gemiddeld genomen, niet als een middel om leerkrachten en scholen verantwoording af te laten leggen voor de kwaliteit van het onderwijs. De gemiddelde scores op de 'School Accounting' schaal, laten zien dat leerkrachten die de kleutertoetsen gebruikten iets negatiever keken naar dit beoogde doel. Ook dit waargenomen verschil is niet significant bevonden ($t$ (115) = -.670, $p$= .504, tweezijdig, 95% CI [-.522, .258]). De Cohens d score van -0,12 geeft aan dat er een klein effect is. Leerkrachten waren het er verder gemiddeld genomen een beetje tot gematigd mee eens dat toetsing iets zegt over het niveau en de vaardigheden van

leerlingen (Student Accounting). Leerkrachten die de kleutertoetsen gebruikten, waren iets positiever. Er is een klein tot matig effect waargenomen (Cohens d = 0.26) en ook het gevonden verschil tussen was hierbij niet significant (*t* (115) = 1,401, *p* = 0,164, tweezijdig, 95% CI [-.104, .606]). Als laatste zagen leerkrachten toetsing enigszins als irrelevant voor het onderwijs. De perceptie van de KT steekproef was iets sterker. Zij vonden toetsing iets minder relevant voor het onderwijs dan de KiB steekproef. De resultaten van de t-toets gaven ook hier geen significant verschil aan (*t* (114) = 1.035, p= .303, tweezijdig, 95% CI [-.155, .495]). De Cohens d score van 0.19 geeft een klein effect aan.

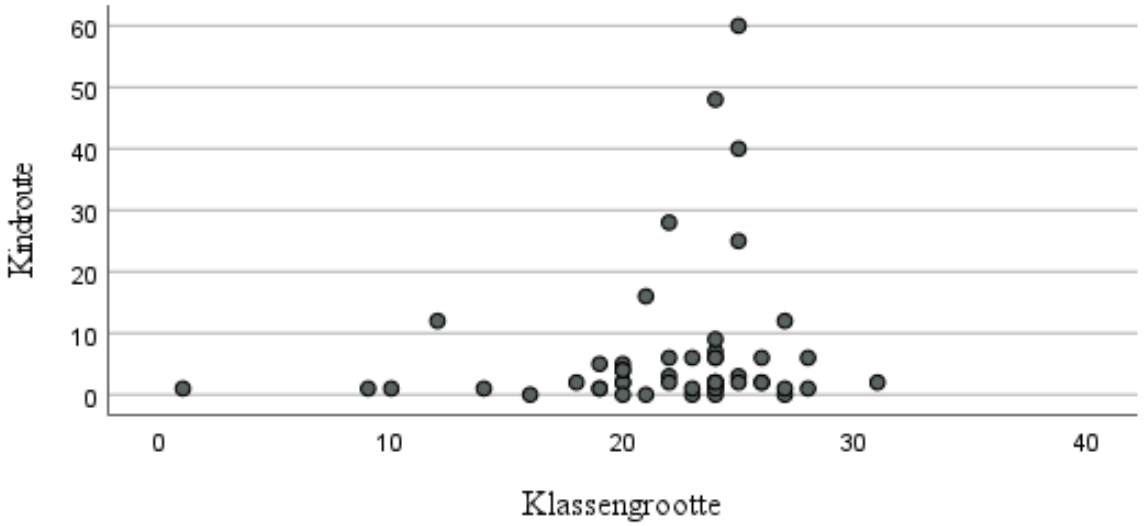
**Gebruik Kindroute**

In Figuur 2 is te zien dat de kindroute het afgelopen kalenderjaar variërend is gebruikt. De meeste leerkrachten gebruikte de kindroute bij <10% van de leerlingen. Dit betekent dat zij meestal een of twee keer de kindroute inzetten (afhankelijk van de klassengrootte). De kindroute werd door acht leerkrachten bij meer dan de 50% van de leerlingen gebruikt. Twee leerkrachten hiervan gebruikte de kindroute meer dan veertig keer. Vier leerkrachten maakten juist geen gebruikt van de kindroute. Er waren drie leerkrachten die de kindroute gemiddeld genomen één keer bij elke leerling in de klas inzetten.

De resultaten van de Spearman correlatie tussen de variabele jaren werkervaring en het gebruikt van de kindroute laten zien dat er een minimale positieve correlatie was tussen beide variabelen. Deze was echter niet significant (*rho*= .017, *p*= .118, eenzijdig). De mate van de inzet van de kindroute kan dus niet verklaard worden vanuit het aantal jaren werkervaring van leerkrachten.

**Figuur 2**

*Spreiding gebruik kindroute*



## Discussie

Een van de doelen van dit onderzoek was onderzoeken in hoeverre leerkrachten het observatie-instrument ‘Kleuter in beeld’ bruikbaar vonden in vergelijking met de oude kleutertoetsen. Uit het onderzoek blijkt dat er weinig verschillen waren tussen de gemiddelde scores op de schalen die de doelen van toetsing meten. Op itemniveau waren echter wel opvallende verschillen zichtbaar. Daarnaast bestonden er verschillende correlaties tussen de schalen die enige duiding behoeven.

Leerkrachten waren, gemiddeld genomen, van mening dat toetsing dient als een middel voor de verbetering van het onderwijs. Toetsing kon volgens leerkrachten daarnaast iets zeggen over het niveau en de vaardigheden van leerlingen. Tegelijkertijd vonden zij niet dat toetsing ingezet moet worden voor het beoordelen van leerkrachten en scholen. Deze percepties waren positief aan elkaar gecorreleerd. Leerkrachten die toetsing inzetbaar vonden voor de verbetering van het onderwijs, vonden toetsing ook inzetbaar om vaardigheden van leerlingen en de kwaliteit van scholen en leerkrachten te meten. De scores op de ‘Improvement’ en de ‘Accountability’ schalen waren, hoewel positief, niet bijzonder hoog. Er lijken ook leerkrachten te zijn die toetsing niet per se als middel zagen voor de verbetering van het onderwijs. De correlatie met de ‘Accountability’ schalen laat zien dat deze leerkrachten tegelijkertijd ook niet positief kijken naar toetsing om de vaardigheden van leerlingen of de kwaliteit en effectiviteit van scholen en leerkrachten te meten. Daarnaast waren er negatieve correlaties tussen de ‘Improvement’ en Accountability’ schalen met de ‘Irrelevance’ schaal. Hoe meer leerkrachten toetsing als doel zagen voor de verbetering van onderwijs, het beoordelen van vaardigheden van leerlingen en beoordeling van scholen; hoe minder zij vonden dat toetsing slecht en onnauwkeurig is en toetsresultaten worden genegeerd. Andersom geldt hetzelfde. Het lijkt erop dat er leerkrachten waren die toetsing relevant vonden voor de bovenstaande doelen, maar er ook een groep was die toetsing niet relevant vond.

De scores op de ‘Improvement’ schaal laten zien dat toetsing door leerkrachten in termen van bruikbaarheid als formatief wordt beschouwd. Tegelijkertijd bevestigen de positieve scores op de ‘Student Accounting’ schaal en de correlatie tussen deze schalen dat leerkrachten toetsing niet uitsluitend als formatief zien, maar ook een summatief. Deze resultaten sluiten aan bij het idee van Taras (2005) dat elk proces van toetsing uiteindelijk op een bepaalde manier tot een summatief doel leidt. Ook als de toetsingsmethode juist beoogd informatie en feedback te geven om het leerproces en lessen te verbeteren. Leerkrachten zagen ‘Kleuter in beeld’ bruikbaar voor dezelfde overkoepelende doelen als leerkrachten die de kleutertoetsen gebruikten. De mate van instemming kwam daarnaast overeen met eerdere onderzoeken waarbij leerkrachten in Nieuw-Zeeland, Queenstown en Canada vergelijkbare scores op de schalen hadden (Brown, 2006, 2013; Daniels et al., 2014). Hoewel er geen significantie verschillen op het schaalniveau te vinden was, zijn er wel belangrijke nuances op itemniveau die relevant zijn.

Leerkrachten vonden dat toetsresultaten te vertrouwen zijn, maar ook dat er rekening moeten worden gehouden met fouten en onnauwkeurigheid in het meten. De perceptie was bij beide stellingen sterker bij leerkrachten die het ‘Kleuter in beeld’ gebruikten. Zij lijken zich ervan bewust te zijn dat er bij observatie sprake kan zijn van vooroordelen en subjectiviteit, zoals Nickerson (1998) beschrijft. Toetsresultaten lijken daarom in de context van het

kleuteronderwijs een nuttige plek te hebben, maar zijn niet allesbepalend. ‘Kleuter in beeld’ werd daarnaast minder gezien als een instrument dat leerkrachten dwingt om les te geven op een manier die tegen hun eigen overtuigingen in gaat. Een verklaring kan zijn dat leerkrachten meer regie hebben over de inzet van dit instrument. Door de keuzevrijheid is het voor leerkrachten beter mogelijk om de inzet van het toetsingsinstrument aan te laten sluiten bij hun eigen overtuigingen. Een andere verklaring is dat leerkrachten observatie zelf passender vinden in het kleuteronderwijs. Leerkrachten die ‘Kleuter in beeld’ gebruikten vonden verder dat toetsing leerlingen minder in categorieën plaatst dan de kleutertoetsen. Bij de kleutertoetsen werden vaardigheidsscores berekend en niveaus toegekend aan de hand van een indeling van A t/m E en I t/m V. In niveau A behoren 25% van de hoogst scorende leerlingen en E 10% van de laagst scorende leerlingen. I stond voor ver boven het gemiddelde en V ver onder het gemiddelde (Lansink & Hemker, 2012). Bij ‘Kleuter in beeld’ zijn deze scores teruggebracht naar niveaus die aangeven of een kind zich op het niveau ‘Midden groep 1’, ‘Eind groep 1’, Midden groep 2’ of ‘Eind groep 2’ (M1, E1, M2, E2) begeeft (Schouwstra & Vloedgraven, 2020). Dit kan verklaren waarom leerkrachten vinden dat ‘Kleuter in beeld’ leerlingen minder in categorieën plaatst. Het kan ook zijn dat de manier van normering dit beïnvloed. ‘Kleuter in beeld’ maakt geen gebruik van normering aan de hand van een landelijke referentiegroep, maar van een absolute normering. De normering is op basis van standaarden die zijn opgesteld door experts. De observaties, opdrachten en activiteiten zijn daardoor niet gebonden aan een bepaalde periode en leerlingen worden niet vergeleken met andere leerlingen in Nederland. Deze verschillen kunnen er tegelijkertijd ook voor gezorgd hebben dat ‘Kleuter in beeld’ als minder oneerlijk werd gezien door leerkrachten. Verder was er een verschil te zien tussen de mate waarin leerkrachten vonden dat de verschillende toetsen hogere denkvaardigheden van leerlingen toetsten. Leerkrachten die ‘Kleuter in beeld’ gebruikten vonden niet dat het observatie-instrument hogere denkvaardigheden meet. Leerkrachten die de kleutertoetsen gebruikten wel. Een verklaring voor het verschil kan liggen in de inhoud van de toetsen. Deze bovenstaande verschillen illustreren de relevantie van de contextuele invloeden zoals Fives en Buehl (2012) beschrijven.

Een tweede doel was onderzoeken hoe de kindroute door leerkrachten werd ingezet. De mate waarin leerkrachten de kindroute inzetten was erg verschillend. De meerderheid van de leerkrachten gebruikte de kindroute echter bij slechts een klein deel (<10%) van de leerlingen. Een inzet van de kindroute van minder dan 10% lijkt in eerste instantie laag. Toch is dit niet geheel onverwacht. De kindroute is tenslotte aanvullend op de leerkrachtroute. Daarnaast was een van de wensen uit het werkveld dat er een kort en krachtig instrument werd ontwikkeld (Schouwstra & Vloedgraven, 2020). Het inzetten van de kindroute vraagt echter om enige extra tijdsinvestering van de leerkracht. Ondanks verschillen in de inzet van de kindroute, kunnen we op basis van de resultaten concluderen dat deze voor leerkrachten wel bijdraagt aan de toetsing van leerlingen. Het was niet duidelijk of leerkrachten de kindroute meerdere keren bij leerlingen hebben toegepast. Daardoor kan niet uitgesloten worden dat leerkrachten de kindroute bij één leerling vaker hebben ingezet. Dit kan echter wel de wijze waarop de resultaten begrepen worden beïnvloeden. Conclusies over de mate waarin de kindroute wordt ingezet, moeten daarom met enige voorzichtigheid worden genomen.

Hoewel het aantal jaren werkervaring niet als verklaring kon worden gezien voor de verschillen in het gebruik van de kindroute, zijn er wel een aantal andere verklaringen

denkbaar. Een verklaring kan liggen in het verschil in het aantal instrumenten (rekenen, taal motoriek en/of sociaal emotioneel) dat leerkrachten gebruikten. Daarnaast waren er een aantal, maar niet alle, leerkrachten die hoog scoorden op de 'Irrelevance' schaal en geen gebruik maakte van de kindroute. Er waren ook enkele leerkrachten die de kindroute niet inzette, en erg laag op de 'School Accounting' schaal scoorden. Hoe leerkrachten kijken naar de bruikbaarheid van de toets, kan dus mogelijk de inzet van de kindroute beïnvloeden. Een andere mogelijkheid is dat het gebruik van de kindroute afhankelijk is van de werkdruk van leerkrachten. Wanneer leerkrachten een hogere werkdruk ervaren, zullen zij mogelijk minder snel geneigd zijn een extra tijdsinvestering te doen. Onderliggende overtuigingen over de betrouwbaarheid van observatie kunnen mogelijk ook invloed hebben gehad op de inzet van de kindroute. De verschillen in het gebruik van de kindroute en de eerder benoemde mogelijke verklaringen brengen aan het licht dat de keuze van de inzet van de kindroute complex is.

**Sterkte- en zwakteanalyse**

De vergelijking tussen de percepties van leerkrachten over toetsen werd vergemakkelijkt doordat bij het verzamelen van de nieuwe data gebruik is gemaakt van hetzelfde instrument dat in het eerdere onderzoek naar de oude kleutertoetsen is gebruikt. Daarnaast is dit instrument, de CoA-III-A, in eerdere onderzoeken ingezet in verschillende jurisdicties, onderwijs niveaus en culturen, waardoor de resultaten van dit onderzoek in een breder referentiekader kon worden gezet en daardoor hielp bij de analyse (Bonner, 2016; Brown, 2013; Daniels et al., 2014). Er was een onderdeel van de CoA-III-A dat niet goed aansloot bij het meten van de doelen van toetsing in het kleuteronderwijs. Het subonderdeel 'toetsing verbeterd het lerend vermogen' is uit de vragenlijst gehaald omdat dit in eerder onderzoek niet passend bleek (Frans et al., 2020). De CoA-III-a sloot dus niet volledig aan op toetsing specifiek bij kleuters. Daarnaast was de Cronbach's Alpha van de 'Student Accounting' schaal laag en vormde de schaal een zwakke schaalstructuur. De items in die schaal meten daardoor mogelijk verschillende constructen en vertoonden weinig samenhang. Dit ging ten koste van de betrouwbaarheid van deze schaal. De beperkte grootte van de schaal (3 items) kan hier invloed op hebben gehad. Dat het formele leren nog niet op de voorgrond staat in het kleuteronderwijs, in vergelijking met het verdere basis- en voortgezet onderwijs kan ook van invloed zijn geweest (Eurydice, 2023). Het toewijzen van een cijfer (item 12) kan bijvoorbeeld logischer klinken voor leerkrachten in de bovenbouw van de basisschool dan voor kleuterleerkrachten. Het toevoegen van meer items die dit construct meten, zoals de niet verkorte versie van de vragenlijst heeft, had mogelijk een meer betrouwbare schaal gegeven. Een andere beperking van de CoA-III-A was dat een item in de 'Irrelevance' schaal niet aan sloot bij de rest van de items in die schaal. De H-coëfficiënten uit de mokkenschaal analyse droegen daarentegen wel aan bij het duiden van de betrouwnaarheid en de interne consistentie van de schalen, die voor de overige schalen goed was.

In de vragenlijst is gecontroleerd of leerkrachten tot de populatie behoorden doormiddel van controlevragen. De werving van respondenten verliep via het netwerk, maar ook via sociale media en open informatie over scholen. Daardoor is het risico op onderzoekersbias beperkt. In combinatie met de anonimiteit van de vragenlijst hinderde dit echter het verzamelen van informatie over de respons (responserate). Daardoor is niet te controleren of meerdere leerkrachten, werkzaam op dezelfde scholen de vragenlijst hebben ingevuld. Omdat dergelijke

leerkrachten onder hetzelfde schoolbeleid werken, kunnen meerdere observaties in het echt slecht één observatie weergeven. Anders gezegd: omdat respondenten vergelijkbare kenmerken of ervaringen hebben gehad, kunnen meerdere observatie mogelijk in de werkelijkheid slechts één observatie zijn. Hierdoor kan de standaardmeetfout onderschat zijn. Omdat de onderzoeksgegevens anoniem zijn verzameld, kon dit niet gecontroleerd worden. Het nadeel hiervan is echter beperkt, omdat er geen significante effecten zijn gevonden. De anonimiteit van de vragenlijst heeft mogelijk wel bijgedragen aan eerlijkere reacties. Er is voorafgaand een powerberekening gedaan om te kijken hoe groot de steekproef moest zijn om een medium tot groot effect aan te tonen. Het aantonen van ten minste een medium effect was echter een gok. Omdat er slechts kleine tot zeer kleine effecten zijn waargenomen, was de steekproef te klein voor een voldoende hoge power. Tegelijkertijd kunnen we ons afvragen in hoeverre het meerwaarde heeft voor theorievorming om voldoende power te hebben bij slechts heel kleine verschillen tussen de gemiddeldes op schalen. Als leerkrachten die 'Kleuter in beeld' gebruikten bijvoorbeeld slechts een heel klein beetje hoger scoorden op de 'Improvement' schaal, dan is deze informatie voor het opzetten van (nieuwe) theorieën over percepties van leerkrachten op toetsing, niet enorm van invloed. Dat het effect wel significant zou zijn bij voldoende power, lijkt in dat geval minder van belang. Er waren een aantal leerkrachten in de nieuwe verzamelde data die niet alle CoA-III-A vragen hebben ingevuld. De invloed op de resultaten lijkt beperkt omdat zijn niet opvallend positief of negatief scoorden op de items die ze wel hebben ingevuld en het slechts om een kleine groep ging. Er was een klein verschil in opleiding tussen de steekproeven. Dit lijkt echter een logisch gevolg van het feit dat er nieuwe leerkrachten komen die de PABO, als huidige leerkrachtopleiding hebben gedaan en leerkrachten de KLOS hebben gevolgd met pensioen gaan.

Het kwantitatieve karakter van het onderzoek, belemmerde tenslotte theorievorming over de inzet van de kindroute. Hoewel werkervaring geen invloed lijkt te hebben op de inzet, zijn er verschillende andere verklaringen benoemd die vanuit de kwantitatieve data of (tijds-) beperkingen niet konden worden bevestigd of weerlegt.

**Implicaties en aanbevelingen voor vervolgonderzoek**

Het soort toetsingsinstrument beïnvloed volgens deze onderzoeksresultaten de manier waarop leerkrachten toetsingsdoelen zien, formatief danwel summatief, niet. Dit betekent echter niet dat het soort toetsingsmethode niet relevant is voor leerkrachten. Observatie werd als eerlijker en nauwkeuriger gezien en gaat minder tegen de overtuigingen van leerkrachten in. Dat leerkrachten meer regie en flexibiliteit hebben, kan hier mogelijk aan hebben bijgedragen. De beslissing van het kabinet om observatie-instrumenten in te zetten, lijkt voor leerkrachten geen onlogische of onnodige keuze (Van Engelshoven, 2018). De observatie-instrumenten doen voor leerkrachten in ieder geval niet onder voor de oude kleutertoetsen. De sterke correlaties tussen de formatieve en summatieve doelen, benadrukken verder dat toetsing, ongeachte het oorspronkelijke doel, ook impliciete oordelen over leerlingen, leerkrachten of scholen kan bevatten. Het is van belang om ons hiervan bewust te zijn bij het gebruik en de ontwikkeling van toetsingsinstrumenten. Het is daarnaast relevant om verder te onderzoeken in hoeverre er beslissingen worden genomen aan de hand van deze oordelen en wat de voorspellende waarde van observatie-instrumenten is.

De waargenomen verschillen in het gebruik van de kindroute benadrukten tenslotte de complexiteit van de keuzes die leerkrachten maken. Die complexiteit van de inzet kon door de beperkte omvang van dit onderzoek niet voldoende onderzocht worden. Om een beter begrip te krijgen over welke factoren de inzet van de kindroute beïnvloeden, zou een kwalitatief vervolgonderzoek hierover waardevol zijn. Daarnaast is het van meerwaarde om verder te onderzoeken in hoeverre de percepties op toetsing van invloed zijn op het gebruik van de kindroute. Hoewel de beperkingen van dit onderzoek het trekken van harde conclusies belemmert, zijn de resultaten hoopvol voor de toetsingspraktijk bij kleuters.

## Referenties


Appl, D. J. (2000). Clarifying the Preschool Assessment Process: Traditional Practices and Alternative Approaches. *Early Childhood Education Journal*, *27*(4), 219–225. https://doi.org/10.1023/B:ECEJ.0000003358.78284.fa

Baghaei, P. (2021). *Mokken Scale Analysis in Language Assessment*. Waxmann Verlag GmbH.

Barnes, N., Fives, H., & Dacey, C. M. (2015). Teachers' Beliefs about Assessment. In H. Fives & M. G. Gill (Eds.), *International Handbook of Research on Teachers' Beliefs* (pp. 229–233). Routledge.

Barnes, N., Fives, H., & Dacey, C. M. (2017). U.S. teachers' conceptions of the purposes of assessment. *Teaching and Teacher Education*, *65*, 107–116. https://doi.org/10.1016/j.tate.2017.02.017

Black, P. J. (Paul J., Harrison, C., Lee, C., Marshall, B., & Dylan, W. (2003). *Assessment for learning: putting it into practice*. Open University Press.

Bonner, S. M. (2016). Teachers' perceptions about assessment: Competing narratives. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (1st ed., pp. 21–39). Routlegde.

Brodie, K. (2013). *Observation, Assessment and Planning in the Early Years - Bringing It All Together*. McGraw-Hill Education. http://ebookcentral.proquest.com/lib/rug/detail.action?docID=1170012

Brown, G. T. L. (2004). Teachers' conceptions of assessment: implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 301–318. https://doi.org/10.1080/0969594042000304609

Brown, G. T. L. (2006). Teachers' conceptions of assessment: Validation of an Abridged version. In *Psychological Reports* (Vol. 99).

Brown, G. T. L. (2013). *Conceptions of assessment: Understanding what assessment means to teachers and students*. Nova Science Publishers.

Brown, G. T. L., & Harris, L. R. (2009). The complexity of teachers' conceptions of assessment: tensions between the needs of schools and students. *Assessment in Education: Principles, Policy & Practice*, *16*(3), 365–381. https://doi.org/10.1080/09695940903319745

Brown, G. T. L., Hui, S. K. F., Yu, F. W. M., & Kennedy, K. J. (2011). Teachers' conceptions of assessment in Chinese contexts: A tripartite model of accountability, improvement, and

irrelevance. *International Journal of Educational Research*, *50*(5–6), 307–320. https://doi.org/10.1016/j.ijer.2011.10.003

Calderhead, J. (1996). Teachers: Beliefs and knowledge. In *Handbook of educational psychology.* (pp. 709–725). Prentice Hall International.

Calderhead, J. (2011). Teachers' conceptions of assessment: Comparing primary and secondary teachers in New Zealand. *Assessment Matters*, *3*, 45–70. https://doi.org/10.18296/am.0097

Certo, J. L. (2006). Beginning teacher concerns in an accountability-based testing environment. *Journal of Research in Childhood Education*, *20*(4), 331–349. https://doi.org/10.1080/02568540609594571

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences Second Edition*.

Condliffe, B. F., & Plank, S. B. (2013). Pressures of the Season: An Examination of Classroom Quality and High-Stakes Accountability. *American Educational Research Journal*, *50*(5), 1152–1182. https://doi.org/10.3102/0002831213500691

Daniels, L. M., Poth, C., Papile, C., & Hutchison, M. (2014). Validating the Conceptions of Assessment-III Scale in Canadian Preservice Teachers. *Educational Assessment*, *19*(2), 139–158. https://doi.org/10.1080/10627197.2014.903654

de Leeuw, E. D. (2012). Counting and Measuring Online. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *114*(1), 68–78. https://doi.org/10.1177/0759106312437290

Deci, E. L., & Ryan, R. M. (2012). Self-Determination Theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (Vol. 1, pp. 416–436). Sage Publications Ltd.

Eurydice. (2023). *Early childhood education and care in the Netherlands*. https://eurydice.eacea.ec.europa.eu/national-education-systems/netherlands/early-childhood-education-and-care

Fives, H., & Buehl, M. M. (2012). Spring cleaning for the "messy" construct of teachers' beliefs: What are they? Which have been examined? What can they tell us? In *APA educational psychology handbook, Vol 2: Individual differences and cultural and contextual factors.* (pp. 471–499). American Psychological Association. https://doi.org/10.1037/13274-019

Frans, N., Post, W. J., Huisman, M., Oenema-Mostert, I. C. E., Keegstra, A. L., & Minnaert, A. E. M. G. (2017). Early identification of children at risk for academic difficulties using standardized assessment: stability and predictive validity of preschool math and language scores. *European Early Childhood Education Research Journal*, *25*(5), 698–716. https://doi.org/10.1080/1350293X.2017.1356524

Frans, N., Post, W. J., Oenema-Mostert, C. E., & Minnaert, A. E. M. G. (2020). Preschool/Kindergarten teachers' conceptions of standardised testing. *Assessment in Education: Principles, Policy and Practice*, *27*(1), 87–108. https://doi.org/10.1080/0969594X.2019.1688763

Gollub, J. P., Bertenthal, M. W., Labov, J. B., & Curtis, P. C. (2002). *Learning and Understanding: Improving Advanced Study of Mathematics and Science in U.S. High Schools (2002) National Academies of Sciences, Engineering, and Medicine. 2002. Learning and Understanding: Improving Advanced Study of Mathematics and Science in U.S. High Schools*. The National Academies Press.

Hanes, B. M. (2010). *Perceptions of Early Childhood Assessment among Early Childhood Educators.*

IBM Corporation. (2020). *Statistical Product and Service Solutions* (No. 28).

Izci, K. (2016). Internal and External Factors Affecting Teachers' Adoption of Formative Assessment to Support Learning. *International Journal of Educational and Pedagogical Sciences, 10*(8), 2800–2807.

Lansink, N., & Hemker, B. (2012). *Wetenschappelijke Verantwoording van de toetsen Taal voor kleuters voor groep 1 en 2 uit het Cito Volgsysteem primair onderwijs*.

Monteiro, V., Mata, L., & Santos, N. N. (2021). Assessment Conceptions and Practices: Perspectives of Primary School Teachers and Students. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.631185

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. https://doi.org/10.1080/09695940701478321

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/https://doi.org/10.1037/1089-2680.2.2.175

Nolan, K. P., & Highhouse, S. (2014). Need for Autonomy and Resistance to Standardized Employee Selection Practices. *Human Performance*, *27*(4), 328–346. https://doi.org/10.1080/08959285.2014.929691

Op den Kamp, M., & Keuning, J. (2012). *Wetenschappelijke verantwoording van de digitale toetsen Rekenen voor kleuters*.

Pellegrini, A. D. (2001). The Role of Direct Observation in the Assessment of Young Children. *Journal of Child Psychology and Psychiatry*, *42*(7), S002196300100765X. https://doi.org/10.1017/S002196300100765X

Prawat, R. S. (1992). Teachers' Beliefs about Teaching and Learning: A Constructivist Perspective. *American Journal of Education*, *100*(3), 354–395. https://doi.org/10.1086/444021

R Core Team. (2021). *R: A language and environment for statistical computing* (4.2.2). https://www.R-project.org/

Remesal, A. (2007). Educational reform and primary and secondary teachers' conceptions of assessment: The Spanish instance, building upon Black and Wiliam (2005). *Curriculum Journal*, *18*(1), 27–38. https://doi.org/10.1080/09585170701292133

Revelle, W. (2021). *Psych: Procedures for Psychological, Psychometric, and Personality Research* (2.1.6). https://CRAN.R-project.org/package=psych

Saeki, E., Segool, N., Pendergast, L., & von der Embse, N. (2018). The influence of test-based accountability policies on early elementary teachers: School climate, environmental stress, and teacher stress. *Psychology in the Schools*, *55*(4), 391–403. https://doi.org/10.1002/pits.22112

Schouwstra, S., & Vloedgraven, J. (2020). *Wetenschappelijke verantwoording Kleuter in beeld-Taal*.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to Nonparametric Item Response Theory* (Vol. 5). Sage Publications.

Sleegers, W. W. A., Proulx, T., & van Beest, I. (2019). Confirmation bias and misconceptions: Pupillometric evidence for a confirmation bias in misconceptions feedback. *Biological Psychology*, *145*, 76–83. https://doi.org/10.1016/j.biopsycho.2019.03.018

Snowling, M. J., Hulme, C., Bailey, A. M., Stothard, S. E., & Lindsay, G. (2012). *Language and Literacy Attainment of Pupils during Early Years and through KS2: Does teacher assessment at five provide a valid measure of children's current and future educational attainments?*

Taras, M. (2005). Assessment–Summative and formative– Some theoretical reflections. *British Journal of Educational Studies*, *53*(4), 466–478. https://doi.org/10.1111/j.1467-8527.2005.00307.x

Thornton-Lang, K. M. (2014). *Observation as a formal assessment tool in early childhood classrooms: A professional development module*. https://scholarworks.uni.edu/grp/238

Tóth, E., & Csapó, B. (2022). Teachers' beliefs about assessment and accountability. *Educational Assessment, Evaluation and Accountability*, *34*(4), 459–481. https://doi.org/10.1007/s11092-022-09396-w

Van der Ark, L. A., & Luijten, M. (2021). *Mokken: Mokken scale Analysys in R* (4.0.0). https://cran.r-project.org/web/packages/mokken/index.html

Van Dijk, M., & Van Geert, P. (2007). Wobbles, humps and sudden jumps: A case study of continuity, discontinuity and variability in early language development. *Infant and Child Development*, *16*(1), 7–33. https://doi.org/10.1002/icd.506

Van Engelshoven, I. (2018). *Uitvoering regeerakkoord t.a.v. kleutertoetsen*.

Vandeyar, S., & Killen, R. (2007). Educators' conceptions and practice of classroom assessment in post-apartheid South Africa. In *South African Journal of Education Copyright* © (Vol. 27, Issue 1).

# Bijlage
## Vragenlijst

Beste participant,

Bedankt voor uw interesse in dit onderzoek. De volgende vragenlijst is onderdeel van een onafhankelijk onderzoek naar attitudes van leerkrachten over toetsing in het kleuteronderwijs. Voor dit onderzoek zijn we op zoek naar leerkrachten van groep 1 en 2 in het (speciaal) basisonderwijs die gebruik maken van het Cito observatie-instrument 'Kleuter in beeld'. Kleuter in beeld kent vier subdomeinen: Taal, Rekenen, Motoriek en Sociaal-Emotioneel. Als u het observatie-instrument voor een of meer van deze subdomeinen gebruikt, kunt meedoen aan dit onderzoek.

Als student kunt u meedoen aan dit onderzoek als u in het vierde jaar van de Opleiding Leraar Basisonderwijs (Pabo) zit en werkt als Leraar in Opleiding. Studenten van de Academische opleiding leraar basisonderwijs (AOLB) kunnen participeren als zij in het derde of vierde leerjaar zitten en stage lopen.

Het invullen van de vragenlijst duurt maximaal **10 minuten**. De verzamelde gegevens worden beveiligd opgeslagen en zullen ten alle tijden anoniem worden verwerkt. Er wordt geen informatie uitgevraagd die herleidbaar is tot specifieke personen. U kunt op elk moment stoppen met uw deelname aan het onderzoek.

Gaat u akkoord met het gebruik van de ingevulde gegevens voor dit onderzoek?

- ○ Ik ga akkoord (1)

Bent u leerkracht van groep 1 en/of 2 in het (speciaal) basisonderwijs?

- ○ Ja (1)
- ○ Nee (2)

Gebruikt u een van de volgende observatie-instrumenten? U kunt meerdere antwoorden aanklikken.

- [ ] Kleuter in beeld - Rekenen (1)
- [ ] Kleuter in beeld - Taal (2)
- [ ] Kleuter in beeld - Motoriek (3)
- [ ] Kleuter in beeld - Sociaal-Emotioneel (4)
- [ ] Ik gebruik **geen** van deze instrumenten (5)

U voldoet helaas niet aan de voorwaarden om deze vragenlijst in te vullen. Toch bedanken we u voor uw interesse en tijd. U wordt nu naar het einde van de vragenlijst geleid.

Met welk geslacht identificeert u zichzelf het meest?

- ○ Man (1)
- ○ Vrouw (2)
- ○ Anders (3)

Wat is uw leeftijd?

- ○ Jonger dan 25 jaar (1)
- ○ 25 - 35 (2)
- ○ 36 - 45 (3)
- ○ 46 - 55 (4)
- ○ 56 - 60 (5)
- ○ 61 jaar of ouder (6)

Hoeveel jaren werkervaring heeft u als leerkracht (in het basisonderwijs)?

| | 0 5 10 15 20 25 30 35 40 45 50 |
|---|---|
| Jaren werkervaring () |  |

Welke opleiding heeft u gevolgd?

- ○ Opleiding Leraar Basisonderwijs (Pabo) (1)
- ○ Kleuter Leidster Opleiding School (KLOS) (2)
- ○ Pedagogische academie (PA) (3)
- ○ Academische opleiding leraar basisonderwijs (AOLB) (4)
- ○ Anders, namenlijk... (5) _____

De volgende stellingen gaan over hoe u naar **toetsing** kijkt.

Neem bij het beantwoorden expliciet het observatie-instrument **'Kleuter in beeld'** in gedachten als het over **toetsen** gaat.

In hoeverre bent u het eens of oneens met de volgende stellingen?

| | Sterk Oneens (1) | Grotendeels Oneens (2) | Beetje eens (3) | Gematigd eens (4) | Grotendeels eens (5) | Sterk eens (6) |
|---|---|---|---|---|---|---|
| Toetsen geeft informatie over hoe goed scholen het doen (1) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen plaatst leerlingen in categorieën (2) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen is een manier om te bepalen hoeveel leerlingen hebben geleerd van het aangeboden onderwijs (3) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen is verweven met de onderwijspraktijk (4) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsresultaten zijn betrouwbaar (5) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen dwingt leerkrachten om op een manier onderwijs te geven die tegen hun overtuiging in gaat (6) | ○ | ○ | ○ | ○ | ○ | ○ |

In hoeverre bent u het eens of oneens met de volgende stellingen?

| | Sterk Oneens (1) | Grotendeels Oneens (2) | Beetje Eens (3) | Gematigd Eens (4) | Grotendeels Eens (5) | Sterk Eens (6) |
|---|---|---|---|---|---|---|
| Leerkrachten toetsen, maar maken weinig gebruik van de resultaten (1) | ○ | ○ | ○ | ○ | ○ | ○ |
| Vanwege meetfouten moet er voorzichtig omgegaan worden met toetsresultaten (2) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen is een accurate indicator van de kwaliteit van een school (3) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen is het toewijzen van een cijfer en/of niveau aan het werk van leerlingen (4) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen stelt vast wat leerlingen geleerd hebben (in de breedste zin) (5) | ○ | ○ | ○ | ○ | ○ | ○ |
| Informatie verkregen van het toetsen, leidt tot aanpassingen in het (dagelijks) onderwijsaanbod (6) | ○ | ○ | ○ | ○ | ○ | ○ |

In hoeverre bent u het eens of oneens met de volgende stellingen?

| | Sterk Oneens (1) | Grotendeels Oneens (2) | Beetje Eens (3) | Gematigd Eens (4) | Grotendeels Eens (5) | Sterk Eens (6) |
|---|---|---|---|---|---|---|
| Toetsresultaten zijn consistent (1) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen is oneerlijk tegenover leerlingen (2) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsresultaten worden opgeslagen en genegeerd (3) | ○ | ○ | ○ | ○ | ○ | ○ |
| Bij toetsen zouden leerkrachten rekening moeten houden met fouten en onnauwkeurigheden in het meten (4) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen is een goede manier om een school te evalueren (5) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen bepaald of leerlingen voldoen aan de normen (6) | ○ | ○ | ○ | ○ | ○ | ○ |

In hoeverre bent u het eens of oneens met de volgende stellingen?

| | Sterk Oneens (1) | Grotendeels Oneens (2) | Beetje Eens (3) | Gematigd Eens (4) | Grotendeels Eens (5) | Sterk Eens (6) |
|---|---|---|---|---|---|---|
| Toetsen meet hogere denkvaardigheden van leerlingen (dus meer dan alleen het reproduceren, begrijpen en toepassen van kennis) (1) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen laat toe dat elke leerling op maat gesneden instructie krijgt (2) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsresultaten zijn te vertrouwen (3) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen belemmert het lesgeven (4) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen heeft weinig impact op het lesgeven (5) | ○ | ○ | ○ | ○ | ○ | ○ |
| Toetsen is een onnauwkeurig proces (6) | ○ | ○ | ○ | ○ | ○ | ○ |

Hoe vaak heeft u in het afgelopen kalenderjaar de kindroute gebruikt?

_____

Uit hoeveel leerlingen bestaat uw klas?

_____

mogen we eventueel contact opnemen voor een gesprek?

- ○ Ja, via het volgende mailadres: (1)

_____

- ○ Nee (2)