rijksuniversiteit groningen

Omgaan met meetnauwkeurigheid Cito-scores

De manier waarop de onzekerheid van Cito-scores wordt meegenomen in de beslissingen en interpretaties van basisschoolleerkrachten.

Marit Nicolai (S3666697)

Academische Pabo

Faculteit GMW

Rijksuniversiteit Groningen

PABA-A412

Begeleider: Dr. N. Frans

Tweede beoordelaar: Dr. S. Parlevliet

Juni 2022

Abstract

Research has suggested that many users have difficulty interpreting the accuracy of test scores. The uncertainty in test scores is often poorly understood by both teachers and parents and thus ignored when interpreting test scores, such as those from Cito. The purpose of this study is to clarify how the uncertainty of Cito-scores is interpreted and thus what effect this has on the decisions teachers make. While teachers may not have a full understanding of the nature of measurement errors, the presentation of measurement errors could lead to a greater awareness of the inaccuracy surrounding test scores compared to a score report that omits measurement errors. This awareness could encourage teachers to gather additional information about a student's abilities. This leads to the following research question: In what way do elementary school teachers take uncertainty in Cito-scores into account in their interpretations and decisions regarding these scores? The research method we used for collecting data was qualitative research. A total of 12 teachers were interviewed by three researchers. Business and private networks of the researchers were used to approach participants. The interviews were open coded and analyzed thematically. The most important result that emerged from the interviews were the answers to the question whether teachers knew what the score interval meant and whether they used it in their teaching. All 12 teachers initially did not know what the scoring interval meant, nor did any of these teachers use the interval for decisions for students. This research design was a qualitative small scale study, the reader should keep in mind that these results are difficult to generalize. But based on this small-scale qualitative study, an important recommendation for practice is to better inform teachers about the scoring interval.

“Assessment literacy verwijst naar het vermogen van leerkrachten om het werk en de prestatiegegevens van leerlingen te onderzoeken, nauwkeurig te begrijpen en om plannen voor de klas en de school te ontwikkelen om de omstandigheden te veranderen die nodig zijn om betere resultaten te behalen’’ (Fullan & Watson, 2000, p. 457). Maar op welke manier onderzoeken leerkrachten het werk van de leerlingen en hoe nauwkeurig wordt dit gedaan? Beoordelen en analyseren behoren tot één van de dagelijkse bezigheden van leerkrachten, met name tijdens de toetsweken. Op veel scholen worden ieder jaar Leerlingvolgsysteem (LVS) toetsen afgenomen (Van der Kleij & Eggen, 2013). De toetsen van Cito zijn één van de meest gebruikte toetsinstrumenten. Het LVS-programma omvat verschillende toetsen die gebruikt kunnen worden om de leervorderingen van leerlingen systematisch in kaart te brengen. De LVS-toetsen zijn primair bedoeld om leerkrachten inzicht te geven in de resultaten van het onderwijs dat is aangeboden. Deze inzichten kunnen vervolgens worden gebruikt om het onderwijs waar nodig aan te passen (Van der Kleij & Eggen, 2013). Door een zorgvuldig constructieproces worden toetsscores vaak beschouwd een waardevolle bron. Over het algemeen worden deze scores beschouwd als zeer betrouwbaar en niet-vooringenomen (Shepard, 2006); ze zijn echter ook onderhevig aan een zekere mate van meetfout (Gardner, 2013). Meetfout kan worden geconceptualiseerd als het verschil tussen de werkelijke of behaalde score en de theoretische ware score tegenhanger (Gardner, 2013). Bij de testscores die worden behaald op een Cito toets, worden om deze reden score intervallen gepresenteerd. De onzekerheid in de testscores van Cito worden weergegeven door middel van het 68%-betrouwbaarheidsinterval rondom de score.

De behoefte aan betrouwbaarheidsintervallen ontstaat, omdat de toets afnemers bij de interpretatie van een testscore de ware score van de leerlingen willen weten. De ware score is de score die een leerling zou behalen als er geen meetfouten aan die test waren verbonden; het is de werkelijke waarde van hetgeen wordt gemeten (Charter & Feldt, 2002). De behaalde score is echter een vertekende schatting van de ware score van de leerling. De behaalde score is feilbaar en leerkrachten dienen dit niet behandelen alsof het de ware score van de leerling is. Het kan verstandig zijn voor toets afnemers om een verkregen score altijd als een benadering te beschouwen; men kan er bijna zeker van zijn dat de verkregen score niet precies gelijk is aan de ware score. Hoewel het geweldig zou zijn om de ware score van de leerling te hebben voor interpretatie, weten we nooit de ware score tenzij de test perfect betrouwbaar is, wat het nooit is. Hoe dan ook, betrouwbaarheidsintervallen geven een bereik van plausibele ware scores om te helpen bij de interpretatie van de score (Charter & Feldt, 2002).

Onderzoek heeft gesuggereerd dat veel gebruikers moeite hebben met het interpreteren van de nauwkeurigheid van deze testscores (Van der Kleij & Eggen, 2013). De onzekerheid in testscores wordt vaak slecht begrepen door zowel leerkrachten als ouders en zodoende genegeerd bij het interpreteren van testscores zoals die van Cito. Als leerkrachten de rapportage van meetfouten negeren, of verkeerd gebruiken, dan kan dat een ernstig probleem zijn, omdat leerkrachten toetsscores nauwkeuriger zouden interpreteren dan ze zijn. Volgens de American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014), zouden uitgevers van toetsen de plicht moeten hebben om testgebruikers te voorzien van foutinformatie die hen in staat stelt om valide conclusies te trekken op basis van testresultaten. Hoewel leerkrachten wellicht geen volledig inzicht hebben in de aard van meetfouten, zou de presentatie van meetfouten kunnen leiden tot een groter bewustzijn over de onnauwkeurigheid rond toetsscores in vergelijking met een score verslag waarin meetfouten zijn weggelaten. Dit bewustzijn zou leerkrachten kunnen stimuleren om extra informatie te verzamelen over de capaciteiten van een leerling (Hopster-den Otter, Muilenburg, Wools, Veldkamp, & Eggen, 2018).

Uit onderzoek van Van der Kleij en Eggen (2013) blijkt dat bij de interpretatie van groei in vaardigheid, in het leerling rapport, opvallend weinig leerkrachten gebruik maakten van het betrouwbaarheidsinterval. Ook eerder onderzoek (Hambleton & Slater, 1997; Zenisky & Hambleton, 2012) over de interpretatie van score rapporten, gaven al aan dat statistische concepten gerelateerd aan betrouwbaarheidsniveaus vaak genegeerd worden door leerkrachten omdat ze dit als niet zinvol zien. Aangezien de interpretatie van gegevens noodzakelijk is voor het adequaat veranderen van omstandigheden om aan de behoeften van leerlingen te voldoen, raakt het aan een van de basisvaardigheden; vaardigheden om de prestaties van leerlingen te beoordelen. Een correcte interpretatie van de scores is een noodzakelijke voorwaarde voor de succesvolle voltooiing van alle fasen van de evaluatiecyclus. Een correcte interpretatie houdt immers rechtstreeks verband met het nemen van een gerechtvaardigd besluit. Leerkrachten nemen bijvoorbeeld beslissingen over de volgende stappen in de instructie, de plaatsing van leerlingen in verschillende instructie groepen of de noodzaak om een leerling extra ondersteuning te geven. Aangezien deze beslissingen serieuze consequenties kunnen hebben voor het onderwijzen en leren, moeten ze worden geïnformeerd door bewijs van hoge kwaliteit (Brookhart & Nitko, 2008). Daarentegen, wanneer een score correct wordt geïnterpreteerd, garandeert dit nog niet dat ook de juiste aanpassingen worden gemaakt op basis van deze scores. Desalniettemin is

assessment literacy niet beperkt tot het correct interpreteren van testresultaten, het gaat ook over het vermogen om kennis over wat leerlingen weten en kunnen, om te zetten in zinvolle instructieve acties (Fullan & Watson, 2000; Mandinach & Jackson, 2012; Popham, 2009).

Verschillende onderzoeken wijzen op misverstanden onder leerkrachten, rond de interpretatie van meetfouten (o.a. Impara, Divine, Bruce, Liverman, & Gay, 1991; Zwick, Zapata-Rivera en Hegarty, 2014). Belangrijke beslissingen op basis van testscores kunnen grote gevolgen hebben voor leerlingen, een voorbeeld hiervan is dat leerlingen mogelijk niet worden toegewezen aan een geschikte instructiegroep en daarom mogelijk niet de instructie krijgen die ze nodig hebben (Goodman & Hambleton, 2004; Newton, 2005; Phelps, Zenisky, Hambleton en Sireci, 2010). Verwarring bij leerkrachten rond het concept van meetfouten kan leiden tot verkeerd geïnformeerde beslissingen met nadelige gevolgen voor leerlingen. Er zou een nauwkeuriger beeld van de leerlingen verkregen kunnen worden als andere, authentieke, gegevensbronnen, zoals observaties van leerlingen of andere testscores, van de leerlingen ook meegenomen worden in deze beslissingen (Brookhart & Nitko, 2008; Mandinach, 2012).

In de huidige score rapporten van Cito wordt de onzekerheid van de testscores weergegeven door een score interval in het leerling rapport. De mate waarin meetfouten rond toetsscores de onderwijs beslissingen van leraren beïnvloeden, is tot nu toe onbekend. Daarnaast is de literatuur hierover al redelijk verouderd. Deze invloed van het score interval op beslissingen van leerkrachten bepaalt echter het nut van het weergeven van meetfouten. In dit onderzoek zal er daarom worden gekeken naar de manier waarop basisschoolleerkrachten die onzekerheid meenemen in hun interpretaties en beslissingen. Het doel van dit onderzoek is om helder te krijgen hoe de onzekerheid van de Cito-scores geïnterpreteerd wordt en welk effect dit dus heeft op de beslissingen die leerkrachten maken. Zodra deze informatie verzameld en geanalyseerd is, kan er ook gekeken worden naar de manier waarop het begrip van die onzekerheid bij leerkrachten verhoogd kan worden.

Dit leidt tot de volgende onderzoeksvraag: Op welke manier nemen basisschoolleerkrachten onzekerheid bij de Cito-scores mee in hun interpretaties en beslissingen met betrekking tot deze scores?

Onderzoeksmethode

Onderzoeksdesign

In dit onderzoek zijn er gegevens verzameld aan de hand van het multiple case study design. Het doel van dit onderzoek was om helder te krijgen hoe de onzekerheid van de Cito-scores geïnterpreteerd wordt door leerkrachten en hoe deze interpretatie meespeelt in de beslissingen die leerkrachten maken. Daarnaast is er samengewerkt met nog twee andere onderzoekers die keken naar verschillende (visuele) weergave vormen van de meetnauwkeurigheid van Cito-scores.

Populatie en steekproef

Voor dit onderzoek zijn basisschoolleerkrachten, door middel van een selecte steekproef, geworven die werken met het Cito LVS. De participanten zijn verworven door schoolstichtingen te benaderen in Noorden van Nederland. Er zijn in totaal twaalf participanten geïnterviewd. Het was de wens om leerkrachten van zoveel mogelijk verschillende lesgroepen, verschillende leeftijden en verschillend geslacht te benaderen. Met de reden dat mannelijke en vrouwelijke leerkrachten, leerkrachten met veel en weinig ervaring en leerkrachten in verschillende lesgroepen van elkaar kunnen verschillen, in de manier waarop ze beslissingen maken op basis van Cito. Daarom was onze wens om leerkrachten met zoveel mogelijk verschillende kenmerken te benaderen, om zo de externe validiteit van het onderzoek te vergroten.

Instrument

Om dit kwalitatieve onderzoek vorm te geven, is er gekozen voor semigestructureerde interviews. Tijdens het interview kon de interviewer doorvragen als de respondent iets interessants zei, of juist vragen om duidelijkheid als de interviewer niet begreep wat er werd bedoeld. Hiermee verkrijg je meer gedetailleerde informatie. Het interviewleidraad met een deel van de gestelde vragen werden vooraf vastgesteld. De vragen in het interview gingen over beslissingen die leerkrachten maken op basis van Cito-scores, de interpretatie van huidige Cito-scores en over verschillende (visuele) weergave vormen van Cito-scores. Allereerst werden er enkele persoonlijke vragen aan de leerkracht gesteld, zoals werkervaring, aantal jaar ervaring met Cito, in welke groep er les wordt gegeven en hoe de gebruiksvriendelijkheid van Cito wordt ervaren. Vervolgens kreeg de leerkracht een Cito score rapport te zien, te vinden in Bijlage 1. Er is gekozen voor dit rapport omdat deze leerling op de laatste score net een IV-score behaald, met een vaardigheidsscore van 128.

Echter, daarbij gegeven een score interval van 126 tot 130 die ook de III-score omvat. Op basis van dit score rapport werden er vragen gesteld over de manier van interpretatie zoals: wat ze zagen op het score rapport, welke beslissingen ze zouden nemen op basis van dit rapport, wat volgens hen de onzekerheid in deze score betekende en hoe dit een rol speelt in hun eigen onderwijs. Het interview werd afgesloten met vragen over andere visuele weergaven van het score interval.

Procedure

De schoolstichtingen en de participanten werden in eerste instantie per mail benaderd via het zakelijke- en privé netwerk van de onderzoekers. Met de participanten die mee wilden werken aan dit onderzoek, werd individueel een afspraak gepland om het interview af te nemen. De interviews werden zoveel mogelijk fysiek afgenomen op de werkplek van de geïnterviewde leerkracht. Wanneer het om praktische redenen niet mogelijk was om het interview fysiek te doen, dan werd er een online afspraak gemaakt via Teams. De interviews vonden plaats in de periode van 20 april tot en met halverwege mei en duurden gemiddeld 20 minuten. Voordat de interviews werden afgenomen is er toestemming van de ethische commissie verkregen. Voorafgaand aan het interview werden participanten mondeling en schriftelijk geïnformeerd over het doel van het onderzoek en werd er middels een toestemmingsformulier toestemming gevraagd om het interview voor onderzoeksdoeleinden op te nemen en te transcriberen. Daarbij werd aangegeven dat de gegevens vertrouwelijk behandeld werden, dat de participanten anoniem zouden blijven en dat ze op elk moment het interview en de opname konden stoppen. De resultaten zullen dus niet te herleiden zijn tot individuele personen. Dit werd gewaarborgd door geen namen of werkplekken te noemen en de leerkrachten om die reden een nummer te geven. Na het onderzoek zullen de transcripties 5 jaar worden bewaard op de beveiligde y-schijf van de universiteit, waarna ze vervolgens verwijderd worden.

Analyse

Zodra de interviews afgenomen waren, werden de interviews getranscribeerd en gecodeerd in Atlas TI. Iedere onderzoeker transcribeerde zijn eigen interview, maar codeerde alle 12 interviews. Echter voordat elke onderzoeker begon met coderen, vond er een intervisie plaats om de manier van coderen goed af te stemmen. Zodra het transcriberen en coderen afgerond was, gingen de onderzoekers weer met elkaar in gesprek over de resultaten. Na het transcriberen werd er gestart met het coderen in Atlas TI. Bij het eerste transcript werden de codes nog heel specifiek per antwoord geformuleerd. Gaandeweg het coderen van alle interviews, ontstonden er meer algemene codes en konden deze codes worden

opgehangen aan verschillende thema's. De thema's zijn gevormd op basis van de vragen die de interviewer stelde. In totaal zijn er voor dit onderzoek 39 codes gevormd op basis van de antwoorden van de 12 leerkrachten. De thema's zijn terug te lezen in de resultaten, Tabel 2. Op deze manier ontstond er een duidelijk overzicht van de respondenten die hetzelfde antwoord hadden gegeven, waardoor de resultaten makkelijker met elkaar en met de theorie vergeleken konden worden. Door de deductieve benadering wisten we op voorhand welke data gegenereerd zou worden en hoe het kader van de analyse eruit zou zien. De resultaten en de focus van de analyse ontwikkelden zich langzaam tijdens het onderzoeksproces (Boeije, 2005). In de resultaten die zijn beschreven na de analyse, is gebruik gemaakt van quotes. Achter de quote staat de leerkracht die deze uitspraak heeft gedaan, met het nummer die de code heeft in het transcript van de betreffende leerkracht.

Resultaten

Naar aanleiding van de afgenomen interviews is Tabel 1 opgesteld. In deze tabel zijn de kenmerken van de participanten weergegeven, die mogelijk van invloed kunnen zijn op de resultaten. De meerderheid (n = 7) van de leerkrachten zit in de leeftijd van 20 tot 29 jaar, 3 leerkrachten in de leeftijd van 30 tot 39 jaar en 2 leerkrachten in de leeftijd van 40 tot 49 jaar. Daarnaast zijn 7 leerkrachten werkzaam in de middenbouw, 4 leerkrachten in de bovenbouw en 1 leerkracht in midden- en bovenbouw. De werkervaring loopt van 1 tot 25 jaar en de ervaring met Cito loopt van een paar maanden tot 23 jaar. Opvallend is dat leerkracht 3 wel veel werkervaring heeft, maar nog maar een paar maanden met Cito werkt. Ze zei hierover dat ze tot die tijd nooit iets hoefde in te voeren dus dat ze er ook niet mee heeft gewerkt. Tot slot hebben 4 leerkrachten tijdens het interview een toelichting gekregen over het score interval, deze leerkrachten gaven aan het anders echt niet te weten.

Tabel 1

Kenmerken participanten

Participant	Geslacht	Leeftijd	Bouw	Werk-ervaring	Ervaring Cito in jaren	Uitleg interval
Leerkracht 1	Man	20 - 29	Middenbouw	5	5	Nee
Leerkracht 2	Man	30 - 39	Midden- en Bovenbouw	22	22	Nee

Leerkracht 3	Vrouw	30 - 39	Middenbouw	10	< 1	Ja
Leerkracht 4	Vrouw	30 - 39	Bovenbouw	12	12	Ja
Leerkracht 5	Man	20 - 29	Bovenbouw	5	2	Nee
Leerkracht 6	Man	20 - 29	Middenbouw	1	1	Ja
Leerkracht 7	Vrouw	20 - 29	Middenbouw	3	3	Ja
Leerkracht 8	Man	20 - 29	Bovenbouw	4	3	Nee
Leerkracht 9	Man	20 - 29	Middenbouw	3	3	Nee
Leerkracht 10	Vrouw	20 - 29	Bovenbouw	1	1	Nee
Leerkracht 11	Vrouw	40 - 49	Middenbouw	20	15	Nee
Leerkracht 12	Vrouw	40 - 49	Middenbouw	25	23	Nee

De codering van de 12 interviews heeft geleid tot verschillende codethema's. Elk thema bevat informatie over beslissingen van leerkrachten op basis van Cito-scores, of informatie over het gebruik van het score interval van Cito. De thema's zijn verwerkt in Tabel 2, met daarbij een quote die elk thema het meest typeert. Het thema 'verwoorden leerling rapport' omvat de antwoorden die leerkrachten gaven op de open vraag of ze konden omschrijven wat ze op het leerling rapport zagen. Het thema 'interpretatie score interval' is opgesplitst in meerdere reacties. De reden hiervoor is dat alle leerkrachten in hun eerste reactie al aangaven dat ze niet wisten wat het score interval betekende. Toch deden enkele leerkrachten, na doorvragen van de interviewer, een poging om het score interval te omschrijven.

Tabel 2

Hoofdthema's: beslissingen en gebruik score interval o.b.v. Cito

Codethema's	Voorbeeld quotes
Beslissingen op basis van Cito algemeen	"Op basis van Cito-scores ga je wel een plan bedenken voor de komende periode van een aantal maanden. En daarbij zet je dan inderdaad wel bepaalde instructie groepen op."
Gebruik andere informatiebronnen	"Ik kijk heel veel naar de lessen naar het gemaakte werk met de methode gebonden toetsen, ja en op basis daarvan kijk ik ook van hoe ik mijn groep verdeel."
Onzekerheid: betekenis en rol?	"Als ik een score niet vindt passen, dan kijk ik even terug van: hoe zijn de Cito's gemaakt, in welke periode was dat?

	Misschien kan ik dan terug kijken dat ik denk: oh er is iets gebeurd bij die leerlingen, ohja, nou, dat zal meegespeeld hebben in de manier waarop zij de Cito's gemaakt hebben.”
Verwoorden leerling rapport	“Ik zie eigenlijk vanaf groep 4 tot midden groep 5 een gemiddelde leerling die ook gemiddeld, of wat op een verwachte manier, groei laat zien. Telkens een mooie groei in vaardigheidsscore en in de laatste toets, de eind 5, dan zie ik de lijn wat afzwakken.”
Interpretatie score interval eerste reactie	“Zo, dat is een goede vraag! Ik heb echt geen idee.. Score interval is… nee.”
Interpretatie score interval tweede of derde reactie	“Dat ja, dat het de score is van het maximum wat ze kunnen scoren.”
	“Misschien de hoeveelheid vragen die zijn gemaakt in de afgelopen aantal jaren? En dat daaraan een score… Ja, ik zou het voor de rest echt niet weten!”
	“Ik denk dat er een marge op zit qua score, dat denk ik. Dat misschien het kind ook iets hoger had kunnen scoren, maar dan binnen die range zeg maar.”
	“Ja, dat zal wel de scores zijn waar ze tussen zitten. Tussen moeten zitten, zou je zeggen.”
Gebruik van score interval	“Dit score interval gebruik ik eigenlijk nooit in mijn interpretaties voor beslissingen, daar kijk ik nooit naar..”
Beslissing op basis van leerling rapport	“Ik zou dit kind nu in de zorggroep plaatsen, omdat als je kijkt naar de afgelopen twee afnames, dat dat steeds… het stijgt wel, maar het stijgt niet genoeg. Dus ik zou denken dat dit kind nu iets meer uitleg nodig heeft dan alleen de basisinstructie. Dat is niet meer voldoende.”

Om te beginnen de resultaten van het thema welke beslissingen de leerkrachten voor hun klas nemen op basis van de resultaten van Cito. Hieruit komt met name naar voren dat leerkrachten de resultaten gaan analyseren, om uit te vinden wie van de leerlingen op welke onderdelen uitvallen. Op basis hiervan geeft de meerderheid van de leerkrachten aan dat ze

een plan bedenken voor de groep, waaronder het maken van instructiegroepen valt. Doordat veel leerkrachten de toets per onderdeel analyseren zijn het vaak geen vaststaande instructiegroepen. Alleen leerkracht 12 geeft expliciet aan dat zij niet op basis van Cito instructiegroepen maakt, maar dat dit puur afhangt van methodewerk. Bijna alle leerkrachten (*n=11*) geven aan dat ze beslissingen vaak niet alleen maar baseren op de Cito-scores. Veel leerkrachten maken ook gebruik van andere informatiebronnen, met name methodewerk en observaties. Het methodewerk representeert de dagelijkse lessen in de klas, veel vakken werken namelijk vanuit een lesmethode ontworpen voor het basisonderwijs. Er waren negen leerkrachten die aangaven dat zij het methodewerk en de observaties ook gebruikten voor het maken van beslissingen in de klas. Er was één leerkracht die alleen methodewerk gebruikt en geen observaties, één leerkracht waarbij gevoel ook een grote rol speelt en één leerkracht waarbij de focus ligt op Cito en niet op andere informatiebronnen.

Wat betreft het voorbeeld leerling rapport verwoordden leerkrachten wat ze zagen als een gemiddeld scorende leerling die op zijn laatste toets afzwakt. Daarbij was er één leerkracht, leerkracht 7, die daaraan toevoegde:

"Maar op zich, ja, dit kan op een extra foutje waarschijnlijk gebaseerd zijn. Het is wel een hoge 4 en nou ja, dit is allemaal ten opzichte van het landelijk gemiddelde." – Leerkracht 7 (13)

Verder was er nog een leerkracht die niet goed kon antwoorden, met de reden dat je er niet bij bent, je weet niet was er is gevraagd, dus was het leerling rapport lastig te interpreteren.

Op basis van dit leerling rapport zou de helft van de leerkrachten (*n=6*) deze leerling zetten in de zorggroep, oftewel instructiegroep. Leerkracht 3 geeft hierover de volgende uitleg:

"Nou als ik zie dat de leerling stagneert en het dus eigenlijk weinig ontwikkeling heeft doorgemaakt en die heeft nou onvoldoende, dan zou ik deze leerling denk ik bij de instructietafel betrekken. Maar als ik zie dat een leerling het heel goed doet en eigenlijk mijn verlengde instructie niet nodig heeft dan kan die leerling ook af en toe losgelaten worden. Dus ik zou niet vast aan de instructietafel, maar misschien per onderdeel afhankelijk, per les afhankelijk… Maar ik zou hem zeker meenemen aan de instructietafel." – Leerkracht 3 (4)

Leerkracht 10 heeft hierover een duidelijke mening. Deze mening is niet gebaseerd op de groei, maar op de score die is gegeven. Haar reactie luidt als volgt:

"Nou een 4 score zit voor mij onder het gemiddelde, dus ik vind gewoon dat daar extra aandacht aan besteed moet worden." – Leerkracht 10 (6)

De andere helft van de leerkrachten geeft aan dat ze de leerling eerst nog in de basisgroep willen houden, maar tijdens instructie de leerling wel goed in de gaten blijven houden. Leerkracht 5 geeft hierover de volgende toelichting:

"Ik zou deze leerling bij de basisgroep houden en tijdens instructies goed in de gaten gaan houden, ook op basis van de data die je al hebt gaan beslissen van, heeft deze leerling bij een bepaald doel wel of geen instructie nodig? Maar ik zou deze leerling niet standaard bij een instructiegroep halen, dus in eerste instantie de basisgroep en mocht het toch nodig zijn bij de instructie extra groep." – Leerkracht 5 (5)

Een opvallende uitspraak die daaraan werd toegevoegd van één leerkracht, leerkracht 12:

"In de basisgroep, omdat ik het nog een hoge 4 vindt en dat is soms echt op één of twee woorden op Cito, want Cito berekent wel heel streng." – Leerkracht 12 (8)

Het volgende thema is de interpretatie van het score interval. Alle 12 leerkrachten geven als eerste reactie dat zij niet weten wat het score interval is. Als gevolg van doorvragen door de interviewer, kwamen er ook tweede en derde reacties naar voren waar leerkrachten toch probeerden te omschrijven wat het score interval zou kunnen betekenen. Hiervan zijn verschillende quotes te zien in Tabel 2. Vier leerkrachten (Leerkracht 2, 8, 10 en 11) benoemen dan toch dat het een soort range is waarbinnen je kunt verwachten dat de leerling scoort. Ook een paar leerkrachten proberen het te omschrijven, maar komen niet in de buurt van de werkelijke betekenis. Leerkrachten 6, 5 en 2 gaven de volgende reacties:

"Dat ja dat het de score is van het maximum wat ze kunnen scoren. Dat dit er dan uitgekomen is, denk ik." – Leerkracht 6 (7)

"Misschien de hoeveelheid vragen die zijn gemaakt in de afgelopen aantal jaren? En dat daaraan een score…. Ja, ik zou het voor de rest echt niet weten!" – Leerkracht 5 (10)

"Dat het, dat het de bedoeling is om.. Ik denk dat ik het nu een beetje snap en dat het score interval eigenlijk een score is, wat verwacht wordt waar je naartoe gaat groeien." - Leerkracht 5 (11)

"*Nou misschien wel wat ze verwachten wat hij gaat scoren*." – Leerkracht 2 (7)

De rest van de leerkrachten gaven bij hun tweede en derde reactie nog steeds aan het echt niet te weten. Alle 12 leerkrachten geven vervolgens aan het score interval niet te gebruiken, omdat ze bij het voorgaande thema ook aan hebben gegeven dat ze, in eerste instantie, niet weten wat het betekent.

"Nee, dat nee, ik heb daar nog nooit naar gekeken naar deze score. Dus die heb ik nog nooit meegenomen in.. Nee, dat die heb ik nog nooit gebruikt!" – Leerkracht 6 (8)

"*Nee, eigenlijk niet dus. Ik kijk er niet eens naar*." – Leerkracht 12 (7)

Het thema wat daarop volgt omvat de reacties van leerkrachten over wat de onzekerheid van een score volgens hen betekent en of dit een rol speelt in hun onderwijs. Leerkracht 1 baseert als enige zijn antwoord op het score interval. Leerkracht 1 heeft echter daarvoor aangegeven dat hij niet weet wat het score interval betekent en dat hij het niet gebruikt, maar binnen dit thema zegt hij het volgende erover:

"*Nou 128 nemen ze als gemiddelde, maar het had ook best 130 kunnen zijn, of 126 denk ik."*- Leerkracht 1 (9)

Daarnaast geeft meer dan de helft van de leerkrachten aan dat de onzekerheid zit in externe factoren. Dingen die zijn genoemd zijn: *de thuissituatie, het is een moment opname, hoe de leerling zich heeft gevoeld die dag, persoonlijke factoren, onnodige of domme fouten.* De factoren die dus te maken hebben met de kenmerken van het kind. De opmerking "*het is een moment opname*" kan zowel geïnterpreteerd worden als kenmerk van het kind, als een kenmerk van de test. Hier is door de leerkracht verder geen duidelijke toelichting op gegeven. Verder zijn er geen specifieke kenmerken van de test benoemd die te maken kunnen hebben met de onzekerheid van een testscore. Daarbij geven drie leerkrachten aan dat de onzekerheid van een score is, dat het beeld wat je van de leerling hebt, niet past bij de score die je hebt gekregen, het geeft niet altijd een reëel beeld. Leerkracht 7 voegt hieraan toe:

"Ik denk wel eens te weinig, omdat ja, je gaat gewoon wel door met de dag. En ja, je hebt dan misschien niet altijd tijd voor hen om te kijken van hé, is er iets of om dat goed te analyseren? Dat merk ik wel, dus dat ik daar gewoon ja te weinig tijd voor hebt op een dag."

- Leerkracht 7 (11)

Conclusie & discussie

Het doel van dit onderzoek was om helder te krijgen hoe de onzekerheid van de Cito-scores geïnterpreteerd wordt en welk effect dit dus heeft op de beslissingen die leerkrachten maken. In dit onderzoek is er getracht antwoord te vinden op de volgende onderzoeksvraag: *Op welke manier nemen basisschoolleerkrachten onzekerheid bij de Cito-scores mee in hun interpretatie en beslissingen met betrekking tot deze scores?* Uit de resultaten is gebleken dat leerkrachten het score interval dat Cito meegeeft bij de scorerapporten, niet gebruiken. Alle 12 leerkrachten wisten in eerste instantie niet wat het score interval betekende en gebruikten dit niet voor hun beslissingen die ze maken op basis van Cito. Daarentegen wisten sommige leerkrachten, na doorvragen van de interviewer, wel een betekenis aan het score interval te verbinden, maar gebruikten dit alsnog niet voor hun beslissingen in de klas.

Leerkrachten die deelnamen aan dit onderzoek analyseren de resultaten van Cito en maken op basis daarvan beslissingen voor hun klas. De leerkrachten keken met name naar de groei in de hele grafiek en benoemde dat de leerling op de laatste toets afzwakte en gegaan is naar een IV-score. De leerkrachten kunnen echter, gezien het score interval, niet stellen dat de leerling per definitie afzwakt omdat ze te weinig informatie hebben. De meerderheid van de leerkrachten uit dit onderzoek maakt op basis van de Cito-scores een plan, waaronder de veelgenoemde instructiegroepen. Een groot deel van de participanten geeft daarbij aan dat ze per specifiek domein van de toets gaan kijken waar leerlingen op uitvallen, dus dat instructiegroepen niet vaststaand zijn. Naast Cito, maakten veel van de participanten ook gebruik van methodewerk en observaties voor hun beslissingen in de klas. De leerkrachten uit dit onderzoek kijken dus met name naar de groei in de grafiek die Cito meegeeft en het verschil in vaardigheidsscores. Ze benoemen dat er met deze leerling iets aan de hand is, dat hij nu onvoldoende scoort en dat de lijn op de laatste toets niet in de verwachting ligt. Het lijkt erop dat veel leerkrachten een IV-score per definitie zien als een onvoldoende. Geen enkele leerkracht benoemt het score interval, de onzekerheid of betrouwbaarheid van een score of merkt op dat het ook een III-score had kunnen zijn. Als leerkrachten wel de betekenis van het score interval kende en hadden betrokken bij hun keuzes, hadden ze die IV-score misschien niet als onvoldoende of zorgelijk beschouwd. De keuze die leerkrachten vervolgens maakten om de leerling in de basisgroep of instructiegroep te plaatsen, was wisselend onder de leerkrachten. Geen enkele keuze was in ieder geval gebaseerd op de onzekerheid van de toets.

De bevinding dat leerkrachten niet weten wat het score interval betekent, sluit aan bij

de verwachting die vooraf is gesteld aan dit onderzoek. Uit, deels verouderd, onderzoek is namelijk ook gebleken dat bij de interpretatie van groei in vaardigheid, in het leerling rapport, opvallend weinig leerkrachten gebruik maakten van het betrouwbaarheidsinterval (Van der Kleij & Eggen, 2013). Ook eerder onderzoek (Hambleton & Slater, 1997; Zenisky & Hambleton, 2012) over de interpretatie van score rapporten, gaven al aan dat statistische concepten gerelateerd aan betrouwbaarheidsniveaus vaak genegeerd worden door leerkrachten omdat dit niet als nuttig wordt beschouwd. Maar, dit betekent natuurlijk niet direct dat deze leerkrachten niks doen met 'onzekerheid', of dit niet in hun achterhoofd houden. Bij de toelichting over wat onzekerheid betekent, benoemen ze vooral de kindkenmerken die kunnen zorgen voor onzekerheid. Met deze factoren proberen de meeste leerkrachten wel rekening te houden. Leerkrachten gaan de toets analyseren en als de uitkomst niet past bij het beeld wat ze hebben van de leerling, betrekken ze deze factoren erbij. Geen enkele leerkracht benoemt factoren van de toets die te maken kunnen hebben met onzekerheid.

Er lijkt bij sommige leerkrachten toch een bepaald besef te zijn van onzekerheid, maar ze lijken het niet goed onder woorden te kunnen brengen. In ieder geval brengen ze de onzekerheid niet in verband met de betrouwbaarheid van een test. Het onderzoek van Van der Kleij & Eggen (2013) sluit hierbij aan, omdat zij ook suggereren dat veel gebruikers moeite hebben met het interpreteren van de nauwkeurigheid van testscores.

Dit onderzoeksdesign was een kwalitatief kleinschalig onderzoek, waardoor de resultaten lastig te generaliseren zijn. Wat betreft de interbeoordelaarsbetrouwbaarheid hebben alle drie de onderzoekers zich gehouden aan het interviewprotocol en is onderling afgestemd op welke manier het interview werd afgenomen. Doordat de interviews door drie verschillende onderzoekers zijn afgenomen zou er wel sprake zou kunnen zijn van interviewbias; beïnvloeding door de interviewer. Het gevolg kan zijn dat niet elke interviewer dezelfde informatie achterhaald heeft bij de participant, of dat de participant wellicht onbewust een bepaalde richting in is gestuurd.

Elke onderzoeker heeft alle 12 interviews gecodeerd. Het codeerschema is met elkaar besproken, maar niet met elkaar vergeleken. De reden hiervoor was, dat voor elke onderzoeker een ander gedeelte van het interview belangrijk was. Toch is de betrouwbaarheid van het codeerschema zoveel mogelijk gewaarborgd door duidelijke codes toe te kennen, dit transparant weer te geven in Tabel 2 en zoveel mogelijk citaten van leerkrachten te gebruiken om de resultaten te onderbouwen.

Doordat er gebruik is gemaakt van het zakelijk- en privé netwerk kunnen de

respondenten de onderzoeker kennen, dit zou mogelijk kunnen leiden tot sociaal-wenselijke antwoorden, waardoor de resultaten wellicht een vertekend beeld kunnen geven. Daarentegen is voorafgaand aan het interview wel expliciet benoemd dat de onderzoekers opzoek waren naar eerlijke antwoorden. Er is dus geen sprake geweest van een willekeurige steekproef, maar een selecte steekproef die wel heeft geleid voor veel variatie in leeftijd, ervaring, geslacht en lesgroep. Ondanks dat dit een kleinschalig onderzoek is dat lastig te generaliseren is, geeft deze selecte steekproef wel een redelijke afspiegeling van de praktijk.

Wat de resultaten verder ten goede doet, is dat de participanten de kans kregen om hun antwoorden toe te lichten en dat de interviewer kon doorvragen naar de achterliggende gedachtegang. Het verhaal van de leerkrachten achter de statistieken wordt zo een stuk duidelijker dan bij kwantitatief onderzoek. Op deze manier zijn de resultaten rijker aan informatie dan bij een vragenlijst. De keuze voor een semi-gestructureerd interview vraagt echter ook meer van de interviewer. Het nadeel hiervan is dat de interviewer zich tijdens de eerste gesprekken meer vastklampt aan het interviewprotocol, dan bij de latere gesprekken. De interviewer heeft dan namelijk een duidelijker beeld van wat hij voor antwoorden kan verwachten en op welke antwoorden er nog verder doorgevraagd kan worden.

Voor vervolgonderzoek is het advies om een grootschaliger kwantitatief onderzoek op te zetten, om zo de resultaten te kunnen generaliseren. Hiervoor zou een willekeurige steekproef van tenminste 250 leerkrachten worden aangeraden. Er zou een vragenlijst ontworpen kunnen worden met vragen over de betekenis en interpretatie van het score interval van Cito. Op deze manier zou je de uitkomsten hierover beter kunnen generaliseren. Een ander advies voor vervolgonderzoek is om een grootschalig experimenteel onderzoek op te zetten. Met een experimentele groep die uitgebreid geïnformeerd is over het score interval en een controlegroep die deze informatie niet heeft gekregen. Met het doel om beslissingen die leerkrachten maken met elkaar te vergelijken. Om te toetsen of leerkrachten met meer kennis over het score interval werkelijk andere, of meer voorzichtige, beslissingen zouden nemen voor hun leerlingen.

Uit dit onderzoek is gebleken dat alle participanten in eerste instantie geen betekenis konden geven aan het score interval van Cito en er is gebleken dat dit score interval geen rol speelt in beslissingen die worden gemaakt voor leerlingen. Deze resultaten geven een duidelijke aanleiding voor verder onderzoek naar dit onderwerp. Al met al is een belangrijke aanbeveling voor de praktijk, op basis van dit kleinschalige kwalitatieve onderzoek, om leerkrachten beter te informeren over het score interval. Als leerkrachten beter geïnformeerd zijn, kan dit stimulerend werken om extra informatie te verzamelen over de capaciteiten van

een leerling (Hopster-den Otter, Muilenburg, Wools, Veldkamp, & Eggen, 2018). Goed geïnformeerde leerkrachten zullen waarschijnlijk beter in staat zijn om valide conclusies te trekken uit de testscores van hun leerlingen.

Literatuurlijst

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Boeije, H. (2005). Analyseren in kwalitatief onderzoek: denken en doen. Amsterdam: Boom.

Brookhart, S. M., & Nitko, A. J. (2008). Assessment and grading in classrooms. Upper Saddle River, NJ: Pearson Education.

Charter, R. A., & Feldt, L. S. (2002). The importance of reliability as it relates to true score confidence intervals. Measurement and Evaluation in Counseling and Development, 35(2), 104–12.

Fullan, M., & Watson, N. (2000). School-based management: Reconceptualizing to improve learning outcomes. School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 11, 453–473 http://dx.doi.org/10.1076/sesi.11.4.453.3561.

Gardner, J. (2013). The public understanding of error in educational assessment. Oxford Review of Education, 39, 72–92. doi:10.1080/03054985.2012.760290

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. Applied Measurement in Education, 17, 145–220. doi:10.1207/s15324818ame1702

Hambleton, R. K., & Slater, S. C. (1997). Are NEAP executive summary reports understandable to policy makers and educators? CSE Technical Report 430. Los Angeles: National Centre for Research on Evaluation, Standards, and Student Teaching.

Hopster-den Otter, D., Muilenburg, S. N., Wools, S., Veldkamp, B. P., & Eggen, T. T. J. H. M. (2018). Comparing the influence of various measurement error presentations in test score reports on educational decision-making. Assessment in Education, 26(2), 123–142.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? Educational Measurement: Issues and Practice, 10(4), 16–18.

Mandinach, E. B., & Jackson, S. S. (2012). Transforming teaching and learning through data-driven decision making. Thousand Oaks, CA: Corwin.

Newton, P. E. (2005). The public understanding of measurement inaccuracy. British Educational Research Journal, 31, 419–442. doi:10.1080/01411920500148648

Phelps, R. P., Zenisky, A., Hambleton, R. K., & Sireci, S. G. (2010). On the reporting of measurement uncertainty and reliability for U.S. educational and licensure tests (OFQUAL 10/4759). London: Office of Qualifications and Examinations.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? Theory into Practice, 48, 4–11 http://dx.doi.org/10.1080/00405840802577536.

Van der Kleij, F., & Eggen, T. J. H. M. (2013). Interpretation of the score reports from the computer program lovs by teachers, internal support teachers and principals. Studies in Educational Evaluation, 39(3), 144–152.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 623–646). Westport: American Council on Education and Praeger.

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. Educational Measurement: Issues and Practice, 31, 21–26 http://dx.doi.org/10.1111/j. 1745-3992.2012.00231.x.

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal - representations of measurement error in test score reports. Educational Assessment, 19, 116–138. doi:10.1080/10627197.2014.903653

Bijlage 1. Interviewschema

Allereerst bedankt voor je deelname aan dit onderzoek. Samen met twee andere studenten doe ik onderzoek naar de interpretatie van meetnauwkeurigheid van Cito-scores. Voor een valide onderzoek is het belangrijk dat je antwoorden zo eerlijk mogelijk zijn en niet sociaal wenselijk. We zijn namelijk op zoek naar je mening en niet naar goede antwoorden.
Om het interview goed uit te kunnen werken, zou ik het graag willen opnemen. Vind je dat goed? Dan start ik nu de opname.
Alle gegevens zullen geanonimiseerd worden en er komen dus geen namen van leerkrachten en scholen in voor. Je antwoorden zijn dus helemaal anoniem. Verder worden de gegevens alleen gebruikt voor dit onderzoek en alle gegevens worden na 5 jaar verwijderd.
Het kan zijn dat ik na een bepaald antwoord nog doorvraag omdat iets me niet helemaal duidelijk is. Heb je nog vragen vooraf?

Om te starten een paar algemene vragen:

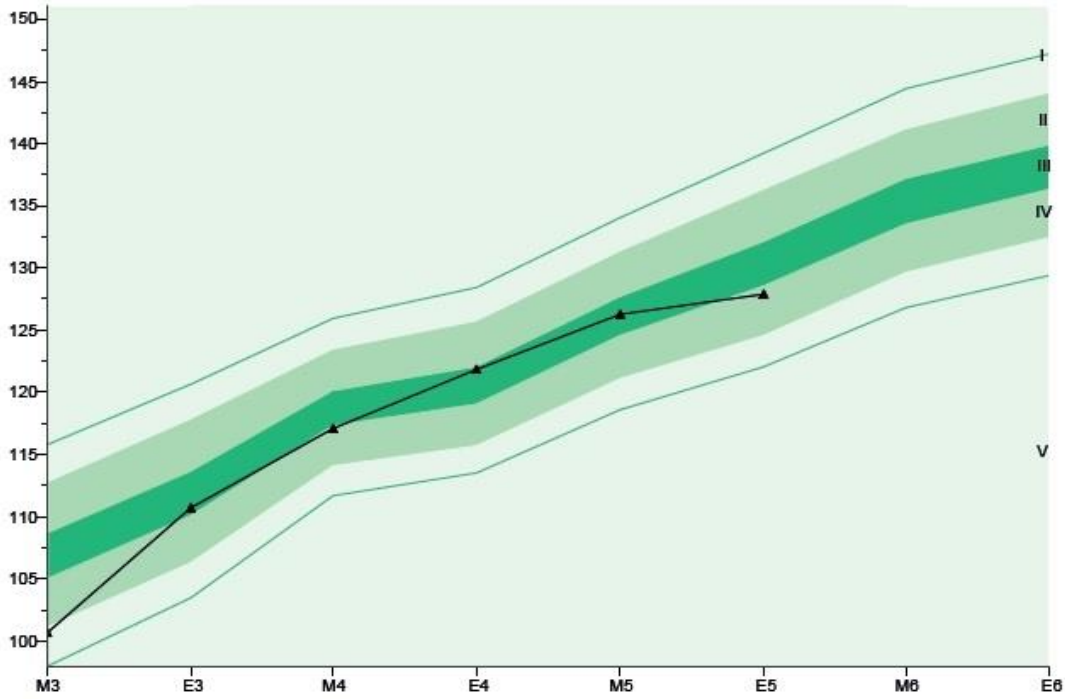
- Hoelang sta je al voor de klas?
- In welke groep geef je momenteel les?
- Hoe lang werk je al met Cito?
- Heb je het gevoel dat je goed met het systeem van Cito kunt werken?
- Welke beslissingen neem je op basis van Cito-scores? (bv instructiegroep)
- Welke andere bronnen (bijvoorbeeld observaties of methode-toetsen) gebruik je verder om beslissingen te maken?

Ik zou je nu graag een aantal vragen willen stellen over een voorbeeld van een leerling rapport. Het gaat om een spelling Cito-score van een willekeurige leerling uit groep 5.
[Nu de leerkracht de afbeelding laten zien]

Leerling rapport spelling

Figuur 1 Voorbeeld van een leerlingrapport

Leerling: **Klaudia Aslan** (5 - 5A) Vergelijking: **Alle leerlingen**
Toets: **Spelling 2011**



Afname datum	Jaar groep	Taak	Toetsscore / Vaard.score	Score interval	Niveau
06-02-2009	4	M3 start + 1	38 / 101	99:102	V
05-06-2009	4	E3 start + 2	37 / 111	109:112	III
09-02-2010	5	M4 start + 2	33 / 117	116:119	IV
21-06-2010	5	E4 start + 2	34 / 122	120:124	III
05-02-2011	5	M5-digi S+2	28 / 126	125:128	III
03-06-2011	5	E5 start + 1	38 / 128	126:130	IV

De volgende vragen gaan over het leerling rapport:

1. Zou je kunnen verwoorden wat je op dit plaatje ziet?
2. Stel dit zou een leerling uit jouw klas zijn, hoe zou je de score op E5 interpreteren?
3. Als je op basis van de score op E5 zou moeten beslissen in welke instructiegroep de leerling komt, wat zou je dan beslissen en waarom?
4. Wat betekent het score interval bij de E5 toets volgens jou (*eventueel aanwijzen*)?
    - En wat betekenen die getallen 126 en 130 dan volgens jou?
5. Hoe neem je dit score interval van de E5 score mee in je interpretatie?

6. Wat betekent volgens jou ‘de onzekerheid’ van een score, en hoe speelt dat een rol in jouw onderwijs (in het algemeen)?
7. En (hoe) neem je dit interval mee in je beslissingen die je maakt op basis van de Cito-scores?

Ik laat nu 4 verschillende plaatjes zien die te maken hebben met de E5 score waar we het net over hadden. Bij elk plaatje stel ik steeds een aantal vragen en dan gaan we door naar de volgende.

Plaatje 1, 2, 3 of 4

1. Wat zie je volgens jou op dit plaatje?
    0. *Leerkracht duidelijk laten verwoorden wat je ziet, beetje sturen als leerkracht vast loopt door vragen te stellen als: wat zegt dit jou?*
    1. *Als leerkracht errorbar niet begrijpt: aanmoedigen*
        - *Wat komt er in je op?*
        - *Wat is je eerst indruk?*
        - *Er is geen goed of fout?*

Als leerkracht <u>echt</u> vastloopt, hier uitleggen wat een score interval is (maar liever niet):
Cito maakt een zo goed mogelijke inschatting van de vaardigheid van de leerling. Het kan zo zijn dat de leerling de toets net iets beter of net iets slechter heeft gemaakt dan dat hij eigenlijk met zijn vaardigheid zou kunnen. In de vaardigheidsscore bij een toets zit dus altijd een foutenmarge. Het score-interval geeft aan dat als de leerling de toets heel vaak opnieuw zou maken, dan zou in 68% van de gevallen de werkelijke vaardigheid van het kind tussen de 126 en de 130 liggen als de leerling 128 punten scoort op de toets.

1. Zou je anders naar de score kijken als je dit plaatje (aanwijzen) erbij had gekregen?
2. Zouden je beslissingen (over bijvoorbeeld een instructiegroep) voor deze leerling veranderen ten opzichte van de beslissingen die je net genomen zou hebben voordat je dit plaatje zag?
3. Als je dit plaatje moet vergelijken met de huidige weergave, waar gaat dan je voorkeur naar uit?
    - Kun je uitleggen hoe dat komt?

Afsluiting

1. Naar welk plaatje gaat je voorkeur uit (plaatje 1, 2, 3 of 4)?
2. Op basis waarvan geef je de voorkeur aan..?

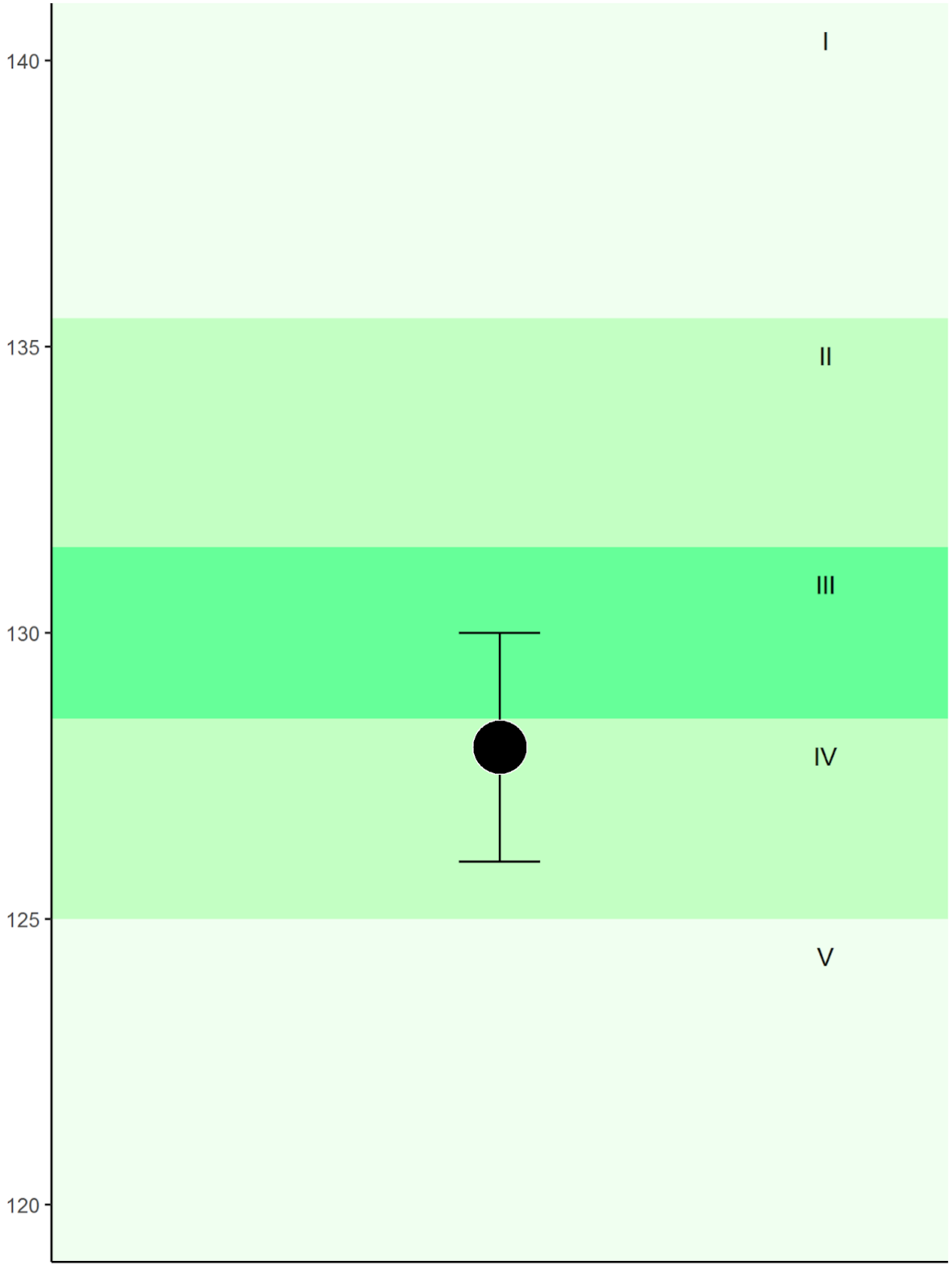
Dat waren de inhoudelijke vragen. Bedankt voor je antwoorden. Is er nog iets dat je kwijt wilt, wat niet aan bod is gekomen?

Dan heb ik nog een aantal korte punten:

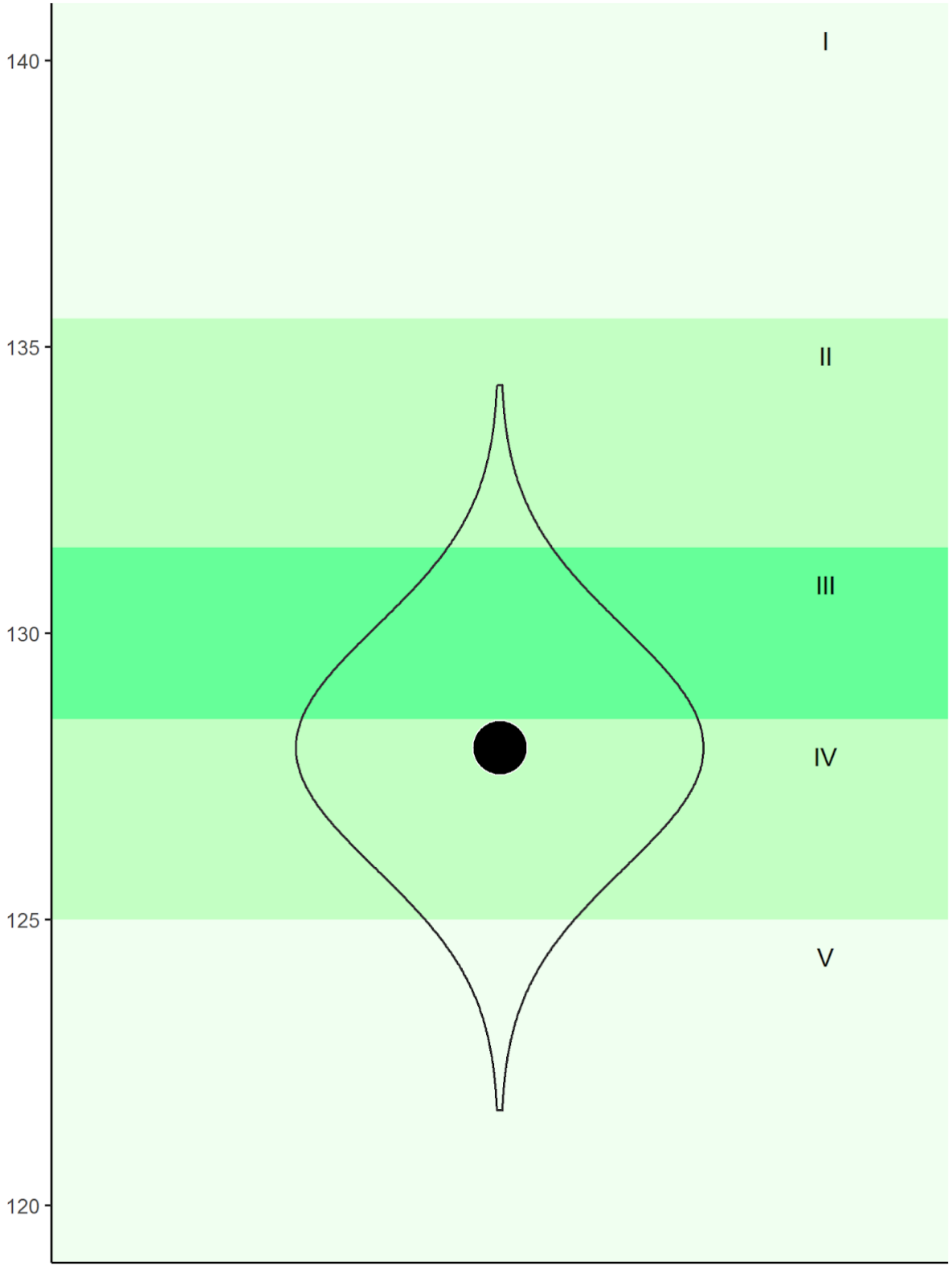
- We gaan de opname helemaal uitschrijven. Wil je daar een kopie van ontvangen?
- Mag ik eventueel contact met je opnemen, mocht er iets onduidelijk zijn?
- Mocht je nog contact met mij willen opnemen, kan dat door mij te mailen. Dan schrijf ik zo mijn e-mailadres even op voor je.
- En tot slot, wil je graag een kopie ontvangen van onze onderzoeksresultaten?

Dan zet ik nu de opname stop. Dankjewel!

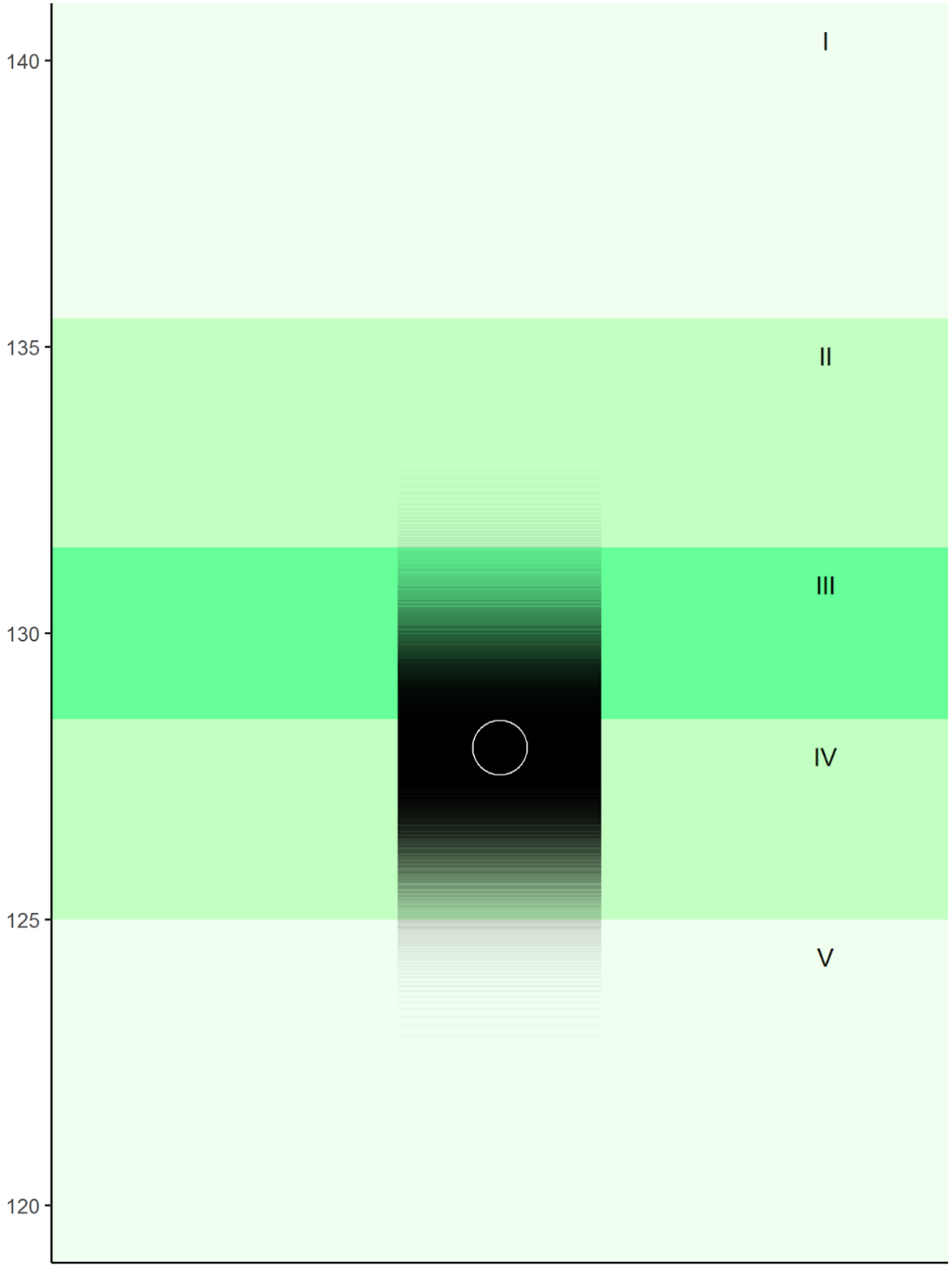
Plaatje 1 (errorbar)



Plaatje 2 (violin plot)



Plaatje 3 (gradient plot)



Plaatje 4 (quantile dotplot)