Omgaan met meetnauwkeurigheid van Cito scores

De interpretatie van verschillende visuele weergavevormen van meetnauwkeurigheid van Cito scores door leerkrachten in het basisonderwijs

Willetta Post (S3733157)
Bachelorwerkstuk academische opleiding leraar basisonderwijs
Faculteit der Gedrags- en maatschappijwetenschappen
Rijksuniversiteit Groningen
PABA-A412
Eerste beoordelaar: dr. N. Frans
Tweede beoordelaar: dr. S. Parlevliet
Juni 2022

Abstract

Cito scores are generally considered as reliable by teachers. In reality there is a 68% confidence interval that indicates that the obtained score involves a certain degree of uncertainty. Visualizing the current score interval could raise awareness of that uncertainty. That is necessary because only 25% of the teachers uses the current score interval. This study looked at the way 12 primary school teachers interpret different visualizations of measurement uncertainty: the *errorbar*, the *violin plot*, the *gradient plot* and the *quantile dotplot*. With semi-structured interviews, the teachers were first presented an example of a student report from a student from group five. Teachers were asked about the interpretation of the score interval displayed in one of the four visualizations of this research. All excerpts from the interview related to the interpretation of the visualizations were coded with the program Atlas TI. The interviews showed that the boundaries of *errorbar* were often interpreted as delineating the area in which the student's score could have been located. The *errorbar* was experienced as clear and well-arranged. The curve of the *violin plot* was frequently interpreted as the mean of the group but sometimes the meaning of it was not very clear because of the large ends. The *gradient plot* was sometimes interpreted as a representation of uncertainty, with it becoming increasingly unlikely to achieve a score as the black color becomes lighter. The *quantile dotplot* seemed difficult to interpret for some teachers because the meaning of the dots was unclear. The results of this study offer the possibility to do further research, for example to find out why teachers appreciate the boundaries of the *errorbar*. With a number of different studies, a visualization can be found that leads to the correct inclusion of uncertainty in the interpretation of Cito scores.

Inleiding

Ongeveer 90% van de basisscholen in Nederland neemt toetsen af volgens het Cito leerlingvolgsysteem (LVS). De toetsen worden een aantal keer per jaar gemaakt door leerlingen van de basisschool en meten de vaardigheid van de leerling voor een bepaald vakgebied. Met het LVS kunnen leerkrachten of andere gebruikers vervolgens toetsresultaten verwerken en automatisch leerlingrapporten, groepsoverzichten en schoolrapporten genereren. In dit proces is nauwkeurige interpretatie van de resultaten van het grootste belang. (Van der Kleij & Eggen, 2013). Besluitvorming op basis van de Cito scores is in het basisonderwijs een dagelijkse bezigheid. Leerkrachten nemen bijvoorbeeld beslissingen over de vervolgstappen in de instructie, de plaatsing van leerlingen in verschillende instructiegroepen of de keuze om een leerling extra ondersteuning te bieden (Hopster-den Otter et al., 2018). Over het algemeen worden dit soort toetsen als zeer betrouwbaar beschouwd (Shepard, 2006), terwijl er in werkelijkheid een onzekerheid bestaat rondom die score (Gardner, 2013). De onzekerheid kan worden omschreven als het verschil tussen de werkelijke score en de geschatte score (Gardner, 2013). Wanneer een meting onafhankelijk wordt herhaald spelen zowel persoonsgebonden factoren als situatie gebonden factoren een rol in de mate van stabiliteit in iemands testscore. Andere factoren zijn enkel gebonden aan de specifieke eenmalig testsessie, daardoor is hun invloed op de herhaalde metingen onvoorspelbaar. De geobserveerde (ofwel geschatte) score kan daarmee worden opgedeeld in een constant of systematisch en een toevallig of niet-systematisch deel (Drenth & Sijtsma, 2006, p. 194). Het systematische deel van de geschatte score is wat we hier aanduiden als de werkelijke score.

Om ervoor te zorgen dat de leerkracht tijdens het interpreteren van scores rekening houdt met het feit dat de toets een schatting van de werkelijke score weergeeft, wordt er gebruik gemaakt van betrouwbaarheidsintervallen die de meetnauwkeurigheid van de score weergeven. Cito maakt gebruik van een 68% betrouwbaarheidsinterval. Een 68% betrouwbaarheidsinterval geeft aan dat 68% van een groep kinderen met een vergelijkbare score, in werkelijkheid een score hebben die valt tussen de grenzen van het betrouwbaarheidsinterval (Charter & Feldt, 2002).

Meijer, Ledoux en Elshof (2011) publiceerden onlangs een rapport over de bruikbaarheid van verschillende leerlingvolgsystemen in het Nederlandse basisonderwijs. De resultaten van dit onderzoek suggereren dat gebruikers van het leerlingvolgsysteem moeite hebben met het interpreteren van de toetsresultaten, wat soms leidt tot onjuiste beslissingen. Daarnaast blijkt het gebruik van de toetsresultaten door leerkrachten beperkt te zijn,

aangezien de interpretatie en analyse van de resultaten voornamelijk uitgevoerd wordt door interne ondersteunende leerkrachten. Ook wist slechts een kwart van de leerkrachten wat het score interval betekent. Het geringe gebruik van het score interval viel specifiek op bij leerkrachten die de groei van de leerling moesten beoordelen (Van der Kleij & Eggen, 2013). Om te zorgen voor de juiste interpretatie is het van belang dat meetnauwkeurigheid goed wordt weergegeven voor leerkrachten zodat ze deze makkelijk kunnen aflezen en er betekenis aan kunnen geven. Dit kan bijvoorbeeld worden gedaan aan de hand van een visualisatie.

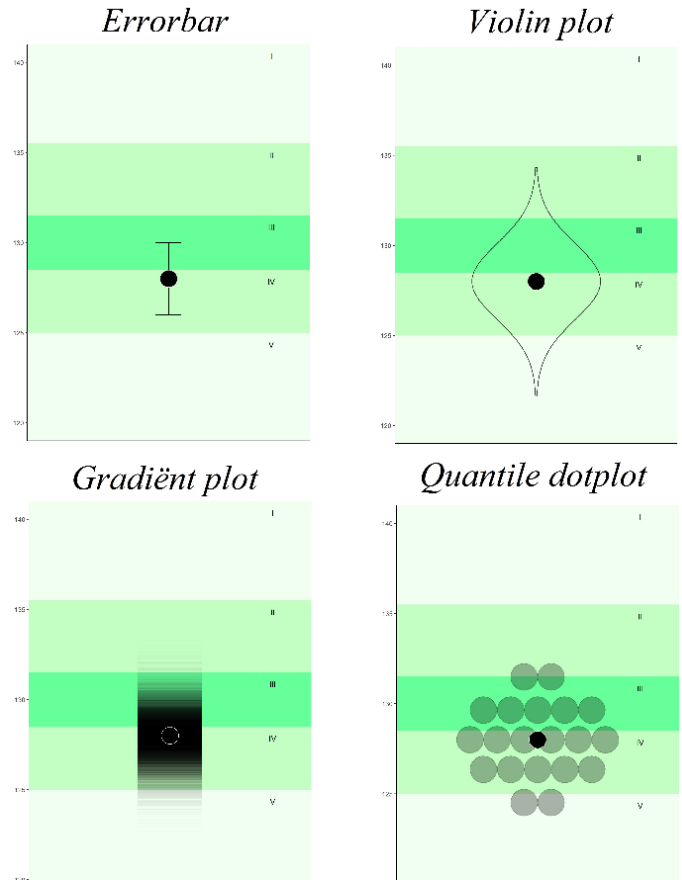
Het meenemen van onzekerheid in de interpretatie van scores kan echter ook leiden tot misvattingen en verkeerde beslissingen, zoals het te hoog of te laag inschatten van een leerling waardoor hij in de verkeerde instructiegroep wordt geplaatst (Hopster-den Otter et al., 2018). Ondanks dat een volledig begrip van onzekerheid bij leerkrachten niet realistisch is, leidt de visualisatie ervan wel tot een groter bewustzijn omtrent onzekerheid van een score ten opzichte van een score zonder visualisatie van de onzekerheid (Hopster-den Otter et al., 2018). Een beter bewustzijn leidt vervolgens weer tot stimulatie van leerkrachten om meer informatie te verzamelen over de vaardigheden van de leerling. Op die manier vindt er besluitvorming plaats op basis van meerdere bronnen. Dat leidt tot verhoging van de validiteit en nauwkeurigheid en kan leiden tot beter onderbouwde beslissingen (Hopster-den Otter et al., 2018).

De juiste interpretatie van onzekerheid is nog belangrijker als de gebruikte toets een relatief grote meetfout bevat, als de behaalde score dicht tegen een bepaalde grenswaarde ligt of als de score bepalend is voor een onomkeerbare beslissing (bijv. het vervolgadvies in groep 8) (Hopster-den Otter et al., 2018). Er bestaat echter geen visualisatie die kan garanderen dat de problemen met interpretatie voor iedereen verbeteren. Vooraf moet dus goed worden nagedacht over de keuze voor het type visualisatie (Padilla et al., 2020). Een combinatie van meerdere visualisaties gebruiken zou nuttig kunnen zijn maar kunnen echter ook leiden tot nieuwe, niet gedocumenteerde problemen (Padilla et al., 2020). Ook Zapata-Rivera et al. (2016) hebben onderzoek gedaan naar het gebruik van visuele weergaves van meetnauwkeurigheid. Uit het onderzoek bleek dat meetnauwkeurigheid beter te begrijpen was voor scores met visuele weergaves van meetnauwkeurigheid ten opzichte van scores zonder visuele weergaves. Wel bleek er extra informatie nodig over de manier waarop de weergave tot stand is gekomen en de betekenis van de verschillende aspecten van de weergave voordat de visuele weergave juist kon worden geïnterpreteerd (Zapata-Rivera et al., 2016).

Vanwege de moeilijkheid die er bestaat rondom het interpreteren van onzekerheid kan er gebruik worden gemaakt van een visualisatie die een grenzen aangeeft (Chance et al., 2004; Hullman et al., 2017). Hiervoor kan bijvoorbeeld een *errorbar* (zie Figuur 1) worden gebruikt. Omdat de *errorbar* een continue weergave toont, is het een geschikte techniek om kwantitatieve gegevens te presenteren (Brodlie et al., 2012; Wainer, 1995). De grootste visuele nadruk ligt bij een *errorbar* op de mate van onzekerheid en de lengte van de balk die wordt bepaald door het type betrouwbaarheidsinterval. Voor een nauwkeurige interpretatie wordt enige statistische kennis van de lezer geëist (Hullman et al., 2011; Zwick et al., 2014). Toch is het een veelgebruikte techniek voor het visualiseren van meetnauwkeurigheid binnen educatieve contexten (Phelps et al., 2010).

Figuur 1

Weergavevormen van onzekerheid



Padilla et al. (2018) stelden echter dat wanneer visuele grenzen, bijvoorbeeld *errorbars* of tekstuele intervallen, worden gebruikt voor continue data, deze grenzen mensen ertoe brengen de data onterecht als categorisch te conceptualiseren. Door een harde grens aan te geven, gaan de kijkers ervan uit dat de specifieke waarde van de grens belangrijk is, vooral wanneer de informatie over hoe de visualisatie tot stand is gekomen ontoereikend is.

Visualisaties van intervallen blijken over het algemeen moeilijk te gebruiken voor zowel experts als leken, zelfs met uitgebreide instructies blijven er fouten bestaan (Belia et al., 2005). In plaats van intervallen te visualiseren, zouden visualisaties die meer eigenschappen weergeven (zoals een *violin plot* (Figuur 1) of *gradiënt plot* (Figuur 1) beter kunnen helpen om de onzekerheid in de data beter te begrijpen. Dit soort visualisaties geven een vollediger beeld van de data doordat eigenschappen zoals de aard van de verdeling en de eventuele uitschieters ook worden weergegeven, deze kunnen verloren gaan met intervallen (Padilla et al., 2020). Visualisaties die frequenties weergeven zijn over het algemeen ook zeer effectief, vooral voor mensen met een lage rekenvaardigheid (Galesic et al., 2009). Gigerenzer (1996) veronderstelde dat onze beslissingen niet toereikend zijn wanneer we verwarrende informatie krijgen, zoals meetnauwkeurigheid die wordt gepresenteerd als percentage (bijv. 10% kans). Frequenties of verhoudingen (bijv. 1 op 10) lijken daarentegen beter te begrijpen (Gigerenzer, 1996).

Kay et al. (2016) creëerden de *quantile dotplot* (Figuur 1) als een alternatief voor het weergeven van meetnauwkeurigheid van een continue variabele. Een *quantile dotplot* laat een verdeling van punten zien waarbij de punten evenredig worden weergegeven met de kwantielen van de verdeling. Elke stip stelt bijvoorbeeld een kans van 5% voor. *Quantile dotplots* zijn getest in verschillende empirische studies, waaruit is gebleken dat ze de variantie van kans inschattingen verminderen in vergelijking met *densityplots*, omdat mensen beter in staat zijn om kleine hoeveelheden te schatten in plaats van oppervlaktes (Kay et al., 2016). Andere studies hebben uitgewezen dat *quantile dotplots* bruikbaarder zijn voor risicovolle beslissingen in vergelijking met interval- en *density plots*, en dat ze significant beter zijn dan tekstuele beschrijvingen van onzekerheid (Fernandes et al., 2018). Correll en Gleicher (2014) vonden dat *violin plots* (waarbij de meetnauwkeurigheid wordt gekoppeld aan de oppervlakte of breedte op een bepaalde y-positie) en *gradiënt plots* (waarbij de meetnauwkeurigheid wordt gekoppeld aan de transparantie) leiden tot meer intuïtieve beoordelingen van de meetnauwkeurigheid van de waarde. Fernandes et al. (2018) vonden in hun onderzoek dat weergave van een *violin- of gradiënt plot* leidt tot kwalitatief betere beslissingen dan weergave van een *errorbar*.

Doel van dit onderzoek is om inzicht te krijgen in de manier waarop leerkrachten in het basisonderwijs bepaalde visuele weergaves van meetnauwkeurigheid interpreteren. De onderzoeksvraag horend bij dit onderzoek is: hoe worden verschillende visuele weergavevormen van meetnauwkeurigheid van Cito scores door leerkrachten in het basisonderwijs geïnterpreteerd?

Methode

Design

De gegevens in dit onderzoek werden verzameld aan de hand van een multiple case study. Doel van dit onderzoek was om inzicht te verkrijgen in de manier waarop verschillende visuele weergaves van meetnauwkeurigheid van Cito scores werden geïnterpreteerd door leerkrachten in het Nederlandse basisonderwijs. Er werd bij dit onderzoek samengewerkt met twee andere onderzoekers die bijdroegen aan het verzamelen van informatie. De twee andere onderzoekers voerden beide ook een eigen onderzoek uit over onzekerheid rondom Cito scores. De doelen van de andere onderzoekers waren om te achterhalen op welke manier basisschoolleerkrachten onzekerheid van Cito scores meenemen in hun interpretaties en beslissingen en wat het verschil in interpretatie van score intervallen is tussen wanneer er een *errorbar* of een schriftelijke weergave getoond wordt.

Steekproef en populatie

Voor dit kwalitatieve onderzoek werden gegevens verzameld door middel van interviews met een selecte steekproef van 12 basisschoolleerkrachten die werken met het leerlingvolgsysteem van Cito. De leerkrachten waren overwegend afkomstig uit het noorden van Nederland, dit was de doelpopulatie waar uiteindelijk uitspraak over gedaan werd. De leerkrachten werden specifiek geselecteerd op basis van de kenmerken geslacht en bouw waarin wordt lesgegeven. Beoogd werd om in deze groep een verdeling te maken waarin een redelijke verdeling bestaat tussen het aantal mannen en vrouwen. Eveneens was het doel om ervoor te zorgen dat er voldoende leerkrachten uit elke bouw (midden- en bovenbouw) aanwezig waren in de steekproef. De onderbouw (groep 1 en 2) werd in dit onderzoek buiten beschouwing omdat er in 2018 werd besloten dat het beleid om te toetsen in het basisonderwijs een aanpassing krijgt zodat er geen schoolse toetsen meer worden afgenomen bij kleuters. In plaats daarvan kan er gebruik worden gemaakt van observaties (Van Engelshoven, 2018). Voor de midden- en bovenbouw werd er onderscheid gemaakt tussen de groepen 3, 4, 5 en 6, 7, 8. Door te selecteren op kenmerken die mogelijk van invloed waren op de uitkomsten (geslacht en bouw) kon er zo breed mogelijk uitspraak gedaan worden over wat mogelijke uitkomsten zouden kunnen zijn voor de doelpopulatie.

Instrument

Er werd gebruik gemaakt van semigestructureerde interviews. De kern van het interview ging over waar leerkrachten beslissingen over maken, waar onzekerheid vandaan komt, hoe deze wordt meegenomen in het maken van beslissingen en wat de weergave van een visualisatie zou kunnen toevoegen aan het interpreteren van onzekerheid rondom een

score. Aan het begin van het interview werden een aantal inleidende algemene vragen gesteld over werkervaring, huidige groep waarin wordt lesgegeven, ervaringen met Cito en bronnen die een rol spelen bij de interpretatie van Cito scores. Vervolgens werden een aantal vragen gesteld over de algemene interpretatie van een leerlingrapport (zie Bijlage E). Dit leerlingrapport bevat een weergave van toetsresultaten van een leerling op een spellingtoets van Cito tot en met het jaar 2011. De score op iedere toets wordt weergegeven met een punt op de grafiek, een toetsscore en een vaardigheidsscore welke wordt uitgedrukt in een score I, II, III, IV of V. Daarbij geven een I en een V de hoogst en laagst scorende 20% van de leerlingen aan. Tevens wordt de onzekerheid rondom de vaardigheidsscore weergegeven aan de hand van een tekstueel score interval. Na de algemene vragen over het leerlingrapport volgden een aantal vragen die specifiek betrekking hadden op de Cito score op de E5 toets van het leerlingrapport. Er werd gekozen voor deze specifieke score vanwege het feit dat deze score zich aan de bovengrens van een IV score bevindt, waarbij het score interval de score omvat die een score III zou uitwijzen, dit zou mogelijk van invloed kunnen zijn op beslissingen die worden genomen op basis van dit scorerapport voor de leerling. Een IV score kan bijvoorbeeld een grenswaarde zijn om een leerling wel of niet in de verlengde instructiegroep te plaatsen. Daarna stelden we een aantal vragen over de interpretatie van onzekerheid weergegeven met een *errorbar* (Bijlage A), een *violin plot* (Bijlage B), een *gradiënt plot* (Bijlage C) en een *quantile dotplot* (Bijlage D). Om te voorkomen dat voorkennis invloed zou hebben op de interpretatie van een weergave werden de weergaves voor iedere leerkracht in een andere volgorde laten zien. Voor alle plaatjes werd steeds gevraagd wat er af te lezen is volgens de leerkracht, of de interpretatie van een score interval anders zou zijn als dit plaatje getoond werd, of dit invloed had op beslissingen die werden gemaakt aan de hand van de Cito score en of de voorkeur uitging naar het behouden van het huidige score interval of het toevoegen van het plaatje aan de huidige weergave. Tot slot werd er gevraagd om een voorkeur voor een van de weergaves van een score interval aan te geven. Het volledige interview protocol is terug te vinden in Bijlage F.

Procedure

Via het sociale netwerk van de onderzoekers werden respondenten per e-mail benaderd om deel te nemen aan het onderzoek. De interviews vonden zoveel mogelijk fysiek plaats op de school van de geïnterviewde leerkracht, tenzij hier geen mogelijkheid voor was. Als de mogelijkheid tot fysieke interviews niet bestond werd er een online afspraak gemaakt. De interviews vonden plaats in de periode van 22 april tot en met 13 mei en duurden ongeveer 20 tot 30 minuten. Op 21 april vond een proefinterview plaats om te meten hoe lang

het interview ongeveer ging duren en waar eventueel nog aanpassingen nodig waren in het interview protocol, het proefinterview duurde ongeveer 20 minuten. Er werden opnames gemaakt van de fysieke en digitale interviews met het programma Teams. De transcripten werden enkel gebruikt voor dit onderzoek en zullen vijf jaar bewaard blijven op de y-schijf van de Rijksuniversiteit Groningen. De opnames werden aan het einde van het onderzoek vernietigd. Voorafgaand aan het interview werden participanten mondeling en schriftelijk geïnformeerd over het doel en hun recht om op elk moment te mogen stoppen met het interview of te mogen aangeven dat de geluidsopname moest worden stopgezet. Daarnaast werd de participanten gevraagd een toestemmingsformulier te ondertekenen voor het opnemen van het interview en het gebruik van de data voor dit onderzoek. De namen van de participanten werden in dit onderzoek vervangen door nummers om de anonimiteit te waarborgen. De ethische commissie PedOn gaf toestemming voor de uitvoering van het onderzoek. De visualisaties (Bijlages A tot en met D) werden ontworpen door scriptiebegeleider Niek Frans in het programma R.

Analyse

De interviews werden getranscribeerd en gecodeerd aan de hand van het programma Atlas TI. Iedere onderzoeker transcribeerde zelf de interviews die door hem waren afgenomen. Alle interviews werden vervolgens door alle drie onderzoekers individueel gecodeerd. Alle fragmenten die betrekking hadden op de interpretatie van de weergaves werden gecodeerd. Na afloop werden de resultaten hiervan nagekeken en besproken met de andere twee onderzoekers. Er werd op een inductieve manier gecodeerd, daarbij werd vooral gelet op het noemen en herkennen van de vaardigheidsscore, de argumentatie voor eventuele veranderingen in interpretatie en beslissingen, het omschrijven van grenzen die betrouwbaarheid aangeven, betekenissen die gekoppeld werden aan specifieke eigenschappen van een weergave en uitspraken over de voorkeur voor weergave van meetnauwkeurigheid van de Cito scores. De beoogde uitkomsten gaven een beeld van de manier waarop verschillende thema’s in uitspraken van leerkrachten aan bod komen bij de interpretatie van Cito scores aan de hand van de getoonde visuele weergaves van meetnauwkeurigheid.

Resultaten

In Tabel 1 is een overzicht te vinden van kenmerken van de participanten. De helft van de participanten was een man, de andere helft vrouw. Zeven leerkrachten vielen in de leeftijdscategorie 20-29, drie in de categorie 30-39 en twee in de categorie 40-49. Vier leerkrachten waren werkzaam in de bovenbouw (groep 6 t/m 8), zeven in de middenbouw (groep 3 t/m 5) en één leerkracht was zowel in de midden als de bovenbouw werkzaam. Er

zat veel variatie in de werkervaring, de minimale werkervaring was één jaar en de maximale 25 jaar. Ook de ervaring in het werken met Cito was erg uiteenlopend. Eén van de leerkrachten werkte minder dan een jaar met Cito maar was al wel 10 jaar werkzaam als leerkracht, eerder hoefde zij nog geen Cito scores in te voeren. De maximale ervaring met Cito was 23 jaar. Vijf leerkrachten hebben tussentijds uitleg gekregen over het begrip score interval omdat ze vastliepen in het geven van antwoorden. Uit de laatste kolom wordt duidelijk in welke volgorde de participanten de verschillende weergaves te zien kregen.

Tabel 1

Kenmerken participanten

Leerkracht	Geslacht	Leeftijds-categorie	Bouw	Werk-ervaring	Ervaring met Cito	Uitleg interval	Volgorde plaatjes*
1	man	20 - 29	midden	5	5	Nee	1-2-3-4
2	man	30 - 39	midden /boven	22	22	Nee	2-3-4-1
3	vrouw	30 - 39	midden	10	< 1	Ja	3-4-1-2
4	vrouw	30 - 39	boven	12	12	Ja	4-1-2-3
5	man	20 - 29	boven	5	2	Ja	1-3-2-4
6	man	20 - 29	midden	1	1	Ja	2-4-1-3
7	vrouw	20 - 29	midden	3	3	Ja	3-2-4-1
8	man	20 - 29	boven	4	3	Nee	4-3-2-1
9	man	20 - 29	midden	3	3	Nee	1-4-3-2
10	vrouw	20 - 29	boven	1	1	Nee	2-1-3-4
11	vrouw	40 - 49	midden	20	15	Nee	3-1-4-2
12	vrouw	40 - 49	midden	25	23	Nee	4-2-1-3

*Noot.** Plaatje 1 is de *errorbar*, plaatje 2 is de *violin plot*, plaatje 3 is de *gradiënt plot* en plaatje 4 is de *quantile dotplot.*

Tabel 2 geeft een overzicht van de onderzochte thema's, de bijbehorende codes en een voorbeeld van een uitspraak van een van de participanten. Voor de thema's werd er onderscheid gemaakt tussen de weergavevormen en de uiteindelijke voorkeur. De codes 'plaatje liever wel erbij' en 'voorkeur huidige weergave' hebben betrekking op het wel of niet

toevoegen van het gevraagde plaatje aan het huidige interval. Het thema ‘voorkeur uit alle vier plaatjes’ geeft een beschrijving van welk plaatje de leerkracht zou kiezen als het interval met cijfers (huidige weergave) zou moeten worden vervangen door één van de plaatjes.

Tabel 2

Code thema’s

Thema	Codes	Voorbeeld quote
Weergavevormen	Beslissingen onveranderd	‘Nee, blijft hetzelfde, de acties zouden hetzelfde zijn.’
	Beslissingen veranderd[1]	‘Ik denk dat ik nu toch meer extra instructie in de klas zou pakken.’
	Beschrijving curve[2]	‘Ja, dat denk ik. En ik denk dat dat een curve is van een leerling waar hij binnen zou moeten vallen.’
	Omschrijving zwarte balk[3]	‘Het ziet eruit als een plaatje die geprint is door een printer die niet goed is.’
	Betekenis bolletjes[1]	‘En die andere rondjes, dat zijn denk ik de categorieën. Waar goed of minder goed is op gescoord.’
	E5 score herkend	‘Ja, de zwarte stip, ik denk dat dat de score is van eind 5 ja.’
	Grenzen betrouwbaarheid	‘En dat hoe minder bolletjes, hoe minder betrouwbaar het wordt, denk ik. Het interval wordt wat kleiner.’
	Interpretatie veranderd	‘Ja misschien wel, ik vind dit wel. Dit plaatje vind ik positiever dan dit.’
	Interpretatie onveranderd	‘Dit had voor mij geen toegevoegde waarde gehad voor het geval.’
	Interval omvat ook andere niveaus	‘Ja, hij zit hier ook tegen niveau 3 aan. Bovenkant van 4’.

	Plaatje liever wel erbij	‘Als ik dit (leerlingrapport) niet had, de grafiek en alleen de getallen, dan had ik liever plaatje twee gehad’
	Voorkeur cijfers	‘Dit (leerlingrapport). Want dat plaatje dat kan ik niet aflezen.’
Voorkeur uit alle vier plaatjes	*Errorbar* (plaatje 1) niet	‘Ja dit bakent het heel sterk af. En, ik denk, ja, kun je dat doen, (…). Ja die duidelijke grenzen, die heb je bij Cito niet.’
	Violin plot (plaatje 2) niet	‘En dit, dit geloof ik niet zo. Dat interval vind ik veel te groot, dus ik kan me niet voorstellen dat dat heel betrouwbaar is.’
	Quantile dotplot (plaatje 4) niet	‘Te veel bolletjes. Dat leidt af van waar het om gaat.’
	Voorkeur *errorbar* (plaatje 1)	‘Plaatje één erbij. Ja, die is gewoon het duidelijkst van hier tot hier is, zeg maar het minimum en het maximale wat de leerling kan halen.’
	Voorkeur *violin plot* (plaatje 2)	‘Ik zou nu voor plaatje twee [*violin plot*] gaan.’
	Voorkeur *gradiënt plot* (plaatje 3)	‘Nou omdat het na jouw uitleg nu wel duidelijk is dat die leerling dan waarschijnlijk daartussen zou kunnen gaan scoren.’
	Voorkeur *quantile dotplot* (plaatje 4)	‘De rondjes die erin staan, want die kunnen iets betekenen. Ja, dat heb ik bij de andere eigenlijk niet. Die vind ik eigenlijk alle 3 gelijk.’

Noot. codes zijn in de analyse uitgesplitst naar weergavevormen.

[1] Code komt alleen bij *quantile dotplot* (plaatje 4) voor
[2] Code komt alleen bij *violin plot* (plaatje 2) voor
[3] Code komt alleen bij *gradiënt plot* (plaatje 3) voor

Over het algemeen gold dat de meeste leerkrachten beslissingen maakten op basis van andere bronnen zoals methodetoetsen en observaties. Daarnaast werden ook eerder behaalde Cito scores meegenomen in de interpretatie om de groei van de leerling te bekijken. Het meenemen van het score interval was op het moment van het interview niet of nauwelijks aan de orde bij alle leerkrachten. Ook werd aangegeven dat een plaatje over onzekerheid in algemene zin van toegevoegde waarde is op het huidige score interval omdat de cijfers niet veel betekenis geven. Een plaatje daarentegen maakt de score, het interval en de bijbehorende niveaus inzichtelijk. Tevens herkenden en benoemden bijna alle leerkrachten vrijwel direct de stip als de E5 score in de vier weergaves.

***Errorbar* (plaatje 1)**

Uit antwoorden van de participanten bleek dat beslissingen op basis van de getoonde Cito score onveranderd zouden blijven als de *errorbar* bij het huidige score interval zou worden gegeven. De grenzen van het interval die de *errorbar* toont werden door de meeste leerkrachten als prettig ervaren. De cijfers 126 en 130 werden herkend en benoemd als de grenzen van het interval door een aantal leerkrachten. Enkele leerkrachten gaven aan dat de interpretatie van de E5 score onveranderd zou blijven vanwege het meenemen van resultaten uit methodetoetsen en observaties bij de interpretatie van Cito scores.

De meeste leerkrachten gaven aan de score op E5 anders te interpreteren als de *errorbar* zouden worden toegevoegd aan het huidige leerlingrapport. Als reden daarvoor werd vaak genoemd dat uit de e*rrorbar* blijkt dat het interval ook andere niveaus omvat en dat de leerling dus ook anders had kunnen scoren.

> Omdat, ja, dit, ik vind dit gewoon. Kijk, van tevoren wist ik niet wat het precies was, dat interval, en ik vind dit eigenlijk wel heel duidelijk, ik denk van, ohja, het kan echt een, nou echt wel een stukje naar boven en een stukje naar beneden. (Leerkracht 3, p. 7)

Ook Leerkracht 11 benoemde dit: “Ja, hij zit hier ook tegen niveau 3 aan. Bovenkant van 4’’ (Leerkracht 11, p. 5). Enkele andere leerkrachten gaven aan dat de interpretatie van de E5 score onveranderd zou blijven.

Negen leerkrachten gaven aan dat ze de e*rrorbar* wel graag erbij zouden willen hebben als aanvulling op de huidige weergave van het score interval. Dit werd beargumenteerd met het feit dat de *errorbar* duidelijk aangeeft in welk niveau de grenzen van het interval zich bevinden zodat het eenvoudiger kan worden afgelezen.

> Ja, en deze eerste grafiek die is als je dan de lijn moet volgen van bijvoorbeeld 130, dan is het heel moeilijk te zien waar die uitkomt. En deze [*errorbar*] is natuurlijk helemaal vergroot en dan kun je precies zien, oh, dit is 130 en dit is zoveel. Dus dat is duidelijker zeg maar. (Leerkracht 3, p. 10)

De overige drie leerkrachten gaven aan toch liever te werken met de huidige weergave van het score interval, Leerkracht 5 zei hierover:

> Omdat je iets meer informatie te zien krijgt dan alleen dit plaatje. Dat is iets.. het geeft iets meer een beeld van wat je gewend bent ook. En hierbij zie je ook niet het verleden van de leerling. En dat vind ik ook wel fijn om te zien om dan ook die onzekerheden misschien een beetje weg te nemen. (Leerkracht 5, p. 6)

Violin plot **(plaatje 2)**

Net als bij de *errorbar* gaven wederom alle leerkrachten aan hun beslissingen niet te veranderen op het moment dat de *violin plot* zou worden toegevoegd aan het huidige score interval. De voornaamste reden was het ongeloof in de weergave vanwege de lijnen die doorlopen tot niveau II en niveau V. Leerkrachten leken het niet realistisch te vinden dat leerlingen zo ver van de behaalde vaardigheidsscore hadden kunnen scoren.

Als antwoord op de vraag wat er zichtbaar wordt in de v*iolin plot* werd vaak verwezen naar de curve. Zo werd de curve door Leerkracht 7 bijvoorbeeld geïnterpreteerd als een groepsgemiddelde.

> Dat buitenste boogje… Ja, ik weet niet zo goed hoe ik het moet uitleggen, maar dat het ongeveer een gemiddelde groep is, denk ik dat dat het breedste stuk is wat leerlingen scoren en dan zie je de twee scores. In de V score is de curve veel smaller, dus dat is een kleinere groep leerlingen denk ik die daarin valt. Dus dat is een soort groepsgemiddelde. (Leerkracht 6, p. 7)

Ook andere leerkrachten benoemden de lijn als een gemiddelde maar vonden het vervolgens lastig om daar een betekenis aan te koppelen. “Is het zo'n gemiddelde lijn? Ik weet het echt niet dit, dit herken ik niet’’ (Leerkracht 10, p. 6).

De onzekerheid van de score die de *violin plot* weergeeft wisten sommige leerkrachten goed te herkennen en te benoemen.

> Nou, ik zie weer een stip bij het niveau waarop het kind zit, scoort, en dan zie je het gaat van breed naar eigenlijk een stuk smaller. En, nou ja, ja, eigenlijk is het ja, zie ik dit weer hetzelfde, denk ik als die andere plaatjes. In het brede gedeelte dus zeer waarschijnlijk dat het kind rond dat niveau ligt. En dat smallere gedeelte waarschijnlijk niet, of die kans is veel kleiner. (Leerkracht 8, pp. 11-12)

Daartegenover stonden leerkrachten die zich geen raad wisten met de weergave. “Dat komt omdat het plaatje niet bekend is. Ik weet ook niet hoe ik hem af moet lezen. (Leerkracht 2, p. 7)

Het grootste deel van de leerkrachten, acht van de 12, gaven aan niets te hebben veranderd aan de interpretatie van de E5 score als de *violin plot* werd getoond naast het huidige score interval. Redenen daarvoor waren het niet begrijpen van de weergave of het onwerkelijk en niet realistisch vinden van de grote uitschieters die de weergave naar boven en beneden heeft.

> Nee. Want dit is wel mooi hoor natuurlijk, maar ja… het zit nu van een hoge II tot een lage V. Dus ik vind dit wel heel random, zeg maar. En ik geloof wel dat jullie dit niet zomaar hebben verzonnen, dat zal wel echt zo zijn denk ik, anders laat je het mij niet zien. Maar als je hier je analyse op moet gaan baseren, daar kun je bijna niet meer mee analyseren. Want misschien is het wel een hoge II. Of toch een hele lage V. Dus of heel goed, of heel slecht. Dus met dit plaatje kan ik niet zo veel. (Leerkracht 4, p. 8)

Slechts vier leerkrachten hebben aangegeven dat het van toegevoegde waarde zou zijn als de *violin plot* wordt toegevoegd aan de huidige weergave van het score interval omdat het meer duidelijkheid geeft dan alleen cijfer en omdat het te maken lijkt te hebben met de onzekerheid van een score.

> Nou, ik zou deze denk ik wel… Klinkt heel stom maar, als ik hem begrijp zou ik denk ik wel bij willen hebben, want het lijkt me wel een hele interessante om…. Deze heeft

> naar mijn idee meer te maken met die onzekerheid, dus dan zou ik deze er wel bij willen hebben als ik hem begrijp. (Leerkracht 9, pp. 7-8)

De andere leerkrachten leken de weergave niet te begrijpen en gaven daarom aan dat het niks zou toevoegen als de *violin plot* zouden worden getoond naast de huidige weergave van het score interval. “Want dat plaatje dat kan ik niet aflezen’’ (Leerkracht 2, p. 7).

***Gradiënt plot* (plaatje 3)**

Aan de zwarte balk in de weergave werden verschillende betekenissen gekoppeld. “Dus hij neigt meer naar de IV, of meer naar de onderkant, meer naar onder het gemiddelde dan naar boven het gemiddelde. Dat zie je bij plaatje 3 [*gradiënt plot*] duidelijker’’ (Leerkracht 1, p. 8). Vooral het zwarte gebied en het vervagen daarvan werd vaak benoemd. Ook in deze weergave werd door veel leerkrachten de E5 score herkend in de weergave als de stip. “Euh, ik denk dat dit de score is, dat is die 128 van die leerling’’ (Leerkracht 7, p. 7). Sommige leerkrachten wisten te benoemen dat het vervagen van de zwarte balk te maken heeft met de onzekerheid van de score en het afnemen van een kans op die score.

> En dat vak is dan wat het ook kan zijn, het score-interval. En dan vervaagt het steeds, dus hoe meer naar de uiteinden, hoe lichter het wordt, hoe onwaarschijnlijker het is dat dat de daadwerkelijke score is. En ik geloof dat dit hetzelfde is als die vorige, maar dan in een ander model. (Leerkracht 4, p. 8)

> Nou, ik neem aan dat dit de… Dat rondje is de score is en dat dat zwarte eromheen dan de betrouwbaarheid is, het gebied waarin het zich bevindt denk ik. (Leerkracht 11, p. 5)

De helft van de leerkrachten gaf aan de score op E5 anders te hebben geïnterpreteerd als de *gradiënt plot* was getoond.

> En hier (huidige weergave) lijkt het wel echt alsof de leerling eronder is waardoor het een onvoldoende is geworden eigenlijk. En een IV is natuurlijk nog geen V-score, dus het is niet een hele dikke onvoldoende. Maar als je die marge in gedachten houdt, dan kijk je er wel anders naar, ja. (Leerkracht 3, p. 6)

De andere leerkrachten die kozen voor een andere interpretatie benoemden net als Leerkracht 3 dat de score nu positiever lijkt uit te vallen als je rekening houdt met het interval die de weergave laat zien. “Nou omdat je nu meer ziet van hij zit tegen de 3 score aan. En hier zie je alleen die dalende lijn” (Leerkracht 11, p. 5). Leerkrachten die ervoor kozen om de interpretatie van de score niet te veranderen vonden de weergave niet duidelijker dan de huidige weergave. “Dus als ik zo een score zou binnenkrijgen, dan had ik daar net zoveel mee gekund als met de score interval op het leerlingrapport, niks eigenlijk” (Leerkracht 12, p. 5).

Ondanks dat de helft van de leerkrachten had aangegeven de score anders te interpreteren als de *gradiënt plot* werd toegevoegd aan de huidige weergave van het score interval zeiden slechts vier leerkrachten dat het iets zou toevoegen als de *gradiënt plot* werd toegevoegd aan de huidige weergave. Uitleg daarvoor was dat deze weergavevorm meer betekenis lijkt te hebben dan enkel de cijfers. “Nou, dit zijn nummers, dat zijn dat, dat zegt niet zoveel. Dit plaatje maakt het duidelijker” (Leerkracht 11, p. 5). De leerkrachten die liever alleen de huidige weergave hebben zeiden dat de *gradiënt plot* onduidelijk of niet prettig af te lezen is. “Dan de kolom. Ik vind deze niet zo niet zo prettig” (Leerkracht 6, p. 11).

***Quantile dotplot* (plaatje 4)**

Bij de *quantile dotplot* werd als enige van de vier weergaves door één leerkracht aangegeven dat beslissingen op basis van de E5 score anders zouden zijn als dit plaatje werd gegeven naast het huidige score interval. “Ik denk dat ik nu toch meer extra instructie in de klas zou pakken. En dat niet, als maar een kleine stijgende lijn zie. Nou ja, voorkomen dat het doorzet” (Leerkracht 12, p. 4). Alle andere leerkrachten zouden weer op dezelfde manier de score interpreteren als voorheen.

> Nou, als ik kijk naar mijn beslissing van zonet, zou ik dat gewoon hetzelfde houden, denk ik. Ik zou gewoon kijken van, hoe doet het kind in de klas? Hoe doet hij zich voor en hoe gaat het met het kind? En op basis daarvan zou ik nu beslissen van nou toch basis of toch intensief? (Leerkracht 8, p. 9)

Over de bolletjes in de *quantile dotplot* werden verschillende uitspraken gedaan. Veel leerkrachten wisten niet of nauwelijks betekenis eraan te geven. Andere leerkrachten interpreteerden de bolletjes als gemiddeldes, kinderen met vergelijkbare scores, andere scores van kinderen uit de klas of categorieën van de spellingtoets.

> Ja… ik zie wel dat dit (wijst y-as aan) de vaardigheidsscores zijn met een bepaalde score. Dus boven de 136 heb je een I. Dit zijn misschien de kinderen van de klas? Dan is het niet zo'n sterke klas. (Leerkracht 4, p. 5)

Een aantal leerkrachten wisten het aantal bolletjes te koppelen aan de betrouwbaarheid rondom de score. "Ja weer hetzelfde. Alleen nu is het in plaats van met een lijn, is met rondjes weergegeven en hoe minder rondjes staan, dan hoe minder groot de kans is dat die score zou worden behaald" (Leerkracht 5, p. 8).

Bijna de helft van de leerkrachten zouden de interpretatie van de E5 score onveranderd laten na toevoeging van de *quantile dotplot*. De leerkrachten die de score wel anders zouden interpreteren hadden als argument hiervoor dat de *quantile dotplot* laat zien dat het interval ook andere niveaus omvat, dus dat het kind ook uit had kunnen komen op een ander niveau op basis van de Cito toets. Van de 12 leerkrachten gaven drie aan het prettig te vinden als de *quantile dotplot* wordt toegevoegd aan het huidige score interval. Eén van hen voegde daar wel aan toe dat het plaatje toegevoegd kan worden mits er vooraf uitleg wordt gegeven over de interpretatie ervan.

> Naar dit plaatje [wijst naar *quantile dotplot*], want dit zegt wel meer wat mij betreft, een score interval, ja als iemand je het uitlegt is het ook prima te doen, maar als het niet uitgelegd wordt dan ja, dan weet je ook niet wat het is dus, maar dan gaat bij voorkeur wel uit naar het plaatje. Het is wel duidelijker. Zeker voor een beginner, denk ik, iemand die begint met het werken met het systeem. (Leerkracht 8, p. 10)

Leerkracht 4 gaf als argument voor het behouden van de huidige weergave dat het huidige interval ook al meer dan genoeg informatie geeft.

> Nee, 126-130. Dit [*quantile dotplot*] is wel visueel ingesteld, dat klopt wel, maar ik denk dan 'ik ben geen kleuter'. Nu ik weet wat het betekent, denk ik, 'ja goh, dat snap ik ook wel'. Dat het er dan net iets boven is of een heel kleine kans dat… Voor mij persoonlijk is dit voldoende. (Leerkracht 4, p. 6)

Voorkeur weergave

Het grootste deel van de leerkrachten, acht van de 12, gaf aan voorkeur te hebben voor de *errorbar* als vervanging van het huidige tekstuele score interval. Leerkrachten gaven

aan dat ze de *errorbar* overzichtelijk en duidelijk vinden onder andere omdat er een boven en ondergrens wordt aangegeven. Deze grenzen lijken makkelijk te interpreteren voor leerkrachten, en werden vaak benoemd als de minimaal en de maximaal te behalen score voor de leerling. “Zijn werkelijke score. In verband met de onzekerheid van een score kan het net zo goed 130 of 126 zijn. Dus in deze range kan zijn daadwerkelijke score zitten” (Leerkracht 4, p.7). “Nou, dat is maximale en het minimale’’ (Leerkracht 10, p. 8). “Het is wat simpeler neergezet dan die andere. Ik vind het andere neemt heel veel plekken in beslag en je ziet allemaal balletjes en lijntjes en dit is gewoon wat makkelijker, inzichtelijk gemaakt” (Leerkracht 6, p. 12). Er was ook een leerkracht die aangaf de *errorbar* juist niet prettig te vinden.

> Ja dit bakent het heel sterk af. En, ik denk, ja, kun je dat doen, en hier zie je dat het een beetje vaag wordt. Ja die duidelijke grenzen, ja, die heb je bij Cito niet, die zijn er niet. Ik denk niet dat je die kunt geven. (Leerkracht 11, p. 7)

Eén leerkracht gaf de voorkeur aan de *violin plot* omdat deze het meeste te maken lijkt te hebben met de onzekerheid van een score.

> Nou, omdat ik zoals ik zeg, ik vond plaatje één vond ik heel overzichtelijk en die lijkt het meest op plaatje drie [*gradiënt plot*] dan in dat opzicht. Dan heeft mijn voorkeur plaatje één [*errorbar*] maar tussen vier [*quantile dotplot*] en twee [*violin plot*], die hebben dus denk ik te maken met die onzekerheid. Dat zou mij meer informatie geven. En dan zou ik dus voor plaatje twee [*violin plot*] gaan. (Leerkracht 9, p. 8)

De grote uiteindes van de *violin plot* werden door een aantal leerkrachten als negatief beschreven.

> Ja, dat zeker. Ja. Maar als je wil weten wat je ermee gaat doen, dan zou ik een plaatje waarbij je voornamelijk III en IV ziet, duidelijker dan wanneer je ook II en V erbij ziet. Want dan zou je ze eigenlijk bijna allemaal erbij kunnen betrekken. (Leerkracht 3, p. 9)

Twee leerkrachten hadden voorkeur voor de gradi*ënt plot* omdat de vervagende grenzen als prettig worden ervaren.

> Doordat het lijkt alsof het zo vaag afgedrukt is, kun je vind ik heel duidelijk zien naar welk niveau het meer neigt… Ja, één [*errorbar*] is op zich ook wel duidelijk hoor. Je ziet de bovenkant van die stok en de onderkant. Je kan in één keer wel zien dat hij verder in de IV zakt dan in de III zeg maar. Maar dat zie je bij deze [*gradiënt plot*] nog iets beter. (Leerkracht 1, p. 9)

Leerkracht 12 koos als enige in dit onderzoek voor de *quantile dotplot* als vervanging van het huidige tekstuele score interval. “De rondjes die erin staan, want die kunnen iets betekenen. Ja, dat heb ik bij de andere eigenlijk niet. Die vind ik eigenlijk alle 3 gelijk” (Leerkracht 12, p. 6). Leerkrachten die de *quantile dotplot* juist niet als voorkeur aangaven leken de weergave te chaotisch en weinigzeggend te vinden.

> Ik word een beetje onrustig van dat dit zes bolletjes zijn, dit vijf en dit vijf, dat twee en dat twee. Ik snap wel waarom, omdat dit die 68 en 32 procent zijn, maar dit [*errorbar*] vind ik veel rustiger ogen. Hier [*errorbar*] zie ik in één oogopslag tussen daar en daar zit waarschijnlijk. (Leerkracht 4, p. 7)

“En dan plaatje 4, dat vind ik echt… nee. Spreekt me ook niet aan. Daar heb ik niks mee. Te chaotisch, ziet er te chaotisch uit” (Leerkracht 2, p. 10).

Discussie

In dit onderzoek werd onderzocht op welke manier leerkrachten in het basisonderwijs bepaalde visuele weergaves van meetnauwkeurigheid interpreteren. Daarbij werd er gekeken naar vier verschillende figuren die de onzekerheid van een score kunnen weergeven. Duidelijke grenzen, zoals bij de *errorbar* werden geïnterpreteerd als prettig wegens de eenvoud en duidelijkheid. Volgens een aantal leerkrachten lijken de duidelijke grenzen goed aan te geven in welk gebied de leerling had kunnen scoren. Wanneer onzekerheid te veel in detail werd weergegeven werd dit door leerkracht soms als onrealistisch bestempeld vanwege het grote gebied in niveaus dat de weergave beschrijft. Dit was bijvoorbeeld bij het zien van de *violin plot* het geval. Leerkrachten wist de onzekerheid van een score vooral te benoemen als de grenzen rondom de score doorlopen tot andere niveaus boven of onder de behaalde score, zoals bij de *violin plot* en de *gradiënt plot*. Daarbij gaven de leerkrachten dat de kans om in het stuk rondom de score te scoren groter is dan de kans om te scoren in de uiteindes van de weergave. Ook de vervaging van bijvoorbeeld de *gradiënt plot* was een eigenschap

die leerkrachten ertoe bracht om een score te kunnen koppelen aan onzekerheid. Er werd benoemd dat hoe lichter het wordt, dus hoe meer je naar de uiteindes van de figuur gaat kijken, hoe meer onwaarschijnlijk het is dat de leerling daar scoort. Wanneer leerkrachten te maken kregen met onbekende eigenschappen van weergaves, zoals de bolletjes van de *quantile dotplot*, werden er verschillende interpretaties gedaan. Sommige interpretaties kwamen in de buurt van de werkelijke betekenis terwijl andere interpretaties niet erg accuraat waren.

Het grootste deel van de leerkrachten gaf aan een voorkeur te hebben voor de *errorbar* als vervanging van de huidige weergave. Voornaamste reden hiervoor waren de duidelijkheid en overzichtelijkheid van de weergave. Slechts een enkeling gaf de voorkeur aan de *violin plot* omdat deze het meeste te maken lijkt te hebben met de onzekerheid rondom een score. Een paar leerkrachten vonden de *gradiënt plot* het meest duidelijk en zouden deze graag als vervanging van het huidige interval zien omdat deze weergave het meest duidelijk maakt naar welk niveau de score neigt. Dit heeft te maken met het feit dat de grenzen van de *gradiënt plot* vervagen en uiteindelijk verdwijnen naarmate de kans op een score afneemt. De *quantile dotplot* werd een enkele keer als voorkeur aangegeven omdat de bolletjes iets zouden kunnen betekenen. Anderen vonden de *quantile dotplot* juist chaotisch en onrustig.

Ondanks dat het niet realistisch is dat leerkrachten onzekerheid volledig begrijpen leidt het weergeven ervan wel tot een groter bewustzijn van het feit dat elke score een bepaalde onzekerheid bevat (Hopster-den Otter et al., 2018). Dit werd ook duidelijk in het onderzoek, na het tonen van weergaves leken leerkrachten te beseffen dat leerlingen ook een andere score hadden kunnen halen. Uit het onderzoek kwam ook naar voren dat sommige leerkrachten de *gradiënt plot* goed wisten te interpreteren aan de hand van de vervaging die de zwarte balk toont. Dit is in overeenstemming met bevindingen van Hopster-den-Otter et al. (2018) die veronderstellen dat vervagende grenzen in een weergave ertoe leiden dat er minder behoefte aan aanvullende informatie is voor het begrijpen van onzekerheid. Vage grenzen zouden ervoor zorgen dat er een besef komt van het feit dat er geen exacte interpretatie van een score bestaat en dat er meer over de onzekerheid van de score wordt nagedacht (Hopster-den-Otter et al., 2018).

Uit het onderzoek van Zapata-Rivera et al. (2016) bleek dat meetnauwkeurigheid beter te begrijpen was als er een visuele weergave werd toegevoegd. Er moest dan echter wel extra informatie worden verstrekt over de figuur voor een juiste interpretatie (Zapata-Rivera et al., 2016). Ook dit kwam naar voren in het onderzoek. Verschillende misvattingen werden gedaan vanwege een gebrek aan verstrekte informatie over de weergaves. De grenzen van de

errorbar leken voor veel leerkrachte betekenisvol te zijn en leidden ertoe dat grenzen onterecht werden gezien als de uiterste waarden van het gebied waarin de leerling had kunnen scoren. Dit komt overeen met de bevindingen van Padilla et al. (2018) die zeggen dat visuele grenzen mensen ertoe brengen data onterecht als categorisch te conceptualiseren en dat het kijkers in de positie brengt om aan te nemen dat de specifieke waarde van de grens van belang is. Weergaves die meer eigenschappen weergaven zouden daarentegen kunnen helpen bij het begrijpen van onzekerheid. Er wordt dan een meer volledig beeld gegeven omdat er rekening wordt gehouden met de verdeling en met uitschieters (Padilla et al., 2020). Dit kwam vooral naar voren bij de interpretatie van de *violin plot* en de *gradiënt plot* waar voor beide weergaves benoemd werd dat de kans om verder van de behaalde score te scores steeds afneemt naarmate je verder naar de uiteindes kijkt. De *quantile dotplot* geeft volgens Kay et al. (2016) een grotere kans op een goede schatting omdat het schatten van hoeveelheden (bolletjes) makkelijker is dan het schatten van oppervlaktes. Uit het onderzoek bleek echter dat leerkrachten moeite hadden met de interpretatie van de betekenis van de bolletjes en dat er verschillende misconcepten werden gedaan. Fernandes et al. (2018) vonden in hun onderzoek dat de weergave van een *violin- of gradiënt plot* leidt tot kwalitatief betere beslissingen dan weergave van een *errorbar*. Dit bleek in het onderzoek terug te komen omdat leerkrachten vanuit de *violin- en gradiënt plot* de onzekerheid van de score op de juiste manier wisten te benoemen terwijl er bij de *errorbar* fouten leken te bestaan in de interpretatie van de grenzen. Volgens Hopster-den-Otter et al. (2018) zou een combinatie van de *errorbar* met de vervagende grenzen van de *gradiënt plot* mogelijk een uitkomst kunnen zijn om onzekerheid weer te geven op een manier die goed te interpreteren is voor leerkrachten.

Tijdens de afname van de interviews merkten de onderzoekers dat leerkrachten op dezelfde manier en steeds korter gingen antwoorden op de vragen die betrekking hadden op de weergaves. Dit kan ermee te maken hebben dat leerkrachten verveeld raakten door de vragen, omdat er bij de weergaves sprake was van herhaling in de vraagstelling. Nadeel hiervan was dat er soms onvolledige of foutieve antwoorden werden gegeven. De weergaves werden echter steeds in andere volgorde laten zien om te voorkomen dat de onderzoeksresultaten over de weergaves ontoereikend zouden zijn. Verder bleek uit het onderzoek dat bij de interpretatie van Cito scores vooral rekening wordt gehouden met voorgaande scores, scores uit methodetoetsen en observaties. Daardoor leken leerkrachten soms te antwoorden dat ze voorkeur hebben voor de huidige weergave. Met de huidige weergave werden echter de cijfers van het score interval bedoeld en niet de andere bronnen

die ook een rol spelen in de interpretatie van scores, in een vervolgonderzoek zou de vraag specifieker gesteld moeten worden. Er zou dan bijvoorbeeld gevraagd kunnen worden of de voorkeur ligt bij het interval in cijfers of bij de weergave als aanvulling op de cijfers.

Een sterk punten van het interview was de opbouw. Doordat er eerst algemene vragen werden gesteld werd er een goed beeld geschetst van de leerkracht, en ontstond er een ontspannen setting. De vragen over het leerlingrapport gaven meer informatie voor de onderzoeker maar ook voor de leerkracht. De plaatjes die daarna getoond werden hadden betrekking op de score die eerder besproken was in het leerlingrapport. Dat maakte het voor de leerkracht duidelijk over welke score de vragen gingen zodat uiteindelijk ook eenvoudiger een vergelijking kon worden gemaakt tussen het huidige score interval en de gegeven weergave.

Het codeschema dat werd gebruikt in dit onderzoek is betrouwbaar en geeft een goede samenvatting van de resultaten. Tijdens het coderen is er rekening mee gehouden dat iedere uitspraak bij een andere weergave hoort. Op die manier zijn er codes ontstaan die specifiek gelden voor één van de vier weergaves. Om het schema meer betrouwbaar te maken zouden codes kunnen worden toegevoegd die een beschrijving geven van algemene uitspraken.

De participanten in de steekproef geven een redelijk beeld van de mogelijke opvattingen van leerkrachten uit het noorden van Nederland in het primair onderwijs. Dit heeft ermee te maken dat er bij het samenstellen van de steekproef rekening is gehouden met de verdeling tussen mannen en vrouwen en de verdeling van de bouw waarin iedere leerkracht werkt. Er waren zowel zes mannen als zes vrouwen in dit onderzoek. Ook in de bouw waarin de leerkrachten werken bestond een redelijke verdeling, vier leerkrachten werken in de bovenbouw, zeven in de middenbouw en één leerkracht werkt zowel in de midden als de bovenbouw. De verdeling in leeftijden was echter niet evenredig, zeven leerkrachten waren 20-29 jaar, drie leerkrachten waren 30-39 jaar en twee leerkrachten waren tussen de 40 en 49. De verdeling in leeftijd kan mogelijkheid van invloed zijn op de resultaten omdat dit verband lijkt te hebben met het aantal jaar ervaring (met Cito). In een vervolgonderzoek is het dus goed om een steekproef samen te stellen waarin ook iedere leeftijdsgroep even sterk wordt vertegenwoordigt.

De vervagende grenzen van de *gradiënt plot* lijken een realistisch beeld te geven van de betrouwbaarheid rondom de behaalde score. Om te kijken of de *gradiënt plot* een goede aanvulling zou zijn op het huidige score interval zou onderzoek gedaan moeten worden naar de mate van verhoging van het bewustzijn van onzekerheid na toevoeging van de *gradiënt* plot. Ook zou er gekeken moeten worden naar de juistheid van de interpretatie van de

weergave. Ook speelt de vraag ‘wat maakt dat een aantal leerkrachten de duidelijke grenzen van de *errorbar* zo belangrijk vinden?’ op. Naar mijn idee vinden leerkrachten het prettig om leerlingen in groepen in te kunnen delen, zodat de lessen op basis daarvan kunnen worden aangepast en ingedeeld. Duidelijke grenzen bieden meer mogelijkheid tot het indelen in groepen ten opzichte van grenzen die niet zo duidelijk afbakenen. Hier zou onderzoek naar gedaan kunnen worden om tevens te achterhalen wat de beweegredenen zijn om de *violin plot* juist af te wijzen vanwege het ontbreken van duidelijke grenzen.

Uit dit onderzoek is duidelijk geworden hoe verschillende leerkrachten de *errorbar, violin plot, gradiënt plot* en *quantile dotplot* interpreteren vanuit een gegeven Cito score met bijbehorend score interval. Er zijn in dit onderzoek bevindingen gedaan omtrent de interpretatie van vier verschillende weergavevormen van een Cito score interval. Zo werd er gevonden dat duidelijke grenzen als prettig worden ervaren en dat vervaging leidt tot het herkennen van betrouwbaarheid in sommige gevallen. Vanuit deze bevindingen zijn er diverse mogelijkheden tot vervolgonderzoek zoals het onderzoeken van het effect van het combineren van visualisaties.

Referenties

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers Misunderstand Confidence Intervals and Standard Error Bars. *Psychological Methods*, *10*(4), 389–396. https://doi.org/10.1037/1082-989x.10.4.389

Brodlie, K. W., & Osoria, R. A. (2012). A review of uncertainty in data visualization. In A. Lopes, J. Dill, R. Earnshaw, D. Kasik, J. Vince, & P. C. Wong (Reds.), *Expanding the frontiers of visual analytics and visualization* (pp. 81–109). London: Springer.

Chance, B., Del Mas, R., & Garfield, J. (2004). Reasoning about sampling distribitions. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–325). Springer.

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), 104–112. https://doi.org/10.1080/07481756.2002.12069053

Correll, M., & Gleicher, M. (2014). Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2142–2151. https://doi.org/10.1109/tvcg.2014.2346298

Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie* (4de editie). Bohn Stafleu van Loghum.

Fernandes, M., Walls, L., Munson, S., Hullman, J., & Kay, M. (2018). Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). New York: ACM.

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, *28*(2), 210–216. https://doi.org/10.1037/a0014474

Gardner, J. (2013). The public understanding of error in educational assessment. *Oxford Review of Education*, *39*(1), 72–92. https://doi.org/10.1080/03054985.2012.760290

Gigerenzer, G. (1996). The Psychology of Good Judgment. *Medical Decision Making*, *16*(3), 273–280. https://doi.org/10.1177/0272989x9601600312

Hopster-den Otter, D., Muilenburg, S. N., Wools, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2018). Comparing the influence of various measurement error presentations in test score reports on educational decision-making. *Assessment in Education: Principles, Policy & Practice*, *26*(2), 123–142. https://doi.org/10.1080/0969594x.2018.1447908

Hullman, J., Kay, M., Kim, Y. S., & Shrestha, S. (2017). Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 446–456. https://doi.org/10.1109/tvcg.2017.2743898

Hullman, J., Rhodes, R., Rodriguez, F., & Shah, P. (2011). Research on graph comprehension and data interpretation: Implications for score reporting. In D. Zapata-Rivera & R. Zwick (Reds.), *Test Score Reporting: Perspectives From the ETS Score Reporting Conference* (pp. 11–45). New Jersey: Princeton University Press.

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016). When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In J. Kaye, A. Druin, C. Lampe, D. Morris, & J. P. Hourcade (Reds.), *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5092–5103). San Jose, Californië: ACM.

Meijer, J., Ledoux, G., & Elshof, D. P. (2011). *Gebruikersvriendelijke leerlingvolgsystemen in het primair onderwijs*. Amsterdam: SCO Kohnstamm Instituut.

Padilla, L., Creem-Regehr, S., Hegarty, M., & Stefanucci, J. (2018). Decision making with visualizations: a cognitive framework across disciplines. *Cognitive Research: Principles and Implications*, *3*(1). https://doi.org/10.1186/s41235-018-0120-9

Padilla, L., Kay, M., & Hullman, J. (2014). Uncertainty visualization. *Wiley StatsRef: Statistics Reference Online*, 1–18. https://doi.org/10.1002/9781118445112

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*. London: Office of Qualifications and Examinations.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Red.), *Educational measurment* (4de editie, pp. 623–646). Westport: American Council on Education and Praeger.

Van der Kleij, F. M., & Eggen, T. J. (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. https://doi.org/10.1016/j.stueduc.2013.04.002

Van Engelshoven, I. K. (2018, juli). *Uitvoering regeerakkoord t.a.v. kleutertoetsen* (Nr. 406). Ministerie van onderwijs, cultuur en wetenschap. https://www.tweedekamer.nl/downloads/document?id=3ebcd10b-d3d3-4825-945a-d70c1853ae41&title=Uitvoering%20regeerakkoord%20t.a.v.%20kleutertoetsen.pdf

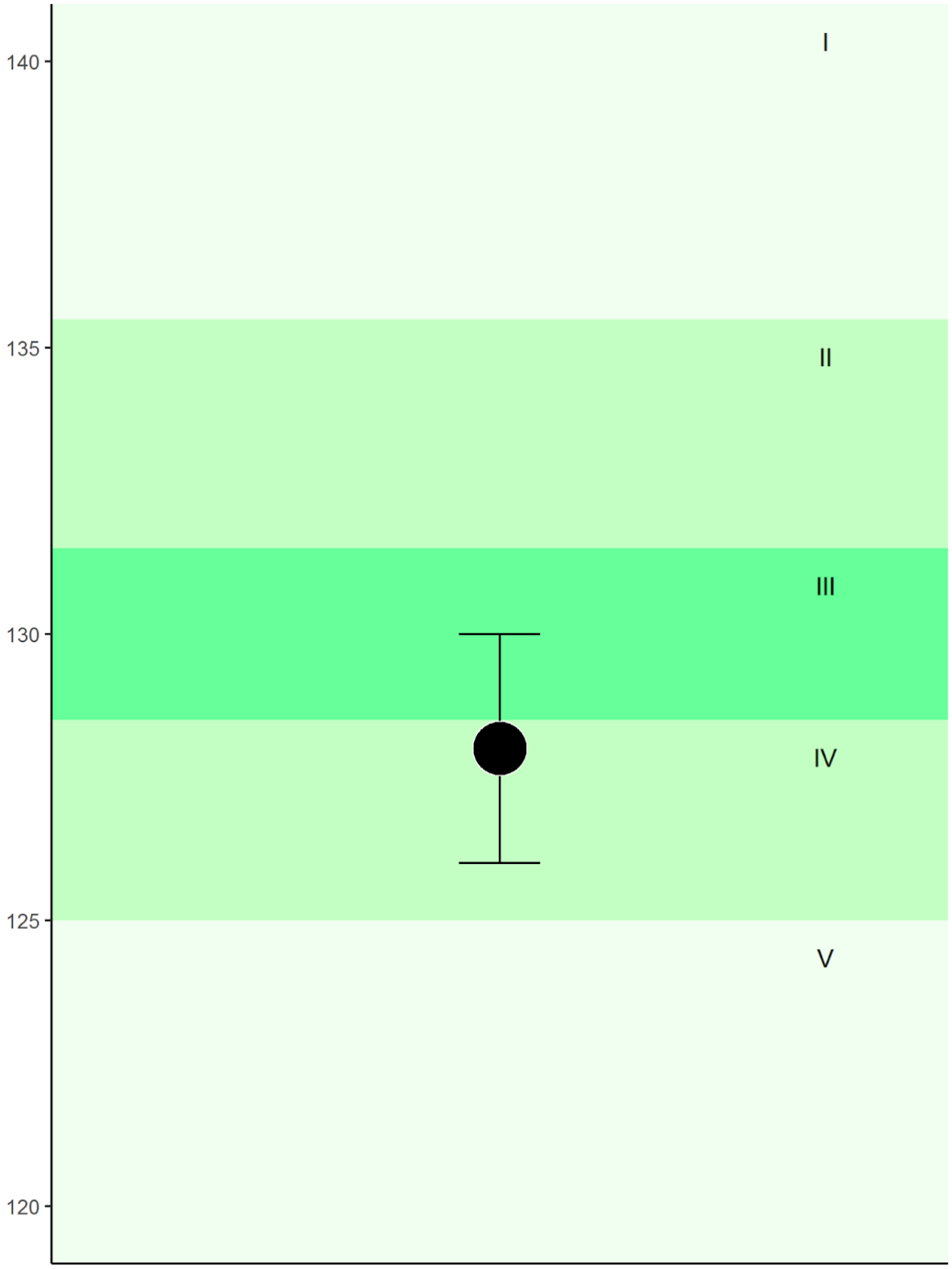
Wainer, H. (1995). Depiciting error. *ETS Research Report Series, 95*(2), 1–14. https://doi.org/10.1002/j.2333-8504.1995.tb01646.x

Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). A review of uncertainty visualization errors: Working memory as an explanatory theory. *Educational assesment, 21*(3), 215–229. https://doi.org/10.1080/10627197.2016.1202110

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing Graphical and Verbal Representations of Measurement Error In Test Score Reports. *Educational Assessment, 19*(2), 116–138. https://doi.org/10.1080/10627197.2014.903653

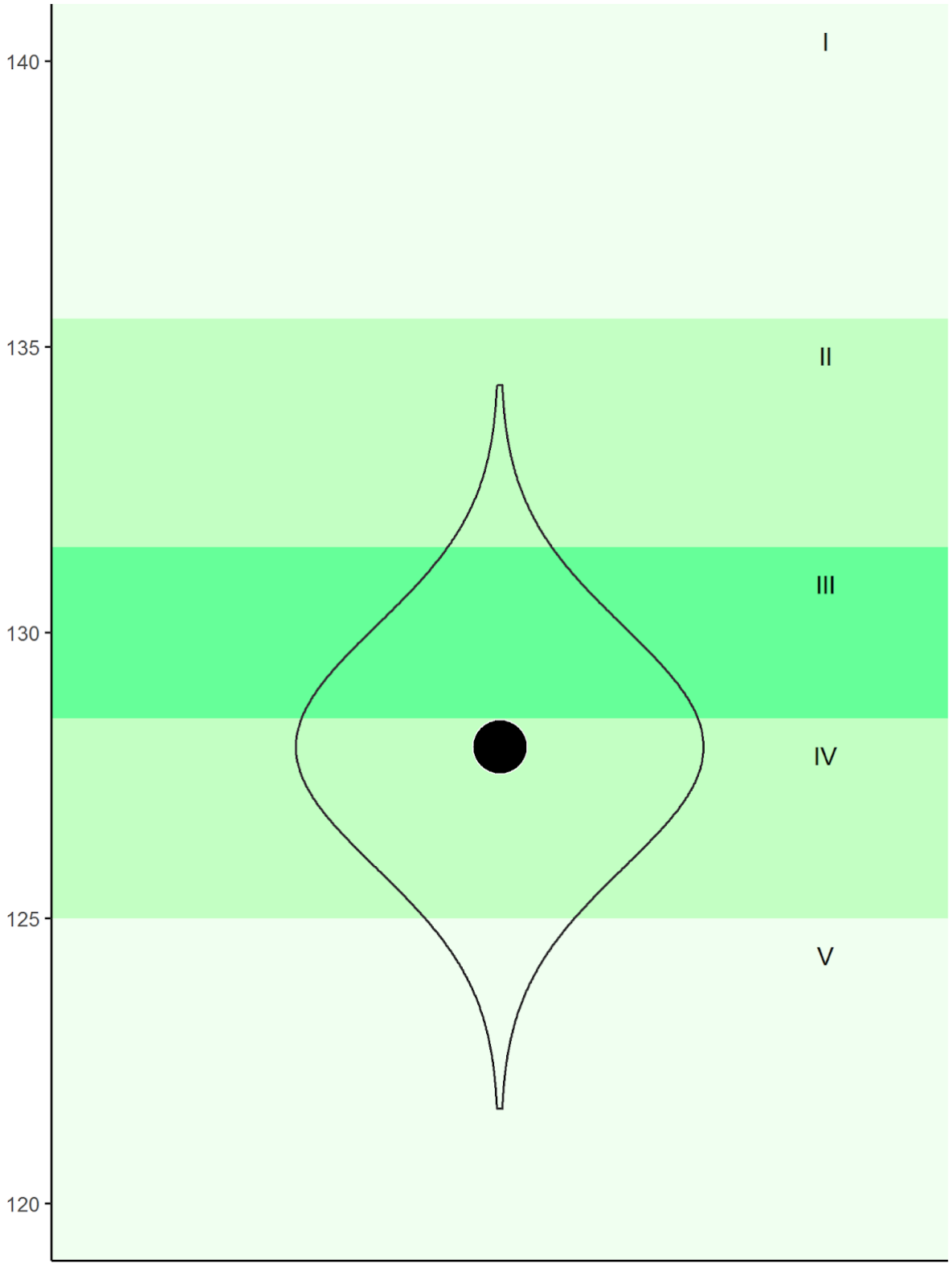
Bijlage A

Plaatje 1 (*errorbar*)



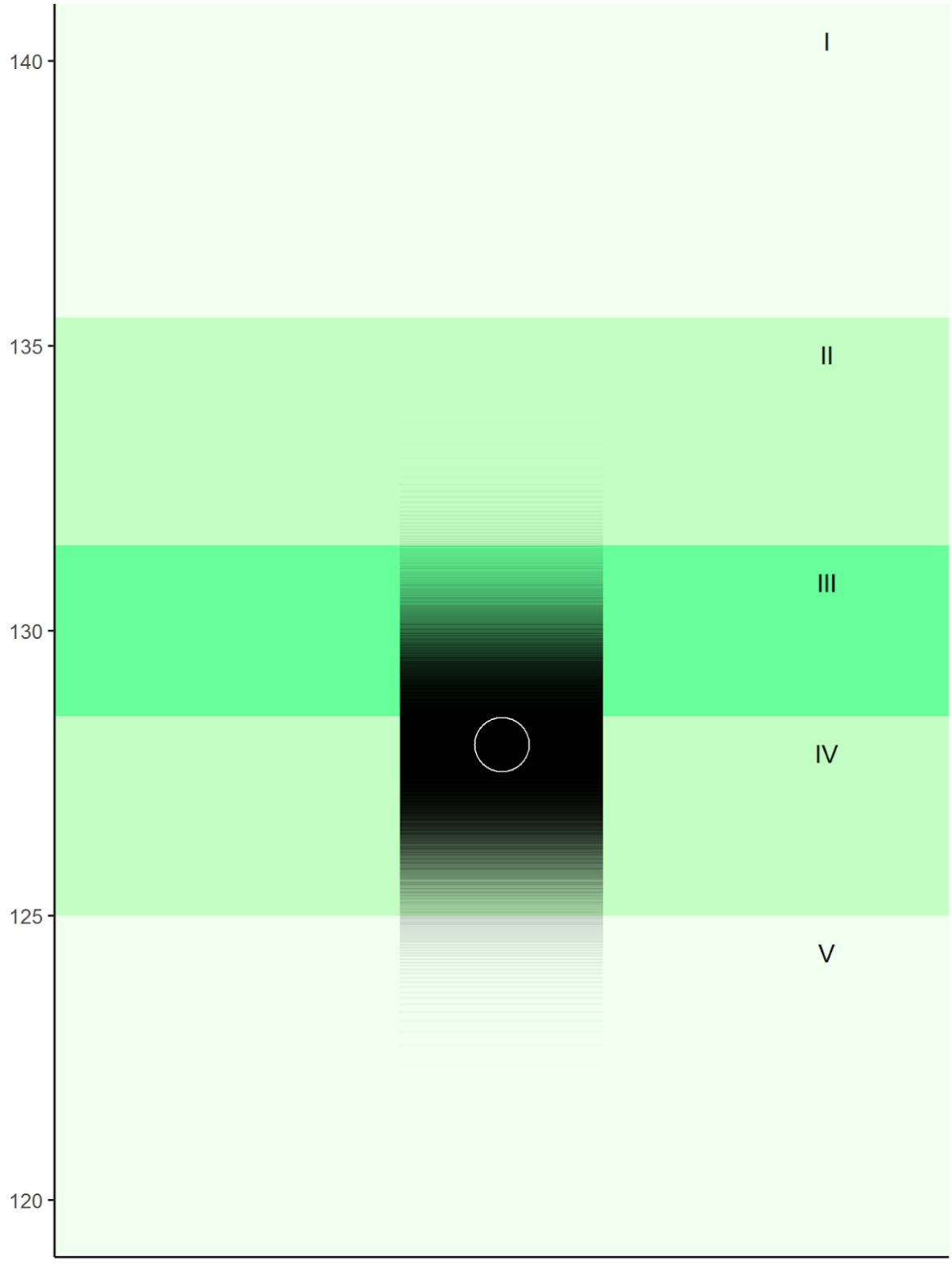
Bijlage B

Plaatje 2 (*violin plot*)



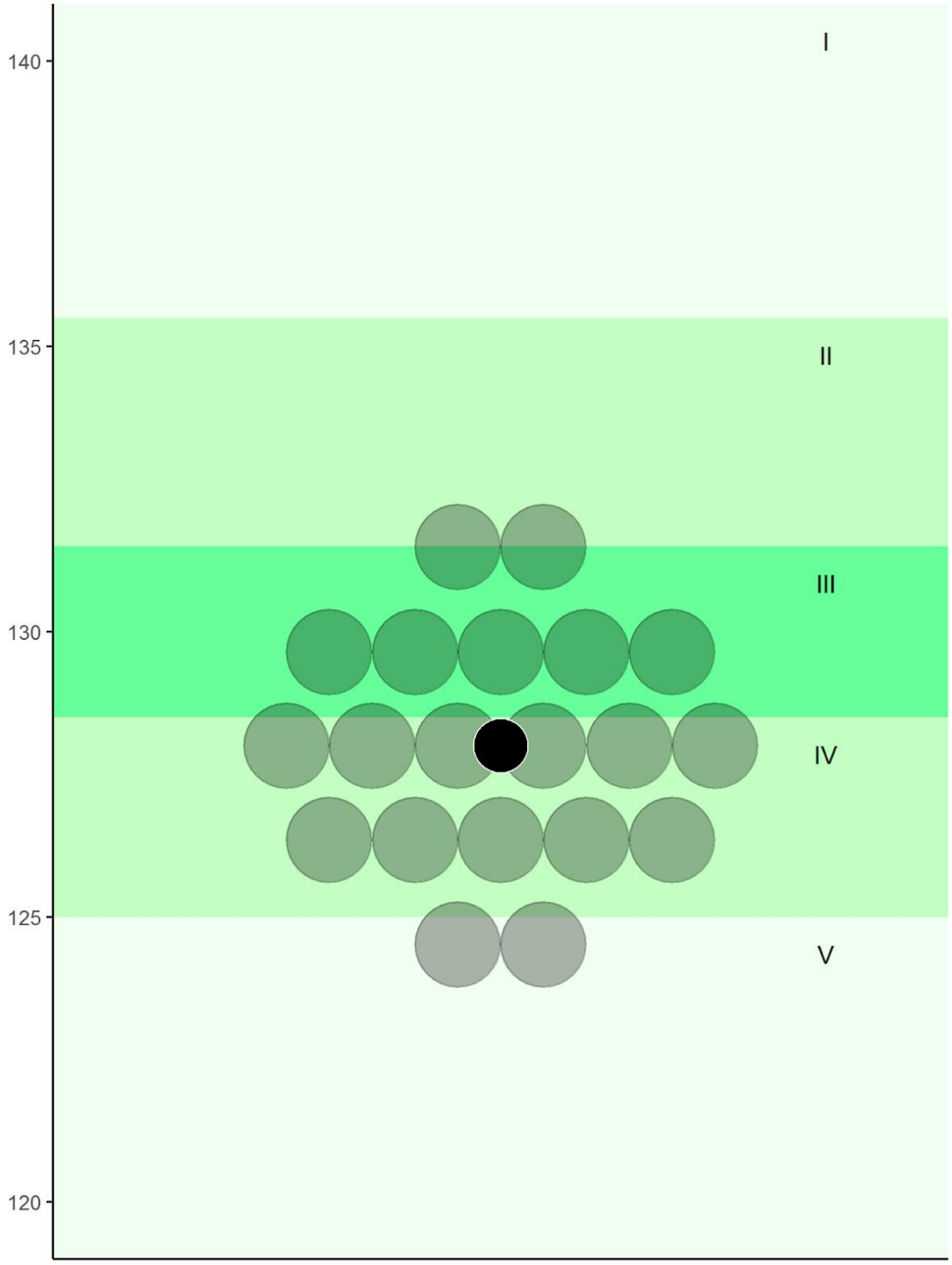
Bijlage C

Plaatje 3 (*gradiënt plot*)



Bijlage D

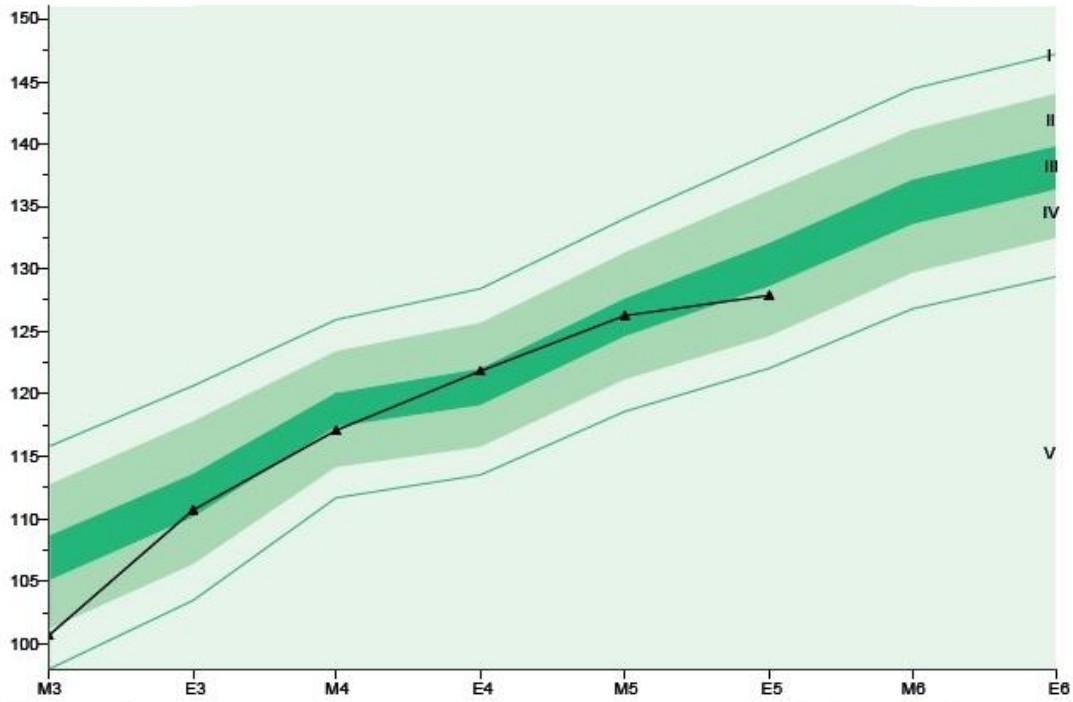
Plaatje 4 (*quantile dotplot*)



Bijlage E

Cito leerlingrapport spelling toets 2011

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

Bijlage F

Interview protocol

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:

1. Hoelang sta je al voor de klas?
2. In welke groep geef je momenteel les?
3. Hoe lang werk je al met Cito?
4. Heb je het gevoel dat je goed met het systeem van Cito kunt werken?
5. Welke beslissingen neem je op basis van Cito scores? (bv instructiegroep)
6. 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 leerlingrapport. Het is een Cito spellingrapport van een Nederlandse leerling uit groep 5.

[Nu de leerkracht de afbeelding van het leerlingrapport (Bijlage E) laten zien]

De volgende vragen gaan over het leerlingrapport:

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 (Bijlage A t/m D) worden nu om en om laten zien in de verschillende volgordes zoals aangegeven]

1. Wat zie je volgens jou op dit plaatje?

Leerkracht duidelijk laten verwoorden wat je ziet, beetje sturen als leerkracht vast loopt door vragen te stellen als: wat zegt dit jou?
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.

2. Zou je anders naar de score kijken als je dit plaatje (aanwijzen) erbij had gekregen?
3. 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?
4. 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 kies je voor de voorkeur van weergave van ……..?

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 meteen 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!