rijksuniversiteit groningen / faculteit gedrags- en maatschappijwetenschappen

Plaatjes kijken voor leerkrachten

Het verschil tussen de interpretatie van een schriftelijk score-interval en een error-bar

Student: S. Wolbers (S2856174)

Begeleider: dr. N. Frans
2e beoordelaar: dr. S. Parlevliet

Rijksuniversiteit Groningen
Faculteit Gedrags- en Maatschappijwetenschappen
Bachelorwerkstuk Pedagogische Wetenschappen (PABA-A412)
Juni 2022

Abstract

Many teachers in primary schools in the Netherlands use the LOVS system to decide upon actions for their students, such as in which instruction group they should be placed. But do those teachers interpret the LOVS data in the right way? Not interpreting the data correctly could have consequences for the students. To gain insight in teachers' interpretation of LOVS data, this multiple case study investigated in which ways the interpretation of primary school teachers differed between a written score-interval and a score interval that was visualized by way of an error-bar when assessing a student report. The study was conducted by interviewing twelve primary school teachers who use LOVS in the Netherlands. The teachers were asked to interpret a written score-interval, and to place a hypothetical student in an instruction group based on that score-interval. After that, teachers were shown a score-interval that was visualized by way of various graphs, amongst which an error-bar, and were asked to indicate whether they would change their decision on the instruction group in which to place the student. The interviews were transcribed, and statements made by the teachers about the score-intervals were coded. Afterwards, important themes that appeared in the interviews were formulated. Results showed that most teachers found the error-bar to be clearer and easier to interpret when compared to the written score interval. Also, teachers were better able to describe the meaning of a score-interval after interpreting the error-bar. Although most teachers indicated that they found the error-bar clearer and easier to interpret, the error-bar did not make the teachers change their decisions on in which instruction group to place the student. Future research could investigate the differences in interpretation between teachers that studied at a university of applied sciences and teachers that studied at a research university.

Inleiding

Op basisscholen in Nederland worden leerlingen in de groepen drie tot en met acht getoetst om hun ontwikkeling te volgen. De meeste scholen gebruiken hiervoor het leerling- en onderwijsvolgsysteem (LOVS), een systeem ontwikkeld door het Centraal Instituut voor Toets Ontwikkeling (Cito) waarbij leerlingen tweemaal per jaar gestandaardiseerde toetsen op het gebied van rekenen en taal maken (Van der Kleij & Eggen, 2013). Deze toetsen meten de reken- en taalvaardigheid van de leerlingen. Nadat de toetsen zijn gemaakt en ingevoerd, genereert LOVS een leerlingrapport met daarop een vaardigheidsscore per toets, welke de beste inschatting vormt die de toets kan maken van de vaardigheid van de leerling in een bepaald leergebied. Ook geeft LOVS het niveau van de leerling weer met een Romeins cijfer van I tot V, waarbij leerlingen die landelijk in de beste 20% scoren een I krijgen, de 20% daarna een II, et cetera. Op een leerlingrapport zijn de opeenvolgende vaardigheidsscores van een leerling weergegeven in een grafiek en met elkaar verbonden door een lijn, waardoor de groei van een leerling te zien is. De testscores die uit LOVS komen kennen echter meetfouten. Dat is niet uniek aan LOVS; meetfouten zijn inherent aan tests. Alle tests hebben meetfouten (Thompson, 2003). Een toets wordt gemaakt door een leerling, maar bij het maken van die toets spelen ook factoren mee die te maken hebben met de persoon en de testsituatie (Drenth & Sijtsma, 2006). Deze factoren spelen mee in hoe een leerling de toets maakt. Meetfouten zorgen op die manier voor een bepaalde onzekerheid in de vaardigheidsscore. Het is belangrijk om deze meetfouten en onzekerheid in een testscore mee te nemen bij de interpretatie van een testscore (Charter & Feldt, 2002), omdat dit ervoor kan zorgen dat leerkrachten betere beslissingen maken naar aanleiding van de toets.

Om de vaardigheid van de leerling beter te kunnen interpreteren geeft LOVS niet alleen een vaardigheidsscore, maar ook een score-interval. Dit score-interval wordt schriftelijk weergegeven door twee cijfers en geeft het 68%-betrouwbaarheidsinterval rondom de score aan, waardoor rekening gehouden kan worden met de meetfouten die inherent zijn aan de test. Wanneer de onzekerheid van scores wordt meegenomen, kan men de score beter interpreteren. Cumming en Finch (2005) spreken ook over de voordelen van het meenemen van betrouwbaarheidsintervallen, maar wijzen ook op de risico's. Zij beschrijven dat velen niet bekend zijn met betrouwbaarheidsintervallen en dat er geen duidelijke richtlijnen zijn voor het interpreteren ervan (Cumming & Finch, 2005). Zelfs veel onderzoekers hebben verkeerde opvattingen over de onzekerheid van scores en weergaves daarvan (Belia et al., 2005). Hoewel Cumming en Finch zich richten op psychologen, beperken deze risico's zich niet tot die beroepsgroep. Andere onderzoekers concluderen dat ook leerkrachten moeite hebben met het

interpreteren van score-intervallen (Meijer et al., 2011; Kennedy-Clark et al., 2020) en dat dit kan resulteren in het maken van de verkeerde beslissingen (Meijer et al., 2011; Van der Kleij & Eggen, 2013), bijvoorbeeld over wat voor soort instructie een leerling nodig heeft of naar welk type vervolgonderwijs een leerling zou moeten gaan. Eerder onderzoek van Hambleton en Slater (1997) liet ook zien dat gebruikers vaak zaken die te maken hebben met betrouwbaarheid links laten liggen wanneer zij een scorerapport interpreteren, omdat ze het niet betekenisvol vinden of omdat ze niet weten wat een score-interval is (Van der Kleij & Eggen, 2013). Het is dus maar de vraag of leerkrachten de meetfouten in de gestandaardiseerde citotoetsen wel meenemen in de interpretatie van de leerlingrapporten. Onderzoek van Meijer et al. (2011), Hambleton en Slater (1997) en Van der Kleij en Eggen (2013) doet vermoeden dat leerkrachten dat misschien niet doen.

Wanneer de onzekerheid van een vaardigheidsscore niet wordt meegenomen bij het interpreteren van de resultaten van een leerling die bovengemiddeld scoort, kan een leerkracht besluiten deze leerling verkorte instructie of gecompliceerder werk te geven. Het kan zijn dat dit besluit niet past bij de werkelijke vaardigheid van de leerling, omdat de werkelijke vaardigheid lager blijkt te liggen dan de vaardigheidsscore weergeeft. Andersom is het mogelijk dat er onterecht een intensief en duur begeleidingstraject wordt gestart omdat de leerling een V-score heeft behaald, terwijl de werkelijke vaardigheid van de leerling dit niet nodig zou maken. Het kan ook zijn dat een leerkracht een besluit neemt naar aanleiding van de groei van een leerling omdat de leerkracht iets waarneemt in de grafiek op het leerlingrapport. Het gevaar van het verkeerd interpreteren van de groei van een leerling is dat een leerkracht (negatieve) groei bij een leerling constateert en daarnaar gaat handelen, terwijl deze (negatieve) groei er in werkelijkheid niet hoeft te zijn.

Uit onderzoek van Demmans, Epp en Bull (2015) blijkt dat visualisering van meetfouten in scores kan helpen bij het interpreteren van testscores. Een voorbeeld van visualisering van meetfouten is de error-bar, die een visuele weergave is van een score-interval. Hopster-Den Otter et al. (2019) kwamen tot de conclusie dat het toevoegen van error-bars aan scores ervoor zorgt dat leerkrachten vaker aangeven meer informatie nodig te hebben dan wanneer er geen weergave wordt gegeven van meetfouten. Hoewel dit geen bewijs vormt voor een beter begrip van meetfouten, geeft dit wel een indicatie dat het toevoegen van een error-bar leerkrachten verder doet kijken dan alleen de vaardigheidsscore. Leerkrachten interpreteren een score met daarbij een error-bar dus anders dan een score waarbij geen interval wordt weergegeven. Waar echter nog geen onderzoek naar is gedaan is of een visuele weergave bij een interval in de vorm van een error-bar anders wordt geïnterpreteerd dan een schriftelijk

weergegeven interval zoals nu in LOVS wordt gebruikt. Het is mogelijk dat de error-bar de interpretatie van het schriftelijke score-interval verbetert. Padilla et al. (2020) lieten namelijk al zien dat visualisaties kunnen resulteren in betere beslissingen. Het is interessant om te onderzoeken of visualisering van het score-interval ook leerkrachten kan helpen om leerlingrapporten uit LOVS beter te interpreteren.

Indien het visualiseren van de score-intervallen die LOVS geeft leerkrachten helpt om de testresultaten beter te interpreteren, kan dit verkeerde beslissingen over de te ondernemen acties naar aanleiding van de score voorkomen. Wanneer een leerkracht het interval goed interpreteert is een leerkracht zich ervan bewust dat de waargenomen (negatieve) groei niet per definitie een accurate weergave hoeft te zijn van de ontwikkeling van de leerling. Met andere woorden: leerkrachten moeten voorzichtig zijn in het doen van uitspraken over de vaardigheid van een leerling. Ook de ontwerper van LOVS, het Cito, onderstreept dit en schrijft dat de verschillen in vaardigheidsgroei tegen de achtergrond van meetfouten moeten worden geïnterpreteerd. Volgens Cito betekent dit dat uitspraken over vaardigheidsgroei wel te doen zijn, maar dat deze uitspraken wel voorzichtig gedaan moeten worden (Tomesen, Engelen & Hiddink, 2019). Uitspraken doen over de groei van een leerling is dus niet eenvoudig. Volgens Cito is de meest simpele manier om te controleren of er daadwerkelijk groei heeft plaatsgevonden het bekijken of er overlap is tussen twee betrouwbaarheidsintervallen. Bij overlap van betrouwbaarheidsintervallen is er geen significant verschil in vaardigheid. Pas als twee betrouwbaarheidsintervallen niet overlappen kan men spreken van een significant verschil in vaardigheid (Hop & Engelen, 2020). Uitspraken doen over de groei van een leerling kan dus wel, mits de scores op de juiste manier worden geïnterpreteerd.

Wanneer een leerkracht in staat is om LOVS-rapporten op de juiste manier te interpreteren, kan de leerkracht beter bepalen welke vervolgstappen hij wil ondernemen op basis van een leerlingrapport. Wanneer het score-interval laat zien dat het niveau van een leerling bijvoorbeeld tussen de II en IV ligt, kan de leerkracht op basis van eigen observaties of de resultaten op methodegebonden toetsen besluiten wat de leerling nodig heeft. Omdat het visualiseren van het score-interval door middel van een error-bar eraan zou kunnen bijdragen dat leerkrachten betere beslissingen maken en de ontwikkeling van leerlingen meer waarheidsgetrouw inschatten, wordt in dit onderzoek getracht een antwoord te formuleren op de volgende vraag: *Op welke manier verschilt de interpretatie van leerkrachten van een interval dat visueel wordt weergegeven door een error-bar ten opzichte van een interval dat schriftelijk wordt weergegeven bij LOVS-scores?*

Methode

Design

Het onderzoek kenmerkt zich als een multiple case study, waarbij drie onderzoekers die aan elkaar verwante onderwerpen onderzoeken samen hebben gewerkt. Bij ieder interview nam één van de onderzoekers het interview af. De drie onderzoekers hebben dezelfde interviewleidraad gebruikt waarin de elementen uit alle drie de onderzoeken aan bod kwamen. Doel van dit onderzoek was het verkrijgen van inzicht in het verschil in de interpretatie van leerkrachten van visueel weergegeven score-intervallen (error-bars) ten opzichte van schriftelijk weergegeven score-intervallen. Een van de andere onderzoekers heeft zich gericht op het meenemen van onzekerheden in beslissingen en interpretaties van citoscores. De derde onderzoeker heeft de interpretatie van verschillende visuele weergavevormen van meetnauwkeurigheid van citoscores onderzocht. Om antwoord te krijgen op de onderzoeksvragen hebben de onderzoekers leerkrachten die gebruikmaken van LOVS eenmalig geïnterviewd.

Steekproef en populatie

In het onderzoek zijn twaalf leerkrachten in het noorden van Nederland die werken met LOVS geïnterviewd. Deze twaalf leerkrachten zijn geselecteerd aan de hand van een selectieve steekproef. Daarbij is gestreefd naar een verdeling waarbij de helft man en de helft vrouw is en waarbij de helft in de onder- en de helft in de bovenbouw werkzaam is. Groep 3 tot en met 5 is daarbij gezien als onderbouw en groep 6 tot en met 8 als bovenbouw. De verdeling in het geslacht en de werkzame bouw van de leerkrachten is aangehouden om een zo valide mogelijke uitspraak te kunnen doen over de doelpopulatie: leerkrachten die werkzaam zijn in het regulier basisonderwijs en ervaring hebben met het door Cito ontwikkelde LOVS. Groep 1 en 2 zijn buiten beschouwing gelaten, omdat daar over het algemeen met observatie-instrumenten wordt gewerkt om de ontwikkeling van leerlingen te volgen en dat een ander soort van testen is (Van Engelshoven, 2018). Vanaf 2023 kunnen scholen geen gebruik meer maken van schoolse toetsen (Rijksoverheid, 2022).

Onderscheid is specifiek gemaakt tussen onder- en bovenbouwleerkrachten, omdat beslissingen over de vervolgstappen die de leerkracht neemt naar aanleiding van de testscore kunnen verschillen. Immers: in de bovenbouw wordt naar een advies over het niveau van vervolgonderwijs toegewerkt, waar een onderbouwleerkracht daar nog niet mee te maken heeft. Bovendien komen in de bovenbouw de zogenaamde functioneringsniveaus en streefniveaus die de leerlingen moeten halen in beeld. In de onderbouw is dat nog niet aan de

orde, omdat daar met alle leerlingen richting het streefniveau wordt gewerkt. Dit leidt ertoe dat leerkrachten in de bovenbouw andere keuzes kunnen maken naar aanleiding van de testscores dan een onderbouwleerkracht. Gezien de procedure was de toegankelijke populatie van het onderzoek geringer en heeft deze zich beperkt tot leerkrachten in het noorden van Nederland binnen het persoonlijke en professionele netwerk van de onderzoekers.

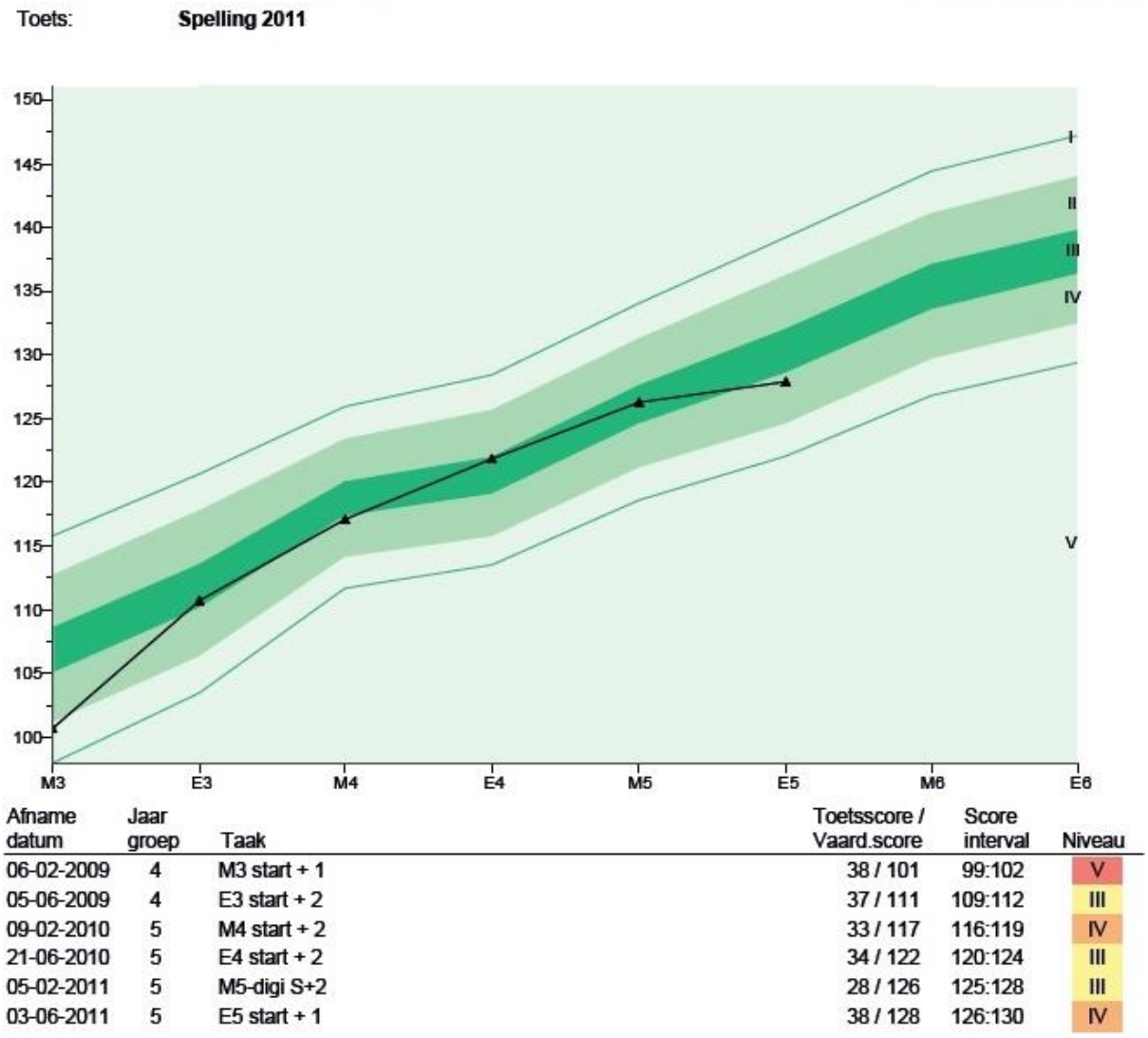
Instrument

De resultaten zijn verkregen aan de hand van een semigestructureerd interview. Dit gaf de onderzoekers houvast en het zorgde ervoor dat de interviews goed met elkaar te vergelijken waren, omdat dezelfde elementen in de interviews aan bod kwamen vanwege de vooraf opgestelde vragen. Het bood echter ook de ruimte voor verduidelijkings- en verdiepingsvragen. Het interview bestond uit drie delen. In het eerste deel is een aantal algemene vragen gesteld waardoor de onderzoekers een beeld kregen van de leerkracht en wat deze met de uitkomst van de LOVS-toetsen doet. In het tweede deel stond een leerlingrapport centraal en is gevraagd naar de interpretatie van het schriftelijk weergegeven score-interval. Het leerlingrapport is weergegeven in Figuur 1. Er is gekozen voor het specifieke leerlingrapport in Figuur 1, omdat de laatste vaardigheidsscore van 128 zich in het IV-gebied bevindt, maar de bovenkant van het score-interval in het III-gebied kwam.

In het laatste deel van het interview kwamen alternatieve visuele weergaves, waaronder de error-bar, aan de orde. Naast de error-bar kregen de leerkrachten een violin plot, een quantile dotplot en een gradient plot te zien. Iedere leerkracht heeft deze visuele weergaves in een andere volgorde onder ogen gekregen. Dit is gedaan om te voorkomen dat het zien van een bepaalde visualisatie invloed zou hebben op de interpretatie van een andere visualisatie. In ieder interview kwam aan de orde: hoe de leerkracht een vaardigheidsscore interpreteert, hoe de leerkracht een schriftelijk weergegeven score-interval interpreteert en in welke instructiegroep de leerkracht deze leerling zou plaatsen op basis van het leerlingrapport en het score-interval. Daarna kwam aan bod hoe de leerkracht verschillende visueel weergegeven score-intervallen interpreteert en of deze visueel weergegeven intervallen de beslissing over de instructiegroep deden veranderen. Aan het einde van het interview is expliciet gevraagd de schriftelijke weergave en de error-bar met elkaar te vergelijken. De volledige interviewleidraad is opgenomen in bijlage 1 en de verschillende visuele weergaven van het score-interval zijn opgenomen in bijlage 2 tot en met 5.

Figuur 1

Leerlingrapport



Procedure

In de selectie van leerkrachten is zowel het persoonlijke als het professionele netwerk van de onderzoekers ingezet. Leerkrachten zijn door de onderzoekers per mail benaderd met de vraag of zij deel wilden nemen aan het onderzoek, waarbij ook het doel van het onderzoek duidelijk is gemaakt: inzicht krijgen in hoe leerkrachten kijken naar onzekerheid in scores. Bij de selectie van leerkrachten is erop gelet dat niet meer dan twee leerkrachten van dezelfde school zijn geïnterviewd. Hiervoor is gekozen omdat het mogelijk is dat leerkrachten van dezelfde school ook dezelfde opvattingen hebben en het in het belang van het onderzoek was om inzicht te krijgen in de opvattingen van verschillende leerkrachten.

De interviews vonden plaats op de school waar de leerkracht werkzaam is. Indien de situatie dat niet toeliet, werden de interviews online afgenomen via Microsoft Teams. Bij iedere

leerkracht nam één van de drie samenwerkende onderzoekers het interview af. De interviews hebben plaatsgevonden tussen 20 april en 13 mei 2022 en duurden ongeveer dertig minuten. Met behulp van het informed consentformulier dat iedere leerkracht heeft getekend is toestemming gevraagd of er een geluidsopname mocht worden gemaakt van het interview, zodat het interview kon worden getranscribeerd en gecodeerd. Ook is toestemming gevraagd voor het gebruik van de data. Het doel van het onderzoek is telkens voor aanvang van het interview mondeling herhaald.

De tijdens het onderzoek verkregen gegevens zijn geanonimiseerd. In de verslaglegging zijn geen namen van respondenten of scholen genoemd. Leerkrachten die geïnterviewd zijn, zijn genummerd, waardoor de resultaten niet herleidbaar zijn tot individuen. Gegevens worden voor een periode van vijf jaar bewaard op de beveiligde Y-schijf van de Rijksuniversiteit Groningen en worden daarna vernietigd, met uitzondering van de geluidsopnames. Deze zijn na transcriptie vernietigd. De ethische commissie van de afdeling Pedagogische en Onderwijswetenschappen van de Rijksuniversiteit Groningen heeft toestemming verleend voor de uitvoering van het onderzoek.

Analyse

Nadat de interviews waren afgenomen zijn deze door de onderzoeker die het interview afnam getranscribeerd. De andere onderzoekers hebben gecontroleerd of dit op de juiste manier is gedaan. Iedere onderzoeker heeft vervolgens elk interview gecodeerd met behulp van het programma Atlas.TI. De onderzoekers maakten gebruik van inductieve codering. Dit gaf de mogelijkheid om zaken die in de interviews naar voren kwamen samen te brengen en zo een conclusie te trekken over de interpretatie van schriftelijke versus visueel weergegeven score-intervallen. In dit onderzoek zijn fragmenten gecodeerd. Daarbij is vooral gelet op hoe de leerkrachten een leerlingrapport en een schriftelijk weergegeven interval interpreteerden, hoe zij een interval met error-bar interpreteerden en wat de verschillen daartussen waren. Ook zijn fragmenten gecodeerd waarin de leerkrachten uitspraken deden over acties die voortvloeiden uit een bepaalde interpretatie en waarin leerkrachten uitspraken deden over het gebruik van citotoetsen. Tot slot zijn uitspraken waarin leerkrachten aangaven dat een error-bar wel of niet verduidelijkt gecodeerd.

De beoogde uitkomst van het onderzoek was een beeld van de verschillen in hoe leerkrachten kijken naar schriftelijk weergegeven intervallen enerzijds en intervallen die weergegeven worden met error-bars anderzijds. Daartoe zijn uitspraken van leerkrachten geclusterd in thema’s. Omdat het onderzoek inzicht beoogde te krijgen in de verschillen in

interpretatie tussen de diverse weergaven van score-intervallen en de gevolgen daarvan, zijn de thema's aan de hand daarvan geformuleerd. Ook is bij het formuleren van de thema's gelet op onderwerpen die in veel interviews aan bod kwamen. Na het coderen en controleren van de interviews zijn de transcripten ter controle toegestuurd aan de leerkrachten die aangegeven hadden dat te willen. Er zijn geen leerkrachten die hebben aangegeven zaken te willen wijzigen. De resultaten die voortkwamen uit de interviews zijn beschreven en ondersteund met citaten van uitspraken die leerkrachten deden. Bij verwijzingen naar uitspraken is steeds het nummer van de leerkracht genoteerd en het nummer van de pagina van het transcript waar de leerkracht de uitspraak deed.

Resultaten

Participanten en thema's

In het onderzoek zijn twaalf leerkrachten geïnterviewd. In Tabel 1 zijn de kenmerken van de verschillende leerkrachten uiteengezet.

Tabel 1

Kenmerken Participerende Leerkrachten

Leerkracht	Bouw	Geslacht	Leeftijd	Werk-ervaring (jaren)	Ervaring LOVS (jaren)	Uitleg interval
1	Onderbouw	Man	20-29	5	5	Nee
2	Onder- en bovenbouw	Vrouw	50+	22	22	Nee
3	Onderbouw	Vrouw	30-39	10	< 1	Ja
4	Bovenbouw	Man	30-39	12	12	Ja
5	Onderbouw	Man	20-29	5	2	Nee
6	Bovenbouw	Man	20-29	1	1	Ja
7	Onderbouw	Vrouw	20-29	3	3	Ja
8	Bovenbouw	Man	20-29	4	3	Nee
9	Onderbouw	Man	20-29	3	3	Nee
10	Bovenbouw	Vrouw	20-29	1	1	Nee
11	Onderbouw	Vrouw	40-49	20	15	Nee
12	Onderbouw	Vrouw	40-49	25	23	Nee

Naast de kenmerken geslacht en bouw is in Tabel 1 de leeftijdscategorie, het aantal jaren werkervaring als leerkracht en het aantal jaren ervaring met het gebruik van LOVS opgenomen. Tevens is per leerkracht aangegeven of in het interview een uitleg is gegeven van de betekenis van een score-interval. Opvallend is dat een groot deel van de leerkrachten in je jongste leeftijdscategorie, tussen de 20 en 29 jaar, valt. Voor de meeste leerkrachten geldt dat hun aantal jaren werkervaring als leerkracht overeenkomt met het aantal jaren ervaring met LOVS. Een duidelijke uitzondering hierop vormt leerkracht 3, die ondanks tien jaar werkervaring minder dan een jaar ervaring heeft met LOVS. Dit is omdat deze leerkracht van school is gewisseld en de vorige school waar deze leerkracht werkzaam was niet met LOVS werkte. Ook leerkracht 5 heeft minder dan de helft van zijn tijd als leerkracht gewerkt met het systeem. De helft van de geïnterviewde leerkrachten is man en vijf van de twaalf leerkrachten is werkzaam in de bovenbouw.

Uit de interviews die met de leerkrachten gehouden zijn kwamen 52 verschillende codes naar voren die samengevoegd zijn onder negen thema's. Tabel 2 geeft een overzicht van de verschillende thema's weer. Onder de codes vielen uitspraken over wat leerkrachten doen met de resultaten van citotoetsen, hoe leerkrachten een leerlingrapport interpreteren, hoe zij naar een schriftelijk score-interval en naar een score-interval weergegeven door een error-bar kijken, welke acties daaruit volgen en of de error-bar het score-interval verduidelijkt. Het volledige codeschema inclusief voorbeelden van uitspraken die bij de verschillende codes horen is opgenomen in bijlage 6.

Tabel 2

Overzicht Belangrijkste Thema's

Thema	Voorbeeld
Gebruik cito algemeen	'Op basis van citoscores ga je wel een plan bedenken voor de komende periode van een aantal maanden. En daarbij zie je dan inderdaad wel bepaalde instructiegroepen inderdaad.'
Interpretatie leerlingrapport	'Ja, het is een hoge IV dus… net onder gemiddeld.'
Actie leerlingrapport	'Maar ik zou hem zeker meenemen aan de instructietafel.'
Interpretatie score-interval	'Ja, dat zal wel de scores zijn waar ze tussen zitten. Tussen moeten zitten, zou je zeggen.'

Actie score-interval	‘Ook al zou het [niveau van de leerling] net op die voldoende [III-score] zitten, dan wil ik eigenlijk dat hij of zij nog steeds een beetje hoger gaat scoren.’
Interpretatie error-bar	‘Ik denk dat dit dat score-interval is wat je daarin ziet. Dat is wat je hier ziet. Dat dit de betrouwbaarheid aangeeft.’
Actie error-bar	‘Nou, ik denk dat ik die leerling alsnog wel in een instructie-groepje mee zou nemen.’
Error-bar verduidelijkt interval wel	‘Dan had ik liever deze er wel bij gehad.’
Error-bar verduidelijkt interval niet	‘Interval zegt me niks en plaatje zegt me ook niks.’

Gebruik leerlingrapport

In alle interviews is aan bod gekomen wat de leerkrachten met een leerlingrapport doen. Negen van de twaalf geïnterviewde leerkrachten gaf in het interview aan de uitslagen van citotoetsen te gebruiken voor het indelen van verschillende instructiegroepen: ‘Nou, wij gebruiken voornamelijk de IV en V-scores voor, nou ja, voor de mindere leerlingen om het zo maar te noemen, dus die komen inderdaad in instructie groepjes en daar schrijven we ook leerling kaarten voor.’ (Leerkracht 6, p. 2)

Om niveaugroepen in te delen, kijken leerkrachten niet alleen naar het niveau dat een leerling haalt. Alle twaalf leerkrachten gaven aan in hun interpretatie van het leerlingrapport ook de ontwikkeling van de leerling mee te nemen. Voor een deel van de leerkrachten speelde de ontwikkeling mee in hun beslissing om de leerling in te delen in een bepaalde niveaugroep.

> [Interviewer: (…) waar zou je deze leerling dan plaatsen?] Dat hangt af van hoe de rest van de groep scoort, maar ik denk aanpak 1. Dus de extra groep zegmaar. (…) [Interviewer: Is dat omdat het een IV is, of omdat er niet genoeg groei is, of…?] Ik denk beide. Het kind scoort nu een IV en een jaar geleden nog een hoge III… (Leerkracht 1, p. 4)

Twee leerkrachten gaven aan de resultaten van de citotoetsen niet te gebruiken voor het indelen van niveaugroepen. Leerkracht 7 en leerkracht 12 gebruiken daarvoor de resultaten die de leerlingen bij de methodegebonden toetsen en de methodelessen behalen.

Interpretatie en gebruik score-interval

Alle twaalf leerkrachten gaven aan het score-interval dat bij de vaardigheidsscore op een leerlingrapport staat op dit moment niet te gebruiken: ‘Heb ik eigenlijk nooit wat mee gedaan.’ (Leerkracht 11, p. 4) Een leerkracht – de leerkracht die minder dan een jaar werkervaring had met LOVS – liet tevens blijken nog nooit van het score-interval te hebben gehoord: ‘Nou dit schema, dit onderste deel, zit bij mijn schema er niet bij.’ (Leerkracht 3, p. 4) Dit had ermee te maken dat deze leerkracht alleen de grafieken opvraagt uit LOVS en niet de tabellen die eronder staan.

Wanneer leerkrachten gevraagd werd naar hoe deze het score-interval rondom de vaardigheidsscore van 128 interpreteerde, gaf iedere leerkracht als eerste reactie niet te weten wat het score-interval was. Bij de vraag wat het score-interval bij de E5-score op het leerlingrapport betekende, zijn de antwoorden van respectievelijk leerkrachten 8 en 9 dan ook vergelijkbaar met de antwoorden die de andere leerkrachten gaven: ‘Zo, Dat is een goede vraag, Ik heb echt geen idee. Score-interval is… nee.’ (Leerkracht 8, p. 6) ‘Nee, die ken ik niet, ken ik niet. Zegt me ook niks.’ (Leerkracht 9, p. 3)

Bij het doorvragen naar wat het score-interval zou kunnen betekenen als ze het goed bekeken, werden verschillende antwoorden gegeven. Vier leerkrachten gaven aan te vermoeden dat het interval de verwachte score van een leerling aangeeft. Leerkracht 7 omschreef dit als volgt: ‘Nou misschien dat je verwacht dat het 130 scoort, maar het scoort 126 dus net iets minder dan wat je verwachting is?’ (Leerkracht 7, p. 5) Leerkracht 4 meende dat het interval het gebied van de IV-score omschreef: ‘Ik denk als je tussen de 126 en de 130 zit qua vaardigheidsscore, dan zit je in de IV.’ (Leerkracht 4, p. 4) Drie leerkrachten bedachten dat het te maken zou kunnen hebben met het aantal gemaakte of het aantal goed beantwoorde vragen.

Drie leerkrachten kwamen met hun interpretatie in de buurt van de juiste betekenis van een score-interval. Zij meenden dat een leerling ook hoger of lager had kunnen scoren op de toets. Geen van hen noemde direct de onzekerheid van testen. Hoewel leerkracht 8 in eerste instantie aangaf geen idee te hebben wat een score-interval betekent, kwam deze wellicht het meest dichtbij: ‘Nee, ik denk dat er een marge op zit qua score, dat denk ik. Dat misschien – het kind had misschien ook iets hoger kunnen scoren, maar dan binnen die range zeg maar.’ (Leerkracht 8, p. 6)

Omdat geen van de leerkrachten het score-interval gebruikte, werd na de interpretatie ervan gevraagd of de beslissing over vervolgstappen anders zou zijn wanneer het score-interval wel was meegenomen. Alle leerkrachten gaven aan dat de beslissing over een instructiegroep

niet zou veranderen wanneer zij het score-interval wel zouden meenemen. Ook wanneer de betekenis van een score-interval wel was uitgelegd.

> [Interviewer: Als je dat nu weet, zou je dan verder dezelfde beslissing maken voor het kind? Als je net vertelde?] Ehm, ja. Want ik vind het nog steeds dat die leerling op het randje tussen een onvoldoende [IV-score] en een voldoende [III-score] zit. Dus ook al zou het net op die voldoende zitten, dan wil ik eigenlijk dat hij of zij nog steeds een beetje hoger gaat scoren. Zodat het een ruime voldoende is en ze het ruim kunnen halen. Anders wordt leerstof ook te moeilijk voor die leerling. (Leerkracht 3, p. 5)

Interpretatie en gebruik error-bar

Wanneer zij de error-bar te zien kregen, legde een aantal leerkrachten de koppeling met het score-interval en het begrip betrouwbaarheid. Dit geldt zowel voor leerkrachten aan wie de betekenis van een score-interval was uitgelegd, zoals leerkracht 4, als voor leerkrachten aan wie de betekenis niet was uitgelegd, zoals leerkracht 8.

> Nou, wat hier [quantile dotplot] bolletjes zijn, zijn hier [error-bar] deze dingen [interval error-bar]. Dus dit is zijn vaardigheidsscore [wijst stip in het midden aan], maar vanwege het score-interval 126-130, kan het net zo goed daar zitten. [Interviewer: Wat kan net zo goed daar zitten?] 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)

> Kijk, van tevoren wist ik niet wat het precies was, dat interval, en ik vind dit [wijst error-bar aan] eigenlijk wel heel duidelijk. Ik denk van ‘oh ja, het [niveau] kan echt wel een stukje naar boven en een stukje naar beneden. (Leerkracht 8, p. 15)

Negen van de twaalf leerkrachten gaven desgevraagd aan de error-bar duidelijker te vinden dan het schriftelijke score-interval. In de transcripten is zeventien keer de code ‘error-bar duidelijker dan score-interval’ toegepast. Als reden dat de error-bar duidelijker is, werd vooral aangevoerd dat het interval gevisualiseerd wordt.

> [Interviewer: kun je ook uitleggen hoe dat komt? Dat je plaatje 1 [het niveau van de leerling] duidelijker vindt maken?] (…) Bij het [leerlingrapport] (…) staan allemaal scores bij interval. Maar dat zegt niet zo veel, die scores. En als je hier naar kijkt, naar [error-bar], dan kun je heel duidelijk zien dat het net onder de lijn van een III-score zit, binnen de IV-score, en dat er een marge zit naar III en naar IV (…) dus dat is duidelijker weergegeven in plaats van in getallen. [Interviewer: Dus het feit dat het een plaatje is maakt het duidelijker dan het getal?] 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 [error-bar] is natuurlijk helemaal vergroot en dan kun je precies zien, oh, dit is 130 en dit is zoveel. Dus dat is duidelijker zegmaar. (Leerkracht 3, p. 10)

Leerkracht 12 noemde ook het punt dat de error-bar een duidelijk minimum en maximum heeft als voordeel: ‘die is gewoon het duidelijkst van hier [wijst onderkant error-bar aan] tot hier [wijst bovenkant van error-bar aan] is, zeg maar het minimum en het maximale wat de leerling kan halen.’ (Leerkracht 12, p. 10) Zes andere leerkrachten gaven ook aan dat ze het relatief smalle interval van de error-bar ten opzichte van andere visualisaties meer duidelijkheid vonden geven. Over een andere visualisatie zei leerkracht 3: ‘Alleen bij [de violin plot] vind ik het onduidelijk eigenlijk dat er ook nog een stukje bij de II en de V zit. (Leerkracht 3, p. 8)

Score-interval en error-bar: geen andere actie

Negen leerkrachten gaven aan dat de error-bar het score-interval verduidelijkt. Drie leerkrachten zeiden ook dat de error-bar de interpretatie van het score-interval deed veranderen en dat zij anders zijn gaan kijken naar de behaalde score van de leerling: ‘[Bij het leerlingrapport dacht ik dat het] echt om een niveau IV ging? Dit interval past bij niveau IV score, hier zie ik dat dat helemaal niet het geval is.’ (Leerkracht 9, p. 5) Dit deed de beslissingen van de leerkrachten over in welke instructiegroep een leerling geplaatst zou worden niet veranderen. Vier leerkrachten gaven op basis van het leerlingrapport en het score-interval aan de leerling in de basisgroep te willen plaatsen. Deze leerkrachten bleven daar na het zien van de error-bar bij. Zes leerkrachten kozen er na het zien van het leerlingrapport en het score-interval voor de leerling in de zorggroep te plaatsen. Hoewel zij soms na het zien van de error-bar anders tegen de score van de leerling zijn gaan aankijken, koos geen van hen ervoor om de

leerling in een andere instructiegroep te plaatsen. Alle leerkrachten bleven na het zien van de error-bar daarmee bij hun beslissing.

Discussie

In dit onderzoek is onderzocht op welke manier de interpretatie van leerkrachten van een schriftelijk weergegeven score-interval en een score-interval weergegeven door een error-bar bij LOVS-scores van elkaar verschillen. Uit het onderzoek blijkt dat de meeste leerkrachten in de steekproef een score-interval weergegeven door een error-bar als duidelijker ervaren dan een schriftelijk weergegeven score-interval. Ook interpreteren de meeste leerkrachten de betekenis van een score-interval beter bij het zien van een error-bar dan wanneer zij een schriftelijk weergegeven score-interval moeten beoordelen. Deze duidelijkheid en betere interpretatie zorgden er bij de geïnterviewde leerkrachten echter niet voor dat hun beslissing over in welke instructiegroep de leerling zou moeten worden geplaatst werd veranderd.

Een aanzienlijk deel van de leerkrachten gebruikt de citotoetsen om leerlingen in te delen in instructiegroepen en laat het handelen dus afhangen van de prestaties van de leerlingen op de citotoetsen. Dit laat het belang van een juiste interpretatie van de resultaten zien. Als de resultaten niet op de juiste manier worden geïnterpreteerd, kan dat dan ook voor veel leerlingen gevolgen hebben.

Het is daarom van belang om uit te zoeken of de error-bar helpt bij het beter interpreteren van het score-interval op een leerlingrapport. De meeste geïnterviewde leerkrachten gaven aan de error-bar duidelijker te vinden dan het schriftelijke score-interval, waarbij vaak als reden werd aangegeven dat dit kwam door het visualiseren van het interval en het hebben van een duidelijke boven- en onderkant. Meerdere leerkrachten koppelden de vaardigheidsscore aan het niveau van de leerling. Doordat de error-bar het score-interval visualiseert was voor de leerkrachten duidelijker te zien dat het niveau van de leerling ook hoger of lager kan liggen dan de vaardigheidsscore en wat dat betekent voor het niveau. Dit is in lijn met de bevindingen van Hopster-den Otter et al. (2019) die concludeerden dat leerkrachten de error-bar waarderen vanwege de exacte presentatie van de uiteinden van het score-interval. Een interpretatiefout die werd gemaakt is dat de boven- en onderkant van de error-bar stonden voor het minimum en maximum dat een leerling zou kunnen halen. Hoewel de boven- en onderkant van de error-bar inderdaad de uiteinden van het score-interval weergeven, betekent dat niet dat de vaardigheid van de leerling daar niet boven of onder kan liggen. Het score-interval geeft immers het 68%-betrouwbaarheidsinterval weer en dat betekent dat 32% van de aannemelijke scores buiten dat interval ligt. Dat de leerkrachten de

error-bar duidelijker vinden dan het schriftelijke score-interval betekent dus niet dat zij het score-interval juist interpreteren.

De meeste leerkrachten die geïnterviewd werden gaven niet alleen aan de error-bar duidelijker te vinden, maar kwamen na het zien van de error-bar ook dichter bij een juiste interpretatie van het score-interval. Veel leerkrachten zagen in dat de error-bar te maken heeft met de onzekerheid van scores. Dat in eerste instantie slechts drie van de twaalf leerkrachten het score-interval enigszins juist kon interpreteren, komt overeen met de conclusies van Van der Kleij en Eggen (2013). Zij toonden aan dat veel leerkrachten niet de kennis hebben om scorerapporten juist te interpreteren. Het is de vraag of de betere interpretatie na het zien van de error-bar komt doordat de error-bar duidelijker is of doordat de error-bar later in het interview aan bod kwam en de onderzoeker al langer met de leerkracht sprak over de onzekerheid van scores.

Leerkrachten gaven bij de error-bar aan dat deze duidelijker was, maar deden dat bij de andere visuele weergaven niet altijd. Dit geeft een indicatie voor dat de betere interpretatie daadwerkelijk te wijten was aan het toevoegen van de error-bar. Aan de andere kant blijkt uit onderzoek van Kannan et al. (2021) dat ouders scorevariabiliteit beter gaan beoordelen wanneer zij meer informatie hebben gehad over onzekerheid van scores en het kan zijn dat dit ook voor leerkrachten geldt. De betere interpretatie kan daarmee ook komen doordat de onderzoeker al uitspraken over onzekerheid had gedaan of doordat bepaalde leerkrachten al andere visualisaties hadden gezien. De leerkrachten waarbij de error-bar als derde of vierde aan bod kwam hadden al geruime tijd met de onderzoeker over de onzekerheid van scores gepraat. Het kan zijn dat een eerder getoonde visualisatie de error-bar verduidelijkte.

Om aan te kunnen tonen dat de betere interpretatie komt door het toevoegen van de error-bar zou onderzoek uitgevoerd kunnen worden waarin twee groepen leerkrachten worden gevraagd om score-intervallen te interpreteren. Daarbij zou één groep een schriftelijk score-interval beoordelen en de andere groep een score-interval gevisualiseerd door een error-bar. Wanneer beide groepen evenveel informatie krijgen en geen andere weergavevormen te zien krijgen, kan uitgesloten worden dat dat de redenen zijn voor de betere interpretatie van een error-bar.

Drie leerkrachten gaven aan dat de error-bar hun interpretatie van het score-interval ook daadwerkelijk deed veranderen, hoewel dat er niet voor zorgde dat leerkrachten hun beslissingen aanpasten. Zij bleven telkens bij hun standpunt over in welke instructiegroep de leerling geplaatst moest worden. Dit duidt erop dat leerkrachten, zelfs wanneer zij voldoende begrip hebben van een score-interval en het score-interval wordt ondersteund door een error-

bar, hun beslissingen baseren op het niveau I-V en de vaardigheidsscore en groei van de leerling. De interpretatie kan dus anders zijn, maar dat zorgt niet voor andere acties. Dit is verrassend, omdat het onderzoek van Padilla et al. (2020) deed vermoeden dat de beslissingen zouden worden verbeterd bij visualisering van score-intervallen. Hierbij moet opgemerkt worden dat de 'juistheid' van beslissingen over de instructiegroep van een leerling subjectief is. Het feit dat de beslissingen niet veranderen betekent niet dat de beslissing minder goed is.

Uitspraken die in dit onderzoek gedaan worden zijn gebaseerd op interviews met twaalf verschillende leerkrachten. Het gestelde doelen waar het gaat om de kenmerken geslacht en bouw zijn gehaald. De leerkrachten zijn qua geslacht precies evenredig verdeeld en hoewel de bouw waarin de leerkrachten werkzaam zijn niet precies half om half is, is zowel onder- als bovenbouw goed vertegenwoordigd. Ook is er diversiteit in hoe lang leerkrachten werken met LOVS en de leeftijd van leerkrachten. Dit maakt dat er is gewerkt met een redelijk diverse groep en dat de steekproef waarschijnlijk een redelijk beeld geeft van de mogelijke opvattingen. Dat de steekproef een redelijk beeld geeft van de verschillende opvattingen wordt ondersteund doordat er na verloop van tijd minder nieuwe opvattingen bij kwamen, hetgeen een indicatie is dat saturatie is bereikt.

Zoals gezegd is er bij een aantal leerkrachten voor gekozen om uit te leggen wat een score-interval inhoudt. De leerkrachten aan wie dit is uitgelegd waren na aanmoediging niet in staat om een schriftelijk of gevisualiseerd interval te interpreteren. Omdat het onderzoek als doel had om inzicht te krijgen in hoe de interpretatie van leerkrachten tussen een schriftelijk en een gevisualiseerd interval verschilt, was het belangrijk dat leerkrachten beide intervallen wel konden interpreteren.

Sommige leerkrachten kwamen zelfstandig in de buurt van een juiste interpretatie van een score-interval. Omdat zij geen uitleg hebben gehad over wat een score-interval is, is het echter mogelijk dat zij op basis van giswerk het score-interval en de visualisaties zijn gaan interpreteren. Hun uitspraken over welke weergavevorm zij het meest duidelijk vonden, moeten daarom met enige voorzichtigheid worden bekeken. Veel leerkrachten konden het score-interval pas op de juiste manier interpreteren nadat de betekenis van een score-interval was uitgelegd. Ook bij deze groep is het onzeker of er veel waarde gehecht kan worden aan de uitspraken over de duidelijkheid van beide weergaves, omdat hun interpretatie is beïnvloed door de gegeven uitleg. Dat veel leerkrachten het score-interval pas na uitleg konden interpreteren, geeft de noodzaak aan voor dat leerkrachten meer leren over het interpreteren van resultaten.

Het interviewprotocol had een logische opbouw, omdat is gestart met een aantal algemene vragen en vervolgens meer specifieke zaken aan bod kwamen. Omdat eerst is gevraagd naar het schriftelijke score-interval en daarna pas naar de error-bar, konden leerkrachten zonder beïnvloeding van een visualisatie aangeven hoe zij een score-interval interpreteerden. Dat daarna werd gevraagd naar de error-bar gaf de ruimte voor leerkrachten om aan te geven of de error-bar het schriftelijke score-interval verduidelijkte. Dat er meer visualisaties aan bod kwamen, maakte wel dat leerkrachten wisten welke vragen er gingen komen en dat er een bepaalde gemakzucht ontstond in het beantwoorden van de vragen. Het is de vraag of de leerkrachten die na drie visualisaties telkens aangaven niet van mening te veranderen over in welke instructiegroep de leerling geplaatst moest worden bij het zien van de vierde visualisatie nog wel goed nadachten over hun beslissing.

De onderzoeker heeft zich ingespannen om zo objectief mogelijk codes toe te wijzen aan uitspraken. Het toewijzen van codes is echter een subjectief proces. Omdat het codeschema door slechts één onderzoeker is toegepast op de interviews en er geen tweede onderzoeker is geweest die hetzelfde codeschema heeft toegepast op dezelfde interviews is niet getoetst of andere onderzoekers tot dezelfde conclusies zouden komen. Het is mogelijk dat een andere onderzoeker bepaalde uitspraken anders zou interpreteren. Codes zijn wel uitsluitend toegewezen aan uitspraken die de onderzoeker duidelijk genoeg achtte. Wanneer een uitspraak niet duidelijk genoeg was, heeft de onderzoeker vervolgvragen gesteld om de uitspraak te verduidelijken. Hierdoor konden codes met meer zekerheid worden toegeschreven.

Ten slotte is er in het onderzoek niet gevraagd naar de opleiding van de leerkrachten. Het ligt in de lijn der verwachting dat het grootste deel van de leerkrachten een hbo-opleiding heeft genoten (de pabo), maar dat is niet gecontroleerd. Het kan zijn dat dit een effect heeft op de mate waarin leerkrachten in staat zijn om data op een juiste manier te interpreteren. Op de pabo wordt weinig tot geen aandacht besteed aan statistiek en daardoor ontbreekt kennis die nodig is om de data goed te kunnen analyseren. Dit sluit aan bij de bevindingen van Kennedy-Clark et al. (2020) die schreven dat leerkrachten een sterke set vaardigheden gerelateerd aan data nodig hebben. Obery et al. (2019) pleiten er dan ook voor om studenten aan lerarenopleidingen te onderwijzen over hoe om te gaan met data. Leerkrachten moeten meer competent worden in het gebruik van data (Schildkamp & Kuiper, 2010). Omdat er op dit moment geen aandacht voor is, is het niet verwonderlijk dat leerkrachten moeite hebben met de interpretatie van score-intervallen. Het zou interessant zijn om te onderzoeken of academische leerkrachten, die universitair opgeleid zijn en zodoende wel statistiek hebben gehad op de opleiding, score-intervallen anders interpreteren en wel meenemen in hun

beslissingen. Onderzoek wijst erop dat leerkrachten die getraind zijn beter in staat zijn om de resultaten uit leerlingvolgsystemen te interpreteren (Staman et al., 2013) en het zou kunnen dat dit zich ook vertaalt naar LOVS. Om dat te kunnen bepalen zou onderzoek opgezet kunnen worden waarin leerkrachten met een pabo-opleiding en leerkrachten met een academische pabo-opleiding met elkaar vergeleken worden.

Al met al geeft dit onderzoek een indicatie dat het toevoegen van een error-bar voordelen zou kunnen bieden ten opzichte van enkel een schriftelijk weergegeven score-interval bij LOVS. De visualisatie en de duidelijke boven- en ondergrens maakt de error-bar duidelijker voor de geïnterviewde leerkrachten en doet hen het score-interval beter interpreteren. Het onderzoek levert echter ook nieuwe vragen op en doet daarom aanbevelingen voor vervolgonderzoek omtrent het uitsluiten van mogelijke alternatieve verklaringen voor een betere interpretatie van error-bars en omtrent de opleiding van leerkrachten.

Referenties

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*(4), 389-396. doi:10.1037/1082-989X.10.4.389

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. doi:10.1080/07481756.2002.12069053

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*(2), 170-180. doi:10.1037/0003-066X.60.2.170

Demmans Epp, C., & Bull, S. (2015). Uncertainty representation in visualizations of learning analytics for learners: Current approaches and opportunities. *IEEE Transactions on Learning Technologies, 8*(3), 242.

Drenth, P. J. D., & Sijtsma, K. (2006). Betrouwbare score en meetfout. In P. J. D. Drenth, & K. Sijtsma (Eds.), *Testtheorie: Inleiding in de theorie van de psychologische test en zijn toepassingen* (pp. 194-202). Houten: Bohn Stafleu van Loghum.

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.

Hop, M., & Engelen, R. (2017). *Wetenschappelijke verantwoording rekenen-wiskunde 3.0 voor groep 6*. Arnhem: Cito. Geraadpleegd op 3 mei 2022, van cito.nl/-/media/files/kennisbank/cito-bv/15-cito-lvs-rekwisk-3-0-gr6-wet-verantwoording.pdf?la=nl-nl

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, 26*(2), 123-142.

Kannan, P., Zapata-Rivera, D., & Bryant, A. D. (2021). Evaluating parent comprehension of measurement error information presented in score reports. *Practical Assessment, Research, and Evaluation, 26*(12), 1-21. doi:https://doi.org/10.7275/rgwg-t355

Kennedy-Clark, S., Galstaun, V., Reimann, P., Martyn, T., Williamson, K., & Weight, J. (2020). Voices on data literacy and initial teacher education: Pre-service teachers' reflections and recommendations. *Australian Journal of Teacher Education, 45*(7), 60-76.

Meijer, J., Ledoux, G., & Dorothé Elshof, M. M. V. (2011). *Gebruiksvriendelijke leerlingvolgsystemen in het primair onderwijs*. Amsterdam: Kohnstamm Instituut.

Obery, A., Sletten, J., Vallor, R. R., & Schmitt-Wilson, S. (2021). Data driven decision making in teacher education: Perceptions of pre-service teachers and faculty who teach them. *Action in Teacher Education, 43*(2), 231-246.

Padilla, L., Kay, M., & Hullman, J. (2020). Uncertainty visualization. *Handbook of computational statistics and data sciene*.

Rijksoverheid. (2022). *Hoe legt de basisschool de prestaties van mijn kind vast?* Geraadpleegd op 1 juni 2022, van https://www.rijksoverheid.nl/onderwerpen/basisonderwijs/vraag-en-antwoord/hoe-legt-de-basisschool-de-prestaties-van-mijn-kind-vast

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education, 26*(3), 482-496. doi:https://doi.org/10.1016/j.tate.2009.06.007

Staman, L., Visscher, A. J., & Luyten, H. (2013). The effects of training school staff for utilizing student monitoring system data. In D. Passey, A. Breiter & A. J. Visscher

(Eds.), *Next generation of information technology in education management* (pp. 3-14). Heidelberg, Duitsland: Springer.

Thompson, B. (2003). Nature of score reliability. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 4-5). Thousand Oaks, California, Verenigde Staten: Sage Publications, Inc.

Tomesen, M., Engelen, R., & Hiddink, L. (2019). *Wetenschappelijke verantwoording LVS-toetsen begrijpend lezen 3.0 voor groep 8*. Arnhem: Cito. Geraadpleegd op 3 mei 2022, van cito.nl/-/media/files/kennisbank/cito-bv/112-cito_lvs-begrijpend-lezen-3-0-gr-8_wet-verantwoording.pdf?la=nl-nl

Van der Kleij, F. M., & 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.

Van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (2014). Towards valid score reports in the computer program LOVS: A redesign study. *Studies in Educational Evaluation, 43*, 24-39. doi:https://doi-org.proxy-ub.rug.nl/10.1016/j.stueduc.2014.04.004

Van Engelshoven, I. K. (2018). *Uitvoering regeerakkoord t.a.v. kleutertoetsen*. Den Haag: De Tweede Kamer.

Bijlage 1 – Interviewleideraad

Allereerst bedankt voor je deelname aan dit onderzoek. Samen met twee andere studenten doe ik onderzoek naar de interpretatie van meetnauwkeurigheid van citoscores. 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 citoscores? (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 leerling rapport. Het gaat om een spelling citoscore van een willekeurige leerling uit groep 5.
[Nu de leerkracht de afbeelding van het leerlingrapport laten zien]

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 citoscores?

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 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 echt 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 kies je voor de voorkeur van weergave van ……..?

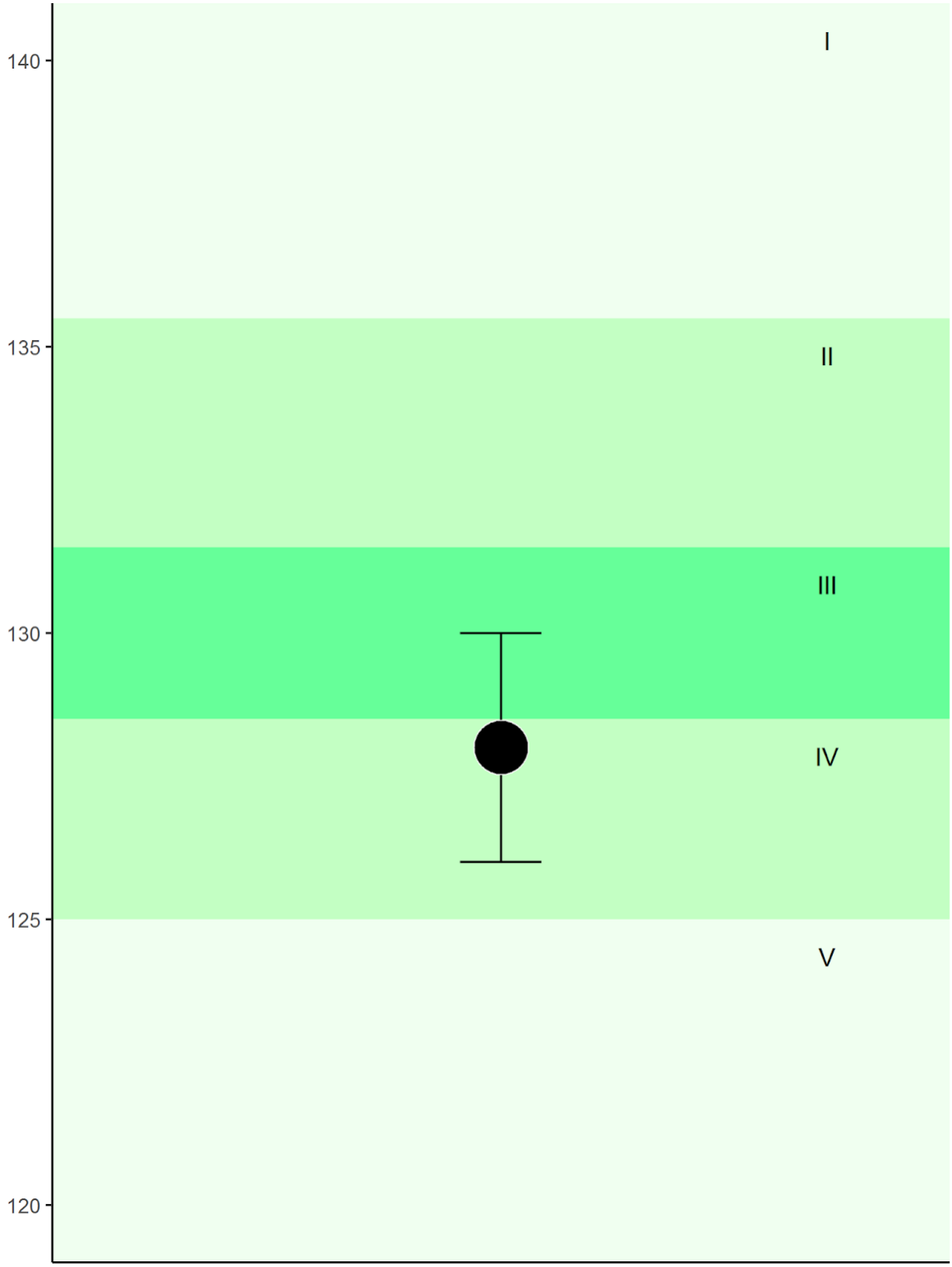
Dat waren de inhoudelijke vragen. Bedankt voor je antwoorden. Is er nog iets dat je kwijt wilt, wat niet aan bod is gekomen?

Dan heb ik nog een aantal korte punten:

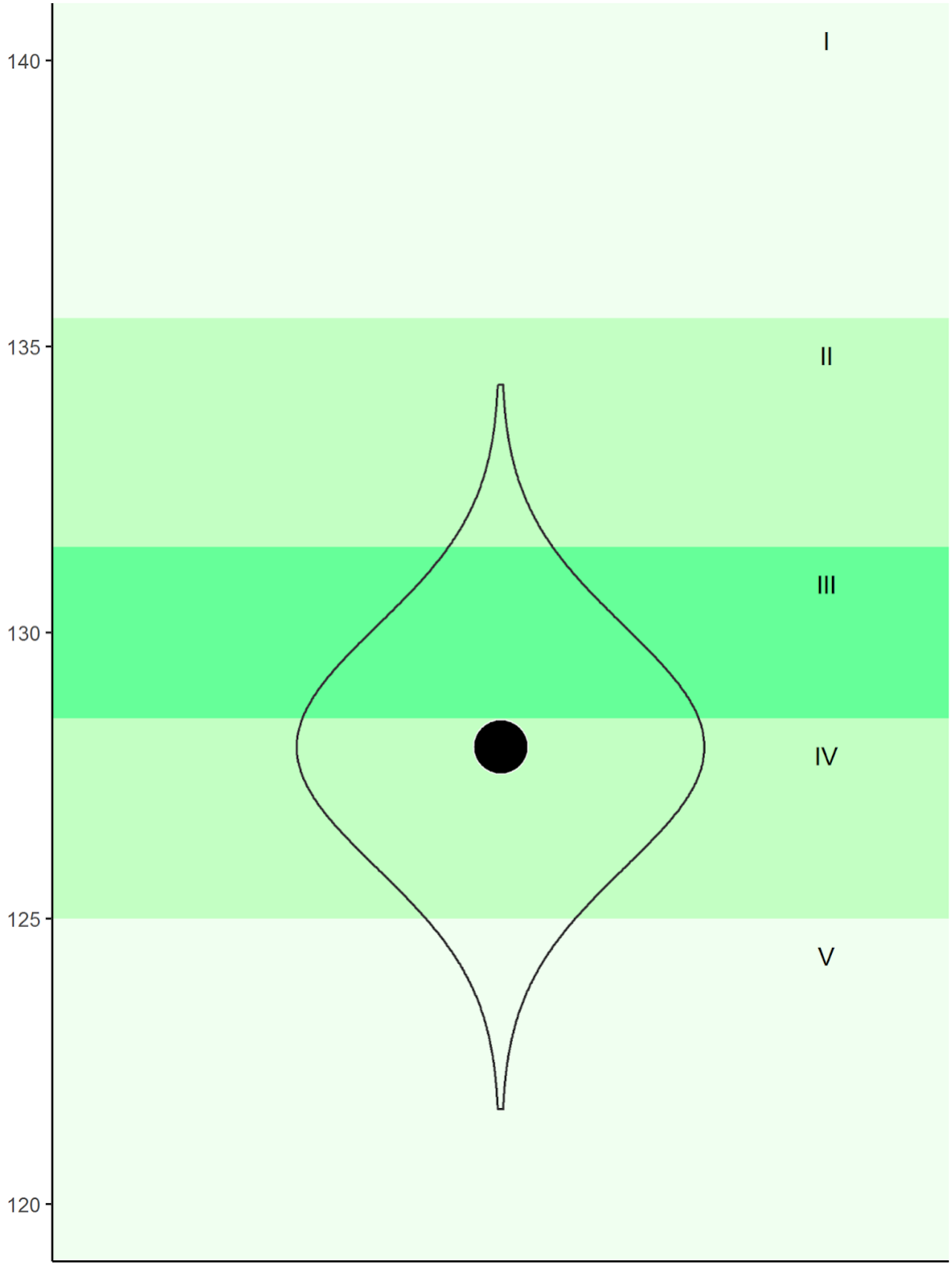
- We gaan de opname helemaal uitschrijven. Wil je daar een kopie van ontvangen?
- Mag ik eventueel contact met je opnemen, mocht er iets onduidelijk zijn?
- Mocht je nog contact met mij willen opnemen, kan dat door mij te mailen. Dan schrijf ik zo meteen mijn e-mailadres even op voor je.
- En tot slot,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

Dan zet ik nu de opname stop. Dankjewel!

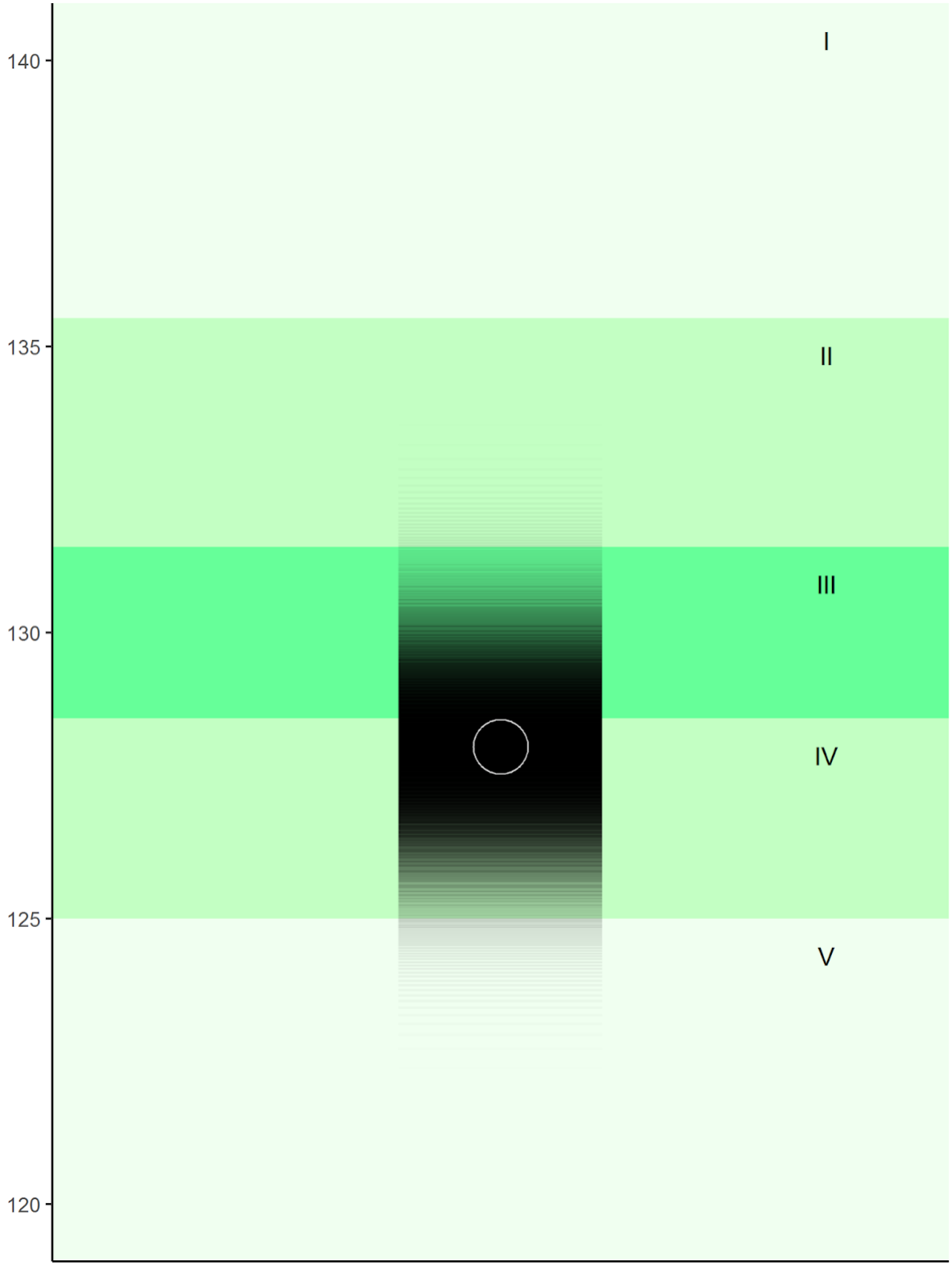
Bijlage 2 – error-bar (plaatje 1)



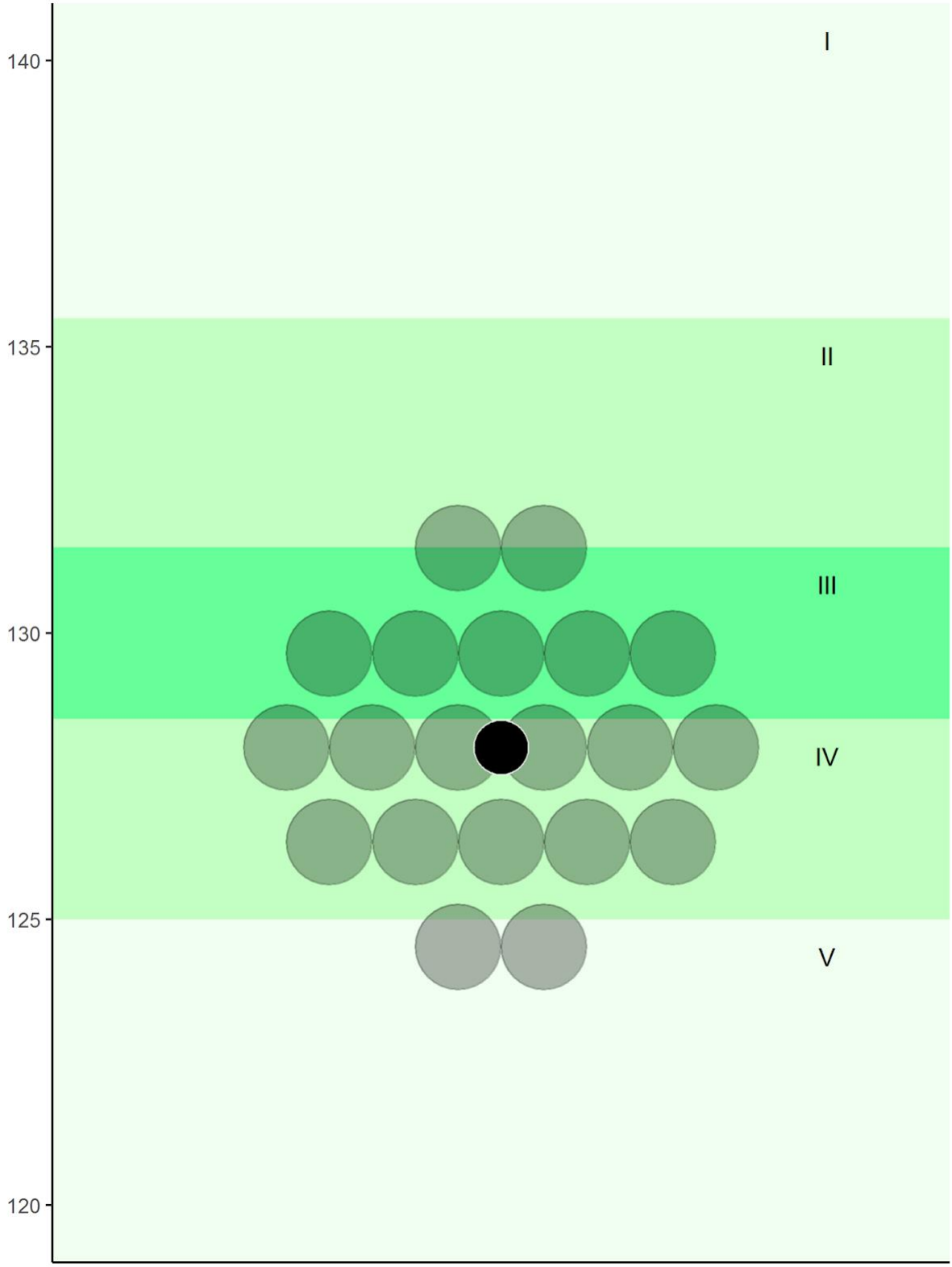
Bijlage 3 – violin plot (plaatje 2)



Bijlage 4 – gradient plot (plaatje 3)



Bijlage 5 – quantile dotplot (plaatje 4)



Bijlage 6 – codeschema

Code	Voorbeeld
Actie error-bar = afhankelijk van context	‘Die zou ik nog wel baseren op wat ik eerder vertelde, dus.’
Actie error-bar = basisgroep	‘Nee, dat zou nog steeds hetzelfde blijven, omdat het een IV-score is.’
Actie error-bar = extra aandacht nodig	[Interviewer: En beslissingen, zouden die anders zijn geworden?] ‘Ook niet.’
Actie error-bar = misschien basisgroep	‘Maar stel dat ik een groep heb met 20 kinderen waarvan er 15 zijn die onvoldoende scoren, dus een IV of een V-score, dan zou ik bij een kind als deze zeggen van ‘nou, dan doe ik jou toch bij aanpak 2, want jij bent er bijna.’
Actie error-bar = zelf nadenken	‘Nee, dat zou nog steeds hetzelfde blijven, omdat het een IV-score is.’
Actie error-bar = zorggroep	‘Nou, ik denk dat ik die leerling alsnog wel in een instructie groepje zou nemen.’
Actie leerlingrapport = afhankelijk van onderwerp	‘Dus ik zou niet vast aan de instructietafel, maar misschien per onderdeel afhankelijk, per les afhankelijk.’
Actie leerlingrapport = basisgroep	‘Zou ik hem nog in de basisgroep houden.’
Actie leerlingrapport = extra aandacht geven	‘Ik zou wel wat extra op dictees inzetten in ieder geval.’
Actie leerlingrapport = zelf nadenken	‘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.’
Actie leerlingrapport = zorggroep	‘Maar ik zou hem zeker meenemen aan de instructietafel.’
Actie score-interval = zorggroep	‘Ja, en daar zou je dan meer aandacht aan besteden in die instructie afhankelijke lessen.’
Cito gebruikt voor handelen	‘Ik deel de kinderen die een bepaalde score hebben gehaald, die deel ik in in een niveaugroep.’
Cito niet gebruikt voor handelen	‘Euh nee, ik gebruik citoresultaten niet als leidraad voor mijn verlengde instructie groepen.’

Error-bar duidelijk vanwege harde grens	[Interviewer: Dat kies je dus op basis van de grens die wordt aangegeven?] ‘Ja, dat is gewoon heel duidelijk eigenlijk, ja.’
Error-bar duidelijker dan score-interval	‘Dus ik zou deze [error-bar] denk ik duidelijker vinden.’
Error-bar is duidelijkste plaatje	‘Ik denk dat ik zou gaan voor nummer 1 [error-bar].’
Error-bar is niet duidelijkste plaatje	‘Ik zou neigen naar deze, plaatje 3 [gradient plot].’
Error-bar verandert interpretatie score-interval	‘Ja eigenlijk, ja wel, want omdat ik hier zoals ik zei, hier dacht ja dat het om een niveau IV ging? Dit interval past bij niveau IV-score, hier zie ik dat dat helemaal niet het geval is.’
Error-bar verduidelijkt niet	‘Interval zegt me niks en plaatje [error-bar] zegt me ook niks.’
Gebruikt score-interval niet	‘Ik kijk nooit naar die score-interval eigenlijk.’
Geen alarmbellen bij een keer IV	‘Dan zou ik nu niet meteen denken van al die paniek in de tent.’
Interpretatie met error-bar verandert niet	[Interviewer: Zou je nu anders naar die E5-score hebben gekeken als je dat score-interval op plaatje 1 [error-bar] mee zou nemen?] ‘Nee.’
Interpretatie error-bar = foute informatie bij cito	‘Ja die duidelijke grenzen, ja, die heb je bij cito niet, die zijn er niet. Ik denk niet dat je die kunt geven.’
Interpretatie error-bar = interpretatie score-interval	‘Ik zie het score-interval tussen de twee zwarte streepjes staan (…) en ik zie die zwarte punt als score die diegene heeft behaald.’
Interpretatie error-bar = interval niveau leerling	‘Dus dat zal wel de onzekerheid van de score zijn, zoals je dat net zegt.’
Interpretatie error-bar = minimum en maximum mogelijk	‘Nou, dat is het maximale en het minimale.’
Interpretatie error-bar = niveau III ook	‘Als ik hem hier zou lezen, zou ik zeggen ‘dit interval past echt bij een niveau IV-score’. Hier zie ik hem ineens voor niveau III ook staan.’

Interpretatie error-bar = onduidelijk	‘Nou, ik vind deze eigenlijk net zo lastig als plaatje 3 [gradient plot].’
Interpretatie error-bar = weet niet	‘Interval zegt me niks en plaatje [error-bar] zegt me ook niks.’
Interpretatie leerlingrapport = groei neemt af	‘En ik zie ook dat de lijn nu wat afvlakt. Je zou willen dat die lijn zou meebewegen met die golfbeweging, maar hij vlakt af.’
Interpretatie leerlingrapport = onvoldoende score	‘En ik zie dat aan het einde deze leerling weer een klein stukje naar beneden gaat, dus weer naar een onvoldoende.’
Interpretatie leerlingrapport = verschilt per vakgebied	‘Nou ja, dat hangt van het vak af, hè? Met spelling doe ik daar op zich niet zo veel mee (…) en met begrijpend lezen kijk ik toch wel echt van nou, waar liggen de hiaten?’
Interpretatie score-interval = aantal vragen gemaakt	‘Misschien de hoeveelheid vragen die is gemaakt in de afgelopen aantal jaren?'
Interpretatie score-interval = aantal vragen goed	‘Oh dat zijn het aantal vragen goed misschien?’
Interpretatie score-interval = gebied niveau	‘Ik denk als je tussen de 126 en 130 zit qua vaardigheidsscore, dan zit je in de IV.’
Interpretatie score-interval = interval niveau leerling	‘Nee, 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.’
Interpretatie score-interval = verwachte score	‘Nou misschien wel wat ze verwachten wat hij gaat scoren.’
Interpretatie score-interval = weet niet	‘Ik zou het niet weten!’
Juiste interpretatie score-interval maar wil niet gebruiken	‘Ik zou dan kijken waar die 126 zit. Maar dan, dat is iets meer naar de onderkant natuurlijk en 130 zit er iets meer aan de bovenkant. (…) Nee, ik zou daar den kik eigenlijk niks mee doen.’
Kleiner interval is duidelijker	‘Dit is iets duidelijker, want hier zitten geen II en V meer bij.’

Meer informatie in plaatje = fijner	‘Dat zou mij meer informatie geven. En dan zou ik dus voor plaatje 2 [violin plot] gaan.’
Minder informatie in plaatje = fijner	‘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.’
Score-interval duidelijker dan error-bar	[Interviewer: Als je moet kiezen tussen [het score-interval] of het laatste plaatje [error bar], welke kies je dan?] ‘De eerste.’
Uitval cito = zorggroep	‘Nou, wij gebruiken voornamelijk de IV en V-scores voor de mindere leerlingen om het zo maar te noemen. Dus die komen inderdaad in instructiegroepjes.’
Verkeerde interpretatie error-bar doet interpretatie leerlingrapport veranderen	‘Ja, want vooral met de gedachte van het score-interval, want nu heb je het idee dat de verwachting van de score ook zo was geweest. En dat blijkt in de grafiek natuurlijk niet zo, dat hij een beetje afdaalt.’
Visualiseren interval duidelijker	‘Als je hiernaar kijkt, naar plaatje nummer 1 [error-bar], dan kun je heel duidelijk zien dat het net onder de lijn van een III-score zit, binnen de IV-score en dat er een marge zit naar III en naar IV.’
Weet niet van bestaan score-interval	‘Nou dit schema, dit onderste deel, zit er bij mijn schema niet bij.’
Weinig ontwikkeling	‘Weinig ontwikkeling zie ik dan eigenlijk.’
Wil error-bar als aanvulling op leerlingrapport	‘Ja, plaatje 1 [error-bar] zou ik eronder willen plakken ja.’
Wil error-bar gebruiken	‘Nou, ik zou dit [error-bar] sowieso meenemen.’
Wil plaatjes niet gebruiken als niet kan interpreteren	‘Nouja, ik vind als je zo’n grafiek krijgt, moet je weten waar je over praat.’