3-6-2022

# Fouten bij het bekijken van meetfouten?

## Hoe leerkrachten en pabo-studenten verschillen in accuratesse van de interpretatie van meetfouten bij Cito-scores

Ruben Boersma (S3720640)

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

Rijksuniversiteit Groningen
Faculteit der Gedrags- en Maatschappijwetenschappen
PABA6002: Bachelorwerkstuk
3 juni 2022

## Abstract

Visual representations of Cito-scores are used to measure academic performance of students in Dutch primary schools. Measurement errors are inherent to any test. Research by Van der Kleij and Eggen (2013) suggests that measurement errors are often being misinterpreted by teachers in Dutch primary school. These misinterpretations could lead to incorrect allocation to instruction groups. Resnick et al. (2018) suggests that students, like teachers-in-training, use effective strategies to analyse data. The purpose of this thesis was to explore if teachers-in-training would differ significantly in their interpretation ability of Cito-scores in comparison to experienced primary school teachers. Using a reanalysis of the original study by Ettema (2021), a correlational design was set. The relation between the interpretation ability and the educational phase of the participant has been researched. 33 primary school teachers and 39 teachers-in-training participated, whose educational phase, sex, age, and the province in which they worked, were identified. All participants were shown ten visual representations of measurement errors of Cito-scores and were asked to estimate the chance of the corresponding student's academic level. The independent-samples t-test concluded that there was no significant difference in the interpretation ability of primary school teachers and teachers-in-training. Although not significantly proven, the data does suggest that teachers-in-training are better able to assess the visual representations of Cito-scores. Future research using a larger sample would support this suggestion. Guiding teachers-in-training to better interpret these visual representations of Cito-scores, would possibly increase their ability to make better estimations of the academic level of their students. In conclusion, while no significant difference in the interpretation ability of measurement errors in visual representations of Cito-scores by teachers-in-training and primary school teachers could be found, the data does suggest that teachers-in-training make more accurate estimates.

## Inleiding

Volgens Cito (z.d.) maakt 85 procent van de scholen in het Nederlandse regulier en speciaal basisonderwijs gebruik van één of meerdere toetsen van het Cito-leerlingvolgsysteem, ook wel LOVS genoemd (Cito, z.d.; Van der Kleij & Eggen, 2013). Met deze landelijke toets worden er op twee meetmomenten in het schooljaar gekeken hoe de leerlingen scoren ten opzichte van een representatieve normgroep (Meijer, Ledoux & Elshof, 2011). Als een leerling op cognitief gebied achterblijft, dan kan de leerkracht de ontwikkeling inzien van de afgelopen jaren (Bunck, Terlien, Van Groenestijn, Toll, & Van Luit, 2017).

Elke test probeert de werkelijke vaardigheidsscore te schatten van de respondent (Charter & Feldt, 2002; Drenth & Sijtsma, 2005), een Cito-toets is een voorbeeld van. Echter zijn alle tests in zekere mate onnauwkeurig, ze kunnen namelijk nooit achterhalen wat de werkelijke score van een respondent is (Kolen, Zeng & Hanson, 1996). Om toch een schatting te doen van de werkelijke score, wordt er een schatting gemaakt op basis van de geobserveerde score (Charter & Feldt, 2002; Drenth & Sijtsma, 2005; Kolen et al., 1996). Omdat er bij testen gewerkt wordt met een schatting van de werkelijke score, zijn meetfouten inherent (Drenth & Sijtsma, 2005). Deze meetfouten zijn willekeurige afwijkingen van de werkelijke score die veroorzaakt worden door toevalligheden die niks te maken hebben met de vaardigheid van een leerling. Bij het voorbeeld van de Cito-toetsen voor basisschoolkinderen kan gedacht worden aan afleiding tijdens het toetsingsmoment, een slechte nachtrust hebben gehad of door het per ongeluk goed gokken van een antwoord. Deze onvoorspelbare gebeurtenissen zijn niet van invloed op de rekenvaardigheid van de leerling, maar wel op de geobserveerde score. De mate waarin meetfouten voorkomen, is in te schatten en weer te geven door middel van de betrouwbaarheid van een test of een betrouwbaarheidsinterval (Charter & Feldt, 2002; Drenth & Sijtsma, 2005). Cito hanteert

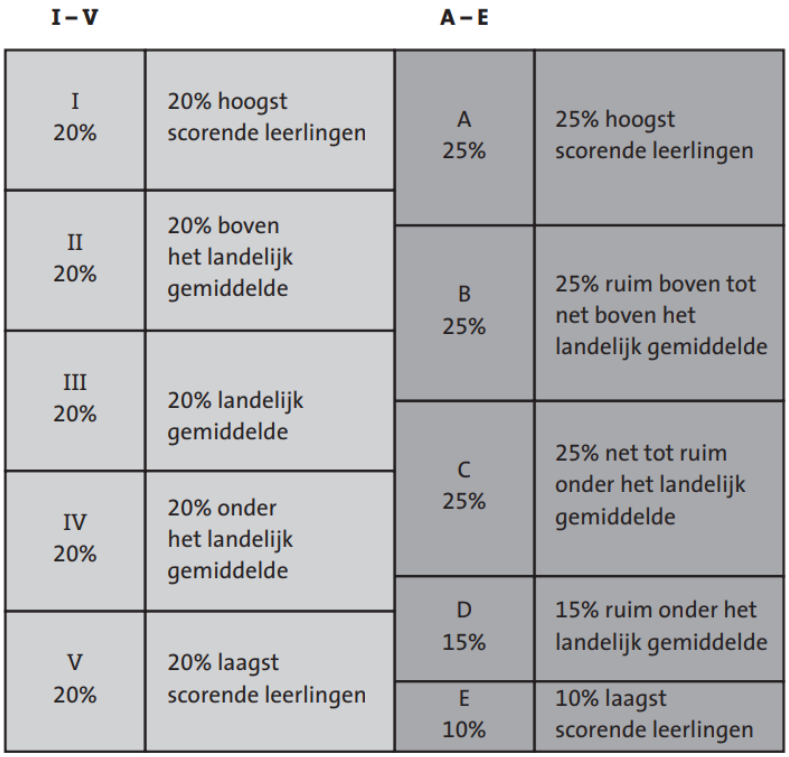
voor het schatten van de vaardigheidsscore van de leerlingen een 68%-betrouwbaarheidsinterval.

Hoewel meetfouten inherent zijn aan elke test, weten veel leerkrachten niet wat meetfouten zijn of passen de kennis die zij hierover hebben niet toe als het gaat om de interpretatie van de geschatte score (Van der Kleij & Eggen, 2013). Verschillende onderzoeken geven hier als voornaamste reden voor dat leerkrachten een gebrek aan kennis en vaardigheden hebben over het gebruik van data om onderwijs af te stemmen op de instructiebehoefte van de leerling (Saunders, 2000; Earl & Fullan, 2003; Van Petegem & Vanhoof, 2004; Kerr, Marsch, Ikemoio, Darilek, & Barney, 2006; Williams & Coles, 2007; Ledoux, Blok, Boogaard, & Krüger, 2009; Zupanc, Urank, & Bren, 2009; Meijer et al., 2011). Uit het onderzoek van Van der Kleij & Eggen (2013) wordt verder duidelijk dat de interpretatie van statistische gegevens en grafieken bij leerkrachten verwarring en verkeerde interpretaties veroorzaakt. Ook het begrip betrouwbaarheidsintervallen was bij de meeste participanten van de focusgroepen niet duidelijk en geen van de participanten gaf aan de betrouwbaarheidsintervallen te gebruiken in de dagelijkse praktijk (Van der Kleij & Eggen, 2013).

Cito hanteert twee verschillende indelingen als het gaat om niveaugroepen, een indeling op basis van de niveaus I tot en met V en een indeling op basis van de niveaus A tot en met E (Cito, 2019). In Figuur 1 zijn beide indelingen op basis van de niveaugroepen te zien. Elke niveaugroep uit deze indeling (I, II, III, IV en V) laat zien hoe de leerling scoort ten opzichte van de landelijke steekproef. Elke Cito-toets kent zijn eigen representatieve landelijke steekproef die jaarlijks wordt aangepast. Voor de rest van de thesis zal de indeling op basis van de niveaus I tot en met V gebruikt worden.

**Figuur 1**

*Cito Indelingen van Niveaugroepen I tot en met V en A tot en met E (Cito, 2019)*



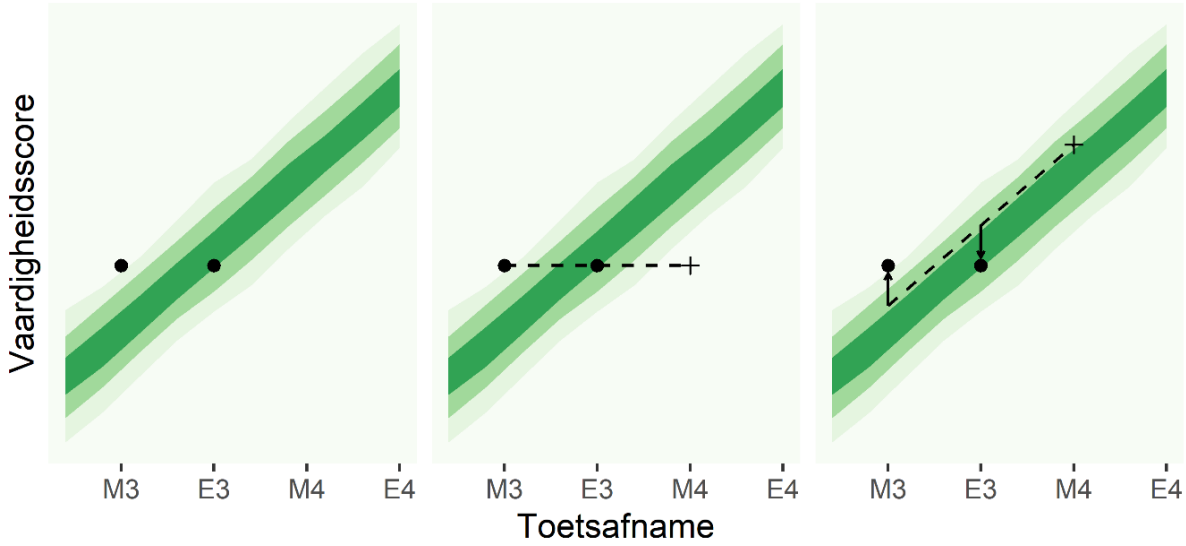
| I – V | | A – E | |
|---|---|---|---|
| I 20% | 20% hoogst scorende leerlingen | A 25% | 25% hoogst scorende leerlingen |
| II 20% | 20% boven het landelijk gemiddelde | B 25% | 25% ruim boven tot net boven het landelijk gemiddelde |
| III 20% | 20% landelijk gemiddelde | C 25% | 25% net tot ruim onder het landelijk gemiddelde |
| IV 20% | 20% onder het landelijk gemiddelde | D 15% | 15% ruim onder het landelijk gemiddelde |
| V 20% | 20% laagst scorende leerlingen | E 10% | 10% laagst scorende leerlingen |

Charter en Feldt (2002) geven aan dat meetfouten wel degelijk een verschil kunnen maken in het onderwijs. Een leerling die zich op het randje begeeft van een IV-score en nog net in een III-score valt, kan als er gekeken wordt naar de foutenmarge, toegewezen worden aan beide scores. De leerling kan vervolgens toegewezen worden aan een instructiegroep, terwijl de daadwerkelijke score van deze leerling hoger ligt dan de score op de Cito-toets liet zien. Van der Bles et al. (2019) geven aan dat gegevens die gepresenteerd worden met onzekerheid, zoals betrouwbaarheidsintervallen, verwarring oproepen bij de lezer. Dit wordt ondersteunt door de eerder aangehaalde bevindingen van Van der Kleij en Eggen (2013). Door deze verwarring worden er volgens Van der Bles et al. (2019) verkeerde interpretaties gemaakt van deze gegevens met als gevolg dat leerlingen mogelijk tot de verkeerde instructiegroep toebedeeld worden.

Frans, Post, Oenema-Mostert en Minnaert (2020) schetsen twee scenario's voor een leerling die een hoge I-score op de M3-toets heeft behaald, te zien in Figuur 2. Als een leerling stilstaat in zijn ontwikkeling en op de E3-toets een lage III-score heeft, kan dit op twee manieren verklaard worden. Als de ontwikkeling van deze leerling daadwerkelijk achteruitgaat, kan er voorspeld worden dat deze leerling een lage V-score zal halen op de M4-toets, dit wordt functiestabiliteit genoemd. Vanwege de betrouwbaarheidsintervallen met bijbehorende meetfouten, kan de daadwerkelijke score op de M3-toets ook een II-score zijn. Wanneer de leerling vervolgens tijdens de E3-toets onder zijn kunnen presteert, kan het daadwerkelijke niveau ook een II-score zijn. Dit betekent dat er voorspeld kan worden dat de leerling op II-niveau blijft en zich doorontwikkeld, dit wordt lineaire stabiliteit genoemd. Een leerkracht die geen rekening houdt met de mogelijkheid van meetfouten zal alleen uitgaan van de functiestabiliteit. Deze leerkracht ziet dat de ontwikkeling van een I-score afzakt naar een lage III-score met als voorspelling dat dit een V-score zou opleveren bij de M4-toets. Een leerkracht die wel rekening houdt met de mogelijkheid van meetfouten zou meer oog hebben voor de lineaire stabiliteit. Hierbij kan de werkelijke score van de leerling zich bevinden op een II-niveau en kan de leerkracht voorspellen dat de leerling op de M4-toets een II-score behaald. Het verschil tussen deze twee leerkrachten is interessant te noemen, aangezien de leerkracht die geen rekening houdt met de mogelijkheid van meetfouten verlengde instructies zal toepassen met het oog op de voorspelde V-score. De leerkracht die wel rekening houdt met de mogelijkheid van meetfouten zal zijn instructie voor deze leerling aanpassen op het beoogde II-resultaat op de M4-toets.

**Figuur 2**

*Visuele Weergave van Functiestabiliteit en Lineaire Stabiliteit (Frans et al., 2020)*




Uit onderzoek van Hopster-den Otter, Muilenburg, Wools, Veldkamp en Eggen (2018) is gebleken dat leerkrachten de visuele weergaven van meetfouten bij testscores verkiezen boven het weglaten van de meetfouten. Ook vonden 30.8% van de respondenten dat visuele weergaven van meetfouten een positieve bijdrage levert aan hun zelfvertrouwen, tegenover 17.8% van de respondenten die minder zelfvertrouwen hierdoor kregen. Ook Padilla, Kay en Hullman (2020) toonden aan dat onzekerheden in data het best ondersteund kunnen worden door middel van visuele weergaven. Hierdoor zou er ook minder verwarring optreden bij de expert die de data probeert te interpreteren, waardoor foutieve interpretaties uitblijven (Van der Bles et al., 2019). Door het gebruik van de visuele weergaven van meetonnauwkeurigheden, werden leerkrachten beter in staat om met deze schattingen om te gaan (Hopster-den Otter et al., 2018). In het onderzoek van Hopster-Den Otter et al. (2018) werd nog niet gekeken in hoeverre leerkrachten in opleiding deze inschattingen kunnen maken op basis van dezelfde visuele weergaven. Uit onderzoek van Resnick, Kastens en Shipley (2018) blijkt dat studenten wel effectieve strategieën gebruiken om visuele weergaven van data te analyseren. Is het zo dat pabo-studenten beter in staat zijn om een

inschatting te maken op basis van een gegeven visuele weergave en verleren leerkrachten in de loop van de tijd deze vaardigheid of speelt ervaring hierin een belangrijke rol. In het onderzoek van der Kleij, Eggen en Engelen (2014) is gebleken dat het ontvangen van training in het gebruik van LOVS niet leidt tot accuratere interpretaties. In de literatuur wordt geen onderzoek gedaan naar de accuratesse van leerkrachten in opleiding in vergelijking met leerkrachten. Door de bevindingen van Resnick et al. (2018) dat studenten effectieve strategieën gebruiken om visuele weergaven van data te analyseren, terwijl de interpretatie van statistische gegevens en grafieken bij leerkrachten verwarring en een verkeerde interpretatie met zich meebrengen (Van der Kleij & Eggen, 2013) kan er gesteld worden dat pabo-studenten accurater zijn in de interpretatie van visuele weergaven dan leerkrachten. Dit wordt ondersteund door het feit dat training in het gebruik van LOVS niet bijdraagt aan accuratere interpretatie van de gegevens (Van der Kleij et al., 2014). De vraag of pabo-studenten beter in staat zijn om te gaan met de interpretatie van het Cito-leerlingvolgsysteem, wordt in de literatuur niet behandeld. Onderzoek kan aantonen of de pabo-studenten een hogere accuratesse hebben dan leerkrachten die al werkzaam zijn in het Nederlandse basisonderwijs.

De onderzoeksvraag die hieruit voortkomt is als volgt: *In hoeverre verschillen leerkrachten en pabo-studenten om een inschatting te maken op basis van een visuele weergave van een Cito-score?*

Het doel van dit onderzoek is om de verschillen van interpretatie van de foutenmarges op Cito-toetsen tussen leerkrachten en studenten in kaart te brengen. Hierdoor kan gekeken worden of leerkrachten de vaardigheden in het begin van hun loopbaan al zouden moeten beheersen of dat ze deze vaardigheid later pas verwerven.

# Onderzoeksmethode

## Design

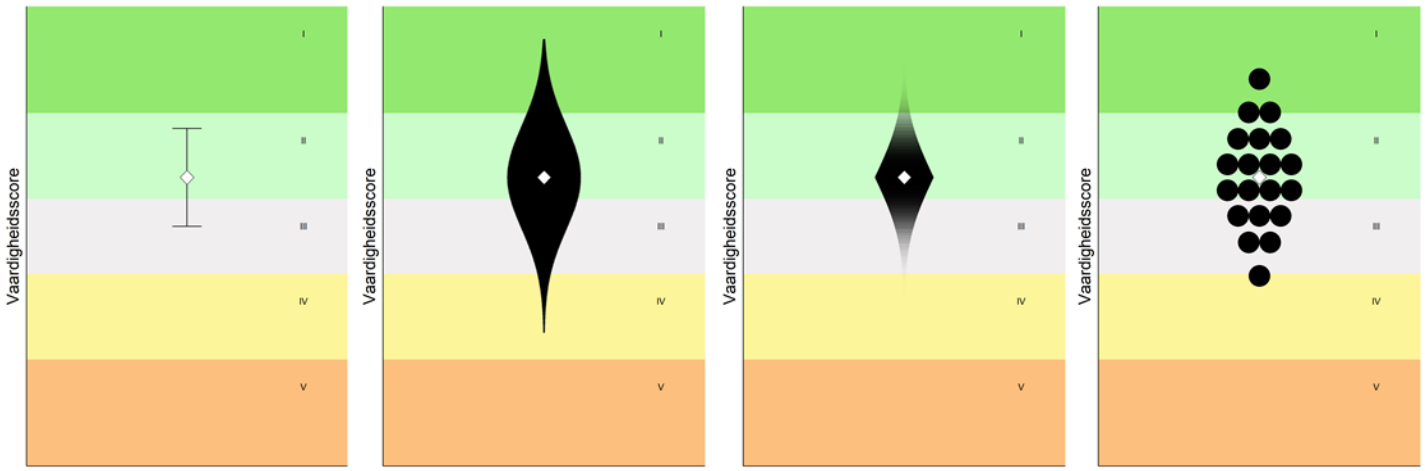
In dit onderzoek werd gewerkt met een bestaande dataset van Ettema (2021). Het doel van het onderzoek van Ettema (2021) was om de vaardigheid van leerkrachten met betrekking tot het inschatten van onzekerheden van testscores in kaart te brengen. Een heranalyse van het onderzoek van Ettema (2021) was gedaan, aangezien het gebruik van relatieve verschillen andere resultaten op kan leveren dan wanneer absolute verschillen gehanteerd worden, zoals in deze heranalyse is gedaan. Dit werd uitgesplitst in vier verschillende visualisaties, namelijk *error-bar, violin plot, gradiënt plot*[1]en *quantile dot plot*. In Figuur 3 zijn deze vier verschillende visualisaties weergegeven. Elke visualisatie liet met een witte ruit de geschatte vaardigheidsscore van een leerling zien. Vervolgens was de onzekerheid weergegeven aan de hand van de vier verschillende visualisaties. De vier weergaven uit Figuur 3 laten dezelfde onzekerheidsmarge zien. Elk gekleurd gebied correspondeert met de niveaugroepen I tot en met V (Cito, 2019).

Het oorspronkelijke onderzoek van Ettema (2021) had een experimenteel design, aangezien de participanten willekeurig één van de vier visualisaties toegewezen kregen. De heranalyse betrof een correlationeel design. Hierbij werd de relatie tussen accuratesse van kansinschattingen vergeleken met de categorische variabele *functie*, wat uitgesplitst werd in pabo-student en leerkracht in het Nederlandse basisonderwijs.

[1] De *gradiënt plot* is een variant op een standaard *gradiënt plot*. De uiteinden lopen taps toe en worden steeds vager. Voor de rest van de thesis wordt onder *gradiënt plot* de bewerkte versie van de *gradiënt plot* verstaan.

**Figuur 3**

*Visualisaties voor Vaardigheidsscore per Visualisatie: Error-bar, Violin Plot, Gradiënt Plot en Quantile Dot Plot.*




**Steekproef en populatie**

De doelpopulatie bestond uit zowel huidige als toekomstige leerkrachten in het Nederlandse basisonderwijs. Voorwaarde hierbij was dat de respondenten bekend zijn met het leerlingvolgsysteem van Cito. Een gelegenheidssteekproef werd gebruikt van leerkrachten in het Nederlandse basisonderwijs en derde- en vierdejaars studenten van de reguliere en academische pabo.

De dataset bevatte 140 respondenten, waarvan 72 respondenten de verschillende visualisaties van onzekerheden hebben gezien. Hiervan zijn er 33 leerkrachten, 35 pabo-studenten en 4 AOLB-studenten. De groep van AOLB- en pabo-studenten zijn samengevoegd. Deze groep wordt vanaf nu aangeduid met ‘studenten’.

**Instrumenten en variabelen**

In het onderzoek waren een aantal variabelen opgenomen. De onafhankelijke variabele *functie* gaf aan of de respondent een leerkracht, een reguliere pabo-student of een academische pabo-student is. De afhankelijke variabele was *accuratesse*, waarbij de accuraatheid van kansinschattingen gemeten werd. Elke respondent ontving vervolgens tien

vragen over één van de vier visuele weergaven van meetnauwkeurigheid in Figuur 3. Aan de hand van één van deze vier type plots diende de respondent verschillende kansinschattingen te maken met een percentage met een minimum van 0 procent en een maximum van 100 procent. Zo kon er van een respondent die het *gradiënt plot* uit Figuur 3 te zien kreeg, gevraagd worden hoe groot de kans geschat wordt dat het daadwerkelijke niveau van deze leerling gelijk zou zijn aan of lager zou zijn dan niveau II. Met de variabele *provincie* is duidelijk gemaakt in welke provincie de respondenten werkzaam zijn of studeren. De verschillende provincies zijn onderverdeeld regio Noord, Oost, Zuid en West. Hierbij worden de provincies Groningen, Friesland en Drenthe onder regio Noord gerekend, Gelderland, Overijssel en Flevoland onder regio Oost, Utrecht, Noord-Holland en Zuid-Holland onder regio West en Zeeland, Noord-Brabant en Limburg vallen onder regio Zuid (Janssen, Verhelst, Engelen, & Scheltens, 2010). Ook de variabelen *leeftijd* en *geslacht* van de respondenten zijn bekend. Daarnaast geeft de variabele *plottype* aan of de respondent vragen heeft gekregen met *error-bars*, *violin plots*, *gradiënt plots* of *quantile dot plots*.

De betrouwbaarheid van de uitkomstmaat van het onderzoek van Ettema (2021) is vastgesteld door gebruik te maken van de Guttman's Lambda-2. Met een waarde van .90 kan gesteld worden dat de uitkomstmaat in deze populatie betrouwbaar is.

**Procedure**

De respondenten werden benaderd via email en sociale media. Deze respondenten werd vervolgens ook gevraagd om de vragenlijst binnen hun eigen netwerk verder te verspreiden. De verspreiding van deze vragenlijst vond plaats tussen december 2020 en januari 2021. Voordat de respondenten deelnamen aan het onderzoek, werden zij schriftelijk op de hoogte gebracht van het doel van het onderzoek, de tijd die het maken van de vragenlijst in beslag zou nemen, dat er geen persoonlijke gegevens verzameld zouden worden en dat de respondenten ten alle tijden zich aan het onderzoek kunnen onttrekken.

Voorafgaand aan het invullen van de vragenlijst werd de respondent schriftelijk gevraagd om toestemming voor de verwerking en het gebruik van gegevens. Voor het uitvoeren van het onderzoek heeft de ethische commissie van Pedagogische- en Onderwijswetenschappen van de Rijksuniversiteit Groningen toestemming verleend.

**Analyse**

De analyses van de resultaten en de bewerking van de bestaande dataset van Ettema (2021) van deze thesis zijn uitgevoerd in ISBM SPSS Statistics 27. Er is gekozen voor een heranalyse en bewerking van de bestaande dataset van het onderzoek van Ettema (2021) om verschillen voor kleine waarden ook goed in kaart te brengen. In het oorspronkelijke onderzoek is gewerkt met ratioscores. Voor dit onderzoek worden echter verschilscores berekend. Wanneer de daadwerkelijke kans op een bepaalde uitkomst 1% bedraagt en de respondent geeft 1.1% als antwoord, dan zal zowel de ratioscore als de verschilscore een afwijking van 0.1 aangeven. Wanneer de daadwerkelijke kans 100% bedraagt en de respondent geeft 90% als antwoord, zal de ratioscore ook een afwijking van 0.1 aangeven, terwijl de verschilscore nu 10 bedraagt. Omdat de relatieve verschillen anders kunnen uitpakken dan de absolute verschillen, is er een heranalyse gedaan.

Om deze verschilscore te berekenen, is vooraf het correcte antwoord per vraag vastgesteld. Aan de hand van scores van de participanten kan er een verschilscore opgesteld worden. Hiervoor wordt de formule $P_{dif_i} = |\hat{p}_i - p_i|$ berekend. $\hat{p}_i$ is hierin de inschatting van een participant op item *i* als proportie en $p_i$ is het goede antwoord van item *i* als proportie. Om een zo accuraat mogelijke score te berekenen, is een verschilscore opgesteld. De absolute waarde zorgt ervoor dat er alleen gekeken wordt naar de afwijking van het werkelijke antwoord en niet naar de eventuele richting.

Om de invloed van uitbijters te verminderen en om tot een normale verdeling te komen, worden vervolgens de logaritmes berekend van de absolute verschilscores. Hierdoor

wordt de $P_{dif_i}$-waarde getransformeerd tot $\log_{10} P_{dif_i}$. Mocht een respondent de juiste inschatting gemaakt hebben ($\hat{p}_i = p_i$) zou log(0) berekend dienen te worden. Dit is echter niet mogelijk. Om dit te verhelpen wordt er een kleine waarde van 0.005 bij de $P_{dif_i}$-waarde opgeteld. Om de interpretatie van 0 staat gelijk aan perfect accuraat te behouden, wordt er vervolgens van dit totaal wederom log(0.005) afgehaald.

De totale formule luidt als volgt: *Accuratesse* = Log ($P_{dif_i}$+ 0.005) - Log (0.005). Hiermee werd voor elke respondent een nieuwe waarde *accuratesse* opgesteld. Hoe lager deze waarde, hoe dichter de kansinschattingen van de respondent bij het goede antwoord zaten.

Voor de beschrijvende statistiek is de verdeling binnen de steekproef in kaart gebracht. Hierbij is de steekproef omschreven aan de hand van de variabelen *geslacht*, *plottype*, *provincie*, *leeftijd* of ze de vragenlijst volledig hebben ingevuld.

Aangezien 1 op de 5 respondenten vroegtijdig de vragenlijst heeft beëindigd, is er een kruistabel met de bijbehorende chi-kwadraatwaarde opgesteld om *geslacht*, *plottype*, *provincie* en *functie* te vergelijken met het wel of niet voltooien van de vragenlijst.

Vervolgens is er een tabel opgesteld met de gemiddelde absolute afwijking per vraag. Hierdoor kon er gekeken worden of de groep leerkrachten en de groep studenten verschilden op basis van de gemiddelde absolute afwijking. Ook de standaarddeviatie is hierbij gerapporteerd. Vervolgens is voor de groep leerkrachten en de groep studenten een histogram opgesteld met een normaalcurve voor de gemiddelde absolute afwijking. Ook de assumpties voor de ongepaarde t-toets werden hier gecontroleerd.

Om de betrouwbaarheid voor de uitkomstmaat te berekenen, is *Guttman's Lambda-2* vastgesteld. Verder is er een kruistabel opgesteld met de bijbehorende chi-kwadraat waarden. Hiermee werd er gekeken of de onderzoeksgroepen onderling goed verdeeld zijn en er

mogelijk geen vertekeningen zichtbaar zijn tussen de groep leerkrachten en studenten. Deze analyse is gedaan door *geslacht, plottype* en *provincie* met de variabele *functie* te vergelijken.

Door gebruik te maken van een ongepaarde t-toets met een 95%-betrouwbaarheidsinterval is gekeken of de gemiddelde accuratesse-waarde van de groep leerkrachten significant verschilde van de groep pabo-studenten. Voor het uitvoeren van de ongepaarde t-toets dient er aan twee assumpties voldaan te zijn. Ten eerste dient er voldaan te zijn aan de assumptie van normaliteit. Hiervoor worden de histogrammen voor de loggetransformeerde accuratessescores van de groep leerkrachten en de groep studenten apart geplot. Hierin wordt ook de normaalcurve toegevoegd. Vervolgens kan gekeken worden of deze groepen normaal verdeeld zijn. Daarnaast dient er voldaan te zijn aan de assumptie van gelijke varianties. Hiervoor worden de standaarddeviaties van beide onderzoeksgroepen berekend en wordt gekeken of deze meer dan factor 2 verschillen. Is dit het geval, dan is er niet voldaan aan de assumptie van gelijke variantie. Naast de ongepaarde t-toets is ook Cohen's *d* berekend, zodat de effectgrootte bepaald kan worden. Een effectgrootte van 1.3 of hoger duidt op een zeer groot effect, tussen de 0.80 en 1.29 duidt het op een groot effect, een score tussen 0.50 en 0.79 duidt op een middelgroot effect, tussen 0.20 en 0.49 wordt er gesproken van een klein effect, tussen -0.19 en 0.19 wordt er gesproken over een verwaarloosbaar effect, tussen -0.49 en -0.20 duidt op een klein negatief effect, et cetera.

Tot slot is er gekeken of er sprake was van een testeffect waarbij respondenten gedurende de test steeds beter werden in het maken van een kansinschatting. Er werd gekeken of een hoger vraagnummer leidde tot een betere accuratesse score, oftewel hoe meer vragen een respondent beantwoordde, hoe accurater de respondent zou kunnen zijn. Voor elke respondent werd een regressielijn opgesteld in een spaghettiplot. Hierin werd het vraagnummer $V_i$ op de x-as en *accuratesse* op de y-as geplaatst. Door middel van lineaire regressieanalyse werd gekeken wat de helling was van de regressievergelijking. Deze

regressievergelijking luidt als volgt: $Accuratesse = a + b * V_i$. Waarbij de constante $b$ door middel van regressieanalyse is geschat. Een negatieve constante $b$ duidt op een testeffect.

## Resultaten

De steekproef bestaat uit 72 respondenten en bestaat uit 5 mannen en 67 vrouwen. Van de 72 respondenten hebben 21 respondenten vragen gehad met *quantile dot plots*, 18 respondenten zagen *gradiënt plots*, 14 respondenten hadden *violin plots* gezien en 19 respondenten zagen *error-bars*. Van de respondenten zijn 42 werkzaam in regio Noord, 10 in regio Oost en 20 in regio West. In regio Zuid zijn er helemaal geen respondenten werkzaam. De gemiddelde leeftijd van de respondenten bedraagt 30;2 jaar met een spreiding tussen de 20 en 61 jaar en de standaarddeviatie bedraagt 12.86. De gemiddelde leeftijd van studenten is rechtsscheef verdeeld en bedraagt 21;6 jaar met een spreiding tussen de 20 en 30 jaar en een standaarddeviatie van 1;10 jaar. De gemiddelde leeftijd voor de leerkrachten is normaal verdeeld en bedraagt 40;4 met een spreiding tussen de 22 en 61 jaar en een standaarddeviatie van 12;10 jaar. Van alle respondenten hebben 57 de vragenlijst volledig ingevuld, 15 respondenten zijn gedurende de vragenlijst afgehaakt. In Tabel 1 wordt gekeken in hoeverre de uitval plaatsvindt in verschillende groepen. Hierin zijn de onderlinge verhoudingen betreft *geslacht*, *provincie* en *plottype* opgenomen, opgesplitst in of de respondent de test volledig heeft afgerond of niet. Hieruit valt op dat het percentage van mensen dat de vragenlijst voortijdig afbreekt bij de *Quantile dot plots* aanzienlijk lager ligt (6.7%) dan het uitvalpercentage van de overige plots. Ook de uitval van studenten is vele malen hoger dan bij leerkrachten. Daarnaast valt op dat de uitval tussen mannen en vrouwen relatief gelijk is, waarbij de kanttekening gemaakt moet worden dat er slechts vijf mannelijke respondenten waren. Tot slot is er veel meer uitval zichtbaar in de regio Oost in vergelijking met regio Noord en West.

**Tabel 1**

*Kruistabel met Chi-kwadraat Waarde voor de Categorische Variabelen Geslacht, Provincie Plottype en Functie, Afgezet Tegen het wel of niet Voltooien van de Vragenlijst*

| Categoriale variabele | | Percentage | | Chi-kwadraat waarde | *p*-waarde |
|---|---|---|---|---|---|
| | | Wel voltooid (*N*=57) | Niet voltooid (*N*=15) | | |
| Geslacht | Man | 7% | 6.7% | .002 | .962 |
| | Vrouw | 93% | 93.3% | | |
| Provincie | Noord | 59.6% | 53.3% | 2.721 | .257 |
| | Oost | 10.5% | 26.7% | | |
| | Zuid | 0% | 0% | | |
| | West | 29.8% | 20% | | |
| Plottype | Quantile dot plot | 35.1% | 6.7% | 5.503 | .138 |
| | Gradiënt plot | 21.1% | 40% | | |
| | Violin plot | 17.5% | 26.7% | | |
| | Error-bar | 26.3% | 26.7% | | |
| Functie | Leerkracht | 50.9% | 26.7% | 2.804 | .094 |
| | Student | 49.1% | 73.3% | | |

**Tabel 2**

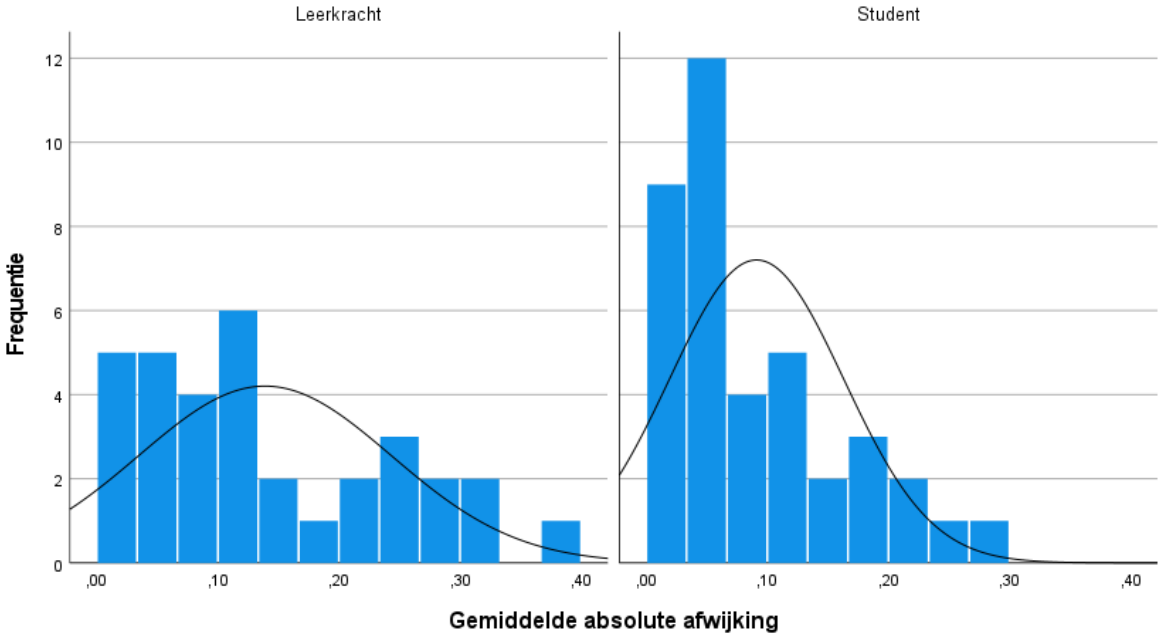
*Gemiddelde Absolute Afwijking van Leerkrachten en Studenten per Vraag*

| **Functie** | | **V1** | **V2** | **V3** | **V4** | **V5** | **V6** | **V7** | **V8** | **V9** | **V10** | **Tot** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Leerkracht** | Gemiddelde | 0.144 | 0.161 | 0.159 | 0.143 | 0.134 | 0.132 | 0.143 | 0.069 | 0.155 | 0.149 | 0.139 |
| | *N* | 33 | 33 | 33 | 33 | 32 | 32 | 30 | 30 | 29 | 29 | 33 |
| | Standaard-deviatie | 0.165 | 0.185 | 0.189 | 0.179 | 0.131 | 0.149 | 0.175 | 0.069 | 0.265 | 0.164 | 0.104 |
| **Student** | Gemiddelde | 0.103 | 0.071 | 0.097 | 0.092 | 0.073 | 0.106 | 0.087 | 0.056 | 0.071 | 0.066 | 0.091 |
| | *N* | 39 | 39 | 33 | 33 | 31 | 31 | 30 | 30 | 28 | 28 | 39 |
| | Standaard-deviatie | 0.120 | 0.094 | 0.132 | 0.165 | 0.069 | 0.117 | 0.136 | 0.074 | 0.132 | 0.082 | 0.072 |

Uit de analyse van Tabel 2 valt op dat de groep studenten op elke vraag een lagere gemiddelde absolute afwijking vertoonden. Uit de groepsgemiddelden blijkt dat de groep studenten een lagere gemiddelde absolute afwijking hebben van 4.8 procentpunten.

**Figuur 4**

*Histogrammen voor Gemiddelde Absolute Afwijking Voor Leerkracht (N=33) en Student (N=39)*




In Figuur 4 zijn de histogrammen met normaalcurve opgenomen voor de gemiddelde absolute afwijking in proporties. Na de logtransformaties waren de verdelingen van leerkracht en student bij benadering normaal verdeeld en weken de standaarddeviaties niet meer dan factor 2 af, hiermee kan gesteld worden dat er aan de assumpties voldaan is. Door de scheve verdelingen en de uitbijters in Figuur 4, is er gebruik gemaakt van de mediaan. De mediaan voor de groep leerkrachten duidt op een afwijking van 11.1 procentpunten en de mediaan voor de groep studenten duidt op een afwijking van 6.5 procentpunten.

De betrouwbaarheid van de uitkomstmaat is vastgesteld door gebruik te maken van de *Guttman's Lambda-2*. Met een waarde van .85 kan gesteld worden dat de uitkomstmaat in deze populatie betrouwbaar is.

**Tabel 3**

*Kruistabel met Chi-kwadraat Waarde voor de Categorische Variabelen Geslacht, Provincie en Plottype, Afgezet Tegen Functie*

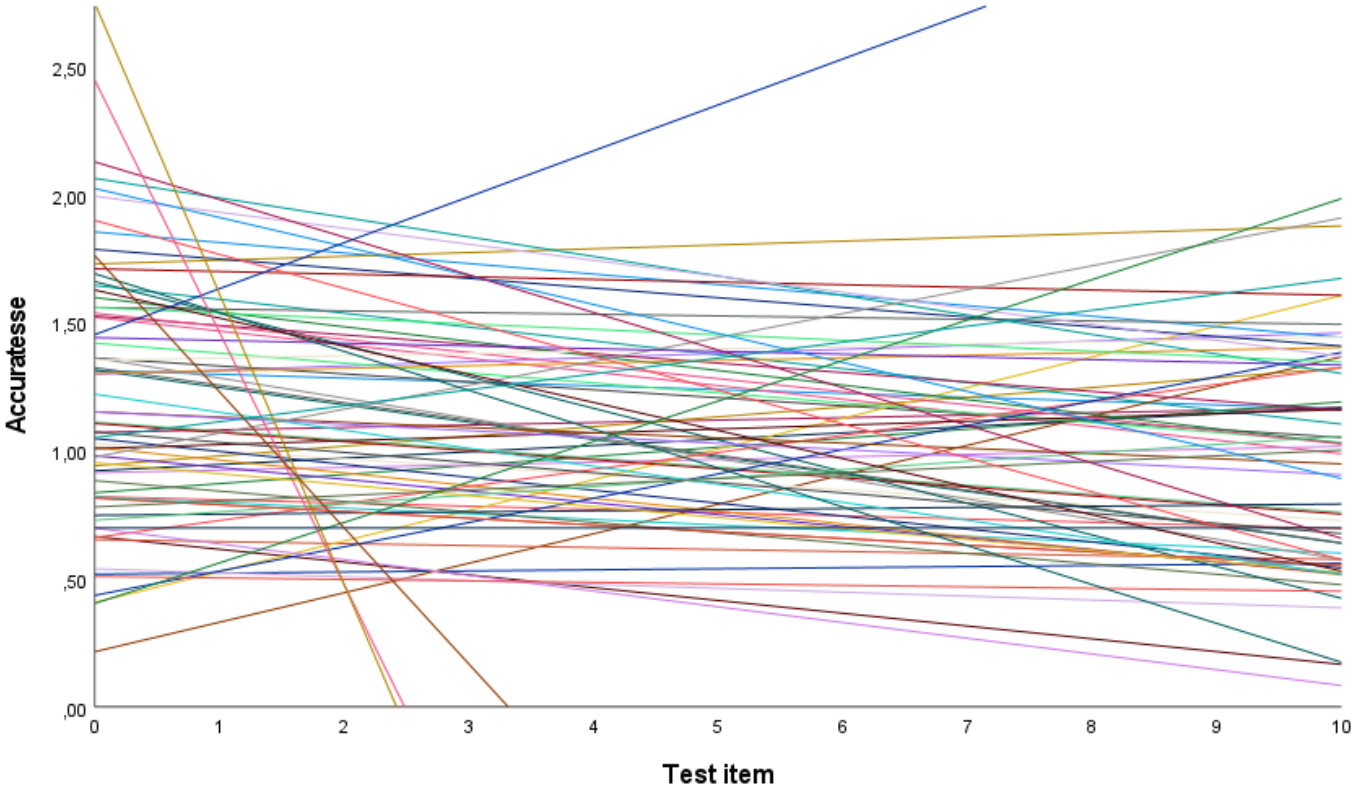
| Categoriale variabele | | Percentage | | Chi-kwadraat waarde | *p*-waarde |
|---|---|---|---|---|---|
| | | Leerkracht (*N*=33) | Student (*N*=39) | | |
| Geslacht | Man | 6.1% | 7.7% | .074 | .786 |
| | Vrouw | 93.9% | 92.3% | | |
| Provincie | Noord | 39.4% | 74.4% | 12.281 | .002 |
| | Oost | 27.3% | 2.6% | | |
| | Zuid | 0% | 0% | | |
| | West | 33.3% | 23.1% | | |
| Plottype | Quantile dot plot | 30.3% | 28.2% | 1.381 | .710 |
| | Gradiënt plot | 24.2% | 25.6% | | |
| | Violin plot | 24.2% | 15.4% | | |
| | Error-bar | 21.2% | 30.8% | | |

In Tabel 3 zijn de onderlinge verhoudingen van de steekproef door middel van een kruistabel weergegeven met de bijbehorende chi-kwadraatscores. Er zijn significant meer studenten uit Noord-Nederland in vergelijking met leerkrachten, bovendien zijn er meer leerkrachten uit Oost-Nederland in vergelijking met studenten $\chi^2(2) = 12.281$, $p = .002$. Er is geen significant verschil in *geslacht* tussen leerkrachten en studenten $\chi^2(1) = .074$, $p = .786$. Ook zijn er geen significante verschillen gevonden betreft *plottype* tussen leerkrachten en studenten, $\chi^2(3) = 1.381$, $p = .710$.

Er is geen significant verschil $t(70) = 1.584$, $p = .118$ in de gemiddelde accuratesse score van leerkrachten ten opzichte van studenten. Het bijbehorende 95%-betrouwbaarheidsinterval is [-0.336,0.294]. Om de effectgrootte te berekenen over het gemiddelde verschil is Cohen's *d* berekend, $d = .347$, wat duidt op kleine verschillen tussen leerkrachten en studenten.

**Figuur 5**

*Spaghetti Plot voor Testeffect voor Accuratesse Score per Participant (N=72)*




Tot slot blijkt uit de lineaire regressieanalyse dat zowel studenten als leerkrachten uit de steekproef steeds beter werden in het maken van de test. De regressiecoëfficiënt voor de groep leerkrachten is -0.023 en voor de groep studenten is deze -0.024. Dit betekent dat de meeste leerkrachten en pabo-studenten uit de steekproef steeds beter lijken te worden in het maken van een kansinschatting naarmate ze meer items beantwoorden. Als dit terug getransformeerd wordt naar een verschilscore, betekent dit dat een respondent na tien items 5 procentpunten beter scoort. In Figuur 5 zijn de visuele weergaven van alle regressievergelijkingen per participant te zien.

## Discussie

Het doel van dit onderzoek was om de verschillen van interpretatie van de foutenmarges op Cito-toetsen tussen leerkrachten en studenten in kaart te brengen. Hierdoor kon gekeken worden of leerkrachten de vaardigheden in het begin van hun loopbaan al beheersen of dat ze deze vaardigheid gedurende hun loopbaan ontwikkelen of verliezen. Uit de ongepaarde t-toets bleek er geen significant verschil te zijn in de gemiddelde accuratesse score tussen leerkrachten en studenten. Op basis van de analyse van de gemiddelden zijn de studenten uit de steekproef 4.8 procentpunten accurater met het geven van een kansinschatting bij een visuele weergave van een Cito-score dan leerkrachten. Dit gegeven is echter niet significant, waardoor deze conclusie niet voor de populatie getrokken kan worden. Met een grotere steekproef is dit mogelijk wel aan te tonen, maar de gevonden resultaten duiden op een klein effect.

Op basis van de resultaten kan gesteld worden dat de groep studenten uit de steekproef accuratere kansinschattingen maakten dan de groep leerkrachten. Dit staat in lijn met de verwachting van Resnick et al. (2018) dat studenten van nature effectieve strategieën gebruiken om visuele weergaven van data te interpreteren. Van der Kleij et al. (2014) stelde dat het aantal jaren ervaring van een leerkracht niet uitmaakte voor het maken van accuratere interpretaties op basis van Cito-scores. Door het ontbreken van significante data, kan er geen uitspraak gedaan worden voor de populatie.

Voor de externe validiteit van het onderzoek zijn er een aantal mogelijke bedreigers. Zo bestaat de steekproef voornamelijk uit leerkrachten en studenten uit Noord-Nederland. Een mogelijke verklaring hiervoor is dat het oorspronkelijke onderzoek (Ettema, 2021) uitgevoerd is door een masterstudent aan de Rijksuniversiteit Groningen, waardoor respondenten in die regio makkelijker te bereiken zijn. Dit maakt dat de steekproef representatiever is voor studenten uit Noord-Nederland dan uit Zuid-Nederland. Er kan echter

vanuit gegaan worden dat pabo-studenten en leerkrachten uit Noord-Nederland gemiddeld gezien niet verschillen in het analyseren van data ten opzichte van pabo-studenten en leerkrachten uit Zuid-Nederland, waardoor de gevonden conclusie alsnog te generaliseren valt. Uit de analyse met betrekking tot de uitval viel op dat participanten die een *gradiënt plot* te zien kregen, veel vaker afhaakten gedurende de vragenlijst. Mogelijk vonden participanten dit type plot lastiger en besloten daarom vroegtijdig de vragenlijst te beëindigen. Dit gegeven is echter niet significant en werd geconcludeerd op basis van kleine aantallen, waardoor deze uitspraak alleen geldt voor de steekproef. Dit betekent dat er een vertekend beeld is tussen de gegevens waar de conclusies op gebaseerd worden en welke gegevens ontbreken. Van de respondenten die afgehaakt zijn, kan ervan uitgegaan worden dat deze respondenten de resterende vragen slechter zouden maken. Door het vroegtijdig beëindigen van de vragenlijst, vallen de resultaten positiever uit. Hierdoor kunnen conclusies op basis van de verschillende visualisatie lastiger gemaakt worden. Daarnaast is een testeffect bij de respondenten aangetoond. De meeste respondenten lijken beter te worden naarmate ze meer items beantwoorden. Dit testeffect geeft aan dat deze vaardigheid mogelijk ook te trainen is.

Wat betreft de interne validiteit kan er gesteld worden dat er voldaan is aan de assumpties van normaliteit en onafhankelijke waarnemingen van de statistische toetsen. Om het effect van uitbijters af te zwakken, is er gebruik gemaakt van een transformatie van de bestaande dataset. Hiervoor is voor elke vraag van de respondenten een $\log_{10}$-score berekend. De betrouwbaarheid van de uitkomstmaat is aangetoond door een Guttman's lambda-2 van .85.

Eventueel vervolgonderzoek met een grotere steekproef zou mogelijk een significant resultaat vinden op de gevonden onderzoeksvraag. Echter is vervolgonderzoek de hoeveelheid tijd, geld en middelen niet waard. Door de resultaten op basis van de steekproef zou een grootschaliger onderzoek waarschijnlijk dezelfde resultaten opleveren, met als enige

verschil dat het dan wel significant aangetoond is. Waar de groep leerkrachten uit de steekproef wijkt gemiddeld 13.9 procentpunten af van de daadwerkelijke score, wijkt de groep studenten 9.1 procentpunten af van de daadwerkelijke score. Aangezien het hier gaat om een gemiddelde kan ervan uit gegaan dat er veel leerkrachten zijn die meer dan 13.9 procentpunten afwijken. Bij de Cito-niveaugroepen I tot en met V wordt er gerekend met intervallen van 20 procent. Waardoor een leerkracht een leerling al vrij snel in een ander niveaugroep toebedeeld dan waar de desbetreffende leerling gezien zijn vaardigheidsscore eigenlijk in thuishoort. Dit kan ervoor zorgen dat een leerling met een III-score op de Cito-toets toebedeeld kan worden aan niveaugroep II of niveaugroep IV, waarbij de instructiestrategie erg verschilt.

Op basis van de gevonden resultaten is het aan te raden om tijdens de lerarenopleidingen in Nederland aandacht te besteden aan meetfouten bij het de interpretatie van vaardigheidsscores van Cito-toetsen. Hierdoor zou het begrip van de leerkrachten in opleiding toenemen en zullen zij accuratere schattingen maken op basis van de opgedane kennis. Met als gevolg dat leerlingen eerder toegewezen worden aan de niveaugroep dat aansluit bij de instructiebehoeften. Aangezien leerkrachten in de meeste gevallen niet weten wat meetfouten zijn of passen zij deze kennis niet toe (Van der Kleij & Eggen, 2013), is het ook noodzakelijk dat zij deze kennis tot zich nemen. Een mogelijke oplossingen hiervoor is het aanbieden van trainingen voor leerkrachten of door gebruik te maken van de kennis die de afgestudeerde pabo-studenten hebben opgedaan tijdens hun studie over dit onderwerp. Hierdoor worden zowel de kennis over en het gebruik van de meetfouten bij cito-scores bij leerkrachten vergroot, waardoor zij het onderwijs beter af kunnen stemmen op de instructiebehoeften van het kind (Saunders, 2000; Earl & Fullan, 2003; Van Petegem & Vanhoof, 2004; Kerr et al., 2006; Williams & Coles, 2007; Ledoux et al., 2009; Zupanc et al., 2009; Meijer et al., 2011).

## Literatuurlijst

Bunck, M. J. A., Terlien, E., van Groenestijn, M., Toll, S. W. M., & Van Luit, J. E. H. (2017). Observing and analyzing children's mathematical development, based on action theory. *Educational Studies in Mathematics : An International Journal, 96*(3), 289–304. https://doi.org/10.1007/s10649-017-9763-6

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

Cito. (2019). *Toetsscore, vaardigheidsscore. . . en dan?* https://www.cito.nl/-/media/files/ve-en-po/cito-flyer-toetsscore-vaardigheidsscore-en-dan.pdf?la=nl-NL

Cito. (z.d.). *Schooladvies: Cito levert een stukje van de puzzel*. Geraadpleegd op 28 februari 2022, van https://www.cito.nl/onderwijs/primair-onderwijs/schooladvies-cito-levert-stukje-puzzel

Drenth, P. J. D., & Sijtsma, K. (2005). *Testtheorie*. Houten: Bohn Stafleu van Loghum.

Earl, L., & Fullan, M. (2003). Using data in leadership for learning. *Cambridge Journal of Education, 33*(3), 383–394.

Ettema, B. (2021). *Visuele weergave van onzekerheid in Cito testscores.* Rijksuniversiteit Groningen.

Frans, N., Post, W., Oenema-Mostert, C., & Minnaert, A. (2020). Signalering met de Cito kleutertoetsen: Ondergemiddeld is niet gelijk aan problematisch. *Tijdschrift voor Orthopedagogiek, 59*(2), 20-27.

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

Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS Rekenen-Wiskunde voor groep 3 tot en met 8.* Arnhem: Cito

Kerr, K. A., Marsch, J. A., Ikemoio, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education, 112*(4), 403–420.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Voorwaardelijke standaard meetfouten voor schaalscores met behulp van IRT. *Journal of Educational Measurement, 33*(2), 129-140. doi: https://doi.org/10.1111/j.1745-3984.1996.tb00485.x.

Ledoux, G., Blok, H., Boogaard, M., & Krüger, M. (2009). Opbrengstgericht werken; over de waarde van meetgestuurd onderwijs. [Data-driven decision making; About the value of measurement oriented education]. Amsterdam, The Netherlands: SCO-Kohn Stamm Instituut.

Meijer, J., Ledoux, G., & Elshof, D. P. (2011). Gebruikersvriendelijke leerlingvolgsystemen in het primair onderwijs. [User-friendly pupil monitoring systems in primary education]. Amsterdam: SCO Kohnstamm Institute.

Padilla, L., Kay, M., & Hullman, J. (2020). Uncertainty Visualization. *Handbook of Computational Statistics and Data Science*

Resnick, I., Kastens, K. A., & Shipley, T. F. (2018). How students reason about visualizations from large professionally collected data sets: a study of students approaching the threshold of data proficiency. *Journal of Geoscience Education*, *66*(1), 55–76. https://doi.org/10.1080/10899995.2018.1411724

Saunders, L. (2000). Understanding schools' use of 'value added' data: The psychology and sociology of numbers. *Research Papers in Education, 15*(3), 241–258.

Van der Bles, A. M., Van der Linden, S., Freeman, A. L. J., Mitchell, J., Galvao, A. B., Zaval, L., & Spiegelhalter, D. J. (2019). Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, *6*(5), [181870]. https://doi.org/10.1098/rsos.181870

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

Van der Kleij, F., 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.

Van Petegem, P., & Vanhoof, J. (2004). Feedback over schoolprestatieindicatoren als strategisch instrument voor schoolontwikkelingen [Feedback about school per formance indicators as a strategic instrument for school development]. *Pedagogische Studiën, 81*, 338–353.

Williams, D., & Coles, L. (2007). Teachers' approaches to finding and using research evidence: An information literacy perspective. *Educational Research, 49*, 185–206 http://dx.doi.org/10.1080/00131880701369719.

Zupanc, D., Urank, M., & Bren, M. (2009). Variability analysis for effectiveness and improvement in classrooms and schools in upper secondary education in Slovenia: Assessment of/for learning analytic tool. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 20*, 89–122 http:// dx.doi.org/10.1080/09243450802696695.