**Omgaan met meetnauwkeurigheid Cito scores: Invloed van verschillende visualisaties op de gemaakte kansinschattingen die leerkrachten maken.**

Student: Elise Boelen (S4129229)

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

Rijksuniversiteit Groningen
Faculteit der Gedrags- en Maatschappijwetenschappen
Bachelor werkstuk Pedagogische Wetenschappen
Juni 2022

## Abstract

About 85% of Dutch primary schools use the Pupil and Education Tracking System (*Leerling Onderwijs VolgSysteem*; LOVS). The progress and results achieved by students are recorded and tracked in the LOVS. Teachers regularly forget that test results for all types of measurements are subject to measurement error. Confidence intervals can help with the interpretation of test results of the LOVS, but teachers indicate that they experience difficulties in interpreting the uncertainty of test results. Previous research has shown that visualization with a confidence interval will influence the educational decisions of the teachers compared to visualization without a confidence interval. The research question explored in this research is: 'to what extent do the different visualizations influence the probability estimates that teachers make when interpreting CITO scores?'. In an online questionnaire, 72 teachers and (academic) pabo students were randomly shown one of four visualizations: Error-bar, Violin plot, Quantile dot plot and Gradient plot. Furthermore we investigated what type of mistakes teachers make in their probability estimates with these visualizations. For this visualization participants were shown ten different scores and were asked to give a probability estimate for each score in relation to a certain cutoff. Regression analysis was used to analyze the accuracy of probability estimates. The results show that in all visualizations participants differ significantly from a perfect probability estimate. The Quantile dot plot differs significantly from the Error-bar. Also the Violin plot differs significantly from the Error-bar when the outliers were excluded from the data. This means that participants with the Quantile dot plot and the Violin plot are significantly more accurate compared to the Error-bar. Gradient plot showed no significant difference from the Error-bar. It is important for test results to be interpreted correctly, because decisions have consequence for teaching and learning in the Dutch primary education.

## Inleiding

Volgens van der Kleij en Eggen (2013) maakt nu 85 procent van de Nederlandse scholen in het regulier en speciaal basisonderwijs gebruik van het Leerling Onderwijs VolgSysteem (LOVS) van Cito. De vorderingen en resultaten die leerlingen behalen worden bijgehouden in het LOVS. In het LOVS staan alle opgaven van een bepaald leergebied op dezelfde vaardigheidsschaal. Hierdoor volg je de ontwikkeling van een leerling door de jaren heen. Dit stelt leerkrachten in staat om een voor- of achteruitgang te zien bij leerlingen op een bepaald leergebied. Voor de volgende leergebieden worden de vorderingen en resultaten bijgehouden in het LOVS: rekenen-wiskunde, spelling en begrijpend lezen (Cito, 2019).

Regelmatig wordt vergeten dat testresultaten van alle soorten metingen onderhevig zijn aan een bepaalde meetfout. Deze meetfout staat voor het toevallige deel in de werkelijke score. De werkelijke score is de score die een leerling behaald als er geen meetfouten in de test zijn opgenomen (Charter & Feldt, 2002). Het toevallige deel is de verschillende factoren die van invloed zijn op de geschatte score. Deze factoren hebben geen verband met de werkelijke score. Het verschil tussen de werkelijke score en de geschatte score is de meetfout. Een voorbeeld hiervan is: op een Cito rekentoets meet de werkelijke score de rekenvaardigheid van een leerling. De geschatte score meet de rekenvaardigheid van een leerling en verschillende factoren die van invloed zijn op de score van een leerling. Een leerling kan een vraag goed gokken. Deze vraag kan niet worden toegeschreven aan de rekenvaardigheid van deze leerling (Sijtsma & Drenth, 2006; Hopster-den Otter et al., 2018; Gardner, 2013).

Leerkrachten krijgen binnen het onderwijs regelmatig te maken met verschillende beslissingen. Voorbeelden van beslissingen zijn: toewijzen van studenten aan een geschikte instructiegroep; beslissen of een student extra ondersteuning nodig heeft en beslissingen over vervolgstappen in instructie aan leerlingen (Van der Kleij & Eggen, 2013; Meijer et al., 2011; Goodman & Hambleton, 2004; Newton, 2005; Phelps et al., 2010). Leerkrachten gaven aan moeite te ervaren met het interpreteren van informatie over de meetnauwkeurigheid van de werkelijke score. Wat mogelijk leidt tot het maken van verkeerde beslissingen (van der Kleij & Eggen, 2013; Ledoux et al., 2009; Meijer et al., 2011; Zwick et al., 2014). Deze beslissingen hebben consequentie voor leren en lesgeven in het Nederlandse basisonderwijs. Het is van belang dat testresultaten op de juiste manier worden geïnterpreteerd waardoor leerkrachten juiste beslissingen maken. Hierdoor wordt tegemoet gekomen aan de

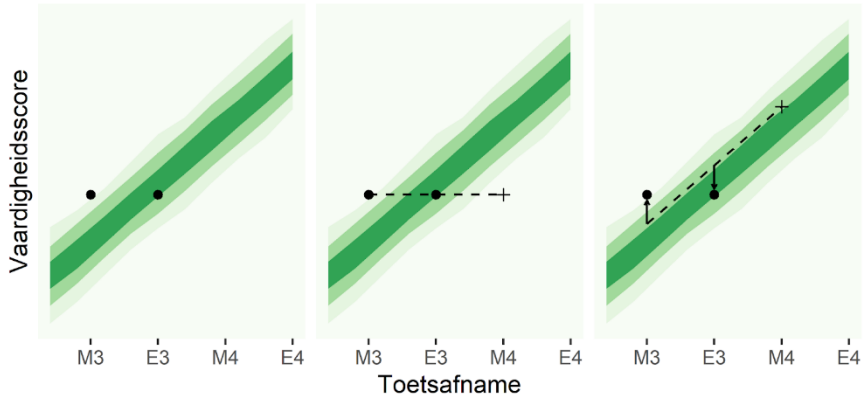
onderwijsbehoeften van leerlingen (Hopster-den Otter et al., 2018; Fullan & Watson, 2000; Vanhoof et al., 2011).

Betrouwbaarheidsintervallen kunnen leerkrachten helpen om de meetnauwkeurigheid van testresultaten te interpreteren. Cito hanteert een 68%-betrouwbaarheidsinterval om leerkrachten te informeren over de meetnauwkeurigheid van de werkelijke score (Janssen et al., 2010). Betrouwbaarheidsintervallen geven leerkrachten de geschatte score met daaromheen een interval. Dit betrouwbaarheidsinterval geeft informatie over de onzekerheid van testresultaten. Het kan voorkomen dat de werkelijke score net buiten het gegeven betrouwbaarheidsinterval valt. Leerkrachten kunnen hierbij de fout maken door een kansinschatting te geven van 0%. Bij een 68%-betrouwbaarheidsinterval is de kans dat de werkelijke score in het betrouwbaarheidsinterval valt niet 0% maar 16%, 16% onder het betrouwbaarheidsinterval en 16% boven het betrouwbaarheidsinterval. Door dit gegeven kunnen leerkrachten mogelijke foute beslissingen maken (Charter & Feldt, 2002).

Ook kan in het LOVS de ontwikkeling van een leerling in de tijd gevolgd worden. Figuur 1 geeft een voorbeeld van twee vaardigheidsscores van een individuele leerling. Wanneer leerkrachten uit gaan van twee perfect accurate scores kan een stagnerende groei geconstateerd worden (tweede paneel). In Paneel 3 wordt uit gegaan van twee scores met daaromheen een meetfout. Hierbij kan geconstateerd worden dat deze leerling een groei laat zien op een bovengemiddeld niveau. Doordat testresultaten onderhevig zijn aan bepaalde meetfouten kunnen dezelfde scores op verschillende manieren geïnterpreteerd worden (Hopster-den Otter et al., 2018; Charter & Feldt, 2002).

**Figuur 1**

*Vaardigheidsscore (y-as) van een M3- en E3-toets (x-as) van een individuele leerling.*

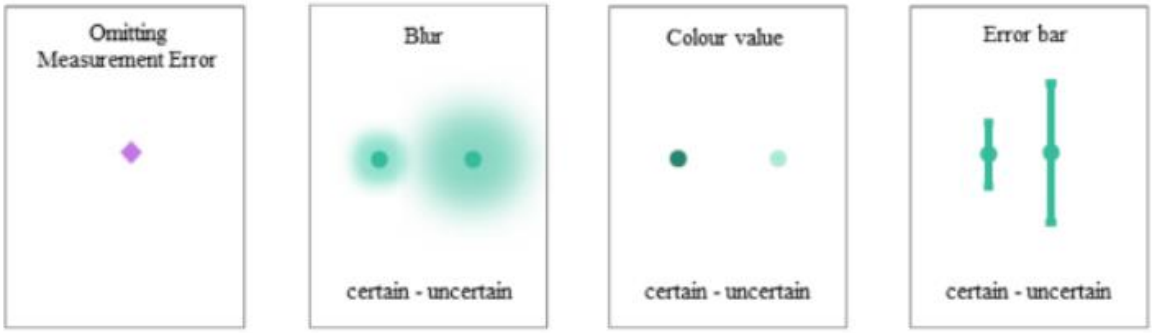



Hopster-den Otter et al. (2018) hebben een mixed-methods onderzoek uitgevoerd om de invloed van verschillende visualisaties op de beslissingen en voorkeuren van leerkrachten

in kaart te brengen. Deze visualisatie zijn *Blur, Colour value en Error-bar* (Figuur 2). Hopster-den Otter et al. (2018) toonden aan dat visualisaties met een betrouwbaarheidsinterval de educatieve beslissingen van leerkrachten beïnvloeden in vergelijking tot visualisaties zonder betrouwbaarheidsinterval. De *Error-bar* zorgde voor een significante grotere vraag naar aanvullende informatie over een leerling. Daarentegen zorgde de visualisatie *Colour value* voor een significante mindere vraag naar aanvullende informatie over een leerling. De visualisatie *blur* verschilde niet significant van de visualisatie zonder betrouwbaarheidsinterval. Een conclusie uit het onderzoek van Hopster-den Otter et al. (2018) is dat een combinatie van de visualisatie *Blur* en *Error-bar* (e.g. een *gradient plot,* zie Figuur 3) een geschikte presentatie kan zijn van een betrouwbaarheidsinterval bij testresultaten. De *Error-bar* heeft een positieve invloed op de beslissingen en voorkeuren van leerkrachten. In combinatie met *Blur,* een natuurlijke associatie van onzekerheid waardoor een exacte interpretatie wordt vermeden.

**Figuur 2**

*De visualisaties Blur, Colour value en Error bar uit het onderzoek van Hopster-den Otter et al., (2018).*

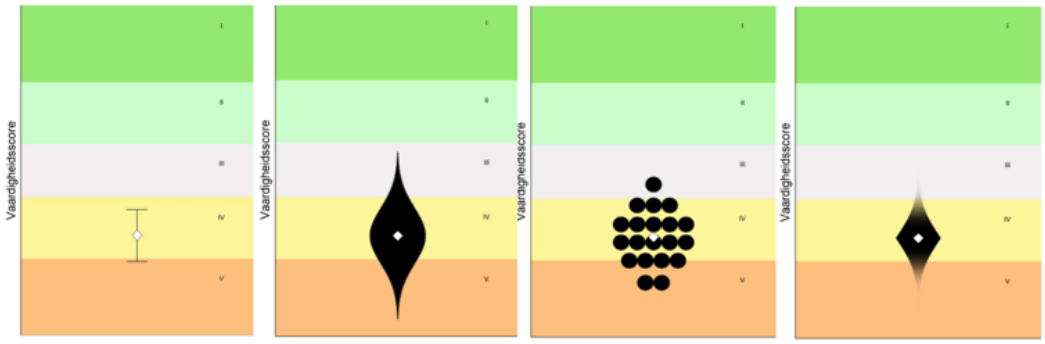



In dit onderzoek wordt gefocust op vier verschillende visualisaties. Deze vier verschillende visualisatie zijn: *Error-bar, Violin plot, Quantile dotplot en Gradiënt plot* (Figuur 3). Eerder onderzoek heeft aangetoond dat de *Error-bar* een positieve invloed heeft op de beslissingen en voorkeuren van leerkrachten (Hopster-den Otter et al., 2018). De *Violin plot* geeft de verdeling van een gespiegelde histogram weer (Hullman et al., 2015). Kay et al. (2016) creëerde een *Quantile dotplot* als alternatief voor het weergeven van onzekerheid voor een continue variabele. Een *Quantile dotplot* is een verdeling waarin punten evenredig worden geclusterd. Elke stip geeft een waarschijnlijkheid van 5% op een bepaalde score weer. Zoals hierboven beschreven kan een *Gradiënt plot* een geschikte presentatie zijn van een betrouwbaarheidsinterval bij testresultaten. De *Violin plot* en *Quantile dotplot* zijn beide

visualisaties die ook een duidelijke verdeling van meetfouten laten zien, terwijl *Blur* en *Colour value* uit het onderzoek van Hopster-den Otter et al. (2018) juist geen duidelijke verdeling van meetfouten weergeven. Een duidelijke verdeling van meetfouten zorgt mogelijk voor een betere kwaliteit van beslissingen (Padilla et al., 2020). In dit onderzoek wordt gekeken of deze verschillende visualisaties van invloed zijn op de kansinschattingen die leerkrachten maken. Door dit te onderzoeken kan gekeken worden of bepaalde visualisaties leerkrachten kunnen helpen testresultaten te interpreteren waardoor leerkrachten juiste beslissingen nemen.

**Figuur 3**

*Voorbeeld Visualisaties voor vaardigheidsscore en meetnauwkeurigheid: Error-bar, Violin-plot, Quantile dotplot en Gradiëntplot.*




Ook wordt onderzocht wat voor soort fouten leerkracht maken in het maken van kansinschattingen. Zoals eerder beschreven is een mogelijke fout: het geven van een kansinschatting van 0% dat de werkelijke score in het betrouwbaarheidsinterval valt. Voor deze foutenanalyse wordt gekeken of bij een bepaalde visualisatie meer verwachte fouten gemaakt worden dan bij een andere visualisatie. Op basis van bovenstaande gegevens ontspringt de volgende onderzoeksvraag: ‘In hoeverre hebben de verschillende visualisaties invloed op de kansinschattingen die leerkrachten maken bij het interpreteren van citoscores?’.

## Methode

### Design

Dit onderzoek had als doel de invloed van visualisaties op de accuratesse van kansinschattingen die leerkrachten maken bij het interpreteren van citoscores te onderzoeken. Dit onderzoek werd uitgevoerd door gebruik te maken van een bestaande dataset (Ettema, 2021). Deze bestaande dataset werd in dit onderzoek geheranalyseerd. In de vorige analyse werd de data bekeken aan de hand van ratio scores. In dit onderzoek werd gekeken naar

verschilscores. Deze heranalyse werd gedaan, omdat bij het gebruik van ratio scores een probleem ontstaan bij het analyseren van kleine kansinschattingen. In de bestaande dataset was gekozen voor een experimenteel between-subject design. Elke participant kreeg willekeurig één van de vier visualisaties van meetnauwkeurigheid toegewezen. De verschillende visualisaties is de onafhankelijke variabele waarbij vier visualisaties van onzekerheid bekeken werden. De accuratesse van kansinschattingen van leerkrachten is de afhankelijke variabele.

**Steekproef en populatie**

De doelpopulatie in dit onderzoek waren Nederlandse leerkrachten of derdejaars (academische) PABO studenten. De Dataset bevat 72 Participanten waarvan 33 leerkrachten en 39 (academische) PABO studenten. Deze participanten kregen één van de vier visualisaties te zien. Door middel van een gelegenheidssteekproef werden de participanten benaderd. De inclusiecriteria voor deze steekproef waren ten eerste dat de participant een basisschoolleerkracht of derde of hogerejaars (academische) PABO student is. Ten tweede moest de participant in Nederland wonen. Ten derde moest de participant bekend zijn met het leerlingvolgsysteem van Cito.

**Instrumenten en variabelen**

De onafhankelijke variabele in dit onderzoek was de visualisaties van meetnauwkeurigheid. Deze vier visualisaties van meetnauwkeurigheid zijn uitgesplitst in vier verschillende visualisaties. Deze vier visualisaties zijn: *Error-bar, Violin-plot, Quantile dotplot en Gradiënt plot* (Figuur 3). De visualisatie gradiënt plot is geen standaard visualisatie. De uiteinden van deze visualisatie lopen taps waardoor zowel de vorm als de '*Blur*' van deze visualisatie de onzekerheid weergeven.

De afhankelijke variabele in dit onderzoek is de accuratesse van kansinschattingen van leerkrachten. Dit werd gemeten door het invullen van een vragenlijst. Voor deze vragenlijst was de Kans Inschatting Werkelijke Score schaal (KIWS-schaal) ontworpen. Door middel van een willekeurige toewijzing kregen de participanten één van de vier visualisaties te zien. Voor deze visualisatie kregen de participanten vijf keer een verschillende vaardigheidsscore te zien, met bijbehorende meetfout, waarvoor de participant een kansinschatting moest geven. Voor de vaardigheidsniveaus in deze visualisaties zijn de percentielgroepen van Cito gehanteerd. De cito scores en bijbehorende meetfout waren gebaseerd op normdata van de Cito rekentoets M5 2012 (Janssen et al., 2010). Bij de toetsen

van het Cito leerlingvolgsysteem worden de uitslagen van de toetsen omgezet in scores. Deze scores variëren van I tot en met V. Het hoogste niveau is I en het laagste niveau is V (Van der Kleij & Eggen, 2013). Figuur 3 toont voorbeelden van de visualisaties zoals participanten deze zagen in de vragenlijst. Bij deze visualisaties werden vragen gesteld als: 'hoe hoog schat u de kans in dat deze leerling een III scoort' of 'hoe hoog schat u de kans in dat deze leerling lager dan een III scoort'. De participanten konden een kans inschatten door een percentage te geven tussen de 0% en 100%. Van dit percentage wordt een proportie gemaakt door dit percentage te delen door 100.

Naast de twee centraal gestelde variabelen werden ook vier andere variabelen meegenomen. Deze variabelen werden meegenomen om de representativiteit van de steekproef vast te stellen. Ook kunnen deze variabelen invloed hebben op de accuratesse van kansinschattingen die leerkrachten maken of een andere verklaring hiervoor vormen. De eerste variabele die werd meegenomen is 'geslacht'. Deze variabele is opgesplitst in 'man' en 'vrouw'. De tweede variabele die werd meegenomen is leeftijd. De participanten konden voor deze variabele hun leeftijd in jaren invullen. De derde variabele die werd meegenomen is provincie. Participanten konden in de vragenlijst invullen in welke provincie zij woonachtig zijn. De variabele provincie werd gespecificeerd in vier groepen. De provincies Groningen, Friesland en Drenthe werden regio 'Noord', de provincies Overijssel, Gelderland en Flevoland werden regio 'Oost', de provincies Noord-Brabant en Limburg werden regio 'Zuid' en de provincies Utrecht, Noord-Holland, Zuid-Holland en Zeeland werden regio 'West' (Janssen et al., 2010). De vierde variabele die werd meegenomen is leerkracht/student. Deze variabele werd opgesplitst in: 'leerkracht' en 'student'. Hierbij zijn de PABO studenten en de academische PABO studenten samengevoegd.

**Procedure**

De dataverzameling vond plaats van december 2020 tot en met januari 2021. Voor dit onderzoek is toestemming verleend van de ethische commissie PedOn. De participanten moesten een online vragenlijst invullen in Qualtrics. Voor het verzamelen van de data zijn de participanten benaderd via mail en sociale media. Op deze manier werden vrienden en familie die voldoen aan de steekproef criteria gemotiveerd om deel te nemen aan het onderzoek. Ook werd het sneeuwbaleffect ingezet. Participanten die de vragenlijst hadden ingevuld, werden gevraagd om de vragenlijst in hun eigen netwerk te verspreiden. Daarnaast waren verschillende PABO opleidingen benaderd om de vragenlijst binnen PABO opleidingen te verspreiden. Voorafgaand aan de vragenlijst kregen de participanten een uitgebreide

informatiebrief met hierin een korte uitleg over het doel van het onderzoek. Ook werd de anonimiteit gewaarborgd door geen persoonlijke gegevens van de participanten te verzamelen waardoor de antwoorden niet te herleiden zijn tot een individu. De vrijwilligheid werd gewaarborgd doordat participanten op elk moment van het onderzoek mochten stoppen. Als laatste is het mailadres van de onderzoekers benoemd, dit voor mogelijke vragen van de participanten over het onderzoek.

**Statistische Analyses**

Zoals eerder beschreven is voor dit onderzoek gebruik gemaakt van een bestaande dataset. Een heranalyse werd gedaan omdat bij het gebruik van ratio scores een probleem ontstaat bij het analyseren van kleine kansinschattingen. Gekeken naar ratio scores waren namelijk bij kleine kansinschattingen al snel grote relatieve verschillen te zien, terwijl het absolute verschil tussen de kansinschatting en de werkelijke kansinschatting een stuk kleiner is. Om dit probleem te voorkomen, werd in dit onderzoek de nadruk gelegd op de absolute verschillen en werd de data bekeken aan de hand van verschilscores.

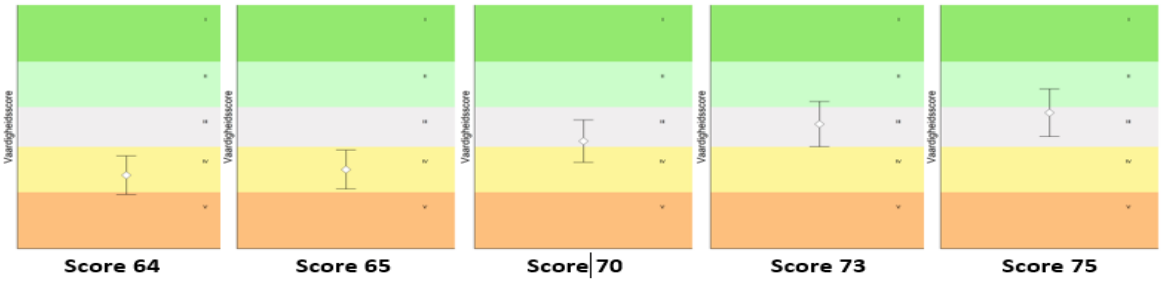
De gegeven antwoorden van de participanten op de vragenlijst zijn verwerkt in het programma IBM SPSS Statistics 27. Om de accuratesse van kansinschattingen per visualisatie te bekijken werd een verschilscore gemaakt van elke kansinschatting. Deze verschilscore is de kansinschatting op item *i* gegeven door de participant ($\hat{p}_i$) minus de daadwerkelijke kans ($P_i$). Van deze verschilscore werd een absolute waarde genomen. Hiermee kan gezien worden hoeveel het antwoord gegeven door de participant afwijkt van het daadwerkelijk antwoord. Met deze variabele werd de beschrijvende statistiek uitgevoerd. Eerst werd een frequentietabel gegeven voor de categorische variabelen ‘geslacht’, ‘leerkracht/student’, ‘visualisatie van meetnauwkeurigheid’ en ‘provincie’. Ook werd gekeken naar missing data bij deze categorische variabelen. Om mogelijke verbanden te ontdekken tussen de visualisaties van meetnauwkeurigheid en de categorische variabele leerkracht/student werd de Chi-kwadraat toets ingezet. De Chi-kwadraat toets werd berekend aan de hand van kruistabellen. Daarnaast werden de kenmerken van de continue variabele ‘leeftijd’ omschreven. Om een verband te ontdekken tussen de visualisaties van meetnauwkeurigheid en de andere continue variabele, leeftijd, werd de ANOVA analyse ingezet. De kenmerken van de variabele accuratesse van kansinschattingen werd in kaart gebracht met een boxplot.

Voor de analyse van de accuratesse van kansinschattingen werd de lineaire regressieanalyse ingezet met een 95% betrouwbaarheidsinterval. Om deze regressieanalyse

uit te voeren werd de verschilscore getransformeerd, omdat de variabele accuratesse van kansinschattingen scheef is verdeeld. Hiervoor werd een logaritme genomen van de verschilscore om de verdeling normaal verdeeld te krijgen. De formule hiervoor is: ($\log_{10}(\hat{p} - p + 0.005) - \log_{10}(0.005)$. Bij $\hat{p} - p$ werd 0.005 bij opgeteld. Dit werd gedaan omdat $\log_{10}(0)$ niet gedefinieerd is. Om de interpretatie dat een score van 0 perfect accuraat is te behouden, werd vervolgens van deze formule $\log_{10}(0.005)$ weer afgehaald. Hierna werden voor alle vragen die de participant heeft beantwoord een gemiddelde score berekend waardoor er één afhankelijke variabele ontstaat. Voor alle analyses werd een significantieniveau gebruikt van .05.

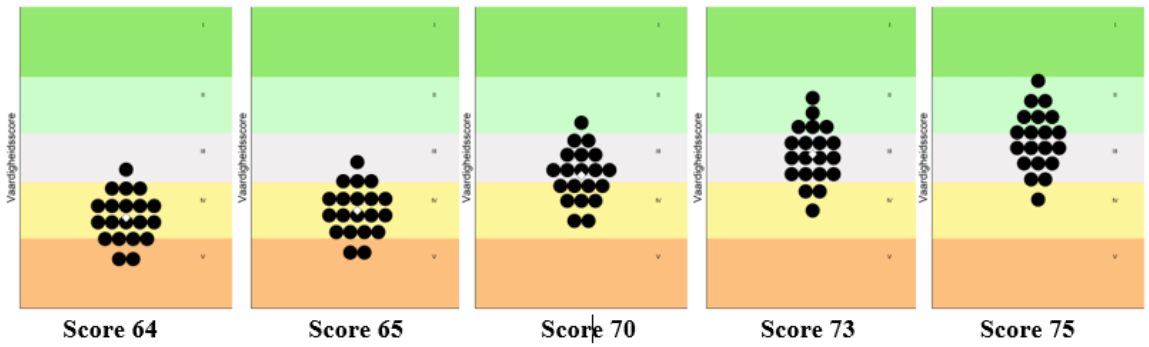
Ook moet de data voldoen aan een aantal assumpties. Ten eerste moeten de waarnemingen in de vier verschillende visualisaties normaal verdeeld zijn. Deze assumptie werd beoordeeld door middel van vier boxplots. Ten tweede moeten de varianties tussen de vier visualisaties vergelijkbaar zijn aan elkaar. Deze assumptie werd beoordeeld door middel van de standaarddeviaties. De kleinste standaarddeviatie is niet meer dan twee keer zo groot dan de grootste standaarddeviatie. Ten derde moeten de waarnemingen onafhankelijk zijn van elkaar. Deze laatste assumptie is in dit onderzoek niet toetsbaar.

**Figuur 4**
*De visualisatie Error-bar per vaardigheidsscore.*




Voor de foutenanalyse werd gekeken naar vier verschillende vaardigheidsscores bij de verschillende visualisaties. In Figuur 4 is een overzicht gegeven van de visualisatie *Error-bar* per vaardigheidsscore. Bij de verschillende vaardigheidsscores werden telkens twee vragen gesteld. Bij een deel van deze vragen konden participanten mogelijk een kansinschatting geven van 0%. Een voorbeeld hiervan is: bij de vaardigheidsscore van 64 zijn de volgende twee vragen gesteld: ‘’In hoeverre is het werkelijke niveau lager dan IV?’’ en ‘’In hoeverre is het werkelijke niveau III?’’. Bij deze twee vragen konden participanten mogelijk een kansinschatting geven van 0%, omdat het betrouwbaarheidsinterval niet binnen niveau III valt en net binnen niveau V valt.

**Figuur 5**
*De visualisatie Quantile dotplot per vaardigheidsscore.*




Bij de *Quantile dotplot* liggen in sommige gevallen de bolletjes precies op de horizontale lijn van de verschillende niveaus. Bijvoorbeeld bij vaardigheidsscore 64 vallen de bolletjes precies op de horizontale lijn tussen niveau III en niveau IV in. Dit is het geval bij de vragen: ‘’In hoeverre is het werkelijke niveau III?’’ (vaardigheidsscore 64), ‘’In hoeverre is het werkelijke niveau lager dan IV?’’ (vaardigheidsscore 65), ‘’In hoeverre is het werkelijke niveau lager dan III?’’ (vaardigheidsscore 73) en ‘’In hoeverre is het werkelijke niveau IV?’’ (vaardigheidsscore 73) (Figuur 5). Om een mogelijk verband te ontdekken tussen de twee gemiddelden, de vier vragen waarbij de bolletjes precies op de horizontale lijn liggen van de verschillende niveaus en de overige zes vragen, werd een *t*-toets uitgevoerd.

## Resultaten

**Tabel 1**
*Frequentietabel van alle categorische variabelen uit de steekproef (N=72).*

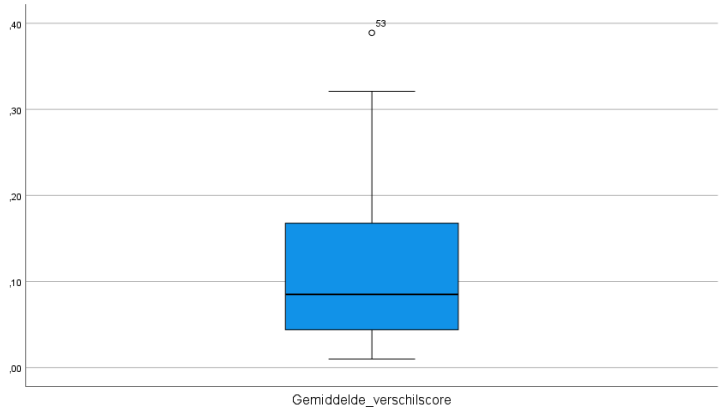
| Variabelen | Frequentie (missing) | Percentage (missing) |
|---|---|---|
| Geslacht | | |
| Man | 5 | 6.9% |
| Vrouw | 67 | 93.1% |
| Leerkracht/student | | |
| Leerkracht | 33 | 45.8% |
| Student | 39 | 54.2% |
| Provincie | | |
| Noord | 33 | 45.8% |
| Oost | 19 | 26.4% |
| Zuid | 0 | 0% |
| West | 20 | 27.8% |
| Visualisaties | | |
| Quantile dot plot | 21 (8) | 29.2% (1.1%) |
| Gradiënt plot | 18 (38) | 25.0% (5.3%) |
| Violin plot | 14 (22) | 19.4% (3.1%) |
| Error-bar | 19 (16) | 26.4% (2.2%) |

In Tabel 1 is te zien dat de steekproef grotendeels uit vrouwen bestond (93.1%) en de steekproef bevatte ongeveer evenveel leerkrachten als studenten. Geen van de participanten komt uit regio Zuid. Bij de visualisatie *Gradiënt plot* is meer missing data dan bij de andere visualisaties. Ook is gekeken of de categorische variabelen, geslacht en leerkracht/student, evenredig zijn verdeeld tussen de vier verschillende visualisaties. De categorische variabelen, geslacht en leerkracht/student, zijn ongeveer evenredig verdeeld tussen de vier verschillende visualisaties. Studenten hebben minder vaak (8.3%) de visualisatie *Violin plot* toegewezen gekregen ten opzicht van bijvoorbeeld de *Error-bar* (16.7%), dit verschil was niet significant $\chi^2(3) = 1.4$, $p = .07$.

De continue variabele leeftijd laat geen uitbijters en missende waarden zien. De leeftijden van de participanten varieert tussen de 20 en 61 jaar, waarbij de meeste participanten een leeftijd hebben tussen de 20 en 40 jaar. De gemiddelde leeftijd van de participanten is 30.2 jaar met een standaarddeviatie van 12.9. Ook is gekeken of een verband ontdekt kan worden tussen de categorische variabele visualisaties van meetnauwkeurigheid en de continue variabele leeftijd. Er was geen significant verschil in de gemiddelde leeftijd van participanten per visualisatie $F(3,68) = .64$, $p = .59$.

**Figuur 6**

*Kenmerken van de variabele accuratesse van kansinschattingen (N=72).*

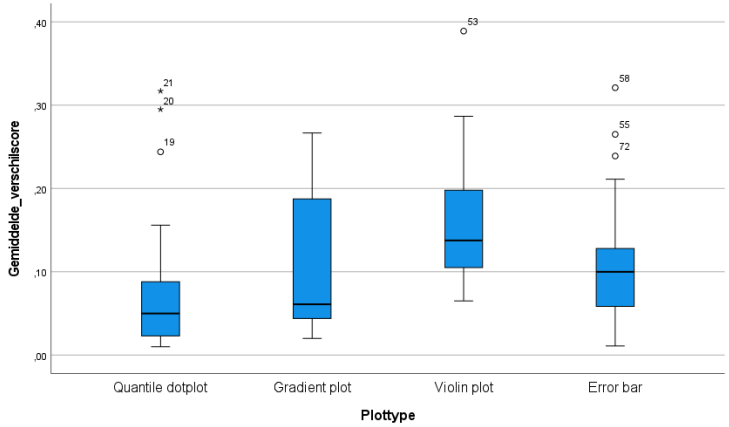



In Figuur 6 is te zien dat de waarnemingen van de variabele accuratesse van kansinschattingen scheef is verdeeld. De mediaan (.09) ligt namelijk dichterbij het minimum dan bij het maximum. De minimale score is .01 en de maximale score is .39. De berekende Guttman's Lambda 2 over de vragenlijst heeft een waarde van .85. De meeste waarden van de accuratesse van kansinschattingen liggen tussen de .80 en 1.40. Ook is in de boxplot een uitbijter aanwezig. Deze participant heeft een hoge verschilscore van .39. Verder valt op dat

deze participant leerkracht is en ouder is dan 50 jaar. Mogelijk heeft deze uitbijter ook een invloed op de verdeling van de waarnemingen.

**Figuur 7**

*Kenmerken van de accuratesse van kansinschatting gesplitst in de vier visualisaties (N=72).*




In Figuur 7 is te zien dat de medianen bij de *Quantile dot plot* (*Mdn* = .05, Range = .01-.32, SD = .09) en *Gradiënt plot* (*Mdn* = .06, Range = .02-.27, SD = .09) lager liggen dan bij de *Violin plot* (*Mdn* = .14, Range = .07-.39, SD = .09) en de *Error-bar* (*Mdn* = .10, Range = .01-.32, SD = .09). De standaarddeviaties tussen de vier verschillende visualisaties zijn nagenoeg gelijk aan elkaar. Hiermee wordt voldaan aan de assumptie van vergelijkbare varianties.

In de boxplots zijn uitbijters te zien bij de *Quantile dotplot, Violin plot en Error-bar*. Deze participanten hebben een score ver boven het gemiddelde en hierbij valt op dat deze participanten voornamelijk leerkrachten zijn. De leeftijden van de participanten variëren tussen de 20 jaar en 61 jaar. Voor de *Quantile dotplot, Violin plot en Error-bar* kan sprake zijn van een normale verdeling. De *Gradiënt plot* lijkt niet helemaal normaal verdeeld. Bij deze visualisatie ligt namelijk de mediaan veel dichterbij het minimum dan bij het maximum. Zoals hierboven beschreven zijn in de boxplots uitbijters te zien. Deze uitbijters hebben mogelijk invloed op de verdeling van de waarnemingen.

**Tabel 2**

*Uitkomsten lineaire regressieanalyse aan de hand van de vier visualisaties (N=72.)*

| Variabelen | Regressiecoëfficiënt | *t*-score | *p-score* | 95%-betrouwbaarheidsinterval |
|---|---|---|---|---|
| Error-bar | 1.11 (intercept) | 14.51 | <.001 | [.96;1.27] |
| Quantile dot plot | -.21 | -2.02 | <.05 | [-,42;-.003] |
| Gradiënt plot | -.04 | -.33 | .74 | [-.26;-.18] |
| Violin plot | .14 | 1.21 | .23 | [-.09;.38] |

In Tabel 2 worden de log-getransformeerde scores gebruikt. Ten eerste lieten alle visualisaties een significante afwijking zien van een perfecte kansinschatting $t$= 9.42 , $p$= <.001. Het intercept van de visualisatie *Quantile dotplot is* .90 (1.11 - .21). Dit houdt in dat participanten voor de *Quantile dot plot* accurater zijn in vergelijking tot de *Error-bar*. Het intercept van de visualisatie *Gradiënt plot* is 1.07 (1.11 - .04). Dit houdt in dat participanten voor de *Gradiënt plot* accurater zijn in vergelijking tot de *Error-bar.* Het intercept van de visualisatie *Violin plot* is 1.25 (1.11 + .14). Dit houdt in dat participanten voor de *Violin plot* minder accuraat zijn in vergelijking tot de *Error-bar*.

De *Violin plot* laat geen significant verschil zien $t$ = 1.21, $p$ = .23. De *Quantile dotplot* laat een significant verschil zien $t$ = -2.02, $p$ < .05. De visualisatie *Gradiënt plot* laat geen significant verschil zien $t$ = -.33, $p$ = .74. Wanneer de uitbijters, omschreven bij Figuur 7, uit de dataset worden gehaald, ontstaan andere verschillen. De *Quantile dotplot* ($t$ = -2.40, $p$ = .02) laten nog steeds een significant verschil zien. De *Gradiënt plot* ($t$ = .72, $p$ = .48) laat nog steeds geen significant verschil zien. Daarentegen laat nu de *Violin plot* ($t$ = 2.001, $p$ = 0.05) een significant verschil zien.

Ook is gekeken wat voor soort fouten leerkrachten maken in hun kansinschattingen bij de verschillende visualisaties. Bij de *Error-bar* wordt twintig keer door de participanten een kansinschatting gegeven van 0%. Namelijk vier keer bij vaardigheidsscore 64, acht keer bij vaardigheidsscore 65, drie keer bij vaardigheidsscore 73 en drie keer bij vaardigheidsscore 75. Dit is meer dan bij de andere visualisaties. Bij de *Quantile dotplot* en *Violin plot* wordt eenmaal een kansinschatting gegeven van 0%. Bij de *Gradiënt plot* wordt geen enkele keer een kansinschatting gegeven van 0%.

Ook is gekeken naar een mogelijke foutenanalyse bij de *Quantile dotplot.* Bij de vier vragen waarbij de bolletjes precies op de horizontale lijn liggen van de verschillende niveaus is het gemiddelde accuratesse van kansinschattingen 1.05 (SD = .41). Bij de overige zes vragen is het gemiddelde accuratesse van kansinschattingen 1.10 (SD = .43). Participanten maken accuratere kansinschatting bij de vier vragen waarbij de bolletjes precies op de horizontale lijn liggen van de verschillende niveaus in vergelijking tot de overige zes vragen. Dit verschil kan als significant worden beschouwd, $t$= 2,86, $p$ = .01.

Bij het analyseren van de genoemde vier vragen valt op dat waarschijnlijk veel participanten de bolletjes hebben geteld. Bij de vaardigheidsscore van 64 kunnen 20 bolletjes

geteld worden. Gekeken naar de vraag, ‘’In hoeverre is het werkelijke niveau III?’’, ligt er één bolletje in vaardigheidsniveau III. Veel participanten hebben dan ook een kansinschatting gegeven van 5%, terwijl de werkelijke kansinschatting 8% is.

## Discussie

In dit onderzoek is gekeken naar de onderzoeksvraag: ‘In hoeverre hebben de verschillende visualisaties invloed op de kansinschattingen die leerkrachten maken bij het interpreteren van citoscores?’. Bij alle visualisaties weken participanten significant af van een perfecte kansinschatting. Tussen de *Error-bar* en de *Quantile dotplot* is een significant negatief verschil gevonden. Dit houdt in dat participanten bij de *Quantile dotplot* accurater zijn in vergelijking tot de *Error-bar.* De range komt zowel bij de *Quantile dotplot* als bij de *Error-bar* uit op .31. Hierdoor kan het verschil in intercepts (-.21) tussen deze twee visualisaties als redelijk groot gezien worden. Tussen de *Error-bar* en de *Violin plot* is ook een significant positief verschil gevonden wanneer de uitbijters uit de data werden geëxcludeerd. Dit houdt in dat participanten bij de *Violin plot* meer afwijken van een perfecte kansinschatting in vergelijking tot de *Error-bar.* Dit verschil kan als een klein verschil gezien worden. Daarentegen liet de *Gradiënt plot* ten opzichte van de *Error-bar* geen significant verschil zien. Op basis van deze gegevens kan aangetoond worden dat de *Quantile dotplot* in vergelijking tot de *Error-bar* leerkrachten mogelijk kan helpen bij het maken van accuratere kansinschattingen.

Van tevoren werd verwacht dat participanten mogelijk een kansinschatting van 0% geven bij vaardigheidsscores die net buiten het betrouwbaarheidsinterval vallen. De resultaten lieten zien dat bij de *Error-bar* participanten 20 keer deze verwachte fout maakten. Op basis van deze gegevens kan aangetoond worden dat bij de *Error-bar* participanten vaker fouten maken bij het interpreteren van vaardigheidsscores die net buiten het betrouwbaarheidsinterval vallen.

Ook is gekeken naar een foutenanalyse voor de *Quantile dotplot.* Bij de vier vragen waarbij de bolletjes precies op de horizontale lijn liggen van de verschillende niveaus, ligt het gemiddelde accuratesse van kansinschattingen significant lager dan bij de overige zes vragen waarbij de bolletjes niet precies op de horizontale lijn liggen van de verschillende niveaus. Wanneer de bolletjes precies op de horizontale lijn liggen, kunnen participanten precies tellen hoeveel bolletjes in een bepaald vaardigheidsniveau liggen. Dit kan verklaren waardoor bij deze vragen het gemiddelde accuratesse van kansinschattingen lager ligt dan bij de overige zes vragen.

Op basis van het onderzoek van Hopster-den Otter et al., (2018) werd verwacht dat de visualisatie *Error-bar* een positieve invloed had op de beslissingen en voorkeuren van leerkrachten. Echter liet dit onderzoek zien dat de *Error-bar* niet zorgt voor accurate kansinschattingen. Ook werd op basis van Hopster-den Otter et al., (2018) verwacht dat de *Gradiënt plot* een geschikte presentatie kan zijn van een betrouwbaarheidsinterval bij testresultaten. Echter konden participanten in dit onderzoek voor de visualisatie *Gradiënt plot* een accuratere kansinschatting maken ten opzichte van de visualisatie *Error-bar,* maar dit verschil werd niet als significant beschouwd. Op basis van het onderzoek van Kay et al., (2016) werd verwacht dat de *Quantile dotplot* zorgde voor meer betrouwbare kansinschattingen. Echter week de *Quantile dotplot* in dit onderzoek significant af van een perfecte kansinschatting. Daarentegen waren participanten bij de *Quantile dotplot* accurater in vergelijking tot de *Error-bar.* Ook kan het tellen van de bolletjes participanten helpen een accuratere kansinschatting te maken. Dit gegeven kan ook verklaren dat de *Quantile dotplot* in vergelijking tot de *Error-bar* leerkrachten mogelijk kan helpen bij het maken van accuratere kansinschattingen. De *Quantile dotplot* en de *Violin plot* zijn beiden visualisaties die een duidelijk verdeling van meetfouten weergeven wat leidt tot een betere kwaliteit van beslissingen in vergelijking tot de *Error-bar* (Padilla et al., 2020). Ondanks dit gegeven toonde dit onderzoek aan dat participanten bij de *Violin plot* minder accuraat waren in vergelijking tot de *Error-bar.*

Een aantal kanttekeningen en sterke punten kunnen geplaatst worden bij de interne validiteit van dit onderzoek. Ten eerste is de variabele ‘geslacht’ niet evenredig verdeeld. Dit kan mogelijk verklaard worden doordat het aandeel vrouwen (87%) dat lesgeeft in het basisonderwijs veruit het grootste aandeel van alle beroepsgroepen in het onderwijs is (Beiro & Ramaekers, 2016). Dit toont aan dat de steekproef uit dit onderzoek een representatieve weergave is van de populatie. De verdeling van de andere categorische variabelen tussen de vier visualisaties zijn ongeveer evenredig verdeeld.

Ook wordt in dit onderzoek voldaan aan de gestelde assumpties. Ten eerste zijn de varianties tussen de vier groepen, *Error-bar, Violin plot, Quantile dotplot en Gradiënt plot,* vergelijkbaar aan elkaar. Ten tweede zijn de waarnemingen in de vier groepen, *Error-bar, Violin plot, Quantile dotplot en Gradiënt plot,* ongeveer normaal verdeeld. Hier moet wel een kanttekening bij geplaatst worden. De waarnemingen bij de *Gradiënt plot* lijken niet helemaal normaal verdeeld. Bij deze visualisatie ligt namelijk de mediaan veel dichterbij het minimum dan bij het maximum. Na de transformatie van de verschilscore is de mediaan dichterbij het midden komen te liggen, maar de mediaan ligt nog steeds dichterbij het

minimum dan bij het maximum. Hierdoor moet voorzichtig omgegaan worden bij interpretatie van de visualisatie *Gradiënt plot.* Aan de laatste assumptie, de waarnemingen moeten afhankelijk van elkaar zijn, kon niet worden getoetst. De meeste participanten zullen deze vragenlijst wel onafhankelijk van elkaar hebben ingevuld, omdat de vragenlijst online naar de participanten is verstuurd. Hierdoor zullen de meeste participanten elkaar niet kennen. Daarentegen zijn ook PABO opleidingen benaderd om binnen de opleiding de vragenlijst te verspreiden. Hierdoor is het mogelijk dat studenten van deze PABO opleidingen elkaar kennen en gezamenlijk de vragenlijst hebben ingevuld waardoor mogelijk niet aan deze assumptie wordt voldaan.

Een sterk punt aan dit onderzoek is de keuze voor een experimenteel between-subject design waardoor participanten door middel van willekeurig toewijzing één van de vier visualisaties te zien kregen. Hiermee had iedere participant evenveel kans op één van de vier visualisaties en kon het toewijzen aan een bepaalde visualisatie niet beïnvloed worden. Het toewijzen aan een bepaalde visualisatie kan mogelijk wel beïnvloed worden door systematische uitval. In totaal konden de 72 participanten gezamenlijk 720 vragen invullen. Van deze 720 vragen zijn 84 vragen niet ingevuld. Onder de groep studenten kwam de meeste uitval voor. Van de participanten waarvan niet alle vragen zijn ingevuld, is wel de gemiddelde accuratesse van kansinschatting berekend. Dit kan mogelijk een bias in de resultaten veroorzaken. Het gemiddelde accuratesse van kansinschattingen van deze participanten lag op .15. Dit gemiddelde ligt niet veel hoger dan het gemiddelde accuratesse van kansinschattingen van alle participanten tezamen waardoor in dit onderzoek het uitvallen van participanten niet voor een grote vertroebeling in de resultaten heeft gezorgd. Verder kwam in de data geen vorm van systematische uitval voor.

Een ander sterk punt aan dit onderzoek is de betrouwbaarheid van de vragenlijst. Om de betrouwbaarheid van deze vragenlijst aan te tonen is de Guttman's Lambda 2 gebruikt. De Guttman's Lambda 2 is berekend op .85. Deze Guttman's Lambda 2 is hoger dan de minimale referentiewaarde van .70.

Ook kan een kanttekening geplaatst worden bij de externe validiteit van dit onderzoek. Allereerst is gebruik gemaakt van een gelegenheidssteekproef. Hierdoor is het moeilijk te stellen of in de genomen steekproef sprake kan zijn van een representatieve doorsnee van de samenleving in Nederland. Door de gelegenheidssteeproef komen de meeste participanten uit regio 'Noord' (45.8%). De rest van de participanten komt uit regio 'Oost' (26.4%) en regio 'West' (27.8%). De steekproef bevat geen participanten uit regio 'Zuid'. Hierdoor kunnen de resultaten mogelijk niet generaliseerbaar zijn naar de doelpopulatie.

Daarentegen wordt niet verwacht dat leerkrachten uit het Zuiden accuratere of minder accuratere kansinschattingen maken. Alle Nederlandse leerkrachten en derdejaars (academische) PABO studenten hebben dezelfde opleiding genoten. Ook wordt op de opleidingen gewerkt aan vergelijkbare leerdoelen. De competenties tussen Nederlandse leerkrachten en derdejaars (academische) PABO studenten zullen waarschijnlijk niet veel verschillen waardoor de resultaten uit dit onderzoek mogelijk wel generaliseerbaar zijn naar de doelpopulatie. Aan de andere kant bieden verschillende hogescholen op hun eigen manier onderwijs aan waardoor onderwijs en dus ook het aanbod aan statistische vakken tussen de hogescholen kan verschillen.

Het fysiek afnemen van de vragenlijst door de onderzoeker kan een aanbeveling zijn voor vervolgonderzoek. Door de vragenlijst fysiek af te nemen kan mogelijk verminderd worden dat participanten niet alle vragen uit de vragenlijst invullen. Ook kan in een fysiek onderzoek de behaalde vaardigheidsscores op een leergebied uit het LOVS van de leerlingen meegenomen worden. Hierdoor kunnen leerkrachten mogelijk de antwoorden op de vragen als essentiëler beschouwen waardoor leerkrachten meer moeite nemen om een accurate kansinschatting te maken dan bij een online vragenlijst. Hier kan ook een kanttekening bij geplaatst worden. Door het meenemen van vaardigheidsscores van leerlingen, kunnen leerkrachten andere informatie uit de klas meenemen. Door deze informatie kan een vertekend beeld ontstaan van de gemaakte kansinschatting. Tevens wordt door het fysiek afnemen van de vragenlijst door een onderzoeker voldaan aan de assumptie van afhankelijkheid.

Ten tweede kan in vervolgonderzoek een vragenlijst gemaakt worden waarin de participanten alle vier de visualisatie te zien krijgen. Hierdoor kan het probleem van variatie van participanteigenschappen tussen de vier visualisatie voorkomen worden. Hierdoor kan met meer zekerheid gesteld worden dat de manipulatie van de onafhankelijke variabele visualisaties van meetnauwkeurigheid het gevonden verschil tussen de verschillende visualisaties heeft veroorzaakt. Ook bied deze aanbeveling als voordeel dat voor vervolgonderzoek minder participanten geworven kunnen worden. Daarentegen zijn de metingen in het onderzoek niet meer onafhankelijk, waardoor meer geavanceerde statistische technieken nodig zijn voor de analyse.

Met dit onderzoek en mogelijk vervolgonderzoek wordt getracht leerkrachten te helpen bij het maken van accuratere kansinschattingen, door te kijken naar welke visualisaties een positieve invloed hebben op de kansinschattingen die leerkrachten maken. Hiermee kunnen verkeerde beslissingen binnen het onderwijs mogelijk verminderd worden.

Dit is van belang, omdat beslissingen in het Nederlandse regulier en speciaal basisonderwijs consequenties hebben voor leren en lesgeven. Door het nemen van juiste beslissingen binnen het regulier en speciaal basisonderwijs kan tegemoet gekomen worden aan de onderwijsbehoeften van leerlingen.

## Literatuurlijst


Beiro, L.F., Ramaekers, M. (2016). *Steeds meer vrouwen voor de klas in het basisonderwijs.* CBS.

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

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

Fullan, M., & Watson, N. (2000). School-based management: Reconceptualizing to improve learning outcomes. School Effectiveness and School Improvement: An In-ternational Journal of Research, Policy and Practice, 11, 453–473. http://dx.doi.org10.1076/sesi.11.4.453.3561.

Gardner, J. (2013). The public understanding of error in educational assessment. Oxford Review of Education, 39(1), 72–92.

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

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.

Hullman, J., Resnick, P., & Adar, E. (2015). Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. Plos One, 10(11), 0142444.

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

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016). When (ish) is my bus? User-centered visualizations of uncertainity in everyday, mobile predictive systems. CHI Conference on Human Factors in Computing Systems.

Kinkeldey, C., MacEachren, A. M., Riveiro, M., & Schiewe, J. (2015). Evaluating the effect of visually represented geodata uncertainty on decision-making: Systematic review,

lessons learned, and recommendations. Cartography and Geographic Information Science, 44, 1–21.

Ledoux, G., Blok, H., Boogaard, M., & Krüger, M. (2009). Data-driven decision making; About the value of measurement oriented education. SCO-Kohn-stamm Instituut.

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., & Hetzler, E. (2005). Visualizing geospatial information uncertainty: What we know and what we need to know. Cartography and Geographic Information Science, 32, 139–160.

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

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

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

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

Sijtsma, K., & Drenth, P.J.D. (2006). *Testtheorie. inleiding in de theorie van de psychologische test en zijn toepassingen.* Bohn Stafleu van Loghum.

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.

Vanhoof, J., Verhaeghe, G., Verhaeghe, J. P., Valcke, M., & Van Petegem, P. (2011). The influence of competences and support on school performance feedback use. Educational Studies.

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