Rapport
2018

De eTIMSS equivalentiestudie

VAN PAPIEREN TOETS NAAR TABLETTOETS, ZIJN ER VERSCHILLEN?

EVA HAMHUIS
CEES GLAS
MARTINA MEELISSEN

UNIVERSITY OF TWENTE.

Samenvatting

In TIMSS (*Trends in International Mathematics and Science Study*) wordt sinds 1995 om de vier jaar wereldwijd het onderwijsniveau in de exacte vakken gemeten aan de hand van (papieren) toets- en vragenlijstafnames bij leerlingen in het primair en secundair onderwijs. In TIMSS-2019 worden de eerste stappen gezet van een papieren dataverzameling naar een uiteindelijk (volledig) digitale afname van toets en vragenlijsten. In het voorjaar van 2017 heeft ongeveer de helft van de deelnemende TIMSS-landen, waaronder Nederland, aangegeven de overstap te willen maken naar een digitale toets- en vragenlijstafname. Gezien de doelstellingen van TIMSS is het van belang dat deze overstap geen invloed heeft op de vergelijking van het onderwijsniveau tussen landen en tussen de afnamejaren. Om deze reden hebben de eTIMSS-landen in het voorjaar van 2017 deelgenomen aan een equivalentiestudie. Hierbij werd de TIMSS-toets bij een kleine groep scholen digitaal afgenomen. Dezelfde leerlingen maakten ook een papieren versie van de toets. In dit rapport wordt verslag gedaan van de resultaten van een vergelijking in de vaardigheden van Nederlandse groep 6 leerlingen tussen de papieren en digitale toetsafname.

Uit het onderzoek komt naar voren dat op basis van *item response analyse* geen verschillen zijn gevonden tussen de papieren toetsafname en de tablet toetsafname. De toetsen lijken overeen te komen op vaardigheidsverdelingen, de itemparameters en op de gemiddelden van de boekjes. Meisjes blijken echter iets beter scoren op de gehele toets en in het bijzonder op de tablettoets. Dit gevonden effect is klein maar noemenswaardig, omdat in voorgaande TIMSS-toetsen jongens beter scoorden. De algehele conclusie van deze studie is dat er geen noemenswaardige verschillen zijn gevonden tussen de toetsafnames. De digitale TIMSS-toets lijkt een geschikte toetsmethode om af te nemen binnen Nederland. Het is echter niet zeker of dit ook voor de andere eTIMSS landen geldt. Om trendvergelijkingen en vergelijkingen tussen landen te kunnen waarborgen, zullen daarom alle eTIMSS landen een *bridge study* (op een extra set scholen een volledig papieren afname) uitvoeren tijdens de hoofdafname van 2019.

Inhoudsopgave

1. Inleiding

In 1958 kwam een groep wetenschappers op gebied van psychologie, sociologie en psychometrie bij elkaar om te discussiëren over onderzoek naar de effectiviteit van onderwijs. Men kwam tot de conclusie dat voor een effectieve evaluatie van onderwijs zowel de input (kenmerken van onderwijssystemen, scholen, leraren en leerlingen) als de output (onderwijsopbrengsten zoals prestaties, onderwijsdeelname en attituden) gemeten zouden moeten worden. Door deze informatie in verschillende landen te verzamelen en tussen landen te vergelijken hoopte men nieuwe inzichten te verkrijgen in wat van belang is voor goed onderwijs. In 1960 werd een pilotstudie uitgevoerd, waarin 13-jarige leerlingen uit 12 landen een internationale toets in verschillende vakgebieden maakten.

Als vervolg op deze pilotstudie werd in 1967 *The International Assocation for the evaluation of Educational Achievement* (IEA) opgericht. Deze organisatie brengt verschillende onderzoeksinstituten en overheden bij elkaar om internationaal vergelijkend onderzoek, evaluatie en innovatie binnen het onderwijs te realiseren. Inmiddels heeft IEA meer dan 30 internationaal vergelijkende onderwijsstudies geïnitieerd waaraan meer dan 100 landen hebben deelgenomen. In deze studies staat een vergelijking tussen het beoogde (wat leerlingen zouden moeten leren), het geïmplementeerde (wat leerlingen leren) en het gerealiseerde curriculum (onderwijsopbrengsten) centraal.

1.1 TIMSS-2019

TIMSS (*Trends in International Mathematics and Science Study*) is de langstlopende IEA-studie. Sinds 1995 meet TIMSS wereldwijd om de vier jaar de leerprestaties van leerlingen in de exacte vakken in het primair en secundair onderwijs. De doelstellingen van TIMSS zijn (Mullis & Martin 2017):

- Vergelijking van landen in de onderwijsopbrengsten in de exacte vakken.
- Vergelijking van landen in de zwakke en sterke punten van hun onderwijssysteem.
- Trendvergelijking binnen landen van de onderwijsopbrengsten in de exacte vakken
- Landen de mogelijkheid bieden een *evidence-based* beleid te voeren ter verbetering van het onderwijs.

De TIMSS-toets is gebaseerd op een raamwerk, waarin de inhoudelijke en cognitieve domeinen en leerstofgebieden beschreven worden die in de toets gemeten moeten worden. Voor elke nieuwe TIMSS-cyclus wordt dit raamwerk geactualiseerd. Naast de toets wordt in TIMSS contextinformatie verzameld door middel van leerling-, leraar-, school- en curriculumvragenlijsten.

Het project wordt gecoördineerd door het *TIMSS & PIRLS International Study Center* van *Boston College.* De Nederlandse deelname aan TIMSS wordt sinds 1995 uitgevoerd door de Universiteit Twente. Nederland heeft sinds 1995 aan alle zes TIMSS-rondes meegedaan, maar vanaf TIMSS-2007 alleen met leerlingen in *grade 4* (groep 6 van het basisonderwijs).

1.2 eTIMSS-2019

Voor TIMSS-2019 is aan de deelnemende landen de mogelijkheid geboden de TIMSS-toets voor de eerste keer digitaal af te nemen (eTIMSS). In TIMSS-2011 en 2015 konden landen er al voor kiezen om de school- en leraarvragenlijst online in te laten vullen. Ongeveer de helft van de landen (waaronder Nederland) was hierin geïnteresseerd en had ook de mogelijkheid om de TIMSS-toets digitaal af te nemen. Omdat een van de belangrijkste doelen van TIMSS trendvergelijkingen zijn, is het van belang dat deze overstap geen invloed heeft op de mogelijkheid om het onderwijsniveau tussen jaren te

kunnen vergelijken. Bovendien moeten de uitkomsten van landen die niet aan deze optie mee kunnen of willen doen, nog steeds te vergelijken zijn met de uitkomsten van de eTIMSS-landen. De eTIMSS-landen namen daarom in 2017 deel aan een equivalentiestudie. Hierbij werd een deel van de toetsopgaven van TIMSS-2015 in rekenen-wiskunde en natuuronderwijs zowel afgenomen op een pc of tablet als op papier. Het belangrijkste doel van de equivalentiestudie is het bijdragen aan de waarborging van de TIMSS-doelen door het identificeren van de mogelijke effecten van de digitale toetsafname ten opzichte van de papieren toetsafname. Voor deze waarborging worden overigens niet alleen de uitkomsten van equivalentiestudie gebruikt. In landen die aan eTIMSS meedoen en daarmee de TIMSS-toets in het hoofdonderzoek in 2019 digitaal afnemen, moeten bij een aanvullende steekproef van ongeveer 50 scholen de papieren versie van de toets afnemen (*bridge study*).

Voorwaarde voor de keuze voor digitale assessment was dat het een duidelijke toegevoegde waarde moest hebben; de digitale toets moet efficiënter en betrouwbaarder de kennis en vaardigheden meten en beter aansluiten op de (individuele) leerbehoeften van leerlingen in de huidige maatschappij (Mullis & Martin, 2017). Tijdens de zogenoemde proefafname in het voorjaar van 2018, zijn daarom geheel nieuwe opgaven uitgetest onder een beperkte steekproef van scholen. Een deel van deze opgaven (de zogenoemde *Problem Solving Inquiry* (PSI) items zijn speciaal voor digitale afname ontworpen. De PSI-taken zijn zo vormgegeven dat ze er visueel aantrekkelijk uitzien met interactieve scenario's die leerlingen voorzien van adaptieve en responsieve manieren om een reeks stappen te volgen om tot een oplossing te komen. Ook zullen deze taken een gelegenheid vormen om het probleemoplossings- of onderzoeksproces van leerlingen te volgen. Het voordeel van het inpassen van PSI's in de toets is dat ze echte wereld en laboratorium situaties simuleren waarbij leerlingen in het oplossingsproces vaardigheden en kennis moeten integreren en toepassen. Het gaat dan om bijvoorbeeld kennis om rekenproblemen op te lossen en wetenschappelijke experimenten uit te voeren. Tegelijkertijd is het ook van belang dat de inhoud van de TIMSS-toets gewaarborgd blijft. Het doel is om de kennis en vaardigheden in rekenen-wiskunde en natuurwetenschappen te meten niet de digitale vaardigheden van leerlingen.

1.3 Onderzoeksvragen

In dit rapport wordt verslag gedaan van de uitkomsten van de equivalentiestudie in het voorjaar van 2017 onder groep 6 leerlingen in Nederland. In de equivalentiestudie van 2017 zijn geen nieuwe opgaven opgenomen, maar is een selectie gemaakt van de opgaven uit TIMSS-2015. De leerlingen zijn in deze pilot tweemaal getoetst, één keer op papier en één keer op een tablet (verschillende versies van de toets). Zoals beschreven ging het in de equivalentiestudie vooral om de waarborging van de TIMSS-doelen door de (prestatie-)effecten te meten van de digitale toets ten opzichte van de papieren toets. De equivalentiestudie is daarnaast gebruikt voor het opdoen van praktische ervaring om een soepele overstap voor landen naar de nieuwe toetsvorm te kunnen bewerkstelligen (Mullis & Martin, 2017). De informatie van de equivalentiestudie is gebruikt om de vormgeving van het toetsinstrument te evalueren en knelpunten in de afnameprocedures te inventariseren.

In dit rapport staan de volgende onderzoeksvragen centraal:

1. *In welke mate zijn er verschillen in vaardigheidsniveau en itemparameters (moeilijkheid en discriminatiegraad) tussen de leerlingen die de papieren TIMSS-toets van 2015 hebben gemaakt en de leerlingen die de papieren equivalentiestudie toets in 2017 hebben gemaakt?*

Voor de equivalentiestudie is een kleine groep scholen benaderd op basis van *convenience sampling.* Aan TIMSS-2015 heeft een representatieve steekproef van scholen deelgenomen (Meelissen & Punter, 2016). Als eerste stap is onderzocht in hoeverre het geschatte vaardigheidsniveau van de leerlingen per boekje in de pilot van 2017 afwijkt van die in TIMSS-2015. Daarnaast zijn de *Differential Item Functioning statistics* (DIF) bekeken. Hiermee is het mogelijk na te gaan in hoeverre de items in 2017 hetzelfde functioneren als in 2015.

2. *Wat zijn de overeenkomsten en verschillen tussen de papieren toets en de tablettoets in vaardigheidsniveau en itemparameters (moeilijkheids- en discriminatiegraad) in de equivalentiestudie?*

Bij de tweede onderzoeksvraag wordt er voor rekenen en natuuronderwijs gekeken of er een verschil zit tussen het geschatte vaardigheidsniveau van leerlingen op de papieren en tablettoets. Met behulp van Item Response Theorie (IRT) wordt ten eerste onderzocht of leerlingen een ander vaardigheidsniveau lijken te hebben en ten tweede of de item parameterschattingen (moeilijkheid van het item en discriminatiegraad) tussen de twee toetsvormen van elkaar verschillen. Vervolgens wordt aan de hand van een kwalitatieve verkenning nagegaan of items die relatief grote verschillen laten bepaalde kenmerken hebben, zoals het behoren tot een bepaald inhoudsdomein of opgavetype.

3. *In hoeverre is geslacht van invloed op de verschillen tussen de papieren toets en de tablettoets prestaties?*

In Nederland presteren jongens over het algemeen beter op de TIMSS-toets dan meisjes. De vraag voor deze vergelijkende studie is of jongens nog steeds beter zijn op zowel de papieren toets als de tablettoets in 2017.

1.4 Opbouw van dit rapport

In dit rapport wordt verslag gedaan van de Nederlandse resultaten van de equivalentiestudie in groep 6 van het basisonderwijs. In het volgende hoofdstuk (hoofdstuk 2) wordt de opzet van de toets en de afname van de toets beschreven. Hier wordt aandacht besteed aan ontwikkeling en opzet van de toets; de steekproef; scholenwerving; dataverzameling; de analysemethodes en kenmerken van de deelnemers.

In hoofdstuk 3 worden de resultaten beschreven. Als eerst wordt nagegaan of de steekproef van de pilot overeenkomt met de representatieve steekproef uit 2015 (onderzoeksvraag 1). Ten tweede wordt de papieren toets vergeleken met de tablettoets. Hiervoor zijn de hoofddomeinen rekenen en natuuronderwijs apart geanalyseerd (onderzoeksvraag 2). Hierna volgt een paragraaf over de mogelijke sekseverschillen in toetsprestaties (onderzoeksvraag 3).

In hoofdstuk 4 worden de conclusies aan de hand van de onderzoeksvragen besproken. Tot slot wordt op de onderzoeksmethode gereflecteerd en worden er aanbevelingen gedaan voor vervolgonderzoek in de discussiesectie.

2. De uitvoering van de equivalentiestudie

2.1 Inhoud en opzet van de toets

Het *TIMSS & PIRLS International Study Center* (ISC) heeft de deelnemende landen vanaf het begin van TIMSS-2019 cyclus betrokken bij de overstap naar digitale toetsing. Begin 2017 kregen de deelnemende landen de keuze om de toets tijdens de proefafname in 2018 en tijdens het hoofdonderzoek van 2019 op papier, op computers of op tablets af te nemen. Landen die kozen voor de digitale toets werden gevraagd of zij ook deel wilden nemen aan de equivalentiestudie in het voorjaar van 2017. Landen hadden de keuze om via usb-sticks de toets af te nemen op computers op de scholen, op externe laptops of op externe tablets. Om niet afhankelijk te hoeven zijn van de ICT-infrastructuur van een school, is in Nederland gekozen voor tablets die door toetsleiders zijn meegenomen naar de scholen.

De toets voor de equivalentiestudie bestond uit een selectie van items uit TIMSS-2015. In het curriculumraamwerk van TIMSS-2015 is een onderverdeling gemaakt naar een inhoudelijke dimensie en een cognitieve dimensie. De cognitieve dimensie heeft betrekking op de handelingen of gedragingen die van leerlingen verwacht worden om een opgave te beantwoorden. De cognitieve domeinen in TIMSS zijn: Weten, Toepassen en Redeneren. Daarnaast zijn er de volgende inhoudelijke domeinen voor rekenen: Getallen, Gegevensweergave en Geometrische vormen en meten. Biologie, Natuur- en scheikunde en Fysische aardrijkskunde vormen de inhoudelijke domeinen voor natuuronderwijs.
De geselecteerde opgaven waren zo verdeeld dat elk domein in de toets gerepresenteerd werd. De mate van representatie was afhankelijk van het belang van het domein in het TIMSS-raamwerk. In Tabel staat de verdeling van items over de inhoudsdomeinen.

Tabel 1
Inhoudsdomeinen rekenen-wiskunde en natuuronderwijs, equivalentiestudie 2017

Rekenen-Wiskunde		**Natuuronderwijs**	
Inhoudsdomein	*Aantal items*	*Inhoudsdomein*	*Aantal items*
Getallen	51	Biologie	43
Gegevensweergave	12	Natuur- en scheikunde	34
Geometrische vormen & meten	27	Fysische Aardrijkskunde	19

In totaal waren er 190 opgaven. Na de afname zijn een aantal items niet meegenomen in de data verwerking en de analyses, omdat er lay-out foutjes of andersoortige foutjes in de items zaten. Uiteindelijk zijn er 186 items gebruikt voor de analyses, waarvan 90 over rekenen-wiskunde en 96 over natuuronderwijs (Tabel 1). De lay-out van de meeste meerkeuze of open vragen kon op de tablets hetzelfde blijven als op papier. Sommige opgaven moesten echter worden aangepast voor de digitale toets. Dit zijn bijvoorbeeld opgaven waarbij op papier lijntjes wordt getrokken van de vraag naar het bijpassende antwoord. Op de tablet moesten de leerlingen de antwoorden slepen naar de juiste velden. De digitale toets werd afgesloten met een paar vragen over hoe de leerlingen de toets hadden ervaren en enkele vragen over hun ervaring met ICT.

2.2 Ontwikkeling van toets en administratieve programmatuur

De internationale coördinatie (IEA Hamburg) biedt landen voor eTIMSS nieuwe ondersteunende software. Ten eerste is er een onlinevertaalsysteem ontwikkeld waarmee de landen de Engelstalige trenditems kunnen vertalen naar de eigen taal en direct kunnen nagaan hoe de items er uit zouden

zien op tablet of pc. Dit maakt het gemakkelijker om de items te controleren op spelling, grammatica en lay-out. Overigens bleek de vormgeving van de toetsopgaven op het tabletscherm (die voor elk land hetzelfde moet zijn ondanks het verschil in taal) eenvoudiger dan de vormgeving van de papieren toetsboekjes. Het tweede nieuwe instrument is een webpage waarmee de dataverzameling gemonitord kan worden. Met dit instrument kan worden gecontroleerd of de toetsdata volledig vanaf de tablet of pc naar de internationale server is geüpload. Het derde instrument is een applicatie waarmee de testversie van toets op de tablets of pc uitgeprobeerd kan worden. Tot slot is er een nieuw online codeersysteem ontwikkeld waarmee de antwoorden op de open toetsopgaven online gescoord kunnen worden. Open opgaven waarop de leerling een kort antwoord moet geven, kunnen voortaan door de computer worden gescoord. Voor de papieren toets is dezelfde procedure gevolgd als in 2015. De scores werden eerst opgeschreven in de toetsboekjes, waarna ze werden ingevoerd in het *data-entry* systeem van TIMSS. Door de online scoring en computer coding is de scoringsprocedure aanmerkelijk efficiënter geworden ten opzichte van de papieren toetsafname.

2.3 Steekproeftrekking

De beoogde populatie voor (e)TIMSS betreft leerlingen die vier jaar formeel onderwijs hebben genoten. Voor Nederland zijn dit leerlingen uit groep 6. Voor de pilot moest elk land zelf ongeveer 25 scholen benaderen en er mocht gebruik worden gemaakt van *convenience sampling*. In Nederland is er in eerste instantie voor gekozen om de scholen op dezelfde wijze te trekken als de steekproef van TIMSS-2015. Op basis van de DUO-bestanden van 2016-2017 is een willekeurige steekproef getrokken gestratificeerd naar gemiddeld leerlinggewicht van de school en *proportional to size*. Voor elke originele school werden gelijktijdig twee reserveschoolen getrokken.

2.4 Scholenwerving

Om de bekendheid van het onderzoek te vergroten is er een oproep geplaatst op de nationale TIMSS website en in (digitale) vakbladen. De geselecteerde scholen voor de equivalentiestudie zijn in maart 2017 schriftelijk uitgenodigd voor het onderzoek. Omdat de wervingsperiode erg kort voor de afnameperiode viel, zijn de scholen een week daarna telefonisch benaderd. Wanneer een geselecteerde school niet deelnam is direct een reserveschool benaderd.

Bij deelname aan het onderzoek werd de leerkracht van groep 6 doorgaans benoemd tot contactpersoon. De contactpersoon was verantwoordelijk voor het aanleveren van de leerlinglijst en contactgegevens en gaf voorkeursdata door voor het afnemen van de toets. Daarnaast was de contactpersoon op de afnamedag het aanspreekpunt op de basisschool. De papieren en tablettoetsen werden afgenomen door de hiervoor getrainde toetsleiders van de Universiteit Twente.

Aan de scholen is een tegenprestatie geboden. De leerlingen kregen na het maken van de tweede toets een presentje. Daarnaast ontvingen de scholen een overzicht van de door de leerling behaalde toetsresultaten. Deze bestond per vakgebied (rekenen en natuuronderwijs) uit drie grafische weergaven: 1) de individuele scores behaald op de papieren toets in 2017 vergeleken met de landelijk gemiddelde van het hoofdonderzoek in 2015. 2) De individuele scores behaald op de papieren toets in 2017 vergeleken met het landelijke gemiddelde op de 2017 papieren toets. 3) De individuele scores behaald op de tablettoets in 2017 vergeleken met de landelijk gemiddelde op de tablettoets van 2017. Het was uiteraard niet mogelijk om op schoolniveau een grafische weergave te maken waarin de papieren toets vergeleken werd met de tablettoets, omdat de leerlingen op papier een andere versie maakten dan op de tablet.

De deelnamebereidheid van de scholen en leerkrachten uit de steekproef was laag. De meest voorkomende reden om af te zien van het onderzoek was dat het te kort dag was en de scholen het organisatorisch niet meer konden inplannen. Daarnaast was een veelvoorkomende reden dat scholen maar aan een beperkt aantal onderzoeken wilden deelnemen, die zij doorgaans al hadden ingepland. Tot slot waren er scholen die aangaven weinig affiniteit te hebben met het onderwerp van de studie. Vanwege de lage medewerkingsbereidheid is vervolgens besloten ook scholen uit het eigen netwerk te benaderen. Hierdoor hebben in totaal 23 scholen meegedaan aan het onderzoek.

De TIMSS-toets bestond uit acht boekjes (versies van de toets) Op basis van de leerling lijsten van de groepen 6 werd met de TIMSS-software de boekjes/modules vooraf aan de leerlingen toegewezen. De toewijzing van welke soort toets (papier of tablet) als eerste moest worden afgenomen, werd eveneens willekeurig door de software bepaald. Op ongeveer de helft van de scholen was de tabletafname als eerste, op de andere helft maakte de leerlingen eerst de papieren toets.

2.5 Dataverzameling

In de periode van mei tot juni 2017 zijn de 23 deelnemende scholen tweemaal bezocht door één of meerdere toetsleiders van de Universiteit Twente. Om de afnamecondities op scholen overeen te laten komen, werden de toetsleiders getraind in het afnemen van de toetsen. Bij de tabletafname moesten de toetsleiders de tablets instellen voor elke leerling en deze samen met kladpapier uitdelen. De toetsleiders vulden het leerling-ID en het wachtwoord in zodat zij zeker wisten dat elke leerling de juiste (vooraf willekeurig toegewezen) toets maakte. De toetsleiders begonnen op de afnamedag met een korte introductie van TIMSS en wat de leerlingen konden verwachten. Daarna gaven zij voor de papieren of tablettoets een klassikale instructie. Dit gaf de leerlingen de gelegenheid om vragen te stellen over hoe de toets werkt. Na de instructie begonnen de leerlingen met het eerste deel. Voor deel 1 hadden de leerlingen 36 minuten de tijd gevolgd door korte pauze van ongeveer 15 minuten. Deel 2 was eveneens 36 minuten. Na het maken van de tablettoets volgende nog een aantal vragen over de toets en het gebruik van tablets en andere digitale middelen op school en thuis. De toetsen verliepen gestructureerd en volgens planning. Af en toe waren er enkele vragen over de inhoud van de toets. Tijdens de toetsafnames bleken de leerlingen goed overweg te kunnen met de tablet en het toetsprogramma van TIMSS.

Voor het vergroten van de betrouwbaarheid van de dataverzameling werden de twee toetsdagen zo gelijk mogelijk gehouden. Zo waren de toetsleider(s), starttijd en pauzeduur zoveel mogelijk hetzelfde. De toetsleider vulde tijdens elke afname het presentieformulier en een debriefing formulier in. Op het debriefing formulier kon het verloop van de toets worden aangeven, zoals (software) problemen of leerlingen die moeite hadden met het gebruik van de tablet.

Na de toetsperiode zijn de open opgaven van de papieren en tablettoets nagekeken aan de hand van het internationaal vastgestelde coderingssysteem van TIMSS-2015. Hiervoor kregen de codeurs een training. Zoals eerder is aangeven konden de tablettoetsopgaven direct online worden nagekeken en is een deel van deze opgaven door de computer gecodeerd. De scores op papieren boekjes zijn ingevoerd in het data-entry systeem van TIMSS.

2.6 Analysemethoden

In dit onderzoek wordt onderzocht of er verschillen zijn tussen de papieren toets en de tablettoets in Nederland. Voor de statistische analyses is gebruik gemaakt van de Item Response Theorie (IRT). Dit is een verzameling van modellen die het antwoordgedrag van personen op meerdere items uit psychologische en onderwijskundige meetinstrumenten beschrijven en (eventueel) verklaren. De

vaardigheden van leerlingen kunnen met IRT geschat worden. Deze zogenoemde latente vaardigheden zijn niet direct observeerbare eenheden die met observeerbare variabelen, zoals testscores gemeten worden (Embretson & Reise, 2000). De IRT-analyses zijn uitgevoerd met het public domain softwarepakket MIRT (Glas, 2010).

Voor de beantwoording van de eerste onderzoeksvraag (vergelijking tussen papieren toets 2015 en papieren toets 2017) zijn de overeenkomstige items van de twee verschillende datasets samengevoegd. Op het eerste gezicht lijkt dit een incompleet data design, omdat verschillende groepen leerlingen verschillende versies van de toets hebben gemaakt. De items kunnen echter aan elkaar gelinkt worden, omdat iedere versie een overlappend blok heeft. Bijvoorbeeld versie 1 heeft blok 1 en 2, versie 2 heeft blok 2 en 3. Dit wordt ook wel een gelinkt data design met overeenkomstige items genoemd (Glas & Geerlings, 2015).

In de eerste analyses is een relatief simpel IRT-model, het *Rasch model* gebruikt. Een deel van de items is als goed of fout gescoord (dichotome scoring, c.q. 0 of 1) en een deel van de items had meerdere antwoord categorieën. Deze items zijn gescoord polytoom gescoord, c.q. als 0, 1, en 2. Voor deze items is een versie van het Rasch model gebruikt voor polytoom gescoorde items. Deze versie is bekend als het partial credit model. In het Rasch model wordt de vaardigheid van elke leerling beschreven met één vaardigheidsparameter (meestal aangeduid met het symbool θ). Elke leerling krijgt na het maken van een test een score die hem of haar ergens op de vaardigheidsschaal, ofwel de θ-schaal plaatst. De kans op een goed antwoord op een dichotome opgave is bijna 0 voor leerlingen met een zeer lage vaardigheid. Elk dichotoom item heeft één moeilijkheidsparameter (aangeduid met het symbool β) (Rasch, 1960). Dit betekent hoe hoger de moeilijkheid van een item, hoe lager de kans op een correct antwoord. De kans op een correct antwoord is daarbij afhankelijk van de latente vaardigheid van de leerlingen. Een leerling die goed is in rekenen heeft meer kans om een rekenopgave correct te beantwoorden dan een leerling die minder goed is in rekenen (Furr & Bacharach, 2008). De polytoom gescoorde items, dat wil zeggen de items die als 0, 1 of 2 gescoord worden, hebben twee itemparameters. Ook hier geldt: hoe hoger de vaardigheid, hoe hoger de verwachte score van een leerling. De latente vaardigheid van een leerling kan met het Rasch model worden geschat door de leerling een test te laten maken met meerdere items van een verschillend moeilijkheidsniveau.

Vervolgens is er opnieuw naar de data gekeken met een meer informatief model, het gegeneraliseerde partial creditmodel (het GPCM-model). Het GPCM-model heeft een extra item parameter; de discriminatieparameter (aangeduid met het symbool α). Met deze parameter wordt de mate van afhankelijkheid tussen de latente vaardigheid en het responspatroon op het item weergegeven. Dit betekent dat de discriminatieparameter beschrijft hoe goed een item onderscheid kan maken tussen leerlingen met een hoog of laag prestatieniveau. De discriminerende waarde van een item geeft de relevantie weer van het kenmerk dat door de test wordt gemeten (Birnbaum, 1968). Met de moeilijkheidsparameter en de discriminatieparameter op testscores kan nauwkeurig geschat worden wat de latente vaardigheden van de leerlingen zijn. Verder is er ook gebruik gemaakt van een complexer model, het 2-dimensionale GPCM-model. Dit model neemt twee latente variabelen mee in plaats van één, die het responsgedrag van de leerlingen kunnen verklaren. Daarnaast zijn er in alle hierboven genoemde modellen vaardigheidsverdelingen opgenomen. Deze zogenaamde marginale vaardigheidsverdelingen modelleren vaardigheidsverschillen tussen groepen leerlingen. Zo zijn er acht marginalen meegenomen zodat de verschillende boekjes met elkaar vergeleken kunnen worden. In de resultatensectie wordt het model met de beste modelfit gepresenteerd. Dit model is gebruikt voor verdere analyses.

Voor dit onderzoek zijn de item parameterschattingen (eigenschappen van de toetsitems) geanalyseerd voor TIMSS-2015 en eTIMSS-2017 op de papieren toets en de tablettoets. Er is nagegaan of de parameterschattingen van de items significant verschillende waarden heeft voor beide toetsmethodes. Hierbij is er gekeken naar de vaardigheidsverdelingen van beide toetsen (θ) en de itemparameters moeilijkheid van een item (β) en discriminerende waarde van een item (α).

2.7 Kenmerken van de deelnemende scholen en leerlingen

De scholen

Van de 60 scholen uit de steekproef hebben in totaal 14 scholen (23%) deelgenomen aan de beide toetsafnames. Dit is aangevuld met 9 scholen uit het eigen netwerk. Bijna alle deelnemende scholen hebben een laag gemiddeld leerlinggewicht en variëren van kleine naar middelgrote scholen. De tabellen B1, B2 en B3 in Bijlage 1, geven een aantal kenmerken weer van de deelnemende scholen.

De leerlingen

In totaal hebben 532 leerlingen zowel de papieren toets als de tablettoets gemaakt. Er deden ongeveer evenveel meisjes mee als jongens. In Tabel 2 staan een aantal leerling kenmerken weergegeven. In de TIMSS vragenlijst wordt aan de leerlingen het aantal boeken thuis gevraagd als een proxy-indicator van het opleidingsniveau van de ouders. Van de leerlingen die aan de pilot hebben deelgenomen geeft 40% aan dat er thuis maximaal één boekenplank met boeken is (25 of minder boeken). In het hoofdonderzoek van TIMSS-2015 was dit 37% (Meelissen & Punter, 2016). Ook in de andere kenmerken wijken de leerlingen in 2017 nauwelijks af van die van 2015. Het merendeel van de leerlingen heeft thuis een eigen tablet of computer ter beschikking. Daarnaast geven bijna alle leerlingen aan dat zij thuis internet hebben. De overgrote meerderheid van de leerlingen kunnen een computer of tablet op school gebruiken en alle scholen hebben volgens de leerlingen de beschikking over internet. De leerlingen in 2017 verschillen onderling behoorlijk in het aantal uren computer- of tabletgebruik. Deze verschillen kunnen te maken hebben met verschillen tussen de scholen in de mate waarin zij digitale leermiddelen gebruiken. Er is echter niet gevraagd naar waar (op school of buiten school) leerlingen computers of tablets gebruiken.

Tabel 2

Kenmerken van de getoetste groep 6 leerling in equivalentiestudie, in aantallen en percentages

	N	%
Aantal boeken thuis (n=524)		
0-10 boeken	56	11
11-25 boeken	150	29
26-100 boeken	195	37
101-200 boeken	74	14
Meer dan 200 boeken	49	9
Wat hebben jullie thuis? *Je eigen computer of tablet* (n=527)		
Ja	424	81
Nee	103	20
Wat hebben jullie thuis? *Internet* (n= 526)		
Ja	513	98
Nee	13	3
Is er op jouw school een computer of tablet die je kunt gebruiken? (n=532)		
Ja	509	97
Nee	14	3
Is er internet op jouw school? (n=532)		
Ja	519	99
Nee	4	1
Hoeveel tijd zit je achter de computer of op een tablet per dag? (n= 532)		
Minder dan 30 minuten	121	23
30 minuten tot 1 uur	185	36
1 tot 2 uur	117	22
2 of meer uren	99	19

3. Resultaten

3.1 Vergelijking hoofdonderzoek 2015 en equivalentiestudie 2017

Vanwege de *convenience sampling* voor de equivalentiestudie kunnen de uitkomsten niet gegeneraliseerd worden naar groep 6 leerlingen in Nederland. Wel is onderzocht in hoeverre de vaardigheidsscores en itemkenmerken van 2017 afwijken van het hoofdonderzoek uit 2015 waarvan de uitkomsten wel representatief zijn voor Nederland (Meelissen & Punter 2016). In de vergelijking zijn alle items opgenomen die in 2015 en 2017 op papier afgenomen zijn.

De gemiddelde scores van de acht versies (boekjes) van de papieren toets uit 2017 is vergeleken met de gemiddelde score van de 'normgroep 2015' (ruim 4600 leerlingen). Uit de analyse komt naar voren dat op twee boekjes na, de populatie parameterschattingen, gemiddelde scores, van de boekjes uit 2017 niet significant verschillen van de populatie parameterschattingen van de steekproef van 2015 (Tabel 3). De effectgrootte van de twee boekjes met een significant verschil is klein. Dit betekent dat de gemiddelde vaardigheid van de leerlingen in 2017 vergelijkbaar is met die van 2015.

Tabel 3
Vergelijking tussen toetsafname hoofdonderzoek 2015 en pilot 2017, per boekje, rekenen en natuuronderwijs

	Gemiddelde	SD	SE (Gemiddelde)	Z-waarde	Effectgrootte*
Normgroep 2015	0,00	1,00	0,000		
Boekje 1	-0,22	0,84	0,11	-2,02*	-0,26
Boekje 2	-0,21	0,87	0,12	-1,78	
Boekje 3	-0,04	0,90	0,12	-0,31	
Boekje 4	-0,28	1,07	0,13	-2,08*	-0,26
Boekje 5	0,11	1,03	0,13	0,09	
Boekje 6	0,02	0,95	0,12	0,02	
Boekje 7	0,03	1,03	0,13	0,03	
Boekje 8	-0,05	1,31	0,17	-0,3	

Noot: *Scores die significant verschillen van de normgroep 2015 worden aangeduid met *(< 0,05). Voor deze verschillen is in de laatste kolom ook de effectgrootte van het verschil met 2015 weergegeven.

Naast de vaardigheidsscores zijn de zogenoemde DIF-statistieken van de items onderzocht. DIF staat voor *differential item functioning*. Verwacht wordt dat wanneer twee groepen op dezelfde vaardigheidsschalen een item beantwoorden, de proportie correcte antwoorden overeenkomen. DIF ontstaat als een item verschillende vaardigheden meet bij verschillend groepen zodat het item andere proporties correct laat zien. In dit geval zijn de groep leerlingen die in 2015 de toets maakten vergeleken met de groep leerlingen die in 2017 de toets maakten.

Op twee items na laten de items geen DIF zien. Voor de twee items die afwijken blijkt de codering van de open antwoorden in 2017 door TIMSS te zijn aangepast. Voor de overige items in de test kan geconcludeerd worden dat de items in de twee toetsjaren hetzelfde functioneren.

3.2 Ruwe scores voor rekenen

In Tabel 4 en 5 staan per boekje het gemiddelde en de standaarddeviatie van de ruwe scores op de papieren en digitale rekentoets. Zoals in hoofdstuk 2 is aangegeven hebben de leerlingen niet twee keer dezelfde boekjes (c.q. dezelfde opgaven in een papieren en digitale versie) gemaakt. Het rekengedeelte bestaat uit ongeveer 20 tot 25 opgaven per toetsboekje. De hoogst behaalde scores en de gemiddelden op een boekje verschillen tussen papier en tablet; in vier boekjes scoren leerlingen gemiddeld hoger op de tablet en in vier boekjes gemiddeld hoger op de papieren toets. Het grootste verschil laat boekje 7 zien, in het voordeel van de papieren afname. In bijlage 2 staat de scoreverdeling grafisch weergegeven.

Tabel 4
Scores op de papieren rekentoets

	Aantal leerlingen	Maximum aantal behaalde punten	Gemiddelde	Standaarddeviatie
Boekje 1	65	15	8,82	3,32
Boekje 2	67	17	9,10	3,64
Boekje 3	68	25	14,07	4,52
Boekje 4	71	26	12,38	5,40
Boekje 5	72	20	11,64	4,38
Boekje 6	66	20	11,23	4,03
Boekje 7	68	23	15,06	4,08
Boekje 8	66	25	13,15	5,30

Tabel 5
Scores op de digitale rekentoets

	Aantal leerlingen	Maximum aantal behaalde punten	Gemiddelde	Standaarddeviatie
Boekje 1	62	19	12,18	4,14
Boekje 2	63	24	12,08	4,75
Boekje 3	66	18	11,00	3,78
Boekje 4	70	20	10,04	4,82
Boekje 5	72	22	13,94	3,93
Boekje 6	66	22	12,76	3,97
Boekje 7	66	16	8,50	3,3
Boekje 8	62	17	8,52	4,12

3.3 Verschillen tussen papieren en digitale rekentoets

Zoals in het methodedeel is aangegeven, zijn er verschillende IRT-modellen geanalyseerd. In Tabel 6 staan de uitkomsten van de vergelijking tussen modellen op basis van de log-likelihood ($\Delta - 2ll$) voor rekenen. Het meest simpele model PCM (het Rasch model) wordt verworpen omdat het GPCM-model significant een betere fit van de data laat zien. Het tweedimensionaal GPCM-model (één dimensie voor de papieren versie en één dimensie voor de digitale versie) verschilt significant van het eendimensionaal GPCM-model en geeft daarom een nog betere fit van de data. Vervolgens zijn acht marginalen (één voor elk boekje) aan het tweedimensionaal GPCM-model toegevoegd. Dit model geeft de beste modelfit. Met dit model worden twee latente vaardigheden geschat op basis van de itemparameters bèta en alfa en zijn tevens de acht toetsboekjes meegenomen in de schatting.

Tabel 6
Vergelijking van de modelfit rekenen

	$\Delta\ df$	Δ-2ll	P
1-dim PCM vs. 1-dim GPCM	187	360	<0,001
1-dim GPCM vs. 2-dim GPCM	1	12	<0,001
2-dim GPCM vs. 2-dim GPCM (8 marginalen)	35	381	<0,001

Vervolgens is de correlatie berekend van de vaardigheidsverdelingen van de twee toetsen op basis van het tweedimensionale GPCM-model. De twee vaardigheidsverdelingen blijken sterk te correleren (r=0,91). In het tweedimensioneel GPCM-model blijft de ranking van de thèta's (vaardigheidsscore van de leerling) zo goed als hetzelfde voor beide toetsen. Het responsgedrag van de leerlingen op de rekentoets wordt het beste verklaard door twee latente vaardigheden, maar deze correleren erg hoog. Dit betekent dat de papieren en tablettoets bijna dezelfde vaardigheden lijken te meten.

Vervolgens is nagegaan in hoeverre de boekjes binnen de toetsmodus en tussen de toetsmodus (papier versus tablet) van elkaar verschillen. Alle boekjes zijn telkens vergeleken met boekje 8 (referentiegroep). Voor deze vergelijking is gebruik gemaakt van de z-toets. Géén enkel boekje laat een Z-waarde zien van kleiner dan -1,96 of groter dan 1,96. Dit betekent dat er geen significante verschillen zijn gevonden binnen de toetsmodus voor rekenen (Tabel 7 en 8).

Tabel 7
Verschillen tussen boekjes voor de papieren rekentoets

	Gemiddelde	SE	Z-waarde
Boekje 1	-0,10	0,29	-0,38
Boekje 2	-0,10	0,30	-0,03
Boekje 3	0,04	0,31	0,13
Boekje 4	-0,34	0,32	-1,05
Boekje 5	0,11	0,30	0,38
Boekje 6	-0,08	0,27	-0,30
Boekje 7	0,10	0,26	0,40
Boekje 8	0,00	1,00	-

Noot: Boekje 8 is gebruikt als referentiegroep.

Tabel 8
Verschillen tussen boekjes voor de digitale rekentoets

	Gemiddelde	SE	Z-waarde
Boekje 1	0,10	0,21	0,51
Boekje 2	0,13	0,27	0,46
Boekje 3	0,41	0,34	1,23
Boekje 4	-0,23	0,38	-0,60
Boekje 5	0,27	0,29	0,92
Boekje 6	0,43	0,43	1,00
Boekje 7	0,03	0,27	0,10
Boekje 8	0,00	1,00	-

Noot: Boekje 8 is gebruikt als referentiegroep.

Daarnaast zijn tussen de toetsmethodes ook geen verschillen gevonden. Hiervoor is een t-toets gedaan. Aangezien de steekproeven voldoende groot zijn kunnen de verkregen waardes benaderd worden als Z-waardes. Er zijn geen Z-waardes gevonden die kleiner zijn dan -1,96 of groter zijn dan 1,96. Er zijn dus geen significante verschillen gevonden tussen gelijknamige boekjes tussen toetsmethodes (zie Tabel 9).

Tabel 9
Verschillen tussen boekjes tussen de groepen papierentoets en tablettoets voor rekenen

	Δ (μ papier –μ tablet)	Δ (SE papier – SE tablet)	Z-waarde
Boekje 1	-0,21	0,35	-0,60
Boekje 2	-0,22	0,40	-0,55
Boekje 3	-0,37	0,46	-0,81
Boekje 4	-0,12	0,50	-0,23
Boekje 5	-0,15	0,41	-0,37
Boekje 6	-0,51	0,50	-1,02
Boekje 7	0,08	0,37	0,22
Boekje 8	-	-	-

Noot: Boekje 8 is gebruikt als referentiegroep.

3.4 Verschil in itemparameters tussen papieren en digitale rekentoets

In deze paragraaf wordt beschreven welke items (opgaven) significant afwijken tussen de papieren en tablet versie op de itemparameters bèta (moeilijkheidsgraad) en alfa (differentiegraad of onderscheidend vermogen). In bijlage 3 staan de bijbehorende tabellen, Tabel B4 en B5 voor de bèta en Tabel B6 en B7 voor de alfa. Hier staan items beschreven met een significant verschil op de bèta en alfa.

In totaal verschillen 13 items significant van elkaar op de bèta (14,5% van alle items). Items met een negatieve Z-waarde worden ‘moeilijker’ bevonden op de tablet. Dit geldt voor negen items. Vier items hebben een positieve Z-waarde deze worden ‘moeilijker’ bevonden op de papieren toets. Naar verhouding zijn er iets meer vragen die ‘moeilijker’ worden bevonden op tablet dan op papier. Naast de bèta is er gekeken of items significant afwijken in hun onderscheidend vermogen. Tien items verschillen significant van elkaar op de alfa (11% van alle items). Zes items hebben een negatieve Z-waarde. Dit betekent dat deze items een hogere discriminatie waarde laten zien op de tablettoets. Vier items hebben een positieve Z-waarde. Deze items hebben een beter discriminerend vermogen op de papieren toets. Vijf items blijken zowel op de bèta als op de alfa te verschillen. Vier van deze items

zijn moeilijker en discrimineren beter in de tablettoets. Het vijfde item is moeilijker op de tablet maar discrimineert beter in de papierentoets.

Van de items die verschillen laten zien tussen papier en tablet in alfa, bèta of beide is vervolgens nagegaan of ze overeenkomstige kenmerken hebben. Van de 13 items die verschillen op de bèta behoren 11 items tot het inhoudsdomein Getallen. Het inhoudsdomein Getallen is echter ook het grootste domein in de TIMSS-toets. Van alle items over Getallen laat 22% een significant verschil zien op de bèta tussen de papieren en de tablettoets, meestal in het nadeel van de tablettoets (8 items hebben hogere moeilijkheidsgraad). Twee van de items met afwijkende bèta's behoren tot het domein Geometrische vormen & meten, dit is slechts 8% van alle items in dit inhoudsdomein.

De items die verschillen op de discriminatie parameter zijn allemaal items uit het inhoudsdomein getallen (20% van alle items in dit inhoudsdomein). Met andere woorden, een relatief groot aantal items binnen dit inhoudsdomein heeft 'anders' reagerende itemparameters in de vergelijking tussen papieren en tablettoets. Tot slot lijkt het erop dat de soort opgave ook een rol speelt; acht van de tien items die verschillen in onderscheidend vermogen zijn open vragen.

3.5 Ruwe scores voor natuuronderwijs

In Tabel 10 en 11 staan per boekje het gemiddelde en de standaarddeviatie van de ruwe scores op de papieren en digitale natuuronderwijstoets. De leerlingen maakten ongeveer 20 tot 25 natuuronderwijs opgaves per toetsboekje. De hoogst behaalde scores en de gemiddelden op een boekje verschillen tussen papier en tablet. Vier boekjes zijn beter gemaakt op de papieren toets en vier boekjes zijn beter gemaakt op de tablettoets. Het grootste verschil laten boekje 7 en 8 zien. Beide boekjes zijn beter gemaakt op de papieren toets. In bijlage 4 staat de scoreverdeling grafisch weergegeven.

Tabel 10
Beschrijvende statistiek natuuronderwijs papier

	Aantal leerlingen	Maximum aantal behaalde punten	Gemiddelde	Standaarddeviatie
Boekje 1	67	22	11,15	3,67
Boekje 2	66	20	12,18	3,96
Boekje 3	68	25	14,44	4,83
Boekje 4	70	23	14,59	4,73
Boekje 5	71	23	13,75	4,67
Boekje 6	65	24	13,54	4,19
Boekje 7	68	26	16,19	5,34
Boekje 8	67	28	17,81	5,93

Tabel 11
Beschrijvende statistiek natuuronderwijs tablet

	Aantal leerlingen	Maximum aantal behaalde punten	Gemiddelde	Standaarddeviatie
Boekje 1	63	24	14,17	4,82
Boekje 2	60	21	14,15	4,06
Boekje 3	62	22	13,03	4,44
Boekje 4	70	21	12,69	4,07
Boekje 5	72	26	16,46	5,38
Boekje 6	66	27	17,41	4,82
Boekje 7	66	20	11,55	4,71
Boekje 8	65	20	11,97	4,00

3.6 Verschillen tussen papieren en digitale natuuronderwijstoets

Zoals in het methodedeel staat aangegeven, zijn er verschillende IRT-modellen geanalyseerd. In Tabel 12 staan de uitkomsten van de vergelijking tussen modellen op basis van de log-likelihood ($\Delta - 2ll$). Hieruit komt naar voren dat het meest simpele model PCM (het Rasch model) kan worden verworpen omdat het GPCM-model significant een betere fit van de data laat zien. Het tweedimensionaal GPCM-model verschilt significant van het eendimensionaal GPCM-model en geeft daarom een betere fit van de data weer. Vervolgens zijn acht marginalen (één voor elk boekje) aan het tweedimensionale GPCM-model toegevoegd. Dit model geeft de beste modelfit weer. Met dit model worden twee latente vaardigheden geschat op basis van de itemparameters bèta en alfa en zijn tevens de acht toetsboekjes meegenomen in de schatting.

Tabel 12
Vergelijking van de modelfit natuuronderwijs

	$\Delta\, df$	Δ-2ll	P
1-dim PCM vs. 1-dim GPCM	215	408	<0,001
1-dim GPCM vs. 2-dim GPCM	1	18	<0,001
2-dim GPCM vs. 2-dim GPCM (8 marginalen)	35	206	<0,001

Vervolgens is de correlatie berekend van de vaardigheidsverdelingen van de twee toetsen op basis van het tweedimensionale GPCM-model. De twee vaardigheidsverdelingen blijken sterk te correleren (r=0,88). In het tweedimensioneel GPCM-model blijft de ranking van de thèta's (vaardigheidsscore van de leerling) zo goed als hetzelfde voor beide toetsen. Het responsegedrag van de leerlingen op de natuuronderwijstoets wordt het beste verklaard door twee latente vaardigheden, maar deze correleren erg hoog. Dit betekent dat de papieren en tablettoets bijna dezelfde vaardigheden lijken te meten.

Vervolgens is nagegaan in hoeverre de boekjes binnen de toetsmodus en tussen de toetsmodus (papier versus tablet) van elkaar verschillen. Alle boekjes zijn telkens vergeleken met boekje 8 (referentiegroep). Voor deze vergelijking is gebruik gemaakt van de z-toets. Géén enkel boekje laat

een Z-waarde zien van kleiner dan -1,96 of groter dan 1,96. Dit betekent dat er geen significante verschillen zijn gevonden binnen de toetsmodus voor natuuronderwijs (Tabel 13 en 14).

Tabel 13

Verschillen tussen boekjes binnen de groep papierenmodus voor natuuronderwijs

	Gemiddelde	SE	Z-waarde
Boekje 1	-0,25	0,24	-1,06
Boekje 2	-0,003	0,26	-0,01
Boekje 3	0,10	0,28	0,36
Boekje 4	0,11	0,28	0,40
Boekje 5	0,24	0,31	0,77
Boekje 6	0,29	0,28	1,04
Boekje 7	0,006	0,27	1,04
Boekje 8	0	1,00	-

Noot. Boekje 8 is gebruikt als referentiegroep.

Tabel 14

Verschillen tussen boekjes binnen de groep tabletmodus voor natuuronderwijs

	Gemiddelde	SE	Z-waarde
Boekje 1	-0,14	0,28	-0,51
Boekje 2	-0,24	0,30	-0,81
Boekje 3	-0,30	0,34	-0,87
Boekje 4	-0,19	0,35	-0,55
Boekje 5	-0,07	0,37	-0,18
Boekje 6	-0,07	0,27	-0,26
Boekje 7	-0,08	0,27	-0,28
Boekje 8	0	1,00	-

Noot. Boekje 8 is gebruikt als referentiegroep.

Daarnaast zijn tussen de toetsmethodes ook geen verschillen gevonden. Hiervoor is een t-toets gedaan. Aangezien de steekproeven voldoende groot zijn en voldoen aan de voorwaarden, kunnen de verkregen waardes gezien worden als Z-waardes. Er zijn geen Z-waardes gevonden die kleiner zijn dan -1,96 of groter zijn dan 1,96. Er zijn dus geen significante verschillen gevonden tussen gelijknamige boekjes tussen toetsmethodes (zie Tabel 15).

Tabel 15

Verschillen tussen boekjes tussen de groepen papierentoets en tablettoets voor rekenen

	Δ (μ papier –μ tablet)	Δ (SE papier – SE tablet)	Z-waarde
Boekje 1	-0,11	0,37	-0,60
Boekje 2	0,24	0,39	-0,55
Boekje 3	0,40	0,44	-0,81
Boekje 4	0,30	0,47	-0,23
Boekje 5	0,17	0,48	-0,37
Boekje 6	0,37	0,39	-1,02
Boekje 7	0,08	0,38	0,22
Boekje 8	-	-	-

Noot. Boekje 8 is gebruikt als referentiegroep.

3.7 Verschil in itemparameters tussen papieren en digitale natuuronderwijstoets

In deze paragraaf wordt beschreven welke items significant afwijken tussen de papieren en tablet versie op de itemparameters bèta (moeilijkheidsgraad) en alfa (differentiegraad). In bijlage 5 staan de bijbehorende tabellen, B8 en B9 voor de bèta en B10 en B11 voor de alfa. Hier staan items beschreven met een significant verschil op de bèta en alfa.

In totaal verschillen 19 items significant van elkaar op de bèta parameter. Dit zijn items met een Z-waarde lager dan -1,96 of hoger dan 1,96. Items met een negatieve Z-waarde worden ‘moeilijker’ op de tablet bevonden. Dit geldt voor zeven items. Items met een positieve Z-waarde worden ‘moeilijker’ bevonden op de papierentoets. Dit geldt voor twaalf items. Naar verhouding zijn er meer vragen die ‘moeilijker’ worden bevonden op de papierentoets. Naast de bèta is er gekeken of items significant afwijken op de alfa (onderscheidend vermogen). Drie items verschillen significant op de alfa. Waarvan één item een beter discriminerend vermogen heeft op de tablet (negatieve Z-waarde) en twee items hebben een beter discriminerend vermogen op papier. Deze twee items horen bij elkaar. Ze verschillen zowel significant op de alfa als op de bèta.

Vervolgens is nagegaan of deze items overeenkomstige kenmerken hebben (Bijlage V, Tabel B9 en B11). Voor de items die verschillen op de bèta behoren vijf tot het inhoudsdomein Biologie (11,6%), acht tot Natuur- en scheikunde (23,5%) en zes tot Fysische aardrijkskunde (31,6%). Opvallend hier is dat alle items die positief significant verschillen behoren tot het inhoudsdomein Fysische aardrijkskunde. Deze items worden moeilijker bevonden op tablet. Wanneer gekeken wordt tot welk cognitief domein de items behoren dan vallen negen binnen het domein ‘toepassing’, vier binnen ‘redeneren’ en zes binnen ’weten’.

Voor de items die verschillen op de alfa, zijn het alle drie meerkeuzevragen uit het cognitief domein toepassen. De eerste heeft als inhoudsdomein Biologie en de andere twee Fysische aardrijkskunde.

3.8 Verschillen in toets prestaties jongens en meisjes

In voorgaande paragrafen is gekeken in hoeverre er verschillen zijn tussen de twee toetsmethodes. In deze paragraaf gaat het om de mogelijke sekseverschillen in prestaties op de papieren toets en de tablettoets. Ten eerste is onderzocht of er een hoofdeffect is van sekse op de hele toets, waarbij er geen onderscheid is gemaakt tussen papier en tablet of naar vakgebied. Vervolgens is nagegaan of er een interactie-effect is tussen sekse en toetsmodus en of het verschilt voor rekenen of natuuronderwijs. Voor deze analyse is gebruik gemaakt van de populatie parameters uit tweedimensioneel GPCM-model.

Als de gemiddelde scores van de meisjes op de toets vergeleken worden met die van jongens, blijkt dat meisje significant hoger scoren (Z-waarde = 2,25, $p < 0,05$). De effectgrootte Cohen's d is 0,13. Op basis van Cohen's classificatie van effectgroottes, is dit een klein effect (Field, 2009). Dit betekent dat het gevonden effect voor een zeer klein deel bijdraagt aan de verklaring van de variantie in het model (ongeveer 1 à 2 procent). De prestaties van jongens en meisje zijn vervolgens vergeleken voor de papierentoets en tablettoets (interactie-effect). Op de papierentoets verschillen meisjes en jongens niet in hun prestaties. Er is echter wel een significant verschil gevonden tussen jongens en meisjes in hun prestaties op de tabletoets. Meisjes scoren significant hoger op de tablettoets (Z-waarde = 2,3, $p < 0,05$). De effectgrootte Cohen's d is 0,15, dit is een klein effect. Meisjes scoren dus iets beter op de tablettoets dan jongens.

Tot slot is nagegaan of het interactie-effect wordt teruggevonden binnen de hoofddomeinen rekenen en natuuronderwijs. Er zijn geen significante verschillen gevonden binnen de hoofdomeinen rekenen en natuuronderwijs wanneer er gekeken wordt naar het effect van geslacht in combinatie met toetsmethode op de toetsprestaties.

4. Conclusie en discussie

In het voorjaar van 2017 hebben ruim 500 leerlingen tweemaal de TIMSS-toets gemaakt uit 2015; één keer op papier en één keer op de tablet. Leerlingen kregen de tweede keer een andere selectie van de in totaal 190 toetsopgaven over rekenen en natuuronderwijs voorgelegd dan de eerste keer. Op de helft van de scholen maakten de leerlingen eerst de papieren toets en een week later de digitale toets. Op de overige scholen was de procedure omgedraaid.

Het voornaamste doel van deze equivalentiestudie was om na te gaan of er prestatieverschillen zijn tussen de twee toetsvormen. Met behulp van Item Response Theorie is de data van de Nederlandse equivalentiestudie geanalyseerd. Voor dit onderzoek zijn drie onderzoeksvragen geformuleerd. Deze worden in de volgende paragraaf beantwoord.

4.1 Conclusie

Onderzoeksvraag 1

In welke mate zijn er verschillen in vaardigheidsniveau en itemparameters (moeilijkheidsgraad en discriminatiegraad) tussen de leerlingen die de papieren TIMSS-toets van 2015 hebben gemaakt en de leerlingen die de papieren equivalentiestudie toets in 2017 hebben gemaakt?

Vanwege de *convenience sampling* en het kleine aantal deelnemende scholen (23) kunnen de uitkomsten niet gegeneraliseerd worden naar alle groep 6 leerlingen in Nederland. Om deze reden zijn het vaardigheidsniveau en het functioneren van de papieren toetsitems van de steekproef uit 2017 (ruim 500 leerlingen) vergeleken met die van de steekproef van het hoofdonderzoek 2015 (bijna 4500 leerlingen). De gemiddelde toetsprestaties van leerlingen uit 2017 blijken vergelijkbaar te zijn met leerlingen uit 2015. Uit de analyse van de DIF-statistieken blijkt dat de items over het algemeen in 2017 en in 2015 hetzelfde functioneerden. Ook de overeenkomsten tussen de achtergrondkenmerken van de leerlingen wijzen erop dat de leerlingen van deze equivalentiestudie niet noemenswaardig afwijken van de representatieve steekproef uit 2015.

Onderzoeksvraag 2

Wat zijn de overeenkomsten en verschillen tussen de papieren toets en de tablettoets in vaardigheidsniveau en itemparameters (moeilijkheids- en discriminatiegraad) in de equivalentiestudie?

Om de tweede onderzoeksvraag te beantwoorden zijn met hulp van IRT-analyses zowel de vaardigheidsverdelingen (thèta) als de eigenschappen van de toetsitems (bèta en alfa) vergeleken tussen de papieren toets en tablettoets. De beste fit voor deze data is een tweedimensioneel GPCM-model. Hieruit blijkt dat zowel rekenen als natuuronderwijs een hoge correlatie laten zien tussen de twee vaardigheidsverdelingen. Dit betekent dat de twee toetsmethodes dezelfde vaardigheden lijken te meten. Het responsgedrag van de leerlingen wordt dus zowel voor rekenen als natuuronderwijs verklaard door twee latente vaardigheden die zeer hoog correleren.

Daarna is onderzocht of de acht toetsboekjes van elkaar verschillen binnen één methode en tussen de methodes. Voor rekenen en natuuronderwijs zijn noch verschillen gevonden tussen de 8 boekjes binnen de methode, noch tussen de methodes (papier of digitaal). Dit betekent dat de verkregen scores overeen lijken te komen, ongeacht de methode waarmee de toets wordt aangeboden.

Op basis van ruwe (gemiddelde) toetsscores leken de toetsen overigens wel te verschillen, vier boekjes waren beter gemaakt in de papieren versie vier boekjes waren beter gemaakt in de digitale versie. Het IRT-model dat onderscheid maakt tussen de vaardigheid van leerlingen en de moeilijkheid en onderscheidingsvermogen van de items laat echter zien dat er geen betekenisvolle verschillen zijn.

Tot slot is onderzocht of de itemparameters (eigenschappen van de toetsitems) van de beide toetsmethodes voor rekenen en natuuronderwijs significant afwijken in moeilijkheidsgraad en differentiatiegraad. Dit blijkt in balans te zijn; er zijn zowel enkele items die afwijken op de bèta en/of alfa in het nadeel van de tablet en items waarvan de bèta's en alfa's lager zijn in de papierenversie. Het is echter wel raadzaam om in het hoofdonderzoek kritisch naar het inhoudsdomeinen Getallen te kijken. Dit inhoudsdomein is oververtegenwoordigd wat betreft verschillen in itemparameters tussen papier en digitaal.

Onderzoeksvraag 3

In hoeverre is geslacht van invloed op de verschillen tussen de papieren toets en de tablettoets prestaties?

Om de derde onderzoeksvraag te beantwoorden is er zowel naar het hoofdeffect van sekse op de toetsprestaties gekeken als naar het interactie-effect van sekse in combinatie met toetsmethode op de toetsprestaties. Het blijkt dat meisjes iets beter scoren op de gehele toets (digitaal en papier en beide vakgebieden samen) dan jongens. Het gevonden effect is echter klein. Daarnaast is er een interactie-effect gevonden van geslacht in combinatie met toetsmethode. Meisjes scoren iets hoger op de tablettoets, ook dit effect is echter klein. Daarnaast is er gekeken of dit verschil naar voren komt binnen één van de hoofddomeinen. Hier is geen verschil gevonden tussen de hoofddomeinen rekenen en natuuronderwijs.

Samengevat lijken de papieren TIMSS toets en de tablet TIMSS toets hetzelfde te meten. Zowel de vaardigheidsverdelingen als de itemparameters komen overeen tussen de papieren en digitale toets. De kleine verschillen die zijn gevonden lijken elkaar uit te middelen. De overgang van papier naar digitale toetsing lijkt in Nederland vooralsnog de vergelijking tussen eTIMSS-2019 en voorgaande TIMSS-cycli niet in gevaar te brengen. Het is echter nog niet zeker dat dit ook voor de andere eTIMSS-landen geldt. Om deze reden zal in alle eTIMSS-landen het hoofdonderzoek van TIMSS worden uitgebreid met de zogenoemde *bridge study*. Naast de scholen voor het digitale hoofdonderzoek worden er extra scholen getrokken voor een papieren afname. Zodat het bij trendvergelijkingen of vergelijkingen tussen TIMSS-landen en eTIMSS-landen mogelijk is om verschillen toe wijzen aan ontwikkelingen in onderwijsniveau en zo nodig te corrigeren voor het effect van de gebruikte toetsmethode.

4.2 Discussie

Aangezien steeds meer leerlingen in Nederland te maken hebben met vormen van digitaal onderwijs sluit de eTIMSS toets goed aan op de leeromgeving van de leerling. Uit deze studie is gebleken dat de leerlingen goed overweg konden met de tablet en de TIMSS-software. Er zijn echter ook nog een aantal verbeteringen mogelijk. De toets geeft nog geen informatie over het antwoordgedrag van leerlingen zoals de tijd die een leerling nodig heeft voor het geven van een antwoord. Verder zou de digitale toets beter kunnen aansluiten op specifieke leerbehoeftes van de leerling. Een voorbeeld hiervan is een groter lettertype of een gesproken instructie voor leerlingen met leesproblemen. Hierdoor kan de

toets nog beter de reken- en natuuronderwijs vaardigheden in kaart brengen en voorkomt het een te grote uitsluiting van leerlingen met bijvoorbeeld dyslexie.

Meisjes hebben de tablettoets significant iets beter gemaakt. Dit is tegengesteld aan de uitkomsten van voorgaande TIMSS-metingen. Zo kwam uit het TIMSS-2015 naar voren dat jongens significant beter scoren op rekenen dan meisjes (Meelissen & Punter 2016). Deze verschillen werden ook gevonden in eerdere TIMSS-studies. Het sekseverschil in deze studie lijkt te zijn omgeslagen in het voordeel van de meisjes. Wellicht veranderen de gevonden resultaten voor sekse weer in het hoofdonderzoek omdat het gevonden effect zeer klein is. Deze bevinding is echter wel interessant om in de gaten te houden voor de hoofdmeting van 2019, zeker omdat in deze meting leerlingen ook getoetst worden met interactieve, samenhangende taken *(Problem Solving Inquiry*). Het is zeker relevant om na te gaan of hierin prestatieverschillen zijn en verschillen in antwoordgedrag zijn tussen jongens en meisjes.

Bijlage I

Tabel B1

Overzicht deelnemende scholen eTIMSS pilot 2017 naar denominatie

	Originele Steekproef		Eerste/Tweede Vervangers		Zelf Geworven		Totaal	
Denominatie	n	%	n	%	n	%	n	%
Openbaar	1		1		1		3	13%
Rooms-Katholiek	3		3		6		12	52%
Protestant-Christelijk	1		3		1		5	22%
Algemeen Bijzonder	1		1		1		3	13%
Totaal	6	26	8	35	9	39	23	100%

Tabel B2

Overzicht deelnemende scholen eTIMSS pilot 2017 naar schoolgrootte

	n	%
1 t/m 100 leerlingen	6	26
101 t/m 200 leerlingen	6	26
201 t/m 300 leerlingen	8	35
301 t/m 400 leerlingen	3	13

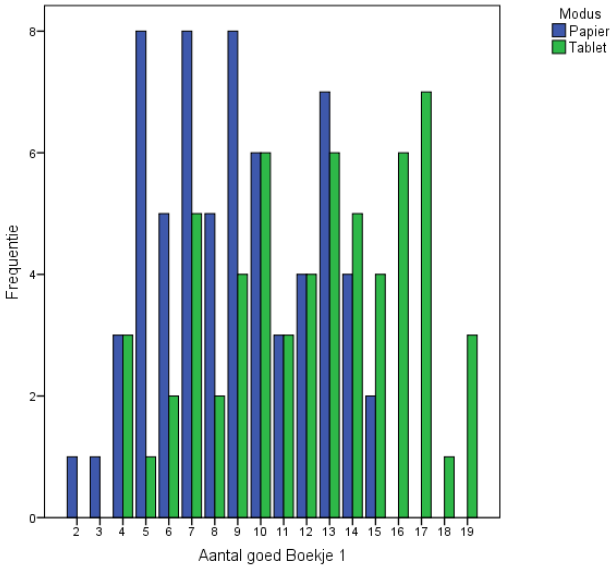
Tabel B3

Overzicht deelnemende scholen eTIMSS pilot 2017 naar omvang vestigingsplaats

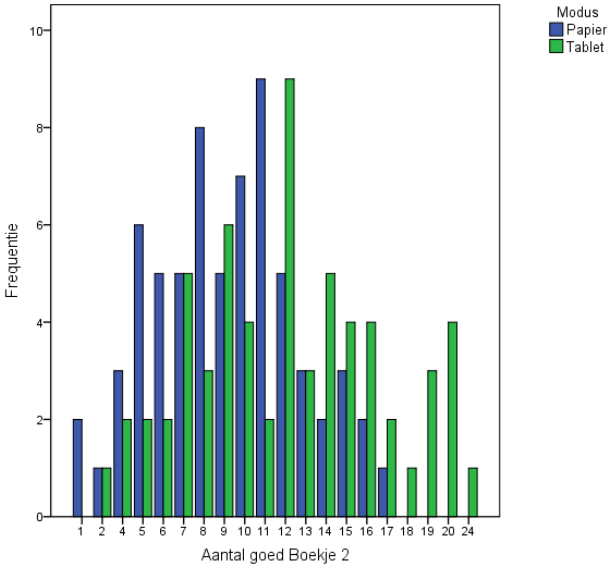
Een plaats of stad met …	n	%
< 15.000 inwoners	9	39
15.000 – 100.000 inwoners	10	44
>100.000 inwoners	4	17

Bijlage II

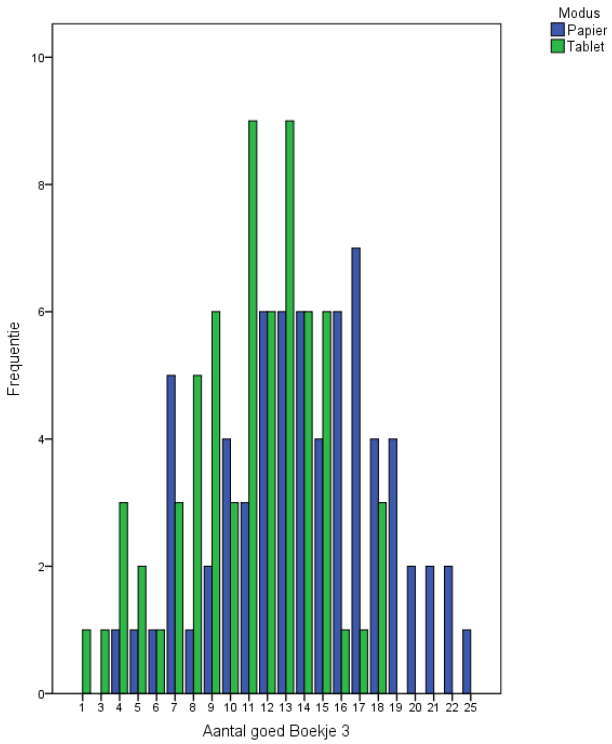
Frequentieverdelingen van de toetsboekjes rekenen geclusterd naar modus



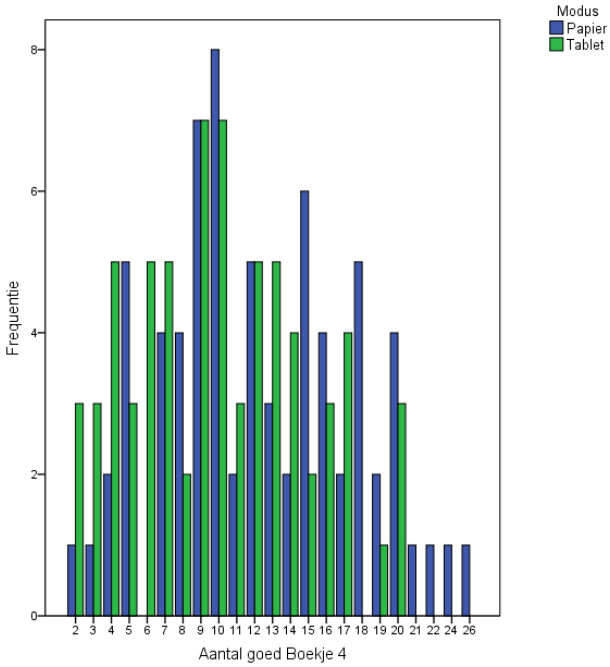
Figuur 1: Frequentieverdeling boekje 1



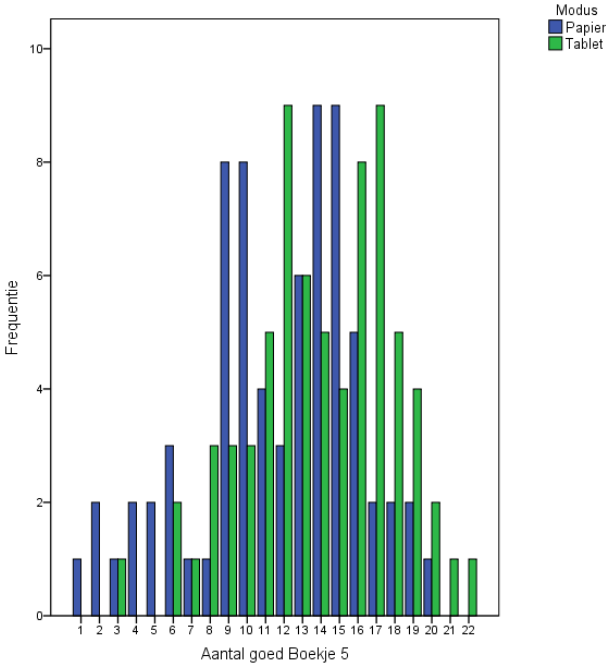
Figuur 2: Frequentieverdeling boekje 2



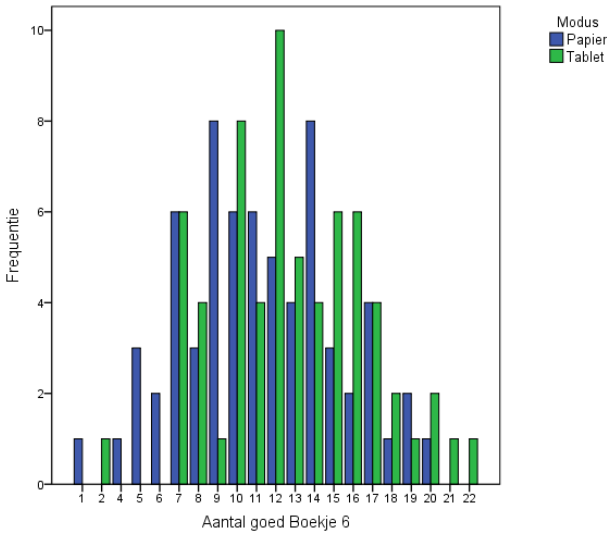
Figuur 3: Frequentieverdeling boekje 3



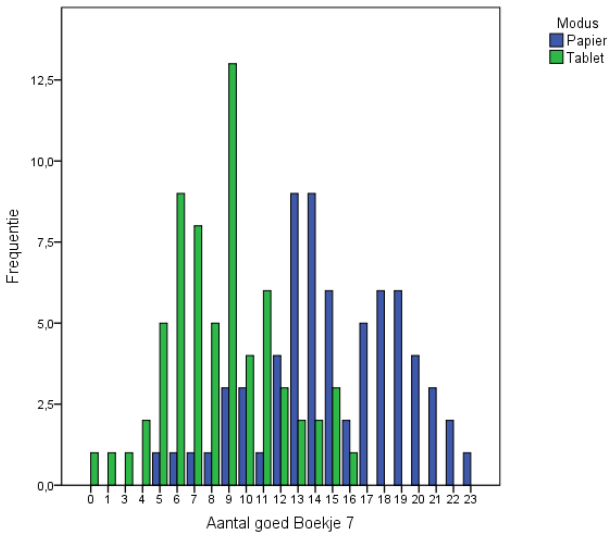
Figuur 4: Frequentieverdeling boekje 4



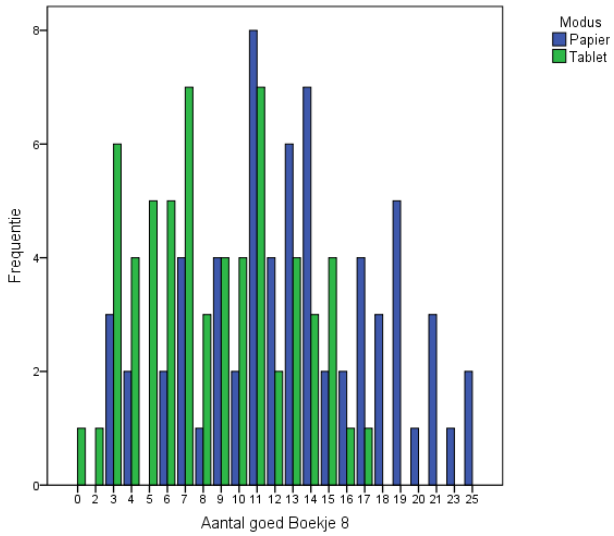
Figuur 5: Frequentieverdeling boekje 5



Figuur 6: Frequentieverdeling boekje 6



Figuur 5: Frequentieverdeling boekje 7



Figuur 6: Freqentieverdeling boekje 8

Bijlage III

Tabel B4

Verschil itemparameters bèta op papier en tablet voor rekenen

Item code	n	Δ (β_{papier}- β_{tablet})	Δ (SE_{papier} – SE_{tablet})	Z-waarde
M061021	122	-4,7	2,27	-2,07*
M061172	125	0,88	0,34	2,56**
M061081B	108	1,47	0,66	2,24*
M051052	130	11,82	4,27	2,77**
M051049	128	-1,16	0,3	-3,89**
M051045	1126	-1,85	0,59	-3,16**
M061018A	122	-1,81	0,45	-3,99**
M061018C	130	-1,54	0,51	-3,02**
M061018D	132	-1,21	0,53	-2,3*
M061039	130	-0,76	0,3	-2,55**
M061207	132	-0,91	0,29	-3,16**
M061049	128	1,97	0,67	2,95**
M051008	130	-1,17	0,33	-3,6**

Noot: Significantieniveau * <0,05; ** <0,01

Tabel B5

Beschrijving van de afwijkende items op bèta rekenen

Item code	Inhoudsdomein	Cognitief Domein	Vraagvorm
M061021	Getallen	Redeneren	Open
M061172	Getallen	Toepassen	Gesloten
M061081B	Geometrische vormen & meten	Toepassen	Open
M051052	Getallen	Weten	Gesloten
M051049	Getallen	Toepassen	Gesloten
M051045	Getallen	Toepassen	Open
M061018A	Getallen	Weten	Gesloten
M061018C	Getallen	Weten	Gesloten
M061018D	Getallen	Weten	Gesloten
M061039	Getallen	Toepassen	Open
M061207	Geometrische vormen & meten	Weten	Gesloten
M061049	Getallen	Toepassen	Gesloten
M051008	Getallen	Redeneren	Gesloten

Tabel B6

Verschil itemparameters alfa papier en tablet voor rekenen

Item code	n	Δ (α_{papier} - α_{tablet})	Δ (SE_{papier} – SE_{tablet})	Z-waarde
M051049	128	-1,07	0,47	-2,26*
M051045	126	-0,93	0,43	-2,19*
M051052	123	-1,09	0,36	-3,05**
M061018A	133	-0,82	0,4	-2,04*
M061018C	130	-0,96	0,38	-2,53**
M061248	127	-0,51	0,21	-2,37*
M061178	130	1,1	0,52	2,11*
M051008	130	1,12	0,4	2,8**
M051031A	127	0,95	0,4	2,34**
M051031B	126	1,5	0,45	3,32**

Noot: Significantieniveau * <0,05; ** <0,01

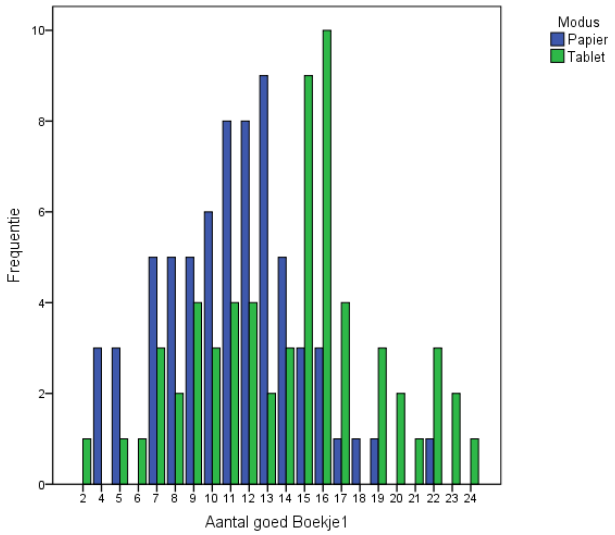
Tabel B7

Beschrijving van de afwijkende items op alfa rekenen

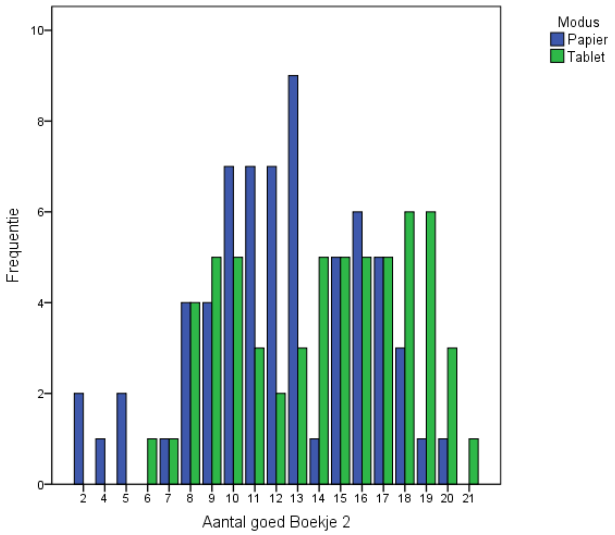
Item code	Inhoudsdomein	Cognitief Domein	Vraagvorm
M051049	Getallen	Toepassen	Meerkeuze
M051045	Getallen	Toepassen	Open
M051052	Getallen	Weten	Meerkeuze
M061018A	Getallen	Weten	Open
M061018C	Getallen	Weten	Open
M061248	Getallen	Redeneren	Open
M061178	Getallen	Weten	Open
M051008	Getallen	Redeneren	Open
M051031A	Getallen	Toepassen	Open
M051031B	Getallen	Toepassen	Open

Bijlage IV

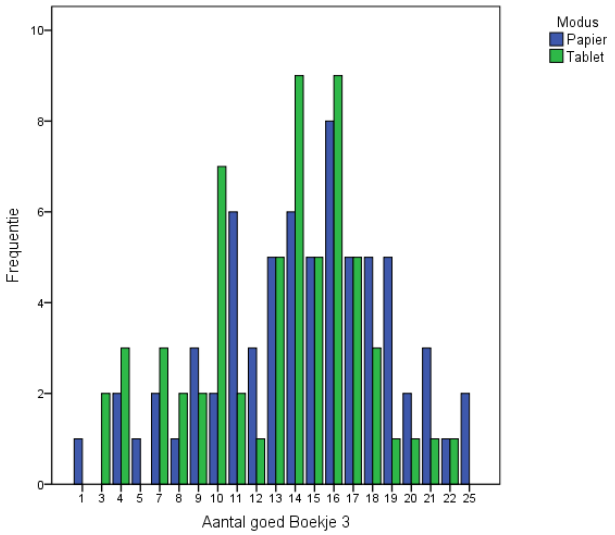
Frequentieverdelingen van de toetsboekjes natuuronderwijs geclusterd naar modus



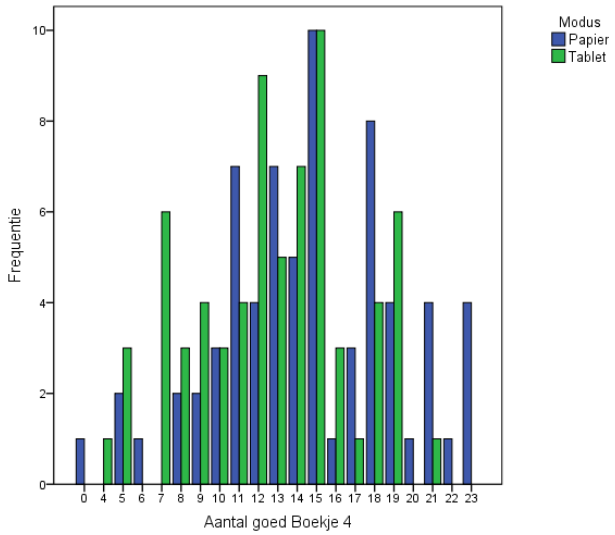
Figuur 9: Frequentieverdeling toetsboekje 1



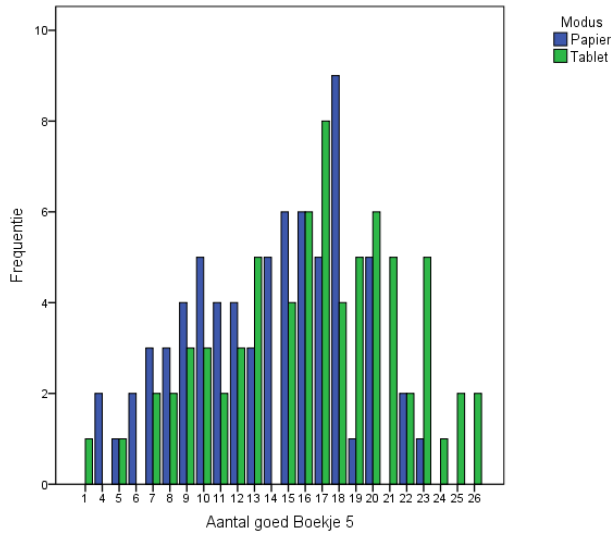
Figuur 10: Frequentieverdeling toetsboekje 2



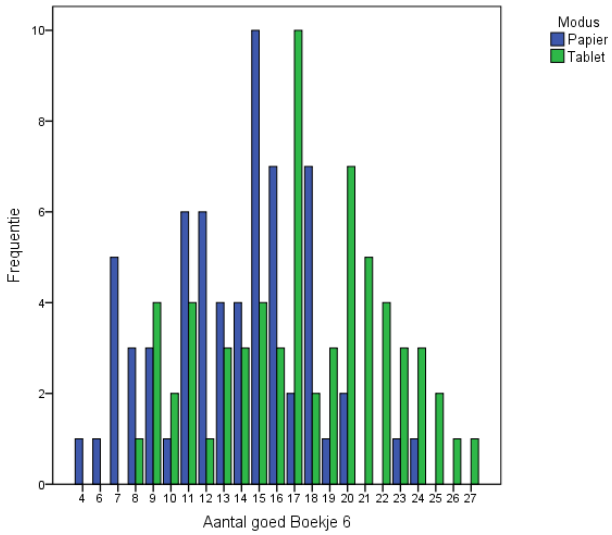
Figuur 11: Frequentieverdeling toetsboekje 3



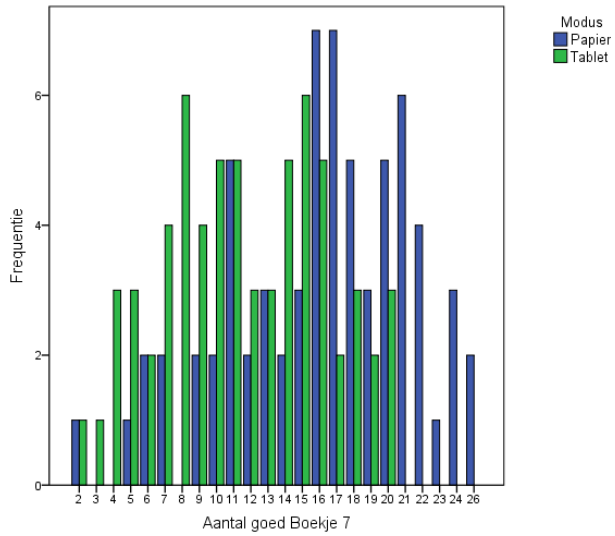
Figuur 12: Frequentieverdeling toetsboekje 4



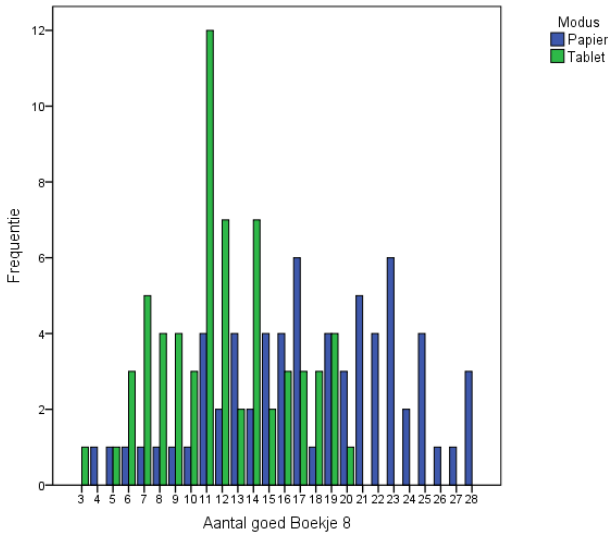
Figuur 13: Frequentieverdeling toetsboekje 5



Figuur 14: Frequentieverdeling toetsboekje 6



Figuur 15: Frequentieverdeling toetsboekje 7



Figuur 16: Frequentieverdeling toetsboekje 8

Bijlage V

Tabel B8

Verschil itemparameters bèta papier en tablet voor natuuronderwijs

Item code	n	Δ (β_{papier}- β_{tablet})	Δ (SE_{papier} – SE_{tablet})	Z-waarde
S051168	131	2,3	1	2,3*
S051142	126	1,18	0,56	2,11*
S061011	130	-0,74	0,36	-2,06*
S061083B	130	0,28	0,07	4,22**
S061083C	130	0,29	0,06	5,27**
S061034	133	-0,63	0,29	-2,16*
S061115A	132	-0,28	0,12	-2,37**
S051138A	137	0,64	0,26	2,45**
S051194	140	0,88	0,15	6**
S051065	139	-0,37	0,18	-2,02*
S061093	134	-2,34	0,43	-5,46**
S061042B	132	0,4	0,15	2,58**
S061041B	133	-0,66	0,27	-2,41**
S061107	128	-0,42	0,19	-2,21*
S061124D	125	-1,81	0,88	-2,04*
S061124E	125	-1,14	0,42	-2,74**
S061116C	125	-0,5	0,08	-6,61**
S061116D	123	-0,77	0,08	-9,94**
S061116E	122	-0,8	0,21	-3,75**

Noot: Significantieniveau * <0,05; ** <0,01

Tabel B9

Beschrijving van de afwijkende items op bèta natuuronderwijs

Item code	Inhoudsdomein	Cognitief Domein	Vraagvorm
S051168	Biologie	Toepassen	Open
S051142	Natuur- en scheikunde	Toepassen	Meerkeuze
S061011	Biologie	Redeneren	Open
S061083B	Natuur- en scheikunde	Weten	Meerkeuze
S061083C	Natuur- en scheikunde	Weten	Meerkeuze
S061034	Natuur- en scheikunde	Toepassen	Open
S061115A	Fysische Aardrijkskunde	Toepassen	Meerkeuze
S051138A	Biologie	Weten	Meerkeuze
S051194	Biologie	Redeneren	Open
S051065	Natuur- en scheikunde	Toepassen	Meerkeuze
S061093	Biologie	Weten	Open
S061042B	Natuur- en scheikunde	Redeneren	Meerkeuze
S061041B	Natuur- en scheikunde	Toepassen	Open
S061107	Natuur- en scheikunde	Redeneren	Meerkeuze
S061124D	Fysische Aardrijkskunde	Weten	Meerkeuze
S061124E	Fysische Aardrijkskunde	Weten	Meerkeuze
S061116C	Fysische Aardrijkskunde	Toepassen	Meerkeuze
S061116D	Fysische Aardrijkskunde	Toepassen	Meerkeuze
S061116E	Fysische Aardrijkskunde	Toepassen	Meerkeuze

Tabel B10

Verschil itemparameter alfa papier en tablet voor natuuronderwijs

Item code	n	Δ (α_{papier}- α_{tablet})	Δ (SE_{papier} – SE_{tablet})	Z-waarde
S061134	131	-1,42	0,56	-2,52**
S061116C	125	2,03	0,58	3,52**
S061116D	123	1,95	0,64	3,04**

Noot: Significantieniveau * <0,05; ** <0,01

Tabel B11

Beschrijving van de afwijkende items op alfa natuuronderwijs

Item code	Inhoudsdomein	Cognitief Domein	Vraagvorm
S061134	Biologie	Toepassen	Meerkeuze
S061116C	Fysische Aardrijkskunde	Toepassen	Meerkeuze
S061116D	Fysische Aardrijkskunde	Toepassen	Meerkeuze

Literatuur

Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists.* Mahwah, New Jersey: Lawerence Erlbaum Associates.

Birnbaum, A. (1968). Some latent trait models. In F.M. Kird & M.R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Field, A. (2009). *Discovering Statistics Using SPSS (3th edition).* London, United Kingdom: SAGE Publications.

Furr, R.M. & Bacharach, V.R. (2008). *Psychometrics: An Introduction (2th edition).* Los Angeles, California: Sage Publications.

Glas, H. (2010). Software Program Multidimensional Item Response Theory (MIRT) [Software]. Enschede, The Netherlands: University of Twente.

Glas, C. & Geerlings, H. (2015) *Research Master Course Psychometrics.* Enschede, The Netherlands: University of Twente.

Meelissen, M., & Punter, A. (2016). *Twintig jaar TIMSS. Ontwikkelingen in leerlingprestaties in de exacte vakken in het basisonderwijs 1995-2015.* Enschede, The Netherlands: University of Twente.

Mullis, I.V.S., & Martin, M.O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks.* Chestnut Hill: Boston College.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

TIMSS & PIRLS International Study Center. (z.d.) About TIMSS & PIRLS International Study Center. Geraadpleegd op 28 februari 2018, van https://timssandpirls.bc.edu/about.html