PEDAGOGISCHE STUDIËN
TIJDSCHRIFT VOOR ONDERWIJSKUNDE EN OPVOEDKUNDE

Tonen de Vlaamse peilingsonderzoeken en internationale studies dezelfde trends in leerlingprestaties?

J. Dockx, L. Pelgrims, K. Aesaert, en R. Janssen

Samenvatting Het Vlaamse onderwijs wordt geconfronteerd met dalende leerlingprestaties in verschillende internationaal vergelijkende studies. Deze studies hanteren daarbij een toetskader dat is gericht op het meten van vaardigheden over verschillende landen en culturen. Hierdoor is het onbekend of een trend voor een bepaalde vaardigheid in een internationale studie gelijk is aan de trend van de vaardigheid zoals deze begrepen wordt in de Vlaamse context. Vlaanderen organiseerde echter tussen 2002 en 2022 jaarlijks peilingsonderzoeken die wel gebaseerd waren op de Vlaamse eindtermen. In deze studie onderzochten we de trends voor lezen en wiskunde volgens de Vlaamse peilingsonderzoeken en vergeleken deze met trends in internationale studies. Daarnaast gingen we de robuustheid na van de trends in internationaal vergelijkende studies door nieuwe itemparameters enkel op basis van Vlaamse gegevens te schatten. De resultaten toonden dat de trends voor de peilingsonderzoeken en internationale onderzoeken gelijkaardig zijn voor overeenkomstige vaardigheden. Wanneer de itemparameters van de internationale studies enkel op basis van de Vlaamse gegevens werden geherkalibreerd, werden de trends meer gelijkaardig in grootte.

Kernwoorden internationaal vergelijkend onderzoek, nationale peilingen, wiskunde, begrijpend lezen, item respons theorie (IRT)

Artikelgeschiedenis
Ontvangen: 1 juni 2023
Ontvangen in gereviseerde vorm: 3 oktober 2023
Geaccepteerd: 23 oktober 2023
Online: 21 december 2023

Contactpersoon
Jonas Dockx,
jonas.dockx@kuleuven.be

Financiering onderzoek
-

Belangen
De auteurs hebben geen belangen te vermelden.

https://doi.org/10.59302/ps.v100i4.18358
2023 (100) 421-447

1 Inleiding

Internationaal vergelijkende studies toonden de voorbije jaren dalende leerlingprestaties in het Vlaamse lager en secundair onderwijs. Zo toonde de *Programme for International Student Assessment* (PISA) de voorbije afnames een daling in de gemiddelde leesvaardigheid, wiskundige geletterdheid en wetenschappelijke geletterdheid van Vlaamse 15-jarigen. Voor wiskundige geletterdheid was Vlaanderen zelfs de grootste daler tussen 2006 en 2018 (De Meyer et al., 2019). De *Trends in International Mathematics and Science Study* (TIMSS) en de *Progress in International Reading Literacy Study* (PIRLS) vonden dan weer een daling in de gemiddelde wiskundevaardigheid en leesvaardigheid van Vlaamse leerlingen in het vierde leerjaar. Voor leesvaardigheid was Vlaanderen ook hier de grootste daler tussen 2006 en 2016, en had in 2016 de vierde laagste score van alle Europese onderwijssystemen (Faddar et al., 2020; Tielemans, Vandenbroeck, Bellens, Van Damme, & De Fraine, 2017).

Internationaal vergelijkende studies hebben echter als beperking dat hun toetsen afgestemd moeten zijn op diverse onderwijssystemen en culturen. Wanneer in deze studies een vaardigheid zoals wiskundige geletterdheid wordt gedefinieerd, dan is dit een internationale interpretatie van de vaardigheid die zich voornamelijk richt op wat er gemeenschappelijk is tussen landen (Hambleton & Zenisky, 2011; Mullis & Martin, 2015, 2017; OECD, 2019a). Een internationaal vergelijkende studie toetst dus niet de inhoud van een vaardigheid zoals ze wordt begrepen binnen één specifiek onderwijssysteem of volgens één curriculum. Een trend in een internationale studie beschrijft dan ook niet noodzakelijk de trend van een vaardigheid zoals ze omschreven wordt binnen de Vlaamse eindtermen en leerplandoelen.

Vlaanderen organiseerde sinds 2002 jaarlijkse peilingsonderzoeken die vanaf 2018 werden uitgevoerd door het Steunpunt Toetsontwikkeling en Peilingen (STEP). In deze peilingsonderzoeken werd onderzocht hoeveel leerlingen de eindtermen bereiken binnen een leer- of vakgebied op het einde van een onderwijsniveau (het lager onderwijs of één van de drie graden in het secundair onderwijs; OECD, 2011, p. 30; STEP, 2023). In tegenstelling tot de internationaal vergelijkende studies, werd het toetsmateriaal ontwikkeld in functie van de Vlaamse eindtermen die voor alle leerlingen van het betreffende onderwijsniveau gelden. Deze peilingsonderzoeken omvatten daarbij ook verschillende herhalingspeilingen, waarin werd onderzocht of het relatieve aantal leerlingen dat de eindtermen bereikt was veranderd. Zo boden de peilingsonderzoeken de mogelijkheid om trends in gemiddelde leerlingprestaties te beschrijven aan de hand van toetsen die zich inhoudelijk richten op de Vlaamse eindtermen en context.

In deze studie bekijken we de gemiddelde trends volgens de peilingsonderzoeken wiskunde en Nederlands lezen tijdens het lager onderwijs en het secundair onderwijs. De overkoepelende vraag is wat de trends zijn in

https://doi.org/10.59302/ps.v100i4.18358

gemiddelde leerlingprestaties volgens de peilingsonderzoeken en of deze trends overeenstemmen met TIMSS, PIRLS en PISA. Daarom vergelijken we in de volgende sectie eerst de toetskaders van de internationaal vergelijkende studies. Daarna gaan we dieper in op de technische aspecten van het beschrijven van gemiddelde trends in leerlingprestaties. Vervolgens bespreken we de opzet van de peilingsonderzoeken. Ten slotte bespreken we de onderzoeksvragen en het studieopzet.

2 Theoretisch kader

2.1 Toetskaders internationaal vergelijkende studies

TIMSS en PIRLS worden internationaal gecoördineerd door de *International Association for the Evaluation of Educational Achievement* (Mullis, Martin, Foy, Hooper, & IEA, 2017; Mullis, Martin, Foy, Kelly, & Fishbein, 2020) en vinden respectievelijk om de vier en vijf jaar plaats. PISA wordt gecoördineerd door de *Organisation for Economic Co-operation and Development* (OECD; OECD, 2019b, pp. 5–6) en vindt om de drie jaar plaats. Hoewel TIMSS, PIRLS en PISA elk internationaal vergelijkende studies zijn, zijn ze niet uitwisselbaar. De studies van de IEA en OECD hebben namelijk verschillende doelen, definiëren de door hun gemeten vaardigheden op een andere wijze en richten zich op een andere soort populatie.

TIMSS en PIRLS hebben als doel om onderwijssystemen te vergelijken in hoe goed hun leerlingen het curriculum voor een bepaalde vaardigheid verwerven. Voor PIRLS is dit het curriculum voor (begrijpend) lezen, en voor TIMSS zijn dit wiskunde en wetenschappen (Mullis & Martin, 2015, 2017). De IEA benadrukt daarbij het belang van cross-culturele validiteit: in alle deelnemende landen moet hetzelfde construct gemeten worden. De ontwikkeling van het toetskader start dan ook met een analyse van het curriculum van elk deelnemend land. De overeenkomstige curriculumonderdelen waarvoor de leerlingen in elk land voldoende leeraanbod hebben gehad vormen de basis van het toetskader en de toetsontwikkeling (Valverde, Bianchi, Wolfe, Schmidt, & Houang, 2002). TIMSS en PIRLS richten zich daarbij op de populatie leerlingen in het vierde leerjaar lager onderwijs. De IEA redeneert namelijk dat leerlingen na vier jaar lager onderwijs voldoende formeel onderwijs hebben genoten in alle onderwijssystemen (Mullis & Martin, 2015, p. 55). Het vierde leerjaar is ook het laatste leerjaar lager onderwijs in enkele onderwijssystemen. Het vierde leerjaar lijkt zo het beste moment om het lager onderwijs tussen onderwijssystemen te vergelijken.

PISA heeft als doel om te beschrijven in welke mate 15-jarigen de kennis en vaardigheden hebben verworven die essentieel zijn voor een deelname

https://doi.org/10.59302/ps.v100i4.18358

aan moderne maatschappijen (OECD, 2019a). Het richt zich op wat het zelf omschrijft als de drie kerndomeinen van kennis en vaardigheden: leesvaardigheid, wiskundige geletterdheid en wetenschappelijke geletterdheid. Bij elke afnamecyclus wordt er voor één van de kerndomeinen een internationaal expertenpanel samengesteld om op basis van de wetenschappelijke literatuur een toetskader te ontwikkelen dat toepasselijk is voor alle deelnemende landen (OECD, 2019a). Voor de twee andere kerndomeinen wordt het kader van de voorgaande afnamecycli gebruikt. PISA richt zich op de populatie van 15-jarigen en dus niet op een specifiek leerjaar. De OECD redeneert namelijk dat leerlingen op deze leeftijd het leerplichtonderwijs in de meeste onderwijssystemen bijna volledig doorlopen hebben en dit het beste moment is om de resultaten het leerplichtonderwijs te vergelijken (OECD, 2019a).

De studies van de IEA en OECD hebben dus elk een andere focus. Hierbij richten TIMSS en PIRLS zich op het lager onderwijs, definiëren de doelpopulatie op basis van leerjaar en beoogt men vaardigheden te meten die gemeenschappelijk zijn over de curricula van de deelnemende onderwijssystemen. PISA richt zich daarentegen op het secundair onderwijs, definieert de doelpopulatie volgens leeftijd en meet vaardigheden die noodzakelijk worden geacht voor de deelname aan moderne maatschappijen.

2.2 Trends in internationaal vergelijkende studies

Wanneer onderwijssystemen herhaaldelijk deelnamen aan PIRLS, TIMSS en PISA, dan is het mogelijk om hun trends in gemiddelde leerlingprestaties over de afnames heen te beschrijven. De IEA en de OECD plaatsen namelijk elke nieuwe afname hun toetsscores op dezelfde meetschaal als de voorgaande afnames. Het beschrijven van trends veronderstelt echter (een mate van) *construct equivalence* over de afnames heen en is niet mogelijk zonder *linking error* (Wu, 2010).

Om *linking error* en *construct equivalence* te begrijpen, beschrijven we eerst hoe de OECD en de IEA de toetsscores van twee opeenvolgende afnamecycli (bijvoorbeeld PISA 2015 en PISA 2018) op eenzelfde meetschaal plaatsen. Alle internationale studies gebruiken hiervoor ankeritems. Dit zijn items die worden afgenomen binnen meerdere afnamecycli (Martin, Mullis, & Hooper, 2017; Martin, von Davier, & Mullis, 2020; OECD, 2019b). De kenmerken van de ankeritems (zoals de moeilijkheidsgraad) worden beschreven aan de hand van itemparameters in *Item Response Theory* (IRT) modellen. In PISA 2018 koos men ervoor om de itemparameters over te nemen uit 2015, waardoor de scores van 2015 en 2018 op dezelfde meetschaal staan (OECD, 2020, Chapter 9). TIMSS 2019 en PIRLS 2016 gebruikten een andere methode, waarbij eerst alle itemparameters vrij geschat werden over de twee laatste afnamecycli (Martin et al., 2017, 2020). Vervolgens werd het lineaire verband berekend tussen de parameterschattingen van de ankeritems binnen de nieuwe afnamecyclus en

https://doi.org/10.59302/ps.v100i4.18358

de voorgaande afnamecyclus. De lineaire functie diende vervolgens om de toetsscores van de twee laatste cycli op dezelfde meetschaal te plaatsen.

Toch is er steeds enige onzekerheid over bij het plaatsen van de toetsscores van verschillende afnamecycli op dezelfde meetschaal. Dit is het gevolg van *linking error*, wat zowel een toevals- als een systematische component omvat. De toevalscomponent van de *linking error* is het gevolg van de schattingsfout van de itemparameters, aangezien de echte waarden onbekend zijn. De OECD en de IEA erkennen deze toevalscomponent van de *linking error* en berekenen deze voor de landgemiddeldes (Martin, Mullis, Foy, Brossman, & Stanco, 2012; Martin et al., 2020, pp. 12.34-12.40; OECD, 2020, Chapter 9). TIMSS, PIRLS en PISA houden echter geen rekening met de systematische component van de *linking error*, welke bestaat uit het verschillend functioneren van items tussen landen en over de afnamecycli binnen landen (Glas & Jehangir, 2013; Monseur & Berezner, 2007; Robitzsch & Lüdtke, 2019; Sachse & Haag, 2017). Wanneer een item verschillend functioneert voor een land, dan is de internationale parameter voor dat land onjuist. Dit kan bijvoorbeeld veroorzaakt worden verschillen in curricula tussen landen en hoe curricula binnen landen veranderen (Robitzsch & Lüdtke, 2019), of verschillen in taal of cultuur (Sachse & Haag, 2017). Kleine veranderingen zoals hoe de toets er uitziet of de itempositie binnen de toets hebben eveneens een invloed. Ook verschuivingen in welk construct gemeten wordt, waar we in de volgende paragraaf beschrijven, is een bron voor systematische *linking error* (Monseur & Berezner, 2007).

Het beschrijven van trends in gemiddelde leerlingprestaties veronderstelt namelijk *construct equivalence*, wat inhoudt dat hetzelfde construct gemeten wordt over de verschillende afnames (Byrne & de Vijver, 2010; Kane, 2017). Internationale studies actualiseren hun constructen echter over de afnames heen. Zo wordt in PISA één van de drie hoofddomeinen bij elke afnamecyclus herbekeken. Zo werden bijvoorbeeld in 2000 aspecten van betrokkenheid en tekstevaluatie toegevoegd aan hun definitie van leesvaardigheid. Naast de geschreven teksten zijn er nu ook digitale of andere teksten (OECD, 2019a). Daarenboven gebruikt PISA nu vooral digitale toetsen, waarbij niet volledig wordt vastgehouden aan de itemparameters van de papieren toetsen. Ook bij TIMSS en PIRLS is er elke afnamecyclus een heranalyse van de curricula van de deelnemende landen en een re-evaluatie van het toetskader (Mullis & Martin, 2015, 2017). PIRLS en TIMSS zijn eveneens aan het overstappen naar een digitale afname, waarbij men niet volledig vasthoudt aan de itemparameters van de papieren afname (Martin et al., 2020, pp. 12.46-12.57). Ten slotte leidt de regelmatige itemverversing ertoe dat niet voor alle items onderzocht kan worden of de huidige items daadwerkelijk hetzelfde construct meten als de oorspronkelijke items die initieel betekenis gaven aan de meetschaal. Trends over verschillende afnamecycli zijn dus mogelijk (deels) toe te schrijven aan verschuivingen in het gemeten construct.

https://doi.org/10.59302/ps.v100i4.18358

2.3 Peilingsonderzoek in Vlaanderen

Peilingsonderzoeken werden sinds 2002 afgenomen om na te gaan hoeveel leerlingen de eindtermen behalen (AHOVOKS, 2022b, 2022a). De eindtermen omvatten de minimumdoelen van het Vlaamse onderwijs, en worden opgesteld voor het lager onderwijs en de drie graden van het secundair onderwijs. Het lager onderwijs heeft één geheel van eindtermen die voor alle leerlingen gelden, terwijl het secundair onderwijs aparte eindtermen heeft voor de verschillende stromen, onderwijsvormen en studierichtingen. De eindtermen ondersteunen de klaspraktijk door enerzijds te beschrijven wat er minstens verworven moet worden, en anderzijds door te beschrijven wat de leerlingen reeds verworven hebben. De eindtermen zouden de basis moeten zijn van het curriculum binnen scholen, wat verder aangevuld kan worden met leerplandoelen, en leerkrachten worden geacht deze eindtermen te kennen.

Tussen 2007 en 2022 waren er jaarlijks twee peilingsonderzoeken, waarvan doorgaans één in het lager onderwijs en één in het secundair onderwijs. Elk peilingsonderzoek richtte zich op een inhoudelijk domein waaronder Nederlands of wiskunde (Carpentier, Costers, Janssen, & Willem, 2019, 2020; Denis, Talloen, Laenen, Janssen, & Aesaert, 2019; Spikic et al., 2022), maar ook domeinen zoals Frans, techniek, of burgerzin. De eerste fase van een peilingsonderzoek bestond uit de ontwikkeling van gevalideerd toetsmateriaal. Dit startte met een analyse van de Vlaamse eindtermen en lesmateriaal. Op basis hiervan werd het toetskader ontwikkeld. Daarbij werden steeds verschillende deeldomeinen (en bijhorende meetschalen) onderscheiden. Zo onderscheidde men voor Nederlands in het lager onderwijs in 2018 lezen, luisteren en schrijven (Denis et al., 2019). Bij wiskunde in het lager onderwijs in 2019 werden dan weer 21 meetschalen opgesteld (Spikic et al., 2022). Elk van deze deeldomeinen was gekoppeld aan één of meerdere eindtermen. Het finale toetskader was vervolgens het uitgangspunt voor de ontwikkeling van de toetsopgaven. Tijdens dit ontwikkelingsproces werden de initiële voorstellen herhaaldelijk beoordeeld en bijgestuurd door extern aangestelde expertenpanels, waarin zowel leerkrachten, onderzoekers, beleidsmakers als andere stakeholders vertegenwoordigd waren. Aansluitend was er ook steeds een pilootstudie en een kalibratiestudie om de kenmerken van de toetsopgaven te onderzoeken, en konden er indien nodig nog aanpassingen gemaakt worden.

Na de ontwikkeling van het gevalideerd toetsmateriaal vond in een tweede fase het feitelijke peilingsonderzoek plaats. Dit startte met de steekproeftrekking binnen de beoogde populatie, waarvan het toetsmateriaal vervolgens werd afgenomen via een onvolledig afnamedesign. Per deeldomein werd een meetschaal opgesteld, waarop zowel de leerlingen als items geplaatst werden. Vervolgens vond de cesuurbepaling plaats, het onderzoek naar welke score (de cesuur) leerlingen minstens moeten behalen op een toets van het

https://doi.org/10.59302/ps.v100i4.18358

peilingsonderzoek om de eindterm(en) van een deeldomein te bereiken. Dit gebeurde aan de hand van de *bookmark standard-setting* methode (Mitzel, Lewis, Patz, & Green, 2001), waarbij een groep van onderwijsexperts aangeven welke toetsopgaven leerlingen correct moeten kunnen beantwoorden om de eindterm(en) te bereiken. Voor elk van de deeldomeinen werd zo'n cesuur bepaald. Het voornaamste resultaat van een peilingsonderzoek was dan ook een schatting van het percentage leerlingen dat de eindterm(en) van elk van de onderzochte deeldomeinen bereikt.

Sommige peilingsonderzoeken waren herhalingspeilingen, zij onderzochten voor een tweede of derde keer hetzelfde inhoudelijke domein als een voorgaand peilingsonderzoek (Carpentier et al., 2019, 2020; Denis et al., 2019; Spikic et al., 2022) en gebruikten daarbij dezelfde meetschaal en cesuur. Net zoals bij het internationaal vergelijkend onderzoek was dit aan de hand van ankeritems. Bij deze herhalingspeilingen werd dan ook de evolutie in het percentage leerlingen dat de eindtermen bereikt gerapporteerd. Bijvoorbeeld, de peilingsonderzoeken Nederlands in het lager onderwijs toonden dat de eindtermen voor het deeldomein lezen behaald werden door 89% in 2007, 92% in 2013, en 84% in 2018. Ook het peilingsonderzoek maakt het beschrijven van trends dus mogelijk.

In tegenstelling tot internationaal vergelijkend onderzoek, waren peilingsonderzoeken dus specifiek gericht op de Vlaamse context. Daarbij was de cesuurbepaling op basis van een expertenbeoordeling van het toetsmateriaal een wezenlijk onderdeel van het onderzoek, en verkreeg men zo een expertgeleide concretisering van de eindtermen. TIMSS, PIRLS, en PISA hadden geen gelijkaardige cesuurbepaling, er zijn immers geen internationale minimumdoelen. Hun *benchmarks* of *levels* werden bepaald aan de hand van markeerpunten op de meetschaal, die werden vastgelegd in functie van de verdeling van de leerlingresultaten zelf (Martin et al., 2017, 2020; OECD, 2020). Wegens het hoofddoel van het peilingsonderzoek werden de peilingsresultaten gerapporteerd aan de hand van het relatieve aantal leerlingen dat de vooropgestelde cesuur bereikt. Voor het beschrijven van trends, maakt het rapporteren aan de hand van de onderliggende vaardigheidsschaal (IRT-schaal) een directere vergelijking met het internationaal vergelijkend onderzoek mogelijk.

PEDAGOGISCHE STUDIËN
https://doi.org/10.59302/ps.v100i4.18358

2.4 Deze studie

Het hoofddoel van deze studie was om de trends in gemiddelde leerlingprestaties te onderzoeken bij de herhalingspeilingen en na te gaan of deze gelijkaardig zijn aan de trends in TIMSS, PIRLS en PISA. De vaardigheden die gemeten werden binnen deze internationale studies waren namelijk niet specifiek gericht op de Vlaamse context, terwijl het peilingsonderzoek dat wel was. Zo werden ook voor de eerste keer de gemiddelde trends volgens

de herhalingspeilingen gerapporteerd aan de hand van hun onderliggende vaardigheidsschaal.

Aansluitend onderzochten we de robuustheid van de trends van het internationaal vergelijkend onderzoek voor het Vlaamse onderwijs. De trends zoals ze tot nu toe gerapporteerd zijn, waren geschat aan de hand van internationale itemparameters voor de IRT-modellen. Deze itemparameters zijn echter onderhevig aan de instabiliteit van itemparameters tussen onderwijssystemen, wat kan leiden tot systematische *linking error* die de trendschatting binnen een onderwijssysteem vertekent (Glas & Jehangir, 2013; Monseur & Berezner, 2007; Robitzsch & Lüdtke, 2019; Sachse & Haag, 2017). Deze vertekening kan echter vermeden worden door unieke itemparameters te schatten voor het onderzochte onderwijssysteem (Sachse, Roppelt, & Haag, 2016).

Voor deze studie gebruikten we voor het lager onderwijs de gegevens van de peilingsonderzoeken Nederlands uit 2007, 2013 en 2018 en wiskunde uit 2009, 2016 en 2021. Voor het secundair onderwijs gebruikten we peilingsonderzoeken wiskunde in de eerste graad A-stroom uit 2009 en 2019 en in de eerste graad B-stroom uit 2008 en 2019. De A-stroom richt zich op leerlingen die het getuigschrift basisonderwijs hebben behaald en bereidt voornamelijk voor op het algemeen secundair onderwijs (ASO) en technisch secundair onderwijs (TSO). De B-stroom richt zich op leerlingen die het getuigschrift niet hebben behaald en bereidt voor op het beroepssecundair onderwijs (BSO). De trends uit deze gegevens werden vervolgens vergeleken met de trends volgens PIRLS tussen 2006 en 2016, volgens TIMSS tussen 2011 en 2019, en volgens PISA tussen 2009 en 2018. We stelden de volgende onderzoeksvragen:

- Welke trends in gemiddelde leerlingprestaties zijn er volgens de peilingsonderzoeken Nederlands (2007-2018) en wiskunde (2009-2021) in het lager onderwijs, en wiskunde in het middelbaar onderwijs (B-stroom: 2008-2019, A-stroom: 2009-2018)?
- Hoe robuust zijn de trends zijn bij TIMSS, PIRLS en PISA wanneer we voor Vlaanderen unieke IRT-itemparameters gebruiken in plaats van de internationale IRT-itemparameters?
- Zijn de trends van overeenkomstige vaardigheden bij de peilingsonderzoeken en de internationale onderzoeken gelijkaardig?

Bij het vergelijken van trends richten we ons op overeenkomstige vaardigheden bij een gelijkaardige populatie. Bijvoorbeeld, we vergelijken de trend voor lezen in het lager onderwijs volgens de peilingsonderzoeken lezen in 2007, 2013 en 2018, met de trend voor lezen volgens PIRLS in 2006 en 2016. Daarbij vergelijken we de trends niet volgens afnamejaar maar het geboortejaar van de leerlingen. Zo waren zowel PIRLS 2016 in het vierde leerjaar als het peilingsonderzoek lezen 2018 in het zesde leerjaar gericht op leerlingen van het geboortejaar 2006. In beide steekproeven werden dus leerlingen

PEDAGOGISCHE STUDIËN
https://doi.org/10.59302/ps.v100i4.18358

met geboortejaar 2006 geselecteerd. Voor wiskunde in het lager onderwijs vergeleken we de trend volgens de peilingsonderzoeken wiskunde in het lager onderwijs met de trend volgens TIMSS. Voor wiskunde in het secundair onderwijs vergeleken we de trend volgens de peilingsonderzoeken wiskunde in de A-stroom en B-stroom met de trend volgens PISA.

3 Methode

3.1 Steekproeven

Dit onderzoek gebruikte de gegevens van negen internationaal vergelijkende studies en tien peilingsonderzoeken. De steekproeven werden reeds uitgebreid besproken in eerdere rapporten en om beknoptheidsredenen geven we hier slechts een samenvatting. Zo werd over de verschillende afnamecycli heen de gewenste deelname van scholen bij de internationale studies steeds bereikt (Martin & Mullis, 2012; Martin, Mullis, & Hooper, 2016; Martin et al., 2017; Martin, Mullis, & Kennedy, 2007; Martin et al., 2020; OECD, 2012, 2014, 2017, 2020). Ook voor de peilingsonderzoeken gold dat de gewenste deelname van scholen steeds werd bereikt (Ameel et al., 2017; Carpentier et al., 2019, 2020; Denis et al., 2019; Gielen et al., 2010; Gielen, Willem, Beringhs, Luyten, & Janssen, 2009; Spikic et al., 2022; Van Nijlen, Denis, Willem, Ameel, & Janssen, 2017).

Zowel de internationaal vergelijkende studies als de peilingsonderzoeken gebruikten een gelijkaardige gestratificeerde tweetrapssteekproeftrekking. In de eerste stap vond stratificatie plaats op schoolniveau. De stratificatievariabelen omvatten doorgaans de provincies, de schoolgrootte, gemiddelde sociaaleconomische status (vaak op basis van de onderwijskansarmoede-indicatoren), studierichtingen (in het secundair onderwijs), onderwijsnet, of het type onderwijs (gewoon of buitengewoon). De internationale studies maakten binnen deze strata gebruik van *probability proportional to size* (PPS) *sampling*, waarbij de kans dat een school werd opgenomen in de eerste stap van de steekproeftrekking proportioneel was aan de schoolgrootte. Bij de peilingsonderzoeken had elke school een even grote kans om geselecteerd te worden. Stratificatie op schoolniveau werd toch als noodzakelijk geacht, omdat leerlingen geclustered zijn binnen scholen en er hierdoor sprake was van een *design-effect* die de effectieve steekproefgrootte verkleinde (Kish, 1965). Stratificatie zal dit effect enigszins verminderen (Martin et al., 2020, pp. 3.1-3.33).

Vervolgens werd in de tweede trap al dan niet een toevallige steekproeftrekking gemaakt van leerlingen of klassen binnen de geselecteerde scholen. Bij TIMSS 2011, TIMSS 2015 en PIRLS 2016 werden alle klassen binnen

Tabel 1

Steekproeven internationale studies en peilingsonderzoeken

Studie		*n* scholen	*n* leerlingen	Populatie	Geboortejaar[a]
Internationale studies					
TIMSS					
	2011	142	4849	4de leerjaar lager onderwijs	2001
	2015	153	5404		2005
	2019	147	4655		2009
PIRLS					
	2006	137	4479	4de leerjaar lager onderwijs	1996
	2016	148	5198		2006
PISA (wiskunde)					
	2009	156	4596	15-jarigen:	1993
	2012	174	5970		1996
	2015	175	5675		1999
	2018	169	4882		2002
Peilingsonderzoeken					
Wiskunde lager onderwijs					
	2009	251	6940	6de leerjaar lager onderwijs	1997
	2016	190	5421		2004
	2021	230	6163		2009
Lezen lager onderwijs					
	2007	105	2862	6de leerjaar lager onderwijs	1995
	2013	108	2794		2001
	2018	119	3120		2006
Wiskunde A-stroom					
	2009	75	3257	2de leerjaar A secundair onderwijs	1995
	2018	104	2985		2004
Wiskunde B-stroom					
	2008	195	5792	Beroepsvoorbereidend leerjaar	1994
	2019	117	3615		2005

[a] De vermelde geboortejaren bij TIMSS, PIRLS en de peilingsonderzoeken gelden voor de normaalvorderende leerlingen. Binnen deze steekproeven waren er ook leerlingen met vertraging of voorsprong die respectievelijk een vroeger of later geboortejaar hadden.

elke school geselecteerd, terwijl bij PIRLS 2006 en TIMSS 2019 twee klassen per school werden opgenomen in de steekproef. Bij PISA 2009, 2012, 2015 en 2018 werden tot 35 leerlingen binnen elke school opgenomen. In de internationale studies met een toevallige steekproeftrekking op leerling- of klasniveau, hadden leerlingen in een grotere school een kleinere kans hadden om opgenomen te worden, wat compenseert voor de grotere kans van de school om opgenomen te worden door PPS *sampling* (Martin et al., 2020; Mullis et al., 2017; OECD, 2020).

PEDAGOGISCHE STUDIËN
https://doi.org/10.59302/ps.v100i4.18358

De peilingsonderzoeken namen steeds alle leerlingen van de beoogde populatie op binnen een deelnemende school.

Tabel 1 toont het aantal scholen en leerlingen dat deelnam aan elke studie die in dit onderzoek wordt opgenomen. Daarnaast toont tabel 1 de beoogde populatie van elk onderzoek, met het geboortejaar.

Ten slotte merken we op dat bij recentere afnamecycli van de internationale onderzoeken het aantal exclusies werden beperkt (Martin et al., 2017, 2020; OECD, 2020). Daarom werden er recent meer leerlingen uit het buitengewoon onderwijs opgenomen, maar dit heeft geen noemenswaardige invloed op het gemiddelde prestatieniveau van Vlaanderen (Tielemans et al., 2017, p. 76). In onze analyses vonden we ook geen noemenswaardige effecten op de trends, en we tonen daarom enkel de trends op basis van de volledige datasets.

3.2 Uitkomsten

Oorspronkelijke vaardigheidsscores TIMSS, PIRLS en PISA
Voor de eerste onderzoeksvraag gebruikten we de oorspronkelijke vaardigheidsscores van TIMSS, PIRLS en PISA. De schaal voor elk van deze studies werd vastgelegd bij de eerste afnamecyclus. Zij kozen daarbij elk voor een gemiddelde van 500 en een standaarddeviatie van 100 over de deelnemende landen (Martin et al., 2017, 2020; OECD, 2020). Deze 500/100 schaal was op zijn beurt weer gebaseerd op een onderliggende IRT-schaal. Voor iedere leerling waren er vijf *plausible values* bij TIMSS, PIRLS en PISA 2009, en 10 *plausible values* bij PISA 2012, 2015 en 2018. In de resultaten refereren we naar de oorspronkelijke vaardigheidsscores als vaardigheidsscores op de internationale schaal.

Nieuwe schatting vaardigheidsscores op basis van Vlaamse gegevens
Voor het schatten van nieuwe vaardigheidsscores gebruikten we de gescoorde itemantwoorden van de Vlaamse leerlingen. De meeste items werden als juist of fout gescoord, of er werd een beoordeling gegeven van de mate van juistheid. De vaakst voorkomende itemtypes waren enkelvoudige meerkeuzevragen, samengestelde meerkeuzevragen en open vragen (Martin et al., 2020, Exhibit 1.5; OECD, 2020, Annex A). Voor de verschillende studies rapporteren we het aantal items die afgenomen werden over de verschillende afnamecycli en de inhoudelijke domeinen die daarbij onderscheiden werden in Tabel 2.

Om de gescoorde itemantwoorden van iedere leerling over de verschillende afnamecycli op één meetschaal te plaatsen (bijvoorbeeld PIRLS in 2006 en 2016), gebruikten we twee IRT-modellen:

- Het twee-parameter model (Martin et al., 2017) voor items waarbij het itemantwoord juist of fout (1 of 0) werd gescoord. In dit model is de kans

Tabel 2

Items en domeinen binnen internationale studies en peilingsonderzoeken

Studie	***n* items**	**Domeinen of deelvaardigheden (n items per domein)**
Internationale studies		
TIMSS 2011, 2015, en 2019	319	Inhoudsdomeinen: Getallen (159), geometrische vormen & meten (102) en data (58). Cognitieve domeinen: Kennen (114), toepassen (137), en redeneren (68).
PIRLS 2006 en 2016	249	Leesdoelen: Leeservaring (127) en informatieverwerving (122). Leesprocessen: Vinden van expliciet vermelde informatie (70), eenvoudige conclusies trekken (77), tekstinterpretatie (67), en tekstevaluatie (35).
PISA wiskunde 2009, 2012, 2015 en 2018	85	Inhoudelijke subschalen: Verandering en relaties (21), vorm en ruimte (21), hoeveelheid (22) en onzekerheid & data (21). Processen: Wiskundig formuleren van situaties (37), gebruik van wiskundige concepten, feiten & procedures (27), en interpreteren, toepassen & evalueren van wiskundige uitkomsten (21).
Peilingsonderzoeken		
Wiskunde lager onderwijs 2009, 2016 en 2021	460	Hoofdrekenen (35), functies & voorstellingswijzen (44), breuken & kommagetallen (45), getalwaarden & gelijkwaardigheden (45), afronden, benaderen en schatten (46), verhoudingen & schaal (47), procent berekenen (54), problemen oplossen bij getallen en bewerkingen (41), betekenisvolle herleidingen (68), en ruimte & ruimtelijke oriëntatie (41).
Lezen lager onderwijs 2009, 2016 en 2021	137	Geen onderscheid in domeinen
Wiskunde A-stroom 2009 en 2018	257	Getalinzicht (26), bewerkingen (20), rekenen met veeltermen (21), algebraïsering (34), evenredigheden (30), omgaan met data (28), meetkundige begripsvorming (24), meetkundige begripsvorming (23), meetkundige procedures: constructies (20), en ruimtemeetkunde (31).
Wiskunde B-stroom 2008 en 2019	251	Breuken optellen & aftrekken (29), geld & functioneel rekenen (54), meetkunde (55), informatieverwerving & -verwerking (58), en meten (55).

https://doi.org/10.59302/ps.v100i4.18358

op een correct itemantwoord een functie van de moeilijkheidsgraad en de discriminatiegraad van het item, en van de vaardigheid van de leerling.
- Het *generalized partial credit model* (Martin et al., 2017) voor items waarbij de mate van correctheid van het antwoord werd beoordeeld (bijvoorbeeld 0, 1 of 2). In dit model zijn de kansen op de verschillende waarden van de schaal een functie van de stap- en discriminatieparameters van het item, en van de vaardigheid van de leerling.

Deze twee IRT-modellen werden gespecificeerd als een unidimensioneel, multidimensioneel (met correlerende factoren), of bifactor model (Reise, 2012), afhankelijk van de beoogde dimensionaliteit bij de oorspronkelijke studies. Het unidimensionele IRT-model gebruikten we voor het peilingsonderzoek lezen (Figuur 1a), en de subschalen voor de peilingsonderzoeken wiskunde in het lager en secundair onderwijs. Het multidimensioneel IRT-model gebruikten we voor TIMSS (Figuur 1b), en PISA. Hiermee volgden we het oorspronkelijke model van deze studies waarbij alle vaardigheden van deze studie worden opgenomen binnen een multidimensioneel IRT-model (Martin et al., 2020, p. 11.5; OECD, 2020, Chapter 9). Het bifactor IRT-model gebruikten we voor de peilingsonderzoeken wiskunde in het lager onderwijs, secundair onderwijs (A- en B-stroom) (Figuur 1c). Zo werd de algemene wiskundevaardigheid, die onderliggend is bij de verschillende deeldomeinen, onderscheiden. Deze drie types modellen werden steeds als *multiple group* modellen gespecificeerd. Het gemiddelde en de variantie van de eerste afnamecyclus werden respectievelijk vastgezet op waardes nul en één, terwijl de gemiddelde en varianties van de latere afnamecyclussen vrij werden geschat. Bij de peilingsonderzoeken, TIMSS en PIRLS werden de itemparameters gelijk gehouden over de afnames. Bij PISA werden de itemparameters ook gelijk gehouden, behalve bij de items die volgens de OECD niet hetzelfde functioneren tussen de digitale en papieren afname (OECD, 2017).

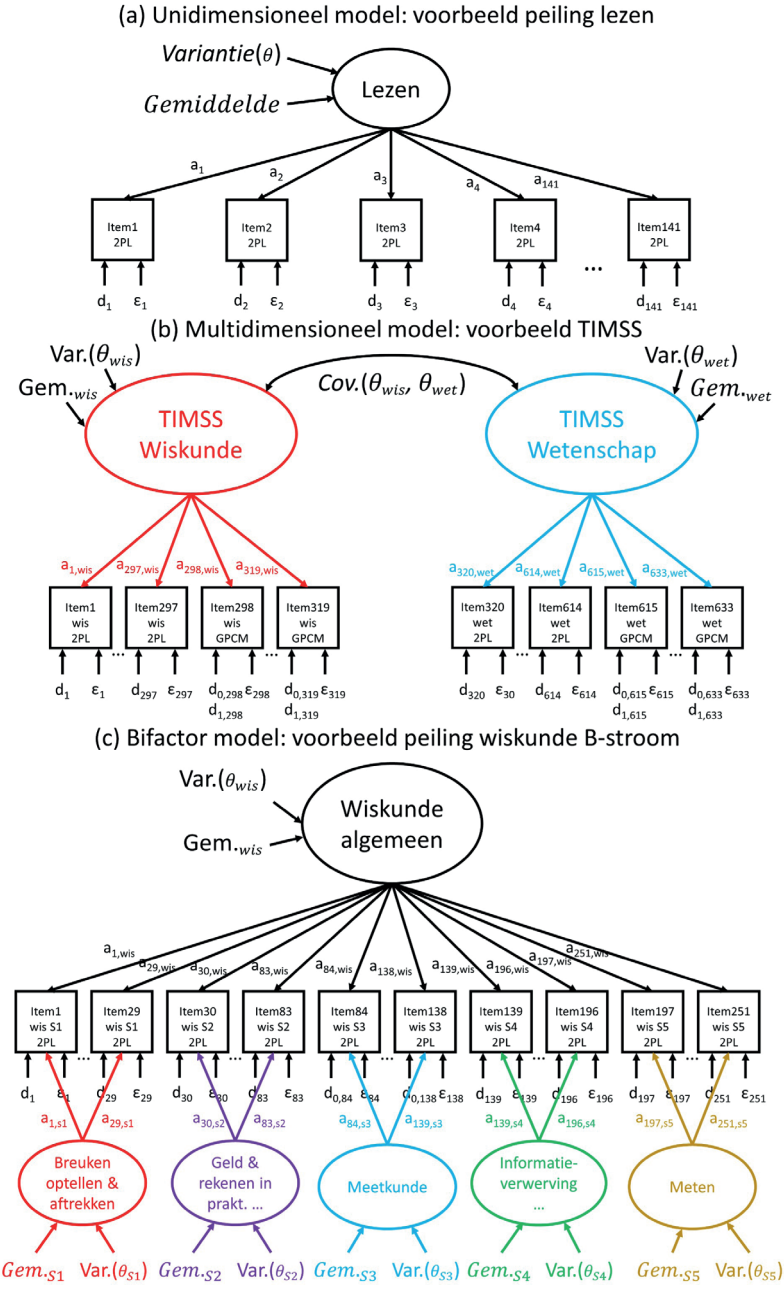
Op basis van deze IRT-modellen en de gescoorde itemantwoorden werden er vervolgens vijf *plausible values* geschat per leerling. Deze *plausible values* laten toe om de meetfout mee onderdeel te maken van de uitkomstanalyses, en tot correctere schattingen van de spreidingen in de vaardigheden en standaardfouten te bekomen (Foy & Yin, 2017; Matthias von Davier, Gonzalez, & Mislevy, 2009, pp. 11, 36). In de resultaten refereren we naar deze nieuwe vaardigheidsscores als vaardigheidsscores op de Vlaamse schaal.

3.3 Uitkomstanalyses

De uitkomstanalyse van elke studie bestond steeds uit een multiple group regressiemodel. Hierbij werd elke afnamecyclus als een groep beschouwd met zijn eigen gemiddelde en variantie. Bijvoorbeeld, bij de uitkomstanalyse van TIMSS waren er drie afnamecycli, en zijn er dus drie gemiddeldes en drie varianties die geschat werden. Hiervoor gebruikten we Mplus 8. Bij elke

Figuur 1

IRT-modellen voor het schatten van plausible values



J. Dockx, L. Pelgrims, K. Aesaert, en R. Janssen

https://doi.org/10.59302/ps.v100i4.18358

analyse werd ook het steekproefontwerp, de clustering van scholen en de stratificatie op schoolniveau, geïncorporeerd in de schatting. Zo gebruikten we bij TIMSS en PIRLS de *Jackknife replicate weights*, en bij PISA de *Fay's balanced replicate* weights via respectievelijke de "JACKKNIFE 2" en "FAY(.5)" methode in Mplus 8 om de standaardfouten te schatten (Muthén & Muthén, 2017). Bij de peilingsonderzoeken gebruikten we de empirische *bootstrap* methode waarbij de scholenclusters en strata geïncorporeerd werden bij het schatten van de standaardfouten. Elke set van plausible values werd apart geanalyseerd en de resultaten werden gecombineerd via de methode van Rubin (1987). Met deze methode wordt de meetfout die eigen is aan testresultaten gemodelleerd en wordt er zo een correctere standaardfout geschat. Om de trend te beschrijven werd het verschil geschat tussen de gemiddeldes van de eerste en laatste afname. Bij een verschil op basis van de internationale schaal, rapporteerden we ook steeds het gestandaardiseerde verschil op basis van de standaarddeviatie van de eerste afnamecyclus als *effect size* (ES). Ten slotte gebruikten we ook de gewichten van TIMSS en PIRLS (HOUWGT) en de gewichten van PISA (FSTUWT).

4 Resultaten

De gemiddelde leerlingprestaties en trends volgens de peilingsonderzoeken voor lezen in het lager onderwijs, de algemene schaal voor wiskunde in het lager onderwijs, en de algemene schaal voor wiskunde in het secundair onderwijs (A-stroom en B-stroom) worden beschreven in Tabel 3. De gemiddelde leerlingprestaties en trends voor lezen in het lager onderwijs volgens PIRLS, wiskunde in het lager onderwijs volgens TIMSS, en wiskunde in het secundair onderwijs volgens PISA (A-stroom en B-stroom) worden beschreven in Tabel 4. In deze tabel worden de resultaten zowel op de internationale schaal als op de Vlaamse schaal getoond. In de volgende paragrafen vergelijken we de resultaten van de peilingen en internationale onderzoeken.

PEDAGOGISCHE STUDIËN
https://doi.org/10.59302/ps.v100i4.18358

4.1 Lezen lager onderwijs

Voor lezen in het peilingsonderzoek vonden we tussen 2007 en 2018 een significant negatief verschil van -0.30. Voor lezen (*reading literacy*) op de internationale schaal in PIRLS vonden we tussen 2006 en 2016 een significant negatief verschil van -21.99 (ES = -0.40). Voor PIRLS op de Vlaamse schaal vonden we tussen 2006 en 2016 een significant negatief verschil van -0.31. Deze trends worden getoond in Figuur 2. Er was geen significant verschil in de trend tussen het peilingsonderzoek en de internationale schaal van PIRLS (delta= 0.09, 95% BI [-0.03, 0.22]), en geen significant verschil met de Vlaamse schaal van PIRLS (delta= 0.01, 95% BI [-0.10, 0.13]).

Tabel 3

Resultaten en trends peilingsonderzoeken

Studie		Gemiddelde	95% B.I.	*SD*
Lezen lager onderwijs				
	2007	0.00	[-0.07, 0.07]	1.00
	2013	0.02	[-0.03, 0.08]	0.88
	2018	-0.30	[-0.36, -0.24]	0.89
	Verschil 2007 en 2018	-0.30	[-0.39, -0.21]	
Wiskunde lager onderwijs				
	2009	0.00	[-0.06, 0.05]	1.00
	2016	-0.18	[-0.23, -0.12]	0.98
	2021	-0.19	[-0.25, -0.13]	0.98
	Verschil 2009 en 2021	-0.19	[-0.27, -0.12]	
Wiskunde A-stroom				
	2009	0.00	[-0.13, 0.13]	1.00
	2018	-0.13	[-0.27, 0.01]	0.97
	Verschil 2009 en 2018	-0.13	[-0.32, 0.06]	
Wiskunde B-stroom				
	2008	0.00	[-0.14, 0.14]	1.00
	2019	-0.25	[-0.33, -0.17]	1.04
	Verschil 2008 en 2019	-0.26	[-0.41, -0.09]	

Tabel 4

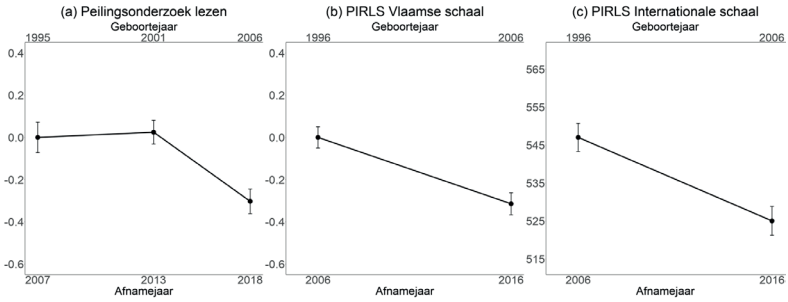
Resultaten en trends internationale studies

Studie		Internationale schaal			Vlaamse schaal		
		Gem.	95% B.I.	SD	Gem.	95% B.I.	SD
PIRLS							
	2006	547.0	[543.3, 550.8]	55.6	0.00	[-0.05,0.05]	1.00
	2016	525.1	[521.3, 528.9]	60.6	-0.31	[-0.37,-0.26]	0.99
	Verschil 2006 en 2016	-22.0	[-25.8, -18.2]		-0.31	[-0.39,-0.24]	
TIMSS							
	2011	549.2	[545.4, 553.0]	59.6	0.00	[-0.05,0.05]	1.00
	2015	545.7	[541.6, 549.7]	60.8	-0.05	[-0.11,0.02]	1.05
	2019	532.4	[528.7, 536.2]	67.5	-0.21	[-0.26,-0.15]	1.10
	Verschil 2011 en 2019	-16.8	[-22.0, -11.5]		-0.21	[-0.29,-0.13]	
PISA Wiskunde A-stroom							
	2009	577.8	[572.1, 583.6]	76.9	0.00	[-0.06, 0.06]	1.00
	2012	570.9	[564.6, 577.2]	82.4	-0.09	[-0.15, -0.02]	0.98
	2015	558.1	[553.0, 563.2]	78.7	-0.16	[-0.22, -0.10]	1.01
	2018	551.8	[546.8, 556.8]	77.6	-0.25	[-0.30, -0.20]	0.99
	Verschil 2009 en 2018	-26.0	[-33.7, -18.3]		-0.25	[-0.33, -0.17]	
B-stroom							
	2009	442.9	[436.2, 449.5]	65.0	0.00	[-0.10, 0.10]	1.00
	2012	423.3	[416.5, 430.1]	69.8	-0.25	[-0.35, -0.15]	1.03
	2015	419.2	[411.5, 427.0]	65.7	-0.29	[-0.40, -0.18]	1.00
	2018	410.8	[400.4, 421.2]	61.6	-0.26	[-0.38, -0.14]	0.93
	Verschil 2009 en 2018	-32.1	[-44.3, -19.9]		-0.26	[-0.42, -0.11]	

https://doi.org/10.59302/ps.v100i4.18358

Figuur 2

Trends lezen lager onderwijs



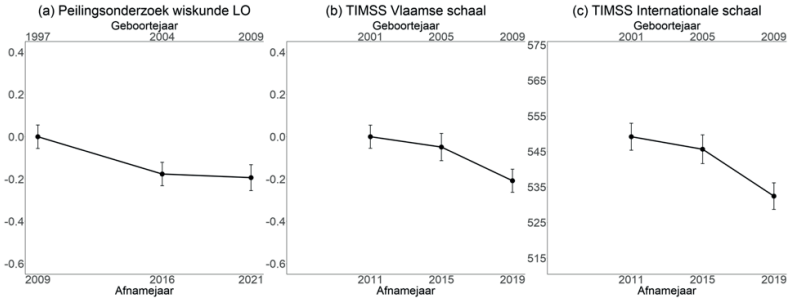
4.2 Wiskunde lager onderwijs

Voor de algemene schaal van wiskunde in het peilingsonderzoek vonden we tussen 2009 en 2021 een significant negatief verschil van -0.19. Wanneer we de tien subschalen apart onderzochten, vonden we voor zes subschalen een significante daling, voor één subschaal een significante stijging, en voor drie subschalen geen significante trend. De gemiddelde trend over de subschalen was -0.23, wat zich binnen het betrouwbaarheidsinterval van de algemene schaal bevond. Voor wiskunde (*mathematics literacy*) op de internationale schaal in TIMSS vonden we tussen 2011 en 2019 een significant negatief verschil van -16.75 (ES = -0.28). Voor TIMSS op de Vlaamse schaal vonden we tussen 2011 en 2019 een significant negatief verschil van -0.21. Deze trends worden getoond in Figuur 3. Er was geen significant verschil in de trend tussen het peilingsonderzoek en de internationale schaal van TIMSS (delta= 0.09, 95% BI [-0.02, 0.19]), en geen significant verschil met de Vlaamse schaal van TIMSS (delta= 0.01, 95% BI [-0.10, 0.13]).

PEDAGOGISCHE STUDIËN
https://doi.org/10.59302/ps.v100i4.18358

Figuur 3

Trends algemene schalen wiskunde lager onderwijs

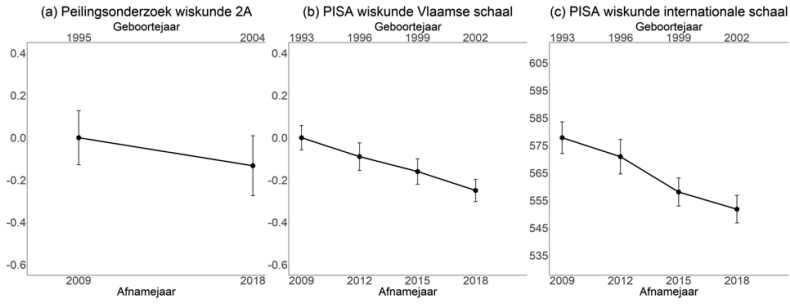


4.3 Wiskunde secundair onderwijs A-stroom

Voor wiskunde in het peilingsonderzoek vonden we tussen 2009 en 2018 een niet-significant verschil van -0.13. Wanneer we de tien subschalen apart onderzochten, vonden we voor één subschaal een significante daling, voor één subschaal een significante stijging, en voor acht subschalen geen significante trend. De gemiddelde trend over de subschalen was -0.07, wat zich binnen het betrouwbaarheidsinterval van de algemene schaal bevond. Voor wiskunde op de internationale schaal in PISA vonden we tussen 2009 en 2018 een significant negatief verschil van -26.01 (ES = -0.34). Voor PISA op de Vlaamse schaal vonden we tussen 2009 en 2018 een significant negatief verschil van -0.20. Deze trends worden getoond in Figuur 4. Er was een significant verschil in de trend tussen het peilingsonderzoek en de internationale schaal van PISA (delta= -0.21, 95% BI [-0.36, -0.06]), en geen significant verschil met de Vlaamse schaal van PISA (delta= -0.12, 95% BI [-0.26, 0.02]).

Figuur 4

Trends algemene schalen wiskunde secundair onderwijs A-stroom



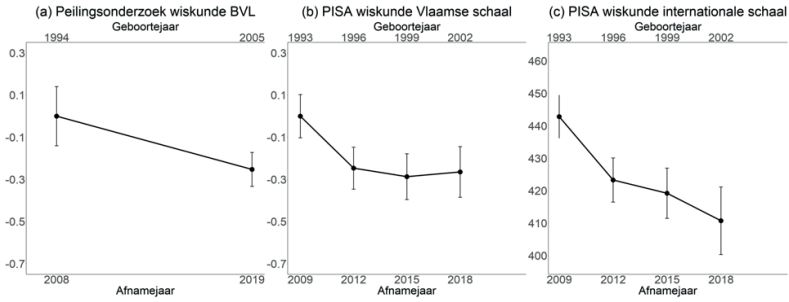
4.4 Wiskunde secundair onderwijs B-stroom

Voor wiskunde in het peilingsonderzoek vonden we tussen 2008 en 2019 een significant negatief verschil van -0.26. Wanneer we de subschalen apart onderzochten, vonden we voor alle vijf subschalen een significante daling, De gemiddelde trend over de subschalen was -0.34, wat zich binnen het betrouwbaarheidsinterval van de algemene schaal bevond. Voor wiskunde op de internationale schaal in PISA vonden we tussen 2009 en 2018 een significant negatief verschil van -32.09 (ES = -0.49). Voor PISA op de Vlaamse schaal vonden we tussen 2009 en 2018 een significant negatief verschil van -0.18. Deze trends worden getoond in Figuur 5. Er was geen significant verschil in de trend tussen het peilingsonderzoek en de internationale schaal van PISA (delta= -0.24, 95% BI [-0.49, 0.01]), en geen significant verschil met de Vlaamse schaal van PISA (delta= -0.01, 95% BI [-0.24, 0.21]).

https://doi.org/10.59302/ps.v100i4.18358

Figuur 5

Trends algemene schalen wiskunde secundair onderwijs B-stroom



5 Discussie

Deze studie richt zich op de vraag wat de trends zijn in gemiddelde leerlingprestaties volgens de peilingsonderzoeken, en of deze gelijk zijn aan de trends volgens TIMSS, PIRLS en PISA. Deze internationale studies beschrijven namelijk trends van vaardigheden zoals ze internationaal begrepen worden, en zijn niet specifiek gericht op de Vlaamse context. Dit wordt vaak als reden aangehaald om hun relevantie voor het Vlaamse onderwijs als beperkt te beschouwen. De peilingsonderzoeken zijn daarentegen wel gericht op de Vlaamse context en eindtermen. Daarnaast vertekenen de internationale IRT-parameters in de internationale studies de trends van individuele onderwijssystemen en wordt aangeraden om IRT-parameters te schatten enkel voor het onderzochte onderwijssysteem (Sachse et al., 2016). Binnen deze studie werden dan ook IRT-parameters geschat enkel op basis van de Vlaamse gegevens.

De peilingsonderzoeken tonen negatieve trends in gemiddelde leerlingprestaties voor lezen en wiskunde in het zesde leerjaar lager onderwijs, en wiskunde in het beroepsvoorbereidend leerjaar. Voor wiskunde op het einde van het tweede leerjaar A is er geen neerwaartse of opwaartse trend. De negatieve trend voor wiskunde in het zesde leerjaar lager onderwijs komt ook tot uiting in een negatieve trend voor zes van de tien subschalen van wiskunde. Slechts één subschaal heeft een positieve trend. Voor wiskunde in het beroepsvoorbereidend leerjaar hebben alle schalen een negatieve trend. Voor wiskunde in het tweede leerjaar A, waar er geen algemene opwaartse of neerwaartse trend is, tonen ook acht van de tien subschalen geen opwaartse noch negatieve trend.

Wanneer we de trends in gemiddelde leerlingprestaties tussen de peilingsonderzoeken en internationale studies vergelijken, dan zien we

een gelijkaardig beeld. Zowel TIMSS als PIRLS tonen dezelfde trends voor respectievelijk wiskunde en lezen in het lager onderwijs als de peilingsonderzoeken. Daarnaast toont ook PISA dezelfde neerwaartse trend voor wiskunde in de B-stroom als de peilingsonderzoeken. Echter, PISA toont een significant neerwaartse trend voor wiskunde in de A-stroom, terwijl de peilingsonderzoeken wel een neerwaartse trend tronen, maar niet-significant. Wanneer we echter de IRT-parameters gebruiken die enkel op Vlaamse gegevens gebaseerd zijn, dan is het verschil tussen de trends volgens PISA en het peilingsonderzoek niet langer significant verschillend. Wanneer we ons enkel richten op de grootteorde van de trends in beide bronnen, dan zien we dat beide bronnen een beperkte neerwaartse trend voor wiskunde vaststellen in de A-stroom.

De consistentie in de trends van de peilingsonderzoeken enerzijds en de internationale studies anderzijds beschouwen we als een indicatie dat beiden, desondanks hun verschillende inhoudelijke en technische kaders, gelijkaardige concepten van onderwijskwaliteit meten. Hoewel we in de inleiding de verschillen in hun opzet benadrukten, zijn de overeenkomsten in de trends niet volledig verbazend. Beiden zijn namelijk ontstaan binnen eenzelfde context, en delen gelijkaardige doelen en veronderstellingen. Zo begonnen de internationale studies van de IEA en de OECD in de jaren 90 als een vorm van kwaliteitsbewaking met de idee dat leerprestaties kwantificeerbaar zijn door geijkte toetsen bij leerlingen af te nemen (Addey, Sellar, Steiner-Khamsi, Lingard, & Verger, 2017; Braun & Singer, 2019). In dezelfde periode werden in Vlaanderen de eindtermen ingevoerd, en kort daarop de peilingsproeven zelf, met ook kwaliteitsbewaking als voornaamste reden (Simons, Kelchtermans, Leysen, & Vandenbroeck, 2016). Ook hier was de onderliggende redenering dat onderwijskwaliteit kwantificeerbaar is door leerprestaties van leerlingen te meten.

Onderzoek in Canada, Duitsland, Engeland en de Verenigde staten toont verder dat de vaardigheden die gemeten worden in internationale studies psychometrisch nauw verwant zijn met hun nationale toetsen (Cartwright, 2012; Ehmke, van den Ham, Sälzer, Heine, & Prenzel, 2020; Hambleton, Sireci, & Smith, 2009; Nissen, Ehmke, Köller, & Duchhardt, 2015; Wagner, Hahn, Schöps, Ihme, & Köller, 2018). Bijgevolg, als toetsen in internationaal vergelijkend onderzoek en nationale studies nauw verwante constructen meten, dan is het niet verbazend dat ze op gelijkaardige trends uitkomen.

Eén opvallend verschil in deze studie is het verschil in de trend voor PISA in de A-stroom op basis van de internationale meetschaal en de meetschaal die enkel op basis van Vlaamse gegevens werd geschat. Ook op basis van de ankeritems die werden afgenomen in 2009 en 2018, vonden we dat de gemiddelde itemscores nauwelijks veranderd zijn. Robitzsch en Lüdtke (2019) merkten al op dat bij PISA de trends van landen op basis van de internationale meetschaal verschillen van een unieke meetschaal per land. Anders dan

https://doi.org/10.59302/ps.v100i4.18358

bij TIMSS en PIRLS, is PISA vanaf 2015 op een volledige digitale afname overgestapt. Hoewel er een statistische correctie plaatsvindt voor enkele items, het toestaan van unieke itemparameters over de afnamemodi bij items met *differential item functioning* (DIF) over de twee afnamemodi, lijkt deze aanpak ontoereikend. Zo werd gevonden dat de digitale afname tot een verlaagde score leidt ten opzichte van de pen-en-papier afname, desondanks de bestaande correctie (Jerrim, Micklewright, Heine, Salzer, & McKeown, 2018; Robitzsch & Lüdtke, 2019). Ter vergelijking, bij TIMSS, dat in 2019 digitaal werd afgenomen in verschillende landen, werd een algemenere correctie toegepast voor de digitale afname door middel van een *linear shift parameter* (Fishbein, Martin, Mullis, & Foy, 2018).

6 Beperkingen

De vergelijking van trends tussen internationaal vergelijkende studies en peilingsonderzoeken gebeurde uiteraard aan de hand van bestaande datasets. De vergelijking was dan ook beperkt tot trends in vaardigheden dewelke gemeten werden in beide soorten studies. We konden hierdoor enkel wiskunde en (begrijpend) lezen vergelijken in het lager onderwijs, en wiskunde in het secundair onderwijs. Op basis van TIMSS kan bijvoorbeeld ook de trend in wetenschappelijke geletterdheid in het lager onderwijs beschreven worden, en bij PISA de trend in lezen en wetenschappelijke geletterdheid in het secundair onderwijs. Bij het peilingsonderzoek zijn er echter geen overeenkomstige herhalingspeilingen om een trend hiervoor te schatten. Omgekeerd, bij het peilingsonderzoek kan ook een trend voor luisteren beschreven wordt in het lager onderwijs, maar deze kan niet geschat worden bij de internationale studies.

Onderzoek op basis van bestaande datasets leidde ook tot andere beperkingen. Zo zijn de peilingsonderzoeken ontworpen rond verschillende subschalen, maar bij TIMSS, PIRLS en PISA wordt de meeste aandacht gegeven aan de algemene vaardigheidsschalen. Door het toetsontwerp van de internationale studies, waarbij elke leerling slechts enkele items per inhoudelijk of cognitief domein ontvangt (Matrix sampling, Mullis & Martin, 2015, pp. 58–66, 2017, pp. 83–91; OECD, 2020, Chapter 9), zijn er vloer- en plafondeffecten wanneer vaardigheidsscores geschat worden voor deze domeinen. Daarnaast is er slechts een beperkt aantal ankeritems over de afnamecycli voor ieder inhoudelijk of cognitief domein. Om deze redenen hebben we voor de subschalen geen nieuwe meetschalen geschat op basis van enkel de Vlaamse data. Een andere beperking betreft het onderzoek naar de samenhang tussen trends en leerlingkenmerken zoals leerlingen hun sociaaleconomische achtergrond. Leerlingkenmerken werden namelijk verschillend gemeten over

de verschillende studies, en soms ook over de afnamecycli binnen een studie, wat een vergelijking ongeschikt maakte. Binnen deze studie hebben we bij de trendschatting dan ook geen onderscheid gemaakt volgens leerlingkenmerken.

7 Conclusie

Bij de recente aandacht voor de dalende leerprestaties in Vlaanderen volgens verschillende internationaal vergelijkende studies, werd vaak de bedenking gemaakt dat deze studies niet specifiek ontwikkeld zijn voor de Vlaamse context en eindtermen. Wanneer we de trends van het Vlaamse peilingsonderzoek echter vergelijken met de trends in internationale studies, dan stellen we vast dat de negatieve trends zich in eenzelfde mate in beide soorten studies voordoen. Voor lezen en wiskunde in het lager onderwijs is er dus een daling in leerprestaties volgens beide bronnen. In het secundair onderwijs is er ook een daling voor de leerlingen in het B-stroom voor leerprestaties in wiskunde. Voor de leerlingen in de A-stroom tonen beide bronnen echter dat de leerprestaties in wiskunde eerder stabiel zijn of licht dalen.

https://doi.org/10.59302/ps.v100i4.18358

Referenties

Addey, C., Sellar, S., Steiner-Khamsi, G., Lingard, B., & Verger, A. (2017). The rise of international large-scale assessments and rationales for participation. *Compare: A Journal of Comparative and International Education, 47*(3), 434–452.

AHOVOKS. (2022a). Onderwijsdoelen modernisering. Retrieved December 19, 2022, from https://www.onderwijsdoelen.be/modernisering

AHOVOKS. (2022b). Peilingen: Wat en waarom? Retrieved December 19, 2022, from https://eind-doelen.be/wat-en-waarom

Ameel, E., Van Nijlen, D., Denis, J., Crynen, M., Van Gorp, K., & Janssen, R. (2017). *Peiling Nederlands (Lezen – Luisteren) in het basisonderwijs 2013 – Brochure*. Leuven: KU Leuven, Steunpunt Toetsontwikkeling en Peilingen.

Braun, H. I., & Singer, J. D. (2019). Assessment for monitoring of education systems: International comparisons. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 75–92.

Byrne, B. M., & de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*(2), 107–132.

Carpentier, N., Costers, S., Janssen, R., & Willem, L. (2019). *Peiling wiskunde in de eerste graad van het secundair onderwijs A-stroom 2018 - Eindrapport*. Leuven: KU Leuven en UAntwerpen, Steunpunt Toetsontwikkeling en Peilingen. Retrieved from www.peilingsonderzoek.be

Carpentier, N., Costers, S., Janssen, R., & Willem, L. (2020). *Peiling wiskunde in de eerste graad van het secundair onderwijs B-stroom 2018 - Eindrapport*. Leuven: KU Leuven en UAntwerpen, Steunpunt Toetsontwikkeling en Peilingen.

Cartwright, F. (2012). *Linking the British Columbia English Examination to the OECD Combined Reading Scale*. Victoria, Canada: British Columbia Ministry of Education.

De Meyer, I., Janssens, R., Warlop, N., Van Keer, H., De Wever, B., & Valcke, M. (2019). *Leesvaardigheid van 15- jarigen in Vlaanderen. Overzicht van de eerste resultaten van PISA2018*. Gent: Universiteit Gent Vakgroep Onderwijskunde. Retrieved from https://www.pisa.ugent.be/index.html

Denis, J., Talloen, W., Laenen, I., Janssen, R., & Aesaert, K. (2019). *Peiling Nederlands in het basisonderwijs 2018 - Brochure*. Leuven: KU Leuven en UAntwerpen, Steunpunt Toetsontwikkeling en Peilingen.

Ehmke, T., van den Ham, A.-K., Sälzer, C., Heine, J., & Prenzel, M. (2020). Measuring mathematics competence in international and national large scale assessments: Linking PISA and the national educational panel study in Germany. *Studies in Educational Evaluation*, *65*, 1–10.

Faddar, J., Appels, L., Merckx, B., Boeve-de Pauw, J., Delrue, K., De Maeyer, S., & Van Petegem, P. (2020). *Vlaanderen in TIMSS 2019. Wiskunde- en wetenschapsprestaties van het vierde leerjaar in internationaal perspectief en doorheen de tijd*. Antwerpen: Universiteit Antwerpen. Retrieved from https://onderwijs.vlaanderen.be/nl/onderzoek/vlaams-en-internationaal-onderwijsonderzoek/internationaal-vergelijkend-onderzoek/trends-in-international-mathematics-and-science-study-timss

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-Scale Assessments in Education*, *6*(1), 1–23.

Foy, P., & Yin, L. (2017). Scaling the PIRLS 2016 achievement data. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 12.1-12.38). Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Retrieved from https://timssandpirls.bc.edu/publications/pirls/2016-methods/P16_MP_Chap12_Scaling_Achievement_Data.pdf

Gielen, S., Van Dessel, K., De Meyst, M., Beringhs, S., Crynen, M., Luyten, B., & Janssen, R. (2010). *Peiling wiskunde in de eerste graad van het secundair onderwijs (A-stroom) 2009 -Eindrapport.* Leuven: KU Leuven Centrum voor Onderwijseffectiviteiten -evaluatie.

Gielen, S., Willem, L., Beringhs, S., Luyten, B., & Janssen, R. (2009). *Peiling wiskunde in de eerste graad van het secundair onderwijs B-stroom 2008 –Eindrapport.* Leuven: KU Leuven, Centrum voor Onderwijseffectiviteit en –evaluatie.

Glas, C., & Jehangir, K. (2013). Modeling Country-Specic Differential Item Functioning. In *Handbook of International large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 97–115). New York City, New York: Chapman and Hall/CRC.

Hambleton, R. K., Sireci, S. G., & Smith, Z. R. (2009). How do other countries measure up to the mathematics achievement levels on the national assessment of educational progress? *Applied Measurement in Education*, *22*(4), 376–393.

Hambleton, R. K., & Zenisky, A. L. (2011). Translating and Adapting Tests for Cross-Cultural Assessments. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-Cultural Research Methods in Psychology,* (pp. 46–74). New York City, NY: Cambridge University Press.

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect'and what has been done about it? *Oxford Review of Education*, *44*(4), 476–493.

Kane, M. T. (2017). Loosening psychometric constraints on educational assessments. *Assessment in Education: Principles, Policy & Practice*, *24*(3), 447–453.

Kish, L. (1965). *Survey sampling*. New York City, NY: John Wiley & Sons.

Martin, M. O., & Mullis, I. V. S. (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Boston, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P., Brossman, B., & Stanco, G. M. (2012). Estimating linking error in PIRLS. IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments,. In *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* (pp. 35–47).

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and Procedures in TIMSS 2015*. Boston, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://timssandpirls.bc.edu/publications/timss/2015-methods.html

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2017). *Methods and Procedures in PIRLS 2016*. Boston, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *PIRLS 2006 Technical Report*. Boston,

MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://timssandpirls.bc.edu/pirls2006/tech_rpt.html
Martin, M. O., von Davier, M., & Mullis, I. V. S. (2020). *Methods and Procedures: TIMSS 2019 Technical Report*. Boston, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://timssandpirls.bc.edu/timss2019/methods
Mitzel, H. C., Lewis, D., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ, NJ: Lawrence Erlbaum.
Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, *8*(3), 323–335.
Mullis, I. V. S., & Martin, M. O. (2015). *PIRLS 2016 Assessment Framework*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education & IEA.
Mullis, I. V. S., & Martin, M. O. (2017). *TIMSS 2019 Assessment Frameworks*. Boston, MA, MA: TIMSS & PIRLS International Study Center, Boston College.
Mullis, I. V. S., Martin, M. O., Foy, P., Hooper, M., & IEA. (2017). *PIRLS 2016 international results in reading.* Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education & IEA.
Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. Boston, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://timss2019.org/reports/download-center/
Muthén, B., & Muthén, L. (2017). *Mplus User's Guide. Eighth Edition.* Los Angeles, CA, CA: Muthén & Muthén.
Nissen, A., Ehmke, T., Köller, O., & Duchhardt, C. (2015). Comparing apples with oranges? An approach to link TIMSS and the National Educational Panel Study in Germany via equipercentile and IRT methods. *Studies in Educational Evaluation*, *47*, 58–67.
OECD. (2011). *School Evaluation in the Flemish Community of Belgium, OECD Reviews of Evaluation and Assessment in Education*. Paris, France: OECD Publishing.
OECD. (2012). *PISA 2009 Technical Report*. Paris, France: OECD. Retrieved from https://www.oecd.org/pisa/pisaproducts/50036771.pdf
OECD. (2014). *PISA 2012 Technical Report*. Paris, France: OECD. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf
OECD. (2017). *PISA 2015 Technical Report*. Paris, France: OECD Publishing. Retrieved from https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf
OECD. (2019a). *PISA 2018 Assessment and Analytical Framework.* Paris, France, France: OECD Publishing.
OECD. (2019b). *PISA 2018 Results (Volume I): What Students Know and Can Do, PISA*. Paris, France: OECD Publishing.
OECD. (2020). *PISA 2018 Technical Report (Forthcoming).* Paris, France: OECD. Retrieved from https://www.oecd.org/pisa/data/pisa2018technicalreport/
Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696.
Robitzsch, A., & Lüdtke, O. (2019). Linking errors in international large-scale assessments:

https://doi.org/10.59302/ps.v100i4.18358

Calculation of standard errors for trend estimation. *Assessment in Education: Principles, Policy \& Practice*, *26*(4), 444–465.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York City, NY: Wiley.
Sachse, K. A., & Haag, N. (2017). Standard errors for national trends in international large-scale assessments in the case of cross-national differential item functioning. *Applied Measurement in Education*, *30*(2), 102–116.
Sachse, K. A., Roppelt, A., & Haag, N. (2016). A comparison of linking methods for estimating national trends in international comparative large-scale assessments in the presence of cross-national DIF. *Journal of Educational Measurement*, *53*(2), 152–171.
Simons, M., Kelchtermans, G., Leysen, J., & Vandenbroeck, M. (2016). *Eindtermen in actie - Werking, doeltreffendheid en toekomst van een beleidsinstrument in Vlaanderen*. Brussel: Departement Onderwijs en Vorming. Retrieved from https://publicaties.vlaanderen.be/view-file/20164
Spikic, S., Goos, M., Denis, J., Costers, S., Janssen, R., & van Renterghem, K. (2022). *Peiling wiskunde in het basisonderwijs 2021 - Eindrapport*. Leuven: KU Leuven & UAntwerpen, Steunpunt Toetsontwikkeling en Peilingen.
STEP. (2021). Website STEP. Retrieved January 30, 2021, from https://peilingsonderzoek.be/en/
Tielemans, K., Vandenbroeck, M., Bellens, K., Van Damme, J., & De Fraine, B. (2017). *Het Vlaams lager onderwijs in PIRLS 2016. Begrijpend lezen in internationaal perspectief en in vergelijking met 2006*. Leuven: Centrum voor Onderwijseffectiviteit en –evaluatie. Retrieved from https://onderwijs.vlaanderen.be/nl/progress-in-international-reading-literacy-study-pirls
Valverde, G. A., Bianchi, L. J., Wolfe, R. G., Schmidt, W. H., & Houang, R. T. (2002). *According to the book: Using TIMSS to investigate the translation of policy intro practice through the world of textbooks*. Dordrecht, the Netherlands: Kluwer Academic Publishers.
Van Nijlen, D., Denis, J., Willem, L., Ameel, E., & Janssen, R. (2017). *Peiling wiskunde in het basisonderwijs 2016 - Eindrapport*. Leuven: KU Leuven, Centrum voor Onderwijseffectiviteit en -evaluatie.
von Davier, Matthias, Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Vol. 2. Issues and methodologies in large scale assessments* (Vol. 2, pp. 9–36). Princeton, NJ: IEA-ETS Research Institute. Retrieved from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf
Wagner, H., Hahn, I., Schöps, K., Ihme, J. M., & Köller, O. (2018). Are the tests scores of the Programme for International Student Assessment (PISA) and the National Educational Panel Study (NEPS) science tests comparable? An assessment of test equivalence in German Schools. *Studies in Educational Evaluation*, *59*, 278–287.
Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15–27.

Auteurs

Jonas Dockx is als postdoctoraal onderzoeker van het FWO werkzaam aan de KU Leuven.
Lore Pelgrims is wetenschappelijk medewerker aan de KU Leuven.
Koen Aesaert is tenure track docent aan de KU Leuven.
Rianne Janssen is gewoon hoogleraar aan de KU Leuven.

Correspondentieadres: dr. Jonas Dockx, Centrum voor Onderwijseffectiviteit- en evaluatie; Adres: Onderwijseffectiviteit en –evaluatie, Dekenstraat 2 – bus 3773, 3000 Leuven, België. E-mail: jonas.dockx@kuleuven.be

Abstract

Do Flemish surveys and international studies show the same trends in student performance?

Flanders (Northern Belgium) has been confronted with declining student performance in international comparative studies. However, these studies aim to assess abilities across different countries and cultures. Accordingly, it is unknown whether a trend for a certain ability in an international study is equal to the trend of that ability as it is understood in the Flemish context. However, between 2002 en 2022, Flanders also organized national assessments that were meant to assess how many students reached the Flemish attainment targets. In this study, we examined the trends for reading and mathematics according to the national assessments and compared them with trends in international studies. Additionally, we checked the robustness of the trends in international comparative studies by estimating new item parameters based on Flemish data only. The results showed that the trends in the national assessment and international studies are similar when corresponding abilities are measured.

Keywords international large-scale assessments, national assessments, mathematics, reading comprehension, item response theory (IRT)