Filip Monté 0592823

Multidisciplinair Instituut Lerarenopleiding

Hoe de complexiteit van wiskunde-items in de tweede graad secundair onderwijs meten?

Een kwantitatieve en kwalitatieve analyse.

Proefschrift ingediend met het oog op het behalen van de graad van
Educatieve Master Wetenschappen en Technologie

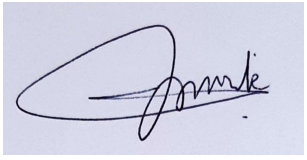
Promotor: Prof. Dr. Bart Windels

Academiejaar: 2022-2023

*Ik verklaar plechtig dat ik de masterproef, "**Hoe de complexiteit van wiskunde-items in de tweede graad secundair onderwijs meten? Een kwantitatieve en kwalitatieve analyse.**", zelf heb geschreven.*

Ik ben op de hoogte van de regels i.v.m. plagiaat en heb erop toegezien om deze toe te passen in deze masterproef.

01 JUNI 2023 *Filip Monté*



Dankwoord

In dit dankwoord wil ik me richten tot enkele mensen die me hebben gesteund tijdens het voltooien van deze masterproef en de studie in zijn geheel.

Ik dank mijn vriendin Sara, en mijn vrolijke dochter Babette (en actieve zoon Maurice), voor de niet aflatende steun in deze studieperiode. Zonder hen was dit niet mogelijk geweest. Ik dank mijn promotor Prof. Dr. Bart Windels voor de hulp en het advies bij deze masterproef, maar ook voor het enthousiasme en het bijbrengen van de vakdidactiek Wiskunde. Ook wil ik alle docenten van de opleiding bedanken voor de inspirerende lessen, de coaching en de reflectiesessies. Tijdens de voorbije twee jaar heb ik ontzettend veel bijgeleerd en heb ik me de overgang naar het onderwijs op geen enkel moment beklaagd.

Vanzelfsprekend dank ik ook mijn (schoon)familie, vrienden en collega's voor de hulp en het luisterend oor dat ze hebben geboden. In het bijzonder vermeld ik Arne De Boeck, Paul Van Cappellen, Didier Deses en Joeri Emmers voor het technisch ondersteunen en het mee nadenken over dit academisch werkstuk.

Specifiek voor deze masterproef wil ik ook de leerkrachten bedanken die de tijd hebben genomen om aan het onderzoek mee te werken. Ondanks hun beperkte tijd hebben zij toch de moeite gedaan om aan mijn onderzoek bij te dragen.

Tot slot dank ik ook de Vlaamse Wiskunde Olympiade vzw voor het beschikbaar stellen van de resultaten van de wiskundeolympiade.

Abstract

Deze masterproef onderzoekt hoe de complexiteit van wiskunde-items voor de tweede graad secundair onderwijs te voorspellen is. Van den Bossche (2022) is nagegaan hoe een complexiteitsschema kon worden gecreëerd en verbeterd. Hier ligt de focus mede op het evalueren van de interbeoordelaarsbetrouwbaarheid, wat door Van den Bossche (2022) werd gesuggereerd.

Op basis van bestaande literatuur en de specifieke eindtermen wiskunde voor de tweede graad secundair onderwijs werd een schema opgesteld dat leraren gebruiken om de complexiteit van wiskunde-items in te schatten. Na een predictieve validatie van het complexiteitsschema vullen wiskundeprofessionals het complexiteitsschema zelfstandig in voor vijf geselecteerde wiskunde-items. De Krippendorff's alpha coëfficiënt meet nadien de interbeoordelaarsbetrouwbaarheid om de onderlinge overeenstemming tussen de beoordelaars na te gaan. Ook wordt de individuele correlatie voor elke beoordelaar gemeten.

De resultaten tonen aan dat het meten van de complexiteit aan de hand van het complexiteitsschema door verschillende beoordelaars een sterke variabiliteit vertoont. Het complexiteitsschema draagt bij om een inschatting over de complexiteit van wiskunde-items te maken maar wanneer beoordelaars zelfstandig de complexiteit beoordelen is de overeenstemming zwak en dus het complexiteitsschema weinig robuust voor een individuele inschatting van de complexiteit. Wanneer de ondersteuning van de beoordelaars ontbreekt, is de correlatie tussen reële en voorspelde complexiteit laag.

De beoordelaars, die het complexiteitsschema invulden, erkennen dat een complexiteitsschema nuttig is en tonen interesse in de uitkomst van de meting. Maar de beoordelaars geven ook aan dat het complexiteitsschema niet toepasbaar is op bijvoorbeeld het evalueren van bewijzen wat ook een onderdeel is van de leerstof in de tweede graad secundair onderwijs. Dit vormt zeker een suggestie naar het vervolgonderzoek waarbij de onderzoeker dan extra aandacht schenkt aan de selectie van de vragen. Ook verdient de selectie en ondersteuning van de beoordelaars in een vervolgonderzoek voldoende zorg om een goede voorspelling van de complexiteit toe te laten.

Inhoudsopgave

1. Inleiding

Karagiannakis et al. (2014) stellen dat "wiskunde wordt gezien als een complex vak dat bestaat uit verschillende domeinen zoals rekenen, probleemoplossing, meetkunde, algebra, waarschijnlijkheidsleer, statistiek". Tegelijk geven Karagiannakis et al. (2014) aan dat wiskunde een verscheidenheid aan basisvaardigheden vereist die verband houden met een gevoel voor kwantiteit, symbolencodering, geheugen en logica vereisen. Cragg & Gilmore (2014) zeggen dat "veel factoren bijdragen aan verschillen in wiskundige bekwaamheid waaronder attitude, werkgeheugen, motivatie, taalvaardigheid en IQ, naast sociale en educatieve factoren". Volgens Karagiannakis et al. (2014) zijn er vele wiskundige moeilijkheden zoals het uitvoeren van wiskundige procedures, het visualiseren van ruimtefiguren of het oplossen van wiskundige problemen. Daarnaast zijn ook andere processen betrokken zoals het onderdrukken van afleidende informatie en ongewenste reacties, of het wisselen tussen verschillende taken (Cragg & Gilmore, 2014). De verzamelnaam wordt "executieve functie" genoemd en Cragg & Gilmore (2014) hebben een model uitgewerkt dat de link tussen executieve functies en wiskundige bekwaamheid aantoont.

Volgens Joskin (2022) is onderwijs van belang voor de ontwikkeling van de samenleving en een sleutelelement voor onze economie. Toch wijzen internationale onderzoeken in de richting van een dalend niveau in de kennis van wiskunde (Joskin, 2022). Op 15 december 2022 sprak Rianne Janssen (KU Leuven) van een code rood voor ons onderwijs gebaseerd op recente peilingsresultaten voor de eindtermen wiskunde afgenomen door de KU Leuven en de UAntwerpen (Vancaeneghem, 2022). Vancaeneghem (2022) schrijft dat meetkunde en getallenleer toch niet zo eenvoudig zijn voor leerlingen van het tweede middelbaar. Ook de basisgeletterdheid, die beschouwd wordt als het minimum om te functioneren in onze maatschappij, staat onder druk. Dit wordt zorgwekkend genoemd.

Internationaal bestaan er verschillende onderzoeken die de wiskundige geletterdheid meten. Een voorbeeld is *the programme for the International Assessment of Adult Competencies* (PIAAC), een onderzoek dat de wiskundige geletterdheid van 16- tot 65-jarigen monitort (Tout, 2020). Tout (2020) beschrijft dat de term "wiskundige geletterdheid" evolueerde tot het raadplegen, gebruiken en kritisch redeneren met wiskundige inhoud, informatie en ideeën die op verschillende manieren worden voorgesteld, om te gebruiken in en toe te passen voor wiskundige uitdagingen van de dagelijkse realiteit. Hiermee wordt ingespeeld op de "21st century skills" en professionele verwachtingen alsook geeft het meer aandacht aan cognitieve processen die horen bij het oplossen van wiskundige uitdagingen (Tout, 2020). Tout (2020)

beschrijft deze processen als (1) situaties wiskundig benaderen en evalueren, (2) handelen naar en gebruiken van wiskunde en (3) evalueren, kritisch reflecteren en beoordelen en ook de wijze van weergeven aanpassen (Tout, 2020). Het scoringssysteem uit ALL (Adult Literacy and Lifeskills), een comparatieve beoordeling van de wiskundige geletterdheid bij volwassenen, wordt verder gebruikt als de manier om de moeilijkheid van items te begrijpen (Tout, 2020; Gal et al., 2005).

Behalve de PIAAC onderzoeken, met de toepassing van complexiteit zoals hierboven beschreven, zijn er ook de PISA onderzoeken waarbij de wiskundige kennis van 15-jarigen wordt gemeten aan de hand van zes variabelen (Turner, 2012). Een eerste variabele is "communicatie" wat betekent de interpretatie van de informatie, die moeilijker is bij complexe vragen en wordt onderverdeeld in vier niveaus (Turner, 2012). Een tweede variabele gaat over het selecteren van de juiste oplossingsstrategie die vanzelfsprekend moeilijker is wanneer verschillende fasen met intermediaire checks worden gebruikt. De derde variabele is het mathematiseren van informatie om die te gebruiken voor de oplossing en bij de vierde variabele speelt de correcte interpretatie een rol die samenhangt met hoe de informatie is weergegeven (Turner, 2012). De vijfde variabele verwijst naar wiskundetaal en wiskundevaardigheden en -procedures. De zesde variabele wordt gebruikt wanneer argumentatie en redenering vereist zijn (Turner, 2012). Al deze variabelen zijn ook te classificeren van minder tot meer complex.

Hoewel er over de kennis van wiskunde veel data beschikbaar zijn, is er over het inschatten van de specifieke complexiteit van wiskunde-items, voor een specifieke graad, tot op heden weinig bekend. Leraren schatten de complexiteit van hun evaluatie doorgaans eerder intuïtief in. Toch heeft het de kracht om een hulpmiddel te zijn dat op een meer objectieve manier toelaat wiskunde te evalueren zodat de leerkrachten de complexiteit van vragen beoordelen en dus hen beter helpt in te schatten hoe complex een wiskundetoets wel is. Om wiskunderesultaten goed te interpreteren, is het van belang om als toets- of vraagontwerper te beseffen waarom een leerling een specifiek wiskunde-item complexer vindt dan een andere. In deze masterproef bekijken we dit vraagstuk vanuit de wiskundige (voor)kennis van de leerling, op basis van de afgevinkte leerplandoelen en bouwen we een complexiteitsschema om de complexiteit van een toetsvraag te bepalen aan de hand van specifieke wiskundige componenten. Door de complexiteit van een toetsvraag te kennen, is de wiskundetoets beter af te stemmen op wat daadwerkelijk wordt geëvalueerd en wordt de interpretatie van de resultaten naar een hoger niveau getild want op basis van de complexiteitsmeting valt het beter in te schatten op welk niveau de leerlingen presteren. Samengevat: De kennis van de complexiteit van een evaluatie

leidt tot een hogere effectiviteit van die evaluatie (Tout et al., 2020). Tout et al. (2020) geeft aan dat “de toepassing van een complexiteitsschema op een evaluatie de test meer betrouwbaar en valide maakt”.

Vanuit deze probleemstelling is het doel van de masterproef tweeledig: Ten eerste is nagegaan hoe de complexiteit van een wiskunde-item voor de tweede graad secundair onderwijs wordt voorspeld. Een eerste onderzoeksvraag luidt als volgt: Hoe kan men de complexiteit van een wiskunde-item voor de tweede graad secundair onderwijs meten? Ten tweede wordt nagegaan dat, wanneer die complexiteit gemeten is, elke leerkracht onafhankelijk van elkaar dezelfde voorspelde complexiteit uitkomt. Een te lage mate van overeenstemming tussen verschillende leerkrachten (onvoldoende interbeoordelaarsbetrouwbaarheid) leidt tot uiteenlopende inschattingen van de complexiteit, en brengt de voorspellende waarde (predictieve validiteit) van het schema ernstig in het gedrang. Een tweede onderzoeksvraag klinkt: Hoe is de overeenstemming tussen verschillende beoordelaars die gebruik maken van het complexiteitsschema?

Deze masterproef is onderverdeeld in vier delen. Deel één is een literatuurstudie waarin rekenvaardigheid, evalueren, meerkeuzevragen, complexiteit meten, een complexiteitsschema opstellen en interbeoordelaarsbetrouwbaarheid worden gekaderd. Deel twee is de beschrijving van de methodologie. Deel drie bevat een beschrijving van de resultaten en in deel vier worden de discussie en conclusies weergegeven.

2. Literatuurstudie

Binnen de sectie van de literatuurstudie worden verschillende zaken besproken die met complexiteit en evaluatie verband houden. Eerst wordt de historiek van het meten van wiskundige geletterdheid kort besproken. Daarna wordt de term 'evaluatie' bekeken om nadien over te gaan op de items van de wiskundeolympiade. Vervolgens wordt nagegaan hoe de complexiteit te meten is en wordt een mogelijk complexiteitsschema voorgesteld. Het laatste onderdeel van de literatuurstudie focust op theorie over de inschatting van de betrouwbaarheid tussen de beoordelaars die het schema invullen.

Hoe rekenvaardigheid meten?

Een grootschalige beoordeling van de rekenvaardigheid werd met het *International Adult Literacy* onderzoek (IALS) in de jaren 1990 uitgevoerd. Dat werd voortgezet met het *Adult Literacy and Lifeskills (ALL)* onderzoek in het midden van de jaren 2000, en in 2011 leidde het finaal tot het *Program for the International Assessment of Adult Competencies (PIAAC)* (Tout, 2020). Wiskundige geletterdheid is een onderdeel van de Adult Literacy and Lifeskills (ALL) studies en het wordt gezien als één van de belangrijke factoren om te bepalen of een populatie zich aanpast en functioneert in de informatiemaatschappij of op het werk (Gal et al., 2005). Volgens Tout (2020) worden met de onderzoeken bewijzen verzameld over de vaardigheden van jongeren en volwassenen en laat het landen begrijpen (1) hoe onderwijs- en opleidingsstelsels deze vaardigheden versterken en (2) hoe dergelijke vaardigheden verdeeld zijn binnen de bevolking. De ontwikkeling van de beoordelingsvragen gebeurt nauwgezet om de kwaliteit en geldigheid van de vragen te waarborgen en het verloopt binnen een statistisch kader die uitgebreide psychometrie en analyse met *item response theory* (IRT) omvatten (Tout, 2020).

Wat is een evaluatie en hoe gebeurt het?

In de educatieve context zorgt men er steeds voor dat een evaluatie adequaat en geloofwaardig is en dat de uitkomst te gebruiken is. Als dat niet het geval is, dan wordt er gesproken van een misevaluatie, eventueel zelfs gelinkt aan de bekwaamheid van de evaluator (Alkin & King, 2017). Alkin en King (2017) stellen dat de interpretatie wordt uitgevoerd door de evaluator en dat die leidt tot een beslissing waarbij wordt aangenomen dat de evaluatiedata, om de beslissing te maken, bruikbaar zijn.

Volgens Dixson & Worrell (2016) is een evaluatie summatief of formatief. Formatieve evaluatie legt de focus op een tussentijdse evaluatie of de leerling het heeft begrepen en of het lesgeven

het doel heeft bereikt (Dixson & Worrell, 2016). Summatieve evaluatie legt de nadruk op het evalueren van de prestatie van een leerling nadat de leeractiviteit heeft plaatsgevonden (Dixson & Worrell, 2016). Een formatieve evaluatie binnen de wiskunde houdt in dat data over de kennis van wiskunde wordt verzameld en vervolgens wordt die kennis gebruikt voor het vormgeven van de volgende instructie-activiteiten (Lyon et al., 2018). Waar formatieve evaluatie bedoeld is om de instructie te evalueren en te verbeteren is summatieve evaluatie eerder retrospectief en gebaseerd op informatie verzameld na een instructie (Dixson & Worrell, 2016; Thurber et al., 2002). De resultaten van een evaluatie zijn dus op verschillende manieren te interpreteren. Het evalueren wordt bijvoorbeeld gebruikt om na een leerproces punten te geven. Men spreekt van "assessment of learning", geïnterpreteerd als "assessment for learning", waarbij de evaluaties valide, transparant en betrouwbaar zijn (Afdeling Onderwijskwaliteitszorg Universiteit Gent, 2022). De inhoud dient afgestemd te zijn op de eindcompetenties van het vak, de cijfers geven correct weer wat het werkelijke beheersingsniveau van de bevraagde is en de evaluatie is transparant naar vorm, spelregels, inhoudelijke verwachtingen en het moment (Afdeling Onderwijskwaliteitszorg Universiteit Gent, 2022).

Wat wiskunde betreft zijn er mogelijk meerdere competenties te evalueren met een evaluatie. Er zijn verschillende competenties binnen wiskunde zoals de competentie om problemen op te lossen, om te redeneren, om procedures toe te passen, om iets abstract wiskundig weer te geven, om elementen te linken of om te communiceren over wiskunde (Boesen et al., 2014). Een evaluatie is bedoeld, of tracht, om alle wiskundige competenties te evalueren en dus niet enkel diegene die eenvoudig te evalueren zijn (Nortvedt et al., 2018). Niss (2002) stelt zelfs dat geen enkel beoordelingsformulier en -instrument voldoende is om het geheel aan competenties betrouwbaar en geldig te beoordelen.

Evalueren is voor verschillende doeleinden te gebruiken, maar doet ook vragen rijzen over welke competentie men evalueert (Nortvedt et al., 2018). Het evaluatieproces bestaat volgens Nortvedt et al. (2018) uit vier sub-processen: (1) welke inhoud evalueren, (2) ontwikkeling van items, (3) uitvoeren van de meting en (4) het duiden van de resultaten. Men stelt dat er enerzijds operationele evaluaties zijn en anderzijds toepassingsgerichte evaluaties. Operationeel evaluaties zijn gericht op het kennen van concepten, strategieën en feiten. Toepassingsgerichte evaluaties zijn gefocust op het gebruik en begrijpen van wiskundige concepten om problemen op te lossen (Thurber et al., 2002).

De Junior Wiskunde Olympiade: Meerkeuzevragen

Tijdens een Junior Wiskunde Olympiade (“JWO”) roepen de leerlingen hun kennis van wiskunde op. In essentie valt dit te vergelijken met het oproepen van voorkennis bij de start van een nieuwe wiskundeles. De JWO maakt gebruikt van meerkeuzevragen. Meerkeuzevragen worden geregeld gebruikt om voorkennis op te wekken (Dochy et al., 1999). Algemeen wordt aangenomen dat de voorkennis van een leerling 'redelijk volledig en correct' is, dus van een redelijke hoeveelheid is, concreet en aanwezig is en goed gestructureerd is (Dochy et al., 1999). Dochy et al. (1999) ontkrachten deze aanname en stippen aan dat leerlingen misconcepties hebben en ongestructureerde kennis bezitten waardoor het leerproces van de leerlingen mogelijk gehinderd wordt. Als er geen voorkennis is, dan kan er ook geen kenniskader zijn, wat op zijn beurt problemen geeft bij het bekrachtigen en beoordelen van nieuwe informatie (Dochy et al., 1999). In deze masterproef is het gebruik van de voorkennis toegepast op vragen uit de JWO die kennis oproepen bij de leerlingen. Voorkennis beïnvloedt de prestaties en dus mogelijk de studieresultaten van de leerlingen (Dochy et al., 1999). Domeinspecifieke kennis is een doorslaggevende voorwaarde om goede wiskundige prestaties te produceren (Schneider et al., 1989). Voorkennis faciliteert de prestatie in positieve zin (Dochy et al., 1999). Ook hoe een wiskunde-item wordt weergegeven, heeft een invloed op de score van een test, omdat ze op een andere manier wordt gesteld dan de leerling gewoon is of omdat ze verschillend wordt aangeleerd (Turner, 2010).

Meerkeuzevragen worden frequent toegepast in educatieve evaluaties omdat het als een effectieve manier van testen wordt beschouwd en het maakt een onmiddellijke meting van vaardigheden, kennis en bekwaamheden mogelijk (Gierl et al., 2017). Deze vragen bestaan uit een stam, verschillende opties en eventueel bijkomende informatie. De stam is duidelijk en nauwkeurig en bevat context, inhoud en vragen die de leerling dient te beantwoorden (Considine et al., 2005). De opties bevatten zowel het goede antwoord als afleiders (Gierl et al., 2017). Een afleider is plausibel en vergelijkbaar met het correcte antwoord maar incorrect. Afleiders evalueren op hun effectiviteit en kwaliteit voor het differentiëren van studenten is aan te raden (Gierl et al., 2017). Zimmaro (2004) stelt voor om per item drie tot vijf opties beschikbaar te stellen. Het ontwikkelen van functionele afleiders via bijvoorbeeld vrije antwoorden-evaluaties dient overwogen te worden waarbij incorrecte antwoorden dan worden gebruikt als afleiders (Ali et al., 2016). Volgens Ali et al. (2016) wordt een zinvolle afleider gekozen door vijf procent of meer van de ondervraagden en wordt die ook meer gekozen door zij die laag presteren dan door zij die hoog presteren. Goed werkende afleiders dragen bij aan de validiteit en de determinerende werking van de test, stellen Ali et al. (2016). De afleiders

hebben een impact op de uitkomst van de test (Ali et al., 2016). Regels voor afleiders waarover eenstemmigheid bestaat zijn het plaatsen van afleiders in een logische en numerieke orde, het gebruik van plausibele afleiders, het vermijden van zinnen zoals “geen enkele van de bovenstaande” of “elk van de bovenstaande”, het gebruik van onafhankelijke afleiders zonder overlap en het vermijden tips te geven naar het juiste antwoord (Gierl et al., 2017).

Hoe de complexiteit meten en waarmee rekening houden?

Zoals eerder aangegeven zegt Tout (2020) dat rekenvaardigheid, de kennis en vaardigheden omvat om effectief wiskundige vragen te beantwoorden en te behandelen. Binnen ALL worden 5 componenten benoemd: (1) de vorm van uitkomst/probleem transparantie, (2) de plausibiliteit van afleiders, (3) de complexiteit van informatie/data, (4) het soort van handeling/vaardigheid en (5) het aantal verwachte stappen (Kirsch & Mosenthal, 1990; Tout et al., 2020). Gal et al. (2005) stellen dat hierbij geen rekening wordt gehouden met de karakteristieken en eigenschappen van de persoon die het item oplost, dat er ook geen rekening wordt gehouden met de onderlinge connectie tussen componenten die een specifieke vraag nog complexer maken, en dat de score dus enkel een schatting is.

Behalve aandacht voor wiskundeactiviteiten is er zeker ook aandacht voor tekstuele aspecten waarbij de moeilijkheidsgraad voor tekst wordt bepaald door de aard van de opdracht (d.w.z. de relatie tussen de tekst en de opdracht die hier opgelost wordt), de structuur en complexiteit van de tekst en de aard van de processen of strategieën die de informatie in de vraag linken aan informatie in de tekst (Tout et al., 2020).

Kirsch en Mosenthal (1990) schrijven dat naast de wiskundige kennis ook de leesvaardigheid bijdraagt aan de complexiteit en dat er variabelen zijn die bepalen of een vraag correct beantwoord wordt. Als er een duidelijke link is tussen wat er gedaan dient te worden en de informatie in een vraag, dan is dit als een makkelijkere vraag te bekijken in vergelijking met een vraag waar de lezer zelf bepaalt wat de kritische informatie is (Kirsch & Mosenthal, 1990).

Het type informatie in de vraag speelt een rol alsook de plausibiliteit van de afleiders (Kirsch & Mosenthal, 1990). Als de plausibiliteit van de afleider hoog is, dan wordt het moeilijker om het correcte antwoord te selecteren (Kirsch, 2001). Volgens Kirsch (2001) speelt documentgeletterdheid ook een rol in de complexiteit en wordt ze daarom ook gebruikt in de constructie van een complexiteitsschema. Een wiskundige prestatie zal dus niet enkel afhangen van de wiskundekennis zelf maar zeker ook van geletterdheid zoals leesstrategieën, woordenschat of begrijpend lezen (*PIAAC Numeracy*, 2009). Voor wat betreft het wiskundige

aspect is het aanbevolen om de link te maken met de leerplandoelstellingen van de leerlingen (Tout et al., 2020).

Wat de specifieke wiskundige complexiteit betreft, spelen volgende zaken een rol: de probleemtransparantie, wat ook te beschrijven valt als de moeilijkheid om het probleem te vertalen, de wiskundeconcepten, het aantal vaardigheden die gebruikt worden en de moeilijkheidsgraad van de vaardigheden (Tout et al., 2020). Sommige vaardigheden zijn moeilijker dan andere en langer of minder lang aangeleerd in functie van het moment dat deze werden gezien, namelijk in de eerste graad of de tweede graad secundair onderwijs (Tout et al., 2020). Als het gaat over specifieke wiskunde-informatie bestaat er meer complexe alsook minder complexe informatie voor de weergave van een getal (bijvoorbeeld breuk versus decimaal) wat zijn effect heeft op de complexiteit van het wiskunde-item (Tout et al., 2020). Verschillende handelingen hebben een andere complexiteit evenals het aantal stappen waarbij een verschil wordt gemaakt tussen het combineren van verschillende stappen en het herhalen van dezelfde stappen (Tout et al., 2020).

Wiskunde en problemen oplossen zijn sterk aan elkaar gelinkte concepten waarbij het probleemoplossend werken dient gezien te worden als het belangrijkste aspect van de wiskunde (Güner & Erbay, 2021). In de jaren 80 was er een didactische stroming die meende dat probleemoplossend vermogen het zwaartepunt van wiskunde op school moest zijn (Schoenfeld, 2016). Uit een onderzoek blijkt volgens Schoenfeld (2016) zelfs dat het wordt gezien als ideale voorbereiding op een olympiade. De redenering omdraaien betekent dat probleemoplossend vermogen bijbrengen het wiskundeonderwijs verantwoordt, zodat dan nieuwe vaardigheden worden ontwikkeld (Schoenfeld, 2016).

Wiskunde-instructie is er op gericht om de leerlingen de kans te geven om een breed gamma aan probleemsituaties en problemen te onderzoeken, variërend van oefeningen tot open probleemstellingen (Schoenfeld, 2016). Het oplossen van problemen vereist zowel cognitieve processen, bijvoorbeeld evaluatie van parameters, als metacognitieve processen, zoals reflectie over vooruitgang (OECD, 2021). Probleemoplossend denken is aan te leren en over wiskunde wordt geleerd aan de hand van probleemoplossend denken (Liljedahl et al., 2016).

Volgens Güner & Erbay (2021) is problemen oplossen een ingewikkelde activiteit die inhoudt dat men denkt op een hoog niveau. Het oplossen van problemen vereist de toepassing van een aantal wiskundige principes en kennis over levensechte, niet-routinematige en open problemen, die op hun beurt helpen bij het leren van wiskunde (Güner & Erbay, 2021). Met bekende

berekeningen lost men routinematige opdrachten op, terwijl niet-routinematige problemen meer uitdagende denkvaardigheden vereisen alsook het toepassen van heuristieken om problemen op te lossen (Güner & Erbay, 2021). Al deze vaardigheden worden enkel bereikt wanneer wiskundige basisvaardigheden en -concepten beschikbaar zijn (Schoenfeld, 2016).

Binnen het probleemoplossend werken zijn er verschillende stappen, stellen Güner & Erbay (2021): het probleem begrijpen, plannen, het plan uitvoeren en het probleem oplossen. Er zijn ook verschillende oplossingsstrategieën. Leerlingen (1) die een diepgaand begrip hebben van wiskundige concepten, (2) die bedreven zijn in probleemoplossende strategieën en technieken, (3) die een positieve houding en geloof ten opzichte van wiskunde en probleemoplossend denken hebben en (4) die het vermogen bezitten om de juiste beslissingen te nemen, zijn leerlingen die succesvol problemen oplossen (Güner & Erbay, 2021). Er zijn verschillende probleemoplossende vaardigheden en een leerling heeft een voorkeur voor vaardigheden die hij of zij eerder gebruikt heeft (Güner & Erbay, 2021).

Er werd ook aangetoond dat de leerkracht een impact heeft op de probleemoplossende vaardigheden van de leerlingen. De strategieën die leraren aan leerlingen aanleren, hangen af van de voorkeur van de leraren (Güner & Erbay, 2021). Wat, zoals eerder gezegd, ook van belang is voor het probleemoplossend vermogen, is de voorkennis van een leerling (Liljedahl et al., 2016). Dit zorgt voor een beter begrip van het probleem en het kiezen van een geschikte strategie om dat probleem op te lossen. Liljedahl et al. (2016) stellen dat het een wisselwerking is van terugkijken en connecties maken met al verworven kennis. Het is afhankelijk van het feit of de leerling het probleem heeft begrepen en daardoor de juiste connecties maakt of potentieel maakt (Liljedahl et al., 2016).

Een laatste onderdeel van wiskundige kennis hier besproken, is redeneren, de zogenaamde wiskundige redeneervaardigheden, zoals beschreven door Siregar et al. (2020). Volgende aspecten zullen deze vaardigheden versterken: hypothesen opstellen, het proces van patronen herkennen, argumenten onderbouwen en conclusies evalueren (Siregar et al., 2020). Steen (1999) stelt dat redeneervaardigheden ontzettend nuttig zijn voor het oplossen van problemen en dat die mogen gezien worden gezien als de basis van wiskunde.

De conclusie luidt dat, gelet op de variatie in nuttige vaardigheden, het een heuse uitdaging is om de complexiteit van wiskunde te meten (Tout et al., 2020).

Complexiteitsschema opstellen

Om de complexiteit in te schatten, wordt er een kader uitgewerkt dat toelaat om in te schatten welke activiteit nu moeilijker is dan een andere activiteit (Gal et al., 2005). Dat kader bepaalt welke vaardigheden worden geëvalueerd, met een focus op de specifieke bekwaamheid voor rekenvaardigheid (Tout, 2020). Binnen dit kader varieert de volledige complexiteitsscore tussen laagste moeilijkheidsscore en hoogste moeilijkheidsscore waarbij elke component een specifieke bijdrage levert aan de totale score (Gal et al., 2005).

Aan alle componenten wordt een mogelijk complexiteitsscore-bereik toegekend dat dan uiteindelijk leidt tot de finale complexiteitsscore op basis van alle gescoorde componenten (Tout et al., 2020). Van den Bossche (2022) heeft deze complexiteitsscore, met een schaal 5 tot 19, vervolgens omgezet in een geschatte moeilijkheidsgraad (M.G.), zie Formule (1), per item.

$M.G. = 1 - \frac{x-5}{19-5}$	Formule (1)

Deze M.G. werd vervolgens gebruikt om de score van een item te schatten wetende dat de maximale score per item vijf is (Van den Bossche, 2022).

Betrouwbaarheid tussen beoordelaars meten

Eenmaal het complexiteitsschema is opgesteld, is het mogelijk om na te gaan wat de overeenstemming tussen de verschillende beoordelaars is, om zo te onderzoeken hoe de complexiteitsinschattingen zich onderling verhouden (Stemler, 2004). Dit bepaalt de betrouwbaarheid van het complexiteitsschema. Stemler (2004) stelt dat een zekere subjectiviteit altijd aanwezig is en daarom dient de betrouwbaarheid adequaat te worden ingeschat.

Gwet (2021) beschrijft dat interbeoordelaarsbetrouwbaarheid talrijke toepassingen heeft voor de evaluatie van inhoud en het meest effectieve ontwerp van een studie is, wanneer elke beoordelaar elk item scoort voor wat betreft het minimaliseren van de standaardfout van de overeenkomstcoëfficiënt. Interbeoordelaarsbetrouwbaarheid zal iets zeggen over de grootte van variabiliteit bij de beoordeling - van bijvoorbeeld complexiteit - tussen verschillende beoordelaars (Landis & Koch, 1977). De interbeoordelaarsbetrouwbaarheid verwijst naar de betrouwbaarheid die wordt beïnvloed door de beoordelaar (Gwet, 2021). Er dient rekening te worden gehouden met zowel de geobserveerde overeenstemming tussen beoordelaars alsook met de op toeval gebaseerde overeenstemming om een inschatting van de betrouwbaarheid te maken (McHugh, 2012). Een lage interbeoordelaarsbetrouwbaarheid geeft aan dat er weinig

overeenstemming is tussen verschillende beoordelaars wanneer vragen naar complexiteit worden beoordeeld (Joly & Hofmans, 2023).

Er wordt in deze masterproef gesproken van validiteit en betrouwbaarheid waarbij validiteit verwijst naar de mate waarin de voorspelde complexiteit overeenstemt met de reële complexiteit en waarbij betrouwbaarheid verwijst naar de mate waarin verschillende beoordelaars overeenstemmen in hun inschatting van de complexiteit (Zapf et al., 2016). Een lage interbeoordelaarsbetrouwbaarheid voor de verschillende categorieën van het complexiteitsschema zal aanleiding geven tot een grote variabiliteit in de validiteit.

3. Data en methode

De methode binnen deze masterproef is opgebouwd volgens verschillende stappen die logisch op elkaar volgen.

Bij de eerste stap werd een bestaand PIAAC-schema voor bepaling van wiskundige geletterdheid gebruikt om de complexiteit van 50 items van verschillende jaargangen van de JWO te meten (Gal et al., 2005; Tout et al., 2020). De Vlaamse Wiskunde Olympiade vzw stemde in om de resultaten van voorgaande jaargangen ter beschikking te stellen. De jaargangen 2017-2018, 2018-2019 en 2019-2020 werden gebruikt. Via die data kon de exacte moeilijkheidsgraad worden berekend op basis van de antwoorden van de deelnemende leerlingen, en worden gecorreleerd aan de berekende complexiteit op basis van het schema om de predictieve validiteit te bepalen. De gebruikte items staan in Tabel 1.

Tabel 1. Selectie van de JWO items voor validatie van het complexiteitsschema.

Jaargang	**Items**
2016-2017	1 – 2 – 3 – 4 – 5 – 11 – 12 – 13 – 14 – 15
2017-2018	1 – 2 – 3 – 4 – 5 – 6 – 7 – 8 – 9 – 10
2018-2019	1 – 2 – 3 – 4 – 5 – 6 – 7 – 8 – 9 – 10
2019-2022	11 – 12 – 13 – 14 – 15 – 16 – 17 – 18 – 19 – 20 – 21 – 22 – 23 – 24 – 25 – 26 – 27 – 28 – 29 – 30

Bij de tweede stap werd het PIAAC-schema verfijnd om het beter te linken aan de wiskundekennis van leerlingen van de tweede graad secundair onderwijs. Deze wiskundekennis voor de tweede graad is terug te vinden op de website https://onderwijsdoelen.be/ op basis van de eindtermen. Aan de hand van een correlatieberekening, in Excel, gebeurde de predictieve validatie waarbij de correlatie tussen de echte moeilijkheidsgraad (op basis van de antwoorden) en de voorspelde moeilijkheidsgraad (op basis van het schema) werd bepaald. Wanneer het schema geschikt was bevonden, op basis van de correlatie, werd overgegaan naar de volgende stap.

Bij de derde stap werd de interbeoordelaarsbetrouwbaarheid van het PIAAC-schema berekend. Hiervoor werd het gecreëerde complexiteitsschema, zowel het originele schema op basis van PIAAC als het verfijnde op basis van PIAAC en de eindtermen wiskunde voor de tweede graad secundair onderwijs, door verschillende beoordelaars ingevuld. Hiervoor werden vijf vragen uit de JWO dataset geselecteerd (Tabel 2).

Tabel 2. Selectie van de JWO items voor evaluatie door beoordelaars.

Jaargang	**Items**
2017-2018	2 –5
2018-2019	1
2019-2022	11 –15

De vragen voor beoordeling van de complexiteit werden willekeurig geselecteerd uit verschillende JWO datasets. Vijftien leraren, waaronder één leraar in opleiding uit de Educatieve Master van de VUB, vulden het schema in. Het criterium voor selectie van de beoordelaars was dat alle beoordelaars ervaring hebben met het doceren en evalueren van wiskunde aan de tweede of derde graad secundair onderwijs. De leerkrachten werden gerekruteerd uit het eigen netwerk van de onderzoeker. Om het complexiteitsschema toe te passen ontvingen zij (1) de geselecteerde vragen met bijhorend antwoord, (2) een begeleidend document hoe het schema diende ingevuld te worden en (3) het in te vullen complexiteitsschema (Zie bijlages 1, 2 en 3). De beoordelaars vulden het schema zelfstandig in. Door het schema in te vullen bepaalden de respondenten voor elk wiskunde-item individueel de complexiteit. Deze waarden werden dan gebruikt om de interbeoordelaarsbetrouwbaarheid te bepalen.

De interbeoordelaarsbetrouwbaarheid werd berekend via Krippendorff's alpha volgens onderstaande Formule (2) en beoordeeld volgens Tabel 3 (Landis & Koch, 1977; Krippendorff, 2011). D_o stelt de geobserveerde onenigheid voor en D_e stelt de verwachte onenigheid voor (Krippendorff, 2011).

$\alpha = 1 - \frac{D_o}{D_e}$	Formule (2)

Krippendorff's alpha laat toe om de interbeoordelaarsbetrouwbaarheid te berekenen voor om het even welke steekproefgrootte van beoordelaars en te beoordelen vragen, en voor om het even welk meetniveau van gegevens (Hayes & Krippendorf, 2007; Zapf, 2016). Bovendien heeft Krippendorff's alpha geen probleem met ontbrekende waarden. Krippendorff's alpha neemt doorgaans een waarde aan tussen 0 en 1, waarbij waarden dichter bij 0 aangeven dat er weinig overeenstemming is tussen de verschillende beoordelaars en waarden dichter bij 1 aangeven dat er een grote mate van overeenstemming is tussen de verschillende beoordelaars.

In extreme gevallen is Krippendorff's alpha een negatieve waarde, wat wijst op systematische vorm van onenigheid tussen de verschillende beoordelaars. Volgens Krippendorff (2018) ligt de minimaal aanvaardbare ondergrens op 0,667. Waarden daaronder geven aan dat de interbeoordelaarsbetrouwbaarheid onvoldoende is. Volgens Zapf et al. (2016) is het mogelijk de Tabel 3 voor de Krippendorff's alpha toe te passen waardoor de interpretatie zeer eenduidig is.

De Krippendorff's alpha coëfficiënten en de bijhorende 95% bootstrapped betrouwbaarheidsintervallen werden berekend met behulp van het icr-pakket in het statistische programma R, zoals beschreven in Hayes & Krippendorff (2007). Omdat de beoordelingsschalen voor de complexiteit een ordinaal meetniveau hebben, werd de optie "metric = ordinal" gespecifieerd in de krippalpha functie (Krippendorff, 2011). Om de 95% bootstrapped betrouwbaarheidsintervallen te berekenen werd de methode beschreven door Krippendorff (2016) toegepast waarbij men gebruikt maakte van 20000 verschillende bootstrap steekproeven.

Aanvullend wordt naast de Krippendorff's alpha coëfficiënt voor elke beoordelaar afzonderlijk ook de correlatie tussen de reële complexiteit en de voorspelde complexiteit berekend via Excel. Zo worden verschillen in de voorspellende kracht van de test (predictieve validiteit) tussen verschillende beoordelaars nagegaan.

Tabel 3. Categorieën voor de evaluatie van de sterkte van overeenkomst (Landis & Koch, 1977; Zapf et al., 2016).

Krippendorff's alpha coëfficiënt	**Sterkte van overeenkomst**
< 0,00	Zwak
0,00 – 0,20	Licht
0,21 – 0,40	Redelijk
0,41 – 0,60	Gematigd
0,61 – 0,80	Substantieel
0,81 – 1,00	Bijna perfect

Bij de vierde en laatste stap wordt gezocht naar mogelijke verklaringen voor de verschillen in de interbeoordelaarsbetrouwbaarheid. De deelnemers werd gevraagd om moeilijkheden of opmerkingen te noteren tijdens het invullen, die vervolgens in het interview worden besproken. Hiervoor worden semigestructureerde interviews (bijlage 7) georganiseerd. De deelnemende

leraren beantwoorden enkele vooropgestelde vragen, en tegelijk is er ook ruimte om vrij te reageren op de toepasbaarheid van het complexiteitsschema. Dit type interview laat dat toe.

4. Resultaten

De eerste stap van het onderzoek is het opstellen van een complexiteitsschema aan de hand van de correlatie tussen de reële complexiteit, op basis van de JWO resultaten, en de voorspelde complexiteit. Voor het originele complexiteitsschema zoals voorgesteld door Tout et al. (2020) wordt een correlatie tussen de reële en voorspelde complexiteit gevonden van 0,91. Voor het gereviseerde complexiteitsschema op basis van de eindtermen van de tweede graad wordt een correlatie gevonden van 0,93.

De tweede stap is de evaluatie van het complexiteitsschema door verschillende beoordelaars. In totaal zijn vijftien beoordelaars bereid gevonden om het complexiteitsschema in te vullen. Elke beoordelaar heeft ervaring met wiskunde en met lesgeven aan de tweede en derde graad secundair onderwijs. Na het invullen van het complexiteitsschema wordt de interbeoordelaarsbetrouwbaarheid bepaald aan de hand van de Krippendorff's alpha coëfficiënt. Tabel 4 en Tabel 5 geven de resultaten van de coëfficiënt voor elke subcategorie van het complexiteitsschema weer. Er is ook telkens het 95% betrouwbaarheidsinterval toegevoegd alsook de sterkte van overeenstemming volgens Tabel 3. Bijlage 4 bevat de code voor de berekening in R, bijlagen 5 en 6 bevatten de data voor de berekening.

Tabel 4: Resultaten Krippendorff's alpha coëfficiënt per subcategorie van de complexiteitsanalyse (origineel complexiteitsschema).

Subcategorie	Krippendorff's alpha coëfficiënt	95% betrouwbaarheidsinterval	Sterkte van overeenstemming
#1	0,124	0,034 – 0,211	Licht
#2	0,158	0,067 – 0,248	Licht
#3	0,542	0,490 – 0,594	Gematigd
#4	0,254	0,156 – 0,350	Redelijk
#5	0,422	0,323 – 0,517	Gematigd
#6	0,368	0,287 – 0,445	Redelijk
#7	0,084	- 0,043 – 0,206	Licht
#8	0,198	0,034 – 0,349	Licht
#9	- 0,017	- 0,126 – 0,089	Zwak
#10	0,496	0,420 – 0,568	Gematigd
#11	0,192	0,075 – 0,306	Licht
#12	0,058	- 0,040 – 0,152	Licht

#13	0,436	0,324 – 0,538	Gematigd
#14	0,207	0,093 – 0,315	Redelijk
#15	0,064	- 0,032 – 0,158	Licht

Tabel 5: Resultaten Krippendorff's alpha coëfficiënt per subcategorie van de complexiteitsanalyse (gereviseerd complexiteitsschema).

Subcategorie	**Krippendorff's alpha coëfficiënt**	**95% betrouwbaarheidsinterval**	**Sterkte van overeenstemming**
#1	0,082	- 0,012 – 0,174	Licht
#2	0,015	- 0,084 – 0,114	Licht
#3	0,521	0,470 – 0,572	Gematigd
#4	0,301	0,204 – 0,395	Redelijk
#5	0,527	0,444 – 0,608	Gematigd
#6	0,319	0,226 – 0,408	Redelijk
#7	0,110	- 0,023 – 0,237	Licht
#8	0,203	0,029 – 0,367	Licht
#9	0,113	0,012 – 0,210	Licht
#10	0,395	0,305 – 0,482	Redelijk
#11	0,319	0,198 – 0,434	Redelijk
#12	0,252	0,169 – 0,333	Redelijk
#13	0,337	0,216 – 0,454	Redelijk
#14	0,086	- 0,040 – 0,207	Licht
#15	0,214	0,123 – 0,299	Redelijk

Uit tabellen 4 en 5 blijkt duidelijk dat de interbeoordelaarsbetrouwbaarheid sterk varieert en voor de meeste subcategorieën als licht tot redelijk is te beschouwen. Er is met andere woorden weinig overeenstemming tussen de beoordelaars onderling over de verschillende categorieën van het complexiteitsschema. Voor elke subcategorie is het 95% betrouwbaarheidsinterval als smal te interpreteren wat erop wijst dat de schatting van de Krippendorff's alpha coëfficiënt vrij precies is (met andere woorden: in andere steekproeven zullen doorgaans gelijkaardige waarden worden bekomen). Wanneer het originele en het gereviseerde schema worden vergeleken, zijn er enkel kleine verschillen merkbaar. Het valt niet te besluiten dat er meer overeenstemming is in het originele dan in het gereviseerde complexiteitsschema of omgekeerd. De wijzigingen zijn

samengevat in Tabel 6. In de meeste gevallen is er geen wijziging, in enkele gevallen wordt één klasse gestegen of gedaald. Bij de revisie is een zwakke sterkte van overeenstemming niet langer aanwezig.

Tabel 6: Vergelijking van Krippendorff's alpha coëfficiënten per subcategorie tussen de originele en gereviseerde complexiteitsanalyse.

Subcategorie	**Krippendorff's alpha coëfficiënt (origineel)**	**Krippendorff's alpha coëfficiënt (revisie)**	**Wijziging klasse**
#1	0,124	0,082	–
#2	0,158	0,015	–
#3	0,542	0,521	–
#4	0,254	0,301	–
#5	0,422	0,527	–
#6	0,368	0,319	–
#7	0,084	0,110	–
#8	0,198	0,203	–
#9	- 0,017	0,113	Zwak naar Licht
#10	0,496	0,395	Gematigd naar Redelijk
#11	0,192	0,319	Licht naar Redelijk
#12	0,058	0,252	Licht naar Redelijk
#13	0,436	0,337	Gematigd naar Redelijk
#14	0,207	0,086	Redelijk naar Licht
#15	0,064	0,214	Licht naar Redelijk

In de derde stap van het onderzoek wordt vervolgens ook de correlatie, tussen de reële gemiddelde score en de inschatting van de complexiteit via het complexiteitsschema, voor elke beoordelaar afzonderlijk geëvalueerd. Uit Tabel 7 blijkt dat de correlatie voor de meeste beoordelaars ook zwak is in vergelijking met de initiële correlatie om de schema's op te stellen (zie predictieve validatie). Waar de initiële predictieve validatie hoog is, is de predictieve validatie voor de meeste beoordelaars laag en in sommige gevallen zelfs negatief. Dit betekent dat het complexiteitsschema onvoldoende robuust is om een goede complexiteitsbeoordeling

toe te laten. De grote verschillen in predictieve validatie liggen in lijn met de lage interbeoordelaarsbetrouwbaarheid die eerder is bepaald.

Alle resultaten wijzen op een zwakke correlatie tussen de reële complexiteit en de berekende complexiteit via het complexiteitsschema. Dit is het geval voor het originele schema en voor het gereviseerde schema.

Tabel 7: Resultaten predictieve validatie aan de hand van correlatiecoëfficiënt tussen voorspelde complexiteit en reële complexiteit.

Beoordelaar	Correlatie (origineel)	Correlatie (revisie)
#13 Els	-0,225	**-0,3063**
#12 Kris	0,1817	**-0,2264**
#11 Nadine	-0,1048	**-0,0490**
#3 Didier	-0,1157	**-0,0474**
#4 Paul	-0,5229	**-0,035**
#7 Ivan	0,0782	0,0062
#5 Ann	0,1829	0,1093
#14 Jenthe	0,2110	0,1664
#6 Riet	0,0891	0,2155
#9 Isabel	-0,2836	0,2226
#2 Piet	0,2984	**0,2533**
#10 Ingeborg	0,0305	**0,3348**
#1 Johannes	0,2862	**0,3743**
#8 Joeri	0,4913	**0,5735**
#15 Joris	0,8098	**0,8098**

Om een verklaring te vinden voor de resultaten werden semigestructureerde interviews afgenomen met de beoordelaars. In totaal hebben dertien beoordelaars de tijd genomen voor een interview. De meeste beoordelaars geven aan dat het invullen van het schema vlotter loopt voor de revisie dan voor het origineel omdat het meer gealigneerd is met de eindtermen van de tweede graad. Het originele schema is uitsluitend gebaseerd op het PIAAC-schema. De revisie is uitgebreid met de specifieke eindtermen voor de tweede graad secundair onderwijs. Uit de resultaten komt daarentegen niet naar voren dat het gereviseerde schema een betere correlatie of interbeoordelaarsbetrouwbaarheid vertoont. Het blijkt bijvoorbeeld dat de eerste

subcategorie van het originele en gereviseerde complexiteitsschema niet noodzakelijk dezelfde inschatting heeft voor beide schema's hoewel er geen wijziging is in de subcategorie-beschrijving en de vragen. Dit toont aan dat andere factoren meespelen bij het inschatten van de complexiteit door verschillende beoordelaars. Het valt te besluiten dat extra begeleiding tijdens het invullen, het zorgvuldig selecteren van beoordelaars, het type vragen en de achtergrond van de beoordelaar een impact hebben op de complexiteitsinschatting.

Wat het type vragen betreft werd in de interviews een aantal keren de opmerking gemaakt dat de selectie van JWO vragen een interessante doch eenzijdige keuze is. De JWO vragen zijn meerkeuzevragen, van nature ingewikkeld en voornamelijk gericht op probleemoplossend denken. Het zijn geen vragen die peilen naar specifieke theoretische kennis zoals bewijzen of feitelijke kennis zoals het geven van een definitie. Dit wordt aangegeven als één van de redenen waardoor leerkrachten het minder toepasbaar zien in de dagelijkse lespraktijk en dus volgens de onderzoeker leidt tot minder affiniteit met de geselecteerde vragen. Ook gaf één beoordelaar mee dat zij ervaring had met enkele van de JWO vragen (gebruikt in de eigen lespraktijk) waardoor de complexiteitsinschatting mogelijk afwijkend is in vergelijking met beoordelaars die geen voorafgaande ervaring hebben met de geselecteerde vragen.

Een andere bedenking bij het type vragen, wordt gemaakt over de oplossingsstrategie. Volgens enkele beoordelaars zijn er meerdere oplossingsstrategieën voor de geselecteerde JWO vragen wat toelaat dat twee beoordelaars ten opzichte van elkaar een andere inschatting hebben over bijvoorbeeld het aantal stappen nodig om een wiskunde-item op te lossen. Dit is potentieel een reden voor de lage betrouwbaarheid van subcategorie vijftien (verwachte aantal bewerkingen/stappen) wat tegelijk de bedenking wekt of het bij inschatting van de complexiteit dient toegestaan te zijn om een opmerking toe te voegen waarom een bepaalde score wordt gekozen.

Met betrekking tot de impact van de achtergrond van leerkrachten komt in de interviews naar voren dat bijvoorbeeld het taalperspectief een rol speelt. Beoordelaars met professionele ervaring in een (taal)diverse leeromgeving geven mee dat zij wat betreft het tekstuele aspect een verschillende inschatting maken omdat zij zich hiervoor verplaatsen in de geest van hun leerlingen. Uit subcategorie 1, de tekstuele categorie bij uitstek, blijkt een lichte overeenstemming wat in lijn ligt met het bovenstaande.

5. Discussie en conclusie

Het doel van de masterproef was, ten eerste, om een complexiteitsschema te ontwerpen dat leraren gebruiken om de complexiteit van wiskunde-items voor de tweede graad secundair onderwijs te meten en, ten tweede, om de interbeoordelaarsbetrouwbaarheid van dit schema te evalueren.

De algemene conclusie van dit onderzoek beantwoordt de onderzoeksvragen die worden gesteld (zie pagina 8 voor de onderzoeksvragen). De complexiteit van een wiskunde-item is te meten met het opgestelde complexiteitsschema maar de overeenstemming tussen verschillende beoordelaars, die gebruik maken van het complexiteitsschema, is laag. Het onderzochte complexiteitsschema is onvoldoende betrouwbaar voor zelfstandig gebruik door leraren in hun dagelijkse lespraktijk op school. De kans is (te) groot dat verschillende leerkrachten tot verschillende beoordelingen zullen komen. Deze overeenstemming tussen beoordelaars is te verhogen mits de beoordelaar(s) voldoende getraind is (zijn), enthousiast en gemotiveerd is (zijn) om het schema accuraat in te vullen. De correlatie tussen de reële complexiteit en de voorspelde complexiteit is bij de meeste leerkrachten zwak. Wat Van den Bossche (2022) al concludeerde wordt hier ontegensprekelijk bevestigd: het evalueren van de complexiteit is een tijdrovende activiteit en uit dit onderzoek blijkt nu ook dat de tijdsbesteding een impact heeft op de inschatting van de complexiteit. Gezien de toepassing van het schema in de lespraktijk als eerder onpraktisch wordt beschouwd omwille van de complexe invulling van het schema zelf, is de keuze van de beoordelaars in een vervolgonderzoek belangrijk en zorgvuldig uit te voeren. Het selecteren van wiskundedeskundigen uit de Vlaamse Wiskunde Olympiade vzw of het team dat werkt aan de Centrale Toetsen zijn te overwegen. Wanneer men voldoende getraind/ervaren is met het type vragen en men het nut van het schema voldoende erkent, dan zal een accuratere inschatting van de complexiteit mogelijk zijn. Het inschatten van de complexiteit van wiskunde-items is complex op zichzelf zoals Tout et al. (2020) het stellen.

Voortbouwend op Van den Bossche (2022) wordt in dit onderzoek gestart met een PIAAC-schema, toegepast op JWO vragen voor de tweede graad secundair onderwijs. De predictieve validiteit van het schema, voor de initiële beoordelaar, is zeer hoog (correlatie van 0,93). Dit is mede te verklaren door het feit dat deze beoordelaar, ook de auteur van deze masterproef, het complexiteitsschema opstelde en sterk vertrouwd is met de vragen door de evaluatie van in totaal 50 vragen. Deze beoordelaar is als zeer getraind te beschouwen voor het invullen van het schema.

Gezien de initieel hoge correlatie mag dus verwacht worden dat dit schema geschikt is om de complexiteit te meten wanneer het wordt ingevuld door andere beoordelaars. Deze conclusie wordt mede versterkt doordat dit schema, zij het licht verschillend en bij een andere doelgroep, ook door Van den Bossche (2022) succesvol werd toegepast.

De vraag is echter of de complexiteitsschema's even valide zijn wanneer ze gebruikt worden door 'doorsnee' leerkrachten zonder ervaring in het beoordelen van complexiteit van wiskunde-items. Een belangrijke voorwaarde voor de validiteit van een beoordelingsinstrument is de interbeoordelaarsbetrouwbaarheid of de mate waarin verschillende beoordelaars op basis van hetzelfde instrument tot gelijkaardige inschattingen komen (Gwet, 2021; Joly & Hofmans, 2023; Stemler, 2004).

Uit de resultaten van de interbeoordelaarsbetrouwbaarheid (Krippendorff, 2011; Krippendorff, 2018) blijkt weinig overeenstemming tussen de complexiteitsbeoordelingen van de verschillende beoordelaars, beoordeeld volgens het schema van Landis en Koch (1977). Dit toont aan dat het bepalen van de complexiteit van wiskunde-items een uitdaging vormt zoals Tout et al. (2020) in hun werk stellen.

Beoordelaars hebben zelfstandig het schema ingevuld zonder begeleiding van de opsteller van het schema. Hierdoor is het aannemelijk dat per subcategorie heel makkelijk een verschillende inschatting wordt gemaakt. Ook het feit dat de beoordelaars werden geselecteerd op basis van een zeer algemene oproep, en dus niet noodzakelijk persoonlijk gelinkt zijn met de onderzoeker, maakt dat het niet vanzelfsprekend is dat beoordelaars honderd procent gemotiveerd zijn om het schema zeer accuraat in te vullen. Een voorbeeld hiervan is de eerste subcategorie van het originele en gereviseerde complexiteitsschema waarbij de inschatting voor verschillende beoordelaars niet noodzakelijk gelijk is voor beide schema's hoewel de subcategorie-beschrijving niet is gewijzigd. Anderzijds geeft dit zeer nuttige info die suggereert dat de inschatting verschilt volgens het moment van de dag wat leidt tot de conclusie dat het complexiteitsschema onvoldoende robuust is om een overeenkomstige complexiteitsinschatting door verschillende beoordelaars toe te laten.

De lage interbeoordelaarsbetrouwbaarheid vertaalt zich in grote verschillen op het vlak van predictieve validiteit. De correlatie tussen de reële complexiteit en de voorspelde complexiteit varieert sterk voor de beoordelaars onderling (Joly & Hofmans, 2023). Een aantal leraren slaagt erin om een goede inschatting te maken en behaalt een hoge correlatie, terwijl andere leraren inschattingen maken die sterk verschillen van de realiteit met een lage correlatie tot gevolg.

De keuze van de JWO vragen, om complexiteit te voorspellen, wordt door de beoordelaars als te eenzijdig gepercipieerd en staat haaks op het feit dat wiskunde uit zeer verschillende domeinen bestaat (Cragg & Gilmore, 2014; Karagiannakis et al., 2014). Er is bijvoorbeeld onvoldoende aandacht voor het peilen naar specifieke theoretische kennis. Volgens de beoordelaars leidt dat tot een lagere toepasbaarheid of bruikbaarheid van het complexiteitsschema in de dagelijkse lespraktijk omdat de leraren zich minder inleven in de keuze van de vragen.

Extra begeleiding tijdens het invullen, het zorgvuldig selecteren van beoordelaars, het type vragen en de achtergrond van de beoordelaar hebben waarschijnlijk een impact hebben op de complexiteitsinschatting. Het is dan ook een suggestie om deze factoren in een vervolgonderzoek te analyseren. Het aantal te evalueren categorieën binnen één onderzoek verminderen of de vragen ter beoordeling aanpassen aan de werkelijke lespraktijk verhoogt zeer waarschijnlijk de intrinsieke interesse van een beoordelaar. Hier moet wel zorgvuldig mee omgesprongen worden gezien verschillende onderzoekers stellen dat de complexiteit door zeer veel factoren worden bepaald (Gal et al., 2005; Kirsch & Mosenthal, 1990; Tout et al., 2020). Het inschatten van complexiteit blijft complex.

De taal blijkt ook een belangrijk aspect en is afhankelijk van de achtergrond van de leerkracht. Deze categorie, subcategorie 1, geeft dit aan en moet in een vervolgonderzoek verder onderzocht worden. Zowel Tout et al. (2020) als Kirsch & Mosenthal (1990) onderschrijven deze conclusie met hun onderzoek.

Het toevoegen van de mogelijkheid tot het maken van een geschreven opmerking bij de keuze van een score brengt meer helderheid over de inschatting en is dus een belangrijke suggestie naar een vervolgonderzoek. Het zorgt voor verrijking van het onderzoek.

6. Literatuurlijst

Afdeling Onderwijskwaliteitszorg Universiteit Gent. (2022). *Evalueren: Hoe doe je dat kwaliteitsvol.* Gent: Universiteit Gent. Opgehaald op 8 maart 2023, van https://onderwijstips.ugent.be/nl/tips/evalueren-hoe-doe-je-dat-kwaliteitsvol/

Ali, S. H., Carr, P. A., & Ruit, K. G. (2016). Validity and Reliability of Scores Obtained on Multiple-Choice Questions: Why Functioning Distractors Matter. *Journal of the Scholarship of Teaching and Learning*, *16*(1), 1–14. https://doi.org/10.14434/josotl.v16i1.19106

Alkin, M. C., & King, J. A. (2017). Definitions of Evaluation Use and Misuse, Evaluation Influence, and Factors Affecting Use. *American Journal of Evaluation*, *38*(3), 434–450. https://doi.org/10.1177/1098214017717015

Boesen, J., Helenius, O., Bergqvist, E., Bergqvist, T., Lithner, J., Palm, T., & Palmberg, B. (2014). Developing mathematical competence: From the intended to the enacted curriculum. *The Journal of Mathematical Behavior*, *33*, 72–87. https://doi.org/10.1016/j.jmathb.2013.10.001

Considine, J., Botti, M., & Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, *12*(1), 19–24. https://doi.org/10.1016/S1322-7696(08)60478-3

Cragg, L., & Gilmore, C. (2014). Skills underlying mathematics: The role of executive function in the development of mathematics proficiency. *Trends in Neuroscience and Education*, *3*(2), 63–68. https://doi.org/10.1016/j.tine.2013.12.001

Dixson, D. D., & Worrell, F. C. (2016). Formative and Summative Assessment in the Classroom. *Theory Into Practice*, *55*(2), 153–159. https://doi.org/10.1080/00405841.2016.1148989

Dochy, F., Segers, M., & Buehl, M. M. (1999). The Relation Between Assessment Practices and Outcomes of Studies: The Case of Research on Prior Knowledge. *Review of Educational Research*, *69*(2), 145–186. https://doi.org/10.3102/00346543069002145

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619

Gal, I., Van Groenestijn, M., Manly, M., Schmitt, M. J., & Tout, D. (2005). Adult numeracy and its assessment in the ALL survey: A conceptual framework and pilot results. *Measuring adult literacy and life skills: New frameworks for assessment,* 137-191.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research*, *87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Güner, P., & Erbay, H. N. (2021). Prospective mathematics teachers' thinking styles and problem-solving skills. *Thinking Skills and Creativity*, *40*, 100827. https://doi.org/10.1016/j.tsc.2021.100827

Gwet, K. L., & Gwet, K. L. (2021). *Analysis of categorical ratings* (Fifth edition). AgreeStat Analytics.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures, 1*(1), 77-89. https://doi.org/10.1080/19312450709336664

Joly, J., & Hofmans, J. (2023). How reliable are personality judgments by political experts? The curious case of Donald Trump. *Personality Science, 4*, e6715. https://doi.org/10.5964/ps.6715

Joskin, A. (2022). *Daling van het onderwijsniveau: de verborgen kosten van de covid-19-pandemie.* Brussel: Federaal Planbureau.

https://www.plan.be/uploaded/documents/202206070754390.PUB_ART_012_EDUC_12646_N.pdf

Karagiannakis, G., Baccaglini-Frank, A., & Papadatos, Y. (2014). Mathematical learning difficulties subtypes classification. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00057

Kirsch, I. S. (2001). The framework used in developing and Interpreting the International Adult Literacy Survey (IALS). *European Journal of Psychology of Education*, *16*(3), 335–361. https://doi.org/10.1007/BF03173187

Kirsch, I. S., & Mosenthal, P. B. (1990). Exploring Document Literacy: Variables Underlying the Performance of Young Adults. *Reading Research Quarterly*, *25*(1), 5. https://doi.org/10.2307/747985

Krippendorff, K. (2011) *Computing Krippendorff's Alpha Reliability*. Philadelphia, PA: Universiteit van Pennsylvania. Opgehaald op 8 maart 2023, van https://repository.upenn.edu/asc_papers/43.

Krippendorff, K. (2016) *Bootstrapping Distributions for Krippendorff's Alpha*. Philadelphia, PA: Universiteit van Pennsylvania. Opgehaald op 8 maart 2023 van https://www.asc.upenn.edu/sites/default/files/2021-03/Algorithm%20for%20Bootstrapping%20a%20Distribution%20of%20Alpha.pdf.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Liljedahl, P., Santos-Trigo, M., Malaspina, U., & Bruder, R. (2016). *Problem Solving in Mathematics Education*. Springer International Publishing. https://doi.org/10.1007/978-3-319-40730-2

Lyon, C. J., Wylie, E. C., Brockway, D., & Mavronikolas, E. (2018). Formative assessment and the role of teachers' content area. *School Science and Mathematics, 118*(5), 144–155. https://doi.org/10.1111/ssm.12277

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica, 22*(3), 276-282.

Niss, Mogens (2002). *Mathematical competencies and the leaning of mathematics: The Danish KOM project.* Roskilde: Roskilde University. Opgehaald op 8 maart 2023, van https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b7b50cfc513371b27ce0b90d4dc19e45b5c7828e

Nortvedt, G. A., & Buchholtz, N. (2018). Assessment in mathematics education: Responding to issues regarding methodology, policy, and equity. *ZDM, 50*(4), 555–570. https://doi.org/10.1007/s11858-018-0963-z

OECD. (2021). *The Assessment Frameworks for Cycle 2 of the Programme for the International Assessment of Adult Competencies*. OECD. https://doi.org/10.1787/4bc2342d-en

PIAAC Numeracy: A Conceptual Framework (OECD Education Working Papers Nr. 35; OECD Education Working Papers, Vol. 35). (2009). https://doi.org/10.1787/220337421165

Schneider, W., Körkel, J., & Weinert, F. E. (1989). Domain-specific knowledge and memory performance: A comparison of high- and low-aptitude children. *Journal of Educational Psychology*, *81*(3), 306–312. https://doi.org/10.1037/0022-0663.81.3.306

Schoenfeld, A. H. (2016). Learning to Think Mathematically: Problem Solving, Metacognition, and Sense Making in Mathematics (Reprint). *Journal of Education*, *196*(2), 1–38. https://doi.org/10.1177/002205741619600202

Siregar, N. C., Rosli, R., & Maat, S. M. (2020). The Effects of a Discovery Learning Module on Geometry for Improving Students' Mathematical Reasoning Skills, Communication

and Self-Confidence. *International Journal of Learning, Teaching and Educational Research*, *19*(3), 214–228. https://doi.org/10.26803/ijlter.19.3.12

Steen, L. A. (1999). *Twenty Questions about Mathematical Reasoning* (ED440849). ERIC. https://files.eric.ed.gov/fulltext/ED440849.pdf

Stemler, S. E. (z.d.). *A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability*. https://doi.org/10.7275/96JP-XZ07

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is Measured in Mathematics Tests? Construct Validity of Curriculum-Based Mathematics Measures. *School Psychology Review, 31*(4), 498–513. https://doi.org/10.1080/02796015.2002.12086170

Tout, D. (2020). Evolution of adult numeracy from quantitative literacy to numeracy: Lessons learned from international assessments. *International Review of Education*, *66*(2–3), 183–209. https://doi.org/10.1007/s11159-020-09831-4

Tout, D., Gal, I., van Groenestijn, M., Manly, M., & Schmitt, M. J. (2020). *PIAAC Numeracy Task Complexity Schema: Factors that impact on item difficulty*. Australian Council for Educational Research. https://doi.org/10.37517/978-1-74286-609-3

Turner, R. (2010). Exploring mathematical competencies. *Research Developments, 24*(24), 5.

Turner, Ross (2012, April 13-17). *Some drivers of test item difficulty in mathematics.* [Conferentie Presentatie]. The Annual Meeting of the American Educational Research Association (AERA), Vancouver, Canada. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1006&context=pisa

Van den Bossche, R. (2022). *De moeilijkheidsgraad van een wiskundetoets voorspellen.* [Thesis Vrije Universiteit Brussel] VUB Campus archief.

Vancaeneghem, J. (2022, 15 december). *De omtrek van een rechthoek is al te veel gevraagd: aanzienlijke groep leerlingen haalt absolute minimum voor*. Het Nieuwsblad. https://www.nieuwsblad.be/cnt/dmf20221214_97824324

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, *16*(1), 93. https://doi.org/10.1186/s12874-016-0200-9

Zimmaro, D. M. (2004). *Writing good multiple-choice exams.* Measurement and Evaluation Center. University of Texas at Austin. Opgehaald op 8 maart 2023, van https://pmm.uinsu.ac.id/wp-content/uploads/2021/07/Panduan-membuat-butir-soal-pilihan-ganda-HOT.pdf.

7. Bijlagen

(1) Geselecteerde wiskunde-items voor de predictieve validatie

(2) Begeleidend document voor de beoordelaars

(3) Ingevuld complexiteitsschema

(4) R Code voor berekening Krippendorff's alpha coëfficiënt

(5) Data voor berekening interbeoordelaarsbetrouwbaarheid (origineel)

(6) Data voor berekening interbeoordelaarsbetrouwbaarheid (revisie)

(7) Semigestructureerd interview – voorbeeld van de template

(8) Toestemmingsformulier