

Universiteit Antwerpen

Instituut voor Onderwijs- en Informatiewetenschappen

**PAARSGEWIJZE VERGELIJKING ALS
EVALUATIETECHNIEK VOOR WISKUNDIG
PROBLEEMOPLOSSEN**

Kristof Vermeiren

Masterproef voorgelegd met het oog op het
behalen van de graad van master in de
Opleidings- en Onderwijswetenschappen

Promotor: prof. dr. S. De Maeyer
Co-promotor: dr. L. Coertjens

Het beoordelen van wiskundige probleemoplossingsvaardigheden: kan het ook anders?

De evaluatiemethodes voor wiskundig probleemoplossen die binnen ons onderwijssysteem gangbaar zijn, worden de laatste jaren alsmaar vaker bekritiseerd. Ze zouden oppervlakkig leergedrag uitlokken, wat op zijn beurt ervoor zorgt dat leerlingen in een verdere school- of werkloopbaan onvoldoende in staat zijn om wiskunde efficiënt toe te passen op nieuwe problemen. Kristof Vermeiren (Universiteit Antwerpen) concludeert in zijn masterscriptie dat er in de methode van ‘paarsgewijze vergelijking’ een waardig alternatief schuilgaat.

Kristof Vermeiren, student in de Opleidings- en Onderwijswetenschappen, liet de wiskundige probleemoplossingsvaardigheden van 58 leerlingen uit het secundair onderwijs beoordelen aan de hand van paarsgewijze vergelijking. Eigen aan deze beoordelingsmethode is dat prestaties van leerlingen onderling vergeleken worden door meerdere leerkrachten. Dit resulteert niet in puntenscores, maar in een rangorde die gaat van slechtste tot beste oplossing. Binnen zijn studie maakte Vermeiren een eerste verkenning omtrent de geschiktheid van paarsgewijze vergelijking door na te gaan in hoeverre bekomen rangordes betrouwbaar zijn en in welke mate beoordelaars hun keuzes baseren op zaken die verband houden met de wiskundige competentie ‘probleemoplossen’.

In zijn studie liet Vermeiren de leerlingen twee wiskundige problemen oplossen, die vervolgens paarsgewijs werden vergeleken. Hij ondervond dat rangordes met hoge betrouwbaarheid konden worden bereikt. Dit betekent dat de beoordelingsmethode een zeer betrouwbaar beeld oplevert van de probleemoplossingsvaardigheden van leerlingen. Daarnaast stelt Vermeiren dat leerkrachten tijdens hun paarsgewijs vergelijkingsproces voornamelijk oog hebben voor relevante zaken: “De argumenten, die de beoordelingswijze van beoordelaars verduidelijkten, bestonden voor respectievelijk 76% (probleem 1) en 83% (probleem 2) uit competentiegerelateerde zaken.” Hierbinnen merkt hij op dat, in vergelijking met wiskundige kennis en vaardigheden, veel meer belang wordt gehecht aan de metacognitieve vaardigheden van leerlingen. Ook deze bevinding spreekt in het voordeel van paarsgewijze vergelijking, aangezien metacognitieve vaardigheden centraal dienen te staan binnen de evaluatie van wiskundige probleemoplossingsvaardigheden.

“Wanneer we alles in rekening nemen, kunnen we besluiten dat paarsgewijze vergelijking als evaluatiemethode voor wiskundige probleemoplossingsvaardigheden zeker veelbelovend is.”, aldus Vermeiren. Wel waren er implicaties dat de mate waarin beoordelaars rekening houden met competentiegerelateerde zaken mogelijks vatbaar is voor de vraagstelling van de toetsvragen. Hiervoor is vervolgonderzoek onontbeerlijk.

Meer weten?

Kristof Vermeiren: kristof.vermeiren@gmail.com of 0494 14 40 39

Promotor prof. dr. Sven De Maeyer: sven.demaeyer@uantwerpen.be of 03 265 49 32

Paarsgewijze vergelijking als evaluatietechniek voor wiskundig probleemoplossen

Kristof Vermeiren

ABSTRACT

De evaluatiemethodes voor wiskundig probleemoplossen die binnen ons onderwijssysteem gangbaar zijn, worden de laatste jaren alsmaar vaker bekritiseerd. Ze zouden oppervlakkig leergedrag uitlokken, wat op zijn beurt ervoor zorgt dat leerlingen in een verdere school- of werkloopbaan onvoldoende in staat zijn om wiskunde efficiënt toe te passen op nieuwe problemen. Deze studie ging de geschiktheid na van paarsgewijze vergelijking (PV) als alternatieve evaluatiemethode voor wiskundige probleemoplossingsvaardigheden. De studie vond plaats in een Vlaamse context, waarbinnen tot nog toe geen onderzoek werd gevoerd naar PV. Om de probleemoplossende competentie van 58 leerlingen in kaart te brengen, werd hen een toets voorgelegd die twee wiskundige problemen bevatte. Vervolgens beoordeelden 14 wiskundigen de geleverde prestaties aan de hand van PV. De analyses tonen aan dat de evaluatie geresulteerd heeft in rangordes van hoge betrouwbaarheid. Daarnaast constateerden we dat vakexperten, tijdens het maken van beoordelingen, hun keuzes voornamelijk baseren op competentiegerelateerde zaken. Er werd met andere woorden een goede inhoudsvaliditeit vastgesteld. Hierbij maakten we wel de bemerking dat sommige onderzoeksresultaten impliceren dat de mate van inhoudsvaliditeit niet uitsluitend kan worden toegewezen aan de beoordelingsmethode. De manier waarop toetsvragen worden gesteld, blijkt immers ook invloed te hebben op de mate waarin beoordelaars leerlingprestaties op een inhoudsvalide wijze vergelijken. Implicaties van verdere resultaten en aanbevelingen voor vervolgonderzoek werden verder bediscussieerd.

INTRODUCTIE

In onze technologisch georiënteerde maatschappij heeft wiskunde de laatste decennia een alsmaar prominentere plaats ingenomen. Enerzijds is er veel vraag naar praktisch bruikbare en concrete wiskunde, en anderzijds is er nood aan meer theoretische en abstracte wiskunde (Secundair onderwijs–Wiskunde–Uitgangspunten, z.d.). In de uitgangspunten voor het secundaire onderwijs, die de Vlaamse overheid vooropstelt voor het vak wiskunde, wordt het belang van een nauwe wisselwerking tussen theorievorming en de bruikbaarheid ervan voor het oplossen van concrete problemen dan ook meermaals in de verf gezet. Vooral deze probleemoplossende vaardigheden nemen alsmaar in belang toe. Door het vlugge tempo waarmee de samenleving verandert, moeten leerlingen voldoende soepelheid ontwikkelen om snel en efficiënt allerhande problemen op te lossen.

Eerder onderzoek heeft aangetoond dat de realisatie van probleemoplossende vaardigheden bij leerlingen geen sinecure is. Heel wat leerlingen verlaten de schoolbanken zonder voldoende in staat te zijn om wiskunde op een efficiënte manier toe te passen op nieuwe problemen waarmee ze in de maatschappij, op het werk of bij verdere studies in contact komen. De transfer van opgedane kennis en vaardigheden naar een verdere school- of werkloopbaan wordt zelfs bij sterk presterende leerlingen vaak niet of onvoldoende bereikt (ACT, 2006).

Jones, Swan en Pollit (2014) leggen een mogelijke oorzaak bij de heersende evaluatievormen in de wiskundeles. Evaluatieactiviteiten verschaffen zowel de leerlingen als de onderwijsgevende waardevolle informatie (Dochy & Gijbels, 2009). De onderwijsgevende krijgt zicht op het feit of leerlingen/studenten de vooropgestelde doelen bereikt hebben en kan aan de hand van deze informatie zijn instructie bijsturen en/of verder plannen. Voor de leerlingen wordt duidelijk wat de verwachtingen van de onderwijsgevende zijn en of ze daaraan beantwoorden. Meetinstrumenten die gehanteerd worden bij wiskundig probleemoplossen kenmerken zich vaak door hun fragmentarisch karakter (Jones et al., 2014). Ze bestaan voornamelijk uit oefeningen waarbij leerlingen scores worden toegekend op basis van (tussen)uitkomsten, tekeningen... die worden afgetoetst aan vooropgestelde criteria (Pollitt, 2004). Bij probleemoplossing zijn zaken zoals het exploratie- en redeneerproces van leerlingen en gekozen oplossingsstrategieën echter meer van belang (Szetela & Nicol, 1992). De fragmentaire schema's, die slechts componenten beoordelen i.p.v. de alomvattende prestatie en redeneerstructuur, belemmeren probleemoplossers om hun ware capaciteiten te tonen. Bovendien moedigen ze leerlingen aan om reproductief te studeren (DfE, 2011). Reproductieve leerstrategieën kunnen opportuun zijn voor eenvoudige leertaken zoals bijvoorbeeld bij het memoriseren van een wiskundige formule. Voor complexe leertaken, zoals het ontwikkelen van wiskundige probleemoplossingsvaardigheden, dringen andere leerstrategieën zich op. Relateren (nieuwe informatie in verband brengen met voorkennis), structureren (via samenvattingen en schema's de globale structuur in het leermateriaal opsporen en verhelderen) en toepassen (zelf toepassingen maken van het geleerde om het geleerde beter te verstaan), zijn voorbeelden van dergelijke relevante leerstrategieën (Verschaffel, Van Dooren, Elen, & Clarebout, 2009). Aangezien het ontwikkelen van wiskundige probleemoplossingsvaardigheden een complexe leertaak is, dienen leerkrachten met tal van zaken rekening te houden binnen hun klascontext. In wat volgt trachten we duidelijkheid te scheppen omtrent de veelzijdigheid van het begrip 'wiskundig probleemoplossen'.

OOGMERK VAN WISKUNDIG PROBLEEMOPLOSSEN

Het oplossen van vraagstukken uit de wiskundeles, het vervolledigen van sudoku puzzels in de krant en het berekenen van kortingen tijdens de koopjesperiode zijn slechts enkele voorbeelden van activiteiten waarbij de term wiskundig probleemoplossen op zijn plaats is. Het begrip kenmerkt zich door zijn complexiteit en veelzijdigheid. Hoe dan ook, als we de kern van probleemoplossing in kaart willen

brengen en de stappen willen beschrijven die doorlopen moeten worden tijdens het oplossingsproces, hebben we nood aan een preciezere en alomvattende definitie van het concept.

Reitman (1965) omschreef een wiskundig probleemoplosser als een persoon die op een wiskundige manier met informatie aan de slag gaat om een bepaald doel te bereiken, maar niet onmiddellijk inziet hoe hij dit moet doen. Bij het doorlopen van het probleemoplossingsproces spelen tal van factoren een belangrijke rol. In de literatuur zijn heel wat artikels terug te vinden van auteurs die elk op hun beurt een poging hebben ondernomen om het probleemoplossingsproces af te bakenen en het concept beter te begrijpen (Polya, 1988; Montague, 1992; Mayer 2003; Royer & Garofoli, 2005; Funke, 2010). Hoewel de complexiteit van het begrip bekrachtigd wordt door deze uiteenlopende invalshoeken, zijn er toch een aantal factoren die we vaak terug zien keren. In wat volgt zullen enkele van deze factoren gaandeweg worden toegelicht.

Probleemsituatie vs. Oefening

Het hoeft geen betoog dat om aan probleemoplossing te kunnen doen, er eerst sprake moet zijn van een probleem. Hoewel hier volgens PISA in het Vlaamse onderwijs nogal licht wordt overgegaan (De Meyer, Warlop, & Van Camp, 2012), is dit toch een zeer belangrijk aandachtspunt: vanaf wanneer kunnen we spreken over een probleemsituatie?

De sleutel bij het bepalen of een gegeven situatie een probleemsituatie is, ligt bij de analyse van hoe iemand reageert op de situatie (Mayer, 1985; Brown & Cocking, 1999; English, 2002). Een eerste voorwaarde om te kunnen spreken over een probleem is dat de persoon zich verantwoordelijk moet voelen om zich te verdiepen in gestelde vragen en vervolgens doelen stelt om de situatie op te lossen. Vervolgens moet de persoon een strategie zoeken of ontwikkelen om de situatie op te lossen. Als er niet meteen een strategie wordt gevonden, hebben we te maken met een probleemsituatie. Wanneer de persoon een strategie hanteert, die hij reeds in soortgelijke contexten heeft gebruikt, wordt de situatie niet aanzien als een probleemsituatie maar als een oefening (Dossey, McCrone, & O'Sullivan, 2006). Met andere woorden, een oefening is een situatie waarbij de persoon vertrouwd is met de kennis en tools die nodig zijn om de situatie op te lossen én in staat is deze kennis toe te passen. We merken op dat het onderscheid tussen probleem en oefening gemaakt dient te worden op basis van het individu. Wat voor de ene persoon een probleem is, is voor de andere een oefening (Mayer & Wittrock, 2006).

Proces van skills, metaskills & will

Om het oplossingsproces van een probleem succesvol te kunnen doorlopen, benadrukt eerder onderzoek de cruciale rol van domeinspecifieke kennis en vaardigheden (Chi, Glaser, & Farr, 1988; Ericsson & Smith, 1991; Mayer, 1992; Smith, 1991). Zo hebben leerlingen bijvoorbeeld bij het berekenen van de oppervlakte van een parallellogram de skills nodig om: de lengte en de hoogte te identificeren, rekenkundige bewerkingen uit te voeren zoals het vermenigvuldigen van de basis met de hoogte, een

oppervlakte-eenheid achter het resultaat te plaatsen... Wat probleemoplossing betreft, wordt onderwijs vaak zo ingericht dat de focus bij dergelijke cognitieve basisvaardigheden ligt. Problemen worden geanalyseerd en men gaat na hoe cognitieve vaardigheden, die nodig zijn om tot de oplossing te komen, kunnen worden opgesplitst om deze vervolgens geïsoleerd en stapsgewijs aan te leren (Mayer, 1998). Hoewel dit misschien wel de meest voor de hand liggende methode lijkt om probleemoplossende prestaties te verbeteren, toont eerder onderzoek duidelijk aan dat werken aan domeinspecifieke kennis en basisvaardigheden niet volstaat (Mayer, 1998). De voornaamste redenen waardoor probleemoplossers falen in hun opdracht zijn: ineffectieve organisatie van het oplossingsproces, foute keuze van oplossingsstrategieën, verkeerd begrip van de probleemsituatie en het gebrek aan monitoring en controle van het oplossingsproces (Victor, 2004). Om transfer mogelijk te maken heeft een probleemoplosser dus niet alleen domeinspecifieke kennis en vaardigheden nodig, maar moet hij ook in staat zijn deze te managen. Kortom, de vraag naar metacognitieve vaardigheden dringt zich op.

Metacognitieve vaardigheden, ook wel eens 'hogere' cognitieve vaardigheden genoemd, gaan over iemands bewustzijn van het eigen denkproces en het vermogen om deze processen te controleren (Hacker & Dunlosky, 2003). Een eerste opdeling hierin werd gemaakt door Flavell (1979), die stelde dat metacognitieve vaardigheden enerzijds bestaan uit metacognitieve kennis en anderzijds uit metacognitieve controle. Metacognitieve kennis verwijst naar de kennis die iemand heeft over zijn eigen cognitieve capaciteiten en het bewustzijn over wanneer het gepast is om deze in te schakelen. Toegepast op wiskundig probleemoplossen betekent dit dat de probleemoplosser inzicht heeft in zijn eigen wiskundige kennis (feiten, algoritmes, heuristieken...) en weet hoe en in welke context het aangewezen is om deze kennis te gebruiken. Metacognitie vergt, naast deze metacognitieve kennis, ook het vermogen om deze kennis effectief te gebruiken tijdens het probleemoplossingsproces. Dergelijk vermogen wordt metacognitieve controle genoemd. Metacognitieve controle/regulatie zorgt ervoor dat de probleemoplosser, a.d.h.v. zijn metacognitieve kennis, cognitieve processen kan reguleren en controleren.

Tot slot onderstrepen Mayer (1998) en Funke (2010) dat het gebruik van skills en metaskills afhangt van motivationele en affectieve factoren. De rol van motivatie bij probleemoplossen kent een lange geschiedenis binnen de onderwijspsychologie. Hoewel theoretici de invloed van deze rol niet steeds benadrukken, is men het er wel over eens dat de bereidheid van de probleemoplosser om zich in te laten met nieuwe omstandigheden integraal deel uitmaakt van de probleemoplossende competentie (Mayer, 1998).

EVALUATIE VAN WISKUNDIGE PROBLEEMOPLOSSINGSVAARDIGHEDEN

Zoals eerder werd aangehaald, dient evaluatie onder andere om de kwaliteit van het leerproces van leerlingen na te gaan. Uiteraard moeten ook de manier van evalueren en de daarbij gebruikte instrumenten aan bepaalde kwaliteitscriteria voldoen (Dochy & Gijbels, 2009). Bij kwaliteitsvolle toetsing dient er sprake te zijn van een goede validiteit en betrouwbaarheid. Betrouwbaarheid heeft betrekking tot de vraag of de resultaten al dan niet vertekend zijn, terwijl validiteit de vraag stelt of we meten wat we willen meten (Dousma, Horsten, & Brants, 1995).

Met het oog op validiteit dienen toetsen, die peilen naar de probleemoplossingsvaardigheden van leerlingen, als doel te hebben om een beeld te krijgen van de drie componenten waaruit de competentie bestaat, nl. domeinspecifieke kennis en vaardigheden, metacognitieve vaardigheden en motivatie (Mayer, 1998). Onderzoekers zijn het erover eens dat men via schriftelijke toetsen, die opgebouwd zijn uit gesloten vraagtypes, geen valide beeld kan vormen van competenties (Darling-Hammond & Adamson, 2010; Rubin, 1996). Daarnaast geven Schoenfeld (1992), Pellegrino, Chudowsky en Glaser (2001) te kennen dat de focus bij een evaluatie van probleemoplossing moet liggen op de beschrijving van (meta)cognitieve acties die geleverd werden tijdens het oplossingsproces. Deze beschrijvingen zijn onmogelijk a.d.h.v. meerkeuze- of korte-antwoordvragen en vergen nu eenmaal een uitgebreid, leerlinggeproduceerd antwoord. Een open vraagtype is dan ook opportuun. De bedenking dient hierbij wel gemaakt te worden dat het feit dat een vraag open is, niet hoeft te betekenen dat deze peilt naar probleemoplossende vaardigheden.

Ondanks geleverde onderzoeksresultaten, wordt nog te vaak teruggegrepen naar fragmentarisch opgebouwde toetsinstrumenten. Deze instrumenten kenmerken zich door de aanwezigheid van een voor leerlingen gekende structuur, die bij hen het gevoel oproept dat ze een gekend algoritme moeten reproduceren in een ‘nieuwe’ situatie (DfE, 2011). De beoordelaar kent vervolgens aan (tussen)uitkomsten scores toe m.b.v. vooropgestelde beoordelingscriteria (Pollitt, 2004). De geschiktheid van dergelijke vorm van absoluut beoordelen bij de evaluatie van competenties werd reeds veelvuldig in vraag gesteld. Onderzoekers maakten de bedenking dat alle relevante criteria op voorhand niet te onderscheiden zijn. Criterialijsten zouden dan ook niet tegemoet komen aan de veelzijdigheid en complexiteit van een competentie (Jones & Alcock, 2013; Sadler, 2009). Bovendien kan in vraag worden gesteld of de gangbare evaluatiemethodes probleemoplossingsvaardigheden meten, dan wel de vaardigheid om leerstof uit het hoofd te leren (DfE, 2011).

Voorgaande bedenkingen impliceren dat de wiskundige probleemoplossingscompetentie van leerlingen moeilijk op een valide manier kan worden beoordeeld aan de hand van de heersende evaluatiemethodes in de wiskundeles. In wat volgt, werpen we ons licht op een mogelijk alternatief.

ALTERNATIEVE MANIER VAN EVALUEREN: PAARSGEWIJS VERGELIJKEN

In dit onderzoek staan we stil bij een alternatieve evaluatievorm: paarsgewijze vergelijking (PV). PV is ontstaan nadat de grondlegger ervan, Louis L. Thurstone (1927), tot het inzicht kwam dat mensen zeer onbetrouwbaar zijn bij het concreet inschatten van fysische eigenschappen zoals gewicht, temperatuur en lengte. Wanneer mensen echter slechts gevraagd worden om een fysische eigenschap van twee items met elkaar te vergelijken, zoals bijvoorbeeld bepalen welke van twee gegeven voorwerpen het zwaarst is, blijken ze veel betrouwbaarder te zijn. Thurstone trok dit door en stelde dat mensen in het algemeen beter zijn in vergelijkingen maken, dan in het vellen van een concreet oordeel. Als we dit toepassen op onderwijs, zou dit kunnen betekenen dat leerkrachten betrouwbaarder zijn in het vergelijken van prestaties tussen leerlingen, dan in het absoluut beoordelen van een welbepaalde prestatie van één leerling.

Het basisprincipe van PV is simpel. Verscheidene beoordelaars krijgen telkens individueel prestaties van twee kandidaten (bv. schriftelijke werkstukken van leerlingen) voorgeschied. Vervolgens is het aan hen om de ene als beter dan de andere te beoordelen. Van scoretoekenning is geen sprake. Gedurende verschillende rondes worden de prestaties telkens in andere paren voorgelegd aan de beoordelaars. Op het einde van de rit resulteren deze beoordelingen in een rangorde van 'slechtste prestatie' tot 'beste prestatie'.

Hoewel PV oorspronkelijk afkomstig is uit de psychofysica, is er binnen de onderwijscontext ook al heel wat onderzoek naar verricht. Pollitt en Murray (1993) waren de eersten die PV als evaluatiemiddel gebruikten binnen het onderwijs. Dit gebeurde bij het toetsen van de spreekvaardigheid voor vreemde talen van studenten uit het hoger onderwijs. Enkele jaren later werd de methode geïntroduceerd bij de verschillende Britse besturen voor examinatie (D'Arcy, 1997). Uit dat onderzoek, dat tevens gevoerd werd binnen het hoger onderwijs, kwam naar voren dat de beoordelaars, afkomstig uit verschillende besturen, op een consistente manier beoordeelden, ongeacht de aard van de taak, het examen of de beoordeelde studentenpopulatie. Dit resulteerde in rangordes met een hoge betrouwbaarheid. Kimbell e.a. (2009) gebruikte PV bij de beoordeling van designprocessen die leerlingen uit het secundair onderwijs hadden doorlopen bij het ontwerpen van een product. Doorheen hun project hielden leerlingen bewijsmateriaal van hun vorderingen bij in een portfolio, dat vervolgens werd beoordeeld met behulp van PV. De resultaten van dit onderzoek toonden aan dat de bekomen rangorde van prestaties een betrouwbaarheidswaarde had van 0,96; een mate van betrouwbaarheid die volgens Pollitt (2012) quasi onhaalbaar is voor de heersende evaluatiemethodes, die zich kenmerken door scoretoekenning van één beoordelaar a.d.h.v. vooropgestelde criteria. Dit onderzoek maakte dan ook meteen duidelijk dat voor het relatief (met elkaar vergelijkend) beoordelen van leerlingen uit het secundair onderwijs, PV een betrouwbaar evaluatiemiddel kan zijn en dat verder onderzoek relevant is.

Wat wiskunde betreft is er nog maar weinig onderzoek gevoerd naar de geschiktheid van PV als evaluatiemethode. Jones, Inglis, Gilmore en Hodgen (2013) stonden stil bij het conceptuele begrip van leerlingen uit het secundair onderwijs bij wiskundige functies en breuken. Ze ondervonden dat PV, in vergelijking met de traditionele evaluatiemethodes, conceptueel begrip van leerlingen op een meer betrouwbare en valide wijze meet.

Hoewel resultaten uit eerder onderzoek de sterke punten van PV als evaluatiemethode voor het meten van competenties reeds meermaals in de verf hebben gezet, is verder onderzoek relevant. Allereerst stellen we vast dat het merendeel van het reeds gevoerde onderzoek plaatsvond binnen eenzelfde (Britse) context van het hoger onderwijs. Daarnaast stellen we ook vast dat onderzoek naar de geschiktheid van PV als evaluatiemethode voor wiskundige competenties nog zeer beperkt is. Deze studie zal dan ook plaatsvinden binnen de context van het Vlaamse secundair wiskundeonderwijs, die tot nog toe niet werd onderzocht.

ONDERZOEKSFOCUS

Met dit onderzoek maken we een verkenning omtrent de geschiktheid van PV als evaluatietechniek voor wiskundige probleemoplossingsvaardigheden. Allereerst gaan we de betrouwbaarheid na van deze evaluatiemethode:

- i. In welke mate levert PV een betrouwbare rangorde op van geleverde prestaties?

Daarnaast willen we een beeld krijgen van de validiteit. Argumenten die beoordelaars spontaan aanhalen om gemaakte beslissingen kracht bij te zetten, zullen onder de loep worden genomen:

- ii. In welke mate levert PV een inhoudsvalide beoordeling op van geleverde prestaties?

METHODOLOGIE

Onderzoekscontext

Dit onderzoek vindt plaats binnen een Vlaamse secundaire school, bij leerlingen uit het tweede leerjaar. Binnen het vak wiskunde leggen in totaal 58 leerlingen uit de A-stroom (10 Latijn, 21 moderne en 27 tso) een schriftelijke toets af die peilt naar hun wiskundige probleemoplossingsvaardigheden. De toets (Appendix A) werd op voorhand opgesteld in samenwerking met twee docenten wiskunde uit een Vlaamse lerarenopleiding. Er werd beroep gedaan op hun expertise om vragen op te stellen conform de noden die Mayer (1998) vooropstelt bij probleemoplossing. Leerlingen krijgen de kans om uitgebreid te antwoorden en tekenen te geven van hun domeinspecifieke kennis en vaardigheden, metacognitieve vaardigheden en motivatie. Dit werd onder andere mogelijk gemaakt door: te werken met open

vraagstellingen, veel antwoordruimte te voorzien en wiskundige problemen aan te bieden die zich lenen tot verschillende oplossingsmethodieken.

De toets bevat twee wiskundige problemen. Hierdoor bekomen we in totaal 116 oplossingen van leerlingen. Om PV mogelijk te maken, zullen enkel oplossingen van hetzelfde probleem met elkaar vergeleken worden. Oplossingen van toetsvraag 1 zullen dus nooit vergeleken worden met oplossingen van toetsvraag 2 (en vice versa).

Beoordelaars

Aangezien we ons binnen een wiskundige context begeven, en er sprake moet zijn van relevante professionele expertise, werden wiskundigen aangesteld als beoordelaar. In totaal werden 14 beoordelaars gevraagd om deel te nemen aan het onderzoek. 10 beoordelaars zijn afgestudeerde wiskundeleerkrachten die momenteel werkzaam zijn in het onderwijs, waarvan vijf in de school waarbinnen dit onderzoek plaatsvindt. De andere vier beoordelaars zijn laatstejaarsstudenten uit een lerarenopleiding. De beoordelaars verschillen onderling dan ook wat betreft ervaring op vlak van het beoordelen van wiskundige probleemoplossingsvaardigheden, gaande van helemaal geen ervaring, tot zeer ervaren.

Door middel van individuele gesprekken werden beoordelaars ingelicht over de werking van het paarsgewijze vergelijkingsplatform (D-PAC). Tijdens het beoordelen zelf werden de beoordelaars voorzien van een statement die de essentie van wiskundig probleemoplossen duidelijk maakt (Appendix B). Een criterialijst werd niet voorzien, aangezien de taak van de beoordelaars er uitsluitend uit bestond om uit twee gegeven oplossingen, de beste te kiezen. Wel werden de beoordelaars verzocht om tijdens het maken van hun beoordelingen de essentie van probleemoplossen in hun achterhoofd te houden.

DATAVERZAMELING

Paarsgewijze vergelijkingsdata

Om oplossingen van leerlingen paarsgewijs te laten vergelijken door de beoordelaars, werd gebruik gemaakt van het digitale platform D-PAC. Binnen dit online-beoordelingsplatform werden de oplossingen van alle leerlingen geüpload, waarna vervolgens beoordelaars in staat werden gesteld om via de computer tijds- en plaatsonafhankelijk paarsgewijze vergelijkingen te maken. Iedere beoordelaar kreeg 84 paren toegewezen, 42 bij iedere toetsvraag, die ze over een tijdspanne van twee weken paarsgewijs dienden te vergelijken. Dit resulteerde bij toetsvraag 1 in 588 volbrachte vergelijkingen. Iedere oplossing werd vergeleken met 20 à 21 oplossingen van andere leerlingen. Bij toetsvraag 2 bekwamen we 518 volbrachte vergelijkingen. Iedere oplossing werd vergeleken met 17 à 18 oplossingen van andere leerlingen.

Kwalitatieve data

Om een antwoord te vinden op onze tweede onderzoeksvraag werd kwalitatieve data verzameld. Na iedere volbrachte vergelijking werd aan de beoordelaars gevraagd om een korte uitleg te noteren, die hun gemaakte keuze verklaart. Er werd besloten om de beoordelaars hiertoe niet te verplichten aangezien dit zou kunnen resulteren in minder intuïtieve antwoorden en bovendien het beoordelingsproces zou kunnen beïnvloeden (Whitehouse, 2012). Daarnaast werd, met het oog op de betrouwbaarheid van de dataverzameling, aan de beoordelaars de mogelijkheid gegeven om hun beoordelingswerk naar wens in te plannen. Bij toetsvraag 1 gaven de beoordelaars uiteindelijk bij 571 van de 588 vergelijkingen een verklaring (97%). Bij toetsvraag 2 gaven de beoordelaars bij 507 van de 518 vergelijkingen een verklaring (98%).

ANALYSE

Analyse van kwalitatieve data

Verklaringen van beoordelaars zijn vaak opgebouwd uit verschillende argumenten. Om deze reden werden de verklaringen opgesplitst in segmenten. Het werken met segmenten maakt het immers mogelijk om complexe, ruwe kwalitatieve data aangaande een bepaald fenomeen op een betekenisvolle manier interpreteerbaar te maken (Boyatzis, 1998). Toegepast op onze studie bestaan segmenten uit (een deel van) een zin of alinea die op een bepaalde manier de beslissing van de beoordelaar verduidelijkt. Deze segmentatie resulteerde voor toetsvraag 1 in 957 segmenten. Hierbinnen vinden we 119 segmenten die dubbel werden gecodeerd. Bij toetsvraag 2 bekwamen we 807 segmenten, waarvan 129 dubbel gecodeerd. Wat de dubbels betreft, gaat het hier over segmenten die taalkundig moeilijk konden worden opgesplitst, maar die wel duidelijk argumenten bevatten die naar verschillende codeercategorieën verwijzen. In dit onderzoek zal verder worden gerekend met respectievelijk de totalen van 957 en 807 argumenten, aangezien de ‘dubbele’ segmenten telkens twee verschillende argumenten bevatten.

Het codeerwerk werd aangevangen op een deductieve manier. Deze werkwijze kenmerkt zich door het feit dat vertrokken wordt vanuit een ruim theoretisch kader (Donche, 2014). Binnen onze studie werd aan de hand van de competentiebeschrijving van Mayer (1998) een codeerschema opgesteld. Argumenten die verwezen naar deze competentiebeschrijving werden ondergebracht in de hoofdcategorie ‘competentiegerelateerde argumenten’. Deze categorie bevat Mayer’s onderverdeling: domeinspecifieke kennis en vaardigheden, metacognitieve vaardigheden en motivatie.

Het gevaar bij een al te deductieve benadering van kwalitatieve data is dat dit theoretisch etnocentrisme in de hand werkt (Donche, 2014). In dit laatste geval, stuurt het theoretisch kader het interpretatieproces, met voornamelijk de nadruk op ‘herkenning’ van wat al gekend is. Aan dit gevaar werd tegemoet gekomen door gaandeweg het coderen, op momenten dat beoordelaars leken te verwijzen naar niet-competentiegerelateerde zaken, op een inductieve manier verder te gaan. De inductieve benadering is

erop gericht om een meest passende theorie te formuleren op basis van de concrete empirische gegevens in het onderzoek (Donche, 2014). Toegepast op ons codeerproces, ontstonden hierdoor nieuwe categorieën die onder werden gebracht in de hoofdcategorie ‘niet-competentiegerelateerde argumenten’. Het uiteindelijke codeerschema is terug te vinden in appendix C.

Om een gepast antwoord te kunnen bieden op onderzoeksvraag 2 werd vervolgens de kwalitatieve data gekwantificeerd. Aangezien we binnen ons onderzoek te maken hebben met een uitgebreide kwalitatieve dataverzameling bij meerdere respondenten, is kwantificering immers een zinvolle analysetechniek om uitspraken te doen ten aanzien van frequentie en zelfs predictie (Donche, 2014). Concreet werd bij iedere categorie van argumenten (bv. domeinspecifieke kennis en vaardigheden, metacognitieve vaardigheden en motivatie) het aantal argumenten opgeteld. Vervolgens werden deze sommen omgezet naar relatieve aantallen. Dit maakte het mogelijk om de verhouding tussen het aantal competentiegerelateerde en het aantal niet-competentiegerelateerde argumenten na te gaan.

Een goede validiteit werd nagestreefd aan de hand van onderzoekerstriangulatie doordat het codeerproces en de analyse door meerdere onderzoekers werd opgevolgd. Daarnaast werd, met het oog op de repliceerbaarheid van het onderzoek, gebruik gemaakt van een logboek waarin tijdens het coderen memo's werden bijgehouden die gedachtes en gemaakte keuzes verduidelijken.

Voor het codeerwerk werd gebruik gemaakt van NVivo 10. De kwantificering van de data werd uitgevoerd in Excel.

Analyse van kwantitatieve data

De vergelijkingen die de beoordelaars maakten, leverden binnen het D-PAC platform statistische data op. Hierop werd vervolgens het statistische ‘Bradley-Terry model’ uitgevoerd. Dit model maakt gebruik van een maximum-likelihood-schatter, wat erop neer komt dat het model een schattingsmethode hanteert die ervoor zorgt dat bekomen parameters, rekening houdend met de steekproefuitkomst, zo aannemelijk mogelijk zijn (Firth, 2005). De belangrijkste parameters die het model oplevert zijn de maten van bekwaamheid voor iedere leerlingenprestatie. Concreet krijgt iedere leerlingenprestatie een waarde, uitgedrukt in logits, toegekend die de kans weergeeft dat deze leerlingenprestatie bij een vergelijking met een willekeurige andere leerlingenprestatie als winnaar uit de bus zou komen. Het spreekt voor zich dat leerlingenprestaties met hogere logit waardes gezien kunnen worden als betere, sterkere prestaties.

Aan de hand van deze logit scores bekomen we een schaal voor ‘wiskundig probleemoplossen’ waarbij 58 leerlingenprestaties gerangschikt staan van slechts naar best. Aangezien er twee toetsvragen werden

gesteld, waarvan de oplossingen vervolgens gescheiden paarsgewijs werden vergeleken, bekomen we twee aparte rangordes.

Om een antwoord te bieden op onderzoeksvraag 1 over de betrouwbaarheid van de opgeleverde rangorde, maken we gebruik van enerzijds de scheidingscoëfficiënt (G) en anderzijds van Cronbach's alpha (α). De scheidingscoëfficiënt geeft de verhouding weer tussen de spreiding van de oplossingen en de meetfout van deze spreiding. Hoge G -waarden gaan gepaard met betrouwbare verschillen tussen oplossingen. Hierbij aanvullend berekenen we Cronbach's alpha om een algemene mate van zekerheid weer te geven wat betreft de positie van de verschillende oplossingen binnen de rangorde. Bij een hoge Cronbach's alpha zijn slechts weinig van de verschillen tussen oplossingen toe te wijzen aan errorvariantie.

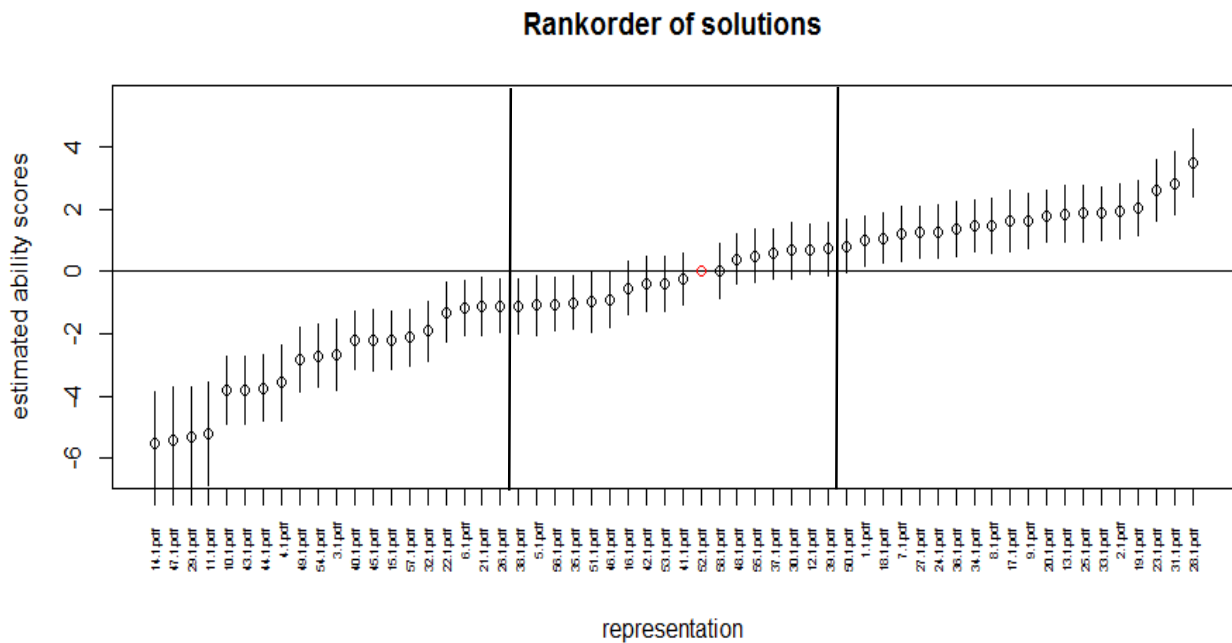
Ook bij deze fase van het onderzoek werd een goede validiteit nagestreefd aan de hand van onderzoekertriangulatie. De statistische verwerking heeft plaatsgevonden in samenwerking met meerdere onderzoekers.

De volledige statistische verwerking werd gedaan met behulp van het statistisch computerprogramma R, waarbinnen het pakket 'BradleyTerry2' werd geïnstalleerd.

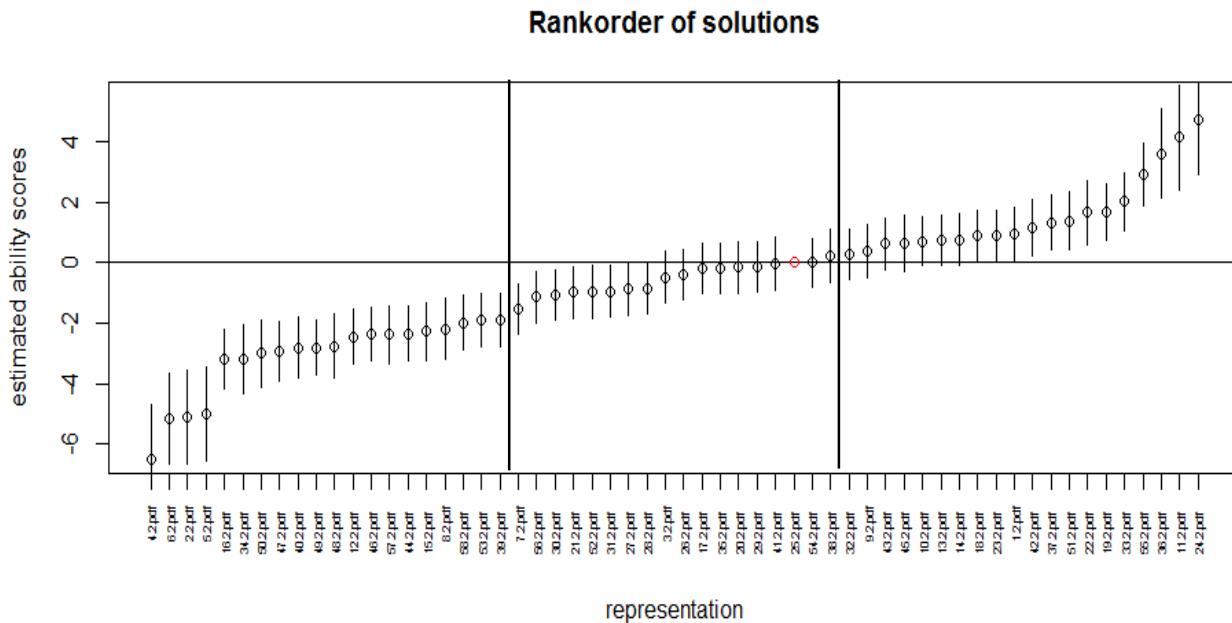
RESULTATEN

Betrouwbaarheid van rangordes

Figuur 1 (toetsvraag 1) en figuur 2 (toetsvraag 2) zijn weergaves van de rangordes die de wiskundige probleemoplossingsvaardigheden van de leerlingen voorstellen. Op de horizontale as werden de verschillende leerlingenrepresentaties gerangschikt volgens hun bijhorende bekwaamheidsscores. Op de verticale as staan de bekwaamheidsscores uitgedrukt in logits. Voor iedere representatie werd de bekwaamheidsscore uitgezet, alsook het bijhorende 95%-betrouwbaarheidsinterval. De bekwaamheidsscore die in het rood is weergegeven, werd gekozen als referentiecategorie. Dit is de representatie die bij een willekeurige vergelijking met een andere representatie evenveel kans heeft om als winnaar, dan als verliezer uit de bus te komen (logit = 0). Deze referentiecategorie werd bepaald op zulke manier dat de totale rangorde de hoogst mogelijke betrouwbaarheid kon bereiken.



Figuur 1. Rangorde toetsvraag 1.



Figuur 2. Rangorde toetsvraag 2.

Allereerst merken we op dat de evaluatie voor beide toetsvragen een mooi gespreide rangorde heeft opgeleverd. Hoewel, rekening houdend met het betrouwbaarheidsinterval, heel wat bekwaamheidsscores overlap creëren, vallen er toch duidelijk representaties van elkaar te onderscheiden. Om interpretatie te vergemakkelijken benoemen we twee representaties onderscheidbaar indien de bovengrens van de bekwaamheidsscore van de mindere representatie even groot of kleiner is

dan de geschatte bekwaamheidsscore van de betere representatie. In alle andere gevallen, nemen we aan dat de twee representaties niet te onderscheiden zijn. Op basis van deze interpretatiemethode vinden we binnen beide rangordes een viertal representaties dat zich onderscheidt van de rest aan de linkerkant van de grafiek. Leerlingen waarvan deze representaties afkomstig zijn, getuigen dan ook het minst van kwaliteitsvolle wiskundige probleemoplossingsvaardigheden. Aan de rechterkant van beide rangordes vinden we opnieuw een groepje van een viertal representaties dat zich onderscheidt van de rest. Deze kunnen gezien worden als de sterkste leerlingenprestaties.

Tussen de twee uiterste groepen bevinden zich heel wat leerlingenprestaties waarvan de bekwaamheidsscores dicht bij elkaar liggen. Desalniettemin zijn er nog duidelijk groepen te onderscheiden. Zo zijn bijvoorbeeld in beide rangordes de bekwaamheidsscores van de 20 meest linkse leerlingenprestaties (links van verticale lijn) duidelijk lager dan de bekwaamheidsscores van de 20 meest rechtse leerlingenprestaties (rechts van verticale lijn). Indien een willekeurige leerlingenprestatie uit de linkse groep vergeleken zou worden met een willekeurige leerlingenprestatie uit de rechtse groep, heeft deze laatste een beduidend hogere kans om als beste oplossing te worden bestempeld.

De overige (middelste) leerlingenprestaties zijn veel moeilijker van elkaar te onderscheiden. Dit zijn representaties die allemaal bij een willekeurige vergelijking met een andere representatie ongeveer evenveel kans zouden hebben om als winnaar, dan als verliezer uit de bus te komen.

Naast het paarsgewijs vergelijken van leerlingenprestaties, stond binnen dit onderzoek de vraag centraal of de bekomen rangordes betrouwbaar zijn. De betrouwbaarheidsmaten van beide rangordes liggen dicht bij elkaar. De scheidingscoëfficiënt (G) bedraagt bij de eerste toetsvraag 2,27. Hierdoor kunnen we stellen dat de 'echte' verschillen tussen de oplossingen 2,27 keer groter zijn dan de verschillen die te wijten zijn aan meetfouten. Bij de tweede toetsvraag ligt de G -waarde amper 0,06 punten lager. De Cronbach's alpha die bij de rangorde van toetsvraag 1 hoort, bedraagt 0,81. We hebben met andere woorden 81% zekerheid dat de positie van de verschillende oplossingen binnen de rangorde gelijk zou blijven indien we de evaluatie zouden overdoen. 19% van de gemeten verschillen is toe te wijzen aan errorvariantie. Ook deze betrouwbaarheidsmaat is quasi gelijk aan die van de rangorde bij toetsvraag 2. Daar bedraagt de Cronbach's alpha 0,80. Aangezien beide waarden boven 0,80 liggen, mogen we spreken over een hoge betrouwbaarheid (Cohen, Manion, & Morrison, 2011). Tabel 1 vat de resultaten samen.

Tabel 1

Betrouwbaarheidsmaten van de rangordes

Betrouwbaarheidsmaat	Toetsvraag 1	Toetsvraag 2
Scheidingscoëfficiënt (G)	2,27	2,21
Cronbach's alpha (α)	0,81	0,80

Inhoudsvaliditeit van de beoordeling

Om uitspraken te kunnen doen over de inhoudsvaliditeit van de beoordeling werden de argumenten van de beoordelaars onder de loep genomen. Allereerst werd een opdeling gemaakt tussen valide en invalide argumenten. Van invalide argumenten was sprake wanneer de voorkeur van de beoordelaar voor één van de twee oplossingen niet werd verduidelijkt, bv. bij argumenten die verwijzen naar de moeilijkheid van de beslissing.

“Er is geen betere oplossing!” [beoordelaar 7, toetsvraag 2, vergelijking 1072]

Bij toetsvraag 1 vonden we 909 valide argumenten (95%). Bij toetsvraag 2 kwamen we 729 valide argumenten (90%). Hierbij merken we wel op dat invalide argumenten regelmatig gepaard gingen met een valide argument.

“Ik kan niet kiezen. Beiden hebben de opdracht fout begrepen.” [beoordelaar 4, toetsvraag 1, vergelijking 213]

Vervolgens werd nagegaan hoeveel van deze valide argumenten gerelateerd zijn aan de competentiebeschrijving die Mayer (1998) vooropstelt bij wiskundig probleemoplossen. Bij toetsvraag 1 waren 76,24% van alle valide argumenten gerelateerd aan de gemeten competentie. Bij toetsvraag 2 lag dit percentage zelfs nog wat hoger, nl. op 82,99%. Opvallend is dat de beoordelaars bij beide toetsvragen duidelijk het meeste belang hechtten aan metacognitieve vaardigheden. Respectievelijk 46,43% en 63,10% van de argumenten verwezen naar metacognitieve vaardigheden van de leerlingen.

“De leerling gebruikt de ene strategie na de andere om dichterbij de oplossing te komen. Hij zoekt verbanden in de mogelijke combinaties en probeert linken te leggen tussen het probleem en eerder geziene leerstof.” [beoordelaar 11, toetsvraag 1, vergelijking 98]

Vervolgens werd er, wat de competentiegerelateerde argumenten betreft, bij beide toetsvragen ook heel wat aandacht besteed aan motivationele aspecten (15,95% & 11,52%).

“Deze leerling zoekt naar een manier om alle mogelijkheden te vinden. Hij toont veel wilskracht voor het vinden van de juiste oplossing.” [beoordelaar 9, toetsvraag 2, vergelijking 1063]

Aan de derde en laatste component van wiskundig probleemoplossen, nl. de domeinspecifieke kennis en vaardigheden van leerlingen konden respectievelijk 13,86% en 8,37% van het totale aantal argumenten gerelateerd worden.

“Aan de linkerkant heeft men het cijferen minder goed onder de knie.” [beoordelaar 4, toetsvraag 1, vergelijking 194]

Zoals eerder aangehaald, konden niet alle valide argumenten worden gerelateerd aan de competentiebeschrijving. Bij toetsvraag 1 bedroeg het aantal niet-competentiegerelateerde argumenten 23,76%. Bij toetsvraag 2 bedroeg dit aantal 17,01%. Bij toetsvraag 1 betekent dit concreet dat bijna een vierde van het totale aantal argumenten dat beoordelaars aanhaalden tijdens het maken van hun keuze, verwijst naar irrelevante zaken. Deze cijfers zijn grotendeels te wijten aan het feit dat beoordelaars heel wat belang hechten aan de juistheid van de (of het aantal) oplossing(en). Bij toetsvraag 1 speelde dit mee voor 20,13%, waardoor deze factor voor beoordelaars belangrijker blijkt te zijn dan de competentiegerelateerde zaken motivatie en domeinspecifieke kennis en vaardigheden.

“De betere oplossing heeft 2 van de 3 vragen juist.” [beoordelaar 10, toetsvraag 1, vergelijking 428]

Bij toetsvraag 2 werden geen concrete deelvragen gesteld, waardoor binnen deze categorie (6,58%) enkel argumenten kwamen te staan die handelden over de mate waarin de leerling het antwoord juist had.

“120 ligt wel dicht bij de juiste oplossing (38), 215 is echt wel veel.” [beoordelaar 12, toetsvraag 2, vergelijking 561]

Deze resultaten zouden kunnen impliceren dat, wanneer een toetsvraag wordt opgesplitst in deelvragen, beoordelaars in verhouding minder rekening houden met competentiegerelateerde zaken. Dit zou tevens de verschillen in eerder vernoemde percentages kunnen verklaren.

Vervolgens merken we op dat bij toetsvraag 2 de niet-competentiegerelateerde categorie ‘Vorm en omvang van de oplossing’ een vrij grote rol heeft gespeeld (10,43%). Een mogelijke verklaring zou kunnen zijn dat, door het gebrek aan concretere deelvragen, de leerlingen meer aangewezen waren op hun vermogen om gestructureerd en netjes te werk te gaan. Dit resulteerde hier en daar in slordige, of net zeer nette oplossingen, die blijkbaar niet onopgemerkt aan de beoordelaars voorbij zijn gegaan.

“Deze leerling heeft overzichtelijk de (meeste) tijdstippen weergegeven die in aanmerking komen. De linkse leerling heeft dit ook geprobeerd, maar voor mij is dit veel minder overzichtelijk.” [beoordelaar 1, toetsvraag 2, vergelijking 792]

Tot slot willen we meegeven dat, net zoals dit bij de invalide argumenten het geval was, niet-competentiegerelateerde argumenten vaak gepaard gingen met competentiegerelateerde argumenten. Dit maakt duidelijk dat eenzelfde beoordelaar regelmatig zijn voorkeur voor een bepaalde leerlingprestatie op zowel relevante, als irrelevante zaken baseert.

“De lln geeft een goed antwoord op vraag 1b en 1c en geeft blijk van goed wiskundig inzicht.”
[beoordelaar 7, toetsvraag 1, vergelijking 806]

Tabel 2 vat alle resultaten samen.

Tabel 2

Relatieve frequentie van de verschillende soorten argumenten

	Toetsvraag 1 (N = 909)	Toetsvraag 2 (N = 729)
Soort argument	Percentage	Percentage
Competentiegerelateerd		
Domeinspecifieke vaardigheden	13,86	8,37
Metacognitieve vaardigheden	46,43	63,10
Motivatie	15,95	11,52
Totaal:	76,24	82,99
Niet-competentiegerelateerd		
Aantal (juiste of foute) antwoorden	20,13	6,58
Vorm en omvang van oplossing	3,52	10,43
Moeilijkheidsgraad van opgeloste vraag	0,11	0,00
Totaal:	23,76	17,01

CONCLUSIE/DISCUSSIE

Dit onderzoek vertrok vanuit de probleemstelling dat leerlingen nog te vaak de schoolbanken verlaten zonder voldoende in staat te zijn om wiskunde op een efficiënte manier toe te passen op nieuwe problemen waarmee ze in de maatschappij, op het werk of bij verdere studies in contact komen. Een mogelijke oorzaak voor deze gebrekkige transfer van opgedane kennis en vaardigheden naar een verdere school- of werkloopbaan werd in eerder onderzoek toegeschreven aan de impact die heersende evaluatievormen in de wiskundeles hebben op het leergedrag van leerlingen (Jones et al., 2014).

Deze studie bouwde verder op voorgaand onderzoek inzake alternatieve evaluatiemethodes. Concreet werd een verkenning gemaakt omtrent de geschiktheid van PV als evaluatiemethode voor wiskundig

probleemoplossen. Het onderzoek vond plaats binnen een Vlaamse secundaire school, bij leerlingen uit het tweede leerjaar, waarbinnen 58 leerlingen voor het vak wiskunde een schriftelijke toets, bestaande uit twee vragen (wiskundige problemen), aflegden die peilde naar hun probleemoplossingsvaardigheden. 14 beoordelaars beoordeelden 116 leerlingenprestaties door in totaal 1106 paarsgewijze vergelijkingen uit te voeren. Bij 1078 van deze vergelijkingen beargumenteerden de beoordelaars hun beslissing.

Om de geschiktheid van de beoordelingsmethode na te gaan, werd stilgestaan bij enerzijds de betrouwbaarheid van de bekomen rangorde (OV1) en anderzijds bij de argumentering van de beoordelaars, om vervolgens uitspraken te doen over de inhoudsvaliditeit (OV2).

De betrouwbaarheid van de rangorde werd nagegaan aan de hand van twee betrouwbaarheidsmaten: de scheidingscoëfficiënt (G) en Cronbach's alpha (α). De resultaten van het onderzoek toonden aan dat deze maten bij zowel toetsvraag 1 ($G = 2,27$; $\alpha = 0,81$) als toetsvraag 2 ($G = 2,21$; $\alpha = 0,80$) hoog genoeg liggen om te kunnen spreken over een rangorde met hoge betrouwbaarheid (Cohen et al., 2011). De veronderstelling van Pollit (2004), dat voor het relatief (met elkaar vergelijkend) beoordelen van leerlingen PV een betrouwbaar evaluatiemiddel kan zijn, wordt door ons onderzoek dan ook bekrachtigd.

De inhoudsvaliditeit werd nagegaan door argumenten van beoordelaars te vergelijken met de competentiebeschrijving van Mayer (1998). Bij toetsvraag 1 haalden beoordelaars in 76,24% van de gevallen competentiegerelateerde criteria aan om hun beslissing te verduidelijken. Bij toetsvraag 2 lag dit percentage nog wat hoger, nl. op 82,99%. Hierbinnen werd in beide gevallen het meeste belang gehecht aan de metacognitieve vaardigheden van leerlingen. Deze bevinding is in lijn met de bemerking die Schoenfeld (1992), Pellegrino, Chudowsky en Glaser (2001) maakten, nl. dat de focus bij de evaluatie van probleemoplossing moet liggen op de beschrijving van metacognitieve acties die leerlingen leveren. Wat de overige competentiegerelateerde argumenten betreft, impliceren de resultaten dat beoordelaars meer belang hechten aan motivatie dan aan domeinspecifieke kennis en vaardigheden. Aangezien onderwijs vaak zo wordt ingericht dat de focus ligt bij cognitieve basisvaardigheden (Mayer, 1998), is dit toch een opmerkelijk resultaat. Beoordelaars lijken, wanneer geen criterialijst voorhanden is, meer belang te hechten aan het tonen van inzet, dan aan het al dan niet kunnen uitvoeren van een deelbewerking. Dit heeft tot gevolg dat PV als evaluatiemethode mogelijks in het voordeel spreekt van gemotiveerde leerlingen. Aan motivatie wordt immers doorgaans binnen traditionele criterialijsten beduidend minder gewicht gekoppeld dan aan kennis en vaardigheden.

Zoals eerder werd verduidelijkt, konden niet alle argumenten worden gerelateerd aan de competentiebeschrijving. Bij toetsvraag 1 verwezen dergelijke argumenten voornamelijk naar de juistheid van de (of het aantal) oplossing(en). Bij toetsvraag 2 was dit veel minder het geval. Toetsvraag 1 was opgebouwd uit verschillende deelvragen. Opmerkelijk hierbij was dat beoordelaars bij deze vraag

zeer regelmatig teruggrepen naar irrelevante, niet-competentiegerelateerde argumenten zoals bijvoorbeeld het verschil in aantal juiste oplossingen tussen leerlingen. Daarnaast merkten we ook op dat niet-competentiegerelateerde argumenten soms handelden rond de vorm en omvang van het antwoord. Dit was vooral merkbaar bij toetsvraag 2. De bedenking kon gemaakt worden dat de vraagstelling het beoordelingsproces van de beoordelaars mogelijks beïnvloed heeft. Deze bemerking kwam voort uit het feit dat toetsvraag 2 geen deelvragen bevat, waardoor leerlingen meer aangewezen waren op hun vermogen om gestructureerd en netjes te werk te gaan. Dat zorgde voor een extra criterium waarop leerlingen konden worden onderscheiden. Dit niet-competentiegerelateerde argument is niet onopgemerkt aan de beoordelaars voorbijgegaan, aangezien zij hieraan meer belang hechtten dan aan bijvoorbeeld domeinspecifieke kennis en vaardigheden.

Het feit dat er toch een beduidend deel van het aantal argumenten niet gerelateerd kon worden aan de competentiebeschrijving impliceert dat beoordelaars doorheen hun beoordelingsproces, naast de gegeven competentiebeschrijving, ook beroep doen op hun eigen expertise, visie en kennis om beslissingen te maken. Hoewel dit an sich volgens Pollit (2012) een sterktepunt is van de beoordelingsmethode, plaatsen we hier toch graag enkele kanttekeningen bij, gezien de mogelijke invloed van de vraagstelling. Zo kan het volgens ons niet de bedoeling zijn dat beoordelaars, vanaf het moment dat een wiskundig probleem uit meerdere deelvragen bestaat, plots gaan teruggrijpen naar irrelevante criteria zoals het aantal juiste of foute antwoorden. Daarnaast bekrachtigen onze bevindingen de veronderstelling dat leerkrachten vaak fragmentarisch te werk gaan door leerlingen scores toe te kennen op basis van (tussen)uitkomsten (Jones et al., 2014).

Wanneer we alles in rekening nemen, kunnen we besluiten dat PV als evaluatiemethode voor wiskundige probleemoplossingsvaardigheden zeker veelbelovend is. Zoals eerder werd aangehaald dient bij kwaliteitsvolle toetsing sprake te zijn van een goede validiteit en betrouwbaarheid (Dochy & Gijbels, 2009). De methode blijkt hieraan tegemoet te komen. Rangordes met hoge betrouwbaarheid konden worden bereikt en de argumenten, die de beoordelingswijze van beoordelaars verduidelijkten, bestonden voor respectievelijk 76% en 83% uit competentiegerelateerde zaken.

De verschillen in argumentaties en bijhorende percentages tussen de twee toetsvragen wekken echter wel de kritische gedachte op dat de inhoudsvaliditeit niet enkel afhangt van de evaluatiemethode. De mate waarin beoordelaars zich laten leiden door niet-competentiegerelateerde zaken lijkt vatbaar te zijn voor de soort vraagstelling die wordt gehanteerd. Deze bevinding levert ons het inzicht op dat de vraagstelling mogelijks bepalend is voor de inhoudsvaliditeit van de beoordeling.

Binnen dit onderzoek werd geopteerd voor een vrij pragmatische aanpak wat betreft de toetsopstelling. Omwille van onze bedenkingen omtrent de rol van de vraagstelling, lijkt het ons nuttig om in vervolgonderzoek eerst dieper in te gaan op het bereiken van een consensus wat betreft de voorwaarden en vormvereisten van een 'ideale' wiskundige probleemoplossingstoets. Hierbij zouden de theoretici in

het veld hun steentje kunnen bijdragen. We zien dit als noodzaak om in vervolgonderzoek meer eenduidigheid te creëren omtrent de rol die PV als evaluatiemethode aan zich speelt inzake inhoudsvaliditeit. Deze basis voor vervolgonderzoek willen we nog aanvullen met verdere aanbevelingen.

Aangezien er binnen onze studie geopteerd werd voor een exploratieve benadering, ontbrak een referentiekader waartegen de empirische antwoorden op onze onderzoeksvragen konden worden afgezet. Vandaar dat het ons nuttig lijkt om in de toekomst de vergelijking te maken tussen PV en andere assessmentmethoden. Dit zou mogelijks relatieve antwoorden kunnen geven op vragen rond aspecten als betrouwbaarheid en validiteit van de evaluatiemethode. Daarnaast lijkt het ons ook nuttig om op termijn de effecten van PV op het leergedrag van leerlingen na te gaan. Deze leergedragingen spelen immers een belangrijke rol bij het al dan niet bereiken van transfer (Verschaffel et al., 2009). Hiervoor zou bijvoorbeeld longitudinaal onderzoek kunnen plaatsvinden om na te gaan of PV gepaste leergedragingen voor wiskundig probleemoplossen uitlokt. Tot slot willen we meegeven dat de data die verzameld werd binnen ons onderzoek nog verder uitgediept kan worden. De data-analyse is beschrijvend van aard gebleven, maar verdere uitdieping was mogelijk geweest. Zo had het bijvoorbeeld interessant geweest om na te gaan of de argumentatie van beoordelaars verschillend was bij winnaars en verliezers. Werden bijvoorbeeld de winnaars op een even inhoudsvalide wijze beoordeeld als de verliezers? Daarnaast had de samenstelling van argumenten ook nog verder onderzocht kunnen worden. Zijn er bijvoorbeeld verschillen merkbaar tussen beoordelaars wat betreft het aantal (al dan niet competentiegerelateerde) zaken waar ze tijdens het maken van een vergelijking rekening mee houden? Door stil te staan bij dergelijke zaken, zou men nog meer inzicht kunnen verwerven in het paarsgewijs vergelijkingsproces van beoordelaars. Er zit hoe dan ook nog een rijkheid van data uit deze studie klaar om uitgepakt te worden.

Om bij enig vervolgonderzoek effectief gebruik te kunnen maken van resultaten uit onze studie, lichten we om te besluiten nog kort enkele beperkingen en bedenkingen van het geleverde onderzoek toe.

Allereerst merken we op dat uitspraken omtrent de betrouwbaarheid van bekomen rangordes gebaseerd werden op betrouwbaarheidsmaten die met enige voorzichtigheid dienen geïnterpreteerd te worden. Hoewel ze in eerder onderzoek zeer frequent werden gebruikt, worden alsmaar meer vragen gesteld naar de geschiktheid ervan binnen de context van PV. Vervolgonderzoek zal hieromtrent duidelijkheid moeten scheppen. Wat het validiteitsonderzoek betreft, dient men voor ogen te houden dat het gekozen theoretische kader van Mayer (1998) bepalend is geweest voor de resultaten. De wiskundige probleemoplossingsvaardigheid wordt in de literatuur door tal van onderzoekers beschreven. Hierdoor hadden de argumenten van beoordelaars binnen dit onderzoek op verschillende wijzen kunnen worden geïnterpreteerd. Tot slot willen we benadrukken dat onze steekproef te beperkt was om grote generalisatie van resultaten en bevindingen te pretenderen.

REFERENTIES

- ACT. (2006). *Ready for College and Ready for Work: Same or Different?* Iowa City: ACT Inc.
- Boyatzis, R.E. (1998). *Transforming qualitative information: Thematic analysis and code development*. London: SAGE Publications.
- Brown, A.L. & Cocking, R.R. (1999). *How people learn: Brain, mind, experience, and school*. Washington: National Academy Press.
- Chi, M.T.H., Glaser, R., & Farr, M.J. (Eds.). (1988). *The Nature of Expertise*. Hillsdale: Erlbaum.
- Cohen, L., Manion, L., & Morrison, K. (2011). *Research Methods in Education (7e druk)*. Abingdon: Routledge.
- D'Arcy, J. (1997). *Comparability studies between modular and non-modular syllabuses in GCE Advanced level biology, English literature and mathematics in the 1996 summer examinations*. Belfast: GCSE.
- Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford: Stanford University, Stanford Center for Opportunity Policy in Education.
- De Meyer, I., Warlop, N., & Van Camp, S. (2012). Probleemoplossend vermogen bij 15-jarigen: Vlaamse resultaten van PISA2012. Opgeroepen op september 14, 2014, van http://www.ond.vlaanderen.be/obpwo/links/pisa/Vlaams_rapport_problemsolving.pdf.
- DfE. (2011). *Independent Evaluation of the Pilot of the Linked Pair of GCSEs in Mathematics - First Interim Report (No. DFE-RR181)*. London: Department for Education.
- Dochy, F., & Gijbels, D. (2009). Evaluatie. In S. Janssens (Eds.), *Leren en Onderwijzen*. Leuven: Acco.
- Donche, V. (2014). *Inleiding in Methoden en Technieken*. Antwerpen: Universiteit Antwerpen.
- Dossey, J.A., McCrone, S.A., & O'Sullivan, C. (2006). *Problem Solving in the PISA and TIMSS 2003 Assessments*. Washington: National Center Press.
- Dousma, T., Horsten, A., & Brants, J. (1995). *Tentamineren*. Groningen: Wolters-Noordhoff.
- English, L.D. (Eds.). (2002). *Handbook of International Research in Mathematics Education*. Mahwah: Lawrence Erlbaum.
- Ericsson, K.A., & Smith, J. (Eds.). (1991). *Toward a General Theory of Expertise*. Cambridge: Cambridge University Press.
- Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical software*, 12(1), 1-12.
- Flavell, J.H. (1979). Metacognitive and cognitive monitoring: A new area of cognitive developmental inquiry. *American Psychologist*, 34, 906-911.
- Funke, J. (2010). Complex problem solving: a case for complex cognition? *Cognitive processing*, 11, 133-142.
- Hacker, D.J., & Dunlosky, J. (2003). Not all metacognition is created equal. *New Directions For Teaching And Learning*, 95, 73-79.
- Jones, I., & Alcock, L. (2013). Peer assessment without assessment criteria. *Studies in Higher Education*, 39(10), 1774-1787.

- Jones, I., Inglis, M., Gilmore, C., & Hodgen, J. (2013). Measuring Conceptual Understanding: The Case of Fractions. In A.M. Lindmeier & A. Heinze (Eds.), *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education*. Kiel: PME.
- Jones, I., Swan, M., & Pollit, A. (2014). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, 13, 151-177.
- Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Martin, F., Davies, D., et al. (2009) *E-scape portfolio assessment: phase 3 report*. London: Technology Education Research Unit Goldsmiths College University of London.
- Mayer, R.E. (1985). Mathematical ability. In R.J. Sternberg (Eds.), *Human Abilities: An Information Processing Approach*. New York: Freeman.
- Mayer, R.E. (1992). *Thinking, Problem Solving, Cognition: Second Edition*. New York: Freeman.
- Mayer, R.E. (1998). Cognitive, Metacognitive and Motivational Aspects of Problem Solving. *Instructional Science*, 26, 49-63.
- Mayer, R.E. (2003). Mathematical problem solving. In J.M. Royer (Eds.), *Mathematical Cognition*. Greenwich: Infoage Publishing.
- Mayer, R.E., & Wittrock, M.C. (2006). Problem solving and transfer. In D. Berliner & R. Calfee (Eds.), *Handbook of Educational Psychology*. New York: Macmillan.
- Montague, M. (1992). The Effects of Cognitive and Metacognitive Strategy Instruction on the Mathematical Problem Solving of Middle School Students with Learning Disabilities. *Journal of Learning Disabilities*, 25, 230-248.
- Pellegrino, J.W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington: National Academy Press.
- Pollitt, A. (2004). *Let's stop marking exams*. Philadelphia: UCLES.
- Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Policy & Practice*, 19, 281-300.
- Pollitt, A., & Murray, N.J. (1993). *What raters really pay attention to*. Cambridge: Cambridge University Press.
- Polya, G. (1988). *How To Solve It*. New Jersey: Princeton University Press.
- Reitman, W.R. (1965). *Cognition and thought*. New York: Wiley.
- Royer, J.M., & Garofoli, L.M. (2005). Cognitive contribution to sex differences. In A.M. Gallagher & J.C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach*. Cambridge: Cambridge University Press.
- Rubin, D.L. (1996). A preface relating alternative assessment, test fairness, and assessment utility to communication. In S. Morreale & P. Backlund (Eds.), *Large scale assessment of oral communication (2nd edition)*. Annandale: Speech Communication Association.
- Sadler, D.R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment and Evaluation in Higher Education*, 34, 159-179.
- Schoenfeld, A.H. (1992). Learning to Think Mathematically: Problem Solving, Metacognition, and Sense Making in Mathematics. In D.A. Grouws (Eds.), *Handbook of Research on Mathematics Teaching and Learning*, 334-370. New York: Macmillan.

Secundair onderwijs–Wiskunde–Uitgangspunten. (z.d.). Opgeroepen op augustus 4, 2014, van <http://www.ond.vlaanderen.be/curriculum/secundair-onderwijs/eerste-graad/vakgebonden/a-stroom/wiskunde/uitgangspunten.htm>.

Smith, M.U. (Eds.). (1991). *Toward a Unified Theory of Problem Solving: Views from the Content Domains*. Hillsdale: Erlbaum.

Szetela, W., & Nicol, C. (1992). Evaluating Problem Solving in Mathematics. *Educational Leadership*, 49(8), 42-45.

Thurstone, L.L. (1927). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology*, 21, 384-400.

Verschaffel, L., Van Dooren, W., Elen, J., & Clarebout, G. (2009). Leren. In S. Janssens (Eds.), *Leren en Onderwijzen*. Leuven: Acco.

Victor, A.M. (2004). *The effects of metacognitive instruction on the planning and academic achievement of first and second grade children*. Chicago: IIP.

Whitehouse, C. (2012). *Testing the validity of judgements about geography essays using the Adaptive Comparative Judgement method*. Manchester: AQA Centre for Education Research and Policy.

APPENDICES

Appendix A: Toets probleemoplossen

Vraag 1:

- a) Vervolledig onderstaand schema door de cijfers 1, 2, 3, 4, 5, 6, 7, 8 en 9 in te vullen, zodanig dat de **som van de drie getallen gelijk is aan 999**. Elk cijfer mag slechts eenmaal worden gebruikt!

$$\begin{array}{r} \square \square \square \\ \square \square \square \\ \square \square \square \\ + \\ \hline 9 \ 9 \ 9 \end{array}$$

- b) Bestaat er slechts één juiste oplossing voor dit probleem? Leg uit.
c) Bestaat er een mogelijke oplossing waarbij het cijfer 1 nergens op de plaats van de honderdtallen staat? Leg uit.

Oplossing vraag 1:

Vraag 2:

Hoeveel keer is de som van de cijfers op een digitale klok gelijk aan zes tussen 00:00 en 12:00? Probeer een oplossing te vinden zonder elk tijdstip te moeten opschrijven.

Oplossing Vraag 2:

Definitief antwoord: keer

Appendix B: Essentie wiskundig probleemoplossen (statement)

"Een wiskundig probleemoplosser is een persoon die op een wiskundige manier met informatie aan de slag gaat om een bepaald doel te bereiken, maar niet onmiddellijk inziet hoe hij dit moet doen."

Het probleemoplossingsproces bestaat uit verschillende componenten. Allereerst zal hij tijdens zijn zoektocht naar de oplossing gebruik maken van **domeinspecifieke basisvaardigheden** zoals hoofdrekenen, cijferen, vergelijking opstellen, tekening/schets maken... Deze vaardigheden an sich bieden echter geen garantie op succes. Een goede probleemoplosser dient in staat te zijn om probleemsituaties correct in te schatten, gepaste oplossingsstrategieën te selecteren, zijn oplossingsproces effectief te organiseren en voldoende te monitoren en te controleren (**metacognitieve vaardigheden**). Tot slot typeren succesvolle probleemoplossers zich door hun hoge mate van bereidwilligheid om zich in te laten met het wiskundig probleem. Ze zijn **gemotiveerd** en geven niet op wanneer ze op hun weg naar de oplossing even vastlopen.

Gebaseerd op voorgaande verklaring, welke van de twee oplossingen getuigt volgens u het meest van goed wiskundig probleemoplossen?

Appendix C: Codeerschema

Relatieve frequentie van de verschillende soorten argumenten

	Toetsvraag 1 (N = 909)	Toetsvraag 2 (N = 729)	
Soort argument	Percentage	Percentage	Voorbeeldquote
Competentiegerelateerd			
Domeinspecifieke kennis en vaardigheden	13,86	8,37	
Antwoordvorming	11,00	6,59	“Hij motiveert de antwoorden ook heel erg duidelijk en gedetailleerd.”
Rekenvaardigheid	1,87	1,51	“De rechtse leerling maakt een rekenfout ($0 + 0 + 3 + 0 = 6$) die cruciaal is.”
Wiskundig taalgebruik	0,44	0,00	“Leerling rechts is een beetje in de war met de term 'honderdtallen'.”
Algemeen	0,55	0,27	“De linker oplossing toont zelf weinig basisvaardigheden.”
Metacognitieve vaardigheden	46,43	63,10	“De rechter leerling dacht al dat hij het systeem had, maar had verder moeten controleren.”
Motivatie	15,95	11,52	“De leerling toont motivatie om het probleem op te lossen.”
Totaal:	76,24	82,99	
Niet-competentiegerelateerd			
Aantal (juiste of foute) antwoorden	20,13	6,58	“Rechter oplossing is fout voor b en c.”
Vorm en omvang van oplossing	3,52	10,43	“Aan de rechterkant is alles minder ordelijk genoteerd.”
Moeilijkheidsgraad van opgeloste vraag	0,11	0,00	“Een goede verklaring vinden voor vraag 1c vond ik net iets moeilijker dan bij vraag 1b.”
Totaal:	23,76	17,01	