Dit project is gefinancierd met steun van de Europese Commissie. Deze publicatie geeft uitsluitend de mening van de auteur weer, en de Commissie kan niet verantwoordelijk worden gehouden voor enig gebruik dat van de daarin opgenomen informatie kan worden gemaakt.

Professionele standaarden voor formatieve beoordeling: een op bewijs gebaseerde benadering

van het project

Formatieve beoordeling bevorderen: van theorie naar beleid en praktijk (FORMAS)

Onderdeel van de

Erasmus+-programma Kernactie 3, Ondersteuning van beleidshervorming, Toekomstgerichte samenwerkingsprojecten

Leonidas Kyriakides, *Departement Opleidings, Universiteit van Cyprus, Cyprus*

Margarita Christoforidou, *Departement Opleidings, Universiteit van Cyprus, Cyprus*

Evi Charalambous, *Departement Opleidings, Universiteit van Cyprus, Cyprus*

Andria Dimosthenous, *Departement Opleidings, Universiteit van Cyprus, Cyprus*

Elena Kokkinou, *Departement Opleidings, Universiteit van Cyprus, Cyprus*

Anastasia Panayiotou, *Departement Opleidings, Universiteit van Cyprus, Cyprus*

Bert Creemers, *Faculteit Gedrags- en Maatschappijwetenschappen, Rijksuniversiteit Groningen, Nederland*

Theodossios Zachariades, *Departement Wiskunde, Nationale en Kapodistrian Universiteit van Athene, Griekenland*

Giorgos Psycharis, *Departement Wiskunde, Nationale en Kapodistrian Universiteit van Athene, Griekenland*

Despina Potari, *Departement Wiskunde, Nationale en Kapodistrian Universiteit van Athene, Griekenland*

Chrissavgi Triantafillou, *Departement Wiskunde, Nationale en Kapodistrian Universiteit van Athene, Griekenland*

Adrie Visscher, *Faculteit Gedrags- en Maatschappijwetenschappen, Universiteit Twente, Nederland*

Jitske de Vries, *Faculteit Gedrags- en Maatschappijwetenschappen, Universiteit Twente, Nederland*

Peter Van Petegem, *Departement Opleidings- en Onderwijswetenschappen, Universiteit Antwerpen, België*

Tine Mombaers, *Departement Opleidings- en Onderwijswetenschappen, Universiteit Antwerpen, België*

Roos Van Gasse, *Departement Opleidings- en Onderwijswetenschappen, Universiteit Antwerpen, België*

Ioannis Ioannou, *Ministerie van Onderwijs en Cultuur, Cyprus*

Liselotte Van de Perre, *Vlaams Agentschap voor Hoger Onderwijs, Volwassenenonderwijs, Kwalificaties en Studieleningen, België*

Christos Milionis, *Ministerie van Onderwijs, Onderzoek en Religieuze Zaken, Griekenland*

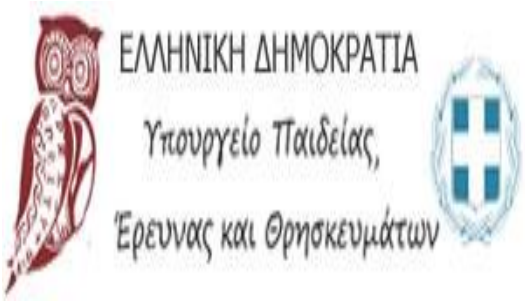
Dionysios Lamprinidis, *Ministerie van Onderwijs, Onderzoek en Religieuze Zaken, Griekenland*

Gijske Mels, *Ministerie van Onderwijs, Cultuur en Wetenschap, Nederland*


University
of Cyprus


Υ.Π.Π.
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ, ΠΟΛΙΤΙΣΜΟΥ,
ΑΘΛΗΤΙΣΜΟΥ ΚΑΙ ΝΕΟΛΑΙΑΣ


HELLENIC REPUBLIC
National and Kapodistrian
University of Athens


ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Παιδείας,
Έρευνας και Θρησκευμάτων


UNIVERSITY
OF TWENTE.


Ministry of Education, Culture and
Science of the Netherlands


University
of Antwerp


Vlaanderen
is onderwijs & vorming

Professionele standaarden voor formatieve beoordeling: een op bewijs gebaseerde benadering

Een van de hoofddoelen van het Erasmus+KA3-project getiteld ‘’Formatieve beoordeling bevorderen: van theorie naar beleid en praktijk (FORMAS)’’ is het genereren van professionele normen voor formatieve beoordeling. Dit verslag presenteert de normen die zijn ontwikkeld op basis van de bevindingen van het FORMAS-project en geeft suggesties over hoe deze kunnen worden gebruikt om de ontwikkeling van onderwijsbeleid te ondersteunen, dat de effectieve implementatie van formatieve beoordeling bevordert.

Eerst wordt de grondgedachte achter de ontwikkeling van de professionele standaarden beschreven. Vervolgens wordt het theoretisch kader gepresenteerd op basis waarvan de normen zijn ontwikkeld en de gebruikte methodologie. Vervolgens worden de professionele standaarden geschetst, die een reeks richtlijnen bieden die als leidraad kunnen dienen voor een effectieve beoordelingspraktijk. Tenslotte worden implicaties voor het onderwijsbeleid getrokken.

1. Grondgedachte

Onderzoekers hebben al lang erkend dat beoordelingsvaardigheden een cruciaal element zijn van een effectieve onderwijspraktijk (Christoforidou et al., 2014; Hopfenbeck, 2018; Panadero et al., 2019). De toenemende erkenning van de behoefte aan docenten met beoordelingsvaardigheden, heeft geleid tot de ontwikkeling van verschillende lijsten met basisbeoordelingscompetenties (bijv. AFT, NCME, & NEA, 1990; Brookhart, 2011; Schafer, 1991; Stiggins, 1995, 1999). Deze lijsten beschrijven beoordelingscompetenties in relatie tot algemene standaarden van beoordelingspraktijken met als doel bij te dragen aan het verhogen van de kwaliteit van de beoordeling van leerlingen en het bevorderen van leren. De tot dusver voorgestelde lijsten beschrijven echter de beoordelingscompetenties in relatie tot de algemene standaarden van de beoordelingspraktijk, zonder details te verstrekken over de specifieke betrokken vaardigheden Bovendien zijn deze lijsten niet in verband gebracht met een specifieke theoretische achtergrond en is er geen empirisch bewijs geleverd dat hun geldigheid ondersteunt. Daarnaast worden recente opvattingen over formatieve toetsing niet effectief aangepakt (Brookhart, 2011). In deze context is het doel van dit verslag om het uitgebreide raamwerk voor het meten van de beoordelingsvaardigheden van docenten, ontwikkeld in het kader van het FORMAS-project (zie Verslag 2: Een uitgebreid raamwerk voor het meten van de beoordelingsvaardigheden van docenten), te gebruiken om een reeks professionele normen en gerelateerde richtlijnen te ontwikkelen. Deze richtlijnen omschrijven een effectieve beoordelingspraktijk, rekening houdend met de dynamische aard van het onderwijs en recente opvattingen over leerlingbeoordeling.

2. Theoretisch kader en methodologie voor het meten van beoordelingsvaardigheden en het genereren van normen bij formatieve beoordeling

Theoretisch kader

Een van de belangrijkste tekortkomingen van eerdere pogingen om professionele standaarden in beoordeling te identificeren, is het gebrek aan theoretische achtergrond op basis waarvan de standaarden zijn ontwikkeld. De professionele standaarden die in deze deliverable worden voorgesteld, zijn afgeleid van het theoretische kader dat is ontwikkeld in het kader van het FORMAS-project voor het definiëren en meten van beoordelingsvaardigheden van docenten (zie *Verslag 2: Een uitgebreid raamwerk voor het meten van beoordelingsvaardigheden van docenten*). Het raamwerk dat in het FORMAS-project is ontwikkeld en gebruikt, houdt rekening met de dynamische aard van het onderwijs en daardoor worden de vaardigheden die bij elke fase van het beoordelingsproces horen, onderzocht. Daarnaast worden beoordelingsvaardigheden gedefinieerd en gemeten in relatie tot het vermogen van docenten om specifieke beoordelingstechnieken te gebruiken om verschillende leerresultaten van leerlingen te meten. Het raamwerk stelt ons daarom in staat om het beoordelingsgedrag van docenten te onderzoeken en specifieke vaardigheden te identificeren die betrokken zijn bij het beoordelen van het leren van leerlingen en hoe deze het leren van leerlingen beïnvloeden. Het FORMAS-raamwerk onderzoekt beoordeling en kijkt naar drie hoofdaspecten van de beoordeling van leerlingen: a) de fasen van beoordeling, b) het gebruik van verschillende beoordelingstechnieken en c) de vijf meetdimensies die zijn afgeleid van eerdere effectiviteitsonderzoeken. Elke beoordelingsfase wordt gedefinieerd op basis van de beoordelingskennis en -vaardigheden die betrokken zijn in de vijf meetdimensies en in relatie staan tot de meest voorkomende beoordelingstechnieken. Voor de lezer wordt hierna een korte beschrijving van de drie aspecten gegeven. Voor meer details over het raamwerk en de specifieke aspecten die zijn onderzocht, worden lezers aangemoedigd om *Verslag 2: Een uitgebreid raamwerk voor het meten van beoordelingsvaardigheden van docenten* te bestuderen.

Het eerste aspect van het voorgestelde kader verwijst naar het beoordelingsproces en identificeert in het bijzonder vijf hoofdfasen die op een alomvattende manier de vaardigheden beschrijven die betrokken zijn bij het proces van beoordelingsontwerp en -praktijk. Deze fasen zijn gebaseerd op de veronderstelling dat effectieve leraren ervoor moeten zorgen dat: (i) geschikte beoordelingsinstrumenten worden gebruikt om geldige en betrouwbare gegevens te verzamelen; (ii) passende procedures bij het beheer van deze instrumenten worden gevolgd; (iii) gegevens die uit de beoordeling voortkomen op een efficiënte manier worden vastgelegd en zonder belangrijke informatie te verliezen; (iv) beoordelingsresultaten worden geanalyseerd, geïnterpreteerd en gebruikt op manieren die het leren van leerlingen kunnen bevorderen; en (v) beoordelingsresultaten worden gerapporteerd aan alle beoogde gebruikers, inclusief ouders en leerlingen, om hen te helpen beslissingen te nemen over hoe de leerresultaten van leerlingen kunnen worden verbeterd. Zonder het sequentiële karakter van de vijf fasen die betrokken zijn bij het proces van ontwerp en uitvoering van beoordeling te verwaarlozen, beschouwt dit raamwerk alle fasen als onderling verbonden en uitwisselbaar en niet als

een stapsgewijs model. Effectieve beoordeling vereist de ontwikkeling van de vaardigheden die nodig zijn in alle vijf fasen.

Het tweede onderzochte aspect heeft betrekking op de evaluatiemethoden die worden gebruikt om het leren van leerlingen te beoordelen (d.w.z. beoordelingstechnieken), aangezien deze een belangrijke rol spelen bij het waarborgen van de kwaliteit en effectiviteit van de beoordeling. Het raamwerk kijkt naar beoordelingsvaardigheden in relatie tot de gebruikte beoordelingstechnieken door rekening te houden met twee belangrijke beslissingen die van invloed zijn op hun selectie: a) de wijze van reageren en b) wie de beoordeling uitvoert. Dit stelt ons in staat te kijken naar beoordelingstechnieken die verschillende manieren van reageren vereisen (dwz schriftelijk, mondeling, uitvoering), maar houdt er tegelijkertijd rekening mee dat deze technieken niet alleen door de docent kunnen worden gebruikt, maar ook door de leerlingen zelf in de vorm van zelf-, peer- en co-beoordeling. Op deze manier sluit het raamwerk aan bij de huidige opvattingen over toetsing als een proces dat de betrokkenheid van leerlingen bij het leerproces vergroot.

Het derde aspect dat wordt gebruikt om de vaardigheden van docenten bij beoordeling te onderzoeken, is gebaseerd op methodologische en theoretische ontwikkelingen op het gebied van onderzoek naar onderwijseffectiviteit (EER). Specifiek worden de volgende vijf dimensies, voorgesteld door het dynamische model van educatieve effectiviteit (Creemers & Kyriakides, 2008) beschouwd: (a) frequentie, (b) focus, (c) fase, (d) kwaliteit en (e) differentiatie. Deze dimensies helpen ons om het functioneren van elk kenmerk van effectieve leraren beter te beschrijven (zie Kyriakides et al., 2021). Frequentie is een kwantitatieve manier om het functioneren van elk effectiviteitskenmerk te meten, terwijl de andere vier dimensies kwalitatieve aspecten van het functioneren van een leerkracht (in dit geval beoordeling) onderzoeken. Beoordeling als een multidimensionale constructie beschouwen, geeft niet alleen een beter beeld van wat leraren effectiever maakt bij het beoordelen van leerlingen, maar helpt ook om meer specifieke strategieën te ontwikkelen voor het verbeteren van de beoordelingspraktijk.

Daarom worden op basis van het FORMAS-raamwerk beoordelingsvaardigheden gedefinieerd en gemeten in relatie tot het vermogen van docenten om specifieke beoordelingstechnieken te gebruiken in de verschillende fasen van het beoordelingsproces. Er wordt rekening gehouden met traditionele en alternatieve beoordelingstechnieken op basis van de manier van reageren en wie de beoordeling uitvoert. Bovendien maakt het aangenomen meetkader de mogelijk rekening te houden met zowel kwantitatieve als kwalitatieve kenmerken van het beoordelingsproces. Het raamwerk is gebruikt om een onderzoeksinstrument te ontwerpen (zie *Verslag 3: De lerarenvragenlijst voor het meten van de beoordelingsvaardigheden van leraren*) om de beoordelingsvaardigheden van leraren te meten. Op basis van de verzamelde gegevens zijn professionele normen voor effectieve beoordeling ontwikkeld. De methodologie die wordt gebruikt om beoordelingsvaardigheden te onderzoeken en professionele normen te identificeren, wordt hierna gepresenteerd.

Methodologie

Door rekening te houden met het hierboven gepresenteerde theoretische kader, werd een lerarenvragenlijst ontwikkeld (zie Verslag 3). De vragenlijst bestond uit 119 items, ontworpen om de beoordelingsvaardigheden van leraren in wiskunde te meten voor de drie aspecten van het theoretische kader (d.w.z. beoordelingsfasen, beoordelingstechnieken, meetdimensies). Elke beoordelingstechniek is onderzocht in relatie tot de vijf fasen van het beoordelingsproces en voor elke fase van het beoordelingsproces is elk van de vijf meetdimensies toegepast. Er is gebruik gemaakt van een vijfpunts Likert-schaal en docenten is gevraagd aan te geven in hoeverre ze zich op een bepaalde manier gedragen tijdens het wiskundeonderwijs in hun klas (voor meer informatie over de docentenvragenlijst zie *Verslag 3: De lerarenvragenlijst voor het meten van de beoordelingsvaardigheden van leraren*).

Om de validiteit van de vragenlijst te onderzoeken, werd in juni 2019 een onderzoek uitgevoerd in de vier landen die deelnamen aan het FORMAS-project (dwz Cyprus, Griekenland, Nederland en België). De data (n= 574) werd geanalyseerd met behulp van het uitgebreide logistieke model van Rasch (Andrich, 1988) om te bepalen in hoeverre de gemeten beoordelingsvaardigheden konden worden herleid tot een algemene eendimensionale schaal (zie Verslag 3 voor meer informatie over het validatieonderzoek). Het uitgebreide logistieke model van Rasch werd toegepast op de hele steekproef van leraren en alle 119 items die betrekking hadden op hun beoordelingsvaardigheden samen, met behulp van het computerprogramma Quest (Adams & Khoo, 1996). Dit model (Andersen, 1977; Wright, 1985) is een uitbreiding van het dichotome Rasch-model naar het geval waarin items meer dan twee antwoordcategorieën hebben en het werd daarom gebruikt om de gegevens te analyseren die voortkwamen uit de antwoorden van leraren op elk vragenlijstitem. Aangezien elk item vijf antwoorden heeft, kan het worden gemodelleerd als vier drempels. Elke drempel heeft zijn eigen moeilijkheidsschatting, en deze schatting wordt gemodelleerd als de drempel waarbij een persoon 50% kans heeft om de ene categorie boven de andere te kiezen (Andersen, 1977). Deze drempels worden berekend in log odds (ook wel logits genoemd) en moeten worden geordend om de afnemende waarschijnlijkheid van elk beoordelingsgedrag weer te geven. Drempels die niet monotoon stijgen, worden als ongeordend beschouwd. De grootten van de afstanden tussen de drempelschattingen zijn belangrijk. Drempelafstanden moeten aangeven dat elke stap een afzonderlijke positie op de variabele definieert en daardoor mogen ze niet te dicht bij elkaar en niet te ver uit elkaar liggen op de logitschaal (Bond & Fox, 2001). Specifiek geven richtlijnen aan dat drempels met ten minste 1,4 logits moeten worden verhoogd (d.w.z. om onderscheid tussen categorieën te laten zien), maar niet meer dan 5 logits (d.w.z. om grote hiaten in de variabele te voorkomen; Linacre, 1999).

Op basis van de analyse van de gegevens zijn vier items verwijderd om een betere aansluiting op het model te krijgen. De herziene versie van de vragenlijst bevatte in totaal 115 items. Tabel 1 illustreert de schaal voor de 115 metingen van beoordelingsvaardigheden met itemmoeilijkheden en docentmetingen die op dezelfde schaal zijn gekalibreerd. De itemdrempelwaarden bleken geordend te zijn van laag naar hoog, wat aangeeft dat de docenten consistent antwoordden met het geordende

antwoordformaat van onze Likert-schaal. De drempelafstanden bleken ook te variëren van 1,8 tot 2,9 logits. Tabel 1 laat zien dat de 115 items van de vragenlijst, die de beoordelingsvaardigheden van leraren meten, goed passen bij het meetmodel, wat wijst op een sterke overeenstemming tussen de 574 leraren op verschillende posities op de schaal, over alle 115 items. Bovendien zijn de vragenlijstitems goed gericht tegen de maatregelen van de leraren, aangezien de scores van leraren variëren van -2,54 tot 2,01 logits en itemmoeilijkheden variëren van -2,34 tot 1,89 logits.

Het Saltus-model werd vervolgens gebruikt om de ontwikkelingsstructuur van de vaardigheden te specificeren. De resultaten boden ondersteuning voor de schaal en ontwikkelingsstructuur van de vaardigheden van leraren bij het beoordelen. In het bijzonder werd vastgesteld dat beoordelingsvaardigheden kunnen worden gegroepeerd in *drie typen/stadia van beoordelingsgedrag*, die op een kenmerkende manier worden onderscheiden en geleidelijk overgaan van gemakkelijkere naar meer geavanceerde vaardigheden. De Saltus-oplossing bleek beter bij de werkelijke gegevens te passen dan het Rasch-model, en biedt een statistisch significante verbetering ten opzichte van het Rasch-model, dat gelijk is aan 391,6 chi-kwadraat-eenheden ten koste van 12 extra parameters (dwz, 4 t-waarden, drie gemiddelden, drie standaarddeviaties en twee onafhankelijke verhoudingen). Tabel 1 geeft de moeilijkheidsgraad van het item weer voor docenten van niveau 1 (d.w.z. kolom 3) en de moeilijkheidsgraad binnen het stadium (d.w.z. kolommen 4 en 5). De schattingen van de Saltus-parameter (d.w.z. t-waarden) worden onderaan de tabel weergegeven.

Tabel 1. Rasch en Saltus parameterschattingen voor de beoordelingsvaardigheden van docenten

Beoordelingsvaardigheden van docenten	**Rasch**	**Impliciete moeilijkheidsgraad binnen het stadium (Saltus)**		
	Alle	**Level 1**	**Level 2**	**Level 3**
Freq Constructie Geschreven	-2,34	-3,27	-3,27	-3,27
Stadium Constructie Geschreven	-2,32	-3,25	-3,25	-3,25
Freq Administratie Geschreven	-2,31	-3,22	-3,22	-3,22
Stadium Administratie Geschreven	-2,29	-3,18	-3,18	-3,18
Focus Constructie Geschreven	-2,28	-3,16	-3,16	-3,16
Freq Verslaglegging Geschreven	-2,27	-3,14	-3,14	-3,14
Freq Analyse Geschreven	-2,26	-3,11	-3,11	-3,11
Stadium Verslaglegging Geschreven	-2,25	-3,09	-3,09	-3,09
Freq Verslag Geschreven	-2,24	-3,06	-3,06	-3,06
Stadium Verslag Geschreven	-2,23	-3,05	-3,05	-3,05
Freq Administratie Mondeling	-2,22	-3,01	-3,01	-3,01
Focus Verslag Geschreven	-2,21	-2,98	-2,98	-2,98
Stadium Analyse Geschreven	-2,19	-2,95	-2,95	-2,95
Freq Constructie Mondeling	-2,17	-2,93	-2,93	-2,93
Freq Verslag Mondeling	-2,15	-2,91	-2,91	-2,91
Stadium Verslag Mondeling	-2,14	-2,88	-2,88	-2,88
Focus Administratie Geschreven	-2,13	-2,85	-2,85	-2,85
Kwaliteit Constructie Geschreven	-1,33	-0,83	-2,77	-2,79
Kwaliteit Administratie Geschreven	-1,31	-0,81	-2,73	-2,76
Freq Verslaglegging Mondeling	-1,28	-0,78	-2,75	-2,71

Freq Constructie Performance	-1,26	-0,76	-2,69	-2,73
Freq Administratie Performance	-1,25	-0,75	-2,66	-2,68
Focus Verslaglegging Geschreven	-1,24	-0,73	-2,61	-2,66
Kwaliteit Verslag Geschreven	-1,22	-0,71	-2,62	-2,64
Kwaliteit Verslaglegging Geschreven	-1,22	-0,69	-2,59	-2,58
Stadium administratie performance	-1,21	-0,67	-2,57	-2,55
Focus analyse geschreven	-1,19	-0,66	-2,55	-2,51
Freq analyse mondeling	-1,18	-0,62	-2,51	-2,49
Kwaliteit analyse geschreven	-1,13	-0,61	-2,48	-2,47
Stadium administratie mondeling	-1,09	-0,59	-2,45	-2,46
Stadium analyse mondeling	-1,08	-0,57	-2,41	-2,44
Focus administratie mondeling	-1,05	-0,55	-2,38	-2,39
Focus analyse mondeling	-1,01	-0,52	-2,36	-2,37
Kwaliteit administratie mondeling	-0,98	-0,49	-2,32	-2,34
Stadium constructie mondeling	-0,95	-0,51	-2,29	-2,28
Focus constructie mondeling	-0,93	-0,47	-2,25	-2,26
Stadium verslaglegging mondeling	-0,91	-0,42	-2,27	-2,21
Focus verslaglegging mondeling	-0,88	-0,46	-2,16	-2,19
Focus verslag mondeling	-0,86	-0,39	-2,12	-2,16
Freq constructie peer	-0,84	-0,33	-2,11	-2,14
Freq administratie peer	-0,81	-0,37	-2,03	-2,12
Freq administratie zelf	-0,79	-0,35	-2,07	-2,11
Stadium constructie peer	-0,72	-0,29	-1,95	-2,09
Stadium administratie peer	-0,71	-0,21	-1,91	-2,07
Freq verslag peer	-0,68	-0,26	-1,82	-2,04
Diff/tie administratie geschreven	-0,66	-0,24	-1,79	-1,95
Diff/tie administratie mondeling	-0,65	-0,19	-1,84	-1,92
Freq analyse peer	-0,64	-0,12	-1,73	-1,88
Freq verslag zelf	-0,63	-0,09	-1,67	-1,85
Freq verslaglegging performance	-0,63	-0,08	-1,66	-1,85
Freq verslag performance	-0,62	-0,07	-1,65	-1,84
Stadium verslaglegging performance	-0,61	-0,07	-1,65	-1,84
Stadium verslag performance	-0,61	-0,06	-1,64	-1,83
Freq analyse performance	-0,60	-0,06	-1,64	-1,83
Stadium analyse performance	-0,60	-0,05	-1,63	-1,82
Kwaliteit administratie prestatie	0,73	1,19	0,11	-1,75
Kwaliteit verslag performance	0,74	1,19	0,11	-1,76
Diff/tie administratie prestatie	0,75	1,20	0,12	-1,77
Kwaliteit constructie performance	0,75	1,20	0,12	-1,77
Kwaliteit verslaglegging performance	0,76	1,21	0,13	-1,78
Kwaliteit constructie mondeling	0,76	1,21	0,13	-1,78
Kwaliteit verslaglegging mondeling	0,79	1,27	0,16	-1,74
Kwaliteit analyse mondeling	0,81	1,23	0,18	-1,72
Freq verslaglegging peer	0,83	1,25	0,19	-1,69
Stadium verslag per	0,84	1,34	0,22	-1,65
Kwaliteit verslag mondeling	0,85	1,38	0,25	-1,63
Focus verslag peer	0,86	1,41	0,27	-1,61
Diff/tie constructie mondeling	0,87	1,42	0,29	-1,57
Diff/tie constructie geschreven	0,89	1,48	0,32	-1,54
Diff/tie analyse mondeling	0,92	1,51	0,35	-1,51

Diff/tie verslag mondeling	0,95	1,55	0,38	-1,48
Diff/tie verslaglegging mondeling	0,98	1,58	0,41	-1,45
Diff/tie verslaglegging geschreven	0,99	1,61	0,43	-1,44
Focus administratie peer	1,03	1,64	0,49	-1,41
Diff/tie verslag geschreven	1,05	1,68	0,51	-1,38
Kwaliteit administratie peer	1,08	1,72	0,53	-1,36
Diff/tie analyse geschreven	1,09	1,76	0,55	-1,34
Diff/tie administratie peer	1,11	1,79	0,59	-1,31
Focus constructie peer	1,14	1,81	0,61	-1,28
Focus administratie zelf	1,16	1,86	0,63	-1,26
Kwaliteit constructie peer	1,17	1,89	0,66	-1,22
Stadium analyse peer	1,19	1,93	0,68	-1,19
Diff/tie constructie peer	1,21	1,96	0,72	-1,16
Focus analyse peer	1,23	1,99	0,75	-1,15
Stadium Verslaglegging peer	1,25	2,03	0,79	-1,09
Focus Verslaglegging peer	1,27	2,06	0,81	-1,07
Kwaliteit analyse peer	1,29	2,09	0,83	-1,05
Kwaliteit Verslaglegging peer	1,31	2,12	0,86	-1,02
Stadium Verslag peer	1,32	2,17	0,89	-0,99
Diff/tie analyse peer	1,33	2,19	0,93	-0,96
Diff/tie verslaglegging peer	1,35	2,23	0,96	-0,94
Freq constructie zelf	1,36	2,25	0,99	-0,91
Stadium constructie zelf	1,37	2,28	1,01	-0,88
Kwaliteit verslag peer	1,38	2,32	1,03	-0,84
Focus constructie zelf	1,39	2,36	1,06	-0,82
Diff/tie verslag peer	1,41	2,38	1,08	-0,79
Stadium administratie zelf	1,42	2,42	1,12	-0,77
Kwaliteit constructie zelf	1,43	2,44	1,15	-0,74
Kwaliteit administratie zelf	1,44	2,48	1,19	-0,72
Freq analyse zelf	1,45	2,51	1,21	-0,69
Freq verslaglegging zelf	1,46	2,56	1,23	-0,67
Focus analyse zelf	1,47	2,57	1,28	-0,63
Stadium analyse zelf	1,49	2,61	1,32	-0,61
Stadium verslag zelf	1,51	2,65	1,35	-0,57
Diff/tie administratie zelf	1,52	2,69	1,39	-0,55
Diff/tie constructie zelf	1,53	2,72	1,41	-0,52
Kwaliteit verslag zelf	1,54	2,76	1,44	-0,49
Kwaliteit verslaglegging zelf	1,55	2,79	1,46	-0,47
Kwaliteit analyse zelf	1,57	2,81	1,48	-0,45
Diff/tie verslaglegging zelf	1,58	2,85	1,53	-0,41
Diff/tie verslag zelf	1,63	2,89	1,55	-0,39
Diff/tie analyse zelf	1,69	2,92	1,57	-0,37
Focus verslaglegging zelf	1,75	2,99	1,59	-0,33
Stadium verslaglegging zelf	1,81	3,01	1,61	-0,29
Focus verslag zelf	1,89	3,05	1,62	-0,26

De Saltus-parameter schat *(τ waarden)*

Item klas	**Geëxamineerde stadium**		
	1	**2**	**3**
1	0*	0*	0*
2	0*	1,79	1,84
3	0*	1,24	3,14

Opmerking 1: Lege lijnen worden gebruikt om de drie stadia van beoordelingsvaardigheden van docenten te scheiden die door clusteranalyse naar voren zijn gekomen.
*Vast op nul voor modelidentificatie.

Als we kijken naar de resultaten in Tabel 1 zien we dat de vijf fasen van het beoordelingsproces die werden gebruikt om de vaardigheden van leraren te meten (zie vorige paragraaf) niet op zichzelf staan, maar integendeel in alle drie de typen stadia bestaan. Dit houdt in dat leraren in alle drie de typen/fasen betrokken zijn bij de dynamische beoordelingscyclus, waarbij hun vaardigheden in elke fase worden gedifferentieerd in termen van complexiteit. Bovendien komen de vijf meetdimensies (d.w.z. frequentie, focus, fase, kwaliteit en differentiatie) niet overeen met één enkele fase, maar zijn ze verspreid over de 3 fasen.

De fasen geïdentificeerd in de 1e fase van het project (validatiestudie) werden gebruikt om beslissingen te nemen met betrekking tot de inhoud en het ontwerp van de TPD die werd geïmplementeerd tijdens de 2e fase van het FORMAS-project (zie Verslag 10 die verwijst naar de impact van de interventie voor het verbeteren van de beoordelingsvaardigheden van docenten die deelnemen aan de TPD en voor de impact ervan op het bevorderen van leerresultaten van leerlingen). De ontwikkelingsschaal werd consistent geïdentificeerd in beide meetperiodes (d.w.z. aan het begin en aan het einde van de interventie), wat verdere empirische ondersteuning bood voor de bevindingen van het validatieonderzoek. Door de twee metingen van deelnemende leraren te vergelijken, werd bovendien waargenomen dat in de gevallen waarin verandering plaatsvond, deze verandering plaatsvond in de richting van het volgende meer veeleisende niveau (d.w.z. van stadium 1 naar stadium 2, of van stadium 2 naar stadium 3). Deze stapsgewijze beweging bevestigt het ontwikkelingskarakter van de onderzochte beoordelingsvaardigheden verder. Bovendien bleek uit de gegevens over de prestaties van leerlingen die tijdens de interventiefase van het FORMAS-project waren verzameld, dat leraren die zich in een hogere beoordelingsfase bevonden, effectiever waren dan docenten die zich in de lagere fasen bevonden. Het feit dat de inhoud van elke fase afzonderlijk was gedefinieerd, maakte het mogelijk om specifieke beoordelingsvaardigheden te identificeren die een grotere impact hebben op de prestaties van leerlingen.

Op basis van de geïdentificeerde classificatie van beoordelingsvaardigheden worden in de volgende paragraaf professionele normen voorgesteld die beschrijven wat docenten geacht worden te kunnen doen met betrekking tot de beoordeling van leerlingen.

3. De professionele normen bij formatieve beoordeling

De hierna gepresenteerde professionele normen zijn ontwikkeld op basis van het theoretische kader en de gegevens die zijn afgeleid van de eerste fase van het FORMAS-project. Zoals hierboven vermeld, suggereerden de analyses van Rasch en Saltus een classificatie van beoordelingsvaardigheden op basis van hun moeilijkheidsgraad. Deze classificatie suggereerde het bestaan van drie verschillende groepen vaardigheden die geleidelijk overgaan van gemakkelijkere naar meer geavanceerde vaardigheden. Er

werd ook ontdekt dat de vijf fasen van het beoordelingsproces in alle drie de fasen naast elkaar bestaan. De ontwikkeling van professionele normen voor de beoordeling van leerlingen op basis van deze resultaten werd gedaan om een beter begrip te krijgen van de vaardigheden die betrokken zijn bij de beoordeling van leerlingen en om te helpen bij het definiëren van specifieke verwachtingen waaraan docenten moeten voldoen met betrekking tot beoordeling. Het is belangrijk op te merken dat hoewel we naar drie normen verwijzen, we onze steekproef van leraren in vier groepen kunnen indelen: a) leraren met Rasch-schattingen onder de moeilijkheidsgraad van fase 1 die de vaardigheden van norm 1 nog niet onder de knie hebben, b) norm 1 docenten die erin zijn geslaagd om de vaardigheden van norm 1 onder de knie te krijgen en acties moeten ondernemen om de vaardigheden van norm 2 te bereiken, c) norm 2 leraren die erin zijn geslaagd om de vaardigheden van norm 2 onder de knie te krijgen en acties moeten ondernemen om hun vaardigheden verder te verbeteren om de vaardigheden van norm 3 onder de knie te krijgen, en d) docenten van norm 3 die alle vaardigheden (ook die van norm 3) onder de knie hebben. Dit houdt in dat norm als cumulatief van aard wordt gezien. Leraren van norm 2 hebben bijvoorbeeld al vaardigheden van norm 1 bereikt en moeten nu actie ondernemen om hun vaardigheden te verbeteren om norm 3 te bereiken. Om een spaarzamere presentatie en betere inspanningen voor hulpverbetering te bereiken, wordt elke norm gepresenteerd in relatie tot de vaardigheden die relevant zijn voor elk van de vijf fasen van het beoordelingsproces (dwz *het construeren/selecteren van beoordelingsinstrumenten/-processen; het beheren van beoordelingsinstrumenten/-processen; het vastleggen van beoordelingsresultaten; het analyseren, interpreteren en gebruiken van beoordelingsresultaten en het rapporteren van de resultaten aan de beoogde gebruikers*). Daarom geeft elke norm per fase van het beoordelingsproces weer wat er van docenten wordt verwacht. Dit houdt in dat alle docenten betrokken zijn bij de vijf fasen van het beoordelingsproces, maar dat ze er niet allemaal in geslaagd zijn hun vaardigheden in dezelfde mate te ontwikkelen. Van leraren die zijn geclassificeerd als werkend naar norm 2, wordt bijvoorbeeld verwacht dat ze zich richten op het verbeteren van hun vaardigheden die relevant zijn voor alle vijf fasen van het beoordelingsproces. Tabel 2 geeft de verdeling weer van de 114 vaardigheden gemeten door de lerarenvragenlijst per professionele norm in relatie tot de vijf fasen van het beoordelingsproces.

Tabel 2. Verdeling van beoordelingsvaardigheden per professionele norm

Normen	Fasen van het beoordelingsproces				
	Beoordelingsinstrumenten/-processen construeren/selecteren	***Administratie van beoordelingstools/processenproces***	***Beoordelingsresultaten vastleggen***	***Analyseren, interpreteren en gebruiken van beoordelingsresultaten***	***Verslaglegging resultaten naar beoogde gebruiker***
Norm 1	Freq Constructie **Geschreven** Stadium Constructie **Geschreven** Focus Constructie **Geschreven** Freq Constructie **Mondeling**	Freq Administratie **Geschreven** Stadium Administratie **Geschreven** Freq Administratie **Mondeling** Focus Administratie **Geschreven**	Freq Verslag **Geschreven** Stadium Verslag **Geschreven** Focus Verslag **Geschreven** Freq Verslag **Mondeling** Stadium Verslag **Mondeling**	Freq Analyse **Geschreven** Stadium Analyse **Geschreven**	Freq Verslaglegging **Geschreven** Stadium Verslaglegging **Geschreven**
Norm 2	Kwaliteit Constructie **Geschreven** Freq Constructie **Performance** Stadium constructie **Mondeling** Focus constructie **Mondeling** Freq constructie **Peer** Stadium constructie **Peer**	Kwaliteit Administratie **Geschreven** Freq Administratie **Performance** Stadium administratie **Mondeling** Focus administratie **Mondeling** Kwaliteit administratie **Mondeling** Stadium administratie **Performance** Freq administratie **Peer** Stadium administratie **Peer** Freq administratie **Zelf** Diff/tie administratie **Geschreven** Diff/tie administratie **Mondeling**	Kwaliteit Verslag **Geschreven** Focus verslag **Mondeling** Freq verslag **Peer** Freq verslag **Zelf** Freq verslag **Performance** Stadium verslag **Performance**	Focus analyse **Geschreven** Freq analyse **Mondeling** Kwaliteit analyse **Geschreven** Stadium analyse **Mondeling** Focus analyse Mondeling Freq analyse **Peer** Freq analyse **Performance** Stadium analyse **Performance**	Freq Verslaglegging **Mondeling** Focus Verslaglegging **Geschreven** Kwaliteit Verslaglegging **Geschreven** Stadium verslaglegging **Mondeling** Focus verslaglegging **Mondeling** Freq verslaglegging **Performance** Stadium verslaglegging **Performance**
Norm 3	Kwaliteit constructie **Mondeling** Diff/tie constructie **Mondeling** Kwaliteit constructie **Performance** Diff/tie constructie **Geschreven** Focus constructie **Peer** Kwaliteit constructie **Peer** Diff/tie constructie **Peer** Freq constructie **Zelf** Stadium constructie **Zelf** Focus constructie **Zelf** Kwaliteit constructie **Zelf** Diff/tie constructie **Zelf**	Focus administratie **Peer** Kwaliteit administratie **Peer** Diff/tie administratie **Peer** Focus administratie **Zelf** Stadium administratie **Zelf** Kwaliteit administratie **Zelf** Diff/tie administratie **Zelf** Kwaliteit administratie **Performance** Diff/tie administratie **Performance**	Stadium verslag **Peer** Kwaliteit verslag **Mondeling** Focus verslag **Peer** Diff/tie verslag **Mondeling** Diff/tie verslag **Geschreven** Stadium Verslag **Peer** Kwaliteit verslag **Peer** Diff/tie verslag **Peer** Kwaliteit verslag **Performance** Stadium verslag **Zelf** Kwaliteit verslag **Zelf** Diff/tie verslag **Zelf** Focus verslag **Zelf**	Kwaliteit analyse **Mondeling** Diff/tie analyse **Mondeling** Diff/tie analyse **Geschreven** Stadium analyse **Peer** Focus analyse **Peer** Kwaliteit analyse **Peer** Diff/tie analyse **Peer** Freq analyse **Zelf** Focus analyse **Zelf** Stadium analyse **Zelf** Kwaliteit analyse **Zelf** Diff/tie analyse **Zelf**	Kwaliteit verslaglegging **Mondeling** Freq verslaglegging **Peer** Diff/tie verslaglegging **Mondeling** Diff/tie verslaglegging **Geschreven** Kwaliteit verslaglegging **Performance** Stadium Verslaglegging **Peer** Focus Verslaglegging **Peer** Kwaliteit Verslaglegging **Peer** Diff/tie verslaglegging **Peer** Freq verslaglegging **Zelf** Kwaliteit verslaglegging **Zelf** Diff/tie verslaglegging **Zelf** Focus verslaglegging **Zelf** Stadium verslaglegging **Zelf**

Belangrijke conclusies komen naar voren wanneer de inhoud van de drie normen nader wordt bekeken. Ten eerste sluiten de afgeleide normen aan bij argumenten over het dynamische karakter van het beoordelingsproces. Zoals hierboven vermeld, staan de vijf fasen van het beoordelingsproces, die werden gebruikt om de vaardigheden van leraren te meten, niet op zichzelf, maar integendeel, ze blijken in alle drie de normen naast elkaar te bestaan. Dit houdt in dat alle docenten betrokken zijn bij de vijf fasen van het beoordelingsproces, maar dat ze er niet allemaal in geslaagd zijn hun vaardigheden in dezelfde mate te ontwikkelen. Zo wordt van alle docenten verwacht dat ze hun vaardigheden in het afnemen van beoordelingen verbeteren, maar zich richten op het ontwikkelen van verschillende, in termen van moeilijkheid en complexiteit, beheervaardigheden voor beoordelingen. Als we bovendien kijken naar de verdeling van vaardigheden die verband houden met de onderzochte beoordelingstechnieken (dwz schriftelijke, mondelinge, prestatie-, peer- en zelfbeoordeling), kunnen we zien dat schriftelijke en mondelinge beoordelingen in alle drie de normen worden gepresenteerd, terwijl de rest van de technieken zijn alleen aanwezig in norm 2 en 3. Deze bevinding suggereert dat het moeilijker is om vaardigheden te ontwikkelen die verband houden met het gebruik van de prestatiebeoordelingstechniek in de wiskunde en met het gebruik van zelf- en groepsbeoordeling. Als we ten slotte kijken naar de toegepaste meetdimensies, kunnen we vaststellen dat de dimensies frequentie, fase en focus aanwezig zijn in alle drie de normen, terwijl de kwaliteits- en differentiatiedimensies alleen aanwezig zijn in de norm 2 en 3. Specifiek komen kwaliteit en differentiatie meestal voor in norm 2 met betrekking tot schriftelijke beoordeling, terwijl ze in norm 3 voorkomen met betrekking tot alle beoordelingstechnieken. Deze bevinding is in lijn met de resultaten van onderzoeken die zijn uitgevoerd om de validiteit van het dynamische model te testen op niet alleen de leerkracht maar ook op schoolniveau (voor een overzicht van deze onderzoeken zie Kyriakides et al., 2021). Hierna volgt een beschrijving van elke voorgestelde professionele norm.

Norm 1: voornamelijk schriftelijke beoordelingen gebruiken om de prestaties in wiskunde te meten voor summatieve doeleinden.

De eerste norm omvat 17 beoordelingsvaardigheden. De meeste van deze vaardigheden (13 van de 17) hebben betrekking op schriftelijke beoordeling, terwijl de rest (4 van de 17) betrekking heeft op mondelinge beoordeling. Dit geeft aan dat mondelinge beoordeling niet systematisch wordt gebruikt om het leren van leerlingen te beoordelen, terwijl andere technieken (d.w.z. prestatie-, zelf- en peerbeoordeling) helemaal niet worden gebruikt. Daarom wordt in norm 1 meestal de nadruk gelegd op het gebruik van schriftelijke beoordelingen. Dit wordt ook ondersteund door het feit dat de frequentie- en fasedimensies van schriftelijke beoordeling voor alle vijf fasen aanwezig zijn. Dit suggereert dat docenten van norm 1 veel gebruik maken van schriftelijke toetsing (frequentie) en het gebruik gelijkmatig verdelen over de vijf fasen (fase). Als het gaat om mondelinge beoordeling, kunnen we zien dat docenten vaak mondelinge beoordelingen ontwerpen, afnemen en opnemen, maar de verkregen resultaten worden niet geanalyseerd en gebruikt om het leren te ondersteunen en de

beoogde gebruikers te informeren. Norm 1 leraren lijken alleen gegevens te gebruiken die voortkomen uit schriftelijke technieken bij het rapporteren van resultaten aan ouders en leerlingen.

Norm 2: Verschillende beoordelingstechnieken gebruiken om prestaties in wiskunde te meten, maar zonder de juiste succescriteria te definiëren en constructieve feedback te geven.

De tweede norm omvat 38 beoordelingsvaardigheden. Vaardigheden die relevant zijn voor alle beoordelingstechnieken zijn aanwezig. Vaardigheden met betrekking tot mondelinge beoordeling komen echter vaker voor dan vaardigheden die verband houden met de andere technieken (13 van de 38). Concreet zijn de fase- en focusdimensie van mondelinge beoordeling in relatie tot alle fasen van het beoordelingsproces aanwezig. Dit suggereert dat leraren van norm 2 niet alleen mondelinge beoordeling toepassen, maar het gebruik ervan ook op de juiste manier verdelen over de vijf fasen (fase) en dat ze hun leerlingen mondeling beoordelen voor meer dan één doel (bijvoorbeeld het identificeren van behoeften van leerlingen, het uitvoeren van zelfevaluatie, het adopteren van zijn/haar langetermijnplanning, met evaluatietaken als uitgangspunt voor het lesgeven). Verder kan worden vastgesteld dat de kwaliteitsdimensie van schriftelijke beoordeling aanwezig is in alle fasen van het beoordelingsproces. Dit suggereert dat de gebruikte schriftelijke beoordeling van goede kwaliteit is wat betreft de eigenschappen van de gebruikte instrumenten (bijvoorbeeld validiteit, betrouwbaarheid, bruikbaarheid, representativiteit) en wat betreft de feedback die wordt gegeven. Bovendien suggereert de aanwezigheid van de kwaliteitsdimensie voor schriftelijke beoordeling gedurende het hele beoordelingsproces dat leraren de beoordeling plannen en gebruiken om formatieve in plaats van summatieve doelen te bereiken. Als we naar prestatiebeoordeling kijken, kunnen we zien dat de frequentie- en fasedimensie van prestatiebeoordeling ook aanwezig is in alle vijf fasen (verwacht de fasedimensie voor Fase 1: hulpmiddelen voor constructie/ selectie/ proces). Dit suggereert dat Norm 2-docenten prestatiebeoordeling vaak gebruiken (frequentie) en het gebruik gelijkmatig verdelen over de vijf fasen (fase). Het feit dat de dimensies focus, kwaliteit en differentiatie ontbreken, suggereert echter dat prestatiebeoordeling niet altijd formatief georiënteerd is (focus), terwijl verbetering van de kwaliteit en de geschiktheid voor verschillende leerlingenbehoeften (differentiatie) vereist is. Vaardigheden met betrekking tot peer- en zelfevaluatie zijn ook aanwezig in Norm 2, wat suggereert dat docenten niet de enigen zijn die verantwoordelijk zijn voor de beoordeling, maar dat ze ook de rol van de beoordelaar verschuiven naar leerlingen. In het bijzonder voeren leraren regelmatig peer- en zelfevaluatie uit en leggen ze de uitgelokte informatie vast. Het feit dat de dimensies kwaliteit en differentiatie niet aanwezig zijn, suggereert echter dat hun praktijken verder kunnen worden verbeterd om meer valide en betrouwbare informatie te verstrekken en om tegemoet te komen aan de individuele behoeften van leerlingen.

Norm 3: Beoordelingstechnieken gebruiken om specifieke en complexere leerdoelen te meten om constructieve feedback te geven, maar zonder leerlingen te betrekken bij het beoordelingsproces en hun beoordelingspraktijk te differentiëren.

De derde norm omvat 60 beoordelingsvaardigheden. Het grote aantal vaardigheden dat is opgenomen in vergelijking met de andere twee normen, ondersteunt de veronderstelling dat een beoordeling van hoge kwaliteit niet gemakkelijk te bereiken is. Kijkend naar de inhoud van norm 3 kunnen we zien dat vaardigheden met betrekking tot alle beoordelingstechnieken aanwezig zijn. Er zijn echter meer vaardigheden die verband houden met peer- en zelfbeoordeling, wat suggereert dat de kwaliteitsimplementatie van deze technieken moeilijker te bereiken is. Specifiek, vaardigheden die relevant zijn voor peer- en zelfevaluatie verschijnen in relatie tot alle fasen van het beoordelingsproces. Dit houdt in dat leraren van norm 3 kwaliteits-peer- en zelfevaluatie gebruiken om de informatie die wordt opgewekt vast te leggen, te gebruiken en te rapporteren om het leren te ondersteunen, terwijl ze tegelijkertijd hun praktijken aanpassen aan de individuele leerbehoeften van leerlingen. Als we kijken naar schriftelijke beoordeling, kunnen we vaststellen dat de differentiatiedimensie over de fasen van het beoordelingsproces (behalve Fase 2: Beheer van beoordelingsinstrumenten/-proces) aanwezig is. Dit suggereert dat hoewel de implementatie van schriftelijke beoordeling gemakkelijker te bereiken is, het differentiëren van schriftelijke beoordeling geen gemakkelijke taak is. Hetzelfde geldt voor mondelinge toetsing aangezien ook kwaliteits- en differentiatieaspecten van mondelinge toetsing aanwezig zijn. Dit suggereert dat leraren van norm 3 mondelinge beoordeling van goede kwaliteit op een meer gestructureerde en geschikte manier gebruiken en hun beoordeling aanpassen aan de gedifferentieerde leerbehoeften van leerlingen. Norm 3-docenten gebruiken prestatiebeoordeling ook op een meer systematische manier en elementen van kwaliteit en differentiatie kunnen worden geïdentificeerd in hun prestatiebeoordelingspraktijk. In het algemeen omvat deze norm alleen kwalitatieve aspecten van alle beoordelingstechnieken in de vijf fasen. Dit kan betekenen dat, hoewel het gemakkelijker is om verschillende technieken in te voeren, het moeilijker is om dit te doen op manieren die de kwaliteit ervan en een gedifferentieerde implementatie in alle fasen van het beoordelingsproces garanderen.

4. Implicaties van de professionele standaarden bij de beoordeling voor het onderwijsbeleid

Een van de doelstellingen van ons Erasmus+KA3-project getiteld ‘’Formatieve beoordeling bevorderen: van theorie naar beleid en praktijk (FORMAS)’’ is het genereren van beleidsrichtlijnen die formatieve beoordeling bevorderen, aangezien onderzoek suggereert dat formatieve beoordelingspraktijken een positieve invloed hebben op de prestaties van leerlingen (Creemers & Kyriakides, 2015; Hattie & Temperley, 2007; Herman et al., 2006; Wiliam et al., 2004). Het project probeert met name beleidsmakers aan te moedigen om het beoordelingsbeleid te hervormen en om ondersteuningsmechanismen voor leerkrachten op te zetten voor de effectieve implementatie van formatieve beoordeling. De ontwikkeling van specifieke professionele standaarden voor formatieve

beoordeling zal naar verwachting de doelstellingen van het project ondersteunen door een basis te bieden voor op theorie gebaseerde en empirisch onderbouwde beleidsbeslissingen.

Met name de voorgestelde professionele normen kunnen beleidsmakers helpen om te verduidelijken wat een goede beoordelingspraktijk is. De kritische beleidsanalyse die in het kader van het FORMAS-project werd uitgevoerd, toonde aan dat er in de deelnemende landen (dwz Cyprus, Griekenland, Nederland en België) geen beleid aanwezig is dat vereist dat leraren deskundig en bekwaam zijn in beoordeling (zie Verslag 1: Een kritische beoordeling van het nationale beleid inzake formatieve beoordeling). Daarom kunnen de voorgestelde normen helpen bij het definiëren van specifieke verwachtingen waaraan leraren moeten voldoen met betrekking tot beoordeling. Dit zal op zijn beurt leraren helpen te begrijpen hoe ze effectieve beoordeling kunnen implementeren in hun dagelijkse lespraktijk, mogelijke tekortkomingen in hun praktijk kunnen identificeren en hen tegelijkertijd verantwoordelijk kunnen houden voor de implementatie van de normen.

Bovendien moet de nodige ondersteuning worden geboden om leraren te helpen hun nadruk te verleggen op effectieve evaluatiepraktijken. De voorgestelde normen kunnen worden gebruikt als leidraad voor beslissingen met betrekking tot de initiële lerarenopleiding en professionele ontwikkeling. Op deze manier zal het onderwijs dat aan leraren wordt aangeboden, zowel in pre- als in-service contexten, inspelen op de professionele behoeften van specifieke groepen leraren, elke keer dat ze worden ondersteund om hun beoordelingspraktijk te verbeteren.

Bovendien suggereert het feit dat de vastgestelde normen zijn gebaseerd op gegevens die aantonen dat leraren gedifferentieerde professionele behoeften hebben met betrekking tot beoordeling, dat ze kunnen worden gebruikt om de identificatie van de specifieke behoeften van leraren bij de beoordeling mogelijk te maken, zodat passende corrigerende maatregelen kunnen worden genomen. Zo kunnen standaarden worden gebruikt om de professionele ontwikkelingsbehoeften van leraren per land en onderwijsfase te identificeren op basis waarvan beslissingen over de nadruk, inhoud en duur van de ondersteuning kunnen worden genomen.

Ten slotte kan de ontwikkeling van professionele standaarden bij formatieve toetsing ook nuttig zijn voor evaluatiedoeleinden. Aangezien de beoordeling van leerlingen wordt erkend als een belangrijke factor voor de effectiviteit van leraren, kunnen normen worden gebruikt om het proces, van het verzamelen van gegevens over de prestaties van een leerkracht voor formatieve evaluatiedoeleinden, te sturen. Het feit dat standaarden een gedetailleerde beschrijving geven van wat leraren zouden moeten kunnen doen met betrekking tot alle fasen van het beoordelingsproces, maakt het mogelijk om specifieke en constructieve feedback te geven. Het biedt ook een basis voor docenten zelf om kritisch te reflecteren op hun beoordelingspraktijk, zodat corrigerende acties kunnen plaatsvinden. Op dit punt is het belangrijk op te merken dat de voorgestelde normen naar verwachting niet zullen worden gebruikt voor summatieve doeleinden (dwz evaluatie door docenten). Bovendien kunnen de voorgestelde normen worden gebruikt voor andere evaluatieve doeleinden, zoals het testen van de impact van verschillende hervormingen op het verbeteren van beoordelingsvaardigheden van docenten in relatie

tot de normen. De normen kunnen bijvoorbeeld worden gebruikt om de effectiviteit te onderzoeken van een professionele lerarenopleiding gericht op de beoordeling van leerlingen of om de impact van nieuw nationaal beleid op de beoordeling van leerlingen te evalueren.

5. Slotopmerkingen

De evaluatie van leerlingen is de afgelopen jaren het middelpunt geweest van verschillende onderwijsinitiatieven. Dit resultaat heeft tot doel beleidsmakers bewust te maken van het introduceren van een empirisch onderbouwde en theorie gestuurde benadering bij de vorming en implementatie van beoordelingsgerelateerd beleid. In het FORMAS-project werden de normen gebruikt om een programma voor professionele ontwikkeling van leraren (TPD) te ontwikkelen dat erkende dat leraren gedifferentieerde professionele behoeften hebben en dat sommigen competenter kunnen zijn in specifieke vaardigheden dan anderen. Voorafgaand aan de TPD vond een eerste evaluatie van de beoordelingsvaardigheden van docenten plaats met behulp van de docentenvragenlijst (zie Verslag 3: De lerarenvragenlijst het meten van de beoordelingsvaardigheden van leraren). De analyse van de verkregen gegevens bevestigde de groepering van beoordelingsvaardigheden op basis van hun moeilijkheidsgraad en het genereren van specifieke beoordelingsnormen. Op basis van deze resultaten werden de inhoud en structuur van de TPD zo ontworpen dat aan deze gedifferentieerde behoeften kon worden tegemoetgekomen, door gedifferentieerde training te bieden aan elke groep leraren op basis van hun eerste evaluatieresultaten. De TPD bleek een positief effect te hebben op zowel de beoordelingsvaardigheden van docenten als de leerresultaten van leerlingen, wat suggereert dat professionele ontwikkelingsinitiatieven die de gedifferentieerde professionele behoeften van docenten erkennen en tegemoet komen aan de beoordeling, een positief effect kunnen hebben op het bevorderen van formatieve beoordelingspraktijken en daardoor betere leerresultaten kunnen bereiken. Daarom kunnen de voorgestelde professionele standaarden als referentiepunt fungeren en de ontwikkeling van onderwijsbeleid sturen op manieren die de effectieve implementatie van formatieve beoordeling ondersteunen.

References

Adams, R. J., & Khoo, S. T. (1996). *Quest: The interactive test analysis system*. Camberwell, Victoria: ACER.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association (AFT, NCME, & NEA). (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues and Practice*, 9(4), 30–32.

Andersen, E. B. (1977). Sufficient statistics and latent trait models, Psychometrika, 42, 69–81.

Andrich, D. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education, 1*(4), 363–378.

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, N. J.: Lawrence Erlbaum Associates, Publishers.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3-12.

Christoforidou, M., Kyriakides, L., Antoniou, P., & Creemers, B. P.M. (2014). Searching for stages of teacher's skills in assessment. *Studies in Educational Evaluation*, *40*, 1-11.

Creemers, B., & Kyriakides, L. (2015). Developing, testing, and using theoretical models for promoting quality in education. *School Effectiveness and School Improvement*, *26*(1), 102-119.

Creemers, B.P.M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: a contribution to policy, practice and theory in contemporary schools.* London and New York: Routledge.

Hattie, J., & Temperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Herman, J.L., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2006). *The nature and impact of teachers' formative assessment practices. CSE Technical Report #703.* National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hopfenbeck, T.N. (2018). Classroom assessment, pedagogy and learning–twenty years after Black and Wiliam 1998. *Assessment in Education*, *25*(6), 545-550.

Kyriakides, L., Creemers, B. P.M., Panayiotou, A., and Charalambous, E. (2021). *Quality and Equity in Education: Revisiting Theory and Research on Educational Effectiveness and Improvement*. London and New York: Routledge

Linacre, J. M. (1999). Understanding Rasch measurement: Estimation methods for Rasch measures. *Journal of Outcome Measurement, 3*(4), 382-405.

Panadero, E., Broadbent, J., Boud, D., & Lodge, J. M. (2019). Using formative assessment to influence self-and co-regulated learning: the role of evaluative judgement. *European Journal of Psychology of Education*, *34*(3), 535-557.

Schafer, W.D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice*, *10*(1), 3-6.

Stiggins, R.J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational measurement: Issues and practice*, *18*(1), 23-27.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, *77*(3), 238-245.

Wiliam, D., Lee, C., Harrison, C., & Black, P. J. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles Policy and Practice, 11*(1), 49-65.

Wright, B.D. (1985). *Additivity in psychological measurement.* In E.E. Roskam (Ed.), Measurement and personality assessment (pp. 101-112). Amsterdam: Elsevier Science Publishers BV.