Categorieënanalyse bij de LOVS-toetsen rekenen-wiskunde

Jan Janssen & Marian Hickendorff
Cito, Arnhem / Universiteit Leiden

1 inleiding

In 2008 is voor het onderdeel Rekenen-Wiskunde een nieuwe rapportagevorm geïntroduceerd, de zogenaamde 'categorieënanalyse'. Met een categorieënanalyse kan worden nagegaan of leerlingen op een bepaald onderdeel van de reken-wiskundetoets meer (of minder) fouten maken dan op grond van hun algemene vaardigheidsniveau verwacht mag worden. Met behulp van de score van een leerling op de totale toets krijgen we een indicatie van zijn of haar rekenvaardigheidsniveau. Op basis van dit algemene niveau verwachten we dat een leerling op bijvoorbeeld het onderdeel meten een bepaald aantal opgaven goed maakt. Bij een categorieënanalyse wordt nagegaan of het verwachte aantal goed gemaakte opgaven per onderdeel overeenkomt met het aantal opgaven dat hij of zij daadwerkelijk goed maakt per onderdeel. Bij de rapportage van het verschil (tussen waargenomen en verwachte scores) wordt aangegeven of dat een klein verschil is dat aan toeval kan worden toegeschreven of dat het een betekenisvol verschil is.
Bij categorieënanalyse wordt gebruikt gemaakt van specifieke eigenschappen van het IRT-meetmodel dat aan de ontwikkeling van de LOVS Rekenen-Wiskunde toetsen ten grondslag ligt. Daarom starten we met het beschrijven van de kenmerken van dit meetmodel alvorens we ingaan op het analyseren van de vaardigheid op categorieniveau.

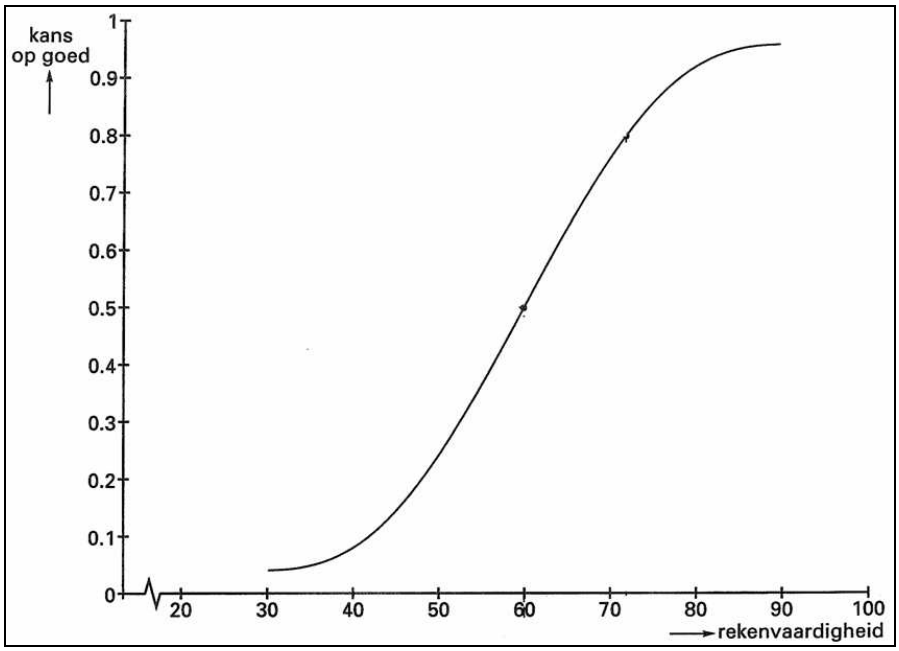
2 ontwikkeling van het LOVS

Bij de ontwikkeling van de LOVS-toetsen brengen we allereerst de doelen in kaart die men bij het reken-wiskundeonderwijs wil realiseren. Dat gebeurt onder andere op basis van het bestuderen van reken-wiskundemethoden, onderzoek- en toetsgegevens. Vervolgens worden opgaven geconstrueerd die voorgelegd worden aan leerlingen. Op basis van de antwoorden van de leerlingen wordt een schaal geconstrueerd die als uitgangspunt bij het rapporteren en interpreteren van de toetsresultaten gebruikt kan worden.
We gaan nu eerst in op de schaalconstructie en het model dat bij het construeren van de reken-wiskundeschaal gebruikt wordt. Kennis van het achterliggende model is nodig om de categorieënanalyse goed te begrijpen.

3 Item Respons Theorie

Bij de Item Respons Theorie (IRT) wordt ervan uitgegaan dat het antwoord van een leerling op een opgave (met andere woorden de respons op een item) wordt bepaald door:
- het vaardigheidsniveau van de leerling op het moment van de meting;
- de moeilijkheid van de opgave.

Met behulp van statistische formules kan bepaald worden welke kans een leerling met een bepaalde vaardigheid heeft om een opgave met een bepaalde moeilijkheid goed op te lossen. Grafisch kan deze relatie als volgt worden weergegeven (fig.1).

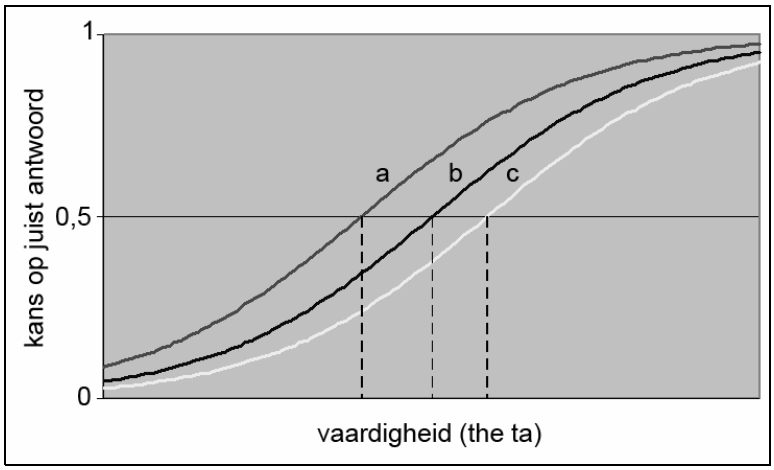


figuur 1: model Item Respons Theorie voor één item

Hiermee hebben we de elementen die van belang zijn voor het rapporteren met behulp van een itemresponsmodel gehad. Dat zijn dus:
- de vaardigheid van een persoon die op de schaal van laag naar hoog kan worden aangegeven;
- de moeilijkheid van de opgaven die van gemakkelijk naar moeilijk kan worden aangegeven;
- de kans dat een leerling met een bepaalde vaardigheid een opgave met een bepaalde moeilijkheid goed maakt.

4 de moeilijkheid van opgaven

De moeilijkheid van een opgave wordt gedefinieerd als de vaardigheid die nodig is om een item met een kans van 50 procent goed op te lossen. In figuur 2 staan de curves van drie items die verschillen in moeilijkheidsgraad. De *a*-curve representeert het gemakkelijkste item en de *c*-curve (de witte curve) is het moeilijkste item van deze drie.

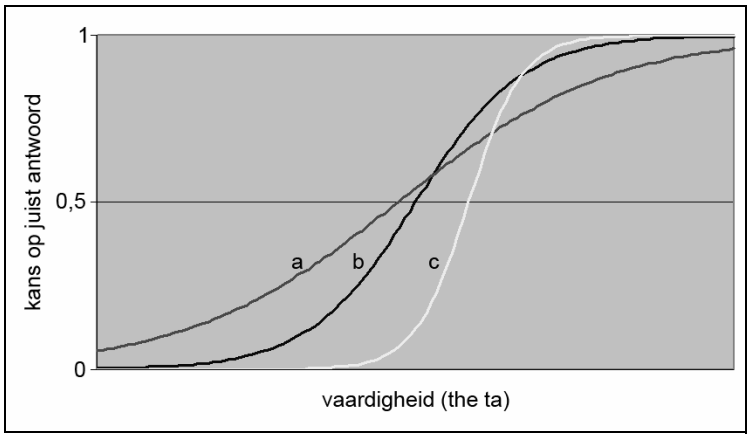


figuur 2: de moeilijkheidsparameter in het IRT-model

5 het discriminerend vermogen van opgaven

De curves, die items representeren, hebben niet allemaal dezelfde vorm. Er zijn relatief steile curves (zie bijvoorbeeld curve *c* in figuur 3) waarbij de kans op succes snel toeneemt, terwijl het niveau op de vaardigheidsschaal slechts weinig toeneemt. Maar er zijn ook relatief vlakke curves waarbij de kans op succes slechts heel geleidelijk toeneemt naarmate het niveau op de vaardigheidsschaal stijgt. De steilheid van de curve vertelt ons iets over het discriminerend vermogen van het item.
In figuur 3 discrimineert item *a* het slechtst en item *c* discrimineert het best. Dat betekent dat item *c* beter onderscheid maakt tussen vaardige en minder vaardige leerlingen dan item

a en *b*.



figuur 3: de discriminatieparameter

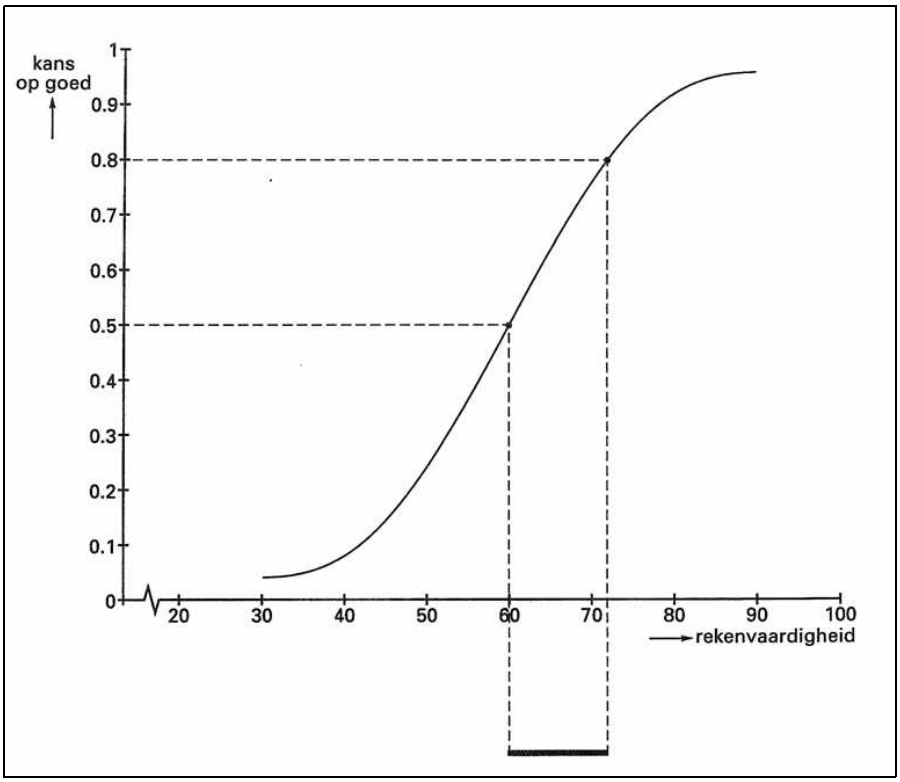
Aan het discriminerend vermogen wordt een getal toegekend dat normaal varieert van ongeveer 1 tot 6. Een discriminerend vermogen van 1 betekent dat het item niet discrimineert en een discriminerend vermogen van 6 dat je te maken hebt met een item dat heel goed onderscheid maakt tussen vaardige en minder vaardige leerlingen. De discriminatieparameter wordt gebruikt bij het bepalen van de gewogen score. Als een leerling een opgave met een discriminatieparameter van 3 goed oplost krijgt hij drie punten. Gewogen scores worden overigens alleen gebruikt als het computerprogramma de score van de leerling bepaalt. In andere gevallen wordt er gewerkt met ongewogen scores. Werken met gewogen scores heeft een grotere betrouwbaarheid als voordeel.

6 representatie van opgaven en het vaardigheidsniveau op de schaal

Van elke opgave kunnen we zo'n curve maken. Alleen, als je met al die curves moet werken, verlies je snel het overzicht over het totaal. Daarom hebben we voor de rapportage de informatie gereduceerd tot balkjes of lijnen die opgaven representeren. De linkerkant van het balkje geeft het $p50$ kanspunt van het item aan en de rechterkant van het balkje het $p80$ kanspunt van de opgave. Dat leidt tot een driedeling:

– Heb je meer dan 80 procent kans op een goed antwoord, dan beheers je die opgave goed.
– Heb je 50 tot 80 procent kans op een goed antwoord, dan beheers je de opgave matig tot goed.
– Heb je minder dan 50 procent kans op een goed antwoord, dan beheers je de opgave

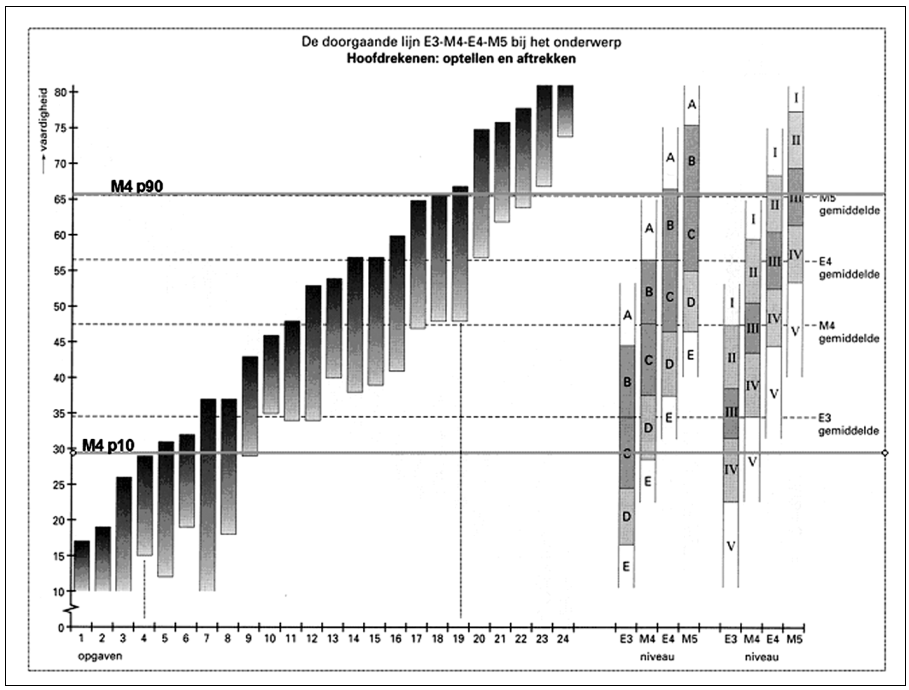
onvoldoende.
In figuur 4 is aangegeven hoe we vanuit een curve tot zo'n balkje komen.



figuur 4: $p50$- en $p80$ kanspunt van een item

In figuur 5 (afkomstig uit de inhoudsverantwoording van de LOVS-toetsen voor groep 4) zijn de balkjes die opgaven representeren niet horizontaal, maar verticaal afgebeeld en staat ook de vaardigheidsschaal van laag naar hoog verticaal weergegeven. Kenmerkend voor de schaal is dat er opgaven op afgebeeld kunnen worden (hier het $p50$ en $p80$ kanspunt van de opgave) en dat de vaardigheid van individuele leerlingen, of zoals hier de vaardigheidsverdeling van groepen leerlingen, op verschillende tijdstippen erop kan worden afge-

beeld.



figuur 5: een gedeelte van de vaardigheidsschaal Rekenen-Wiskunde

In bovenstaande figuur is met een stippellijn het vaardigheidsniveau van de gemiddelde *E*3-leerling, *M*4-leerling, *E*4-leerling en *M*5-leerling (valt samen met de *M*4 *p*90-lijn) aangegeven. Met behulp van de schaal kan vooruitgang van leerlingen in de tijd in kaart gebracht worden.
Het combineren van het afbeelden van opgaven met het afbeelden van het vaardigheidsniveau van een leerling maakt inhoudelijke interpretatie mogelijk.

In figuur 5 zijn de vaardigheidsniveaus van de percentiel-10 leerling en de percentiel-90 leerling medio groep 4 gemarkeerd. De percentiel-10 leerling beheerst de eerste 6 opgaven goed of nagenoeg goed, de opgaven 7 en 8 matig en de opgaven 9 tot en met 24 onvoldoende. De percentiel-90 leerling beheerst de eerste 19 opgaven goed, de opgaven 20 tot en met 22 matig en de opgaven 23 en 24 onvoldoende. Bekijken we de bijbehorende opgaven dan kunnen we concluderen dat de percentiel-10 leerling medio groep 4 voornamelijk opgaven goed beheerst die tot de leerstof van groep 3 horen en dat de percentiel-90 leerling al de moeilijkste opgaven (zoals 39 + 25) van de leerstof van groep 4 goed beheerst.
In het vervolg van dit artikel gaan we nu in op de wijze waarop we de vaardigheid van de leerling kunnen analyseren op categorieniveau. Daarbij maken we gebruik van de specifieke eigenschappen van het IRT-meetmodel.

7 van traditionele foutenanalyse naar categorieënanalyse

Bij traditionele foutenanalyses wordt er per leerling een overzicht van het aantal foute antwoorden per categorie gemaakt. Op basis van die overzichten neemt men beslissingen over de extra te besteden aandacht aan categorieën bij leerlingen.
Traditionele foutenanalyses komen in tal van varianten voor. Van overzichten met een indeling in enkele hoofdcategorieën tot zeer gedetailleerde overzichten met veel categorieën die slechts enkele opgaven tellen. De meeste varianten hebben als gemeenschappelijk kenmerk dat een absoluut criterium als uitgangspunt wordt genomen.
In het voorbeeld van figuur 6 zien we dat leerling 1 van de tien opgaven over getallen er zes goed heeft gemaakt. Daarmee voldoet deze leerling niet aan het gestelde criterium (zeven opgaven goed maken). Bij het onderdeel vermenigvuldigen en delen heeft de leerling ook minder opgaven goed gemaakt dan het gestelde criterium. Uit deze informatie wordt dan afgeleid dat extra aandacht moet worden besteed aan de onderdelen getallen en vermenigvuldigen en delen.

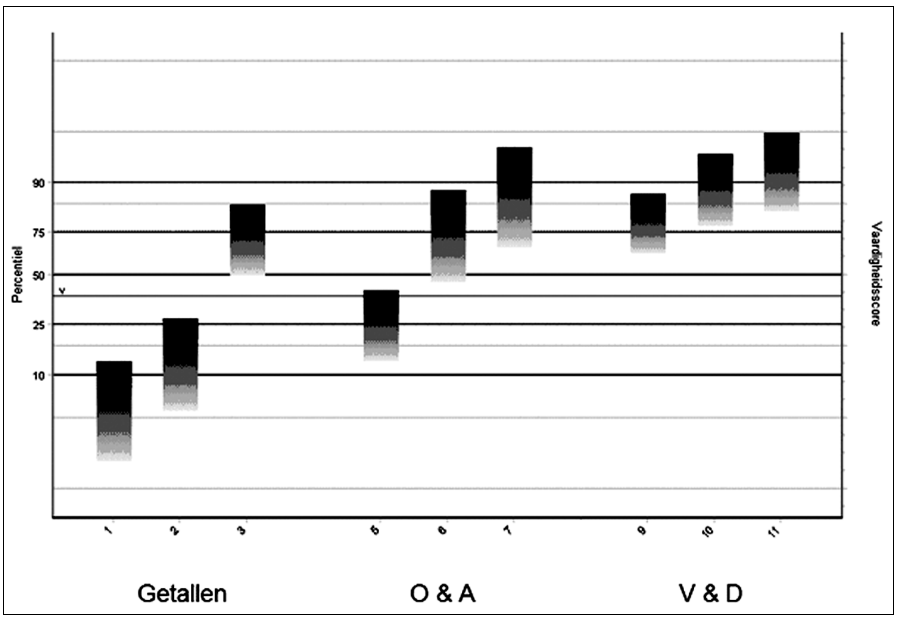
	Aantal opgaven in toets	Criterium	Leerling 1	Leerling 2	Leerling 3
Getallen	10	7	6*		
+ en – 12	12	8	8		
× en :	10	7	5*		
Meten	8	5	5		

figuur 6: traditionele foutenanalyse

Met deze benadering is het niet mogelijk een goede diagnose te maken. Er wordt immers geen rekening gehouden met het vaardigheidsniveau van de leerling en de moeilijkheidsgraad van de opgaven. Daarom hebben we een alternatief ontwikkeld dat met behulp van de computer uitgevoerd kan worden en dat categorieënanalyse wordt genoemd.

Uitgaande van figuur 7 verduidelijken we het principe achter de categorieënanalyse. In deze figuur zijn van elk van de categorieën Getallen, Optellen en aftrekken (O&A) en Ver-

menigvuldigen en delen (V&D) drie items met hun $p50$- en $p80$-kanspunten afgebeeld.



figuur 7: vaardigheidsniveau in relatie tot categorieen

Van een leerling met een vaardigheidsniveau dat in figuur 7 is aangeduid met v hebben we over zijn score op de drie items van het onderdeel getallen een andere verwachting dan over zijn score op de drie items van het onderdeel vermenigvuldigen en delen. Bij het onderdeel getallen heeft een leerling immers bij twee van de drie opgaven een kans groter dan 80 procent om die goed op te lossen en bij het onderdeel vermenigvuldigen en delen heeft deze leerling bij alle drie de items minder dan 50 procent kans op een goed antwoord. Uit het psychometrisch model waarmee binnen het LOVS wordt gewerkt, kan worden berekend wat de verwachte score is op elk van de categorieën, gegeven de totaalscore op de toets. Deze verwachte score kan weer worden omgezet in een percentage van de maximale score op elk van de categorieën.

8 functie van categorieënanalyse en de categorieën bij de toetsen

De functie van categorieënanalyse is nagaan of het niveau van een leerling op onderdelen significant afwijkt van het niveau dat bij de verschillende onderdelen op basis van het algemene vaardigheidsniveau van de leerling wordt verwacht. Uit de vaardigheidsscore die de leerling op de toets behaalt en het daarbij behorende vaardigheidsniveau weten we of we met een sterke of zwakke leerling te doen hebben. Met behulp van de categorieënanalyse kunnen we nagaan of een leerling, gegeven zijn algemene niveau, evenwichtig presteert op de verschillende categorieën van de toets.
In figuur 8 wordt een overzicht gegeven van de categorieën die we bij de verschillende LOVS-toetsen Rekenen-Wiskunde onderscheiden.

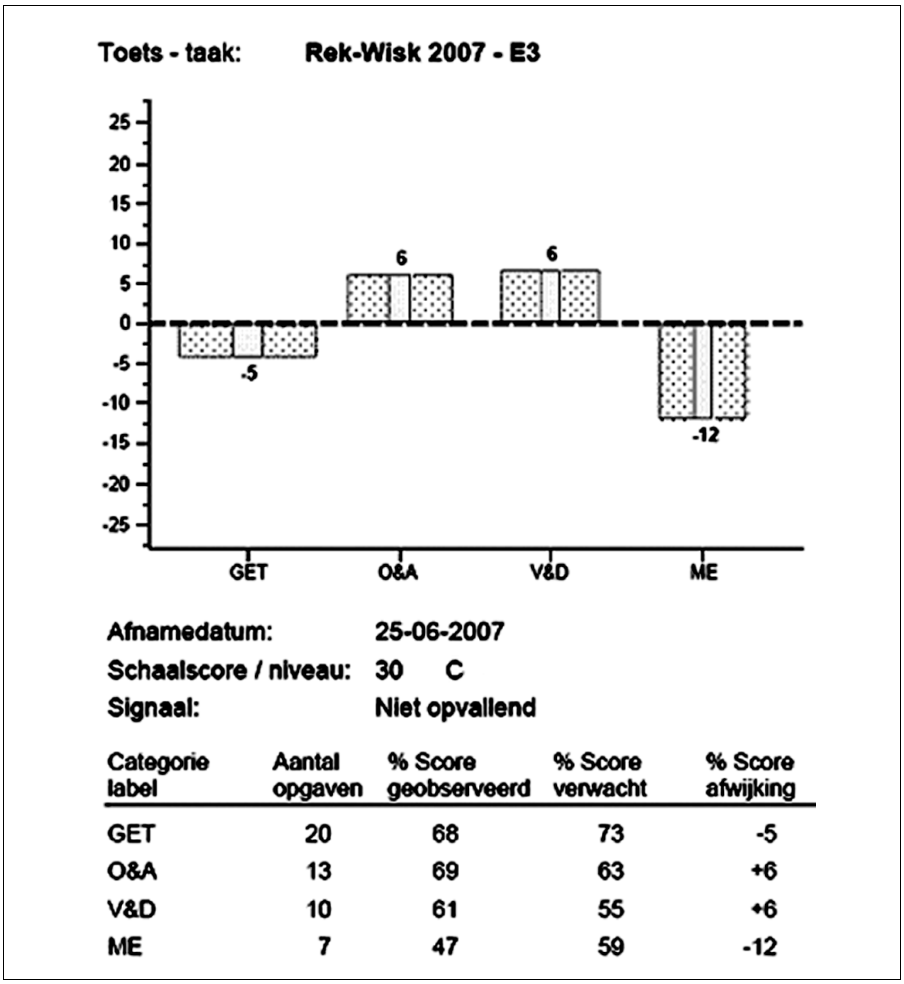
	M3	M3	M4	E4	M5	E5	M6	E6	M7	E7
Getallen										
Optellen en aftrekken										
Vermenigvuldigen en delen										
Hoofdrekenen										
Bewerkingen met papier										
Meten, Tijd en Geld										
Meten										
Tijd en Geld										
Verhoudingen, Breuken en Procenten										
Aantal categorieën	4	4	4	4	5	5	6	6	6	6

figuur 8: overzicht categorieën bij de toetsen

9 categorieënanalyse

Bij de categorieënanalyse wordt met behulp van statistische formules (Chi-kwadraat) een afstandsmaat bepaald. Elke categorie levert een eigen bijdrage aan de afstandsmaat. De bijdrage hangt af van het verschil tussen de geobserveerde en de verwachte score en van het aantal items dat de betreffende categorie telt. De bijdrage is groter naarmate het aantal items en/of het verschil geobserveerd–verwacht toeneemt. De gegevens over het aantal items dat de categorie telt en het verschil geobserveerde–verwachte score, worden in de tabel van een profielweergave vermeld (fig.9). In de grafiek van deze figuur (een scoreprofiel van een leerling met niveau C) geeft de hoogte van de kolom het verschil aan tussen

geobserveerde en verwachte score. De oppervlakte (het egale gedeelte) van de kolommen komt overeen met de bijdrage van elke categorie aan de afstandsmaat. Hoe groter de oppervlakte, des te groter de bijdrage. Op basis van de grootte van de afstandsmaat zijn er drie mogelijkheden: het signaal 'niet opvallend', 'opvallend' of 'zeer opvallend' kan worden gegeven.



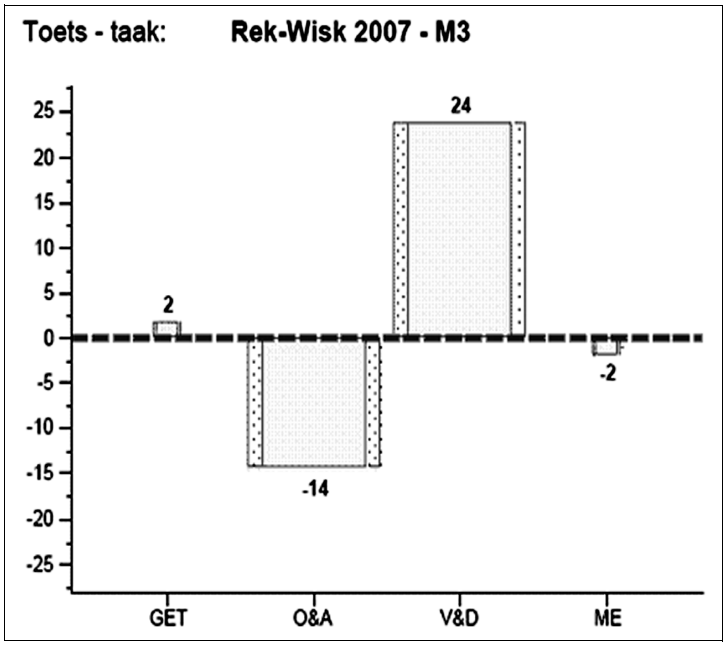
Categorie label	Aantal opgaven	% Score geobserveerd	% Score verwacht	% Score afwijking
GET	20	68	73	-5
O&A	13	69	63	+6
V&D	10	61	55	+6
ME	7	47	59	-12

figuur 9: voorbeeld: niet opvallend profiel

Voor de beste 10 procent en de zwakste 10 procent van de leerlingen genereert het programma geen profiel. Categorieënanalyse is daar niet zinvol, omdat de 10 procent beste leerlingen nauwelijks fouten maken en voor de 10 procent zwakste leerlingen geldt dat zij zoveel fouten maken dat we moeten concluderen dat zij, gezien hun vaardigheid, beter een toets van een vorig afnamemoment kunnen maken.

Uit analyses is gebleken dat 10 procent van de leerlingen een ‘opvallend’ of ‘zeer opvallend’ profiel heeft. Bij een ‘opvallend’ profiel behoort het profiel tot de hoogste 10 procent afstandsmaten, maar niet tot de 5 procent hoogste. Met een ‘zeer opvallend’ profiel hebben we te maken als het profiel behoort tot de 5 procent hoogste afstandsmaten.
In figuur 9 geeft het gestippelde gedeelte in de kolommen aan hoe ver de leerling van een ‘opvallend’ profiel afzit. In deze figuur is het gestippelde deel van de staven groot in vergelijking met het egaal gekleurde deel. Dit betekent dat bij deze leerling de afstandsmaat erg klein is en nog behoorlijk ver af ligt van het kritische punt waarbij we van een ‘opvallend’ profiel gaan spreken.

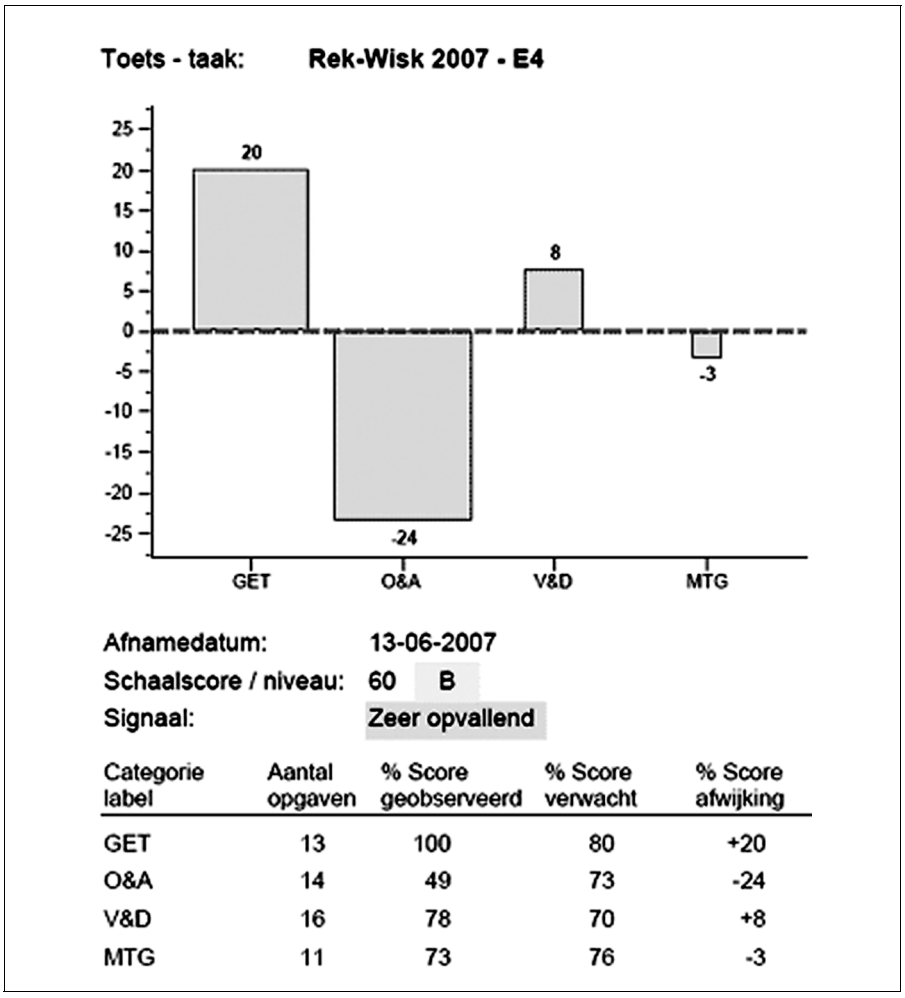
In het profiel van de leerling in figuur 10 zijn de gestippelde stroken relatief klein in vergelijking met het egaal gekleurde deel van de staven. Dit betekent dat we het profiel nog steeds als niet opvallend aanmerken, maar dat de afstandsmaat bij deze leerling behoorlijk dicht in de buurt komt bij de kritische grens waarbij we van een ‘opvallend’ profiel spreken.



figuur 10: voorbeeld: niet opvallend profiel -2-

In figuur 11 staat een voorbeeld van een ‘zeer opvallend’ profiel. Het profiel wordt ‘zeer opvallend’ genoemd omdat de afstandsmaat (die verder niet in de figuur wordt vermeld)

behoort tot de 5 procent hoogste afstandsmaten.



Categorie label	Aantal opgaven	% Score geobserveerd	% Score verwacht	% Score afwijking
GET	13	100	80	+20
O&A	14	49	73	-24
V&D	16	78	70	+8
MTG	11	73	76	-3

figuur 11: voorbeeld: zeer opvallend profiel

Aan de oppervlakte van de staven zien we dat de afstandsmaat vooral bepaald is door de categorieën 'Getallen' (GET) en 'Optellen en aftrekken' (O&A). Het gaat om een relatief goede leerling (niveau B), die op de categorie 'Getallen' behoorlijk beter presteert dan op grond van zijn prestatie op de toets als geheel kan worden verwacht. Maar op de categorie 'Optellen en aftrekken' presteert de leerling in belangrijke mate onder de verwachting. Bij de categorie 'Vermenigvuldigen en delen' en 'Meten' presteert de leerling conform verwachting.

Bij de interpretatie moeten we voorzichtig zijn: het zou kunnen zijn dat bij deze leerling

het inzicht in en het begrip van optellen en aftrekken is achtergebleven. Het zou echter ook zo kunnen zijn dat de leerling bijvoorbeeld bij de aftrekopgaven steeds eenzelfde fout maakt en dat aandacht voor dit aspect snel een vooruitgang voor die leerling zou betekenen. Uit de figuur kunnen we onmogelijk aflezen wat de juiste interpretatie is. De functie van de profielanalyse is het signaleren van een opvallende of zeer opvallende onevenwichtigheid in het profiel die om nadere analyse vraagt.
De classificaties 'opvallend' en 'zeer opvallend' zijn puur beschrijvend en zijn op theoretisch-statistische gronden bepaald. Normaal gesproken kan voor ongeveer 10 procent van de leerlingen zo'n beschrijving verwacht worden. Zijn er in een klas echter veel meer dan 10 procent opvallende en zeer opvallende profielen dan is het aan te bevelen de profielen samen te bekijken. Als bijvoorbeeld alle opvallende profielen een onderprestatie laten zien op de categorie 'Meten' kan dit erop wijzen dat dit onderdeel in het onderwijs niet erg goed is overgekomen.
Zo kan categorieënanalyse zowel onevenwichtige prestaties op individueel niveau als op groepsniveau aan het licht brengen.

literatuur

Computerprogramma LOVS. Arnhem: Cito.
Handleiding Computerprogramma LOVS. Arnhem: Cito.
Janssen, J., F. Scheltens & J.M. Kraemer (2004-2009). *LOVS Rekenen-Wiskunde, Inhoudsverantwoording bij de toetspakketten voor groep 3 tot en met 7.* Arnhem: Cito.
Verhelst, N. (2007). *Profielanalyse met Item Respons Theorie.* Arnhem: Cito.