Foutenanalyse en constructie van afleiders voor meerkeuzevragen: een empirisch onderzoek

E. De Corte en A. Verkens
Afdeling Didactiek en Psychopedagogiek, K.U. Leuven

Samenvatting

De redactie van afleiders voor meerkeuzevragen stelt de itemschrijver voor een aantal problemen. In dit artikel wordt een empirisch onderzoek beschreven dat uitgaat van de idee dat bepaalde van deze problemen adequaat opgelost kunnen worden door foutenanalyse in het proces van itemredactie te integreren. De volgende conclusies komen hierbij naar voren.

1. *Ervaren itemschrijvers zijn voldoende in staat om onderscheid te maken tussen leerdoelen waarvoor, bij instrumentalisering in meerkeuzevraagvorm, relatief veel afleiders kunnen worden geconstrueerd en leerdoelen waarvoor dit minder het geval is.*
2. *Om zo reëel mogelijke afleiders te redigeren, is het nuttig de items eerst in open-vraagvorm aan de leerlingen voor te leggen teneinde een foutenінventarisatie door te voeren, welke als basis kan fungeren voor de redactie van de afleiders. Deze stellingname steunt niet alleen op de empirische bevindingen maar ook op theoretische overwegingen in verband met het remediëren van tekorten bij de leerlingen.*
3. *Meer beperkt tot het domein van leerdoelen waarop het onderzoek werd uitgevoerd, m.n. leerdoelen van het rekenonderwijs einde basisschool, wordt volgende conclusie gesteld. Wanneer men te weinig tijd of mankracht heeft om een foutenanalyse te maken van alle leerdoelen uit een toets, doet men er goed aan zich te beperken tot die leerdoelen waarvoor de itemschrijvers menen dat zij veel afleiders kunnen bedenken. Immers voor deze leerdoelen stelt men slechts een matige inhoudelijke overeenstemming vast tussen de afleiders voorgesteld door itemschrijvers en de fouten van de leerlingen.*

1. *Inleiding*

Tijdens de voorbije vijftien jaar werd in het onderzoek en in de onderwijspraktijk de behoefte gevoeld aan degelijke instrumenten voor regelmatige evaluatie van het rendement van het onderwijs. Aansluitend bij deze reële noden werd door het Centrum voor Psychopedagogisch en Didactisch Onderzoek (C.P.D.O. verbonden aan de Afdeling Didactiek en Psychopedagogiek van de K.U. Leuven) een onderzoeksproject opgezet met als hoofddoel: de ontwikkeling van wetenschappelijk gefundeerde en praktisch bruikbare evaluatiemethoden[1]. Onderhavig onderzoek werd uitgevoerd in het kader van dit project.

De opgaven uit objectieve evaluatieproeven en studietoetsen worden in toenemende mate aangeboden in meerkeuze-vraagvorm. Bij de redactie van deze items dient de itemschrijver bepaalde formele en inhoudelijke kwaliteiten na te streven, opdat de meerkeuzevragen voldoende begripsvaliditeit zouden bezitten. Tot de formele kenmerken van het item behoren de eigenschappen die inherent zijn aan het gebruik van de meerkeuze-vraagvorm en tevens relatief onafhankelijk zijn van de inhoud (bijv. aantal alternatieven, lengte van de alternatieven)[2].

De inhoudelijke samenstelling van het item, i.c. van de itemstam en van de alternatieven (juist antwoord en afleiders), wordt in belangrijke mate bepaald door de concrete doelstelling waarvan het item een instrumentalisering is. De inhoudelijke samenstelling van de afleiders vormt echter een bijzondere moeilijkheid voor de meeste itemschrijvers. Een veel gebruikte werkwijze bestaat erin dat de redacteur zelf mogelijke afleiders ontwerpt die naar zijn oordeel plausibel zijn. Om aan de constructie van de afleiders een meer gefundeerde basis te geven, werd door de auteurs voorgesteld om hierbij de didactische foutenanalyse in te schakelen[3]. Deze werkwijze bestaat o.m. uit het inventariseren via openvragen van de meest frequent voorkomende fouten van de leerlingen. Vervolgens worden deze fouten, na een didactische foutenanalyse, opgenomen als afleiders in de overeenkomstige meerkeuzevragen.

In dit artikel wordt een onderzoek gerapporteerd

waarin de bijdrage van de foutenanalyse tot de constructie van afleiders empirisch bestudeerd werd.

2. *Probleemstelling*

Bij de redactie van meerkeuzevragen stelt men vast dat itemconstructeurs slechts een beperkt aantal onjuiste alternatieven of afleiders kunnen bedenken[4]. In de praktijk treft men veelal items aan met drie afleiders. Aan deze algemene vaststelling kan volgende nuancering aangebracht worden. Voor bepaalde concrete leerdoelen menen ervaren itemschrijvers gemakkelijk een meerkeuzevraag te kunnen bedenken met drie of meer goede afleiders, terwijl zij dit voor andere doelen moeilijker achten. Met goede afleiders wordt bedoeld, dat er overeenstemming is tussen deze afleiders en de fouten die men aantreft bij de leerlingen die de betreffende leerdoelstelling niet hebben verworven na het verstrekte onderwijs.

Men kan zich afvragen welke de geldigheid is van het zojuist aangeduide onderscheid dat itemschrijvers menen te kunnen maken tussen leerdoelen. Met name rijzen volgende vragen: 1. is dit onderscheid in overeenstemming met het feitelijk antwoordgedrag van de leerlingen, in die zin dat zij op de overeenkomstige open vragen respectievelijk *meer en minder verschillende fouten* produceren; 2. is er vervolgens ook *inhoudelijke overeenstemming* tussen de foutieve antwoorden van de leerlingen en de afleiders geconstrueerd door itemschrijvers?

Om deze vragen empirisch te benaderen werd een domein van leerdoelen afgebakend. Het betreft 374 leerdoelen van het rekenonderwijs op de lagere school (niveau einde zesde leerjaar), die deel uitmaken van een systematisch opgestelde inventaris van doelstellingen.[5] Uit deze inventaris werden alleen de leerdoelen uit de rubrieken Getallenkennis (G), Hoofdbewerkingen (H), Maateenheden en schaalverdelingen (M) in het onderzoek opgenomen. Uit praktische overwegingen was beperking nodig. Er werd voor de genoemde rubrieken geopteerd, omdat ze de basisnoties van het rekenonderwijs bevatten.

Aansluitend bij de algemene vaststelling geformuleerd in het begin van deze paragraaf, werd aan drie deskundigen gevraagd de 374 leerdoelen te klasseren in de categorieën 'weinig afleiders' en 'veel afleiders'. De categorieën werden als volgt omschreven en dienovereenkomstig door de beoordelaars gehanteerd. In de categorie 'weinig afleiders' treft men de leerdoelen aan waarvoor de itemschrijvers menen dat het *moeilijk* is om drie goede afleiders te construeren. We nemen het aantal drie als begrenzing in functie van het vooraf gekozen type van meerkeuzevraag, m.n. items met vier alternatieven. De categorie 'veel afleiders'' groepeert de doelen waarvoor, steeds volgens het oordeel van de itemschrijvers, bij de instrumentalisering in meerkeuzevragen *vrij gemakkelijk* drie of meer goede afleiders kunnen geschreven worden.

Het klasseren van de leerdoelen in de categorieën 'weinig – veel afleiders' is in belangrijke mate mede afhankelijk van de ervaring van de itemschrijvers betreffende de foutieve oplossingen die men bij de leerlingen kan aantreffen. Aan het klasseren zelf werd voldoende tijd besteed en er werd hieromtrent een hoge mate van overeenstemming bereikt tussen de drie deskundigen. We zijn niet gekomen tot een volstrekt replicabele klassering, maar personen die voldoende vertrouwd zijn met de inventaris leerdoelstellingen en tevens enige ervaring hebben op het gebied van het itemschrijven, kunnen zeker tot een voldoende graad van overeenstemming komen met onze klassering, d.i. een globale overeenstemming van tenminste 75 %.

In paragraaf drie onderzoeken we of het onderscheid dat itemschrijvers maken tussen leerdoelen waarvoor bij de instrumentalisering in meerkeuzevraagvorm weinig en veel afleiders kunnen geconstrueerd worden, overeenstemt met het antwoordgedrag van de leerlingen in die zin dat zij op de overeenkomstige openvragen respectievelijk weinig en veel verschillende fouten produceren. Vervolgens wordt in paragraaf vier gepeild naar de mogelijkheden om met een foutenanalyse, beperkt tot een inventarisatie van de fouten van de leerlingen door middel van openvragen, tenminste drie goede afleiders te redigeren voor de overeenkomstige meerkeuzevragen. Tot slot wordt in paragraaf vijf onderzocht of er inhoudelijke overeenstemming is tussen de foutieve oplossingen van de leerlingen en de afleiders die geconstrueerd worden door itemschrijvers.

3. *Empirische gegevens over het klasseren van de leerdoelen*

3.1 *Beschrijving van de gevolgde werkwijze*

De toetsing die hier aan bod komt heeft betrekking op de vraagstelling naar de overeenstemming tussen de rubricering van de leerdoelen in de categorieën 'weinig – veel afleiders' en het aantal verschillende fouten die men voor de overeenkomstige doelen bij de leerlingen inventariseert met openvragen. Als hypothese werd gesteld dat voor de leerdoelen die

geklasseerd werden in de categorie 'weinig afleiders' (W.A.-doelen), de leerlingen op de overeenkomstige openvragen (W.A.-items) slechts een gering aantal verschillende fouten zullen produceren. Voor de leerdoelen geklasseerd als 'veel afleiders' (V.A.-doelen) verwachtte men dat de leerlingen op de overeenkomstige openvragen veel verschillende fouten zullen maken (V.A.-items). Volgende werkwijze werd bij de uitvoering van de toetsing gevolgd.

Na het rubriceren van de leerdoelen in de categorieën 'weinig – veel afleiders', werd overgegaan tot de samenstelling van een gestratificeerde steekproef. Bij de toepassing van deze steekproeftrekking werden de leerdoelen uit beide categorieën gegroepeerd volgens de verschillende leerstofeenheden of strata waartoe ze in de inventaris van leerdoelstellingen behoren. In functie van het aantal doelen in de verschillende strata en van het totaal aantal doelen dat om praktische redenen in de proeven kon opgenomen worden, werd de steekproeftrekking uitgevoerd. Voor de drie rubrieken samen (G/H/M) werden 86 leerdoelen in het onderzoek betrokken. Vervolgens werden deze doeleinden geïnstrumentaliseerd in de open-vraagvorm. De afname van de openvragen, gebundeld in drie proeven, gebeurde bij 152 leerlingen uit het zesde leerjaar van het basisonderwijs. Per item werden het aantal verschillende fouten en hun frequentie bepaald.

Bij de uitvoering van de statistische toetsingen werd gebruik gemaakt van de toetsingsprocedure van Wilcoxon[6]. De gestelde hypothese werd getoetst voor de drie leerstofrubrieken samen en voor de afzonderlijke rubrieken. Een toetsing voor de verschillende rubrieken samen is zinvol omdat in de meeste evaluatieproeven en schooltoetsen de verschillende rubrieken eveneens gegroepeerd worden. De toetsing per rubriek is bedoeld om eventuele preciseringen te kunnen aanbrengen bij de over-all toetsing.

3.2 *Resultaten*

De resultaten van de toetsing van de overeenstemming tussen de rubricering van de leerdoelen en het aantal verschillende fouten bij de leerlingen worden gegeven in tabel 1.

Voor de gegroepeerde gegevens van de drie rubrieken werd een significante z-waarde (+2.58) op het 1 %-niveau verkregen. De toetsingen binnen de afzonderlijke rubrieken leverden geen significante resultaten op voor de leerstofrubrieken G en M. In de rubriek H trad daarentegen wel een significant verschil op tussen de W.A.-items en de V.A.-items

Tabel 1. z-waarden voor de W.A.-V.A.-toetsing binnen elke leerstofrubriek en voor de drie rubrieken samen wat het aantal verschillende fouten per item betreft (toets van Wilcoxon)

Toetsing	Rubrieken			Totaal
	G	H	M	
W.A. – V.A.	+0.99	+1.74*	+0.95	+2.58**

* $p < .05$ (rechtseenzijdig getoetst)
** $p < .01$ (rechtseenzijdig getoetst)

wat de centrummaat van het aantal verschillende fouten betreft.

Het door de itemschrijvers gemaakte onderscheid tussen 'W.A.- en V.A.-doelen verwijst naar de mogelijkheid om drie of meer afleiders te redigeren. Dit onderscheid stemt volgens de uitgevoerde toetsing globaal overeen met het verschil in aantal fouten tussen de W.A.- en de V.A.-items. Dit verschil is echter van een andere grootte. Voor de W.A.-items noteerden we 10.5 als mediaan van het aantal verschillende fouten per item en voor de V.A.-items een mediaan van 23.5. Bij een eerste inspectie van de onderzoeksgegevens is dit groot aantal verschillende fouten in beide categorieën in belangrijke mate toe te schrijven aan de fouten die slechts éénmaal werden gemaakt. Dergelijke fouten zijn weinig bruikbaar als afleiders, omdat zij veelal teruggaan op een individueel oorzakenpatroon dat niet is terug te vinden in andere groepen leerlingen.

In welke mate het onderscheid tussen beide categorieën items is toe te schrijven aan de invloed van het aantal fouten met frequentie één, wordt in volgende toetsing onderzocht. Voor elke openvraag werd een index berekend, i.c. de verhouding van het aantal verschillende fouten met frequentie één tot het aantal verschillende fouten die frequenter voorkomen. Deze indexen werden voor de drie rubrieken samen en vervolgens binnen elke leerstofrubriek, gegroepeerd in de categorieën W.A.-items en V.A.-items. Om het eventuele onderscheid tussen beide categorieën na te gaan werd de toets van Wilcoxon gebruikt. Tabel 2 bevat de resultaten van deze toetsingen.

Uit tabel 2 blijkt dat er voor de drie rubrieken samen, alsook voor de rubrieken G en M tussen de W.A.- en de V.A.-items geen significant verschil is wat het aantal verschillende fouten met frequentie één betreft. Alleen in de rubriek H hebben de V.A.-items grotere indexen dan de W.A.-items. Dit betekent dat

Tabel 2. z-waarde voor de toetsing W.A. – V.A. binnen elke rubriek en voor de drie rubrieken samen wat relatief aantal fouten met frequentie één betreft (toets van Wilcoxon)

Toetsing	Rubrieken			Totaal
	G	H	M	
W.A. – V.A.	−1.17	+2.09*	− 0.32	− 0.41

* $p < .05$ (rechtseenzijdig getoetst)

voor de V.A.-items, in vergelijking met de W.A.-items, het aantal fouten met frequentie één groter is ten opzichte van het aantal fouten met een frequentie groter dan één. Deze tendens is hoofdzakelijk toe te schrijven aan de items over cijferrekenen waar tal van individuele fouten optreden.

3.3 *Interpretatie van de resultaten*

De resultaten voor de drie rubrieken samen, zoals aangegeven in tabel 1, laten toe te besluiten dat de V.A.-items een significant groter aantal verschillende fouten opleveren dan de W.A.-items. Op grond van de tweede toetsing mogen we aannemen dat dit onderscheid tussen beide categorieën items niet louter kan worden toegeschreven aan het eventueel groter aantal fouten met frequentie één onder de V.A.-items. Een toename van het aantal fouten met frequentie één gaat immers gepaard met een toename van het aantal fouten met een hogere frequentie. Deze conclusie is geldig indien men de verschillende leerstofrubrieken als een geheel beschouwt.

Hieruit mogen we besluiten dat het onderscheid dat door de deskundigen gemaakt werd tussen de leerdoelen waarvoor bij de instrumentalisering in meerkeuze-vraagvorm weinig en veel afleiders kunnen worden geconstrueerd, globaal genomen beantwoordt aan het antwoordgedrag van de leerlingen die op de overeenkomstige openvragen weinig en veel verschillende fouten produceren.

4. *Bruikbaarheid van de foutenanalyse bij de inventarisatie van fouten als afleiders voor meerkeuzevragen*

4.1 *Overzicht van de gevolgde werkwijze*

Volgens de itemschrijvers is het voor bepaalde leerdoelen moeilijk om drie afleiders te bedenken (W.A.-leerdoelen). Men kan zich hier afvragen of een foutenanalyse, i.c. een fouteninventarisatie, een oplossing kan bieden. Dit vermoeden steunt op de medianen van het aantal verschillende fouten per item. In vorige paragraaf werden deze medianen berekend voor de W.A.- en de V.A.-items. Beide medianen overtreffen in ruime mate het streefaantal drie. Bij de berekening van deze medianen werden echter ook de fouten opgenomen die slechts eenmaal voorkwamen in de onderzoeksgroep. Al zijn deze fouten belangrijk vanuit didactisch oogpunt (remediërend onderwijs), toch zijn ze, zoals reeds gezegd, niet bruikbaar als afleiders omdat ze teruggaan op een individueel oorzakenpatroon dat veelal niet is terug te vinden in een parallelle onderzoeksgroep.

In verband met de vraag of de fouteninventarisatie voor alle items drie afleiders kan aanbrengen, werden de items uit het onderzoek gedichotomiseerd in functie van een vooraf bepaalde minimumeis t.a.v. het aantal verschillende fouten. Deze eis is nogmaals gebaseerd op het type van meerkeuzevraag waarvoor werd geopteerd, nl. een meerkeuzevraag met drie afleiders. Een item wordt als een plus-item aangeduid, indien voor dit item drie of meer verschillende en bruikbare foutieve oplossingen geïnventariseerd werden. Als min-items duiden we deze aan waarvoor minder dan drie verschillende en bruikbare fouten werden vastgesteld. Een fout is bruikbaar als ze tenminste bij twee leerlingen uit de onderzoeksgroep voorkomt.

Bij dit dichotomiseren van de items per leerstofrubriek werd tevens de reeds eerder doorgevoerde rubricering aangehouden. Het betreft de rubricering in W.A.- en V.A.-items. Voor de W.A.-items werden de overeenkomstige leerdoelen door de geconsulteerde deskundigen gekenmerkt als leerdoelen waarvoor zij bij de instrumentalisering in meerkeuze-vraagvorm moeilijk drie afleiders kunnen redigeren; deze beoordelaars meenden dat zij bij de instrumentalisering van de V.A.-leerdoelen gemakkelijk drie of meer afleiders kunnen bedenken. Deze rubricering werd in dit onderzoek over de bruikbaarheid van de fouteninventarisatie behouden, om richtlijnen te kunnen opsporen zodat, bij eventuele beslissingen omtrent de toepassing van een inventarisatie, het oordeel van de itemschrijvers eveneens in aanmerking kan worden genomen.

Tabel 3 bevat een overzicht van de verdeling van de plus- en min-items over de categorieën W.A. en V.A.

Zowel voor de categorie W.A. als voor de categorie V.A. stellen we een overwicht van de plus- op de min-items vast. Hiervan werd een verdere statistische analyse uitgevoerd. Eerst werd binnen elke

Tabel 3. Verdeling van de plus- en de min-items over de categorieën W.A. en V.A. voor de verschillende leerstofrubrieken

Leerstofrubriek	Categorie	Itembeoordeling	
		plus-item	min-item
G	W.A.	9	3
	V.A.	15	2
	Totaal	24	5
H	W.A.	8	5
	V.A.	17	4
	Totaal	25	9
M	W.A.	3	0
	V.A.	18	0
	Totaal	21	0
G + H + M	W.A.	20	8
	V.A.	50	6
	Totaal	70	14

categorie (W.A. en V.A.) het eventuele overwicht van de plus- op de min-items nagegaan met behulp van de tekentoets[7]. Vervolgens werden de W.A.-items vergeleken met de V.A.-items naar het eventuele overwicht van het aantal plus- op het aantal min-items. Om dit verschil te toetsen voor de drie rubrieken samen en voor de afzonderlijke rubrieken, werd hier opnieuw de toets van Wilcoxon gebruikt.

4.2 *Resultaten*

4.2.1 *Vergelijking van het absoluut aantal plus- en min-items per categorie*

Om het eventuele overwicht van het aantal plus-items op de min-items te toetsen, werd de tekentoets gebruikt. Als hypothese werd gesteld dat er in elke categorie meer plus- dan min-items zijn. Tabel 4 bevat de resultaten van de verrichte toetsingen.

Voor de gegevens uit de drie rubrieken samen, stellen we zowel voor de V.A.-categorie als voor de W.A.-categorie een significant groter aantal plus- dan min-items vast. Bij de rubrieken G en H wordt voor de categorie W.A. geen significant verschil genoteerd tussen het aantal plus- en min-items.

Tabel 4. Toetsingsresultaten betreffende het verschil tussen het aantal plus- en min-items per categorie (W.A. en V.A.) (tekentoets)

Leerstofrubriek	Categorie	Resultaat tekentoets
G	W.A.	n. sign.
	V.A.	sign. + > −
H	W.A.	n. sign.
	V.A.	sign. + > −
M	W.A.	—
	V.A.	sign. + > −
Totaal	W.A.	sign. + > −
	V.A.	sign. + > −

Wegens het gering aantal W.A.-items in rubriek M was de toetsing binnen deze categorie niet zinvol. Voor de categorie V.A. stellen we wel een significant verschil vast in elk van de drie rubrieken; deze categorie bevat dus telkens meer plus- dan min-items.

4.2.2 *Vergelijking van de categorieën W.A. en V.A. naar het relatieve overwicht van de plus- op de min-items*

Bij het onderzoek van de categorieën W.A. en V.A. naar het relatieve overwicht van de plus- op de min-items, stelden we als hypothese dat de V.A.-items een groter overwicht van de plus-items hebben in vergelijking met de W.A.-items. Tabel 5 bevat de z-waarden voor elke rubriek en voor de drie rubrieken samen.

Tabel 5. z-waarden voor de toetsing W.A. – V.A. binnen elke rubriek en voor de drie rubrieken samen wat het relatieve overwicht van de plus- op de min-items betreft (toets van Wilcoxon)

Toetsing	Rubrieken			Totaal
	G	H	M	
W.A. – V.A.	+0.88	+1.20	—	+2.05*

* $p < .05$ (rechtseenzijdig getoetst)

Voor de drie rubrieken samen bevat de categorie V.A. significant meer plus- dan min-items ten opzichte van de categorie W.A. De rubriek M kwam

niet in aanmerking voor een toetsing aangezien de categorieën W.A. en V.A. geen min-items bevatten. Er werd geen significant resultaat vastgesteld voor de rubrieken G en H.

4.3 *Interpretatie van de resultaten*

Het vermoeden dat een fouteninventarisatie nuttig is wanneer de itemschrijvers moeilijk drie afleiders kunnen bedenken (W.A.-leerdoelen), moet thans gerelativeerd worden. We stellen nl. vast dat de inventarisatie, zowel voor de V.A.-items als voor de W.A.-items, niet alle items kan voorzien van drie bruikbare afleiders. Deze items werden in de toetsingen aangeduid als min-items(zie de vergelijking van het absoluut aantal plus- en min-items per categorie). Uit de tweede toetsing die in deze paragraaf werd beschreven (zie de vergelijking van de categorieën W.A. en V.A. naar het relatieve overwicht van de plus- op de min-items), kan men besluiten dat in de categorie V.A. significant meer plus- dan min-items optreden in vergelijking met de categorie W.A. Dit sluit aan bij de algemene vaststelling betreffende de lagere mediaan voor het aantal verschillende fouten in de W.A.-categorie ten opzichte van de V.A.-categorie.

Als algemene conclusie stellen we dat wanneer de constructie van afleiders voor meerkeuzevragen wordt voorafgegaan door een inventarisatie van foutieve oplossingen, er altijd leerdoelen zullen zijn waarvoor geen drie zinvolle afleiders kunnen worden opgesteld. Deze situatie zullen we echter meer aantreffen onder de W.A.-leerdoelen dan onder de V.A.-leerdoelen. Het lijkt ons voorbarig om hieruit te besluiten dat een foutenanalyse kan worden achterwege gelaten bij de constructie van meerkeuzevragen. Reeds elders hebben we een theoretische verantwoording gebracht om het gebruik van de foutenanalyse te integreren in het item-redactieproces[8].

Tabel 6. Overeenstemming tussen afleiders uit meerkeuzevragen en meest frequente foutieve oplossingen van leerlingen per rubriek

Rubriek	Aantal items	Graad van overeenstemming			
		0	1	2	3
G	20	5	8	7	–
H	22	8	7	6	1
M	13	4	3	5	1
Totaal	55	17	18	18	2

Al biedt de foutenanalyse geen volledige oplossing voor het nagestreefde aantal afleiders per item, deze werkwijze heeft alleszins een belangrijke inbreng bij de redactie van inhoudelijk relevante afleiders. Dit zijn afleiders die overeenstemmen met de meest frequent voorkomende fouten van de leerlingen. In paragraaf 5 wordt de mate van overeenstemming nagegaan tussen de afleiders geconstrueerd door itemschrijvers en de fouten die werden geïnventariseerd met openvragen bij de leerlingen.

5. *Inhoudelijke overeenstemming tussen de fouten uit de foutenanalyse en de afleiders geconstrueerd door itemschrijvers*

Bij de constructie van de openvragen voor dit onderzoek hebben we getracht om zoveel mogelijk vragen te construeren die parallel zijn met reeds bestaande meerkeuzevragen uit de itembank van het Centrum voor Psychopedagogisch en Didactisch Onderzoek[9]. Voor een aantal items was dit niet mogelijk omdat we de doelstelling-validiteit van de openvragen niet in het gedrang wilden brengen omwille van het nagestreefde parallellisme. De afleiders in de bedoelde meerkeuzevragen werden door de betrokken itemschrijvers zelf ontworpen en dus niet via toepassing van openvragen verkregen.

In deze paragraaf gaan we de overeenstemming na tussen de drie afleiders uit de bestaande meerkeuzevragen en de drie meest frequent voorkomende fouten van de leerlingen op de parallelle openvragen uit het onderzoek. In de tabellen wordt gesproken van een overeenstemming nul voor een item, wanneer van de drie meest frequent voorkomende fouten er geen enkele terug te vinden is onder de afleiders van de overeenkomstige meerkeuzevraag. Er is sprake van een overeenstemming 1 voor een item, als van de drie meest frequent voorkomende fouten er één is terug te vinden onder de afleiders; analoog wordt gesproken van een overeenstemming 2 en 3.

Tabel 6 bevat de gegevens voor de drie leerstofrubrieken geordend naar de graad van overeenstemming.

Voor meer dan de helft van de items die in aanmerking werden genomen, stellen we een overeenstemming 1 of 2 vast. In de categorie overeenstemming 3 treffen we slechts twee items aan. Ongeveer een derde van de items behoort tot de categorie overeenstemming 0. Aangezien slechts een gedeelte van de items uit de steekproef werd opgenomen voor het onderzoek naar de inhoudelijke overeenstem-

ming tussen afleiders en fouten, komt de representativiteit van de steekproefgegevens in het gedrang. Daarom kunnen statistische generalisatietechnieken hier geen toepassing vinden. Niettemin geven we in tabel 7 een ordening van de verkregen gegevens in functie van de klasseringscategorieën W.A. en V.A., omdat dit tot een interessante vaststelling leidt die oriënterend kan zijn voor verder onderzoek.

Tabel 7. Overeenstemming tussen afleiders uit meerkeuzevragen en meest frequente foutieve oplossingen van leerlingen per rubriek en per categorie (W.A. – V.A.)

Rubriek	Aantal items	Categorie		Graad van overeenstemming			
		Code	Aantal	0	1	2	3
G	20	W.A.	7	2	2	3	–
		V.A.	13	3	6	4	–
H	22	W.A.	9	3	1	4	1
		V.A.	13	5	6	2	–
M	13	W.A.	1	1	–	–	–
		V.A.	12	3	3	5	1
Totaal	55	W.A.	17	6	3	7	1
		V.A.	38	11	15	11	1

Als algemene tendens kunnen we aangeven dat ongeveer twee derden van de V.A.-items een overeenstemming 0 of 1 hebben, daar waar dit slechts voor ongeveer de helft van de W.A.-items geldt. Hieruit kunnen we afleiden dat de itemschrijvers bij de instrumentalisering van de leerdoelen uit de categorie V.A. in vergelijking met de categorie W.A., minder goed afleiders kunnen bedenken die overeenstemmen met de meest frequent voorkomende fouten. De V.A.-leerdoelen bieden de itemschrijvers meer mogelijkheden voor de constructie van afleiders. Dit groter aantal beschikbare potentiële afleiders kan de selectie die de itemschrijvers hierop moeten doorvoeren bemoeilijken. De beschikbare gegevens laten toe volgende hypothese te formuleren: het aantal afleiders dat itemschrijvers kunnen bedenken is omgekeerd evenredig met de graad van overeenstemming.

6. *Conclusies*

Het empirisch onderzoek reveleerde dat deskundige beoordelaars bekwaam zijn om een onderscheid te maken tussen leerdoelen waarvoor men bij de instrumentalisering in meerkeuze-vraagvorm respectievelijk weinig of veel afleiders kan redigeren. Deze conclusie steunt op volgende bevinding: de instrumentalisering in de open-vraagvorm van de leerdoelen uit de categorie 'weinig afleiders' leverde significant minder verschillende fouten op bij een representatieve groep leerlingen dan deze uit de categorie 'veel afleiders'. Er zijn echter aanwijzingen dat er, voor de onderzochte leerstofrubrieken, slechts een matige inhoudelijke overeenstemming is tussen de afleiders die itemschrijvers kunnen bedenken en de meest frequent voorkomende fouten. Deze overeenstemming is het geringst voor de leerdoelen uit de categorie 'veel afleiders'.

Wanneer men over weinig tijd en mankracht beschikt om voor alle leerdoelen uit een evaluatieproef een fouteninventarisatie op te zetten, dan dient men bij voorkeur de leerdoelen uit de categorie 'veel afleiders' hieraan te onderwerpen. In een dergelijke situatie verkrijgt men op deze wijze de best mogelijke waarborgen voor de kwaliteit van de afleiders. Vanzelfsprekend geniet het de voorkeur om alle leerdoelen via de open-vraagvorm in een foutenanalyse te betrekken, omdat de inhoudelijke samenstelling van de afleiders een belangrijke functie heeft bij de opbouw van het remediërend onderricht[10].

De gedane vaststellingen bieden tevens perspectieven met betrekking tot de opleiding van itemschrijvers. In dit verband dient vooral aandacht

besteed te worden aan de foutenanalyse, die de itemschrijvers behulpzaam kan zijn bij de inhoudelijke samenstelling van afleiders.

Aantekeningen en literatuur

1. Voor verdere gegevens over dit project, zie: E. De Corte, *Didactische evaluatie van het onderwijs*. (Studia paedagogica, Nieuwe reeks, 1.) Leuven, Univ. Pers Leuven, 1973, p. 68 e.v.
2. W. Lans en G. J. Mellenberg, Constructie en beoordeling van items: formele aspecten. – In: A. D. de Groot, R. F. van Naerssen, e.a., *Studietoetsen: construeren, afnemen, analyseren*. Den Haag, Mouton, 1969, p. 65–89.
3. Zie hierover: E. De Corte en A. Verkens, Didactische foutenanalyse: voorstel van een systematische methode. *Ped. Tijdschr./Forum v. Opvoedk.*, 1976 (1), 553–568. E. De Corte en A. Verkens, Foutenanalyse: didactische bijdrage tot de constructie en de interpretatie van afleiders in meerkeuzevragen. *Ped. Tijdschr./Forum v. Opvoedk.*, 1977 (2), 21–31.
4. A. D. de Groot, R. F. van Naerssen, e.a., *o.c.*, p. 8.
5. De leerdoelen zijn afkomstig uit: E. De Corte, M. Deriemaeker, G. Janssens, J. Jong en G. Tistaert, *Leerdoelstellingen van het rekenonderwijs op de basisschool, niveau einde zesde leerjaar*. Leuven, J. B. Wolters, 1974, 178 pp. De methode die gevolgd werd bij het samenstellen van de inventaris, wordt uitvoerig beschreven in E. De Corte, *Inventariseren van actueel geldende leerdoelen. Ontwikkeling van een empirische methode*. (Studia paedagogica, Nieuwe reeks, 4) Leuven, Univ. Pers. Leuven, 1975, X – 85 pp.
6. Een parametervrije toets is hier aangewezen omdat men met deze toets niet hoeft te onderstellen dat de populaties normaal verdeeld zijn en door eenzelfde standaardafwijking gekarakteriseerd zijn. De toets van Wilcoxon werd toegepast volgens de werkwijze beschreven in: J. C. Spitz, *Statistiek voor psychologen, pedagogen en sociologen*. Amsterdam, Noord-Hollandse Uitg. Mij., 1968 (3de herziene druk), p. 336–343. Deze toets is alleen gevoelig voor verschillen tussen de populaties met betrekking tot de centrummaat. Spitz specificeert niet over welke centrummaat het juist gaat.
7. H. de Jonge en G. Wielenga, *Statistische methoden voor psychologen en sociologen*. Groningen, Wolters, 1963 (2de druk), p. 213–216.
8. Zie: E. De Corte en A. Verkens, Foutenanalyse: didactische bijdrage tot de constructie en de interpretatie van afleiders in meerkeuzevragen. *Ped. Tijdschr./Forum v. Opvoedk.*, 1977 (2), 21–31.
9. Zie hierover: G. Tistaert, Functionele betekenis van itembanking voor didactische doeleinden. *Tijdschr. Opvoedk.*, 1973–74 (19), 211–228. De meerkeuzevragen zijn opgenomen in: G. Tistaert (Ed.), *Meerkeuzevragen voor de evaluatie van het rekenonderwijs, niveau einde zesde leerjaar*. Leuven, J. B. Wolters, 1975, 2 delen, 189 + 198 pp.
10. Zie: E. De Corte en A. Verkens, o.c., p. 26–27.

Curricula vitae

E. De Corte (1941), doctor in de pedagogische wetenschappen (1970), hoogleraar aan de K.U. Leuven in het Departement Pedagogische Wetenschappen, Afdeling Didactiek en Psychopedagogiek met als voornaamste onderwijsopdrachten pedagogische psychologie (bij pedagogiek- en psychologiestudenten) en didactiek (in de lerarenopleiding).

Adres: Pedagogisch Instituut, Dekenstraat 28–30, B–3000 Leuven.

A. Verkens (1949), onderwijzer (1968), licentiaat in de pedagogische wetenschappen, afstudeerrichting Psychopedagogiek (1976).

Adres: Steenweg op Halle 19, B–1684 Leerbeek.