

EEN VERGELIJKENDE STUDIE VAN SCORINGSCriteria
VOOR MEERVOUDIGE CLASSIFICATIE:
EEN DWARSDOORSNEDE- EN EEN LONGITUDINAAL
ONDERZOEK
[A COMPARATIVE STUDY OF SCORING CRITERIA
USED IN MULTIPLE CLASSIFICATION TASKS: A CROSS-SECTIONAL
AND A LONGITUDINAL STUDY]

Johannes KINGMA & Kees P. VAN DEN BOS

University of Alberta
Department of Psychology

&

Rijks Universiteit Groningen
Vakgroep Orthopedagogiek van het in zijn
ontwikkeling belemmerde kind

The relative merits of three types of scoring criteria for multiple classification tasks — the Genevan judgment-plus-explanation criterion, the counter-suggestion criterion, and the non-Genevan judgment-only criterion — have been evaluated in two studies. The 487 children in the cross-sectional study, aged 4 to 11, completed ten multiple classification tasks at a single test administration. The 275 children in the longitudinal study, aged 4 to 7, were administered the same battery of tasks three times with an interval of four months. Scored in accordance with each of the three criteria the battery proves an internally consistent test for multiple classification. In each case, factor analysis yields a single factor and the factor structures show a very high degree of similarity. Furthermore, these factor structures remain stable in the longitudinal study, so that the multiple classification tasks refer to one and the same underlying concept for all three scoring criteria. Both cross-sectional and longitudinal findings indicate that major increments in multiple classification scores occur in the primary school grades 1 and 2, whereas development in preschoolers is minimal. Implications of these findings are discussed. The point is made that, in spite of the excellent psychometric properties of all three criteria, the selection of scoring criteria in future multiple classification research still has to be based on a firm theoretical framework.

Sedert de zestiger jaren is er in de cognitieve ontwikkelingsliteratuur

De hier gerapporteerde research werd voor een deel gesubsidieerd door De Nederlandse Organisatie voor Zuiver Wetenschappelijk Onderzoek (Z.W.O., subsidienr. H 56 - 314) en voor een deel door de National Science and Engineering Research Council (Grant # A 1235). Dit artikel kwam mede tot stand door het uitvoerige commentaar van een reviewer (Dr. Luc Goossens) op eerdere manuscript-versies. De auteurs zijn deze reviewer erkentelijk.

een grote belangstelling te bespeuren naar studies over Piagetiaanse taken (Brainerd, 1978). Een veelvuldig gebruikte soort taak betreft de meervoudige classificatie. In deze taken bestaat de opgave voor het kind erin om in een eenvoudige matrix op basis van rij- en kolomsgewijze gelijkenissen het ontbrekende plaatje uit een reeks alternatieven te kiezen. Echter uit inspectie van een groot aantal studies naar meervoudige classificatie blijkt dat er geen consensus bestaat over het te gebruiken scoringscriterium.

In de ontwikkelingspsychologische literatuur zijn ruwweg drie scoringscriteria te onderscheiden voor de meervoudige classificatie taken: het beoordeling-alleen-, het beoordeling-plus-uitleg- en het tegensuggestie criterium (Kingma, 1981). Het eerste criterium houdt in dat alleen de vraag wordt gesteld welk van de alternatieve plaatjes in het lege vakje van de matrix hoort. Als het correcte plaatje wordt aangewezen dan wordt volgens dit criterium de oplossing als correct bestempeld. Het beoordeling-alleen criterium wordt vrijwel uitsluitend door onderzoekers buiten de Geneefse school gebruikt (zie bijvoorbeeld Almy, 1970; Bell, 1975; Bingham-Newman & Hooper, 1974; Camp, 1975; Farnum, 1976; Hamel, 1974; Hooper, Sipple, Goldman & Swinton, 1979; Kingma, 1981; Lagattuta, 1974; Odom, 1978; Overton & Brodzinsky, 1972; Sigel, 1971). De andere twee criteria zijn gebruikt door Piaget en medewerkers van de Geneefse school (Piaget & Inhelder, 1967; Inhelder & Sinclair, 1969; Inhelder, Sinclair & Bovet, 1974; Pascual Leone & Bovet, 1966) en slechts door enkele onderzoekers buiten de Geneefse school (zie bijvoorbeeld Lovell, Mitchell & Everett, 1962; Smedslund, 1962, 1967a, 1967b). Bij het toepassen van het beoordeling-plus-uitleg criterium wordt een oplossing van een meervoudige classificatietask als correct beschouwd, wanneer het kind naast een correcte keuze ook in staat is een logische verklaring hiervoor te geven, waaruit moet blijken dat het kind de verschillende dimensies van de objecten in de matrix in zijn/haar keuze heeft betrokken (Piaget & Inhelder, 1967). Het tegensuggestie criterium is een uitbreiding van het beoordeling-plus-uitleg criterium. Wanneer een kind zowel een correcte keuze als een logische verklaring heeft gegeven, wordt een tegensuggestieve vraag gesteld, die luidt of er niet een ander plaatje is in de reeks van alternatieve plaatjes dat net zo goed of zelfs nog beter past in de matrix dan het plaatje van de eerste keuze (voor toepassingen van dit criterium zie Inhelder & Sinclair, 1969; Inhelder, Sinclair & Bovet, 1974; Pascual Leone & Bovet, 1966; Smedslund, 1962). Echter in de Geneefse studies worden deze twee criteria niet tegelijk gebruikt (Piaget & Inhelder, 1967). Omdat de ver-

schillende onderzoekers hun studies weliswaar op Piagets theorie baseren, maar desondanks verschillende criteria hanteren voor meervoudige classificatie, dient de vraag naar de functionele equivalentie van de drie scoringscriteria (hun onderlinge uitwisselbaarheid) te worden beantwoord.

De twee te rapporteren studies hadden enigszins verschillende, maar complementaire doelstellingen. Studie 1 was opgezet om via een dwarsdoorsnede benadering voor de drie verschillende criteria normatieve gegevens te verkrijgen over de ontwikkeling van meervoudige classificatie. Studie 2 behelsde een longitudinaal onderzoek naar de ontwikkeling van meervoudige classificatie. Hiertoe werden drie maal gedurende één schooljaar (met intervallen van vier maanden) kleuters en kinderen uit de klassen 1 en 2 van de lagere school meervoudige classificatietaken voorgelegd. Het doel hiervan was na te gaan of transversale verschillen in meervoudige classificatie bevestigd werden door de longitudinale veranderingen binnen een termijn van acht maanden en dit voor de drie scoringscriteria.

STUDIE 1

METHODE

Proefpersonen. De 487 proefpersonen in Studie 1 waren als volgt in 9 leeftijdsniveaus ondergebracht: leeftijdsniveau (1) 4-jarigen ($n = 39$ met een gemiddelde van 4;4 jaar, $SD = 3$ mnd); (2) 5-jarigen ($n = 42$ met een gemiddelde van 5;1 jaar, $SD = 3$ mnd); (3) oudere 5-jarigen ($n = 44$ met een gemiddelde van 5;8 jaar, $SD = 2$ mnd); (4) 6-jarigen ($n = 51$ met een gemiddelde van 6;4 jaar, $SD = 6$ mnd); (5) 7-jarigen ($n = 65$ met een gemiddelde van 7;5 jaar, $SD = 5$ mnd); (6) 8-jarigen ($n = 55$ met een gemiddelde van 8;4 jaar, $SD = 5$ mnd); (7) 9-jarigen ($n = 60$ met een gemiddelde van 9;6 jaar, $SD = 5$ mnd); (8) 10-jarigen ($n = 64$ met een gemiddelde van 10;7 jaar, $SD = 5$ mnd) en (9) 11-jarigen ($n = 67$ met een gemiddelde van 11;7 jaar, $SD = 6$ mnd). De kinderen bezochten vier kleuterscholen en drie lagere scholen te Groningen. De meesten van hen waren afkomstig uit de sociale middenklasse.

Materiaal. Het testmateriaal bestond uit twee proefitems en tien meervoudige classificatie items. In al deze taken zijn plaatjes van objecten in eenvoudige matrices ondergebracht (zie Figuur 1). Een plaatje ontbreekt in zo'n matrix. De taak van het kind is om op basis van gelijkheden en ongelijkheden tussen de plaatjes, het ontbrekende plaatje uit een reeks

van alternatieve plaatjes (rechterzijde Figuur 1) te kiezen. Het is duidelijk dat het subject om tot een juiste oplossing te komen (2.5 voor item 2; 5.2 voor item 5 en 9.3 voor item 9) meer dan één dimensie in ogenschouw moet nemen.

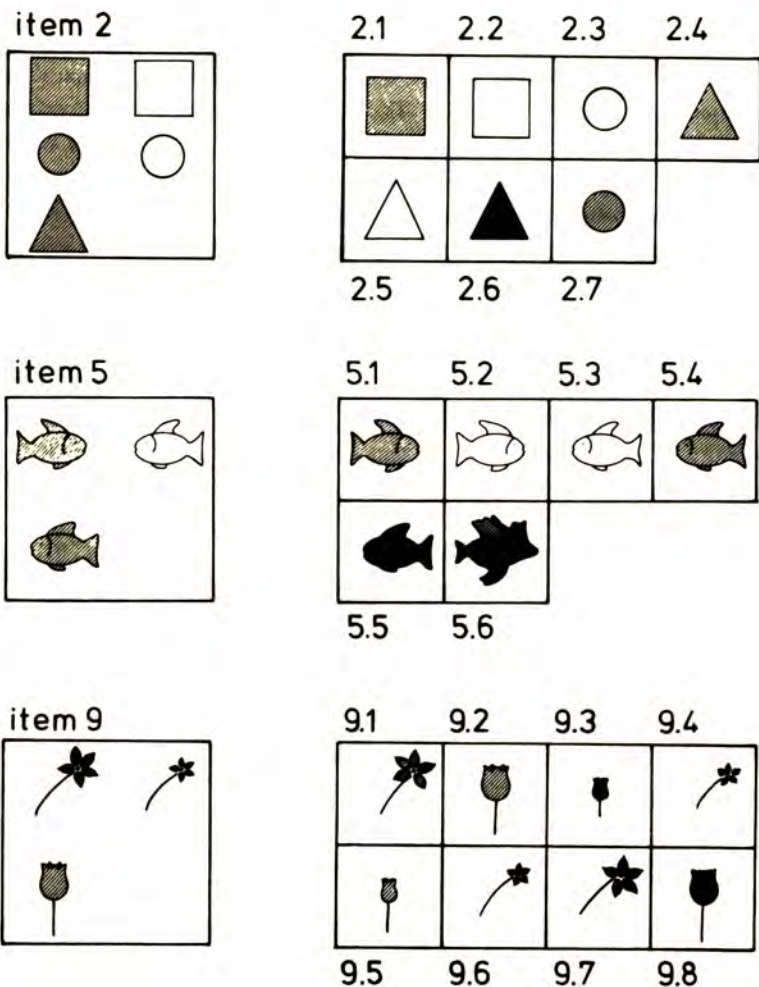


Fig. 1. — Enige voorbeelden van meervoudige-classificatietaken.

Het eerste proefitem bestond uit een 2×2 matrix en een reeks van zes alternatieve plaatjes, en het tweede proefitem bestond uit een 2×3 matrix en eveneens een reeks van zes alternatieve plaatjes. De tien test items bestonden uit zeven 2×2 matrices en twee 2×3 matrices en een

3 × 3 matrix, terwijl het aantal alternatieve plaatjes varieerde per item tussen vier en acht. Elk item was gedrukt op een afzonderlijk vel papier (30 cm breed en 25 cm lang), zoals weergegeven in Figuur 1.

Procedure. De opgaven werden individueel afgenomen door zes ervaren proefleiders. Uit elke leeftijdsgroep werd ongeveer hetzelfde aantal kinderen willekeurig aan elke proefleider toegewezen (door middel van toevalstabellen). Voordat met de testafname werd begonnen, kreeg het kind twee proefopgaven voorgelegd om hem/haar aan de vragen en het stimulus materiaal te laten wennen. De proefleider legde het principe van de meervoudige classificatie opgaven uit door te wijzen op de overeenkomsten en verschillen van de objecten in de matrix en door te benadrukken dat er één plaatje in de matrix ontbrak. Foutieve antwoorden werden niet gecorrigeerd, maar als het kind een incorrect plaatje had gekozen of zijn/haar keuze niet van argumenten wist te voorzien, wees de proefleider nogmaals op de overeenkomsten en verschillen in de dimensies van de objecten in de matrix. Vervolgens werden de 10 meervoudige classificatie opgaven in een willekeurige volgorde afgenomen. In iedere taak werd het kind gevraagd om het juiste plaatje te kiezen uit de reeks van alternatieve plaatjes, dat paste in het lege vakje van de matrix. Hierna vroeg de proefleider het kind om uitleg van de gemaakte keuze: „waarom?“, of „hoe weet je dat?“ Daarna werd de tegensuggestie geïntroduceerd met de vraag: „Kun je een plaatje aanwijzen dat nog beter of net zo goed past in het lege vakje als het plaatje dat je zoëven hebt aangewezen?“

Gedurende de testafname werd het kind door twee proefleiders onafhankelijk geobserveerd. De ene proefleider, waaraan het kind was toegewezen, testte het kind en hij/zij schreef op een antwoordformulier de uitleg die het kind gaf voor zijn/haar keuze (beoordeling). De andere proefleider (die apart zat van het kind en de eerste proefleider) noteerde eveneens de beoordeling, de uitleg en of het kind bestand was tegen de tegensuggestie. Deze dubbele procedure had als doel om de interbeoordelaars betrouwbaarheid vast te stellen. Voordat met het onderzoek in Studie 1 werd begonnen, ontvingen de proefleiders training in het afnemen van de tests en het coderen van de argumenten. Voordat met het eigenlijke testwerk werd begonnen, waren alle proefleiders getraind tot een betrouwbaarheids criterium van meer dan 95 % overeenkomst in het coderen van de argumenten.

Scoring. Drie scoringssystemen werden op de data toegepast: het beoordeling-alleen criterium, het beoordeling-plus-uitleg criterium en

het tegensuggestie criterium. Volgens het *beoordeling-alleen* criterium kreeg een meervoudige classificatie een score één wanneer het kind het correcte plaatje had aangewezen in de reeks van alternatieven. Een opgave kreeg een score nul bij een keuze van een incorrect alternatief. Bij toepassing van het *beoordeling-plus-uitleg* criterium werd aan een meervoudige classificatie opgave een score één toegekend wanneer het kind zowel een correct oordeel als een correcte uitleg van dat oordeel had gegeven. Uit de uitleg moest blijken dat de relevante dimensies in ogenschouw waren genomen, zoals de relevante kenmerken van de objecten in zowel de rij als de kolom van de matrix. Een meervoudige classificatie opgave kreeg een score nul wanneer een onjuiste beoordeling of een inadequate uitleg was gegeven. Bij het hanteren van het *tegensuggestie-criterium* kreeg een meervoudige classificatie opgave een score één wanneer een score één was toegekend volgens het *beoordeling-plus-uitleg* criterium én wanneer het kind zijn/haar keuze *niet* veranderde na de tegensuggestieve vraag. In andere gevallen werd de meervoudige classificatie een score nul gegeven. Samenvattend: elk kind kreeg per taak dus drie scores. Voor elk van de drie scoringssystemen was de maximum totaalscore voor een kind dus 10.

RESULTATEN

Alvorens met de analyse van de resultaten werd begonnen, werden voorafgaande inspecties op de data uitgevoerd. Allereerst werd voor elk paar proefleiders de interbeoordelaars betrouwbaarheid berekend. Voor de beoordeling (de keuze van het juiste plaatje) was 100 % overeenstemming, dit werd eveneens geconstateerd voor het feit of het kind tegen de tegensuggestie bestand was. Daarentegen was 97 % overeenstemming tussen de beoordelaars over het correct zijn van de uitleg voor een correcte keuze van het in de matrix ontbrekende plaatje. De totaalscores werden vervolgens berekend voor elk van de leeftijdsniveaus. De frequentie verdelingen van de gemiddelde totaalscores waren vanaf leeftijdsniveau 1 tot en met leeftijdsniveau 9 positief scheef verdeeld (monotoon stijgend). Daarom werden non-parametrische statistische tests geschikt bevonden voor deze gegevens. Voor de totaalscores volgens elk van de drie scoringscriteria werd geen significant proefleidereffect geconstateerd van de zes groepen kinderen die aan de zes proefleiders waren toegewezen (Kruskal-Wallis H varieerde tussen 1.21 tot 2.03 voor de drie scoringscriteria, $p > .10$). Omdat er geen systematische verschillen werden geconstateerd tussen jongens en meisjes op hun totaalscores

volgens elk van de drie scoringscriteria (Mann-Whitney *U* test, *z*-waarden varieerden tussen 0.88 en 1.11, $p > .10$), refereren de hier verder beschreven resultaten naar de gecombineerde scores van de jongens en de meisjes.

Moeilijkheidsgraad

Voor iedere leeftijdsgroep werden per meervoudige classificatie-opgave de proporties correcte antwoorden (*p*-waarden) voor elk van de drie scoringscriteria berekend. Voor het criterium *beoordeling-alleen* gaf de analyse van de *p*-waarden van de 10 individuele meervoudige classificatie items aan, dat er slechts bij leeftijdsniveau 9 (de 11-jarigen) sprake was van een plafond-effect (*p*-waarde $\geq .90$) voor alle items, terwijl bij de leeftijdsniveaus 7 en 8 (respectievelijk 9 en 10-jarige leeftijd) dergelijke effecten werden geobserveerd voor, respectievelijk, drie en zes items. Bij leeftijdsniveau 1 (de 4-jarigen) werd voor drie items een bodem-effect (*p*-waarde $\leq .10$) gevonden, terwijl de *p*-waarden voor de overige items op dit leeftijdsniveau varieerden van .13 tot .35. Wat betreft het *beoordeling-plus-uitleg* scoringscriterium waren de *p*-waarden voor de individuele items aanzienlijk kleiner dan volgens het bovengenoemde criterium. Op leeftijdsniveau 1 tot en met 3 (de 4-jarigen tot en met de oudere 5-jarigen) varieerden de *p*-waarden van .00 tot .05, terwijl op leeftijdsniveau 4 (de 6-jarigen) nog een bodem-effect werd gevonden voor zeven items. Vanaf leeftijdsniveau 5 (de 7-jarigen) tot 9 (de 11-jarigen) werden bovendien geen plafond-effecten geconstateerd. Op leeftijdsniveau 9 varieerden de *p*-waarden tussen .59 en .86. De *p*-waarden voor de individuele items gescoord volgens het *tegensuggestie* criterium vertoonden voor de leeftijdsniveaus 1 tot en met 4 eveneens een bodemeffect voor *alle* items, terwijl in de overige leeftijdsgroepen geen plafond-effecten werden vastgesteld (de *p*-waarden van de items waren kleiner dan de waarden volgens het *beoordeling-plus-uitleg* criterium). De betekenis van de analyse van de *p*-waarden is dat zij inzicht geeft hoe goed de 10-item test differentieert in de onderzochte groep. Het blijkt dus dat de differentiatie van de 10-item test sterk afhangt van het scoringscriterium dat wordt gehanteerd. Voor wat betreft het *beoordeling-alleen* criterium, toont de *p*-waarde analyse aan dat de test goed differentieert, zowel in de kleuterschool als in de hogere klassen van de basisschool (leeftijdsniveaus 7 en 8). Echter, voor de twee Geneefse criteria komt naar voren dat de test te moeilijk is voor de kleuterschool (alleen bodem-effecten), maar dat zij goed differentieert in de lagere school (geen plafond-effecten).

Op het eerste gezicht lijkt het beoordeling-alleen criterium een test-batterij voor meervoudige classificatie op te leveren die in tegenstelling tot de twee Geneefse criteria geschikt lijkt voor de kleuterschool en alle klassen van de lagere school. Echter, voor wat betreft de kleuterschool (leeftijdsniveau 1 tot en met 3) dient dan wel de kanttekening te worden gemaakt dat de gemiddelde scores niet boven de gokkans-verwachte waarden uitkomen. Immers, het aantal alternatieven in de reeks plaatjes varieerde van vier tot acht in de tien classificatieopgaven (2 items met 4 alternatieve plaatjes, 3 items met 6, 2 items met 7, 3 items met 8). Op basis van gokkans is de verwachte waarde van de tien items $(1/4 \times 2) + (1/6 \times 3) + (1/7 \times 2) + (1/8 \times 3) = 1.67$. De leeftijdsniveaus 1 tot en met 3 (zie Tabel 1) hebben volgens dit criterium een gemiddelde score van respectievelijk 1.4, 1.7 en 1.8 (voor de longitudinale Studie 2 zijn de gemiddelden respectievelijk 1.1, 1.8 en 1.8 voor deze leeftijdsniveaus op de eerste afname), hetgeen nauwelijks afwijkt van de op gokkans verwachte waarde. Hieruit mogen we concluderen dat ook voor het criterium beoordeling-alleen de testbatterij minder geschikt lijkt voor de kleuterschool. De testbatterij volgens dit criterium is geschikt voor alle klassen van de lagere school, omdat de gemiddelde scores in de leeftijdsniveaus 5 tot en met 9 duidelijk boven de op gokkans verwachte waarde uitkomen.

Psychometrie

Ten behoeve van de betrouwbaarheidsbepaling werden voor ieder scoringscriterium item-rest correlaties en de Kuder Richardson interne consistentie index (KR_{20}) berekend. Volgens Nunnally (1978) zijn een $KR_{20} > .80$ en item-restcorrelaties van $r > .50$ bevredigend. Bij het beoordeling-alleen criterium werd een $KR_{20} = .91$ gevonden en de item-rest correlaties varieerden van .51 tot .77. Ook bij het beoordeling-plus-uitleg en het tegensuggestie criterium werden hoge KR_{20} waarden gevonden (respectievelijk .93 en .94) en de item-rest correlaties varieerden respectievelijk van .63 tot .70 en van .61 tot .84.

Factoranalyse

Met behulp van factoranalyse kan de structurele relatie tussen de 10 meervoudige classificatietaken worden onderzocht. Factoranalyse (principale componenten) op de totaalscore van de ppn van de gehele steekproef ($n = 487$) volgens het beoordeling-alleen criterium, gaf een eerste hoofdcomponent met 57 % verklaarde variantie te zien. De

overige hoofdcomponenten hadden eigenwaarden kleiner dan 1, en op de grond van Kaisers (1961) criterium werd geen Varimax rotatie toegepast. Gelijksortige resultaten werden verkregen bij de factoranalyse van de scores volgens het beoordeling-plus-uitleg en het tegensuggestie criterium. Hier werden eerste hoofdcomponenten, die resp. 62 % en 67 % van de variantie verklaarden, gevonden en de eigenwaarden van de tweede en verdere hoofdcomponenten waren steeds kleiner dan 1. Nunnally's (1978) gezichtspunt toepassend, mogen we concluderen dat de tien meervoudige-classificatietaken voor ieder van de scoringscriteria een intern consistente test leveren. Vervolgens werd de overeenstemming tussen de factorstructuren bij de drie scoringscriteria onderzocht. Hier-toe werd paarsgewijs op de factorladingen matrices een orthogonale procrustes rotatie volgens het kleinste-kwadraten criterium toegepast, waarna de congruentie-coëfficiënt van Tucker (zie Ten Berge, 1977a) verkregen werd. De waarden van deze coëfficiënt varieerden van .96 tot .99, hetgeen betekent dat de factorstructuren voor de drie scoringscriteria een grote mate van overeenkomst opleveren (zie Haven & Ten Berge, 1977; Ten Berge, 1977b).

Criteriumverschillen

Analyses van gemiddelden (Tabel 1) gaven aan, dat op elk leeftijds-niveau de gemiddelden van de tien meervoudige classificatie items bij het hanteren van het criterium beoordeling-alleen significant hoger waren (Wilcoxon z -waarden, per leeftijdsniveau, varieerden van $z = 3.68$ tot $z = 6.48$, $p < .01$) dan bij de andere criteria. De gemiddelde scores bij het beoordeling-plus-uitleg criterium waren significant hoger dan bij het tegensuggestie criterium op de leeftijdsniveaus 4 tot en met 9 (Wilcoxon z -waarden varieerden van 3.14 tot 4.29, $p < .01$). Op de lagere leeftijds-niveaus werden uiteraard voor de twee laatstgenoemde Geneefse criteria,

Tab. 1. — Gemiddelde scores van de tien meervoudige classificatietaken als functie van leeftijdsniveaus en scoringscriteria voor de Dwarsdoorsnede Studie I ($n = 487$)

scoringscriterium	leeftijdsniveaus								
	1	2	3	4	5	6	7	8	9
beoordeling-alleen	1.4	1.7	1.8	2.8	5.8	7.1	8.2	8.7	8.9
beoordeling-plus-uitleg	0.0	0.1	0.1	0.7	3.6	4.7	6.6	7.5	7.5
tegensuggestie	0.0	0.0	0.1	0.7	2.2	3.3	5.0	6.0	6.1

vanwege de eerder gerapporteerde bodem-effecten, geen significante verschillen aangetroffen. In het algemeen dan, blijkt het tegensuggestie cri-

terium (TS) strenger (minder antwoorden worden als correcte uitingen van meervoudige classificatie beschouwd) dan het beoordeling-plus-uitleg criterium (BPU), terwijl het laatste criterium weer strenger is dan het beoordeling-alleen criterium (BA). Het is evident dat $TS = BPU +$ nog iets anders, en dat $BPU = BA +$ nog iets anders.

Leeftijdsverschillen

Voor elk leeftijdsniveau was per verschillend scoringscriterium de gemiddelde score over de tien meervoudige-classificatietaken berekend (zie Tabel 1). Bij het beoordeling-alleen criterium zijn de gemiddelde scores van de leeftijdsniveaus 1 en 3 bijna identiek, terwijl van leeftijdsniveau 4 tot en met leeftijdsniveau 9 een geleidelijke toename van de gemiddelden waargenomen kan worden, hetgeen een transversale ontwikkelingstrend suggereert. Ook voor het beoordeling-plus-uitleg- en het tegensuggestiecriterium wezen toetsingen van gemiddelden (Tabel 1) uit, dat vanaf leeftijdsniveau 3 tot leeftijdsniveau 8 transversale ontwikke-

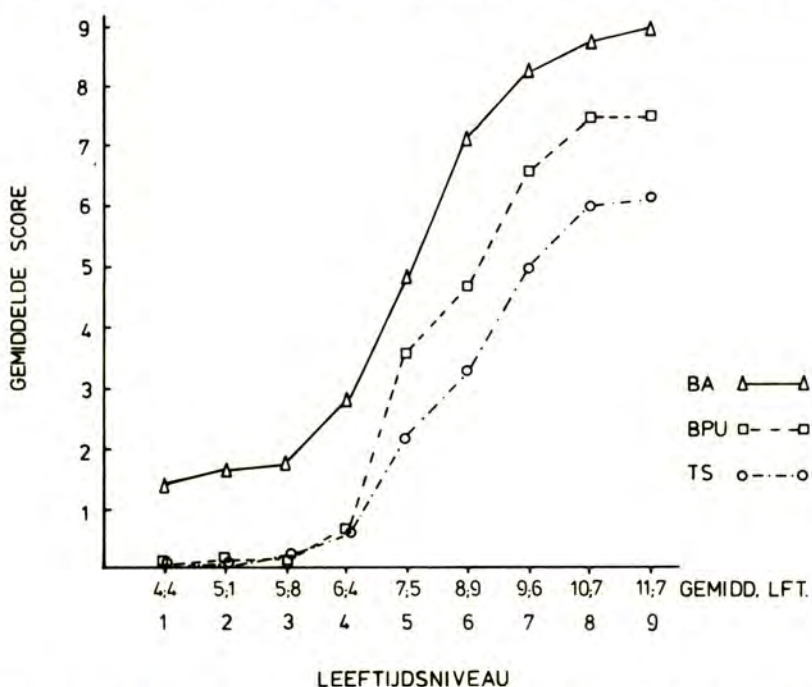


Fig. 2. — Een grafische weergave van de ontwikkelingscurves volgens de drie scoringscriteria voor de leeftijdsniveaus 1 tot en met 9 uit het dwarsdoorsnede onderzoek (BA = Beoordeling Alleen; BPU = Beoordeling-Plus-Uitleg; TS = Tegen Suggestie).

lingstrends bestaan. De resultaten voor de drie criteria geven dus aan, dat de ontwikkeling van meervoudige classificatie zich vooral in de eerste klassen van de lagere school voordoet en niet in de kleuterschoolleeftijd. In de hoogste jaren van de lagere school lijkt de ontwikkeling te vertragen (de trend zwakt wat af). In de vorige paragraaf werd aangegeven dat de drie scoringscriteria verschillen in strengheid. Toch blijken de transversale ontwikkelingspatronen van de drie criteria vrijwel identiek qua vorm. Dit wordt in Figuur 2 grafisch weergegeven.

Samenvattend kunnen we stellen dat ieder van de drie scoringscriteria een intern consistente test voor meervoudige classificatie vormt en bovendien overeenkomstige factorstructuren voor de verschillende scoringscriteria. Het blijkt dat het tegensuggestie criterium strenger is dan het beoordeling-plus-uitleg criterium, terwijl het laatste weer stringenter is dan het beoordeling-alleen criterium. De transversale ontwikkelingscurves vertonen echter qua vorm een grote overeenkomst.

STUDIE 2

METHODE

Proefpersonen. De 275 proefpersonen in Studie 2 waren als volgt verdeeld over 5 leeftijdsniveaus: (1) 4-jarigen ($n = 43$) met een gemiddelde van 4;6 jaar, $SD = 4$ mnd); (2) 5-jarigen ($n = 54$ met een gemiddelde van 5;0 jaar, $SD = 2$ mnd); (3) de oudere 5-jarigen ($n = 50$ met een gemiddelde van 5;8 jaar, $SD = 2$ mnd); (4) 6-jarigen ($n = 61$ met een gemiddelde van 6;6 jaar, $SD = 5$ mnd); en (5) 7-jarigen ($n = 67$ met een gemiddelde van 7;6 jaar, $SD = 5$ mnd). Voor de goede orde, deze steekproef vertoonde geen overlap met de steekproef uit Studie 1. De kinderen uit Studie 2 bezochten zes kleuterscholen en vier lagere scholen te Groningen. De meesten van hen waren afkomstig uit de sociale middenklasse.

Materiaal. Het testmateriaal bestond uit dezelfde twee proefitems en tien meervoudige classificatie items als in Studie 1 (zie beschrijving aldaar).

Procedure. De opgaven werden op dezelfde wijze afgenomen als in Studie 1 en dezelfde zes ervaren proefleiders uit Studie 1 namen de tests af. Ook werden voor iedere testafname (drie keer per jaar, met een interval van 4 maanden) de kinderen willekeurig aan een proefleider toegewezen (met de restrictie dat elke proefleider per leeftijdsniveau ongeveer hetzelfde aantal kinderen moest testen). Evenals in Studie 1 werd het kind gedurende de testafname onafhankelijk geobserveerd door twee proef-

leiders, zodat voor elke testafname de interbeoordelaars betrouwbaarheid kon worden berekend. De eerste testafname werd in augustus aan het begin van het schooljaar uitgevoerd, de tweede testafname vond vier maanden na de eerste afname plaats in januari en de laatste testafname werd vier maanden na de tweede uitgevoerd in de maand mei.

Scoring. Evenals in Studie 1 werden in deze longitudinale Studie 2 de drie verschillende scoringscriteria gebruikt: het beoordeling-alleen, het beoordeling-plus-uitleg- en het tegensuggestie criterium (zie beschrijving aldaar).

RESULTATEN

Allereerst werd een aantal voorafgaande inspecties op de data uitgevoerd. De interbeoordelaars betrouwbaarheid voor elk paar proefleiders werd voor iedere testafname berekend. Een perfecte overeenstemming werd gevonden voor de beoordeling (de keuze van het plaatje) en eveneens voor de reactie op de tegensuggestieve vraag voor alle drie testafnames. De interbeoordelaars-betrouwbaarheid voor de interpretatie van de uitleg voor een correcte keuze varieerde voor de drie testafnames tussen 95 % en 97 % overeenstemming. De gemiddelde totaalscores vertoonden van leeftijdsniveaus 1 tot en met 5 een min of meer monotoon stijgende verdeling, vandaar dat in de verdere analyse van de resultaten van deze studie non-parametrische statistische tests worden gebruikt. Er werden geen significante proefleider-effecten gevonden voor de scores van de zes groepen kinderen, die per testafname aan de zes proefleiders waren toegewezen, ongeacht welk van de drie scoringscriteria werd toegepast (Kruskal-Wallis H -waarden varieerden tussen 0.56 en 1.98 over de verschillende testafnames en de drie scoringscriteria, $p > .10$). Geen systematische verschillen tussen jongens en meisjes werden geconstateerd voor hun totaalscores op de tien meervoudige classificatietaken per testafname en per scoringscriterium (Mann-Whitney U test, z -waarden varieerden tussen 0.41 en 1.30, $p > .09$). Daarom worden in de nu verder beschreven resultaten de scores van de jongens en de meisjes gecombineerd.

Moeilijkheidsgraad

De p -waarden voor de tien meervoudige classificatie opgaven gescoord volgens het criterium *beoordeling-alleen* vertoonden op de drie testafnames geen plafond-effect (p -waarde $> .90$) in een van de leeftijdsni-

veaus. Bij leeftijdsniveau 1 (de vierjarigen) werd voor vier items een bodem-effect geconstateerd op de eerste en tweede testafname, terwijl op de derde testafname voor drie items bodem-effecten werden vastgesteld. Voor wat betreft het *beoordeling-plus-uitleg* criterium werd voor de leeftijdsniveaus 1 tot en met 3 voor alle items een bodem-effect gevonden voor de drie testafnames, terwijl geen plafond-effecten werden geconstateerd voor de leeftijdsniveaus 4 en 5 op de drie opeenvolgende testafnames. Bij toepassing van het *tegensuggestie* criterium werden eveneens voor de leeftijdsniveaus 1 tot en met 3 voor alle items bodem-effecten gevonden. Leeftijdsniveau 4 vertoonde op de eerste en tweede testafname bodemeffecten voor acht items, terwijl op de derde testafname vier items een bodemeffect hadden. Voor leeftijdsniveau 5 werd op de drie testafnames voor één item een bodem-effect vastgesteld, terwijl de overige 9 items geen plafond-effect vertoonden.

In het algemeen stijgen de *p*-waarden van de items in de opeenvolgende test-afnames voor alle drie criteria in de leeftijdsniveaus 4 en 5 (de 6- en 7-jarigen). Deze toename werd ook geconstateerd voor het criterium *beoordeling-alleen* voor de leeftijdsniveaus 1 tot en met 3, terwijl voor de twee Geneefse criteria de toename van de *p*-waarden in deze lagere leeftijdsniveaus te vewaarlozen is op de drie opeenvolgende testafnames. Opnieuw blijkt uit de analyse van de *p*-waarden, evenals in ons dwarsdoorsnede onderzoek, dat de differentiatie van de 10-item test sterk afhangt van het scoringscriterium dat wordt gehanteerd. Het blijkt dat bij toepassing van het criterium *beoordeling-alleen*, de *p*-waarden van alle tien meervoudige classificatie items toenemen in de opeenvolgende testafnames. Eenzelfde conclusie werd getrokken in het dwarsdoorsnede onderzoek. De bevinding in het longitudinale onderzoek dat voor de twee Geneefse scoringscriteria voor de leeftijdsniveaus 1 tot en met 3 nauwelijks een toename van de *p*-waarden werd geconstateerd (maar wel voor leeftijdsniveaus 4 en 5) is eveneens in overeenstemming met het resultaat van de analyse in het dwarsdoorsnede onderzoek, namelijk: dat deze twee Geneefse criteria een test opleveren die te moeilijk is voor de kleuterschool, maar wel goed differentieert op de lagere school. Met andere woorden: de resultaten van de *p*-waarden analyse van het dwarsdoorsnede onderzoek en die van het longitudinale onderzoek convergeren.

Psychometrie

Voor de drie afzonderlijke opeenvolgende afnames van de tien meer-

voudige classificatie taken bij het beoordeling-alleen criterium, varieerden de gevonden KR_{20} waarden van .86 tot .89. De item-rest correlaties varieerden bij dit criterium van .52 tot .79. Bij het beoordeling-plus-uitleg criterium varieerden de KR_{20} waarden en de item-rest correlaties respectievelijk tussen .87 en .90, en tussen .53 en .82. Bij het tegensuggestie criterium varieerden de KR_{20} waarden en de item-rest correlaties respectievelijk tussen .87 en .91, en tussen .51 en .87. Ook in het longitudinale onderzoek werden (evenals in het dwarsdoorsnede onderzoek) hoge interne consistente tests verkregen voor ieder van de drie criteria in de drie opeenvolgende testafnames.

Factoranalyse

Factoranalyse van de totaalscore (over tien taken) bij het beoordeling-alleen criterium voor alle ppn ($n = 275$) gaven op alle drie afnames voor de tweede en verdere hoofdcomponenten eigenwaarden kleiner dan 1 te zien. De eerste hoofdcomponenten op de drie afnames verklaarden daarentegen resp. 50 %, 58 % en 60 % van de variantie. Gelijksortige resultaten werden verkregen bij het beoordeling-plus-uitleg en het tegensuggestie criterium. Bij het eerstgenoemde criterium verklaarden de eerste hoofdcomponenten voor de drie afnames resp. 52 %, 55 % en 50 % van de variantie. Bij het laatstgenoemde criterium waren deze percentages respectievelijk 52 %, 52 % en 57 %. We kunnen concluderen, dat bij elk van de scoringscriteria op ieder van de afnamemomenten binnen één jaar, de tien taken te zamen een vrij intern consistente en homogene test vormen.

Vervolgens werd, net als in Studie 1, de overeenkomst onderzocht tussen de factorstructuren voor de verschillende scoringscriteria, in dit geval per afname. Op de eerste afname varieerden de berekende coëfficiënten van Tucker (1951) op de criteriumparen van .90 tot .96. Gelijksortige resultaten werden verkregen op de tweede en derde testafname. Hierna werd de overeenkomst nagegaan tussen de factorstructuren voor de verschillende afnames per scoringscriterium. Bij het beoordeling-alleen criterium varieerden de berekende coëfficiënten van Tucker op de afname-paren van .97 tot .98. Gelijksortige hoge coëfficiënt-waarden (variërend van .96 tot .99) werden gevonden bij de andere scoringscriteria.

Deze resultaten tonen een grote mate van overeenkomst in factorstructuren aan zowel voor de verschillende scoringscriteria per afname, als voor de verschillende afnames per scoringscriterium. Opnieuw blijkt

dat deze longitudinale bevindingen een bevestiging vormen voor de resultaten van ons dwarsdoorsnede onderzoek, waarin op één tijdstip een grote mate van overeenkomst werd geconstateerd tussen de factorstructuren volgens de drie scoringscriteria.

Criteriumverschillen

Tabel 2 geeft een overzicht van de gemiddelde scores voor de tien meervoudige classificatietaken. De analyses toonden aan dat op elk leeftijdsniveau de gemiddelden van de tien meervoudige classificatie-items bij het criterium beoordeling-alleen significant groter waren (Wilcoxon z -waarden varieerden van $z = 3.75$ tot $z = 8.64$, $p < .01$, voor de drie testafnames) dan voor de testafnames per leeftijdsniveau van de andere criteria. De gemiddelde scores volgens het beoordeling-plus-uitleg criterium waren significant groter dan die voor het tegensuggestie criterium, voor alleen leeftijdsniveau 5 op alle drie testafnames (Wilcoxon z -waarden lagen tussen $z = 2.79$ en $z = 3.23$, $p < .01$), terwijl voor de derde afname een significant verschil werd gevonden voor leeftijdsniveau 4 (Wilcoxon $z = 2.46$, $p < .01$). Tevens gaven de analyses aan, dat gedurende een periode van ongeveer acht maanden (de periode van de drie testafnames) voor de leeftijdsniveaus 1 tot en met 3 geen significante verschillen tussen de twee Geneefse criteria worden geconstateerd, maar dit is vanwege de eerder vastgestelde bodem-effecten van de items dan

Tab. 2. — Gemiddelde scores van de tien meervoudige classificatietaken als functie van leeftijdsniveaus en scoringscriteria voor de Longitudinale Studie 2 ($n = 275$)

scoringscriterium	leeftijdsniveaus				
	1	2	3	4	5
beoordeling alleen					
1 ^e afname	1.1	1.8	1.8	2.9	5.8
2 ^e afname	1.6	1.8	2.2	4.2	6.5
3 ^e afname	1.8	1.9	2.7	5.6	7.3
beoordeling-plus-uitleg					
1 ^e afname	0.0	0.0	0.0	1.0	3.7
2 ^e afname	0.0	0.2	0.4	2.2	4.0
3 ^e afname	0.2	0.2	0.9	3.8	4.8
tegensuggestie					
1 ^e afname	0.0	0.0	0.0	0.8	2.3
2 ^e afname	0.0	0.0	0.4	1.5	2.8
3 ^e afname	0.0	0.1	0.7	2.5	3.5

ook niet verwonderlijk. Wel kunnen we stellen dat de resultaten van dit longitudinale onderzoek een ondersteuning vormen voor de eerdere

bevindingen uit ons dwarsdoorsnede onderzoek, namelijk, het beoordeling-alleen criterium is het minst stringente criterium, terwijl het tegen-suggestie criterium het meest stringente criterium is, en dat het beoordeling-plus-uitleg criterium een middenpositie inneemt. Eveneens blijkt uit de longitudinale bevindingen dat in een periode van ongeveer acht maanden de onderlinge relatie tussen de drie scoringscriteria stabiel blijft.

Leeftijdsveranderingen

Vervolgens werd per scoringscriterium en per leeftijdsniveau de longitudinale vooruitgang over de testafnames geanalyseerd (zie Tabel 2). Voor het criterium beoordeling-alleen waren voor de leeftijdsniveaus 1 tot en met 3, geen significante vooruitgangen tussen de testafnames te bespeuren (Wilcoxon z -waarden varieerden tussen 0.22 en 1.21, $p > .10$). Significante vooruitgang in de gemiddelde scores op de tien meervoudige classificatietaken volgens het criterium beoordeling-alleen werd geconstateerd tussen de drie opeenvolgende testafnames voor leeftijdsniveau 4 en tussen de eerste en tweede testafname voor leeftijdsniveau 5. Voor geen van de twee Geneefse criteria werd een significante longitudinale vooruitgang op de gemiddelde scores vastgesteld voor de leeftijdsniveaus

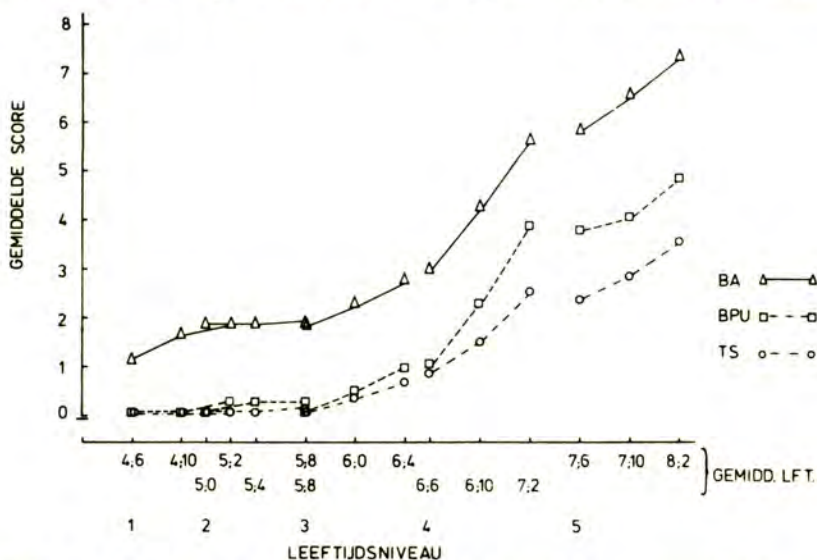


Fig. 3. — Een grafische weergave van de ontwikkelingscurves per scoringscriterium voor de drie testafnames voor de leeftijdsniveaus 1 tot en met 5 uit de longitudinale studie (BA = Beoordeling Alleen; BPU = Beoordeling-Plus-Uitleg; TS = Tegen Suggestie).

1 tot en met 3. Voor het beoordeling-plus-uitleg criterium was voor leeftijdsniveau 4 in de opeenvolgende testafnames een significante vooruitgang te bespeuren, terwijl voor leeftijdsniveau 5 de vooruitgang tussen de eerste en derde testafname significant was. Voor het tegensuggestie criterium werd alleen tussen de eerste en derde testafname bij de leeftijdsniveaus 4 en 5 een significante longitudinale vooruitgang vastgesteld. In Figuur 3 worden deze vooruitgangen grafisch weergegeven.

Alhoewel de drie criteria qua stringentie verschillen, is uit Figuur 3 af te lezen dat qua vorm een vrijwel identiek ontwikkelingspatroon wordt verkregen voor meervoudige classificatie. De statistische toetsing ondersteunde het grafische beeld dat er niet of nauwelijks een longitudinale vooruitgang wordt geboekt voor de leeftijdsniveaus 1 tot en met 3, en wel voor leeftijdsniveaus 4 en 5, met name tussen de eerste en derde testafname (een tijdsbestek van ongeveer acht maanden). Een en ander betekent een belangrijke convergentie met de resultaten van de transversale Studie 1. Ook wanneer men in de kleuterschool en de klassen 1 en 2 van de lagere school de kinderen gedurende het grootste deel van het schooljaar volgt, doen de belangrijkste evoluties inzake meervoudige classificatie zich voor op de lagere school en niet in de kleuterklassen.

Vergelijking van de Figuren 2 en 3 en/of van de Tabellen 1 en 2 (respectievelijk in horizontale en in verticale richting te lezen) laat bovendien een grote overeenkomst zien tussen de gemiddelden die in de twee studies verkregen werden voor de jongste leeftijdsniveaus. Meer specifiek: de gemiddelde leeftijden van de vijf eerste leeftijdsniveaus in het longitudinale onderzoek na 8 maanden, komen bijzonder goed

Tab. 3. — Een overzicht van de gemiddelde scores op de tien meervoudige classificatietaken per leeftijdsgroep uit de longitudinale studie op de derde testafname en de gemiddelde scores voor de leeftijdsniveaus 2 tot en met 6 uit het dwarsdoorsnede onderzoek

criterium	leeftijdsniveaus									
	1 ^a	(2) ^b	2 ^a	(3) ^b	3 ^a	(4) ^b	4 ^a	(5) ^b	5 ^a	(6) ^b
beoordeling- alleen	1.8	(1.7)	1.9	(1.8)	2.7	(2.8)	5.6	(5.8)	7.3	(7.1)
beoordeling- plus-uitleg	0.2	(0.1)	0.2	(0.1)	0.9	(0.7)	3.8	(3.6)	4.8	(4.7)
tegen- suggestie	0.0	(0.0)	0.1	(0.1)	0.7	(0.7)	2.5	(2.2)	3.5	(3.3)

Noot. Het leeftijdsniveau gemarkeerd met een a verwijst naar een niveau uit het longitudinale onderzoek; een b verwijst naar het bijbehorende leeftijdsniveau uit het dwarsdoorsnede onderzoek.

overeen met de gemiddelde leeftijden van de kinderen uit de leeftijdsniveaus 2 tot en met 6 uit het dwarsdoorsnede onderzoek. In Tabel 3 wordt daarom een overzicht gegeven van de gemiddelde scores van deze corresponderende leeftijdsniveaus. Paarsgewijze toetsingen per criterium van verschillen tussen de scores van de corresponderende leeftijdsniveaus leverde geen significante verschillen op (Mann-Whitney U test; z -waarden varieerden tussen 0.00 en 0.23, $p > .10$). De beide benaderingen (dwarsdoorsnede en longitudinaal) leveren dus zeer vergelijkbare gemiddelde scores op van de leeftijdsniveaus 2 tot en met 6.

Ook om een andere reden zijn deze resultaten interessant. Omdat de resultaten van longitudinaal onderzoek vaak opgedreven worden door de herhaalde afname van taken (praktijk-effect), werd door Kingma (1981, 1983b) in zijn longitudinale onderzoek een controlegroep per testafname gebruikt om deze effecten te evalueren. Kingma vond geen aanwijzingen dat de kinderen door herhaald testen praktijk-effecten vertoonden. Gezien de overeenkomst tussen de transversale en longitudinale bevindingen van de huidige studies, durven we te stellen dat hier evenmin praktijk-effecten in het longitudinale onderzoek waarschijnlijk zijn.

DISCUSSIE

De belangrijkste doelen van de twee gerapporteerde studies waren het beantwoorden van de vragen over (a) de functionele equivalentie van de drie scoringscriteria die veelvuldig worden gebruikt in onderzoek naar de ontwikkeling van meervoudige classificatie, en (b) het controleren van de resultaten van de dwarsdoorsnede-studie via een studie waarin de longitudinale methode werd toegepast. De voornaamste onderzoeksresultaten waren: (1) de 10-item test voor meervoudige classificatie, gescoord volgens de drie criteria, heeft een hoge interne consistentie; (2) de tien classificatie-opgaven verwijzen voor alle drie scoringscriteria naar één onderliggend concept (één hoofdcomponent); (3) er bestaat een grote mate van overeenkomst tussen de factorladingen-matrices van elk der drie criteria op de (ene) afname in het dwarsdoorsnede-onderzoek, en (4) op elke afname in het longitudinale onderzoek; (5) de leeftijdsverschillen en -veranderingen die in de twee studies werden aangetoond, vertoonden een grote overeenkomst. Gezien deze overeenkomst, kan gesteld worden dat de ontwikkeling van het meervoudige-classificatie begrip zich niet in de kleuterschool maar wel in de eerste jaren van de lagere school voordoet en kunnen de gemiddelden voor de jongste leef-

tijdsniveaus (2 tot en met 6) in Tabel 1 als (ruwe) normen voor Nederlandse kinderen gebruikt worden. Door het ontbreken van longitudinale gegevens in de hogere klassen van de lagere school, is de afzwakking in de ontwikkeling die zich daar lijkt voor te doen, slechts als een voorlopige bevinding te beschouwen. Derhalve kunnen de gemiddelden voor de drie hoogste leeftijdsniveaus in Tabel 1 slechts als voorlopige aanduidingen gehanteerd worden.

Met betrekking tot uitkomst (1) kan worden toegevoegd, dat de interne consistentie zeer stabiel bleef over de opeenvolgende batterij-afnames in de longitudinale studie (Studie 2) bij alle scoringscriteria. Dus, zelfs bij het kleinere leeftijdsbereik van de longitudinale studie, wordt, vergeleken met het dwarsdoorsnede onderzoek, de interne consistentie van de tien taken nauwelijks beïnvloed. Voor wat betreft het beoordeling-alleen criterium is dit resultaat in overeenstemming met de bevindingen uit zowel voorafgaand dwarsdoorsnede onderzoek (Kingma, 1981, 1983a, 1983b, 1984; Kingma & Koops, 1983) als longitudinaal onderzoek (Kingma, 1981, 1983b). Voor zover we weten is de interne consistentie van de twee Geneefse criteria niet of nauwelijks onderzocht in eerder dwarsdoorsnede onderzoek naar meervoudige classificatie. Een uitzondering is de studie van Kingma (1981) die voor de twee Geneefse scoringscriteria eveneens een hoge interne consistentie vaststelde. Longitudinaal onderzoek naar meervoudige classificatie volgens de twee Geneefse criteria is ons niet bekend, vandaar het belang van onze bevinding dat tijdens de opeenvolgende testafnames in het longitudinaal onderzoek de interne consistentie van de tien meervoudige classificatie opgaven stabiel bleef bij toepassing van deze twee scoringscriteria.

Voor wat betreft de onderzoeksuitkomsten (2), (3) en (4), die een hoge overeenkomst in factorstructuur aangaven bij de drie criteria en tevens stabiliteit in de longitudinale studie, mogen we concluderen dat bij alle drie scoringscriteria de tien meervoudige classificatie-opgaven naar eenzelfde onderliggend concept verwijzen. Met andere woorden: de drie criteria convergeren qua factoriële validiteit. Dit is een opmerkelijk resultaat omdat in eerder onderzoek met andere Piagetiaanse taken (zoals, bijvoorbeeld, conservatie-, seriatie- en transitiviteits-opgaven) de Geneefse criteria zowel onduidelijke factorstructuren opleveren als inconsistente tests (Kingma, 1981, 1982, 1983c, 1983d, 1984; Kingma & Koops, 1985). Alhoewel de twee Geneefse criteria stringenter zijn dan het beoordeling-alleen criterium, is er een grote mate van convergentie voor de drie scoringscriteria te bespeuren. In de eerste plaats is de testbatterij

voor alle drie criteria intern consistent en geschikt voor alle klassen van de lagere school. In de tweede plaats vertonen de ontwikkelingsverlopen volgens elk van de drie criteria een overeenkomst qua vorm, dat wil zeggen, dat met toename van de leeftijd een toenemend aantal classificatie-opgaven correct wordt opgelost. De keuze tussen de drie scoringscriteria zal daarom op andere dan zuiver psychometrische overwegingen gebaseerd moeten worden.

Aansluitend bij de discussie over het criteriumprobleem in de neo-Piagetiaanse research, zouden volgens Brainerd (1973) bij toepassing van het Geneefse criterium beoordeling-plus-uitleg zogenaamde Type II fouten (iemand als non-classificeerder bestempelen, terwijl hij/zij hier toe wel in staat is) in het experiment worden geïntroduceerd. Om niet onnodig stringent te zijn en om bovendien Type II fouten te vermijden, opteert Brainerd voor toepassing van het minst restrictieve criterium beoordeling-alleen om het minimaal noodzakelijke bewijs te verkrijgen dat het kind de opgave aankan. Anderzijds wordt door andere onderzoekers die pleiten voor toepassing van een Geneefs criterium (zie bijvoorbeeld Reese & Schack, 1974) geponeerd, dat het beoordeling-alleen criterium gevoeliger is voor Type I fouten (iemand als classificeerder bestempelen, terwijl hij/zij dat niet is). Waarschijnlijk hebben beide kampen een deel van het gelijk. Immers wij constateerden reeds (zie de paragraaf „moeilijkheidsgraad” van Studie 1) bij toepassing van het criterium beoordeling-alleen, dat de scores van de kinderen uit de kleuterschool niet boven de op gokkans verwachte waarde uitstegen. In feite zou men bij toepassing van het criterium beoordeling-alleen dus steeds moeten nagaan of de gemiddelde scores groter zijn dan op basis van gokkans is te verwachten. Nu zou men verder kunnen aanvoeren dat men de scores dient te corrigeren voor gokkans, door bijvoorbeeld per leeftijdsniveau de gemiddelde score te verminderen met de op basis van gokkans verwachte waarde. Het gevaar bestaat dan echter dat het criterium beoordeling-alleen onnodig stringenter wordt gemaakt, omdat niet alle kinderen noodzakelijkerwijs op alle items gokken. Om een gokparameter per item en per leeftijdsgroep te bepalen zijn de technieken van de klassieke testtheorie ongeschikt (zie bijvoorbeeld Lord, 1980), maar de parametrische item-respons modellen bieden deze mogelijkheden wel (Hulin, Drasgow & Parsons, 1983; Lord, 1980).

Alhoewel de psychometrische technologie al zover is gevorderd dat per item een gokparameter kan worden geschat, is het voor zover ons bekend niet mogelijk om met de beschikbare modellen Type II fouten voor het beoordeling-plus-uitleg criterium vast te stellen. Hiertoe dienen eerst

nieuwe modellen te worden ontwikkeld om de parameters per item voor deze fouten te kunnen schatten. Ondanks het mogelijk vóórkomen van deze fouten zijn echter de psychometrische eigenschappen (in het licht van de klassieke testtheorie) voor geen van de criteria nadelig beïnvloed.

Voor wat betreft het tegensuggestie criterium toonden we aan, dat dit criterium weer stringenter was dan het beoordeling-plus-uitleg criterium. Hoewel het tegensuggestie criterium een betrouwbare test voor classificatie produceerde, kan men zich op theoretische gronden toch afvragen of het een adequaat criterium is. Kohnstamm (1967) heeft aangegeven dat het weliswaar vanuit het gezichtspunt van de proefleider een goed teken is dat het kind bestand is tegen de tegensuggestie, maar dat het strikt genomen geen informatie geeft omtrent het classificatiebegrip van het kind. Verder is er in diverse recente studies aangevoerd, dat het veranderen van gedachten na een tegensuggestie misschien niets van doen heeft met classificatie-begrip, maar eerder met opvattingen omtrent sociale wenselijkheid, een volgzame houding, etc. (zie bijvoorbeeld Brainerd, 1973). Dit alles onderstreept dus, dat, in weerwil van de bevinding in deze studie dat alle drie criteria intern consistente tests voor meervoudig classificeren verschaffen, het nodig zal blijven een criteriumkeuze hecht-theoretisch te funderen.

REFERENTIES

- Almy, M. (1970). *Logical thinking in the second grade*. Colombia University: Teachers College Press.
- Bell, Y. R. (1975). *The effect of perceptual support on the performance of preoperational, transitional and operational children on multiple classification tasks*. Unpublished Ph.D. dissertation. The University of Nebraska.
- Bingham-Newman, A. M., & Hooper, F. H. (1974). Classification and seriation instructions and logical task performance in the preschool. *American Educational Research Journal*, 11, 379-390.
- Brainerd, C. J. (1973). Judgments and explanation as criteria for the presence of cognitive structures. *Psychological Bulletin*, 79, 172-179.
- Brainerd, C. J. (1978). *Piaget's theory of intelligence*. Englewood Cliffs, NJ: Prentice Hall.
- Camp, D. L. (1975). *An investigation of six major logical groupings of concrete operational thought*. Unpublished Ph.D. dissertation, The University of Iowa.
- Farnum, J. E (1976). *An investigation of analogous processes in concept attainment through manipulations designed to focus attention on either extension or intension in a multiplicative classification task*. Unpublished Ph.D. dissertation, The University of Rochester.

- Hamel, B. R. (1974). *Children from 5 to 7. Some aspects of the number concept*. Rotterdam, The Netherlands: Rotterdam University Press.
- Haven, S., & TenBerge, J. F. M. (1977). Tucker's coefficient of congruence as a measure of factorial invariance: An empirical study. *Heymans Bulletin*, 290 Ex, Department of Psychology, University of Groningen, The Netherlands.
- Hooper, F. H., Sipple, T. S., Goldman, J. A., & Swinton, S. S. (1979). A cross-sectional investigation of children's classificatory abilities. *Genetic Psychology Monographs*, 99, 41-89.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.
- Inhelder, B., & Sinclair, H. (1969). Learning cognitive structures. In P. Mussen, J. Langer, & M. Covington (Eds.), *Trends and issues in developmental psychology*. New York: Holt, Rinehart & Winston.
- Inhelder, B., Sinclair, H., & Bovet, M. (1974). *Apprentissage et structures de la connaissance*. Paris: Presses Universitaires de France.
- Kaiser, H. F. (1961). A note on Guttman's lower bound for the number of common factors. *Journal of Statistical Psychology*, 14, 1-2.
- Kingma, J. (1981). *De ontwikkeling van quantitative en relationale begrippen bij kinderen van 4-12 jaar*. (Dissertatie R.U. Groningen). Lisse: Swets & Zeitlinger.
- Kingma, J. (1982). A criterion problem: The use of different operationalizations in seriation research. *Perceptual and Motor Skills*, 55, 1303-1316.
- Kingma, J. (1983a). Piagetian tasks and traditional intelligence as predictors of performance on addition and subtraction tasks in primary school grades one and two. *Journal of Psychology*, 55, 39-53.
- Kingma, J. (1983b). The development of seriation, conservation and multiple classification: A longitudinal study. *Genetic Psychology Monographs*, 108, 43-67.
- Kingma, J. (1983c). Seriation, correspondence and transitivity. *Journal of Educational Psychology*, 75, 763-771.
- Kingma, J. (1983d). Some behavioral characteristics of the partial seriators reconsidered. *Journal of General Psychology*, 108, 231-247.
- Kingma, J. (1984). Criterion problems in conservation research reconsidered from a psychometric point of view. *Journal of General Psychology*, 111, 109-129.
- Kingma, J., & Koops, W. (1983). Piagetian tasks, traditional intelligence and achievement tests. *British Journal of Educational Psychology*, 53, 258-290.
- Kingma, J., & Koops, W. (1985). Betrouwbaarheid en validiteit van conservatie criteria. *Psychologica Belgica*, 25, 17-31.
- Kohnstamm, G. A. (1967). *Piaget's analysis of class inclusion: Right or wrong*. The Hague, The Netherlands: Mouton.
- Lagattuta, J. A. (1974). *An investigation of the developmental parallelism among four Piagetian operational groupings: Additive and multiplicative relations*. Unpublished Ph.D. dissertation, University of Maryland.
- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

- Lovell, K., Mitchell, B., & Everett, I. R. (1962). An experimental study of the growth of some logical structures. *British Journal of Psychology*, 53, 175-180.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw Hill.
- Odom, R. D. (1978). A perceptual-salience account of décalage relations and development change. In L. S. Siegel & C. J. Brainerd (Eds.), *Alternatives to Piaget*. New York: Academic Press.
- Overton, W. G., & Brodzinsky, D. (1972). Perceptual and logical factors in the development of multiplicative classification. *Developmental Psychology*, 6, 104-109.
- Pascual Leone, J., & Bovet, M. (1966). L'apprentissage de la quantification de l'inclusion et la théorie opératoire. *Acta Psychologica*, 25, 1-17.
- Piaget, J., & Inhelder, B. (1967). *La genèse des structures logiques élémentaires*. Neuchâtel, Switzerland: Delachaux et Niestlé.
- Reese, H. W., & Schack, M. L. (1974). Comment on Brainerd's criteria for cognitive structures. *Psychological Bulletin*, 81, 67-69.
- Sigel, G. E. (1971). The development of classificatory skills in young children: A training program. *Young children*, 170-184.
- Smedslund, J. (1962). The acquisition of conservation of substance and weight in children: VII conservation of discontinuous quantity and the operations of adding and taking away. *Scandinavian Journal of Psychology*, 3, 69-77.
- Smedslund, J. (1967a). Determinants of performance on double classification tasks. I. *Scandinavian Journal of Psychology*, 8, 88-96.
- Smedslund, J. (1967b). Determinants of performance on double classification tasks. II. *Scandinavian Journal of Psychology*, 8, 97-101.
- TenBerge, J. F. M. (1977a). Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, 42, 267-276.
- TenBerge, J. F. M. (1977b) *Optimizing factorial invariance*. Unpublished Ph.D. dissertation, State University of Groningen, The Netherlands.
- Tucker, L. R. (1951). *A method for synthesis of factor analytic studies*. Personal Research Section No. 904. Washington, D.C.: Department of the Army.

K. P. Van den Bos
Vakgroep Orthopedagogiek
Rijks Universiteit Groningen
Oude Boteringestraat 1
9712 GA Groningen
Nederland

Received September 1986

of

J. Kingma
Dept. of Psychology
University of Alberta
Edmonton, Alberta
Canada T6G 2E9