Drie redenen om te toetsen in het onderwijs

E. WARRIES

Stichting Research Instituut voor de Toegepaste Psychologie aan de Universiteit van Amsterdam

Inleiding

Binnen de staf van het R.I.T.P.[1] wordt sinds het begin van 1970 regelmatig in werkbijeenkomsten gesproken over het juiste gebruik van studietoetsen in het onderwijs. Tijdens deze gedachtenwisseling, waarbij ook het C.I.T.O.[2] frequent geraadpleegd is, werd door de deelnemers enkele malen de veronderstelling geuit dat slecht gebruik van toetsen veroorzaakt kan worden door betrekkelijke onbekendheid met de materie, waarbij men zich, om te beginnen, het doel van de toetsing niet helder bewust is. Met het oog op het doel waarvoor men ze gebruikt zullen studietoetsen sterk kunnen verschillen in manier van construeren, afnemen, scoren en analyseren. Zou men in het onderwijs, zonder zich rekenschap te geven van de redenen om te toetsen, grijpen naar beschikbare toetsen en toetsconstructie-methoden, ook al worden deze aangeboden door vakmensen, dan zou men daarmee fouten maken.

Om die mogelijke fouten te vermijden is het onderstaande geschreven[3]. Wij zijn in Nederland nog in het stadium waarin objectieve studietoetsen en schooltoetsen nog maar net bezig zijn de scholen en universiteiten binnen te komen. In deze omstandigheden leek het mij opportuun een bijdrage te leveren aan de doordenking van toetsdoelen door een zowel kritische als informatieve bijdrage. Nadrukkelijk wordt daarbij tevoren aangetekend, dat juist in deze beginsituatie in Nederland toets-situaties en de problemen die zich daarin voordoen nog niet definitief op een rij gezet kunnen worden. De persoonlijke meningen die hieronder volgen, moeten over twee jaar misschien wel weer anders geformuleerd worden. De 'vijf risico's' waarmee het artikel eindigt, worden dan ook nadrukkelijk aan de lezer als voorlopige schets van de gevarenzône aangeboden, door mij vanuit een eigen visie waargenomen, zij het genoteerd tegen de achtergrond van de gedachtenwisseling met de collega's.

Begripsbepaling. Professionele toetsen. Fasen.

De meeste docenten en studenten[4] weten uit eigen ervaring wat een studietoets is of kan zijn. De lezer die uitvoerige of formele definities van studietoetsen wil vinden, kan terecht in de engelstalige en tegenwoordig ook nederlandse handboeken voor docenten of toetsconstructeurs[5]. Voor goed begrip zij hoogstens gezegd dat ik onder een objectieve studietoets hier in elk geval versta:

een toets 'met een ingebouwde puntentelling', bestaande uit opgaven[6] waarin de probleemstelling of vraag vergezeld is van twee of meer antwoorden waarvan er slechts één goed is. Voor verder goed begrip moet nog gezegd worden dat het hier gaat over het toetsen van zaken die op school geleerd zijn. Er bestaan natuurlijk ook tests die gemaakt zijn voor het meten van de algemene ontwikkeling, studiezin, leervermogen, aanleg, intelligentie of capaciteiten, maar die zijn hier niet aan de orde. Dit artikel behandelt uitsluitend studietoetsen of schooltoetsen, die in de school worden gebruikt om prestaties, en niet bijv. het prestatievermogen, te meten. Op grond van het toegepaste fabricageproces is de aldus gedefinieerde studietoets nog te onderscheiden in drie soorten, namelijk de teacher-made toetsen en de professionele toetsen en als derde

soort wat ik noem de half-professionele toetsen.

Teacher-made toetsen worden door onderwijzers, leraren of hoogleraren samengesteld zonder voortdurende samenwerking met toetsconstructeurs van professie. In het algemeen zullen deze door docenten gemaakte toetsen niet voldoen aan de technische en soms wel perfectionistisch aandoende eisen die in de handboeken worden geformuleerd. Dat behoeft geen bezwaar te zijn zolang bij het nemen van beslissingen op grond van de uitslagen deze beperking maar beseft wordt. Er zijn trouwens in Nederland wel scholen in het v.o. waar men met groeiende deskundigheid een verzameling van steeds verbeterde items opbouwt. Voor de gewone vakdocent of voor een groep vakdocenten blijft daarbij natuurlijk altijd de beperking gelden dat men niet zoveel tijd in de constructie kan investeren als de professionele toetsconstructeurs doen.

Professionele toetsen worden vervaardigd door heterogeen samengestelde teams die veel meer hulpmiddelen en financiële armslag hebben dan de docent die ten dele in zijn vrije tijd met toetsconstructie bezig is. Zo beschikt het nog niet zo lang werkende C.I.T.O. te Arnhem over een snel groter wordende staf van psychometrici en over statistisch en organisatorisch geschoolde medewerkers die tezamen met vakdocenten toetsen construeren. Zowel de voorbereiding van de 'toetsprogramma's' als de constructie van in de winkel verkrijgbare 'kant-en-klaar toetsen'[8] vragen een enorme investering aan hooggekwalificeerde man-uren. Daar staat uiteraard tegenover dat de grote aantallen examinandi en het belang van de beslissingen waartoe de uitslagen soms leiden, een dergelijke investering ruimschoots rechtvaardigen. Dit is evident in het geval van toetsprogramma's zoals de Schooltoetsen Basisonderwijs, de Brugklas-eindtoetsen en de toetsen voor eindexamens. Het nagenoeg ontbreken van in de boekwinkel verkrijgbare coöperatieve toetsen die in de school op een zelf te bepalen tijdstip worden afgenomen, houdt waarschijnlijk verband met het ontbreken van een examennoodzaak op dit gebied.

Alle professionele toetsen behoren een handleiding te hebben waarin aan de toetsgebruiker verantwoording wordt afgelegd over de manier van samenstellen, over de steekproef van studenten waarop de toets is afgenomen, over de kwantitatieve eigenschappen van de items, enz.

De laatste jaren komen er boekjes met multiple-choice opgaven op de markt die als 'toets' worden aangeboden of als zodanig in elk geval door de koper worden beschouwd, hoewel ze niet zijn samengesteld volgens de vaktechnische regels die men daarvoor stelt. Deze boekjes, die uiteraard ook geen handleiding hebben waarin de vereiste verantwoording wordt afgelegd door gekwalificeerde personen, zijn uit het oogpunt van de testconstructeur eigenlijk maar 'half' professioneel, hoewel ze door de koper vaak als volwaardige toets worden opgevat. In het beste geval kunnen ze worden benut als 'itemverzamelingen' die door de docent geraadpleegd worden om zelf proefwerken uit samen te stellen.

Zoals hierboven al aangeduid, zal het gebruiksdoel van de toets richtinggevend kunnen zijn bij de te volgen werkwijze in de vier fasen: constructie, afname, scoring en analyse van de toets. Dit kan men aldus verklaren.

Tijdens het *construeren* kan worden bepaald en geëffectueerd hoe moeilijk de items zullen zijn, of kennis dan wel leervermogen getoetst moet worden, of de toets moet differentiëren tussen knappe en minder knappe leerlingen en bij welke van de in gebruik zijnde lesmethoden men zal aansluiten. Het *afnemen* van studietoetsen is in de meeste gevallen gebonden aan een standaardinstructie die voor studenten en docenten richtlijnen bevat over de gang van zaken vóór en tijdens de toets. Deze regels kunnen met het gebruiksdoel verschillen in de manier waarop men de hoeveelheid werktijd aangeeft, het probleem van raden bij niet weten behandelt, de geheimhouding van opgaven behandelt, enz. Zo kunnen ook de regels voor de

scoring van de geleverde prestaties verschillen per toets. Ervaren testgebruikers beginnen bijna altijd met een ruwe score die verkregen wordt door het aantal goede antwoorden te tellen. De belangrijke omzetting echter van deze ruwe score in een (school-)cijfer, een standaardscore of percentielscore houdt verband met de reden waarvoor getoetst wordt; ze kan verschillen per toets[10]. De *analyse* van de antwoorden en de scores der studenten achteraf verschaft statistische gegevens waardoor de toets, met het oog op het gebruiksdoel, kan worden herzien. Ook kunnen opgaven uit meerdere toetsen worden samengevoegd tot een nieuwe toets met speciale eigenschappen.

De drie doeleinden waarvoor getoetst wordt in het onderwijs

Proefwerken, examens, objectieve en niet objectieve prestatiebeoordelingen hebben als gemeenschappelijk kenmerk dat de student een studieprestatie leveren moet en dat zijn prestatie een beoordeling, cijfer of score, oplevert. Daarmee houdt dikwijls de overeenstemming op en gaan de verschillen in toepassing overheersen. In het onderwijs worden toetsen gebruikt om een groot aantal redenen. Op de 'variabelen van de examensituatie' is gewezen door De Groot[11] die daarbij ondermeer onderscheid maakt tussen toelatingsexamen, eerste examen, tussen-examens en afsluitend examen. Dezelfde schrijver noemt elders[12] vier mogelijkheden voor het gebruik van studietoetsen, te weten voor (I) predictie, (II) selectie, (III) evaluatie en (IV) operationalisering van doelstellingen. In het onderstaande wordt niet ingegaan op de funktie van studietoetsen in de voorspelling van schoolsucces, omdat deze toepassing nog buiten het eigenlijke onderwijs valt in ons land en misschien meer op het terrein van de beroepskeuze-adviseur en de psycholoog ligt.

Studietoetsen ter concretisering van onderwijsdoelstellingen zijn uitermate belangrijk, maar komen in dit artikel ook niet aan de orde, omdat deze toepassing op dit moment nog behoort tot de eerder aangeduide zaken, die misschien over twee jaar actueel zullen zijn geworden en zozeer gemeengoed zijn dat ze prioriteit gaan genieten op het problemenlijstje. In dit artikel heb ik het alleen over het selectief gebruik van studietoetsen, en over evaluatief gebruik. Onder evaluatie zou ik daarbij liever allereerst denken aan het evalueren van de studieresultaten van groepen (schoolklas, school, wijk, regio) en – in afwijking van De Groot – de 'individuele evaluatie' bij voorkeur opvatten en betitelen als 'diagnose' of 'didactische diagnose'. Met de term 'diagnostisch toetsen' wordt aangesloten bij het zich ontwikkelende spraakgebruik in ons land[12a]. Anderzijds wordt het Engelse 'evaluation' ook minder vaak voor individuele metingen gebruikt. Doorgaans doelt men op iets als de empirische vaststelling van de mate waarin bepaalde doeleinden in feite gerealiseerd zijn – in een groep.

Een scherp onderscheid tussen het toetsen voor *selectie*, voor *evaluatie* en voor (didactische) *diagnose* lijkt de eerste jaren voldoende om verkeerde toepassingen door definitorische misverstanden te voorkomen. Ook buiten Nederland worden deze drie trouwens al heel lang beschouwd als de voornaamste funkties van 'educational measurement'[13].

Het toetsen voor selectie

Selectie is te beschouwen als een institutionele activiteit, die de toelating regelt tot onderwijs[14], waarbij naar mijn mening het belang van de onderwijsinstelling eerder de aandacht krijgt, dan dat van de individuele kandidaat die toegelaten wil worden. Het instituut heeft eén goede naam op te houden, wil het peil van het onderwijs handhaven of probeert kandidaten te weren die stagnerend zullen werken. Selectie, als technisch probleem onder meer beschreven door Wiegersma[15], heeft in het verleden vooral psychologen intensief bezig gehouden. De kunst van het goed selecteren is in hoge mate afhankelijk van het goed prediceren van later succes in opleiding of baan. Studietoetsen kunnen daarbij een belangrijk aandeel hebben, omdat het resultaat van voorafgaand onderwijs nu eenmaal een zekere voorspelling inhoudt over succes in

volgend onderwijs.

De Groot spreekt in 'Vijven en Zessen' niet alleen over deze correlatie met later succes maar ook en vooral over de functie van studietoetsen als men wil komen tot niet-rekbare maatstaven in de beoordeling van schoolprestaties. Hij spreekt ook van een 'selectiedrempel', die min of meer constant te houden is over de jaren. Een absolute drempel aan kennis is overigens naar mijn mening niet zo makkelijk vast te stellen en te handhaven als in het onderwijs wel eens verondersteld wordt. Institutionele selectie zal zich dan ook eerder bedienen van een correlatie-model, waarbij door middel van relatief meten[16] de beste kandidaten worden toegelaten, dan dat men een eenmaal gestelde drempel handhaaft. Overigens bestaat er onder onderzoekers een tendentie om de research, die zich richt op steeds betere selectie, minder interessant te achten, omdat perfecte voorspellingen niet bereikbaar zijn in de werkelijkheid van het onderwijs. Onderwijskundige psychologen richten zich de laatste tijd dan ook liever op een betere begeleiding zodat eenmaal toegelaten kandidaten zonder uitval de eindstreep kunnen halen, dan dat zij zich langer bezighouden met de perfectionering van een selectie-apparaat dat weliswaar redelijk goed is vergeleken met elk ander hulpmiddel maar waarvan de waarde altijd weer door de gebruikers overschat wordt. Misschien dat met deze toewending naar begeleiding de realisering van het door velen na De Groot bepleite 'onderwijs met selectievrije perioden' snel dichterbij komt. Dat neemt niet weg dat studietoetsen zich, ook al door hun meeteigenschappen, goed lenen voor selectief gebruik.

Toetsingen voor selectief gebruik zijn, naar ik meen, door drie eigenaardigheden in meerdere of mindere mate gekenmerkt:

1. Bij de constructie zal de toetsmaker de neiging hebben opgaven te kiezen die niet rechtstreeks aansluiten bij het voorafgaande opleidingsprogramma. Hij kan dit doen omdat, behalve kennis, ook eigenschappen als intelligent handelen, niet direct op school aangeleerd inzicht, en algemene ontwikkeling als voorspellers van later schoolsucces zijn te beschouwen.
2. Bij de constructie sluit de toetsmaker zich aan bij de examinatorentraditie dat vragen die gemakkelijk zijn (empirisch: door meer dan 75% goed worden beantwoord) beter niet gesteld kunnen worden. Hij kiest tamelijk moeilijke vragen uit.
   Ook zal bij analyse van de antwoorden van de leerlingen bij selecterende toetsen gestreefd worden naar een zo groot mogelijke statistische variantie van de scores, omdat een selecterende toets in de groep kandidaten een scheiding dient aan te brengen tussen de bokken en de schapen: zij 'die het niet kunnen' tegenover zij 'die het wel kunnen'. Theoretisch is aannemelijk gemaakt[17] dat als alle items van een toets stuk voor stuk gemiddeld door circa de helft van de kandidaten fout worden beantwoord, het scorebereik van de toets het grootst is.
3. De omzetting van de ruwe scores van de leerlingen in standaardscores vindt plaats nadat de prestaties van leerlingen gerangschikt zijn. De uiteindelijke standaardscore of percentielscore van een leerling hangt dus af van de relatieve positie in de groep van medekandidaten. Bij selectieve toetsen is deze werkwijze adequaat, omdat het aanbrengen van de caesuur tussen zakken en slagen aldus vereenvoudigd wordt.

Samenvattend kan men dus zeggen dat constructie, scoring en analyse van een 'toets bestemd voor selectie' zal leiden tot een toets die (a) tamelijk moeilijk is (*b*) niet specifiek hoeft te zijn voor het genoten onderwijs en (*c*) gemakkelijk een vergelijking van de prestaties der kandidaten toelaat.

Het evaluerend gebruik van studietoetsen

Evaluatie, als empirische bepaling van de mate waarin beoogde resultaten bereikt zijn, is in ons land nog tamelijk onbekend. Er is een publikatie van Souren e.a.[18] over de stand van het

lager onderwijs in Noord-Brabant en er is het verslag van Wiegersma en Groen[19] over de nederlandse deelname aan een internationaal vergelijkend onderzoek over resultaten van wiskunde-onderwijs, maar verder is mij uit de nederlandse literatuur weinig bekend over het op grote schaal evaluerend of evaluatief toepassen van studietoetsen in het onderwijs. Zoals door mij elders betoogd[20] is het te verwachten, dat binnen enkele jaren de gemeenschap, gemeentes, schoolverenigingen, het Rijk, een beroep zullen doen op toetsconstructeurs om mee te werken aan grote evaluatieprogramma's. Programma's als 'National Assessment' in de Verenigde Staten leveren grote hoeveelheden vergelijkende informatie aan steden en staten[21].

Het evaluerend gebruik van toetsen behoeft zich trouwens niet alleen nationaal of regionaal af te spelen; men kan ook per klas het gegeven onderwijs evalueren. Niet alleen in research-toepassingen[22] maar ook in de dagelijkse praktijk van het onderwijs is het mogelijk het gegeven onderricht te evalueren en vervolgens bij te sturen op die punten waar kennelijk nog misverstanden of kennis-tekorten heersen in de klas. Onder anderen Van Calcar en De Bruyne[23] wijzen op de mogelijkheden van een snelle fouten-analyse voor de docent door middel van objectieve toetsen.

Het komt mij voor, dat het gebruik van toetsen voor het evalueren van de studieprestaties van groepen, in de huidige ontwikkeling gekenmerkt behoort te zijn door het volgende:

1. Bij de constructie van evaluatieve toetsen zal bijzondere aandacht moeten worden gegeven aan de representativiteit van de vragen voor de bestudeerde stof. De toetsconstructeurs zullen zich terdege moeten verdiepen in het leerplan of de lesmethode van de scholen waarvoor de betreffende toets is bestemd. Niet alleen omdat op deze wijze een eerlijke vergelijking mogelijk is tussen groepen leerlingen, maar ook omdat anders de kans bestaat dat andere variabelen (zoals bijvoorbeeld de algemene ontwikkeling) een te grote rol gaan spelen in vergelijking met het eigenlijke leereffect.
2. Bij de constructie zal erop gelet dienen te worden dat bij bepaalde onderdelen van de stof de items een zekere differentiatie in moeilijkheid vertonen. Dit is noodzakelijk omdat anders geen juist beeld wordt verkregen in die gevallen waar een item in een bepaalde groep van scholen als te moeilijk beschouwd moet worden.
3. Bij de constructie moet men erop bedacht zijn dat de gebruikte terminologie in de opgaven, en de toelichtingen daarop, aansluit bij het gegeven onderwijs. Dit speelt bijvoorbeeld in de wiskunde waar dezelfde sleutelbegrippen door verschillende termen worden aangeduid.
4. De uiteindelijke scoring van de resultaten zal een gedifferentieerd beeld per groep moeten opleveren. Hiermee is bedoeld dat voor verschillende deelaspecten, mogelijk vertegenwoordigd door 10 of minder items, een score wordt toegekend zodat voor elke geëvalueerde groep als het ware een profiel van prestaties ontstaat.
5. De uiteindelijke scoring zal zich moeten richten op een afwegen van de groepsprestatie tegen zowel locale normen als grotere, misschien landelijke normgroepen.
6. De afname van de gehele toets kan gespreid worden over verschillende groepen leerlingen. Het is niet noodzakelijk dat elke leerling een gehele toets of zelfs alle items van een deeltoets maakt. Om praktische redenen zal het waarschijnlijk beter zijn dat individuele leerlingen een selectie van items uit de gehele toets maken. Om praktische redenen zal het evenzeer soms noodzakelijk zijn, hele klassen te laten meedoen, hoewel niet alle gevallen uit die klas mee hoeven te tellen in de uiteindelijke scoring.

Samenvattend kan men dus zeggen dat de constructie en scoring van een toets bestemd voor evaluatie zal leiden tot een toets die (*a*) over het geheel niet al te moeilijk is, (*b*) representatief is

voor het gegeven onderwijs, (*c*) gemakkelijke vergelijkingen toestaat zowel tussen de prestaties in de geëvalueerde groep als tussen de prestaties van een bepaalde groep en andere groepen, (*d*) gespreid kan worden in afname over steekproeven van leerlingen, (*e*) bij klassikale afname directe feed-back levert aan de docent. (In dat geval zal dikwijls sprake zijn van zelf gemaakte, teacher-made, toetsen.)

Het diagnostisch of didactisch gebruik van studietoetsen

Het spraakgebruik dat bezig is te ontstaan in onderwijskringen wijkt hier enigszins af van de medische betekenis van het woord diagnose dat slaat op een gedifferentieerd onderzoek naar alle mogelijke afwijkingen. Ook psychologen zullen onder diagnostisch onderzoek van een leerling vaak iets anders verstaan dan docenten die met toetsen hebben leren werken. Het onderzoek van de psycholoog zal zich, met name bij het opsporen van oorzaken van leermoeilijkheden, richten op milieuomstandigheden, persoonlijkheidskenmerken, en het meten van niet op school leerbare capaciteiten die het intelligent handelen bepalen. Hoewel het mogelijk is dat de psycholoog daarbij van bestaande objectieve studietoetsen gebruik zal maken, wordt dit type diagnostisch gebruik hier niet behandeld. Het gaat hier over het gebruik van studietoetsen in de school bij de gewone gang van zaken als er nog geen sprake is van leermoeilijkheden in de min of meer medische betekenis van het woord.

Het diagnostisch gebruik van studietoetsen zoals hier bedoeld is gekenmerkt door het zoeken naar (nog niet ontdekte) tekorten. Daar komt dan onverbrekelijk bij een individuele follow-up: op grond van de leerlingscores worden maatregelen genomen om het niet-gekende in te halen, te repeteren of aan te vullen.

Er is zowel sprake van professioneel als van niet professioneel diagnostisch gebruik in Nederland. In een ontwikkelingsprojekt dat door het Algemeen Pedagogisch Studiecentrum gecoördineerd wordt en dat wordt uitgevoerd aan de Rijks Scholengemeenschap te Schagen wordt gebruik gemaakt van door de leraren zelf geconstrueerde diagnostische toetsen. In het eerste Interimverslag[24] wordt melding gemaakt van het diagnostisch gebruik van zulke toetsen die een score opleverden op grond waarvan de leerlingen zelf herhalingsstof of keuzestof kozen. Deze toetsen werden gebruikt om gedifferentieerd onderwijs binnen heterogene brugklassen mogelijk te maken. Ook in een reisverslag van J. Timmer[25] wordt gesproken over een dergelijk gebruik van studietoetsen in een Zweeds wiskundeprojekt. Het systematisch gebruik van professionele didactische studietoetsen in het basisonderwijs wordt in ons land thans vooral gepropageerd door een groep wetenschappelijke onderzoekers en in de praktijk werkzame psychologen, die zich onder meer verenigd hebben in een werkgroep met de naam 'Werkgroep van Vier tot Zeven'. De werkzaamheden van onder meer deze groep, die zijn kontakten heeft zowel in Enschede als in Haarlem, Utrecht en elders, hebben onlangs geleid tot de publikatie van een toetshandleiding van Van Calcar en De Bruyne[23]. De door Van Calcar beschreven toetsen dienen onder meer om bepaalde fundamentele vaardigheden in het lezen van de kinderen in de eerste klas van het basisonderwijs te meten. De uitslagen van deze professionele toetsen geven een indicatie van bepaalde leesmoeilijkheden op grond waarvan de onderwijzer of onderwijzeres vervolgens correctiemaatregelen kan nemen. Nadrukkelijk wordt bij dit soort toetsen dan ook verondersteld, dat men de beschikking heeft over oefenmateriaal of 'verrijkingsmateriaal'[25a] waardoor de hiaten in de kennis van de kinderen opgevangen kunnen worden. In het voortgezet onderwijs is het mogelijk dat de leerlingen zelf toegang hebben tot dit materiaal, zodat zij zelf herhalingsstof of oefenmateriaal kunnen uitkiezen, nadat zij hun eigen score hebben bepaald en geconstateerd hebben dat ze de zaak nog niet voldoende beheersen.

Dit is een kenmerkend verschil van diagnostische toetsen met elk ander meetmiddel: Bij diagnostische toetsen gaat het niet om een registratie van de scores maar gaat het erom dat de

leerling uiteindelijk tot *beheersing* van de stof komt. In vele gevallen zal een registratie van de scores in het geheel niet plaatsvinden.

De diagnostische toetsen, die tot nu toe bij ons bekend zijn, zijn bijna altijd gekenmerkt door het volgende:

1. De constructie van diagnostische toetsen is gericht op een zo goed mogelijke representatie van alle onderwijskundig relevante aspecten. Die relevante aspecten kunnen zowel te maken hebben met bepaalde veel voorkomende leermoeilijkheden, met frequent voorkomende vergissingen als met de logische opbouw van de stof. Vooral in het geval van de jonge kinderen in het basisonderwijs zal dikwijls een staalkaart van veel voorkomende fouten terug te vinden zijn in de betreffende diagnostische toets.
2. In het begin van het lager onderwijs zal de reeks van gemeten aspecten dikwijls van fundamenteler karakter zijn en misschien ook 'professioneler' constructeurs behoeven dan in het voortgezet onderwijs. In het voortgezet onderwijs zal de toets, dikwijls bestaande uit items behorende bij specifieke onderdelen van de stof, met enige moeite door de vaksectie van het betreffende vak zelf te maken zijn.
3. Diagnostische toetsen zonder aanbevelingen voor herhalingsstof of oefenmateriaal hebben géén zin.
4. Zeker bij toetsen in het begin van het basisonderwijs zullen eventuele standaardscores uitgedrukt moeten worden in de normen van sub-groepen die gedefinieerd zijn aan de hand van geografisch gebied, beroepenniveau van de school of beroepenniveau van de ouders van de leerling.
5. Bij de constructie van didactische of diagnostische toetsen kan de samenwerking met de didacticus van zeer groot belang zijn.

Samenvattend zou ik willen zeggen dat de constructie van een toets bestemd voor didactische diagnose behoort te leiden tot een toets die (*a*) niet moeilijk maar in principe voor iedereen haalbaar is, (*b*) bijzonder nauw aansluit bij het gegeven onderwijs, (*c*) eventuele vergelijking toestaat met goed gedefinieerde normgroepen. Zo'n toets leidt niet tot klassifikatie van leerlingen maar tot correctiemaatregelen in individuele gevallen.

Het oneigenlijk gebruik van studietoetsen: vijf risico's

In het voorgaande is, ten dele beschrijvend, ten dele betogend, aangegeven, hoe de drie redenen om studieprestaties te toetsen in het onderwijs, kunnen worden gekenmerkt. In het kort komt het er misschien op neer, dat selecterend gebruik van toetsen vooral vergelijking beoogt, dat didactisch of diagnostisch gebruik gekenmerkt is door het werken met absolute maatstaven van beheersing waaraan voldaan moet worden, en dat voor evaluatieve toetsen vooral de representativiteit van de opgaven van belang is. Voor alle school- en studietoetsen geldt trouwens, dat bij de samenstelling en het gebruik ervan nauwgezet afgewogen moet zijn wat de bedoeling van de toetsing is en daaraan voorafgaand uiteraard wat de doelstellingen van het onderwijs waren.

Ter afsluiting van dit artikel geef ik hieronder vijf hypothetische risico's die de docent loopt bij de drie soorten van toetsingen zoals ik die hierboven heb beschreven. Aan de lezer wordt overgelaten te beoordelen in hoeverre deze risico's in feite bestaan, dan wel in de toekomst misschien zullen gaan ontstaan.

1. *Men toetst niet meer wat geleerd is, maar men toetst de aanleg*

Welke van de drie redenen om te toetsen er ook is, bij elke toetsing in de school gaat het om wat geleerd is en niet om het leervermogen. De toetsconstructeur blijft echter altijd het risico lopen dat hij zich zal richten op het meten van vaardigheden die niet op school worden geleerd. In het gunstigste geval kan daarbij de nadruk komen te liggen op wat de E.T.S. lange jaren als

'developed ability' heeft aangeprezen, ten koste van de inhoudelijke aansluiting van de toets bij het leerprogramma. In het ernstigste geval zal men helemaal niet meer kijken naar wat geleerd is, maar zich geheel richten op de capaciteit om te leren.

2. *Men construeert selecterende toetsen terwijl men meent te evalueren of te diagnosticeren:*

De constructie en scoringsmethoden van selecterende toetsen zijn gemakkelijk over te nemen in oorspronkelijk niet-selecterende metingen. Elders[16] heb ik er op gewezen dat deze vorm van m.i. onjuist gebruik van studietoetsen steunt op de psychologische testtheorie. Het 'selectief maken' van didactische of van evaluerende toetsen kan zeer geleidelijk gaan.

In het eerste geval b.v. via een beginsel van aanvankelijk vrijstellende examens die later als te makkelijk worden beschouwd. In het tweede geval door het verwerpen van evaluerende toetsen die men na verloop van tijd als onbevredigend ervaart omdat de groepsprestaties 'te hoog' worden.

3. *Men diagnosticeert niet meer, maar groepeert*

Didactische toetsen, die optimaal functioneren als ze gebruikt worden om alle studenten – in wat voor heterogene groep ook – tot een zeker niveau van beheersing te brengen, kunnen worden gebruikt in een vroegtijdig toegepast systeem van 'setting'. Diagnostische toetsen immers maken het mogelijk leerlingen min of meer te klasseren in plaats van ze te verwijzen naar herhalingsstof of oefenmateriaal. Een dergelijke klassifikatie kan niet laat genoeg plaatsvinden. Het risico bestaat dat groeperen te vroeg gebeurt en de verdere loopbaan van de leerling nadelig beïnvloedt.

4. *Men evalueert niet meer voor feed-back, maar voor beoordeling*

Het is heel eenvoudig, op grond van de uitslag van evaluerende toetsen, een docent, een klas, een groepje leerlingen, te beoordelen en eventueel te veroordelen – en het daarbij te laten. Tekorten die geconstateerd worden met evaluerende toetsen kunnen worden opgevat als tekortkomingen van de leerlingen en niet als een te corrigeren hiaat in de kennis van de groep. Men kan dus die correctie achterwege laten. Misschien omdat men geen tijd heeft. Misschien zelfs omdat de docent of de vaksectie de uitslagen van de evaluerende toets begripsmatig alleen maar vermag te zien als normbepalend: kandidaten voor m.a.v.o.-III, m.a.v.o.-IV, h.a.v.o. of gymnasium.

5. *Men selecteert teveel – omdat het zo gemakkelijk gaat*

Het grote risico van selecterende toetsen is naar de mening van de schrijver van het bovenstaande eenvoudigweg, dat ze te dikwijls gebruikt zullen worden. De constructie van goede of althans goedselecterende toetsen is al gauw bevredigend: als men de toetsen maar moeilijk genoeg maakt, is er altijd een caesuur tussen 'slechte' en 'goede' studenten aan te brengen. Dat een dergelijke gedachtengang door de schrijver soms expliciet en soms stilzwijgend verworpen wordt, is uit het bovenstaande genoegzaam gebleken, menen wij.

Noten

1. Aan dit beraad in het Research Instituut voor de Toegepaste Psychologie aan de Universiteit van Amsterdam (directeur: prof. dr. A. D. de Groot) hadden de volgende medewerkers deel: drs D. J. Bos, dr C. van Calcar, drs Sj. Sandbergen, drs H. Stroomberg, J. Timmer, dr E. Warries (directeur R.I.T.P.), drs H. Wesdorp. Het beraad vond plaats in een werkgroep met de titel 'Kritisch Gebruik Studietoetsen'.
2. Aan het gesprek over goede voorlichting in de werkgroep Kritisch Gebruik Studietoetsen werd in 1970 deelgenomen door mevrouw drs M. Peeters-Sips van de psychometrische afdeling van het Centraal Instituut voor Toetsontwikkeling. Met drs J. Solberg, directeur van het C.I.T.-O. had ik een voor mij verhelderend gesprek over een eerdere versie van dit artikel hetgeen tot

wijziging van de tekst leidde. Over de constructie en het gebruik van toetsen is ook verder in het personele vlak veelvuldig contact tussen C.I.T.O. en R.I.T.P.

3. Ik heb mij daarbij georiënteerd op het Nederlandse onderwijs en op de huidige toestand. Zo zijn Amerikaanse begrippen als guidance en counseling, met bijbehorende tests, in ons land nauwelijks bekend. Toetsen die als eerste funktie hebben, leerlingen te adviseren welke stroom of welk leerpakket ze zullen kiezen, zijn voorlopig in ons land nog geen gemeengoed.
4. In dit artikel wordt de term 'docenten' gebruikt om leraren, onderwijzers, hoogleraren aan te duiden. Zowel 'studenten' als 'leerlingen' slaat op al diegenen die studeren, bezig zijn gestructureerde kennis op te doen, dus ook de leerling in het basisonderwijs.
5. Het beste voor docenten geschreven handboek is *Measuring Educational Achievement* van professor Robert L. Ebel, in 1965 uitgekomen. Over toetsconstructie in Nederland is onder meer geschreven door de experts dr R. R. Gras *Studietoetsen voor moderne talen*, dissertatie, (Wolters, 1967), drs Sj. Sandbergen 'Studietoetsen: Noodzaak, nut en vereisten' in *Vernieuwing van Opvoeding en Onderwijs*, 27, 263, 14–28, aug. 1968, drs K. D. Thio 'Studietoetsen in Nederland: projecten en activiteiten', *De Psycholoog*, IV, 7, 379–393 (1969) en dr E. Warries 'Proefwerken met Vierkeuze-vragen' in *Ped. Stud.* 46, 1, 22–41 en 46, 3, 161–165, 1969. Voorjaar 1970 is het eerste nederlandse handboek uitgekomen: *Studietoetsen*, bij Mouton, geschreven door professor dr A. D. de Groot en dr R. F. van Naerssen.
6. Toetsconstructeurs van professie spreken bij voorkeur van 'items' op z'n nederlands uitgesproken.
8. In 'De Psycholoog', IV, 7, 379–393, 1969, geeft drs K. D. Thio van het C.I.T.O. de volgende beschrijving van deze twee wijzen van beschikbaarstelling:
   – kant-en-klare toetsen: toetsen, die in de gedaante van afgeronde, voor gebruik gerede opgavenboekjes, onder toevoeging van handleiding, normschalen, eventueel met antwoordbladen en met de mogelijkheid van scoringsservice door een of ander bureau, op de leermiddelenmarkt worden gebracht. Dit is dus de 'klassieke' manier van tests uitgeven.
   – 'testing programs' (misschien het best te vertalen met: 'doorlopende toetsingsprogramma's'): een continuproject van produktie en afname van toetsen, dat meestal in opdracht van beleidsinstanties door professionele instituten wordt uitgevoerd op grote schaal.
10. De Groot en Van Naerssen gaven het boek 'Studietoetsen' als ondertitel 'Construeren, afnemen, analyseren'. Vanwege het grote belang dat ik ben gaan hechten aan de wijze van *scoren*, heb ik in dit artikel consequent de term 'scoren' tussen de klassieke trits gevoegd.
11. Prof dr A. D. de Groot, *Vijven en Zessen*, blz. 68–73, Wolters, 1966.
12. In het eerder genoemde 'Studietoetsen', hoofdstuk 3.

12a. Zie b.v. E. F. Vroom 'De subtoetsen voor stillezen van de Amsterdamse Schooltoets 1969' *Ped. Stud.* 1970, 47, 205–216.

13. Henry Dyer, aangehaald door Winton H. Manning, 'The functions and uses of educational measurement', p. 12–22, *Proceedings of the 1969 Invitational Conference on Testing Problems*, Educational Testing Service 1969.
14. Ook toetsen bij eindexamens en overgangs-examens zou ik als selectieve toepassingen willen beschouwen. Men zou kunnen aanvoeren dat de eindexamens van onze scholen voor voortgezet onderwijs niet passen in dit schema. Daartegen valt dit in te brengen: Eindexamens geven in ons onderwijsbestel zeer vaak recht op toelating tot ander onderwijs en zijn uit dien hoofde te beschouwen als toelatingsselecties.
    Zelfs voor die beroepsopleidingen waarvan het eindexamen als definitieve afsluiting beschouwd kan worden, geldt toch dat het diploma betekent dat men wordt toegelaten tot een groep van gekwalificeerde beoefenaren: een selectiesituatie waarbij deze laatste groep dikwijls, gezien of ongezien, zijn invloed uitoefent.
15. Prof. dr S. Wiegersma: *Selectieproblemen.* De beoordeling van geschiktheid voor functie, studie en beroep. Swetz & Zeitlinger, 1964.
16. Zie dr E. Warries 'Het relatief meten van leerprestaties in het onderwijs', *Nederlands Tijdschrift voor Psychologie.* Speciaal nummer over Didakometrisch en Psychometrisch Onderzoek, XXV, 6, 429–439, juni 1970.
17. Door Harold Gulliksen, geciteerd door Robert L. Ebel, p 75, *Measuring Educational Achievement*, Prentice Hall 1965.
18. Souren C. J. M. H. e.a. *Rapport over een onderzoek naar de stand van het gewoon lager onderwijs*

in Noord-Brabant. Prov. Bestuur van Noord-Brabant 1957.

19. S. Wiegersma en M. Groen, *Resultaten van wiskunde-onderwijs*, Wolters-Noordhoff, 1968.
20. Verslag van een werkbezoek aan Educational Testing Service, rapport R.I.T.P., 1970.
21. Zie ondermeer hierover: Appendix A in *Evaluation as Feedback and Guide*, Fred. T. Wilhelms, ed., Washington 1967.
22. Zoals bijvoorbeeld beschreven door A. D. de Groot en medewerkers in *Bewegingsmeetkunde*, Empirische Studies over Onderwijs No. 11, Wolters-Noordhoff, 1968.
23. *Algemene Handleiding voor Toetsen*, voorlopige versie, Ped. Centrum Enschede en R.I.T.P., 1970.
24. Zie '*Interimverslag Projekt Schagen*', Onderwijskundig Studiecentrum, Amsterdam, januari 1970.
25. J. Timmer: Het I.M.U.-projekt, een systeem voor het geven van wiskundeonderwijs op de Zweedse lagere school. Verslag van een 4-daagse studiereis naar Zweden. *R.I.T.P. rapport* 1969.

25a. Zie b.v. 'Oefenmateriaal voor het aanleren van een aantal begrippen, bestemd voor leerlingen uit de eerste klassen van het basis-onderwijs.' T. W. de Wilde en mej. H. M. Plegt, Pedag. Centrum Enschede, zonder jaartal. Bij de Stichting Schooladviesdienst te Utrecht is in gebruik een 'Schema voor hulpprogramma aan leerlingen die vertragingen vertonen bij het aanvankelijk lezen' aansluitend op de methode 'Zo leren lezen' van F. B. Caesar, aansluitend op de klassikale toets.