Hoe kun je het beste meten of een leerling een tekst begrijpt?

Een vergelijkend onderzoek naar vier methoden

Judith Kamalski, Ted Sanders, Leo Lentz en Huub van den Bergh[1]

In het voortgezet onderwijs is het gebruikelijk om tekstbegrip in de moedertaal te toeten aan de hand van meerkeuzevragen en open vragen. Er bestaan echter ook andere methoden om tekstbegrip te meten, bijvoorbeeld de gatentekst en de sorteertaak (het plaatsen van concepten in groepjes, op basis van de verbanden die in de tekst gelegd zijn). In dit artikel doen de auteurs verslag van een onderzoek naar de kwaliteit van vier begripstoetsen, toegespitst op de kenmerken 'betrouwbaarheid' en 'validiteit'. De resultaten suggereren dat de sorteertaak een betere methode is om tekstbegrip te meten dan de welbekende meerkeuzevragen en open vragen.

Het is niet eenvoudig een goede toets te maken om tekstbegrip in de moedertaal te meten. Dat merken onderzoekers, die de invloed van bepaalde tekstkenmerken op begrip willen meten, maar dat merken leraren en examenconstructeurs ook.

Tekstbegrip meten betekent dat de tekstrepresentatie die de lezer geconstrueerd heeft, geoperationaliseerd moet worden. Volgens de nieuwste inzichten in de cognitieve linguistiek en onderzoek naar discourse processing (Kintsch, 1998; Singer, 1990) kunnen drie niveaus van mentale representatie onderscheiden worden. Het *oppervlakteniveau* betreft grammatica en woordbetekenis. Het tweede, iets dieperliggend niveau, is de *betekenisrepresentatie*: het toekennen van betekenis aan afzonderlijke zinnen. Het laatste niveau, het *situatiemodel*, betreft de diepste en meest complexe vorm van tekstbegrip: nieuwe informatie uit de tekst moet geïntegreerd worden met de voorkennis die de leerling al heeft over het onderwerp. Informatie uit diverse zinnen wordt geïntegreerd én verbonden met de bestaande kennisstructuur in het hoofd van de leerling (zie ook Land en Sanders, te verschijnen).

Het meten van dit situatiemodel is belangrijk als we willen vaststellen hoe goed een leerling een tekst nu begrepen heeft. Land, Sanders, Lentz en Van den Bergh (2002) pasten voor het eerst taken om situatiemodellen te meten toe op een educatieve context. Uit dit en uit ander eerder onderzoek blijkt echter dat het meten van deze representaties een ingewikkelde taak is. Iedere tekst is anders en dus is ieder situatiemodel anders. Bovendien heeft iedere leerling andere bestaande kennisstructuren en die beïnvloeden het situatiemodel ook. Er bestaan daarom veel verschillende manieren en methoden om begrip op

het diepste niveau te meten en die worden intuïtief toegepast, zonder precies te weten *wat* we nu eigenlijk aan het meten zijn.

In dit artikel doen we verslag van een onderzoek dat bedoeld is om enkele tekstbegripmethoden te vergelijken op betrouwbaarheid en validiteit. De centrale vraag is dus: "hoe kan tekstbegrip in de moedertaal het beste gemeten worden?"

De onderzochte tekstbegripmethoden

Vier verschillende methoden zijn onderzocht. We onderzoeken of deze methoden de representatie van het situatiemodel van de tekst meten: meten ze de relaties die leerlingen door het lezen van de tekst hebben gelegd? Twee methoden zijn traditioneel en worden veel gebruikt in een educatieve context: meerkeuze/open vragen en een gatentekst. Twee methoden zijn nieuw en recent in Amerika ontwikkeld: de sorteertaak en de mental model taak. Hieronder zullen we de methodes kort bespreken en voorbeelden van vragen geven. De voorbeelden hebben steeds betrekking op de volgende passage uit het Eindexamen Nederlands 2003, tijdvak 2.[2]

> Door betere scholing en tal van informatiebronnen hebben mensen tegenwoordig meer kennis van verschillende professies, wat wel protoprofessionalisering wordt genoemd. Zo voelt de leek zich veel meer dan vroeger deskundig op medisch gebied, door allerlei medische programma's en rubrieken. Dit heeft geleid tot afgenomen respect voor en meer wantrouwen en agressie jegens artsen. Welnu, er is dankzij de media ook sprake van een politieke protoprofessionalisering van de burgers. Dat heeft niet zozeer geleid tot meer democratie, als wel tot meer wantrouwen en ongenoegen.

Meerkeuze en open vragen van het Cito

In de onderzoeksliteratuur wordt deze methode *question asking* genoemd (Kintsch, 1998). De onderzoeker of de docent stelt een aantal vragen, waarop de leerling ofwel een open antwoord kan formuleren ofwel moet kiezen uit een aantal voorgelegde antwoorden. Gemeten wordt dan onder andere of een leerling belangrijke verbanden in de tekst gelegd heeft. Een vooraf opgesteld antwoordmodel stelt de onderzoeker of de leraar in staat deze antwoorden te beoordelen. Een voorbeeld van een open vraag over bovenstaande passage is: Leg uit waarom protoprofessionalisering tot wantrouwen zou leiden.

Cloze test ofwel 'gatentekst'

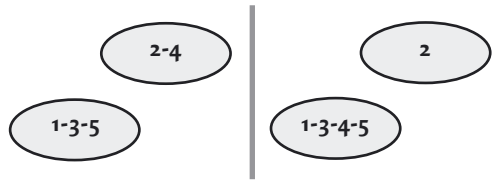
Een cloze test (Taylor, 1953) bestaat uit een tekst, waarbij ieder *n*de, in dit geval het 8e woord, is weggelaten. De leerling moet dan op de open plek een woord invullen waarvan hij denkt dat het daar hoort. Alleen letterlijk correcte woorden worden goed gerekend.[3] Van deze taak is op voorhand niet helemaal duidelijk op welk niveau tekstbegrip gemeten wordt: heeft een leerling genoeg aan het oppervlakteniveau of aan de betekenisrepresentatie of moet toch het diepere situatiemodel ingeschakeld worden?

Sorteertaak ofwel 'groepjes maken'

Sorteertaken werden al lang gebruikt in de psychologie, maar niet specifiek om tekstbegrip te meten, tot het onderzoek van McNamara, Kintsch, Songer en Kintsch (1996). Leerlingen zien na het lezen van de tekst een twintigtal sleutelbegrippen uit de tekst en krijgen de instructie om deze concepten in groepjes te plaatsen, op basis van de verbanden die in de tekst gelegd zijn. Er mogen zoveel groepjes gemaakt worden als de leerlingen zelf willen en de groepjes mogen zo klein of groot zijn als de leerlingen zelf willen. Bij het nakijken wordt een scoremodel gebruikt waarin groepjes gemaakt

zijn op basis van de tekst. Een leerling kan bijvoorbeeld alle oorzaken bij elkaar in één groepje zetten en alle gevolgen in een ander groepje of hij kan juist bij iedere oorzaak een gevolg zetten. Beide strategieën zijn goed. Begrippen die in het verkeerde groepje terechtkomen worden fout gerekend en van de maximale score afgetrokken. Op deze manier wordt wederom getoetst of een leerling belangrijke verbanden uit de tekst heeft weten te leggen. Het voorbeeld hieronder laat een gedeelte van de sorteertaak zien. De volledige sorteertaak bevatte twintig begrippen, hieronder zijn er slechts vijf weergegeven.

1. protoprofessionalisering
2. internationale politiek
3. leken weten meer van professies
4. respect voor deskundigen neemt af



Figuur 1: Een voorbeeld van een ingevulde sorteertaak

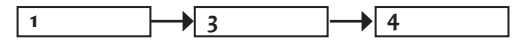
De mentale representatie van tekstpassage 1 zou onderscheid moeten kunnen maken tussen begrip 2 en de andere begrippen. In dit voorbeeld is de sortering links slechts gedeeltelijk correct: alleen het groepje met 1,3 en 5 is correct, het andere groepje relateert twee begrippen die in de tekst niet gerelateerd zijn. De rechtse sortering zou het maximale aantal punten opleveren.

Mental model taak ofwel 'hokjes invullen'

Dit type taak is eveneens bekend uit de psychologie en wordt pas recent toegepast op teksten (McNamara, 2001). Wederom werkt men met sleutelbegrippen, een twintigtal, uit de tekst. De leraar of de onderzoeker maakt op basis van de tekst een stroomschema, een schema met hokjes en pijltjes die zo bijvoorbeeld causale relaties uit de tekst aangeven of opsommingrelaties. De leerling moet dan het juiste begrip in het juiste hokje plaatsen. Eigenlijk is deze vraag de gestuurde vorm van de veel vrijere sorteertaak, waarbij de leerling helemaal vrij is om de begrippen op zijn manier te categoriseren. Bij de mental model taak zijn de relaties namelijk al weergegeven. Ook hier wordt ieder fout geplaatst begrip fout gerekend en afgetrokken van de maximale score. Hierbij wordt gemeten of een leerling de relaties gelegd zijn die belangrijk zijn voor het begrip van de redenering in de tekst.

Een voorbeeld. In de tekst wordt beschreven dat er ongenoegen heerst bij burgers over de Nederlandse politiek. In de tekst komt een reeks van oorzaken en gevolgen voor. Deze reeksen zijn als volgt weer te geven: **OORZAAK ⇒ ONTWIKKELING ⇒ GEVOLG.** Maak op basis van onderstaande begrippen reeksen zoals deze.

1. leken weten meer van professies
2. internationale politiek
3. respect voor deskundigen neemt af
4. minder respect voor de politiek



Figuur 2: Voorbeeld van een ingevulde mental model taak

De goede reeks zou zijn 1, 3, en dan 4. In de mental model taak zoals die gebruikt is in het experiment, stonden wederom twintig begrippen; in dit voorbeeld is de taak versimpeld.

Kenmerken van betrouwbare en valide tekstbegripmethoden

Wanneer is een methode nu 'betrouwbaar en valide'? Om dit objectief te kunnen vaststellen, zijn voorafgaand aan het onderzoek enkele criteria opgesteld.

Allereerst is er het probleem van de 'interne betrouwbaarheid'. De diverse afzonderlijke vragen moeten allemaal tekstbegrip meten en moeten dus een hoge onderlinge samenhang hebben. Het mag niet van toeval afhangen of een leerling het antwoord op vraag 1 weet en dan ook het antwoord op vraag 2: als iemand vraag 1 weet, dan moet de kans ook groot zijn dat hij vraag 2 goed beantwoordt. Het meetinstrument hiervoor is een statistische maat genaamd Cronbach's alpha (Cronbach, 1951). Het betreft een maat tussen 0 en 1. Daarbij geldt: hoe hoger, hoe meer samenhang. Algemeen kan gesteld worden dat vanaf 0.7 een methode intern betrouwbaar wordt genoemd (Van Wijk, 2000): de afzonderlijke vragen hangen dan genoeg met elkaar samen om te concluderen dat ze samen één concept meten.

Ten tweede nemen we aan dat de vier methoden allemaal tekstbegrip meten. De diverse methoden moeten dus overeenkomst vertonen. Dit wordt gemeten door correlaties tussen methoden uit te rekenen en hier geldt ook weer dat dit wordt uitgedrukt in een getal tussen 0 en 1: hoe hoger, hoe meer samenhang (Cronbach, 1951; Cronbach & Meehl, 1955; Holleman, 2000).

Ten derde moet iedere methode volgens ons in staat zijn om goed onderscheid te maken tussen twee verschillende niveaus op school. Daarom doen zowel leerlingen uit 4vwo als uit 6vwo mee aan dit onderzoek. Een goede methode moet volgens ons duidelijk het verschil tussen deze twee groepen aan kunnen geven: 6vwo-ers moeten gemiddeld beter presteren dan 4vwo-ers. Ook moeten 6vwo-ers constanter presteren: zij hebben meer training gehad en de vaardigheid om teksten te lezen is beter ontwikkeld. Daardoor maakt het minder verschil welke tekst ze nu precies lezen.

Ten vierde is het belangrijk om aan te tonen dat deze methoden inderdaad begrip meten en niet een andere factor die er toevallig heel erg op lijkt. IQ is één van die factoren: de resultaten van een leerling zijn natuurlijk wel gerelateerd aan zijn of haar IQ, maar het mag niet hetzelfde zijn. Een goede tekstbegripmethode meet namelijk begrip en geen IQ. Hetzelfde geldt voor voorkennis en attitude. Er moet samenhang zijn tussen de scores op deze verschillende constructen, maar geen totale overlap.

Het onderzoek

Aan dit onderzoek hebben 400 leerlingen deelgenomen, 200 4vwo-ers en 200 6vwo-ers, verdeeld over vier Nederlandse scholen. Het experiment duurde twee lesuren, waardoor iedere leerling op twee verschillende momenten is getoetst. Iedere leerling heeft vier teksten voorgelegd gekregen, met drie verschillende taken en één taak dubbel. Bijvoorbeeld: tekst één met meerkeuzevragen, tekst twee met een sorteertaak, tekst drie met een mental model taak en tekst vier weer met een sorteertaak. Daarnaast vulden ze een IQ test in en vragen over voorkennis en attitude. Ook hebben wij leerlingen gevraagd naar hun mening over de methoden.

De resultaten

Allereerst het woord aan de leerlingen: wat vonden zij zelf van de methoden? Ze klaagden in ieder geval het meest over de gatentekst. Commentaren als 'hoe moet ik nou weten wat de schrijver hier voor woord gebruikte?' waren niet van de lucht. De traditionele eindexamenvragen riepen weinig reactie op. De sorteertaken waren het populairst: 'eigenlijk een soort puzzeltje', maar ook de mental model vraag deed het niet slecht bij de leerlingen. Ondanks het feit dat de leerlingen geen training in de nieuwe methoden hadden gehad, begrepen ze zonder problemen wat de bedoeling was. De leerlingen merkten op, dat ze bij de sorteertaak het gevoel hadden dat ze niets fout konden doen en dat ze dat een prettig gevoel vonden.

Per methode zullen hieronder de uitkomsten van het onderzoek kort op een rijtje gezet worden. Hiervoor gebruiken we de criteria die hierboven uiteengezet zijn. Allereerst wordt in een schema (Tabel 1) de statistische informatie gegeven, vervolgens wordt dit uitgelegd in de tekst.

(Meerkeuze en open) vragen

Deze methode bleek intern niet betrouwbaar (gemiddeld over de teksten: α=0.3). Wat betreft de variatie: de score van 0 betekent dat de variatie tussen de leerlingen bijna helemaal afhankelijk bleek van de tekst. Het is dus eerder de tekst die het resultaat bepaalt, dan de capaciteiten van het kind. Er is geen samenhang met de resultaten op de andere methoden: de meerkeuzevragen lijken iets anders te meten dan de andere methoden. IQ, voorkennis en attitude verklaren bij alle methoden samen 6% van de variatie. Dit wil zeggen dat er wel invloed is van deze constructen als de leerlingen tekstbegripvragen moeten beantwoorden, maar dat deze invloed klein is.

Cloze test

Deze methode is wel intern betrouwbaar (gemiddeld over de teksten: α=0.8). Er is redelijk veel samenhang met de sorteertaak (r=0.46) en veel samenhang met de mental model taak (r=1.00) De cloze toets kan

Methode	Criterium 1: betrouwbaarheid	Criterium 2: correlaties	Criterium 3: variatie door leerling		Criterium 4: hoeveel procent wordt verklaard door IQ, attitude en voorkennis?
			4vwo	6vwo	
MK	0.3	-	0	0	
CL	0.8	1.00 met MM, 0.46 met SO	0.26	0.03	6% voor alle methoden samen
SO	0.65	0.46 met CL, 0.50 met MM	0.42	0.42	
MM	0.5	0.50 met SO, 1.00 met CL	0.34	0.10	

Tabel 1: De resultaten van het vergelijkend tekstbegriponderzoek (MK= meerkeuze vragen, CL= cloze test, SO= sorteertaak, MM= mental model taak)

onderscheid maken tussen 4 en 6vwo, maar in 6vwo is de cloze toets niet meer in staat om voldoende onderscheid tussen de leerlingen te maken. Dit is te zien in de tabel: in 6vwo is nog maar 3 procent van de variatie te verklaren door capaciteiten van het kind en 97% is afhankelijk van de tekst. IQ, voorkennis en attitude spelen een kleine rol bij het beantwoorden van deze taak, want slechts 6% van de variatie wordt bepaald door deze factoren.

Sorteertaak

De interne betrouwbaarheid van de sorteertaak is in onze ogen nog acceptabel (gemiddeld α=0.65 over alle teksten). Er is redelijke samenhang met de cloze testen (r=0.46) en met mental model taken (r=0.50). Het onderscheidend vermogen tussen 4 en 6vwo is prima. De variatie is zowel in 4vwo als in 6vwo prima: een gedeelte van het resultaat is te verklaren door de tekst en een deel door de leerling. Voorkennis, meningen en IQ spelen een kleine, maar ondergeschikte rol (6% verklaarde variatie).

Mental model taak

De betrouwbaarheid is hier α=0.5 en dat is aan de lage kant. Er is veel samenhang met de cloze testen (r=1.00) en redelijk veel samenhang met de sorteertaken (r=0.50), wat aangeeft dat deze methoden dus hetzelfde construct meten. De mental model taak kan goed onderscheid maken tussen de twee niveaus 4 en 6vwo. Ook hier zijn voorkennis, attitude en IQ van belang, maar we meten duidelijk iets anders met de methode, omdat slechts 6% van de variatie verklaard kan worden met behulp van deze factoren.

Conclusie en discussie

De resultaten van dit experiment lijken erop te wijzen dat de sorteertaak de meest geschikte taak zou kunnen zijn om tekstbegrip te meten: op geen van onze criteria hebben we deze taak kunnen afwijzen. De vraag welke methode op de tweede plaats komt, is moeilijker te beantwoorden. De andere taken hebben allemaal voor- en nadelen. De cloze taak heeft als voordeel dat de taak intern erg betrouwbaar is, maar als nadeel dat hij in 6vwo de verschillen tussen leerlingen niet goed kon weergeven. De mental model taak heeft als nadeel een relatief onbetrouwbaar resultaat, maar is wel geschikt voor beide groepen leerlingen. De (open en gesloten) tekstbegripvragen komen bij dit experiment als slechtste optie uit de bus. Ze scoorden het laagst op interne betrouwbaarheid, op samenhang met andere methoden en weinig van de variatie tussen de leerlingen bleek toe te wijzen aan de capaciteiten van de leerling.

Zijn deze resultaten uit dit onderzoek generaliseerbaar? Een belangrijke kanttekening die we moeten plaatsen is, dat in dit onderzoek vier teksten zijn gebruikt, wat te weinig is om te concluderen dat de resultaten voor *alle* teksten gelden. Daarnaast zijn de gebruikte teksten persuasief van aard. Het zou kunnen zijn dat voor meer informerende teksten het resultaat anders zou uitpakken. Toch zijn er ook aanwijzingen voor de relevantie van situatiemodellen in informerende teksten (zie bijvoorbeeld Kamalski, Lentz en Sanders, 2004; Land, Sanders, Lentz en van den Bergh, 2002).

Welke betekenis hebben deze resultaten nu voor tekstbegriptoetsing in het voortgezet onderwijs, met de voorgaande beperkingen in acht genomen? Nieuw theoretisch en empirisch werk, zoals in dit experiment, suggereert dat het zeer de moeite waard is om systematisch na te gaan wat de mogelijkheden zijn voor recent ontwikkelde situatie model taken in examens, zoals de sorteertaak. Dit type taak lijkt betrouwbaar en valide. In dit experiment scoren de sorteertaken aanmerkelijk beter op betrouwbaarheid en validiteit dan de klassieke examenvragen.

Noten

1. Een uitgebreider verslag van dit onderzoek is te vinden in Kamalski, Sanders, Lentz en van den Bergh (ter publicatie aangeboden). How to measure situation model representations. On the validity of text comprehension tasks.
2. De tekst is gebaseerd op de rede van Prof. dr. A.C. Zijderveld op 11 mei 2002 gepubliceerd in het opinieweekblad Vrij Nederland.
3. Later zijn ook varianten van de cloze test ontwikkeld waarbij ook synoniemen goed gerekend werden of een m.c.-vorm waarbij meerdere mogelijkheden werden gegeven (zie bijvoorbeeld Jochems en Montens 1987). Deze vormen correleren echter sterk en daarom is in dit experiment de voorkeur gegeven aan de simpelste scoringsmethode.

Literatuur

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-333.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 280-302.

Holleman, B. (2000). *The forbid/allow asymmetry. On the cognitive mechanisms underlying wording effects in surveys.* Amsterdam: Rodopi.

Jochems, W., & Montens, F. (1987). De multiple-choice cloze toets als algemene taalvaardigheidstoets. *Tijdschrift voor Onderwijsresearch, 12*, 133-143.

Kamalski, J., Lentz, L., & Sanders, T. (2004). Coherentiemarkering in informerende en persuasieve teksten. Een empirisch onderzoek naar cognitieve en affectieve effecten. *Tijdschrift voor Taalbeheersing, 26 (2)*, 85-104.

Kintsch, W. (1998). *Comprehension. A paradigm for cognition.* Cambridge: CUP.

Land, J., & Sanders, T. (te verschijnen). Lezen. Leuk én leerzaam? Over tekstbegrip op het Vmbo. In Raukema, A. & Schram, D. (Red.), *Tekstcomplexiteit*.

Land, J., Sanders, T., Lentz, L., & Bergh, van den, H. (2002). Coherentie en identificatie in studieboeken. Een empirisch onderzoek naar tekstbegrip en tekstwaardering op het vmbo. *Tijdschrift voor taalbeheersing, 4*, 281-302.

McNamara, D. S. (2001). Reading both High-Coherence and Low-Coherence Texts: Effects of Text Sequence and Prior Knowledge. *Canadian Journal of Experimental Psychology, 55*, 51-62.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1-43.

Singer, M. (1990). *Psychology of Language. An introduction to sentence and discourse processes.* Hillsdale, NJ: Erlbaum.

Taylor, W. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly, 30*, 414-438.

Wijk, C. van (2000). *Toetsende Statistiek: Basistechnieken. Een praktijkgerichte inleiding voor onderzoekers van taal, gedrag en communicatie.* Bussum: Coutinho.