Maaike van den Haak,
Menno de Jong en Peter Jan Schellens

Hardopdenkprotocollen en gebruikersonderzoek

Volledigheid en reactiviteit van de synchrone hardopdenkmethode

Samenvatting

De hardopdenkmethode heeft zich ontwikkeld tot een gangbare onderzoeksmethode voor uiteenlopend lees- en schrijfonderzoek. In dit artikel staat de validiteit van de hardopdenkmethode als pretestinstrument centraal. Discussies over de validiteit van hardopdenkprotocollen betreffen de reactiviteit van de methode en de volledigheid van de verbalisaties. De vraag wordt beantwoord in hoeverre de aard van het testobject van invloed is op de reactiviteit van de methode en de volledigheid van de protocollen. Daartoe zijn synchrone en retrospectieve hardopdenkprotocollen vergeleken voor twee typen internet-applicaties: een online bibliotheekcatalogus en een gemeentelijke website. In beide gevallen blijken synchrone hardopdenkprotocollen minder geverbaliseerde problemen te bevatten dan retrospectieve protocollen. De reactiviteit van de methode blijkt samen te hangen met het testobject: in de bibliotheekcatalogus leidde de opdracht om synchroon hardop te denken tot meer observeerbare fouten in de taakuitvoering, op de gemeentelijke website niet. Deze verschillen kunnen worden herleid tot verschillen in taken tussen de beide testobjecten.

1. Inleiding

De hardopdenkmethode is een dominante benadering geworden voor het registreren van gebruiksprocessen van communicatiemiddelen, met name websites en instructieve documenten. Het doel van een dergelijke registratie is vaak het pretesten van de communicatiemiddelen: het opsporen van mogelijke gebruikersproblemen en het op basis daarvan formuleren van revisiesuggesties. In grote lijnen komt de methode erop neer dat proefpersonen aan bepaalde taken worden gezet, met als extra opdracht om tijdens de taakuitvoering voortdurend hun gedachten te verbaliseren. De hardopdenkprotocollen, die bestaan uit een weergave van zowel de acties als de verbalisaties van proefpersonen, leveren gedetailleerde procesgegevens op over de wijze waarop gebruikers te werk gaan met het betreffende communicatiemiddel. Deze procesgegevens zorgen ervoor dat de onderzoeker precies te weten komt waar zich problemen voordoen, zodat deze vervolgens in een revisieronde verholpen kunnen worden.

Al sinds de vroege jaren '80 is de methode in de Verenigde Staten gebruikt om allerlei documenten te evalueren. Aanvankelijk kregen proefpersonen slechts de opdracht om een tekst hardop te lezen en hun gedachten daarbij te verwoorden (Flower, Hayes & Swarts, 1983; Swaney e.a., 1981). De

begrijpelijkheid van de tekst stond in dergelijk onderzoek centraal. Daarna is er in toenemende mate gewerkt met specifieke taken die proefpersonen aan de hand van een instructieve tekst moesten verrichten (Schriver, 1987; Jansen & Steehouder, 1989). In die gevallen gaat het, naast begrijpelijkheid, ook om de vindbaarheid en de praktische toepassing van informatie. Dieli (1986) beschrijft en vergelijkt hardopdenkonderzoek mét (*user protocols*) en zonder (*reader protocols*) specifieke taken en concludeert dat beide vormen van hardopdenkonderzoek inderdaad verschillende soorten resultaten opleveren. Meer recent vestigen Lentz & Pander Maat (2003) opnieuw de aandacht op de mogelijke waarde van hardopdenkprotocollen zonder taken. Zij beschrijven enkele studies waarin de hardopdenkmethode zonder taken wordt vergeleken met twee andere pretestmethoden (de plus-en-minmethode en het computerprogramma Focus) en concluderen dat lezersprotocollen in ieder geval geschikter zijn om begripsproblemen op te sporen dan beide andere pretestmethoden. Over de hele linie heerst er grote tevredenheid over het gebruik van hardopdenkonderzoek voor het pretesten van communicatiemiddelen. Aangenomen wordt dat de geconstateerde problemen goede voorspellers zijn van de problemen die echte gebruikers in het dagelijks leven zullen gaan ondervinden.

De opkomst van het internet heeft de hardopdenkmethode als pretestinstrument alleen maar populairder gemaakt. Teksten die via het internet worden aangeboden, krijgen er bijna per definitie een zware selectiedimensie bij (het zoeken en navigeren op de website), die de aandacht voor veel andere soorten gebruikersproblemen lijkt te verdringen. Op natuurlijke wijze komen bij de evaluatie van websites twee afzonderlijke onderzoekstradities bij elkaar: die van tekstevaluatie (De Jong & Schellens, 1995) en usability testing (Nielsen, 1993; Dumas & Redish, 1999). Binnen de traditie van de usability testing golden hardopdenkprotocollen al van oudsher als de dominante evaluatiemethode. Een en ander heeft overigens wel geleid tot een onderbelichting van de tekstuele en visuele *content* van websites en een overmatige aandacht voor navigatieproblemen.

In de context van usability testing heeft de validiteit van hardopdenkprotocollen nauwelijks ter discussie gestaan. Integendeel, de methode wordt doorgaans als onbetwistbaar criterium gebruikt om andere evaluatiemethoden mee te vergelijken. Daarbij wordt sterk geleund op het standaardwerk van Ericsson & Simon (1993), waarin op basis van een groot aantal empirische studies wordt geconcludeerd dat hardopdenkgegevens, onder een aantal stringente voorwaarden, in principe valide zijn. Zo mag de taak zelf niet te moeilijk zijn voor de proefpersoon (anders zal deze stoppen met hardop denken), maar mag het anderzijds ook niet gaan om een min of meer geautomatiseerde taak. Ook wordt er onderscheid gemaakt tussen drie niveaus van verbaliseren. De eerste twee niveaus (het verklanken van informatie die al in dezelfde vorm in het korte termijn geheugen aanwezig is en gedachten die als enige transformatie vertaald moeten worden in verbale informatie) geven volgens Ericsson & Simon geen problemen. Alleen wanneer proefpersonen meerdere stappen moeten zetten om van hun gedachten naar verbalisaties te komen (dit wordt het derde niveau genoemd), levert hardopdenken volgens Ericsson & Simon geen valide gegevens meer op. Hierbij is te denken aan het filteren van de verbalisaties, het geven van verklaringen voor het eigen gedrag of het binnenhalen van kennis uit het lange termijn geheugen. Daarnaast geven Ericsson & Simon strenge gedragsregels voor de proefleider, die de proefpersoon er zonodig wel aan moet herinneren hardop te denken ("keep talking"), maar op geen enkele andere manier mag interveniëren in het proces.

De voornaamste discussie binnen het domein van de usability testing lijkt te gaan over de vraag in hoeverre het theoretisch raamwerk van Ericsson & Simon eigenlijk van toepassing is op de praktijk van usability testing. Wright & Monk (1991) deden een studie waarin ze een strikte uitvoering van de hardopdenkmethode (exact volgens de regels van Ericsson & Simon opgezet) probeerden te vergelijken met een vrijere variant. De vergelijking viel in het water doordat geen van hun proefleiders in de strikte conditie zich voldoende aan de richtlijnen hield. Tot vergelijkbare conclusies kwamen Boren & Ramey (2000) in een veldonderzoek naar de werkwijze van usability professionals: de gedragsregels van Ericsson & Simon worden niet nageleefd in de praktijk van usability testing. Boren & Ramey betogen dat dit niet per se slecht hoeft te zijn, omdat de doelstellingen van een usability test sterk afwijken van die van veel ander hardopdenkonderzoek. Niet het hele proces, maar de problemen die zich voordoen, zijn het centrale aandachtspunt. Ook is er doorgaans het nodige te observeren, waardoor de afhankelijkheid van de verbalisaties minder groot is. Als een alternatief voor de traditionele benadering stellen Boren & Ramey een "speech communication"-benadering voor, waarin de proefleider in sommige (beregelde) gevallen wel mag ingrijpen in het proces en ook meer interageert met de proefpersoon.

Buiten de context van usability testing is er doorlopend aandacht geweest voor de validiteit van hardopdenkprotocollen. Diverse studies hebben aangetoond dat het gebruik van de hardopdenkmethode in bepaalde contexten kan leiden tot vertekeningen. Proefpersonen kunnen hun taken beter of juist slechter uitvoeren dankzij de opdracht hardop te denken. Proefpersonen die slechter presteren, hebben te lijden onder een te zware cognitieve belasting. In een eerdere studie zagen we dat dergelijke effecten ook in een usability test kunnen optreden: bij de evaluatie van een online bibliotheekcatalogus stelden we vast dat hardopdenkende proefpersonen meer observeerbare fouten maakten en minder taken correct uitvoerden dan stilwerkende proefpersonen (Van den Haak, De Jong & Schellens, 2003). Ericsson & Simon (1993) veronderstellen dat proefpersonen minder gaan verbaliseren wanneer hun taak te moeilijk wordt. Dat kan zo zijn, maar er blijkt in de praktijk soms ook een negatief effect te zijn op de kwaliteit van de taakuitvoering. Proefpersonen die beter presteren, profiteren van een mogelijk faciliterende bijdrage van de hardopdenkopdracht: het feit dat ze gedwongen worden hardop te denken, zorgt ervoor dat ze zorgvuldiger of anderszins beter te werk gaan dan normaal. Dit lijkt bijvoorbeeld het geval in leesprocessen (Silvén & Vauras, 1992; Loxterman, Beck & McKeown, 1994; Kucan & Beck, 1997). Ook de manier waarop proefpersonen te werk gaan, kan door het hardopdenken worden beïnvloed. Janssen, Van Waes & Van den Bergh (1996) lieten bijvoorbeeld zien dat de opdracht om hardop te denken het schrijfproces (geoperationaliseerd als het pauzegedrag) van proefpersonen beïnvloedde, met name als het ging om een complexe schrijftaak. Mumma, Draguns & Seibel (1993) stelden vast dat hardopdenkende proefpersonen anders (efficiënter) te werk gingen bij het toekennen van persoonlijkheidskenmerken aan mensen dan proefpersonen die stil werkten. Knoblich & Rhenius (1995) vonden in hun onderzoek dat de opdracht om hardop te denken invloed had op de manier waarop proefpersonen een koelinstallatie bedienden.

Dergelijke studies roepen vragen op over de validiteit van de hardopdenkmethode. Deze lijkt in principe contextafhankelijk te zijn: in het ene geval leidt hardopdenken tot een betere taakuitvoering, in het andere geval tot een minder effectieve taakuitvoering en in verschillende situaties blijken proefpersonen anders te werk te gaan. We kunnen op grond van zulk onderzoek natuurlijk niet direct concluderen dat de validiteit van hardopdenkprotocollen als pretestmethode dus ook problematisch zal zijn, maar het onderzoek vestigt wel de aandacht

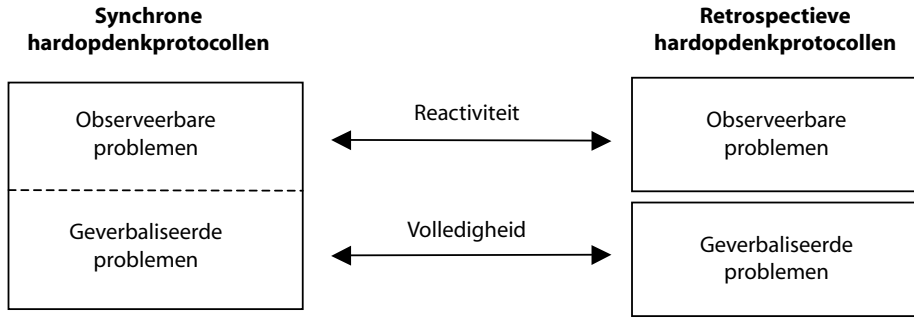
op de mogelijkheid dat er validiteitsproblemen optreden. Voor een usability test zouden alle hierboven beschreven afwijkingen een bedreiging vormen. Blijkbaar kunnen de resultaten van een hardopdenktest een (positief dan wel negatief) vertekend totaalbeeld geven van de gebruiksvriendelijkheid van een communicatiemiddel. Het percentage correct uitgevoerde opdrachten hoeft niet te corresponderen met het succespercentage onder stilwerkende gebruikers. Ook kan het hardopdenken problemen in een communicatiemiddel veroorzaken en andere problemen verhullen. Usability professionals hoeven hier natuurlijk niets van te merken: de uitkomsten van een usability test zullen er in hun perceptie niet minder overtuigend om zijn.

Het al dan niet optreden van vertekeningen doordat proefpersonen hun gedachten moeten verbaliseren, lijkt voor een belangrijk deel samen te hangen met de (deel)taken die zij moeten uitvoeren. Bij elk empirisch tegenbewijs tegen de validiteit van de hardopdenkmethode verwijzen Ericsson & Simon (1993) ofwel naar onregelmatigheden in proefleidergedrag of hardopdenkinstructies ofwel naar onderzoeksdoelstellingen die de type 1- en type 2-verbalisaties te boven gaan. Algemene uitspraken over de validiteit van de hardopdenkmethode liggen onzes inziens niet voor de hand. In plaats daarvan zou een uitvoerige verkenning van de relatie tussen taaksoorten en de validiteit van hardopdenkprotocollen een essentiële stap zijn in de verdere ontwikkeling van de methode. De typologie van verbalisaties zoals Ericsson & Simon (1993) die gebruikten, biedt in dit opzicht nog te weinig houvast.

Ook binnen de context van usability testing lijkt een bezinning op de geschiktheid van de methode op allerhande deeltaken op zijn plaats. In de literatuur over human-computer interaction wordt de suggestie gewekt dat het testen van software en het testen van een website feitelijk op hetzelfde neerkomt, ondanks de grote verschillen tussen een regulier softwarepakket en een website. Hoewel hardopdenkprotocollen in principe op dezelfde manier kunnen worden gebruikt voor beide testobjecten, is het goed denkbaar dat het anders gesteld is met de validiteit van de methode. Ook in het onderzoek naar tekstevaluatie is vooralsnog weinig rekening gehouden met de invloed van het testobject. In de overzichten die tot nu toe gemaakt zijn over het methodologische onderzoek naar tekstevaluatie (De Jong, 1998; De Jong & Schellens, 1995; 2000), zijn de inzichten uit de human-computer interaction onverkort meegenomen. Het is wederom de vraag of dat gerechtvaardigd is.

In dit artikel bespreken we twee studies waarin de resultaten van synchrone en retrospectieve hardopdenkprotocollen met elkaar vergeleken worden. Bij synchrone hardopdenkprotocollen moeten proefpersonen direct tijdens de taakuitvoering hun gedachten verbaliseren; bij retrospectieve hardopdenkprotocollen verrichten proefpersonen eerst in stilte de taken en verbaliseren ze hun gedachten achteraf, terwijl ze naar een video-opname van hun taakuitvoering kijken. De twee methoden werden al door Nielsen (1993) als concurrenten beschreven. Recent onderzoek waarin de verbalisaties van retrospectief hardopdenkende proefpersonen werden afgezet tegen de oogbewegingen tijdens de taakuitvoering, geeft nadere ondersteuning voor deze, wat bewerkelijker, vorm van hardopdenkonderzoek: de verbalisaties achteraf bleken goed overeen te komen met de oogbewegingen die de proefpersonen tijdens de taakuitvoering hadden gemaakt (Guan e.a., 2006). Een vergelijking van beide vormen van hardopdenken levert een beeld op van de reactiviteit van de synchrone hardopdenkmethode en de volledigheid van de verbalisaties van proefpersonen tijdens de taakuitvoering (zie Figuur 1). Bij synchrone hardopdenkprotocollen worden observeerbare en geverbaliseerde problemen tegelijkertijd verzameld; bij retrospectieve hardopdenkprotocollen zijn de twee soorten problemen verdeeld over twee ronden: de observeerbare problemen tijdens het stil

werken, de geverbaliseerde problemen tijdens het retrospectief hardopdenken. Een vergelijking van de observeerbare problemen betreft de reactiviteit van synchroon hardopdenken: de taakuitvoering wordt in de retrospectieve conditie immers in het geheel niet beïnvloed door een extra hardopdenktaak voor de proefpersonen. Een vergelijking van de geverbaliseerde problemen heeft betrekking op de volledigheid van synchrone hardopdenkprotocollen: de verbalisaties van de proefpersonen in de retrospectieve conditie staan immers niet onder druk van de belasting die de taakuitvoering op zich al met zich meebrengt.



Figuur 1: *Schematisch overzicht van het onderzoek*

Het gaat in dit onderzoek niet alleen om een vergelijking van de twee methoden, maar ook om een vergelijking van testobjecten. We kijken of het onderscheid tussen synchrone en retrospectieve hardopdenkprotocollen op dezelfde manier werkt voor een online bibliotheekcatalogus en een gemeentelijke website. De eerste studie betrof de evaluatie van de bibliotheekcatalogus van de Universiteit Utrecht (Van den Haak, De Jong & Schellens, 2004) en was een uitgebreide replicatie van een onderzoek dat we eerder in het *Tijdschrift voor Taalbeheersing* beschreven (Van den Haak, De Jong & Schellens, 2003). De tweede studie had betrekking op de website van de gemeente Haarlem (Van den Haak, De Jong & Schellens, 2007). In beide studies vergeleken we de twee typen hardopdenkprotocollen overigens ook nog met een derde variant, namelijk *constructive interaction* (waarbij proefpersonen in tweetallen aan de taken moesten werken), maar die laten we in dit artikel buiten beschouwing. Voordat we de opzet en de resultaten van ons onderzoek beschrijven, zullen we in de volgende paragraaf eerst ingaan op de kenmerkende verschillen die optreden tijdens het werken met de twee testobjecten. De verschillen worden gerelateerd aan verwachtingen ten aanzien van de reactiviteit van de hardopdenkmethode en de volledigheid van de verbalisaties in de protocollen.

2. Gebruikerstaken in de online catalogus en op de website

In de bibliotheekcatalogus en op de website gaven we de proefpersonen taken die kenmerkend waren voor het betreffende communicatiemiddel. In de online bibliotheekcatalogus moesten de proefpersonen in vijf opdrachten met diverse zoektechnieken werken om allerlei publicaties te vinden. Ze moesten zoeken op auteur en op onderwerp. Ze moesten combinaties van zoektermen gebruiken, selecteren op taal en op jaar van publicatie en zoek-

termen trunceren. In de gemeentelijke website moesten de proefpersonen in vijf scenario's in totaal twaalf deeltaken verrichten, die betrekking hadden op het kopen van een huis, het aanbieden van afval, het inschrijven bij de gemeente als inwoner, het betalen van gemeentelijke belastingen en het verkrijgen van een parkeervergunning. Bij een vergelijking van de taakuitvoering in beide studies komen we tot enkele kenmerkende verschillen.

Het werken met de online bibliotheekcatalogus kan getypeerd worden als complex maar overzichtelijk. Literatuur zoeken in een bibliotheekcatalogus is voor studenten een erkend lastige taak. In ons geval moesten de proefpersonen ook nog zoeken in een catalogus waar ze niet aan gewend waren (de proefpersonen studeerden aan de Universiteit Twente; de catalogus was van de Universiteit Utrecht). De complexiteit van de taak had vooral te maken met twee aspecten: (a) een gebrek aan voorkennis over informatiezoeken, en (b) problemen met de toepassing van de eigen voorkennis op een nieuw systeem. In beide gevallen moesten zij zich een mentaal model proberen te vormen van de opzet van de catalogus. Hoewel de vijf taken geheel afzonderlijk van elkaar konden worden verricht, is het goed denkbaar dat er bepaalde leereffecten zouden optreden. Dat laatste had te maken met de overzichtelijkheid: het aantal knoppen waaruit de proefpersonen moesten kiezen, was in wezen beperkt.

Het werken met de website van de gemeente Haarlem was in principe minder complex. Het ging om minder specialistische taken, die in principe aansloten op de dagelijkse leefwereld van de proefpersonen. Maar de taken waren aan de andere kant ook minder overzichtelijk. In plaats van één interface waarmee alle taken verricht moesten worden, moesten de proefpersonen navigeren over de hele site. Een belangrijke taak was steeds uit te vinden waar de gewenste informatie "verstopt" zat. Daarbij kregen de proefpersonen te maken met diverse menu's: op één webpagina stonden soms horizontaal en verticaal verschillende menu's waaruit gekozen moest worden en als er een keuze was gemaakt kwam op het lagere niveau vaak weer een ander menu tevoorschijn. Leereffecten waren in deze context zo goed als onmogelijk. Elk scenario stelde andere eisen aan de proefpersonen.

Het grootste verschil tussen beide testobjecten betrof het belang van leesactiviteiten in de taakuitvoering. Van lezen, browsen of skimmen was in de online catalogus nauwelijks sprake. Alleen als de proefpersonen ervoor kozen de online help te raadplegen, moesten ze een hoeveelheid tekst verwerken. Op de gemeentelijke website, daarentegen, waren lees-, browse- en skimactiviteiten juist erg dominant. Dat gold in de eerste plaats voor de *content* van de website, die uiteindelijk gelezen moest worden. Maar dergelijke activiteiten waren evenzeer van belang voor alle menu's, waarvan de linknamen meer waren dan lexicale aanduidingen van in principe eenduidige begrippen. Linknamen als "Leven & wonen" en "Komen & gaan" lijken op kopjes in een reguliere tekst, waarbij er een voortdurende spanning is tussen informativiteit, inhoudelijke dekkendheid, bondigheid en aantrekkelijkheid.

Op basis van de complexiteit van de online bibliotheekcatalogus was het te verwachten dat hier gemakkelijk problemen met cognitieve belasting zouden optreden. Proefpersonen moeten graven in hun herinnering naar de functies die normaal gesproken beschikbaar zijn in een bibliotheekcatalogus en moeten deze functies relateren aan de knoppen op de interface. Het is goed denkbaar dat er bij een dergelijk testobject problemen optreden als de proefpersonen ook nog hardop moeten denken. Dat bleek in ieder geval in een vorige studie, waarin een andere online bibliotheekcatalogus centraal stond (Van den Haak, De Jong & Schellens, 2003). Het is ook voorstelbaar dat leereffecten in die situatie uitblijven die er bij stilwerkende proefpersonen wel zouden zijn. Dat leidde tot de hypothese:

H1 In de online bibliotheekcatalogus maken proefpersonen in de synchrone conditie meer observeerbare fouten dan proefpersonen in de retrospectieve conditie.

Op basis van eerder onderzoek naar de invloed van hardopdenken op het leesproces verwachtten we bij de gemeentelijke website juist het tegenovergestelde (Silvén & Vauras, 1992; Loxterman, Beck & McKeown, 1994; Kucan & Beck, 1997). Verondersteld kan worden dat de opdracht om hardop te denken leidt tot een meer systematische werkwijze, waarbij moeilijk of niet te verbaliseren deelactiviteiten als browsen of skimmen worden vervangen door lezen. Daarnaast zou ook de vertraging van de taakuitvoering, die gepaard gaat met synchroon hardopdenken, kunnen leiden tot een minder impulsieve aanpak. Dergelijke mogelijke effecten van hardopdenken werden niet meegenomen in het overzicht van Ericsson (1988), die zich beperkte tot de effecten van het hardop lezen en het verbaliseren na elke zin in de tekst, en die op grond daarvan aanmerkelijk optimistischer was over de beperkte effecten van de hardopdenkmethode op tekstbegrip. Dat leidde tot de volgende hypothese:

H2 Op de gemeentelijke website maken proefpersonen in de synchrone conditie minder observeerbare fouten dan proefpersonen in de retrospectieve conditie.

Wat betreft de verbalisaties verwachtten we geen interactie-effect tussen de hardopdenkvariant en het testobject. Hier verwachtten we dat de retrospectieve hardopdenkmethode in beide gevallen meer ruimte zou geven aan de proefpersonen om problemen te verbaliseren. Onze laatste hypothese was derhalve:

H3 Zowel in de online bibliotheekcatalogus als op de gemeentelijke website verbaliseren proefpersonen in de synchrone conditie minder problemen dan proefpersonen in de retrospectieve conditie.

3. Onderzoeksopzet

In deze paragraaf geven we een beknopt overzicht van de opzet van het onderzoek. Een uitgebreidere weergave is te vinden in de afzonderlijke verslaglegging van beide studies (Van den Haak, De Jong & Schellens, 2004; 2007).

De online bibliotheekcatalogus van de Universiteit Utrecht werd getest met 40 studenten als proefpersonen: 20 in de synchrone hardopdenkconditie, en 20 in de retrospectieve hardopdenkconditie. De proefpersonen kregen vijf taken met de catalogus. In de synchrone conditie moesten ze tijdens de taakuitvoering hun gedachten verbaliseren. In de retrospectieve conditie bestond het onderzoek uit twee stappen: eerst moesten de proefpersonen stilwerkend de taken met de catalogus verrichten; vervolgens moesten ze hun gedachten verbaliseren terwijl ze naar een video-opname van hun taakuitvoering keken.

De website van de gemeente Haarlem werd eveneens met 40 studenten geëvalueerd, die verdeeld werden over de twee condities. In dit geval kregen de proefpersonen vijf scenario's voorgeschoteld, die elk bestonden uit twee tot vier taken. In totaal moesten ze twaalf vragen aan de hand van de website beantwoorden.

In beide gevallen richtte de analyse zich primair op de detectie van mogelijke gebruikersproblemen. Daarbij maakten we onderscheid tussen drie typen problemen. Het eerste type

bestond uit de *observeerbare* problemen: dat waren problemen waarbij de proefpersoon in de fout ging zonder dat dit leidde tot een vorm van verbalisatie. De detectie van deze problemen was gebaseerd op een vergelijking van het ideale handelingsverloop met de gebeurtenissen op het scherm. Het kon bijvoorbeeld zijn dat een proefpersoon een verkeerde zoekterm invulde of een verkeerde link aanklikte. Goed beschouwd zouden er voor dergelijke problemen helemaal geen hardopdenkgegevens nodig zijn. Het tweede type problemen bestond uit de *geverbaliseerde* problemen: dat waren problemen die niet zichtbaar waren op het scherm, maar die wel door een proefpersoon onder woorden werden gebracht. Het kon bijvoorbeeld zijn dat een proefpersoon duidelijk maakte te twijfelen tussen meerdere links of in verwarring te zijn geraakt. Het derde type bestond uit *gemengde* problemen: dat waren problemen waarbij de proefpersoon zowel een observeerbare fout had gemaakt als geverbaliseerd had. Dit is enigszins een restcategorie, omdat de relatie tussen verbalisaties en handelingen verschillende vormen kon aannemen. Het kon zijn dat een proefpersoon eerst een probleem verbaliseerde, waarna het in de daaropvolgende actie ook zou blijken. Maar het gebeurde ook dat een proefpersoon eerst een fout maakte en er vervolgens over begon te praten. Het kwam ook voor dat de foutieve handeling en de verbalisatie van het probleem op verschillende momenten in het proces plaatsvonden. Vanwege dit gemengde karakter hebben we besloten om deze laatste categorie in onze analyses buiten beschouwing te laten en ons te concentreren op de zuivere observeerbare en geverbaliseerde problemen.

4. Resultaten

De hoofdeffecten van het testobject en van de hardopdenkconditie en de interactie-effecten tussen beide zijn getoetst in een multivariate variantieanalyse met het aantal observeerbare problemen en het aantal geverbaliseerde problemen als afhankelijke variabelen. De resultaten van deze analyse zijn te vinden in Tabel 1 (gemiddelde scores en standaarddeviaties) en Tabel 2 (toetsing).

Een eerste bevinding is dat het aantal observeerbare problemen per proefpersoon significant verschilde tussen de twee testobjecten. In de online bibliotheekcatalogus werden, ongeacht de hardopdenkconditie, significant meer observeerbare fouten gemaakt dan in de gemeentelijke website. De eta^2 van .17 duidt op een groot effect. Dit correspondeert met de aard van de twee testobjecten: bij het gebruik van een website moeten relatief meer mentale en minder fysieke handelingen worden verricht. Wel moet daarbij worden vastgesteld dat het aantal geverbaliseerde problemen niet verschilde tussen de beide testobjecten.

Tabel 1: *Gemiddelde aantallen problemen per proefpersoon in de vier condities*

	Observeerbare problemen		**Geverbaliseerde problemen**	
	Synchroon	**Retrospectief**	**Synchroon**	**Retrospectief**
Online bibliotheekcatalogus	5.5 (2.5)	3.1 (1.7)	1.7 (2.1)	3.4 (2.3)
Gemeentelijke website	1.8 (1.9)	2.8 (2.8)	2.0 (2.1)	3.4 (2.5)

Noot: Standaardafwijking (SD) tussen haakjes

Tabel 2: *Toetsing van de verschillen (Manova)*

	***F*-waarde**	**Vrijheidsgraden**	**Significantie**	***Eta²***
Overall-effect testobject	8.230	2.75	$p<.001$	.18
Overall-effect hardopdenkconditie	4.628	2.75	$p<.05$	.11
Overall interactie-effect	6.266	2.75	$p<.005$	.14
Effect testobject op:				
- observeerbare problemen	15.166	1.76	$p<.001$	.17
- geverbaliseerde problemen	.062	1.76	n.s.	
Effect hardopdenkconditie op:				
- observeerbare problemen	1.858	1.76	n.s.	
- geverbaliseerde problemen	9.296	1.76	$p<.005$	.11
Interactie-effect op:				
- observeerbare problemen	11.611	1.76	$p<.001$	.13
- geverbaliseerde problemen	.062	1.76	n.s.	

Vervolgens blijkt dat de hardopdenkconditie van invloed was op het aantal geverbaliseerde problemen. In de retrospectieve hardopdenkconditie werden meer problemen geverbaliseerd dan in de synchrone hardopdenkconditie. Wederom duidt de eta^2 (.11) op een aanzienlijk effect. Met deze bevinding is de derde onderzoekshypothese, die betrekking had op de volledigheid van synchrone hardopdenkprotocollen, bevestigd. Er blijken achteraf problemen in het proces te zitten die tijdens de taakuitvoering niet geverbaliseerd worden door de proefpersonen. En deze constatering is onafhankelijk van de aard van het testobject. Onzes inziens even interessant is de constatering dat er geen verschil is tussen de condities in aantallen observeerbare problemen. Er zijn, in overeenstemming met onze bespreking van de literatuur, geen eenduidige uitspraken te doen over de reactiviteit van synchrone hardopdenkprotocollen.

Wanneer we kijken naar de interactie-effecten, blijkt dat het aantal observeerbare problemen samenhangt met het samenspel tussen testobject en hardopdenkconditie. Ook hier is de eta^2 aanzienlijk (.13). In het geval van de online bibliotheekcatalogus lijkt het gebruik van de synchrone hardopdenkmethode problemen te veroorzaken bij de proefpersonen. Dit wordt bevestigd door een t-test: proefpersonen in de synchrone conditie hebben significant meer observeerbare problemen in de catalogus dan proefpersonen in de retrospectieve conditie (t=3.601, df=38, $p<.001$, Cohen's d=1.14). Daarmee is de eerste onderzoekshypothese bevestigd. In het geval van de gemeentelijke website is de tendens juist in de tegenovergestelde richting: het gebruik van de synchrone hardopdenkmethode lijkt problemen te voorkomen bij de proefpersonen. Dit verschil is echter niet significant (t-test, t=-1.365, df=38, p=.180). De tweede onderzoekshypothese is daarmee niet bevestigd, maar de onderliggende veronderstelling dat de reactiviteit van de synchrone hardopdenkmethode samenhangt met de aard van het testobject blijft wel overeind.

5. Discussie

In de context van formatief evaluatieonderzoek wordt de hardopdenkmethode doorgaans gezien als een valide benadering, die veel van de nadelen van methoden op basis van zelfrapportage (zoals de plus-en-minmethode) niet heeft. Een onmiskenbaar voordeel van de methode, althans wanneer er gewerkt wordt met taken, is dat we niet hoeven te vertrouwen op oordelen en inschattingen van de proefpersonen, maar dat we ook kunnen zien of ze daadwerkelijk met het communicatiemiddel uit de voeten kunnen. Het huidige onderzoek plaatst enkele vraagtekens bij de zekerheden die we meenden te hebben. Zo blijkt dat de problemen die proefpersonen tegenkomen in sommige gevallen worden veroorzaakt doordat zij tijdens de taakuitvoering hardop moeten denken. Deze bevinding correspondeert met de resultaten van een vorige studie, waarin synchrone en retrospectieve hardopdenkprotocollen werden vergeleken voor de evaluatie van een andere online bibliotheekcatalogus (Van den Haak, De Jong & Schellens, 2003). In die vorige studie vonden we zelfs dat ook het percentage correct uitgevoerde taken verschilde: synchroon hardopdenkende proefpersonen waren daarin significant minder succesvol dan retrospectief hardopdenkende proefpersonen. In het huidige onderzoek vinden we geen verschil in succespercentage, wat mogelijk is toe te schrijven aan het geringere aantal taken dat de proefpersonen moesten verrichten. Tot zover bevestigt het onderzoek de resultaten van de vorige studie. Maar uit de vergelijking met de gemeentelijke website blijkt in de huidige studie dat de reactiviteit van de synchrone hardopdenkmethode contextafhankelijk is. De opdracht om hardop te denken had op de website niet dezelfde uitwerking als in de bibliotheekcatalogus. De synchroon hardopdenkende proefpersonen presteerden weliswaar niet significant beter dan de retrospectief hardopdenkende proefpersonen, maar het is op basis van onze resultaten wel aannemelijk dat er testobjecten en opdrachten zullen zijn waarbij het hardopdenken faciliterend werkt voor de taakuitvoering.

Deze constatering vraagt om vervolgonderzoek dat moet leiden tot een raamwerk waarin een taaktypologie wordt verbonden aan assumpties over de reactiviteit van het hardopdenken. Daarvoor is een taakanalyse nodig waarin het gebruik van websites, interfaces en instructieve teksten zoals handleidingen of formulieren wordt uitgesplitst in eenduidige deeltaken. In zorgvuldig afgebakende experimenten kan vervolgens worden nagegaan welke invloed het hardopdenken heeft op de taakuitvoering. Dergelijk onderzoek zou, binnen het domein van de formatieve evaluatie maar mogelijk ook daarbuiten, substantieel bijdragen aan de theorievorming over de waarde en beperkingen van het hardopdenken als onderzoeksmethode.

Een andere bevinding is dat de synchrone hardopdenkmethode leidt tot minder geverbaliseerde problemen dan de retrospectieve hardopdenkmethode. Dit leidt tot vragen over de aard van synchrone en retrospectieve hardopdenkprotocollen. Een voor de hand liggende verklaring voor eventuele verschillen zou zijn dat de proefpersonen in de retrospectieve conditie meer los komen van de taakuitvoering. Toch lijkt het recente onderzoek van Guan e.a. (2006) dit niet te bevestigen: de retrospectieve verbalisaties betroffen passages waar de proefpersonen tijdens de taakuitvoering daadwerkelijk aandacht aan besteedden, zoals bleek uit hun eye-tracking data. Met andere woorden: ook in retrospectief hardopdenkonderzoek blijken proefpersonen behoorlijk getrouw hun eigen taakuitvoering te volgen. Gedetailleerd tekstanalytisch onderzoek naar de verschillen tussen verbalisaties in synchrone en retrospectieve hardopdenkprotocollen zou in dit opzicht erg waardevol zijn.

De bovengenoemde vragen hebben primair een theoretische inslag. Maar er zijn ook nog allerlei praktische vragen die evenzeer van belang lijken voor de validiteit van de methode. De discussie over proefleidergedrag die door Boren & Ramey (2000) is aangezwengeld, zal op basis van empirisch onderzoek tot een ontknoping moeten komen. Aan welke richtlijnen moet een proefleider zich precies houden in de "speech communication"-benadering en wat zijn de effecten van zulk proefleidergedrag? Ook de instructie die aan hardopdenkende proefpersonen wordt gegeven en de oefening of demonstratie die de proefpersonen aan het begin van een sessie krijgen voorgeschoteld, zijn uiterst relevant. Het verbaliseren van gedachten is niet iets wat alle proefpersonen intuïtief kunnen. De systematische ontwikkeling van een onderzoekscontext die proefpersonen optimaal voorbereidt op de hardopdenktaak lijkt praktisch gesproken de meest dringende kwestie waar onderzoek naar gedaan moet worden.

Dat de talrijke vragen met betrekking tot (synchrone en retrospectieve) hardopdenkprotocollen vooralsnog onbeantwoord zijn, is wellicht te wijten aan de robuustheid van de methode en aan het onmiddellijk herkenbare nut van de resultaten die deze oplevert. In de drie studies die we tot nu toe hebben verricht, waren er bijvoorbeeld geen systematische verschillen te ontdekken in de lijsten met problemen die op basis van synchrone en retrospectieve hardopdenkprotocollen werden gevonden. En de meeste problemen die we vonden leken zonder meer de moeite waard (wat ook bleek uit een analyse waarin experts de ernst en aannemelijkheid van de problemen moesten beoordelen). Deze beide kenmerken hebben mogelijk bijgedragen tot de mainstream kennis over usability testing die zich al een decennium lang nauwelijks ontwikkelt. In adviesliteratuur over formatief evaluatieonderzoek en op congressen wordt al jarenlang steeds dezelfde informatie over hardopdenkonderzoek herhaald. Om op dit gebied een stap verder te komen zal er meer gedetailleerd en systematisch onderzoek gedaan moeten worden naar de werking van de hardopdenkmethode.

Een belangrijke implicatie van het onderzoek dat we hier beschreven, is dat we voorzichtig moeten zijn om methodologische inzichten die met het ene testobject zijn verkregen, te vertalen naar andere onderzoeksobjecten. Het onderzoek naar interfaces en software vertoont grote verschillen met de praktijk van tekst- of website-evaluatie. Inzichten uit die hoek moeten derhalve met de nodige voorzichtigheid worden geïnterpreteerd. Denk bij deze inzichten niet alleen aan de werking van hardopdenkprotocollen, maar mogelijk ook aan heuristische evaluatie, cognitive walkthroughs en scenario-evaluatie (Nielsen & Mack, 1994) en aan de optimistische inschattingen over aantallen proefpersonen die nodig zouden zijn voor een betrouwbaar usability test resultaat (Nielsen, 1994; Virzi, 1992). Een bijkomende overweging is dat veel onderzoek uit deze hoek methodologisch niet sterk is (Gray & Salzman, 1998) en dat de verslaglegging ervan vaak zo bondig is dat er veel te raden overblijft over de precieze opzet van het onderzoek. We kunnen dergelijke studies beter als inspiratie zien voor eigen onderzoek naar tekst- en website-evaluatie dan als volwaardige en toepasbare bron van kennis.

Bibliografie

Boren, M.T., & Ramey J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication, 43*, 261-278.

Dieli, M. (1986). *Designing successful documents: An investigation of document evaluation methods*. Diss. Carnegie-Mellon University. Pittsburgh.

Dumas, J.S., & Redish, J.C. (1993). *A practical guide to usability testing*. Norwood, NJ: Ablex.

Ericsson, K.A. (1988). Concurrent verbal reports on text comprehension: A review. *Text, 8*, 295-325.

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis. Verbal reports as data*. Cambridge, MS: MIT Press.

Flower, L., Hayes, J.R., & Swarts, H. (1983). Revising functional documents: the scenario principle. In: P.V. Anderson, R.J. Brockmann & C.R. Miller (Eds.), *New essays in technical and scientific communication. Research, theory and practice* (pp. 41-58). Farmingdale, NY: Baywood.

Gray, W.D., & Salzman, M.C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13*, 203-261.

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In: *Proceedings of the SIGCHI conference on Human Factors in computing systems, Montréal, Québec, Canada* (pp. 1253-1262). New York: ACM Press.

Haak, M. van den, Jong, M. de, & Schellens, P.J. (2003). Hardopdenkprotocollen als pretestmethode: Synchroon en retrospectief hardopdenken vergeleken. *Tijdschrift voor Taalbeheersing, 25*, 236-252.

Haak, M.J. van den, Jong, M.D.T. de, & Schellens, P.J. (2004). Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. *Interacting with Computers, 16*, 1153-1170.

Haak, M.J. van den, Jong, M.D.T. de, & Schellens, P.J. (2007). Evaluation of a municipal Web site: Three variants of the think-aloud method compared. *Technical Communication, 54* (te verschijnen).

Jansen, C.J.M., & Steehouder, M.F. (1989). *Taalverkeersproblemen tussen overheid en burger. Een onderzoek naar verbeteringsmogelijkheden van voorlichtingsteksten en formulieren*. Diss. Rijksuniversiteit Utrecht. 's-Gravenhage: SDU.

Janssen D., Waes, L. van, & Bergh, H. van den (1996). Effects of thinking aloud on writing processes. In: C.M. Levy (ed.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 233-250). Mahwah, N.J.: Lawrence Erlbaum.

Jong, M. de (1998). *Reader feedback in text design. Validity of the plus-minus method for the pretesting of public information brochures*. Dissertation University of Twente. Amsterdam/Atlanta, GA: Rodopi.

Jong, M. de, & Schellens, P.J. (1995). *Met het oog op de lezer. Pretestmethoden voor schriftelijk voorlichtingsmateriaal*. Amsterdam: Thesis.

Jong, M. de, & Schellens, P.J. (2000). Toward a document evaluation methodology. What does research tell us about the validity and reliability of evaluation methods? *IEEE Transactions on Professional Communication, 43*, 242-260.

Knoblich, G., & Rhenius, D. (1995). Zur Reaktivität Lauten Denkens beim komplexen Problemlösen [Reactivity of thinking aloud during complex problem solving]. *Zeitschrift für experimentelle und angewandte Psychologie*, *XLII*, 419-454.

Kucan, L., & Beck, I.L. (1997). Thinking aloud and reading comprehension research: Inquiry, instruction, and social interaction. *Review of Educational Research, 67*, 271-299.

Lentz, L., & Pander Maat, H. (2003). Waarom het lezersprotocol zo'n goede methode is om begripsproblemen op te sporen. *Tijdschrift voor Taalbeheersing, 25*, 202-220.

Loxterman, J.A., Beck, I.L., & McKeown, M.G. (1994). The effects of thinking aloud during reading on students' comprehension of more or less coherent text. *Reading Research Quarterly, 29*, 353-367.

Mumma, G.H., Draguns, J.G., & Seibel, R. (1993). Reactive affects of concurrent verbalization in person perception tasks. *European Journal of Social Psychology, 23*, 295-311.

Nielsen, J. (1993). *Usability engineering*. Boston, MA: Academic Press.

Nielsen, J. (1994). Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies, 41*, 385-397.

Nielsen, J., & Mack, R.L. (1994). *Usability inspection methods*. New York: John Wiley.

Schriver, K.A. (1987). *Teaching writers to anticipate the reader's needs. Empirically based instruction*. Diss. Carnegie-Mellon University. Pittsburgh.

Silvén, M., & Vauras, M. (1992). Improving reading through thinking aloud. *Learning and Instruction, 2*, 69-88.

Swaney, J.H., e.a. (1981). *Editing for comprehension: improving the process through reading protocols.* Technical report no. 14. Document Design Project, Carnegie-Mellon University. Pittsburgh, PA.

Virzi, R.A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors, 34*, 457-468.

Wright, P.C., & Monk, A.F. (1991). A cost-effective evaluation method for use by designers. *International Journal of Man-Machine Studies, 35*, 891-912.