Maaike van den Haak
Menno de Jong en
Peter Jan Schellens

Hardopdenkprotocollen als pretestmethode

Synchroon en retrospectief hardopdenken vergeleken

Samenvatting

Hardopdenkprotocollen zijn een veelgebruikte pretestmethode. Toch is de methodologische kennis over de methode nog beperkt. In dit artikel wordt een vergelijking beschreven van synchrone en retrospectieve hardopdenkprotocollen voor de evaluatie van een online bibliotheekcatalogus. De beide methoden zijn op drie aspecten vergeleken: probleemdetecties, taakuitvoering en proefpersoonervaringen. De methoden leveren vergelijkbare verzamelingen gebruikersproblemen op, maar deze komen anders tot stand. Bij retrospectief hardopdenkonderzoek zijn de verbalisaties van proefpersonen cruciaal. Bij synchroon hardopdenkonderzoek is het leeuwendeel van de problemen observeerbaar in het proces en wordt er minder geverbaliseerd. Proefpersonen gaan vaker in de fout en zijn minder succesvol in de taakuitvoering. Dit roept vragen op over de reactiviteit van synchroon hardopdenkonderzoek, met name bij complexe taken.

*1. Introductie**

Hardopdenkprotocollen zijn een veelgebruikte methode voor de formatieve evaluatie van software, interfaces, websites en instructieve documenten. De methode houdt in dat gebruikers uit de doelgroep een aantal taken met behulp van het te evalueren communicatiemiddel uitvoeren en daarbij voortdurend hun gedachten verbaliseren. De methode heeft face validity, omdat de data die ermee verkregen worden het feitelijke gebruik weerspiegelen, en dus verder gaan dan het oordeel van proefpersonen over de gebruikersvriendelijkheid. Hardopdenkonderzoek is ingebed in een lange en gerespecteerde onderzoekstraditie, gericht op de cognitieve processen van proefpersonen tijdens de uitvoering van een breed scala aan taken – zoals schaken, probleemoplossen, schrijven, lezen en besluitvorming – met de monografie van Ericsson & Simon (1993) als belangrijke mijlpaal. Wanneer hardopdenken wordt gebruikt als pretestmethode, gaat het echter niet primair om inzicht in de cognitieve processen, maar om zicht op de kwaliteit van communicatiemiddelen of artefacten. In de afgelopen jaren zijn er diverse handboeken verschenen met uitvoerige instruc-

ties voor het uitvoeren van een hardopdenk usability test (Nielsen, 1993; Rubin, 1994; Dumas & Redish, 1999; Barnum, 2002).

De adviezen die in deze handboeken worden gegeven, worden echter nog nauwelijks ondersteund door methodologisch onderzoek (zie voor een overzicht De Jong & Schellens, 2000, 2002). Verschillende onderzoeken, waaronder dat van Jansen & Steehouder (1989), maken aannemelijk dat de hardopdenkmethode in combinatie met andere evaluatietechnieken kan leiden tot effectievere communicatiemiddelen, maar de bijdrage van het hardopdenkonderzoek is daarin niet geïsoleerd en de stap van hardopdenkresultaten naar een tekstrevisie wordt op geen enkele manier verantwoord. Recent onderzoek van Boren & Ramey (2000) maakt bovendien duidelijk dat de richtlijnen die Ericsson & Simon (1993) geven voor hardopdenkonderzoek in de praktijk niet worden opgevolgd en, zo betogen de auteurs, gezien de doelstellingen van een usability test wellicht ook niet zo strikt gehanteerd hoeven te worden. Ook woedt er in de literatuur nog steeds een discussie over de volledigheid en de mogelijke reactiviteit van hardopdenkprotocollen bij verschillende soorten taken, waarbij de resultaten van empirische vergelijkingen tussen hardopdenkende en stilwerkende proefpersonen de uitgangspunten van Ericsson & Simon (1993) lang niet altijd ondersteunen (zie bijvoorbeeld Russo, Johnson & Stephens, 1989; Short e.a., 1991; Loxterman, Beck & McKeown, 1994; Janssen, Van Waes & Van den Bergh, 1996). Hardopdenkende proefpersonen voeren hun taak soms anders, beter of slechter uit dan proefpersonen die stil mogen werken. Er zijn, met andere woorden, veel meer onzekerheden rondom hardopdenkprotocollen dan in de adviesliteratuur over usability testing wordt gesuggereerd.

Het onderzoek waarover in dit artikel wordt gerapporteerd, maakt deel uit van een groter onderzoeksproject gericht op de waarde en beperkingen van een aantal varianten van hardopdenkprotocollen als pretestmethode. De Jong & Schellens (1995) maken in hun overzicht van methoden onderscheid tussen twee varianten van deze onderzoeksmethode: hardopleesonderzoek (zonder taken) en hardopwerkonderzoek (met taken). Wij richten ons op de laatste variant. We beschrijven een eerste experiment waarin synchrone en retrospectieve hardopdenkprotocollen worden vergeleken bij de evaluatie van een online bibliotheekcatalogus. Retrospectieve hardopdenkprotocollen, ook wel 'retrospective testing' (Nielsen, 1993) of 'aided subsequent verbal protocol' (Henderson e.a., 1995) genoemd, verschillen in één opzicht van de (gebruikelijke) synchrone hardopdenkprotocollen: de proefpersonen denken niet hardop tijdens de taakuitvoering, maar voeren de taken in eerste instantie stil uit, en verbaliseren hun gedachten pas achteraf, aan de hand van een videoopname van hun taakuitvoering.

Theoretisch zijn er zowel voor- als nadelen aan het gebruik van retrospectieve hardopdenkprotocollen in plaats van synchrone hardopdenkprotocollen. Een voordeel betreft de mogelijke afname van reactiviteit: omdat de proefpersonen hun taken volledig op de eigen wijze en in het eigen tempo kunnen uitvoeren, zullen hun werkwijze en hun succes bij de taakuitvoering een betere afspiegeling vormen van hun normale taakuitvoering. Bij synchrone hardopdenkprotocollen is eerder sprake van reactiviteit: proefpersonen zouden bijvoorbeeld beter kunnen presteren dan normaal als gevolg van een meer gestructureerd werkproces, of juist slechter als gevolg van een te grote werkbelasting (Russo, Johnson & Stephens, 1989). Een tweede voordeel betreft de mogelijkheid om ook werktijden te meten. In de literatuur over usability testing worden de werktijden per taak vaak gehanteerd als indicator voor de gebruikersvriendelijkheid van een interface of website. Bij retro-

spectieve hardopdenkprotocollen kan een tijdsmeting zonder problemen als extra bron van informatie worden meegenomen; bij synchrone hardopdenkprotocollen ligt dat minder voor de hand, omdat van hardopdenken algemeen wordt aangenomen dat het de taakuitvoering in een per proefpersoon wisselende mate vertraagt. Een derde voordeel is dat proefpersonen de mogelijkheid hebben om te reflecteren op hun ervaringen tijdens de taakuitvoering, wat hen ertoe zou kunnen brengen om verbanden te leggen tussen individuele problemen en te zoeken naar meer structurele oorzaken.

Naast voordelen heeft het gebruik van retrospectieve hardopdenkprotocollen ook een aantal mogelijke nadelen. Een nadeel is dat de proefpersoonsessies aanmerkelijk langer duren dan bij synchrone hardopdenkprotocollen: de tijd verdubbelt omdat de proefpersonen niet alleen hun taken uitvoeren, maar deze ook nog achteraf bekijken. Een tweede en belangrijker nadeel is dat er gemakkelijk vertekeningen kunnen optreden in de achteraf geverbaliseerde gedachten. Het is onwaarschijnlijk dat proefpersonen zich na afloop nog al hun gedachten tijdens de taakuitvoering kunnen herinneren. Ericsson & Simon (1993) stellen dat belangrijke informatie verloren gaat in retrospectief onderzoek, wat bevestigd wordt door verschillende andere studies (e.g., Russo, Johnson & Stephens, 1989; Teague, De Jesus & Nunes-Ueno, 2001). Veel hangt echter af van de stimuli die de proefpersonen krijgen bij het verwoorden van hun gedachten. Bij retrospectief hardopdenken aan de hand van een video-opname hebben de proefpersonen meer aanknopingspunten dan bij ongeholpen retrospectieve hardopdenkprotocollen. Vertekeningen kunnen ook ontstaan wanneer proefpersonen besluiten bepaalde gedachten te verbergen, gedachten te verzinnen of te wijzigen, vanwege zelfpresentatie of sociale wenselijkheid. Hoewel dergelijke vertekeningen ook kunnen optreden bij synchrone hardopdenkprotocollen, hebben retrospectief hardopdenkende proefpersonen meer gelegenheid om hun verbalisaties te redigeren. Toch zijn ook zij voortdurend gebonden aan de gebeurtenissen zoals die op de videoband zijn vastgelegd, wat de bandbreedte om gedachten achteraf aan te passen beperkt maakt.

In de literatuur over usability testing worden synchrone en retrospectieve hardopdenkprotocollen vaak beschreven als vergelijkbare alternatieven (zie bijvoorbeeld Nielsen, 1993). Toch zijn er nog nauwelijks harde empirische gegevens beschikbaar over de twee hardopdenkvarianten. Verschillende onderzoekers hebben naar eigen zeggen een vergelijking tussen synchrone en retrospectieve hardopdenkprotocollen gemaakt, maar vergelijken synchrone hardopdenkprotocollen in feite met zuiver retrospectief onderzoek, waarbij de proefpersonen geen stimuli hebben gekregen bij het achteraf verbaliseren van hun gedachten (Branch 2000; Kuusela & Paul, 2000; Taylor & Dionne, 2000).

Tot dusver hebben slechts twee studies daadwerkelijk een vergelijking gemaakt tussen retrospectieve en synchrone hardopdenkprotocollen. Hoc & Leplat (1983) hebben de twee varianten gebruikt om een cognitief proces te onderzoeken, waarbij proefpersonen een set van letters moesten ordenen op een computerscherm met behulp van een beperkt aantal commando's. In de retrospectieve conditie werd de proefpersonen eerst gevraagd een beschrijving van hun werkproces te geven zonder stimuli, waarna ze vervolgens hardop moesten denken terwijl ze een log file met de stappen in hun taakuitvoering bekeken. Hoc & Leplat concluderen dat retrospectief onderzoek zonder stimuli beter vermeden kan worden vanwege de verstoringen en gaten in de protocollen, maar dat de retrospectieve en synchrone hardopdenk protocollen vergelijkbare resultaten opleveren. Hierbij moet wel worden opgemerkt dat zowel de taak die de proefpersonen moesten uitvoeren (deze leek op een puzzel) en de analyse van de resultaten (die meer gericht was op strategie dan op de

gevonden problemen) niet overeenkomen met die van een usability test.

Bowers & Snyder (1990) hebben de twee hardopdenkvarianten vergeleken in een usability test gericht op het gebruiken van meerdere windows op een computerscherm. Zij vonden geen significante verschillen met betrekking tot taakuitvoering en de benodigde tijd hiervoor, maar concludeerden wel dat de retrospectieve hardopdenkconditie beduidend minder verbalisaties opleverde, en dat deze verbalisaties bovendien vaak verschilden van de synchrone verbalisaties: ze waren meer gericht op verklaringen en minder op procedures. Hoewel deze resultaten zeker interessant zijn, heeft de studie een belangrijk nadeel: er wordt geen aandacht geschonken aan de hoeveelheid en soorten problemen die de proefpersonen in beide condities naar voren brachten. Omdat probleemopsporing de belangrijkste functie van usability testing is, is een cruciaal aspect niet in de vergelijking van de methodes meegenomen.

In dit artikel presenteren we een onderzoek dat gericht is op de waarde van synchrone versus retrospectieve hardopdenkprotocollen door een vergelijking te maken tussen de twee varianten met het oog op usability testing. Drie onderzoeksvragen staan centraal:

- Zijn er verschillen tussen synchrone en retrospectieve hardopdenkprotocollen met betrekking tot het aantal en type gevonden problemen?
- Zijn er verschillen tussen synchrone en retrospectieve hardopdenkprotocollen met betrekking tot taakuitvoering?
- Zijn er verschillen tussen synchrone en retrospectieve hardopdenkprotocollen met betrekking tot de ervaringen van de proefpersonen tijdens de test?

2. Opzet van het onderzoek

Testobject. Als testobject voor deze studie is gekozen voor een online bibliotheekcatalogus. Een dergelijke catalogus combineert de kenmerken van een zoekmachine met die van een website: hij heeft een taakgericht karakter, vereist de nodige online navigatie, en is vaak complex van aard, met name voor beginnende gebruikers. Online catalogi zijn met andere woorden een dankbaar object voor usability tests. Dat blijkt ook uit de literatuur op het gebied van bibliotheek- en informatiewetenschappen, waarin evaluatieonderzoek een regelmatig terugkerend thema is (bijvoorbeeld: Campbell, 2001; Battleson, Booth & Weintrop, 2001; Norlin & Winters, 2002).

De catalogus die in dit onderzoek is gebruikt, is de online catalogus van de Vrije Universiteit (UBVU). Deze is een aantal jaar geleden geïntroduceerd, en is sindsdien niet veranderd. Figuur 1 toont de homepage van de catalogus: een simpele layout, met een zoekscherm in het midden en links negen keuzeknoppen. Deze knoppen betreffen de standaard zoekmogelijkheden die catalogi doorgaans bieden, zoals eenvoudig of uitgebreid zoeken, sorteren, of bladeren. Zoals bij de meeste catalogi het geval is, heeft de UBVU ook een helpfunctie, die hulp biedt bij het gebruiken van de catalogus.

Hoewel de UBVU-catalogus in eerste instantie bestemd is voor studenten en medewerkers van de Vrije Universiteit, kunnen ook externe bezoekers toegang krijgen tot de catalogus, uiteraard met uitzondering van interne onderdelen zoals 'lenen' of 'reserveren'. Alle informatie in de catalogus is zowel in het Nederlands als in het Engels beschikbaar, behalve de helpfunctie: deze is alleen in het Engels te raadplegen.

Figuur 1. *Homepage van de UBVU-bibliotheekcatalogus*

Proefpersonen. Aan het experiment namen 40 studenten deel van de opleiding Toegepaste Communicatiewetenschap aan de Universiteit Twente. Alle studenten zaten in het tweede of derde jaar van hun studie en hadden al enige ervaring met het gebruik van online bibliotheekcatalogi. Geen van allen was echter bekend met de UBVU-catalogus. Als zodanig waren ze geschikte proefpersonen: als student behoorden ze tot de doelgroep van de catalogus maar als UT-student hadden ze geen voorkennis over en ervaring met de UBVU-catalogus.

De deelnemers zijn geworven door middel van posters en emailberichten, en kregen een financiële vergoeding voor hun deelname. Er waren geen criteria met betrekking tot geslacht of leeftijd. Uiteindelijk namen 5 mannen en 35 vrouwen deel aan het experiment, in de leeftijd van 18 tot 24. Deze 40 deelnemers werden gelijkmatig over de twee condities verdeeld: er waren geen verschillen tussen de twee groepen met betrekking tot geslacht, leeftijd en ervaring met online bibliotheekcatalogi.

Taken. Om de UBVU-catalogus te evalueren met synchrone en retrospectieve hardopdenkprotocollen werden zeven taken geformuleerd, die samen de belangrijkste functies van de catalogus omvatten. De taken konden onafhankelijk van elkaar worden uitgevoerd. De complete set taken was als volgt:

1. Zoek hoeveel publicaties de UBVU-catalogus bevat over het onderwerp 'communicatie'.
2. Zoek hoeveel publicaties de UBVU-catalogus bevat over het onderwerp 'taal' of 'interactie'.

3. Zoek hoeveel publicaties de UBVU-catalogus bevat die geschreven zijn door de auteur A. Hannay.
4. Zoek van welke auteur de UBVU-catalogus de meeste publicaties bevat over het onderwerp 'popmuziek'.
5. Zoek hoeveel Nederlandstalige publicaties de UBVU-catalogus bevat over het onderwerp 'Shakespeare'.
6. Zoek hoeveel publicaties de UBVU-catalogus bevat over het onderwerp 'telecommunicatie' die gepubliceerd zijn vanaf 1999.
7. Zoek hoeveel publicaties de UBVU-catalogus bevat over het onderwerp 'web-' (d.w.z. website, webwinkel, webcommunicatie) binnen de context van het internet.

Taak 1 tot en met 4 hadden betrekking op de zoekfunctie (eenvoudig en uitgebreid) en de sorteerfunctie van de catalogus. Taak 5 en 6 waren gericht op het inperken van zoekresultaten (op taal en op jaar van publicatie). Taak 7 had betrekking op truncatie, een bibliotheekfunctie die vergelijkbaar is met de zogenaamde 'wild card' zoekoptie.

Vragenlijsten. Naast de zeven taken kregen de proefpersonen in beide condities ook twee vragenlijsten voorgelegd. De eerste vragenlijst werd overhandigd vóór aanvang van het experiment, en bevatte vragen over de achtergrondgegevens van de proefpersonen. Naast demografische gegevens (leeftijd, geslacht en opleiding) werd gevraagd of de proefpersonen al eerder met een online bibliotheekcatalogus hadden gewerkt en of ze een cursus op dit gebied hadden gevolgd. Verder werd nagegaan of ze kennis hadden van veel voorkomende catalogusfuncties, zoals zoeken, bladeren, etc.

De tweede vragenlijst werd na afloop van het experiment aan de proefpersonen voorgelegd en betrof de ervaringen van de proefpersonen tijdens het experiment. Drie onderwerpen stonden hierbij centraal: (1) hoe hebben de proefpersonen het synchroon of retrospectief hardopdenken ervaren, (2) hebben de proefpersonen naar eigen indruk anders gewerkt dan ze normaal zouden doen, en (3) hoe hebben de proefpersonen de aanwezigheid van de onderzoeker en de opnameapparatuur ervaren? De proefpersonen moesten hun ervaringen weergeven op vijfpuntsschalen met semantische differentialen. Daarnaast bood de vragenlijst ruimte voor eventuele opmerkingen.

Onderzoeksprocedure. Het onderzoek bestond uit 40 individuele sessies, die allemaal in dezelfde testruimte plaatsvonden. Tijdens elke sessie werden video-opnamen gemaakt van het computerscherm en werd de stem van de proefpersoon opgenomen. Daarnaast was de onderzoeker aanwezig in de ruimte om te observeren en aantekingen te maken. Gemiddeld werkten de proefpersonen 20 minuten aan de zeven taken.

De synchrone hardopdenksessies verliepen als volgt. Na binnenkomst vulde de proefpersoon de eerste vragenlijst (over achtergrondkenmerken) in. Vervolgens kreeg de proefpersoon instructies over de gang van zaken tijdens het onderzoek en werden de taken overhandigd. De instructies, die met het oog op uniformiteit van papier werden voorgelezen, luidden als volgt: 'Denk hardop terwijl je de taken uitvoert en doe gewoon alsof de onderzoeker niet aanwezig is. Je kunt haar niet om hulp vragen, maar ze zal je wel eraan herinneren hardop te blijven denken als je een tijdje stilvalt. Vergeet verder niet dat we niet jou, maar de catalogus testen. Als er iets misgaat, ligt dat dus aan de catalogus, en niet aan jou.'

Nadat de proefpersoon alle zeven taken had uitgevoerd, kreeg hij/zij de tweede vragenlijst (over de ervaringen tijdens het experiment).

De retrospectieve hardopdenksessies begonnen, net als de synchrone sessies, met de eerste vragenlijst en een instructie. De proefpersoon kreeg dezelfde zeven taken voorgelegd, maar moest in dit geval deze taken in stilte uitvoeren. Wederom was het niet geoorloofd om hulp van de onderzoeker te vragen. Na voltooiing van de taken kreeg de proefpersoon een video-opname van de taakuitvoering te zien, met de vraag om hierop tijdens het lopen van de band commentaar te geven. Daarbij kon de opname niet worden stilgezet. Tot slot moest de proefpersoon weer de tweede vragenlijst invullen.

Verwerking van de onderzoeksgegevens. Nadat de veertig sessies waren voltooid, werden de verbalisaties van de proefpersonen uitgeschreven en werd ook hun navigatie binnen de site genoteerd. Uit deze navigatie en de overige handelingen van de proefpersonen werden de problemen gedestilleerd die tijdens het gebruik van de catalogus optraden. Elke handeling die afweek van het optimale handelingsverloop (dat wil zeggen: het minste aantal vereiste handelingen) bij een taak werd als probleem gemarkeerd. Daarnaast werd in de transcripten van de verbalisaties gekeken naar verbale signalen die op problemen duidden, zoals uitingen van twijfel, onwetendheid of irritatie.

De totale set aan data werd als volgt geanalyseerd. Eerst werd vastgesteld hoeveel problemen er in beide condities naar voren kwamen. Daarna werd bij iedere probleemdetectie gekeken naar de wijze waarop deze aan het licht was gekomen: door observatie van de handelingen, door analyse van de verbalisaties, of door een combinatie van observatie en verbalisaties. Tot slot werden de problemen inhoudelijk gecategoriseerd. Omdat voor de combinatie van hardopdenkgegevens en bibliotheekcatalogi nog geen standaardlijst met probleemcategorieën bestaat, zijn op basis van een globale indeling van het zoekproces en een analyse van de ontdekte gebruikersproblemen, de volgende vijf probleemtypen gehanteerd:

Lay-outproblemen:	De proefpersoon kan bepaalde informatie op het scherm van de catalogus niet of moeilijk vinden.
Terminologieproblemen:	De proefpersoon begrijpt de terminologie in de catalogus niet.
Data-invoerproblemen:	De proefpersoon weet niet hoe hij/zij een zoekopdracht moet invoeren (bijv. het invoeren van een zoekterm of het gebruik van de dropdown menu's).
Volledigheidsproblemen:	In de catalogus ontbreekt informatie die nodig is voor het effectief gebruik ervan.
Feedbackproblemen:	De catalogus geeft geen relevante feedback op (verkeerd) uitgevoerde zoekopdrachten.

Naast deze vijf probleemtypen hadden de proefpersonen ook af en toe technische problemen, zoals een internetverbinding of browser die niet werkte. Deze problemen zijn niet in de analyse meegenomen.

Met betrekking tot de taakuitvoering van de proefpersonen is gekeken naar twee indicatoren: het met goed gevolg voltooien van de opdrachten en de tijd benodigd voor het voltooien van de opdrachten. Deze indicatoren zijn zowel per taak als voor de gehele taakuitvoering (alle zeven taken) bekeken.

3. Resultaten

Paragraaf 3.1 beschrijft het aantal en soort problemen dat in beide condities naar voren is gekomen. In paragraaf 3.2 worden de resultaten met betrekking tot de taakuitvoering beschreven. In paragraaf 3.3 wordt, ten slotte, ingegaan op de ervaringen van de proefpersonen tijdens het onderzoek, zoals gemeten in de tweede vragenlijst.

3.1 Aantal en soort gevonden problemen. De analyse van de veertig sessies leverde in totaal 72 verschillende problemen op. Sommige van deze problemen werden door bijna alle (30 tot 35) proefpersonen gevonden, maar meer dan de helft van het totaal aantal verschillende problemen werd gevonden door vijf of minder van de 40 proefpersonen. Dit geeft aan dat er een behoorlijk aantal individuele problemen was: problemen die weliswaar door sommige proefpersonen werden gevonden, maar die door de meeste proefpersonen niet als problematisch werden ervaren.

Tabel 1 geeft een overzicht van het gemiddeld aantal gevonden problemen per proefpersoon. Daarbij is meteen onderscheid gemaakt naar de manier waarop de problemen aan het licht zijn gekomen: door observatie, door verbalisaties, of door een combinatie van beiden. Uit de tabel blijkt dat er geen significant verschil was in het aantal ontdekte problemen per proefpersoon in beide condities. Maar er waren wel duidelijke verschillen in de manier waarop de problemen aan het licht kwamen.

Tabel 1. *Aantal problemen per proefpersoon met synchrone en retrospectieve hardopdenkprotocollen, ontdekt door observatie, verbalisaties of een combinatie van beide*

	Synchroon		Retrospectief		
	Gemiddeld	**SD**	**Gemiddeld**	**SD**	**Significantie**
Observatie	6.7	2.2	4.0	2.0	$p<0.001$
Verbalisaties	0.5	0.7	4.5	3.4	$p<0.001$
Observatie & verbalisaties	6.7	4.0	5.1	2.2	n.s.
Totaal	13.9	3.3	13.6	4.1	n.s.

De retrospectieve hardopdenkconditie leverde aanzienlijk meer geverbaliseerde problemen op (t-test, $t=5.168$, $df=38$, $p<0.001$, Cohen's $d=1.29$). De proefpersonen in deze conditie noemden gemiddeld 4,5 problemen die niet door observatie konden worden vastgesteld (tegenover slechts 0,5 geverbaliseerde problemen in de synchrone hardopdenkconditie). Dit verschil kan worden verklaard doordat proefpersonen in de retrospectieve conditie simpelweg meer tijd hadden om problemen te verbaliseren. Zij hoefden per slot van rekening pas na afloop van de taken commentaar te geven op hun werk, waardoor ze zich volledig konden richten op het evalueren van de catalogus, en naast taakgerelateerde problemen ook andere problemen konden bespreken, vaak geformuleerd als waardeoordelen. De proefper-

sonen in de synchrone hardopdenkconditie moesten tegelijkertijd werken en hardopdenken, en konden daarom minder aandacht geven aan het verbaliseren van problemen. Als gevolg daarvan noemden zij voornamelijk problemen die direct met hun taakuitvoering te maken hadden. Dit komt ook tot uiting in het aantal problemen dat door een combinatie van verbalisaties en observatie aan het licht kwam: 93% van alle verbaliseringen in de synchrone hardopdenkconditie kwam overeen met een observeerbaar probleem in de taakuitvoering; in de retrospectieve hardopdenkconditie was dit percentage maar 54%.

Een tweede significant verschil tussen beide condities betreft het aantal problemen dat door non-verbale indicatoren, d.w.z. puur door observatie, aan het licht is gekomen. Zoals tabel 1 aangeeft, bracht de synchrone hardopdenkconditie beduidend meer geobserveerde problemen aan het licht dan de retrospectieve hardopdenkconditie (6,7 versus 4,0; t-test, $t=4.083$, $df=38$, $p<0.001$, Cohen's $d=1.63$). Blijkbaar gingen proefpersonen in de synchrone hardopdenkconditie vaker in de fout tijdens de taakuitvoering dan proefpersonen in de retrospectieve conditie. Dit verschil kan worden verklaard door de verschillende werkbelasting in beide condities: terwijl de proefpersonen in de retrospectieve hardopdenkconditie zich uitsluitend met hun taken hoefden bezig te houden, moesten de proefpersonen in de synchrone conditie hierbij ook hun gedachten verbaliseren. Het is goed denkbaar dat deze extra belasting een negatieve uitwerking had op de taakuitvoering van de proefpersonen.

Om inzicht te krijgen in de typen problemen die in beide condities naar voren kwamen, werden alle probleemdetecties gecategoriseerd. Bij wijze van voorbeeld geeft tabel 2 een indruk van concrete gebruikersproblemen zoals die in de verschillende probleemcategorieën zijn ondergebracht.

Tabel 2. *Voorbeelden van probleemtypen uit de hardopdenkprotocollen*

Lay-out	De proefpersoon heeft moeite om de knop voor gevorderd zoeken te vinden op de homepage van de catalogus De proefpersoon kan de namen van co-auteurs niet vinden in de zoekresultaten van de catalogus
Terminologie	De proefpersoon begrijpt de betekenis van het woord 'limieten' niet. De proefpersoon begrijpt de betekenis van het woord 'truncatie' niet.
Data-invoer	De proefpersoon heeft moeite met het gebruik van booleaanse operatoren. De proefpersoon weet niet hoe data in het 'jaar'-venster moeten worden ingevoerd.
Volledigheid	De namen van auteurs ontbreken in de zoekresultaten. De helpfunctie geeft alleen informatie in het Engels, niet in het Nederlands.
Feedback	De catalogus geeft geen foutmelding als de proefpersoon een fout maakt. De catalogus geeft niet aan hoe de zoekresultaten geordend zijn (op jaar, op eerste auteur, etc.)

Tabel 3 geeft een overzicht van de typen problemen die in de beide condities zijn ontdekt. Er waren op dit punt geen significante verschillen: alle vijf probleemtypen werden in vergelijkbare aantallen in beide condities gevonden. Terminologie en data-invoer werden door de proefpersonen in beide condities als meest problematisch ervaren.

Tabel 3. *Aantal problemen uit verschillende categorieën per proefpersoon in de synchrone en retrospectieve hardopdenkconditie*

	Synchroon		Retrospectief		
	Gemiddeld	**SD**	**Gemiddeld**	**SD**	**Significantie**
Lay-out	2.9	1.2	2.6	1.3	n.s.
Terminologie	4.1	1.5	4.1	2.0	n.s.
Data-invoer	4.9	1.2	4.9	1.2	n.s.
Volledigheid	1.1	0.9	1.2	0.6	n.s.
Feedback	1.0	1.0	0.9	0.6	n.s.
Totaal	13.9	3.3	13.6	4.1	n.s.

De analyses tot nu toe waren gericht op de algemene trends in de resultaten, waarbij individuele problemen buiten beschouwing bleven. Een vergelijking van de lijsten met problemen die in beide condities naar voren kwamen, geeft een indruk van de mate van overlap tussen de synchrone en de retrospectieve hardopdenkconditie. Van de in totaal 72 verschillende problemen werd 47% gevonden in beide condities, 31% alleen in de synchrone hardopdenkconditie, en 22% alleen in de retrospectieve hardopdenkconditie. Meer overlap bestaat wanneer gekeken wordt naar de overlap in probleemdetecties. We kijken dan steeds per probleemdetectie of die ook in de andere conditie naar voren is gekomen. Tabel 4 laat zien dat 89% van alle probleemdetecties betrekking heeft op problemen die door proefpersonen in beide condities werden gevonden.

Tabel 4. *Percentage problemen die uniek zijn voor één van beide hardopdenkcondities*

	Uniek synchroon	**Uniek retrospectief**	**Ontdekt in beide**
Lay-out	10	12	78
Terminologie	1	6	93
Data-invoer	6	2	92
Volledigheid	11	4	84
Feedback	8	2	91
Totaal	6	5	89

Het algemene beeld dat ontstaat, is dat synchrone en retrospectieve hardopdenkprotocollen vergelijkbaar zijn wat betreft het aantal en de typen gevonden problemen. De twee methoden verschillen echter met betrekking tot de wijze waarop deze problemen naar voren komen: synchrone hardopdenkprotocollen brengen meer problemen aan het licht die tijdens de taakuitvoering te observeren zijn; retrospectieve hardopdenkprotocollen resulteren in meer problemen aan de hand van de verbalisaties van proefpersonen. De verbalisaties spelen een ondergeschikte rol in de synchrone hardopdenkprotocollen. Dit resultaat is opmerkelijk, omdat men er bij de hardopdenkmethode als usability test juist van uitgaat dat de verbale protocollen cruciaal zijn voor het ontdekken van problemen. Uit het huidige experiment blijkt echter dat de verbale protocollen niet zozeer nieuwe problemen aan het licht brengen, maar voornamelijk dienen ter ondersteuning van de problemen die ook waarneembaar zijn. Het feit dat observeerbare problemen significant meer voorkomen in de synchrone hardopdenkconditie kan, zoals eerder beschreven, worden verklaard door de zwaardere werkbelasting van de proefpersonen. Dat is een eerste aanwijzing voor de reactiviteit van de (synchrone) hardopdenkmethode in dit experiment. Om deze reden is het interessant om te kijken of de dubbele werkbelasting ook invloed heeft gehad op de taakuitvoering van de proefpersonen.

3.2 Taakuitvoering. De taakuitvoering van de proefpersonen in beide condities is bekeken aan de hand van twee indicatoren: de succesvolle voltooiing van de zeven opdrachten en de tijd die nodig was om deze opdrachten te voltooien. Tabel 5 geeft een overzicht van de resultaten van beide indicatoren. Met betrekking tot de benodigde tijd zijn er geen significante verschillen gevonden, noch per taak noch voor de gehele set van zeven taken. Blijkbaar had het hardop denken in de synchrone conditie geen vertragend effect op de taakuitvoering van de proefpersonen. De dubbele werkbelasting in de synchrone hardopdenkconditie had echter wel invloed op de mate van succes bij het voltooien van de taken. De proefpersonen in deze conditie waren significant minder succesvol in het correct voltooien van de complete takenset dan de proefpersonen in de retrospectieve conditie (t-test, $t=2.252$, $df=38$, $p<0.05$, Cohen's $d=0.71$). Dit resultaat correspondeert met de eerdere bevinding dat de synchrone hardopdenkprotocollen ook meer waarneembare problemen bevatten dan de retrospectieve protocollen. Dit bevestigt het eerdere vermoeden van reactiviteit van de synchrone hardopdenkprotocollen. Hierbij moet worden opgemerkt dat de proefpersonen in het algemeen moeite hadden met het uitvoeren van de opdrachten: gemiddeld genomen werd slechts 40% van de taken met succes voltooid. De lastigste taak (taak 7) werd door slechts één van de 40 proefpersonen met succes volbracht; de eenvoudigste taak (taak 4) door 38 van de 40 proefpersonen. In de discussieparagraaf zullen we hier nader op ingaan.

Tabel 5. *Taakuitvoering in de synchrone en retrospectieve hardopdenkcondities*

	Synchroon		Retrospectief		
	Gemiddeld	**SD**	**Gemiddeld**	**SD**	**Significantie**
Aantal taken dat succesvol is afgerond	2.6	1.0	3.3	1.0	$p<0.05$
Totale tijd benodigd voor de zeven taken	21.1	5.7	19.6	5.0	n.s.

3.3 Ervaringen van de proefpersonen. De vragenlijst met betrekking tot de ervaringen van de proefpersonen tijdens het onderzoek richtte zich op drie aspecten: de ervaringen met synchroon of retrospectief hardopdenken, de werkwijze bij de taakuitvoering en de aanwezigheid van onderzoeker en opnameapparatuur.

Over hun ervaringen met (synchroon of restrospectief) hardopdenken moesten de proefpersonen op vijfpuntsschalen aangeven of zij deze activiteit moeilijk, onaangenaam, vermoeiend, onnatuurlijk of tijdrovend vonden. Omdat deze vragen samen geen betrouwbare schaal vormden, werd elke vraag individueel geanalyseerd. Deze individuele analyses (zie tabel 6) laten zien dat er geen significante verschillen waren in de oordelen van de proefpersonen over het hardopdenken (t-test). Over het algemeen oordeelden de proefpersonen redelijk neutraal, met scores rond het midden van de vijfpuntsschaal. Voor de synchrone hardopdenkconditie betekent dit dat de reactiviteit van de methode, zoals die in de vorige paragrafen naar voren kwam, niet als zodanig door de proefpersonen zelf werd ervaren.

Tabel 6. *Proefpersoonervaringen met betrekking tot het hardopdenken*

	Synchroon		Retrospectief		
	Gemiddeld	**SD**	**Gemiddeld**	**SD**	**Significantie**
Moeilijk – gemakkelijk	2.4	0.8	2.7	1.2	n.s.
Onprettig – prettig	2.7	0.8	2.9	1.0	n.s.
Wel – niet vermoeiend	3.4	1.0	3.8	1.4	n.s.
Onnatuurlijk – natuurlijk	3.4	0.9	3.0	1.5	n.s.
Wel – niet tijdrovend	3.2	1.2	3.2	1.1	n.s.

NB: Scores op een vijfpuntsschaal (1 = negatief, 5 = positief)

De proefpersonen werd vervolgens gevraagd om in te schatten in hoeverre hun werkwijze verschilde van een normale werkwijze: sneller of langzamer, gerichter of minder gericht, etc. De resultaten, die in tabel 7 zijn weergegeven, bevatten wederom geen significante verschillen tussen de synchrone en de retrospectieve hardopdenkconditie. In beide condities waren de proefpersonen van mening dat hun werkwijze slechts weinig verschilde van hun normale manier van werken. Na een hercodering van de antwoorden om elke afwijking van de normale werkwijze (naar beide kanten van de schaal) vast te stellen (waarbij de middelste score als 0 werd gecodeerd en de extremen als 2), bleek dat de acht variabelen een betrouwbare schaal vormden (Cronbach's alpha = 0.84). Op deze schaal bleken de proefpersonen in de retrospectieve hardopdenkconditie, naar hun eigen oordeel, significant meer afwijkend te hebben gewerkt tijdens dan de proefpersonen in de synchrone hardopdenkconditie (met een gemiddelde afwijking van 0.33 versus 0.29; t-test, $t=2.242$, $df=38$, $p<.0.05$, Cohen's $d=0.72$).

Dit betekent dat, in tegenstelling tot de eerdere bevindingen met betrekking tot probleemdetecties en taakuitvoering, de proefpersonen in de retrospectieve hardopdenkconditie meer reactiviteit ervoeren dan de proefpersonen in de synchrone conditie. Dit kan echter te maken hebben met het tijdstip waarop de proefpersonen in de retrospectieve hardopdenkconditie de vragenlijst invulden. Zij deden dit na afloop van het bekijken en becommentariëren van hun video-opname, en het is goed voorstelbaar dat de per definitie onnatuurlijke taak van het achteraf commentaar geven in hun beoordeling van de taakuitvoering is mee genomen.

Tabel 7. *Proefpersoonervaringen met betrekking tot hun werkwijze in de test, vergeleken met hun normale werkwijze*

	Synchroon		Retrospectief		
	Gemiddeld	**SD**	**Gemiddeld**	**SD**	**Significantie**
Sneller – langzamer	2.7	0.7	2.3	0.8	n.s.
Meer – minder gericht	2.6	0.6	2.1	0.9	n.s.
Meer – minder geconcentreerd	3.3	0.6	3.5	0.9	n.s.
Meer – minder vasthoudend	2.6	0.9	2.7	0.9	n.s.
Meer – minder succesvol	3.0	0.5	2.9	0.7	n.s.
Meer – minder prettig	3.2	0.5	3.4	0.6	n.s.
Meer – minder oog voor fouten	2.6	0.7	2.2	0.7	n.s.
Meer onstpannen – meer gespannen	3.4	0.6	3.7	0.5	n.s.

NB: Scores op een vijfpuntsschaal (3 = geen verschil met de normale werkwijze)

Het laatste deel van de vragenlijst betrof de aanwezigheid van de onderzoeker en het gebruik van opnameapparatuur. De proefpersonen werd eerst gevraagd aan te geven in hoeverre ze het onplezierig, onnatuurlijk of storend hadden gevonden dat de onderzoeker tijdens het experiment aanwezig was. Vervolgens werden dezelfde vragen nog een keer gesteld, maar dan met betrekking tot het gebruik van de opnameapparatuur. Voor alledrie de aspecten van de testsituatie kon een voldoende betrouwbare schaal (op basis van twee vragen) worden gevormd (Cronbach's alpha = 0.66 voor 'onplezierig', 0.81 voor 'onnatuurlijk' en 0.62 voor 'storend'). De resultaten staan in tabel 8. De scores met betrekking tot '(on)plezierig' en '(on)natuurlijk' waren noch negatief noch positief, en verschilden niet significant tussen de twee condities. De scores met betrekking tot '(niet) storend' waren tamelijk positief in beide condities, maar de proefpersonen in de synchrone hardopdenkconditie vonden de testsituatie nog minder storend dan de proefpersonen in de retrospectieve hardopdenkconditie (t-test, t=2.368, df=33.4, p<.05, Cohen's d=0.75).

Dit laatste verschil kan wederom worden verklaard door het tijdstip waarop de proefpersonen in de retrospectieve hardopdenkconditie de vragenlijst invulden. Een tweede verklaring ligt in het feit dat de aanwezigheid van de onderzoeker in het eerste gedeelte van de retrospectieve conditie minder functioneel was dan tijdens de synchrone conditie. Ook zou het achteraf bekijken van hun eigen proces confronterend kunnen zijn voor de proefpersonen. Tot slot zou ook de werkdruk van de proefpersonen een verklaring kunnen zijn. De proefpersonen in de synchrone hardopdenkconditie moesten gelijktijdig taken uitvoeren en hardopdenken, waardoor ze minder aandacht konden besteden aan de onderzoeker en de opnameapparatuur.

Tabel 8. *Proefpersoonervaringen met betrekking tot de testsituatie: de aanwezigheid van de onderzoeker en opnameapparatuur*

	Synchroon		Retrospectief		
	Gemiddeld	**SD**	**Gemiddeld**	**SD**	**Significantie**
Onplezierig	2.8	0.3	2.7	0.8	n.s.
Onnatuurlijk	2.9	0.7	3.1	1.3	n.s.
Storend	4.3	0.6	3.7	0.9	p<.05.

NB: Scores op een vijfpuntsschaal (1 = negatief, 5 = positief)

Al met al ondersteunen de oordelen van de proefpersonen de validiteit van zowel de synchrone als de retrospectieve hardopdenkprotocollen. De vragen resulteerden in overwegend positieve oordelen voor beide evaluatiemethoden. Uit de vragenlijst kwam echter een aantal verschillen naar voren tussen beide condities die slecht corresponderen met de gegevens betreffende de probleemdetecties en de taakuitvoering uit paragraaf 3.1 en 3.2. De proefpersonen in de retrospectieve hardopdenkconditie ondervonden meer reactiviteit als gevolg van de testsituatie, en vonden deze situatie storender dan de proefpersonen in de synchrone hardopdenkconditie. Dit kan op een reëel verschil tussen beide methoden duiden, maar het is ook aannemelijk dat dit verschil is veroorzaakt door een achteraf minder gelukkige keuze in de onderzoeksprocedure (waarbij de proefpersonen in de retrospectieve hardopdenkconditie de vragenlijst niet direct na de taakuitvoering, maar pas na het becommentariëren van de video-opname invulden).

4. Discussie

Het hier gerapporteerde onderzoek toont aan dat er zowel overeenkomsten als verschillen zijn tussen synchrone en retrospectieve hardopdenkprotocollen. De verschillen die tussen beide condities gevonden zijn, geven een nieuwe kijk op de validiteit van hardopdenkprotocollen voor usability testing. Hoewel beide methoden vergelijkbaar waren wat betreft de aantallen en de typen probleemdetecties, verschilden ze significant in de wijze waarop deze probleemdetecties tot stand kwamen. De synchrone hardopdenkconditie leverde significant meer problemen op die puur op basis van observatie naar voren kwamen. De retrospectieve hardopdenkconditie resulteerde in significant meer problemen die niet te observeren waren, maar alleen door de verbalisaties van proefpersonen aan het licht kwamen. Deze resultaten maken duidelijk dat synchroon hardopdenkonderzoek een meer getrouwe representant is van een strikt taakgerelateerde usability test, terwijl retrospectief hardopdenkonderzoek een breder scala aan gebruikersreacties lijkt op te leveren. Dit komt overeen met de bevindingen uit het onderzoek van Bowers & Snyder (1990), waarbij proefpersonen in de retrospectieve hardopdenkconditie allerlei verklaringen en suggesties verwoordden, terwijl de proefpersonen in de synchrone hardopdenkconditie zich vaak beperkten tot het beschrijven van hun handelingen. Om de waarde van de feedback van beide methodes te bepalen is verder onderzoek nodig naar de predictieve validiteit van synchrone en retrospectieve hardopdenkprotocollen: Hoe belangrijk zijn de gevonden problemen? Zijn er veel false alarms, met name in de geobserveerde problemen bij synchroon hardopdenkende proefpersonen en in de geverbaliseerde problemen in retrospectieve hardopdenkprotocollen?

Met betrekking tot het gebruik van synchrone hardopdenkprotocollen brengen de resultaten van dit onderzoek twee belangrijke zaken aan het licht. De eerste is de bijzonder beperkte bijdrage van de proefpersoonverbalisaties aan de opbrengst (in aantal gevonden gebruikersproblemen) van de usability test. De verbalisaties van proefpersonen resulteerden nauwelijks in de detectie van nieuwe problemen, en dienden voornamelijk ter ondersteuning of verklaring van problemen die ook te observeren waren in de proefpersoonhandelingen. Dit kan op zich ook een belangrijke bijdrage zijn, zeker voor de diagnose en het op waarde schatten van de gevonden problemen. Maar toch moet worden vastgesteld dat de verbalisaties in de synchrone hardopdenkconditie een minder belangrijk onderdeel van de usability test waren dan doorgaans in handboeken over usability testing wordt gesuggereerd.

Een tweede en nog belangrijker observatie is dat de synchrone hardopdenkprotocollen reactiviteit veroorzaakten in de usability test. Dit komt overeen met eerdere bevindingen van Russo, Johnson & Stephens (1989), die de validiteit van hardopdenkprotocollen bestudeerden voor verschillende cognitieve taken en concludeerden dat het hardopdenken de taakuitvoering zowel kon hinderen als bevorderen. Deze observatie weerspreekt de resultaten van Bowers & Snyder (1990), die geen verschil vonden in de taakuitvoering van synchroon en retrospectief hardopdenkende proefpersonen.

In het hier gerapporteerde onderzoek had het hardop denken een consistent en plausibel negatief effect op de taakuitvoering. De extra taak om tijdens de taakuitvoering gedachten te verbaliseren zorgde ervoor dat de proefpersonen meer fouten maakten en minder succesvol waren in het verrichten van de zeven taken. Aan de hand van dit resultaat lijkt de twijfel gerechtvaardigd of de taakuitkomst in een synchrone hardopdenktest een correcte

indicatie geeft van de gebruikersvriendelijkheid van een communicatiemiddel, en of de problemen die in hardopdenkonderzoek worden gevonden per definitie ook echte gebruikersproblemen zijn. Onderzoek naar de predictieve validiteit, zoals omschreven door De Jong & Schellens (2000, 2002), van hardopdenkresultaten is dus geen overbodige poging om vast te stellen wat feitelijk al bekend is, maar een belangrijke stap in verder onderzoek naar de reactiviteit van de methode. Er bestaat ten slotte een reële mogelijkheid dat een probleem gevonden in een synchrone hardopdenktest (gedeeltelijk) is veroorzaakt door de gehanteerde methode zelf. Vergeleken met stilwerkende proefpersonen levert de hardopdenkmethode weliswaar meer probleemdetecties op, maar deze zijn grotendeels toe te schrijven aan extra problemen in de taakuitvoering, en niet zozeer aan de verbalisaties zelf.

Of deze constatering al dan niet schadelijk is, staat overigens nog ter discussie. De meeste usability tests hebben tot doel gebruikersproblemen te identificeren en beoordelen, en het is vol te houden dat het een voordeel van synchroon hardopdenken is dat dergelijke problemen kennelijk gemakkelijker aan het licht komen. In die interpretatie is de drempel om taken succesvol te verrichten alleen wat hoger. Natuurlijk is zo'n positieve draai aan de resultaten alleen te geven als uit onderzoek blijkt dat deze extra geobserveerde problemen in de praktijk corresponderen met reële gebruikersproblemen.

De meest plausibele verklaring voor de reactiviteit van synchrone hardopdenkprotocollen ligt in de werkbelasting van de proefpersonen: de moeilijkheidsgraad van de taken kan een cruciale factor in dit onderzoek zijn geweest. De gegevens met betrekking tot de taakuitvoering maken duidelijk dat de zeven taken die de proefpersonen voorgelegd kregen, erg moeilijk waren. De cognitieve belasting van de taken in combinatie met de extra belasting van het hardopdenken lijkt een negatief effect te hebben gehad op zowel de verbalisaties als de taakuitvoering van proefpersonen. De gaten in de verbalisaties worden onderschreven door Ericsson & Simon (1993, p.91), die stellen dat proefpersonen mogelijk stoppen met verbaliseren wanneer ze te zwaar cognitief belast zijn. Het negatieve effect op de taakuitvoering wordt echter niet eenduidig verklaard door de bestaande literatuur (Russo, Johnson & Stephens, 1989; Ericsson & Simon, 1993). In tegendeel, er zijn ook studies die aantonen dat het synchroon hardopdenken juist een positief effect op taakuitvoering heeft (Loxterman, Beck & McKeown 1994). Het zou dan ook interessant zijn om de relatie tussen taakcomplexiteit, verbalisatie en taakuitvoering van proefpersonen in een synchrone hardopdenktest nader te onderzoeken.

Een laatste opmerking betreft de generaliseerbaarheid van het huidige onderzoek. Het gaat hier om een eerste vergelijkende studie, waarbij slechts één object was betrokken. Een belangrijk kenmerk van de UBVU-catalogus en de taken gebruikt in dit onderzoek is dat er in de wijze waarop de proefpersonen met de computer werkten veel te observeren viel. Het zou interessant zijn om te kijken of vergelijkbare resultaten naar voren komen bij applicaties met een minder duidelijk waarneembaar gebruikersproces. Een replicatie van dit onderzoek met documentatie, websites of interfaces met een meer open taakdomein zou een nuttige follow-up zijn om synchrone en retrospectieve hardopdenkprotocollen verder te onderzoeken.

Al met al duiden de resultaten van dit onderzoek erop dat synchrone en retrospectieve hardopdenkprotocollen kunnen worden beschouwd als gelijkwaardige maar duidelijk verschillende evaluatiemethoden. Een sterk, en nieuw argument dat voor het gebruik van retrospectieve hardopdenkprotocollen pleit, is dat ze minder gevoelig voor de invloed van taakcomplexiteit zouden kunnen zijn, zowel met het oog op reactiviteit als met betrekking

tot de volledigheid van de verbalisaties. In richtlijnen voor hardopdenkonderzoek wordt vaak gesteld dat de onderzoeker taken dient te formuleren met een gemiddelde moeilijkheidsgraad, zodat de deelnemers noch in een automatisch werkproces vervallen noch belast worden met een te zware cognitieve belasting. Bij usability testing zijn deze richtlijnen echter niet altijd haalbaar. Tenslotte liggen de kwaliteit van het testobject en de selectie van realistische taken doorgaans niet in de handen van het usability test team.

Noten

***** Dit artikel is een Nederlandse bewerking van Van den Haak, De Jong & Schellens (2003).

Bibliografie

Barnum, C.M. (2002). *Usability testing and research*. New York: Longman.

Battleson, B., A. Booth & J. Weintrop (2001). Usability testing of an academic library web site: a case study. *Journal of Academic Librarianship, 237*, 188-198.

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

Bowers, V.A. & H.L. Snyder (1990). Concurrent versus retrospective verbal protocols for comparing window usability. *Proceedings of the Human Factors Society 34th Meeting, 8-12 October 1990* (pp.1270-1274). Santa Monica: HFES.

Branch, J.L. (2000). Investigating the information-seeking processes of adolescents: The value of using think alouds and think afters. *Library & Information Science Research, 22*, 371-392.

Campbell, N. (ed.) (2001). *Assessment of library-related Web sites: Methods and case studies*. Chicago: LITA.

Dumas, J.S. & J.C. Redish (1999). *A practical guide to usability testing*. Revised edition. Exeter: Intellect.

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

Haak, M.J. van, M.D.T. de Jong & P.J. Schellens (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an onlike library catalogue. *Behaviour & Information Technology, 22,* 339-351.

Henderson, R.D., M.C. Smith, J. Podd & H. Varela-Alvarez (1995). A comparison of the four prominent user-based methods for evaluating the usability of computer software. *Ergonomics, 38*, 2030-2044.

Hoc, J.M. & J Leplat (1983). Evaluation of different modalities of verbalization in a sorting task. *International Journal of Man-Machine Studies, 18*, 283-306.

Jansen, C.J.M. & M.F. Steehouder (1989). *Taalverkeersproblemen tussen overheid en burger. Een onderzoek naar verbeteringsmogelijkheden van voorlichtingsteksten en formulieren*. Dissertatie Universiteit Twente, Enschede. 's-Gravenhage: Sdu.

Janssen, D., L. van Waes & H. van den Bergh (1996). Effects of thinking aloud on writing processes. In C.M. Levy & S. Randell (eds.), *The science of writing. Theories, models, individual differences, and applications* (pp.233-250). New Jersey: Lawrence Erlbaum.

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

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

Jong, M. de & P.J. Schellens (2002). Tekstevaluatie. Onderzoek naar de validiteit van probleemopsporende methoden. *Tijdschrift voor Taalbeheersing, 24*, 146-166.

Kuusela, H. & P. Paul (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology, 113*, 387-404.

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

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

Norlin, E. & C.M.I.Winters (2002). *Usability testing for library websites: a hands-on guide*. Chicago: American Library Association.

Rubin, J. (1994). *Handbook of usability testing: How to plan, design, and conduct effective tests*. New York: John Wiley.

Russo, J.E., E.J. Johnson & D.L. Stephens (1989). The validity of verbal protocols. Memory *&* *Cognition, 17*, 759-769.

Short, E.J., S.W. Evans, S.E. Friebert & C.W. Schatschneider (1991). Thinking aloud during problem solving: Facilitation effects. *Learning and Individual Differences, 3* (2), 109-122.

Taylor, K.L. & J.P. Dionne (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 29*, 413-425

Teague, R., K. De Jesus & M. Nunes-Ueno (2001). Concurrent vs. post-task usability test ratings. *Proceedings of the Conference on Human Factors and Computing Systems, 31 March – 5 April 2001* (pp.289-290). Seattle, WA: ACM SIGCHI.