MENNO DE JONG EN
PETER JAN SCHELLENS

Tekstevaluatie. Onderzoek naar de validiteit van probleemopsporende methoden

1. Inleiding[*]

Samenvatting

Hoewel het nut van het pretesten van concepteksten algemeen geaccepteerd lijkt, is onderzoek naar de waarde van verschillende methoden voor tekstevaluatie nog tamelijk schaars. Dit artikel geeft een overzicht van onderzoek dat is verricht naar de validiteit van probleemopsporende evaluatiemethoden. Er worden twee onderzoekslijnen besproken. In onderzoek naar de predictieve validiteit wordt direct of indirect nagegaan of de lezersproblemen die een evaluatiemethode aan het licht brengt, waardevol zijn met het oog op revisie van de tekst. In onderzoek naar de congruente validiteit wordt nagegaan wat de verschillen en overeenkomsten zijn in de problemen die met verschillende methoden worden opgespoord.

Onder tekstschrijvers en communicatieprofessionals is het besef gegroeid dat het testen van teksten in een conceptfase, ook wel formatieve evaluatie genoemd, een belangrijke bijdrage kan leveren aan de effectiviteit van communicatie. Er is inmiddels ook een flink aantal methoden beschikbaar waaruit men in de praktijk kan kiezen als men een concepttekst, een opzet voor een advertentiecampagne, een website of een interface wil testen (voor overzichten zie De Jong & Schellens, 1995 en 1997; De Jong & Heuvelman, 1999.) Internationaal zijn vooral methoden om de gebruikersvriendelijkheid ('usability') van software, interfaces en technische documentatie te testen in de belangstelling komen te staan, zoals blijkt uit een flink aantal handboeken op dit gebied (Nielsen, 1993; Dumas & Redish, 1993; Rubin, 1994; Lindgaard, 1994; Velotta, 1995; Faulkner, 2000; Barnum, 2001).

In het verlengde van deze belangstelling vanuit de praktijk is de ontwikkeling en validering van evaluatiemethoden op de Nederlandse onderzoeksagenda verschenen. In 1996 wijdde het *Tijdschrift voor Taalbeheersing* al een themanummer aan onderzoek naar methoden voor tekstevaluatie (Renkema & Schellens, 1996). In hun overzicht van onderzoek op het gebied van tekstontwerp besteden Schellens & Maes (2000: 176-183) apart aandacht aan het Nederlandse onderzoek naar methoden voor tekstevaluatie. Dat onderzoek heeft tot dusver geleid tot diverse nieuwe evaluatiebenaderingen, zoals de functionele analyse (Lentz & Pander Maat, 1993), het CCC-evaluatiemodel (Renkema, 1994) en het computerprogramma Focus (Lentz & De Jong, 2000; De Jong & Lentz, 2001). Maar daarnaast heeft het bijgedragen aan de validering van evaluatiemethoden: er is een lijn van onderzoek ontstaan naar de waarde en beperkingen van bestaande en nieuw ontwikkelde evaluatieme-

thoden. In dit artikel geven we een overzicht van valideringsonderzoek naar formatieve evaluatiemethoden, zowel in Nederland als daarbuiten.

Een recent onderzoek op het gebied van mens-computer interactie laat goed zien waar het in zulk onderzoek om draait. Molich e.a. (1999) vergeleken de testresultaten van acht teams met ervaring in usability testing en één studententeam, die alle de opdracht kregen om de Hotmail-website van Microsoft te evalueren. De resultaten bleken zeer inconsistent. Het aantal problemen in de website dat de teams signaleerden, varieerde van 10 tot 150. Verder werd 75% van de gesignaleerde problemen slechts door één team gesignaleerd. Dergelijke resultaten roepen uiteraard vragen op over de betrouwbaarheid en validiteit van de gebruikte evaluatiemethoden. Zouden we niet meer overeenstemming mogen verwachten in de resultaten van de teams? Wat zegt bij zulke uiteenlopende resultaten een individuele usability test nog over de kwaliteit van en de problemen in de site?

Zoals uit het bovenstaande al blijkt, willen we ons in dit artikel niet beperken tot de evaluatie van teksten in strikte zin. Naar onze mening is het niet zinnig om scherpe grenzen te trekken tussen de evaluatie van papieren teksten, teksten op beeldscherm en teksten in een meer of minder rijke (audio-)visuele omgeving. Wel gaat het ons steeds om communicatiemiddelen waarin de verbale component een dominante rol speelt. Het kan daarbij gaan om lesmateriaal, interfaces, handleidingen, voorlichtingsmateriaal, etc.

We beperken ons in dit overzicht tot probleemopsporende methoden. Dat zijn methoden die de detectie en diagnose van mogelijke lezersproblemen beogen. Dat impliceert doorgaans een kwalitatieve en exploratieve onderzoeksaanpak. Aan kwantitatief toetsende methoden, waarmee een totaalbeeld van de kwaliteit van een document wordt verkregen in de vorm van een score (bijvoorbeeld voor de begrijpelijkheid of aanvaarbaarheid van de geboden informatie), besteden we hier geen aandacht. Methoden zoals leesbaarheidsformules, de cloze test en het meten van taakuitvoeringstijden blijven dus buiten beschouwing. In de appendix bij dit artikel is een lijst opgenomen met een korte omschrijving van de evaluatiemethoden die in de tekst worden genoemd.

De vraag waar het in het onderzoek naar de waarde van probleemopsporende pretestmethoden uiteindelijk om gaat is: hoe kan een *tekstontwerper of schrijver in de loop van het tekstontwerpproces optimaal profiteren van een formatieve evaluatie van conceptteksten?* In de dagelijkse praktijk moet een professionele tekstschrijver daartoe een geschikte methode kiezen, beslissen over het soort en het aantal respondenten dat hij gaat inzetten, en de vaak grote hoeveelheid data die hij verzamelt, interpreteren en vertalen in revisies van de tekst. Al deze keuzes beïnvloeden de opbrengst van de evaluatie en dus de zin van de evaluatie-inspanningen. Onderzoek naar methoden voor tekstevaluatie is erop gericht empirische ondersteuning te verschaffen voor dergelijke keuzes.

Tot op zekere hoogte vertonen alle beschikbare methoden enige *face validity*. Voor elke tekstevaluatiemethode kan een min of meer plausibele rationale worden gegeven en in de praktijk kan vrijwel elke poging om conceptteksten systematisch te evalueren vruchtbaar zijn. Alleen al het feit dat een tekst intensieve aandacht krijgt in een evaluatieproces zal vaak nieuwe en waardevolle inzichten opleveren. Vandaar dat onderzoek met name gericht moet zijn op de relatieve kracht en zwakte van de beschikbare evaluatiemethoden. Onderzoek naar de validiteit van methoden is erop gericht objectieve en empirische evidentie te verschaffen voor de veronderstellingen waarop de methoden zijn gebaseerd. In dit artikel bespreken we twee onderzoekslijnen:

- *Onderzoek naar de predictieve validiteit van methoden.* Traditioneel verwijst validiteit in de

methodologische literatuur naar de mate waarin een methode resultaten oplevert die vrij zijn van systematische vertekeningen. De vraag daarbij is of een methode werkelijk meet wat zij pretendeert te meten. Toegepast op methoden voor probleemopsporende evaluatie impliceert dat twee belangrijke vragen: (1) komen de problemen die een methode opspoort overeen met de problemen die lezers werkelijk hebben, en (2) ziet de methode geen werkelijke problemen over het hoofd?

- *Onderzoek naar de congruente validiteit van methoden*. Als methoden hetzelfde beogen te meten, moeten ze ook dezelfde resultaten opleveren. In de praktijk is dat bij methoden voor formatieve evaluatie zelden het geval. Vergelijkingen tussen probleemopsporende methoden kunnen betrekking hebben op verschillen en overeenkomsten in de resultaten van methoden: in aantallen problemen, soorten problemen en overlap tussen problemen die verschillende methoden aan het licht brengen. Op die manier kan een antwoord worden gevonden op de vraag voor welk type problemen welke methode bij welk type document de beste resultaten oplevert.

2. Predictieve validiteit

Zoals de term 'predictieve validiteit' al aangeeft, gaat het hierbij om de vraag of methoden een juiste voorspelling opleveren van de problemen die lezers in een tekst ervaren. Normaal gesproken worden in een onderzoek naar de predictieve validiteit van een methode de resultaten afgezet tegen de resultaten van een zogenaamd 'criteriuminstrument', waarvan de validiteit al min of meer vast staat. In het geval van probleemopsporende evaluatiemethoden is een dergelijk instrument echter niet beschikbaar. Vandaar dat drie andere typen van onderzoek worden uitgevoerd om licht te werpen op de predictieve validiteit van methoden: pretest-en-revisie-experimenten, probleembeoordelingsstudies en onderzoek met gemanipuleerde teksten.

2.1 Pretest-en-revisie-experimenten De meest gekozen onderzoeksopzet om de predictieve validiteit van evaluatiemethoden te onderzoeken maakt gebruik van revisie als tussenstap. De onderzoeksvraag is of revisie op basis van evaluatieresultaten leidt tot een aantoonbare verbetering van de kwaliteit van de tekst. Om dit vast te stellen wordt de kwaliteit van de oorspronkelijke en een gereviseerde versie van de tekst experimenteel vergeleken. Dat kan ofwel in een onderzoeksopzet met onafhankelijke groepen, waarin verschillende groepen proefpersonen worden toegekend aan de originele en de gereviseerde versie, ofwel in een opzet met een paarsgewijze vergelijking, waarin proefpersonen een voorkeur moeten uitspreken voor één van beide versies.

In de loop der jaren zijn er heel wat pretest-en-revisie-experimenten uitgevoerd, leidend tot de globale conclusie dat een formatieve evaluatie zoden aan de dijk zet. In de meeste studies hadden evaluatie en revisie positieve effecten op de kwaliteit van de onderzochte teksten. Voor zes onderzoeken geldt dat deze algemene conclusie niet kan worden toegeschreven aan één specifieke methode. Jansen, Klatter & De Vet (1991), Jansen & Steehouder (1989), Sanders e.a. (1994), Schriver (1997, p. 444-462) en Wright (1994) baseerden hun revisies op een combinatie van expertanalyse en lezersfeedback, terwijl Allwood & Kalén (1997) drie verschillende soorten lezersonderzoek gebruikten als input voor de revisie van een handleiding (te weten het onderstrepen van moeilijke fragmenten, het noteren

van vragen en hardop-denkprotocollen). De overige 17 revisieonderzoeken werpen enig licht op de verdiensten van specifieke evaluatiemethoden. In sommige van de hier genoemde onderzoeken worden de effecten van meerdere evaluatiemethoden met elkaar vergeleken. Tabel 1 geeft een overzicht van de resultaten voor zes soorten methoden:

- Tekstgerichte methoden (bijvoorbeeld richtlijnen en checklists).
- Expertgerichte methoden (bijvoorbeeld een review door doelgroep- of genredeskundigen).
- Doelgroepgerichte methoden waarin uitsluitend de effecten van tekstgebruik centraal staan (bijvoorbeeld begripstests).
- Doelgroepgerichte methoden waarmee het proces van tekstgebruik in kaart wordt gebracht (bijvoorbeeld hardop-denkprotocollen).
- Doelgroepgerichte methoden waarmee oordelen van tekstgebruikers worden verzameld (bijvoorbeeld de plus-en-minmethode).
- Doelgroepgerichte methoden waarmee de effecten en het proces van het tekstgebruik en de oordelen van tekstgebruikers geïntegreerd worden bestudeerd (bijvoorbeeld een één-op-één evaluatie van lesmateriaal).

Tabel 1. *Resultaten van pretest-en-revisie-experimenten*

Type methode	**Geen effect**	**Gemengd effect**	**Positief effect**
Tekstgericht: richtlijnen	Duffy & Kabance (1992)		*Renkema (2000)
Expertgericht	*Weston e.a. (1997)	*Swaney e.a. (1992)	Davidove & Reiser (1991)
Doelgroepgericht: effecten tekstgebruik			*Baker (1970); Gropper & Lumsdaine (1961)
Doelgroepgericht: proces tekstgebruik			*Swaney e.a. (1992)
Doelgroepgericht: tekstbeoordeling	Ahlschwede (1970)	*De Jong (1998); *De Jong & Rijnks (te verschijnen); *Renkema (2000)	*Lentz & De Jong (2000); Nathanson & Henderson (1980); Micklos & Bishop (1982); Vroom (1987)
Doelgroepgericht: geïntegreerd			Davidove & Reiser (1991); *Kandaswamy e.a. (1976); *Wager (1983); *Weston e.a. (1997)

* onderzoek met enige vorm van controle over de rol van revisor

Er zijn slechts twee onderzoeken verricht naar het effect van tekstgerichte methoden. Duffy & Kabance (1982) onderzochten of een evaluatie aan de hand van traditionele richtlijnen voor leesbaar schrijven aan de begrijpelijkheid van teksten zou bijdragen. Hun conclusie was negatief. Omdat de richtlijnen in dit onderzoek beperkt bleven tot tamelijk basale kenmerken op woord- en zinsniveau, is het niet gerechtvaardigd hieruit conclusies te trekken voor heuristieken en richtlijnen in het algemeen. Renkema (2000) onderzocht de waarde van een evaluatie van zakelijke brieven aan de hand van zijn CCC-model, en kwam tot positieve bevindingen: de revisies die studenten op grond van hun eigen CCC-evaluatie produceerden, werden door een team van tekstdeskundigen gemiddeld hoger gewaardeerd dan de oorspronkelijke versies van de brieven. De effectiviteit van de brieven voor de doel-

groep bleef in dit onderzoek echter buiten beschouwing. Onderzoek met het CCC-model kan overigens, vanwege het zeer globale, weinig sturende karakter van het model, ook worden gezien als representant van een expertgerichte evaluatiebenadering.

Voor expertgerichte methoden werden in drie onderzoeken uiteenlopende resultaten verkregen. Weston e.a. (1997) vonden geen significante verbetering na revisie van lesmateriaal op basis van het commentaar van drie inhoudelijke en drie doelgroepexperts. Swaney e.a. (1992) stelden vast dat een evaluatie door een tekstontwerpteam van vier soorten functionele documenten leidde tot een verhoogde effectiviteit in twee van de vier gevallen. Ten slotte vonden Davidove & Reiser (1991) dat revisie van lesmateriaal gebaseerd op feedback door negen docenten leidde tot hogere testresultaten bij leerlingen. In deze drie onderzoeken werden de effecten van expertcommentaar ook vergeleken met de feedback van lezers. De resultaten van Weston e.a. en Swaney e.a. laten zien dat feedback van de doelgroep wellicht een krachtiger middel is om een tekst op de behoeften van lezers toe te snijden dan expertcommentaar: waar met behulp van expertcommentaar geen effecten werden behaald, lukte dit met lezersfeedback wel. Davidove & Reiser vonden, evenals Golas (1983), echter geen verschillen in effect tussen de feedback van experts en lezers.

Twee wat oudere studies naar de effecten van doelgroepgerichte methoden richtten zich op de vraag of de resultaten van begripstesten kunnen bijdragen aan de revisie van communicatiemiddelen (Gropper & Lumsdaine, 1961; Baker, 1970). In beide onderzoeken werden significante verbeteringen vastgesteld.

Slechts in één studie werd gebruik gemaakt van doelgroepgerichte methoden waarin het proces van het tekstgebruik werd gevolgd. Swaney e.a. (1992) stelden in hun eerdergenoemde onderzoek vast dat een verzekeringspolis was verbeterd door revisie op basis van hardop-denkprotocollen van lezers die de opdracht kregen om de geboden informatie daadwerkelijk te gebruiken.

Zeven onderzoeken waren gericht op de effecten van pretestonderzoek waarin de doelgroep om een tekstbeoordeling werd gevraagd. De resultaten liepen uiteen. Ahlschwede (1970) stelde slechts een marginale verbetering vast van een brochure na revisie op basis van een voorloper van de plus-en-minmethode. De Jong (1998, p. 47-91) vond in zijn onderzoek met zes brochures dat lezers een voorkeur hadden voor de versies die op basis van plus-en-minmethode waren herzien. Verder was de effectiviteit van vijf van de zes brochures in een aantal opzichten toegenomen. De Jong & Rijnks (te verschijnen) evalueerden een op grond van plus-en-minresultaten herziene brochure opnieuw met de plus-en-minmethode, en stelden vast dat het aantal genoemde problemen in de brochure niet was afgenomen, maar dat er wel een duidelijke verschuiving was in de aard van de problemen. Renkema (2000) vond in een onderzoeksopzet waarin studenten een beperkte plus-en-minevaluatie moesten uitvoeren (met slechts vier proefpersonen), een significante verbetering wanneer de herschrijvers lezersfeedback gebruikten die door medestudenten was verzameld, maar geen significante verbetering wanneer de herschrijvers hun eigen lezersfeedback verzamelden. Nathenson & Henderson (1980) en Micklos & Bishop (1982) vonden positieve effecten van het gebruik van beoordelingsmethoden, respectievelijk verkregen met feedbackvragen in de tekst en met de signaled stopping technique. Hetzelfde geldt voor Vroom (1987), met de plus-en-minmethode, en Lentz & De Jong (2000), met het door hen ontwikkelde computerprogramma Focus.

Ten slotte werden in vier studies de effecten onderzocht van een combinatie van verschillende soorten lezersdata, waarbij lesmateriaal werd uitgeprobeerd op individuele leerlingen en/of in kleine groepen (Kandaswamy, Stolovitch & Thiagarajan, 1976; Wager, 1983; Davidove

& Reiser, 1991; Weston e.a., 1997). In alle gevallen werden significante verbeteringen geboekt op grond van de doelgroepgerichte methoden.

Al met al wordt het in de adviesliteratuur veronderstelde belang van een pretest onder potentiële lezers door deze bevindingen onderstreept. Voor tekst- en expertgerichte methoden levert het beschikbare onderzoek een minder eenduidig beeld op, maar ook voor het nut daarvan bestaan aanwijzingen.

Er moet echter een belangrijk voorbehoud worden gemaakt bij de hierboven beschreven resultaten. Een potentiële zwakte van validiteitsonderzoek met revisie als tussenstap gaat schuil in de bijdrage van de revisoren of herschrijvers. Een onderzoeksopzet die is gebaseerd op revisie moet er op een of andere manier voor zorgen dat de gevonden effecten kunnen worden toegeschreven aan de gegevens uit de evaluatie. In slechts negen van de pretest-en-revisie-experimenten is een poging gedaan om de rol van de revisor in zekere mate te controleren:

- Baker (1970), Golas (1983), Kandaswamy, Stolovitch & Thiagarajan (1976), Lentz & De Jong (2000) en Renkema (2000) gebruikten een onderzoeksopzet met meervoudige revisie, waarin verschillende herschrijvers de evaluatieresultaten gebruikten om een tekst te herzien.
- In het onderzoek van Wager (1983) werden de revisies in relatie tot de feedback door experts gecontroleerd op hun geschiktheid.
- Swaney e.a. (1992) en Weston e.a. (1997) kozen voor een design waarin revisies op grond van lezersfeedback werden vergeleken met revisies waarin geen gebruik van feedback werd gemaakt.
- De Jong (1998) en De Jong & Rijnks (te verschijnen) leverden een schriftelijke verantwoording van alle revisiebeslissingen, waarin steeds expliciet een verband werd gelegd tussen lezersproblemen en veranderingen in de tekst.

2.2 Probleembeoordelingsstudies Een tweede manier om de predictieve validiteit van probleemopsporende evaluatiemethoden te onderzoeken bestaat uit een nader onderzoek naar het belang van de opgespoorde problemen. Dat kan door hun waarschijnlijkheid vast te stellen ('Is het waarschijnlijk dat de beoogde lezers het opgespoorde probleem daadwerkelijk zullen ondervinden?') en door de ernst van de gevolgen van het probleem in te schatten ('In hoeverre zal het opgespoorde probleem de effectiviteit van de tekst bedreigen?') (Nielsen, 1994; Lentz & De Jong, 1996).

Het belang van opgespoorde lezersproblemen kan op twee manieren worden bepaald. De eerste mogelijkheid bestaat uit de constructie van een test waarmee de lezersproblemen worden geverifieerd in een nieuwe, bij voorkeur grotere steekproef van beoogde lezers. Dit type onderzoek richt zich op de waarschijnlijkheid van problemen en zegt niet noodzakelijk iets over de ernst ervan. We zijn deze benadering tot zover in geen enkel onderzoek tegengekomen. Een tweede mogelijkheid bestaat erin experts (of ervaren lezers uit de doelgroep) te vragen de ernst van problemen te beoordelen. Van dit type onderzoek komen we een aantal voorbeelden tegen, maar die waren doorgaans niet primair gericht op het verkrijgen van inzicht in de validiteit van methoden. Ze trachtten een onderscheid te maken tussen belangrijke en onbelangrijke problemen met het oog op andere onderzoeksdoelen, zoals het bepalen van de gewenste steekproefomvang (Lewis, 1994; Nielsen, 1994; Virzi, 1992) of de mate van consensus onder professionals bij de beoordeling van lezerscommentaar (Lentz & De Jong, 1996).

De Jong (1998, p.93-105) gebruikte expertoordelen als een additionele bron van informatie over de validiteit van de plus-en-minmethode voor het testen van brochures. De resultaten bleken moeilijk te interpreteren door een gebrek aan overeenstemming tussen de participerende deskundigen. De experts waren het er weliswaar over eens dat een substantieel deel van de lezersfeedback betrekking had op belangrijke lezersproblemen (gemiddeld zo'n 37% van de genoemde problemen in een brochure), maar zij waren het in het geheel niet eens over de vraag welke problemen de belangrijkste waren. Een meerderheid van de lezersproblemen kreeg daardoor een nietszeggende neutrale gemiddelde score, ongeveer in het midden van een vijfpuntsschaal. Dit kan worden beschouwd als een belangrijk nadeel van dit type valideringsonderzoek.

2.3 Onderzoek met gemanipuleerde teksten Een derde manier om de predictieve validiteit van evaluatiemethoden te onderzoeken gaat uit van de 'error detection'-benadering. In dit type onderzoek worden lezers geconfronteerd met gemanipuleerde teksten waarin bepaalde soorten problemen zijn ingebouwd. De vraag is vervolgens of de lezers de ingebouwde problemen inderdaad zullen opmerken. Deze benadering is vaak gebruikt in theoretisch onderzoek naar comprehension monitoring, waarin wordt nagegaan in hoeverre lezers zich bewust zijn van de begripsproblemen die ze tegenkomen (Baker, 1985, 1989).

Er zijn in totaal vijf onderzoeken geweest waarin de validiteit van pretestmethoden met behulp van gemanipuleerde teksten is onderzocht. Twee daarvan hadden betrekking op de plus-en-minmethode. Schuurs, Van den Bergh & Verhoeven (1991) onderzochten de waarde van drie varianten van de methode voor het opsporen van vier soorten problemen (spelling, grammatica, structuur en afstemming op de lezer). Deze problemen bevonden zich, achteraf gezien, wellicht te veel buiten het domein van gangbaar pretestcommentaar. Het is dan ook niet verwonderlijk dat de detectiepercentages aan de lage kant waren: de proefpersonen ontdekten gemiddeld tussen de 43% (spelling) en de 8% (tekststructuur) van de aangebrachte problemen. Het onderzoek leidde ook tot onverwachte resultaten: zo leverde een specifiek op publiekgerichtheid toegespitste variant van de plus-en-minmethode juist relatief weinig commentaar op over de afstemming op de lezer. Daarnaast bleken de varianten van de plus-en-minmethode verschillend te werken bij de twee in het onderzoek betrokken teksten. Ook in het tweede onderzoek naar varianten van de plus-en-minmethode, van Pander Maat (1996), waren de detectiepercentages laag. In dit onderzoek werden vijf soorten lezersproblemen aangebracht in medicijnenbijsluiters: (a) problemen met medische termen, (b) problemen met 'quantifiers' (woorden of constructies die niet-cijfermatige indicatie van een hoeveelheid, kans of tijdsduur geven), (c) problemen met betrekking tot medische condities, (d) problemen met ontbrekende informatie en (e) inconsistenties in de tekst. De detectiepercentages per categorie bleven gemiddeld onder de 50%; zeer lage detectiepercentages werden gevonden voor problemen met quantifiers (6%) en ontbrekende informatie (2%).

De drie overige studies, die wat het gebruikte onderzoeksmateriaal betreft op elkaar voortbouwden, hadden betrekking op het pretesten van vragenlijsten. In vragenlijsten werden de volgende vijf typen problemen ingebouwd: (a) suggestieve vragen, (b) dubbele vragen, (c) ambigue vragen, (d) inadequate termen, en (e) ontbrekende antwoordalternatieven. Er werden drie evaluatiemethoden gebruikt, telefonische interviews, face-to-face interviews en hardop-denkprotocollen. Hunt, Sparkman & Wilcox (1982) vonden zeer lage detectiepercentages voor de eerste vier categorieën (van 3 tot 5%) en een iets hoger percentage voor ont-

brekende antwoordalternatieven (33%). Met relatief hoog opgeleide proefpersonen bereikten Diamantopolous, Reynolds & Schlegelmilch (1994) aanzienlijk hogere detectiepercentages (variërend van 16% voor dubbele vragen tot 48% voor inadequate termen). Ten slotte vergeleken Reynolds & Diamantopoulos (1998) de effectiviteit van een persoonlijke (face-to-face) en een onpersoonlijke (schriftelijke) evaluatiemethode om de vijf soorten problemen te ontdekken. Zij stelden vast dat de persoonlijke evaluatie effectiever was dan de onpersoonlijke, met een gemiddelde detectiescore van 32 tegenover 20%. Het verschil kon grotendeels worden toegeschreven aan de categorie 'inadequate termen'.

De 'error-detection'-aanpak heeft het voordeel dat de onderzoeksopzet rechtlijniger en eenvoudiger is dan bij de andere twee soorten onderzoek naar predictieve validiteit. De benadering strookt echter niet helemaal met de vooronderstellingen van probleemopsporende methoden, namelijk dat de evaluatie resulteert in nieuwe en verrassende inzichten in lezersproblemen. Bovendien is, in principe, aanvullende evidentie nodig om aan te tonen dat de gemanipuleerde problemen daadwerkelijk reële lezersproblemen zijn. Pander Maat (1996) voerde een extra controle uit op de problemen die hij in de bijsluiter opnam, hetgeen hem ertoe bracht bij nader inzien drie van de veertien aangebrachte problemen buiten beschouwing te laten. In de andere vier onderzoeken ontbrak een dergelijke check en lijken de aangebrachte problemen duidelijker aan te sluiten op het professionele perspectief van de tekstschrijver of de onderzoeker dan op dat van de beoogde gebruikers. Het is, met andere woorden, de vraag of het wel fair is om van de doelgroep te verwachten dat ze de aangebrachte problemen aanwijst. Een eigenaardigheid van de beschikbare onderzoeken is verder dat ze niet rapporteren over het percentage problemen dat door ten minste één van de participanten is ontdekt. In plaats daarvan wordt gerapporteerd over het gemiddelde percentage ontdekte problemen per proefpersoon. In dat opzicht dragen de onderzoeken, ondanks hun pretenties, niet echt bij aan de validering van probleemopsporende evaluatiemethoden.

2.4 Conclusie Het onderzoek naar de predictieve validiteit van evaluatiemethoden bevestigt de basisgedachte dat een probleemopsporende evaluatie waardevol is. Er is echter nog weinig bekend over de predictieve validiteit van specifieke methoden. De drie gangbare soorten validiteitonderzoek hebben alle hun voor- en nadelen, en kunnen elkaar om die reden aanvullen. Pretest-en-revisie-experimenten kunnen worden gebruikt om een totaalindruk te verkrijgen van de mogelijke opbrengsten van een methode. Probleembeoordelingsstudies kunnen in principe nuttig zijn om belangrijke en onbelangrijke problemen van elkaar te onderscheiden. Onderzoek met gemanipuleerde teksten vormt een nuttige aanvulling omdat het uitgaat van een verzameling problemen die een evaluatiemethode volgens de onderzoeker, aan het licht moet kunnen brengen in plaats van de daadwerkelijke output van een methode.

3. Congruente validiteit

Onderzoek naar de congruente validiteit van evaluatiemethoden is gericht op overeenkomsten en verschillen in de evaluatieresultaten. Vier aspecten kunnen in de vergelijking aan de orde zijn: (1) het aantal ontdekte problemen, (2) de aandacht die aan verschillende soorten problemen wordt besteed, en (3) de mate waarin de ontdekte problemen overeen-

komen. Onderzoek naar de congruente validiteit kan zich ook rechtstreeks richten op de bijdrage die verschillende methoden leveren aan de effectiviteit van teksten. Omdat het dan feitelijk gaat om een vergelijking van de predictieve validiteit van methoden, is het beschikbare onderzoek in deze categorie in de vorige paragraaf verwerkt.

Het aantal ontdekte problemen geeft een ruwe indicatie van de waarde van een methode. De onderliggende veronderstelling - 'hoe meer hoe beter' - is eerder gerechtvaardigd voor methoden waarbij proefpersonen de tekst gebruiken als hulpmiddel bij het uitvoeren van een concrete taak, dan bij methoden waarbij proefpersonen de tekst niet gebruiken maar beoordelen. In het laatste geval kunnen de proefpersonen in hun beoordelingsrol geneigd zijn ook weinig reële problemen te signaleren om te laten zien dat ze hun taak als proefpersoon serieus nemen. Het belang van het aantal ontdekte problemen wordt ook gerelativeerd door onderzoek waaruit blijkt dat bij een gelijk aantal gesignaleerde problemen het belang van die problemen uiteen kan lopen (De Jong & Schellens, 2001a) en dat er geen eenduidige relatie is tussen het aantal gesignaleerde problemen en de kwaliteit van een brochure (De Jong & Rijnks, te verschijnen).

De aandacht die besteed wordt aan verschillende soorten problemen geeft een indicatie van de evaluatiestandaards die in verschillende methoden worden opgeroepen (zie ook Baker, 1985, 1989). Probleemcategorieën worden daarbij op verschillende manieren gedefinieerd, afhankelijk van het soort documenten dat wordt geëvalueerd en de onderzochte methoden. Bij het pretesten van vragenlijsten onderscheiden Presser & Blair (1994) de categorieën 'respondent semantic' (problemen van respondenten met de interpretatie van een vraag), 'respondent task' (problemen met het beantwoorden van een vraag), 'interviewer' (problemen van interviewers met het stellen van een vraag of het coderen van een antwoord) en 'analysis' (problemen van de onderzoeker met de interpretatie van de resultaten). In een ander onderzoek hanteren Oksenberg, Cannell & Kalton (1991) een vergelijkbare, maar wat verder uitgewerkte indeling. In een aantal andere onderzoeken werd de volgende typologie van De Jong (1998) gebruikt (Elling, 1997; De Jong & Lentz, 1996, 2001; De Jong & Rijnks, te verschijnen; De Jong & Schellens, 1998, 2001ab; Lentz & De Jong, 1997; Sienot, 1997):

- *Begripsproblemen*: lezers ondervinden interpretatie- of toepassingsproblemen door onduidelijke informatie, moeilijke zinsbouw, lastige woordkeus.
- *Acceptatieproblemen*: lezers twijfelen aan de juistheid van informatie of zijn het niet eens met waardeoordelen of adviezen in de tekst.
- *Waarderingsproblemen*: lezers geven de voorkeur aan een andere formulering van informatie zonder dat er sprake is van een begrips- of acceptatieprobleem.
- *Structuurproblemen*: lezers hebben problemen met de structuur van de aangeboden informatie of met de wijze waarop die structuur is gemarkeerd.
- *Relevantieproblemen*: lezers vinden informatie overbodig of te uitgebreid.
- *Volledigheidsproblemen*: lezers vragen om meer informatie over een topic.
- *Problemen met de lay-out*: lezers hebben problemen met de grafische vormgeving of de illustraties in een document.
- *Correctheidsproblemen*: Lezers signaleren een fout in grammatica, spelling of interpunctie of geven aan dat andere tekstuele conventies zijn geschonden.

Wanneer de ene methode bij dezelfde tekst bijvoorbeeld meer begripsproblemen oplevert en de andere methode meer acceptatieproblemen, leiden we daaruit af dat de ene methode een andere evaluatiestandaard oproept dan de andere.

De mate van overlap tussen de ontdekte problemen van twee methoden geeft het meest

gedetailleerde beeld van overeenkomsten en verschillen tussen methoden. Dergelijk onderzoek kan leiden tot conclusies over de mate waarin methode A de resultaten van methode B voorspelt of de mate waarin methode A en B elkaar aanvullen. De vaststelling of lezersproblemen identiek of verschillend zijn, kan in de praktijk echter problemen geven, zeker wanneer de vergeleken methoden verschillende soorten feedback opleveren, bijvoorbeeld bij een vergelijking tussen de resultaten van een usability test en die van een expertevaluatie met behulp van heuristieken. Lavery, Cockton & Atkinson (1997) stellen een gebrek aan zorgvuldigheid in dit opzicht vast in de beschikbare literatuur. Zij stellen daarom een standaardstructuur voor de rapportage van gebruikersproblemen voor om de vergelijking van individuele problemen te vergemakkelijken.

Hieronder beschrijven we het beschikbare onderzoek naar de congruente validiteit van evaluatiemethoden. Door de grote variëteit aan vergeleken methoden is het onmogelijk een compleet en gedetailleerd overzicht van de resultaten te geven. In plaats daarvan bespreken we de volgende clusters van onderzoek:

- onderzoek naar expert- en doelgroepgerichte methoden,
- onderzoek naar tekstgebruiks- en tekstbeoordelingsmethoden,
- onderzoek naar individuele en groepsgewijze beoordelingsmethoden.

3.1 Feedback door experts vergeleken met feedback door lezers De vraag of experts – met of zonder hulp van heuristieken, checklists of andere hulpmiddelen – andere feedback geven dan de beoogde lezers, is relatief vaak onderzocht (Desurvire, 1994; Dieli, 1986; John & Marks, 1997; De Jong & Lentz, 1996; Lentz & De Jong, 1997; Lentz & Pander Maat, 1992, 1993; Mack & Montaniz, 1994; Nielsen, 1994; Pander Maat, 1996; Presser & Blair, 1994; Renkema & Wijnstekers, 1997; Schriver, 1997, p.444-462; Weston, 1987). Het laatstgenoemde onderzoek is overigens lastig te interpreteren. Weston (1987) concludeert dat experts en lezers verschillende feedback geven op lesmateriaal, maar de experts en lezers kregen volledig verschillende vragen voorgelegd. Als gevolg hiervan zijn de resultaten niet zonder meer toe te schrijven aan de participanten in de evaluatie. In ander onderzoek is voor een beter interpreteerbare opzet gekozen. Experts en lezers krijgen dan een vergelijkbare evaluatietaak of experts wordt gevraagd de lezersproblemen te voorspellen.

Een aantal onderzoekers ging na in hoeverre experts in staat zijn de resultaten van een doelgroepgerichte evaluatie te voorspellen. Bij het pretesten van vragenlijsten vonden Presser & Blair (1994) een matige overlap tussen de resultaten van expertevaluatie en een aantal doelgroepgerichte methoden (Yule's Q = 0.36). In drie andere onderzoeken werd vastgesteld dat individuele experts - professionele schrijvers, intermediairs of inhoudelijk deskundigen - niet in staat waren meer dan 15% van de lezersproblemen uit een plus-en-min-evaluatie te voorspellen, en dat de voorspelbaarheid nauwelijks toeneemt bij een selectie van ernstige of relatief vaak genoemde lezersproblemen (De Jong & Lentz, 1996; Lentz & De Jong, 1997; Pander Maat, 1996). Mack & Montaniz (1994) kwamen tot eenzelfde conclusie - met percentages van 18 tot 22% - voor software ontwerpers en usability deskundigen die de gebruikersproblemen in een interface trachtten te voorspellen. Schriver (1997, p. 454) vond hogere predictiepercentages voor een ontwerpteam (dus niet voor individuele experts) in een iteratief ontwerpproces van handleidingen: 50% van de gebruikersproblemen in de oorspronkelijke handleiding werd correct voorspeld, tegenover 19% van de problemen in de gereviseerde handleiding. Dieli (1986) vond een nog hoger percentage van 72% voor een

team van tien tekstontwerpers die gezamenlijk een handleiding evalueerden, maar zij vergeleek probleemgebieden in plaats van probleembeschrijvingen, hetgeen uiteraard eerder tot correcte voorspellingen leidt (een passage die door experts en lezers om verschillende redenen als problematisch wordt gezien, gold in Dieli's onderzoek als een correcte voorspelling). Al met al lijkt de conclusie gerechtvaardigd dat de feedback van lezers uit de doelgroep ook voor experts vaak verrassende inzichten oplevert in de problemen die lezers ervaren in een tekst.

Tegenover de vraag of experts lezersproblemen kunnen voorspellen, staat de vraag of experts mogelijk extra feedback leveren, die met een doelgroepgerichte methode niet aan het licht komt. Zelfs als experts expliciet de opdracht krijgen zich te beperken tot het voorspellen van lezersproblemen, correspondeert 40 tot 70% van hun voorspellingen niet met de problemen die door lezers worden ondervonden of geformuleerd. (Dieli, 1986; John & Marks, 1997; De Jong & Lentz, 1996; Lentz & De Jong 1997; Pander Maat, 1996). Lentz & De Jong (1997) vonden een significant verschil tussen soorten experts: professionele schrijvers formuleerden aanzienlijk meer nieuwe problemen dan intermediairs. Helaas is er tot nu toe geen onderzoek gedaan naar de validiteit van de 'false alarms' die experts afgeven. We weten dus niet of de extra problemen die experts signaleren, waardevolle aanvullingen zijn of onechte problemen die voor de doelgroep niet gelden.

In een aantal studies is nagegaan wat het effect is van richtlijnen of checklists die de evaluatie van experts moeten ondersteunen (Desurvire, 1994; Dieli, 1986; John & Marks, 1997; Lentz & Pander Maat, 1992; Nielsen, 1994; Mack & Montaniz, 1994; Renkema & Wijnstekers, 1997). In vergelijking met open, ongerichte evaluaties door experts, rapporteert Dieli (1986) hogere predictiepercentages (85% van de probleemgebieden) als experts de door haar ontwikkelde revisiefilters gebruiken, en lagere percentages (35%) bij gebruik van traditionele schrijfrichtlijnen. Het inzetten van heuristieken voor usability testing kan dus vruchtbaar zijn, maar alles hangt kennelijk af van het precieze soort ondersteuning dat experts wordt geboden. De overige onderzoeken leveren zowel positieve als negatieve resultaten op. Lentz & Pander Maat (1992, 1993) vonden dat experts die met hun op een functionele evaluatie gebaseerde checklist werkten, alle problemen konden voorspellen die in een kleinschalige pretest met onder meer de plus-en-minmethode waren gevonden. Daarbij dient te worden aangetekend dat hun checklist was toegesneden op een zeer specifieke tekstsoort (voorlichtingsteksten over subsidiemogelijkheden). Nielsen (1994) vond dat experts die werkten met een heuristiek voor de evaluatie van interfaces, meer dan 80% van de interfaceproblemen ontdekten die in een usability test onder gebruikers aan het licht waren gekomen. En ook Renkema & Wijnstekers (1997) komen, in een vergelijking van plus-en-min- en expertevaluaties met behulp van het CCC-model, tot een hoge mate van overlap: in twee steekproeven uit de revisievoorstellen van studenten vonden zij een overlap van respectievelijk 88 en 63%. Hierbij is echter de vraag welke rol het 'filter' van de revisiebeslissingen precies heeft gespeeld. Daartegenover staan resultaten van Desurvire (1994), die vaststelde dat experts die werkten met heuristieken of scenario's, nooit meer dan 50% van de interfaceproblemen noemden die gebruikers in een usability test ondervonden. Bovendien misten ze vaker de serieuze problemen dan de ondergeschikte hindernissen in een interface. Mack & Montaniz (1994) vonden predictiepercentages van 13 tot 15% voor experts met een scenariobenadering.

John & Marks (1997) benaderden dezelfde vraag van de andere kant: zij onderzochten of problemen die door experts naar voren werden gebracht - met behulp van één van vijf ver-

schillende hulpmiddelen (heuristieken en verschillende vormen van scenario's - werden bevestigd in een usability test. Meer dan de helft van de gesignaleerde problemen was niet terug te vinden in de resultaten van de usability test. In het onderzoek van Nielsen (1994), waarin experts een interface evalueerden aan de hand van een heuristiek, was dat percentage nog iets hoger (58%). De auteurs lijken te verschillen in de duiding van de gevonden resultaten. John & Marks karakteriseren de niet-bevestigde problemen als verspilde moeite. Zij hanteren de resultaten van de usability test met andere woorden als criteriuminstrument, waarvan de validiteit op voorhand wordt aangenomen. Nielsen beargumenteert dat de extra problemen die experts aandragen juist zeer waardevol kunnen zijn (zonder dit overigens in zijn onderzoek ook na te gaan).

In aanvulling op verschillen in percentages ontdekte problemen, kunnen systematische verschillen tussen experts en lezers worden ontdekt door de gehanteerde 'evaluatiestandaards' te vergelijken. De Jong & Lentz (1996) en Lentz & De Jong (1997) vonden dat experts in vergelijking met lezers meer aandacht besteedden aan de presentatie van informatie en minder aan de inhoud. Ook Dieli (1986) stelde vast dat experts relatief veel aandacht besteedden aan stilistische aspecten. Pander Maat (1996) vond dat lezers bij de evaluatie van een medicijnenbijsluiter meer aandacht besteedden aan begripsproblemen, terwijl experts (in dit geval apothekers en communicatiestudenten) meer feedback gaven op de aanvaardbaarheid, de structuur en de volledigheid van geboden informatie. Hij merkt overigens ook op dat experts en lezers binnen probleemcategorieën verschillende problemen kunnen vaststellen. Met betrekking tot de aanvaardbaarheid van informatie gaven lezers vooral aan informatie te wantrouwen of adviezen niet serieus te nemen, terwijl experts meer feitelijke rectificaties gaven. Daarnaast ontkrachten de studies van Lentz & Pander Maat (1993) en Renkema & Wijnstekers (1997) Schriver's (1989) veronderstelling dat tekstgerichte methoden nauwelijks feedback opleveren op globale en structurele tekstkenmerken. Zowel de functionele analyse als het CCC-model blijkt op deze punten zeker niet minder commentaar te genereren dan een doelgroepgerichte pretest. De Jong & Schellens (2001b) laten in een vergelijking tussen de plus-en-minmethode en een functionele analyse zelfs zien dat de kracht van een functionele analyse vooral betrekking heeft op de selectie en structurering van informatie; juist voor het opsporen van lokale brochureproblemen, bijvoorbeeld met begrijpelijkheid en waardering van informatie, lijkt de tekstgerichte methode veel minder geschikt.

Samenvattend kunnen we concluderen dat expertgerichte en doelgroepgerichte evaluatie eigenlijk twee heel verschillende dingen zijn. Hulpmiddelen als heuristieken of scenario's kunnen experts helpen om een document vanuit het lezersperspectief te evalueren, maar de empirische gegevens die zulke benaderingen ondersteunen zijn vooralsnog schaars. Een veelbelovende alternatieve benadering om doelgroep- en expertgerichte evaluatie dichter bij elkaar te brengen, schuilt in de opleiding van schrijvers. Onderzoek van Schriver (1992) en Couzijn (1995; Couzijn & Rijlaarsdam 1998) toonde aan dat leerling-schrijvers – respectievelijk studenten en middelbare scholieren - kunnen profiteren van een systematische confrontatie met feedback van lezers. Zij kregen daardoor meer voeling met de behoeften van hun lezers en worden gaandeweg beter in het voorspellen van lezerscommentaar.

Het is naar onze mening niet zinnig om de evaluatie door experts uitsluitend te zien als een alternatief voor een doelgroepgerichte evaluatie. Experts kunnen hun eigen bijdrage leveren aan de kwaliteit van documenten en de resultaten van doelgroepgerichte evaluatie

aanvullen. Zij zullen bijvoorbeeld een beter overzicht hebben over de doelstellingen van een document en over de alternatieven in het tekstontwerp. Bovendien is het doorgaans onmogelijk om alle aspecten van een document en alle mogelijke invalshoeken van lezersgroepen in een lezersgerichte evaluatie te betrekken.

3.2 Tekstgebruik versus tekstbeoordeling Congruente validiteit is ook aan de orde in vergelijkend onderzoek naar de waarde van tekstgebruiksmethoden (zoals hardop-werkonderzoek) in vergelijking met methoden waarin zelfrapporterende proefpersonen in een beoordelaarsrol worden geplaatst (zoals de plus-en-minmethode). Beide benaderingen hebben in principe hun voor- en nadelen. Enerzijds wordt verondersteld dat tekstgebruiksdata superieur zijn waar het om daadwerkelijke problemen bij taakuitvoering gaat. Zelfrapporterende proefpersonen realiseren zich mogelijk niet altijd dat zij bijvoorbeeld een begripsprobleem hebben, of kunnen besluiten een probleem maar niet onder woorden te brengen om geen slechte indruk te maken (of juist problemen verzinnen om de onderzoeker tevreden te stellen). Ook kunnen ze problemen vergeten doordat de rapportage plaatsvindt na afronding van een leesproces of taakuitvoeringsproces. Aan de andere kant kunnen zelfrapporterende proefpersonen feedback geven op een breder scala aan tekstkenmerken, zoals aanvaardbaarheid, volledigheid, waardering en relevantie. Enkele onderzoekingen bieden empirische ondersteuning voor deze vooronderstellingen.

De resultaten uit vijf studies bieden enige ondersteuning voor de superioriteit van gebruiksdata. Henderson e.a. (1995) vergeleken (retrospectieve) hardop-denkprotocollen met data uit toetsaanslagregistratie, een vragenlijst en een interview bij het testen van software. De hardop-denkmethode bracht meer problemen aan het licht dan de toetsaanslagregistratie en de vragenlijst. Er waren geen significante verschillen met de interviews. Allwood & Kalén (1997) testten gebruikershandleidingen met hardopdenk-protocollen, het onderstrepen van moeilijke passages en het opschrijven van vragen die tijdens het lezen rijzen. Er werden marginaal significante verschillen gevonden tussen hardop denken en onderstrepen. In een kleinschalig onderzoek vergeleken Medley-Mark & Weston (1988) een individuele hardop denkende proefpersoon met een duo dat de gerezen problemen met lesmateriaal al doende tezamen besprak en een kleine groep van drie proefpersonen die stil en zonder onderbreking werkten en achteraf werden geïnterviewd. De hardop denkende proefpersoon was het meest productief, de groep van drie identificeerde gezamenlijk het minste en de minst gedetailleerde problemen. Bischoping (1989) stelde vast dat de problemen die interviewers na een focusgroepdiscussie over een conceptvragenlijst rapporteerden, niet goed overeenkwamen met de problemen die konden worden aangewezen in gedragsprotocollen die van interviewers en respondenten werden gemaakt. Ten slotte vonden Gillham & Buckner (1997) een discrepantie tussen het gedrag van gebruikers van een multimedia encyclopedie op CD-ROM en hun rangordening van belangrijke kenmerken. Hoewel de proefpersonen veel tijd besteedden aan het spelen met de multimediale mogelijkheden van de CD-ROM, achtten ze het belang van die mogelijkheden niet erg groot. Parallel daaraan vonden ze de tekstinhoud van cruciaal belang, terwijl ze relatief weinig tijd besteedden aan het lezen van tekstuele informatie.

Een onderzoek van Sienot (1997) bevestigt de veronderstelling dat zelfrapportage een breder scala aan problemen aan het licht kan brengen dan tekstgebruiksmethoden. Sienot vergeleek de plus-en-minmethode en hardop-denkprotocollen bij de evaluatie van Websites. De plus-en-minmethode bracht significant meer problemen aan het licht dan de hard-

opdenk-protocollen, vooral omdat er meer waarderingsproblemen werden geformuleerd. In hetzelfde experiment onderzocht Sienot ook of de evaluatieresultaten werden beïnvloed door de opdracht die de proefpersonen kregen. De helft van de proefpersonen kreeg een aantal gerichte opdrachten, de andere helft kon vrij surfen over de Websites. De proefpersonen zonder gerichte taak brachten meer problemen aan het licht, met name betreffende de volledigheid en de relevantie van informatie. Aan de andere kant rapporteerden zij significant minder begripsproblemen dan de proefpersonen met gerichte opdrachten. Dit laatste resultaat correspondeert min of meer met de resultaten van een vergelijking van gebruikersprotocollen (hardop-denkprotocollen met taken) en lezersprotocollen (hardop-denkprotocollen zonder taak) door Dieli (1986). Volgens Dieli richtten de gebruikers met taak zich vooral op toegankelijkheid en toepasbaarheid, terwijl de lezers zonder taak zich richtten op de betekenis van de geboden informatie.

Een vergelijkend onderzoek van De Jong & Lentz (2001) naar de resultaten van twee typische tekstbeoordelingsmethoden, namelijk de plus-en-minmethode en het computerprogramma Focus (waarin proefpersonen hun commentaar op teksten invoeren door passages aan te klikken, probleemcategorieën te kiezen en probleemomschrijvingen in te typen), laat zien dat er mogelijk een gradatie is binnen de tekstbeoordelingsmethoden. De proefpersonen met de plus-en-minmethode leken de tekst meer vanuit een gebruikersperspectief te benaderen dan de proefpersonen met Focus. Dat uitte zich in meer aandacht voor de volledigheid van de geboden informatie en minder aandacht voor waardering, spelling en interpunctie.

3.3 Individuele versus groepsgewijze dataverzameling De Jong & Schellens (1996, 1998) vergeleken de resultaten van de gebruikelijke individuele variant van de plus-en-minmethode met een groepsgewijze variant waarin de bespreking van plussen en minnen niet individueel maar in focusgroepen plaats vindt. Op basis van algemene literatuur over focusgroepen kunnen een aantal vooronderstellingen worden geformuleerd over de dynamica van de groepsdiscussie. Zo kunnen we er vanuit gaan dat de deelnemers elkaars meningen beïnvloeden, commentaar zullen achterhouden, of zullen voortborduren op het commentaar van anderen. Dergelijke verschijnselen werden inderdaad gevonden. Het eerdergenoemde onderzoek van Bischoping (1989) liet ook zien dat deelnemers in een focusgroep niet alle problemen in een vragenlijst naar voren brengen die ze eerder hadden genoteerd.

De hoofdvraag in het onderzoek van De Jong & Schellens betrof echter de invloed van het groepsproces op de probleemdetectie. Er werden geen verschillen gevonden in het aantal problemen dat per proefpersoon naar voren werd gebracht. Er was echter wel een verschil in de soorten problemen die aan het licht traden. De focusgroep bracht participanten ertoe meer feedback te leveren op de brochure als geheel en minder op problemen op woordniveau in te gaan. Bovendien formuleerden participanten in de focusgroepen meer acceptatieproblemen en minder begripsproblemen dan individuele proefpersonen. Op basis van deze resultaten lijkt de groepsgewijze variant van de plus-en-minmethode meer geschikt wanneer de overtuigingskracht van de concepttekst in het geding is en de individuele variant meer geschikt wanneer de pretest meer gericht is op begrijpelijkheid van informatie.

3.4 Conclusie Het beschikbare onderzoek naar de congruente validiteit van evaluatiemethoden laat zien dat de gebruikte methode een sterke invloed heeft op het aantal en de soort problemen die aan het licht worden gebracht. Dat geldt met name voor de vergelij-

king tussen expertgerichte en doelgroepgerichte methoden. Maar ook binnen de categorie doelgroepgerichte methoden worden verschillen gevonden, met name tussen tekstgebruiksdata en data op basis van zelfrapportage door proefpersonen en tussen individuele en groepsgewijze dataverzameling.

Duidelijk is geworden dat een expertgerichte evaluatie normaal gesproken andere resultaten zal opleveren dan een doelgroepgerichte evaluatie. Toch lijkt het zinvol om nader onderzoek te doen naar de voorspelbaarheid van pretestresultaten bij verschillende teksttypen en/of bij verschillende doelgroepen: het overall-beeld van moeilijk te voorspellen lezersproblemen zou op die manier kunnen worden genuanceerd en toegespitst. Toekomstig onderzoek zal ook duidelijk moeten maken in hoeverre experts door hulpmiddelen als heuristieken en scenario's kunnen worden geholpen bij het anticiperen op lezersproblemen. Ook zal toekomstig onderzoek in kaart moeten brengen wat de waarde van expertevaluaties is: waarin schuilt de expertise van de communicatieprofessional en tekstschrijver, en in hoeverre levert die expertise inzichten op over de publieksafstemming van teksten die doelgroepgerichte methoden over het hoofd zien?

Over verschillen in opbrengst tussen tekstgebruiks- en tekstbeoordelingsmethoden bestaan in de adviesliteratuur over pretesten duidelijke aannames: de eerstgenoemde methoden worden als superieur beschouwd voor het ontdekken van taakgerelateerde problemen (selectie, begrip en toepassing), terwijl de laatstgenoemde methoden een breder scala aan problemen aan het licht kunnen brengen. Deze aannames krijgen enige bevestiging in de onderzoeksliteratuur, maar het onderzoek op dit punt is slechts matig ontwikkeld. Voor de keuze tussen tekstgebruiks- en tekstbeoordelingsmethoden is het belangrijk te weten welke soorten selectie-, begrips- en toepasssingsproblemen gemakkelijk onopgemerkt blijven in een evaluatie door zelfrapporterende proefpersonen, en hoe dat komt. Aan de andere kant is het zinnig om te onderzoeken hoe belangrijk de detectie van een breder scala aan lezersproblemen is voor de effectiviteit van verschillende tekstgenres.

Ook de tot nu toe beperkte resultaten over individuele versus groepsgewijze dataverzameling vragen om meer onderzoek. Omdat de huidige bevindingen zijn gebaseerd op de evaluatieresultaten met slechts één brochure, ligt replicatieonderzoek voor de hand, met andere persuasief georiënteerde brochures maar wellicht ook met informatieve en instructieve teksten.

In algemene zin is er meer onderzoek nodig naar de congruente validiteit van methoden om een beter zicht te krijgen op de effecten van methoden op de hoeveelheid en de aard van verzamelde feedback. Dat onderzoek kan uiteindelijk worden geplaatst in beschouwingen over typen methoden (zoals hierboven is gedaan), maar voor de evaluatiepraktijk zijn de vergelijkingen van specifieke methoden wellicht nog belangrijker. Om evaluatiedesigns beter af te kunnen stemmen op de doelen van een formatieve evaluatie is ook onderzoek naar de effecten van varianten van evaluatiemethoden van belang.

4. Discussie

Het in dit artikel besproken onderzoek maakt duidelijk dat tekstevaluatie een onderzoeksgebied in beweging is. In het overzichtsartikel van Schriver (1989) moest het overgrote deel van de citaties nog worden geleend uit andere disciplines; ruim tien jaar later is er in verschillende communicatiedisciplines een behoorlijk aantal methodologisch geo-

riënteerde publicaties verschenen waarin specifiek wordt ingegaan op de waarde en beperkingen van tekstevaluatiemethoden.

De lijst met relevante onderzoekspublicaties over tekstevaluatie was zelfs nog langer geweest als we dit overzicht niet hadden beperkt tot de validiteit van methoden. In aanvulling op het hier gepresenteerde onderzoek zijn er nieuwe onderzoekslijnen te zien met betrekking tot de betrouwbaarheid van pretestresultaten ('Bij hoeveel proefpersonen kunnen we spreken van een min of meer stabiele en uitputtende lijst van lezersproblemen?'), de verschillen in feedback bij verschillende steekproefsamenstellingen ('Geven hoger opgeleiden andere feedback op een tekst dan lager opgeleiden?') en de manier waarop evaluatieresultaten worden gebruikt bij de revisie van een tekst, waarbij fasen als de detectie van problemen, de diagnose, de inschatting van het belang van problemen en de vertaling van problemen in revisievoorstellen worden doorlopen. Voor een overzicht van dergelijke studies verwijzen we naar De Jong & Schellens (2000).

Ondanks het beschikbare onderzoek zijn er nog veel vragen onbeantwoord. En door de ontwikkeling van nieuwe methoden komen er ook steeds nieuwe vragen bij. Heel dringend is bijvoorbeeld het onderwerp van website-evaluatie geworden, waarbij bestaande methoden uit verschillende disciplines (zoals de communicatiekunde en de mens-computer interactie) moeten worden aangepast en uitgeprobeerd voor de evaluatie van informatie die via het Internet wordt aangeboden. Op het gebied van de congruente validiteit dienen zich hierbij nieuwe onderzoeksthema's aan, zoals de voorspellende waarde van een evaluatie van papieren prototypes en de verschillen en overeenkomsten tussen laboratorium- en real-life evaluatieonderzoek (De Jong & Heuvelman, 1999).

In de toekomst zullen we in het Twentse onderzoek de aandacht richten op enkele veelbelovende methoden voor tekst- en website-evaluatie waar slechts beperkt onderzoek naar verricht is. Bij expertgericht onderzoek denken we aan de functionele analyse, gebruikersscenario's en heuristieken. Deze drie benaderingen bieden experts verschillende vormen van ondersteuning: respectievelijk informatie over de intenties van de tekst, realistische gebruikersperspectieven op de geboden informatie en evaluatiecriteria op grond van beschikbare kennis over tekstontwerp. Het lijkt zinvol om in onderzoek na te gaan welke gevolgen deze benaderingen hebben voor de feedback die experts geven. In hoeverre worden expertbeoordelingen beïnvloed door dergelijke vormen van ondersteuning? Gaat het commentaar van experts meer lijken op de feedback die lezers uit de doelgroep geven? En welke vorm van expertevaluatie draagt het meest bij aan de effectiviteit van teksten? Bij het onderwerp heuristieken zien we overigens interessante aansluitingsmogelijkheden bij taalbeheersingsonderzoek naar globale versus analytische opstelbeoordeling.

Bij doelgroepgericht onderzoek denken we vooralsnog aan twee speerpunten, gericht op een typische tekstgebruiksmethode en een typische tekstbeoordelingsmethode. In de eerste plaats loopt er een promotieonderzoek naar de waarde en beperkingen van (varianten van) de hardop-denkmethode voor het evalueren van websites, brochures en handleidingen. Met name voor teksten met een instructieve functie lijkt de methode een voor de hand liggende keus, maar het ondersteunende onderzoek is tot nu toe uiterst beperkt. In de tweede plaats zullen we, in samenwerking met de Universiteit Utrecht, onderzoek gaan opzetten naar de validiteit van het computerprogramma Focus, dat bedoeld is om op efficiënte wijze oordelen van lezers op teksten te verzamelen. De uiteindelijke vraag is, net als in eerder onderzoek naar de plus-en-minmethode, in hoeverre de resultaten van Focus bij-

dragen aan de effectiviteit van teksten (en in de toekomst ook van websites). Het uiteindelijke doel van deze onderzoeksinspanningen is de communicatieprofessional te voorzien van een scala aan evaluatietechnieken waarvan de waarde in onderzoek is bewezen.

Noot

* Dit artikel is een bewerking van een eerder gepubliceerd overzicht (De Jong & Schellens, 2000). Daarin wordt een overzicht gegeven van de Engelstalige onderzoeksliteratuur over formatieve evaluatie. Behalve op onderzoek naar de validiteit van methoden gaan we in dat artikel ook in op onderzoek naar de samenstelling en omvang van steekproeven en naar de implementatie van pretestresultaten in de fase van revisie.

Bibliografie

Ahlschwede, M.P. (1970). *Pre-testing a publication for low-income homemakers.* N.C. State University, Dept. Agricultural Information, Raleigh, NC, Tech. Rep. no. 10.

Allwood C.M. & T. Kalén (1997). Evaluating and improving the usability of a user manual. *Behaviour & Information Technology*, 16, 43-57.

Baker, E.L. (1970). Generalizability of rules for empirical revision. *AV Communication Review*, 18, 300-305.

Baker, L. (1985). How do we know when we don't understand? Standards for evaluating text comprehension. In D.L. Forrest-Pressley, G.E. MacKinnon and T.G. Waller (Eds.), *Metacognition, Cognition, and Human Performance* (pp.155-206). New York: Academic Press.

Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review*, 1, 3-38.

Barnum, C.M. (2001). *Usability testing and research*. New York: Longman.

Bischoping, K. (1989). An evaluation of interviewer debriefing in survey pretests. In C. Cannell, L. Oksenberg, G. Kalton, K. Bischoping and F.J. Fowler (Eds.), *New Techniques for Pretesting Survey Questions* (pp.15-29). University of Michigan, Survey Research Center, Ann Arbor, MI, Tech. Rep. HS 05616.

Couzijn, M. (1995). *Observation of Writing and Reading Activities. Effects on Learning and Transfer.* Dissertatie Universiteit van Amsterdam.

Couzijn, M. & G. Rijlaarsdam (1998). Learning to write by reader observation and written feedback. In G. Rijlaarsdam, H. van den Bergh and M. Couzijn (Eds.). *Effective Teaching and Learning of Writing. Current Trends in Research*. (pp.224-252). Amsterdam: Amsterdam University Press.

Davidove, E.A., & R.A. Reiser (1991). Comparative acceptability and effectiveness of teacher-revised and designer-revised instruction. *Educational Technology Research & Development*, 39, 29-39.

Desurvire, H.W. (1994). Faster, cheaper!! Are usability inspection methods as effective as empirical testing? In J. Nielsen and R.L. Mack (Eds.). *Usability Inspection Methods* (pp.173-202). New York: John Wiley.

Diamantopoulos, A., N. Reynolds & B. Schlegelmilch (1994). Pretesting in questionnaire design: The impact of respondent characteristics on error detection. *Journal of the Marketing Research Society*, 36, 295-314.

Dieli, M. (1986). *Designing Successful Documents: An Investigation of Document Evaluation Methods*. Dissertation Carnegie Mellon University: Pittsburgh, PA.

Duffy, T.M. & P. Kabance (1982). Testing a readable writing approach to text revision. *Journal of Educational Psychology,* 74, 733-748.

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

Elling, R. (1997). Revising safety instructions with focus groups. *Journal of Business and Technical Communication*, 11, 451-468.

Faulkner, X. (2000). *Usability engineering*. New York: Palgrave.

Gillham, M., & K. Buckner (1997). User evaluation of hypermedia encyclopedias. *Journal of Educational Multimedia and Hypermedia*, 6, 77-90.

Golas, K.C. (1983). Formative evaluation effectiveness and cost: Alternative models for evaluating printed instructional materials. *Performance & Instruction Journal*, 22 (5), 17-19.

Gropper, G.L., & A.A. Lumsdaine (1961). *Studies in televised instruction: The use of student response to improve televised instruction: A summary report*. Metropolitan Pittsburgh Educational Television Stations WQED-WQEX/ American Institutes for Research, Pittsburgh, PA, Tech. Rep. no. 7.

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.

Hunt, S.D., R.D. Sparkman & J.B. Wilcox (1982). The pretest in survey research: Issues and preliminary findings. *Journal of Marketing Research*, 19, 269-273.

Jansen, C., S. Klatter & D. de Vet (1991). Formulierenonderzoek bij de Informatiseringsbank. *Communicatief*, 4, 189-204.

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.

John, B.E. & S.J. Marks (1997). Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology*, 16, 188-202.

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

Jong, M. de & A. Heuvelman (1999). De formatieve evaluatie van voorlichtingssites op het World Wide Web. Een inventarisatie van benaderingen. In A.A. van Ruler e.a. (Red.). *Jaarboek Onderzoek Communicatiemanagement 1999* (pp.117-135). Alphen aan den Rijn: Samsom.

Jong, M.D.T. de & L.R. Lentz (1996). Expert judgments versus reader feedback. A comparison of text evaluation techniques. *Journal of Technical Writing and Communication*, 26, 507-519.

Jong, M. de, & L. Lentz (2001). Focus: Design and evaluation of a software tool for collecting reader feedback. *Technical Communication Quarterly*, 10, 387-401.

Jong, M. de, & D. Rijnks (te verschijnen). Dynamics of iterative reader feedback. An analysis of two successive "plus-minus" evaluations (Artikel aangeboden ter publicatie.)

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 (1996). Interviews of groepsgesprekken? Een vergelijkend onderzoek naar twee varianten van de plus-en-minmethode. *Taalbeheersing*, 18, 339-350.

Jong, M. de, & P.J. Schellens (1997). Reader-focused text evaluation: An overview of goals and methods. *Journal of Business and Technical Communication*, 11, 402-432.

Jong M. de, & P.J. Schellens (1998). Focus groups or individual interviews? A comparison of text evaluation approaches, *Technical Communication*, 45, 77-88.

Jong, M. de, & P.J. Schellens (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.

Jong, M. de, & P.J. Schellens (2001a). Readers' background characteristics and their feedback on documents: The influence of gender and educational level on evaluation results. *Journal of Technical Writing and Communication*, 31, 267-281.

Jong, M. de, & P.J. Schellens (2001b). Optimizing public information brochures. Formative evaluation in document design processes. In: D. Janssen & R. Neutelings (Eds.), *Reading and writing public documents* (pp.59-83). Amsterdam: John Benjamins.

Kandaswamy, S. H.D. Stolovitch & S. Thiagarajan (1976). Learner verification and revision: An experimental comparison of two methods. *AV Communication Review*, 24, 316-328.

Lavery, D., G. Cockton & M.P. Atkinson (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology,* 16, 246-266.

Lentz, L. & M. de Jong (1996). Argumenteren over lezersproblemen: Is consensus haalbaar? *Taalbeheersing*, 18, 351-367.

Lentz, L. & M. de Jong (1997). The evaluation of text quality: Expert-focused and reader-focused methods compared. *IEEE Transactions on Professional Communication*, 40, 224-234.

Lentz, L., & M. de Jong (2000). Hoger en lager opgeleiden en hun feedback op een voorlichtingstekst. In: R. Neutelings, N. Ummelen & A. Maes (Red.), *Over de grenzen van de taalbeheersing. Onderzoek naar taal, tekst en communicatie* (pp.317-325). Den Haag: Sdu.

Lentz, L. & H. Pander Maat (1992). Evaluating text quality: Reader-focused or text-focused? In H. Pander Maat & M. Steehouder (Eds.). *Studies of Functional Text Quality* (pp.101-114). Atlanta, GA: Rodopi.

Lentz, L., & H. Pander Maat (1993). *Wat mankeert er aan die tekst? De evaluatie van voorlichtingsteksten over subsidieregelingen*. Amsterdam: Thesis.

Lewis, J.R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors*, 36, 369-378.

Lindgaard, G. (1994). *Usability Testing and System Evaluation: A Guide for Designing Useful Computer Systems*. London: Chapman & Hall.

Mack, R. & F. Montaniz (1994). Observing, predicting, and analyzing usability problems. In J. Nielsen & R.L. Mack (Eds.). *Usability Inspection Methods* (pp.195-339). New York: John Wiley.

Medley-Mark, V. & C.B. Weston (1988). A comparison of student feedback obtained from three methods of formative evaluation of instructional materials. *Instructional Science*, 17, 3-27.

Micklos, D. & W. Bishop (1982). *The measurement of redundancy and its effects in science communication*. Paper presented at the Annual Meeting of the Association for Education in Journalism and Mass Communication, Athens, OH, July 25-28, 1982.

Molich, R., A. Damgaard Thomsen, B. Karyukina, L. Schmidt, M. Ede, W. van Oel & M. Arcuri (1999). *Comparative evaluation of usability tests.* [http://www.dialogdesign.dk/cue.htm].

Nathenson, M.B. & E.S. Henderson (1980). *Using Student Feedback to Improve Learning Materials*. London: Croom.

Nielsen, J. (1993). *Usability Engineering*. New York: Academic Press.

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R.L. Mack (Eds.). *Usability Inspection Methods* (pp.25-62). New York: John Wiley.

Oksenberg, L., C. Cannell & G. Kalton (1991). New strategies for pretesting survey questions. *Journal of Official Statistics*, 7, 349-365.

Pander Maat, H. (1996). Identifying and predicting reader problems in drug information texts. In T. Ensink & C. Sauer (Eds.). *Researching Technical Documents* (pp.17-48). Groningen: University of Groningen, Dept. of Speech and Communication.

Presser, S. & J. Blair (1994). Survey pretesting: Do different methods produce different results? *Sociological Methodology*, 24, 73-104.

Renkema, J. (1994). *Taal mag geen belasting zijn. Een onderzoek-in-burger naar brieven van ambtenaren*. Den Haag: Sdu.

Renkema, J. (2000). Pretesten testen: De CCC-analyse en de beperkte plus-en-minmethode vergeleken. In: R. Neutelings, N. Ummelen & A. Maes (Red.), *Over de grenzen van de taalbeheersing. Onderzoek naar taal, tekst*

en communicatie (pp.371-380). Den Haag: Sdu.

Renkema, J., & P.J. Schellens (Red.) (1996). Themanummer Tekstevaluatie. *Taalbeheersing,* 18 (4), 305-382.

Renkema, J., & M. Wijnstekers (1997). Doelgroep-onderzoek of bureau-analyse? In H. van den Bergh, D. Janssen, N. Bertens & M. Damen (Red.), *Taalgebruik ontrafeld* (pp.365-373). Dordrecht: Foris.

Reynolds, N. & A. Diamantopoulos (1998). The effects of pretest methods on error detection rates: Experimental evidence. *European Journal of Marketing,* 32, 480-498.

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

Sanders, T., J. Sanders, J. Renkema & C. van Wijk (1994). Praktijkonderzoek naar tekstkwaliteit. Een standaardvoorbeeld. *Communicatief*, 7 (4), 13-22.

Schellens, P.J., & A.A. Maes (2000). Tekstontwerp. In: A. Braet (Red.), *Taalbeheersing als communicatiewetenschap. Een overzicht van theorievorming, onderzoek en toepassingen* (pp.154-188). Bussum: Coutinho.

Schriver, K.A. (1989). Evaluating text quality: The continuum from text-focused to reader-focused methods. *IEEE Transactions on Professional Communication*, 32, 238-255.

Schriver, K.A. (1992). Teaching writers to anticipate readers' needs. A classroom-evaluated pedagogy. *Written Communication*, 9, 179-208.

Schriver, K.A. (1997). *Dynamics in document design. Creating text for readers*. New York: John Wiley.

Schuurs, U., H. van den Bergh & G. Verhoeven (1991). De invloed van pretestinstructies op de bruikbaarheid van de ontlokte oordelen. In: M.M.H. Bax & W. Vuijk (Red.), *Thema's in de taalbeheersing* (pp.370-379). Dordrecht: ICG.

Sienot, M. (1997). Pretesting web sites. A comparison between the plus-minus method and the think-aloud method for the World Wide Web. *Journal of Business and Technical Communication*, 11, 469-482.

Swaney, J.H., C.J. Janik, S.J. Bond & J.R. Hayes (1992). Editing for comprehension: Improving the process through reading protocols. In E.R. Steinberg (Ed.). *Plain language: Principles and practice* (pp.173-203). Detroit, MI: Wayne State University Press.

Velotta, C. (Ed.) (1995). *Practical approaches to usability testing for technical documentation*. Arlington, VA: STC.

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

Vroom, B. (1987). Publieksonderzoek met behulp van de plus-en-min methode. *Tijdschrift voor Taalbeheersing*,9, 256-271.

Wager, J.C. (1983). One-to-one and small-group formative evaluation: An examination of two basic formative evaluation strategies. *Performance & Instruction Journal*, 22 (5), 5-7.

Weston, C.B. (1987). The importance of involving experts and learners in formative evaluation strategies. *Canadian Journal of Educational Communication*, 16(1), 45-58.

Weston, C., C. Le Maistre, L. McAlpine & T. Bordonaro (1997). The influence of participants in formative evaluation on the improvement of learning from written instructional materials. *Instructional Science*, 25, 369-386.

Wright, A.D. (1994). The value of usability testing in document design. *The Bulletin of the Association for Business Communication*, 57 (1), 48-51.

Appendix: Omschrijving van genoemde evaluatiemethoden

Begripstest: Deelnemers krijgen tekstbegripsvragen over de tekst die zij gelezen hebben. Het resultaat is een totaalindruk van de begrijpelijkheid van de tekst, maar daarnaast kan een tekstbegripstest helpen om specifieke passages aan te wijzen waar de lezers begripsproblemen hebben.

CCC-evaluatie: Experts evalueren een tekst aan de hand van een matrix van drie criteria (correspondentie, consistentie en correctheid) en vijf tekstniveaus (teksttype, inhoud, opbouw, formulering en presentatie). Een evaluatie aan de hand van het CCC-model kan worden gezien als een vorm van heuristische evaluatie.

Eén-op-één evaluatie van lesmateriaal: Individuele leerlingen worden geobserveerd terwijl ze werken met lesmateriaal. De evaluator overlegt daarbij met de deelnemers over hun ervaringen en hun suggesties om het lesmateriaal te verbeteren.

Focus: Een computerprogramma waarmee lezers uit de doelgroep problemen kunnen signaleren in teksten. Voor iedere probleemdetectie moeten deelnemers met de muis het betreffende tekstfragment selecteren, een probleemcategorie kiezen en een probleemomschrijving intypen.

Focusgroepen: Lezers uit de doelgroep bespreken hun oordelen over een tekst in een groepsdiscussie (bijvoorbeeld van 4 tot 6 personen) met een gespreksleider. Om de aandacht van de deelnemers op de tekst te blijven richten, is een combinatie met de plus-en-minmethode handig: de deelnemers zetten individueel hun plussen en minnen, en bespreken hun ervaringen vervolgens groepsgewijs.

Functionele analyse: Een vorm van expertevaluatie aan de hand van drie opeenvolgende stappen: (1) een inventarisatie van relevante doelgroepsegmenten, (2) de vaststelling van relevante tekstfuncties voor deze doelgroepsegmenten, en (3) de koppeling van evaluatiecriteria aan de onderscheiden tekstfuncties. Voor teksten met gelijke functies kan een generieke "functionele analyse checklist" worden opgesteld.

Hardop-denkprotocollen: Lezers uit de doelgroep krijgen de taak om een tekst te gebruiken en daarbij hardop te lezen en hun gedachten te verbaliseren. Bij instructieve teksten krijgen de deelnemers realistische taken, en richt de methode zich met name op de problemen bij de selectie en de toepassing van de geboden informatie (in dit geval wordt ook gesproken van de 'hardop-werkmethode'). Hardop-denkprotocollen worden echter ook zonder taken gebruikt om teksten te evalueren.

Heuristische evaluatie: Experts evalueren een tekst aan de hand van een lijst met aandachtspunten, richtlijnen of evaluatiecriteria. Deze heuristieken zijn gebaseerd op onderzoeksresultaten en/of op praktijkkennis. Met name voor de evaluatie van websites is inmiddels een enorme diversiteit aan evaluatieheuristieken beschikbaar.

Plus-en-minmethode: Lezers uit de doelgroep krijgen de opdracht om een tekst te lezen en plussen en minnen in de kantlijn te zetten. In een interview worden vervolgens de motieven voor de gezette plussen en minnen achterhaald.

Retrospectieve hardop-denkprotocollen: Lezers uit de doelgroep krijgen een aantal realistische taken die ze met een tekst moeten verrichten. Hun verrichtingen worden op video opgenomen. Vervolgens wordt de video-opname afgespeeld, en krijgen de deelnemers de vraag om de gedachten en motieven die zij tijdens de taakuitvoering hadden te verwoorden.

Scenario-evaluatie: Experts evalueren een tekst aan de hand van een aantal realistische gebruikersscenario's. Zij volgen de route van gebruikers in verschillende situaties door een tekst of website, en inventariseren de mogelijke problemen die zij tegenkomen bij met name de selectie en de toepassing van informatie. De experts nemen dus de rol aan van surrogaat-gebruikers. Een verwante evaluatiebenadering in de mens-computerinteractie is de 'cognitive walkthrough'.

Signaled stopping technique: Lezers uit de doelgroep krijgen de taak om een tekst te lezen en om eventuele leesonderbrekingen te registreren. In een interview worden de redenen voor de onderbrekingen achterhaald.