NTNU
Det skapende universitet

En sammenligning av RTA og CTA:
testsituasjonens innvirkning på resultatet

Øyvind Aarøe

Master i informatikk
Oppgaven levert: Mai 2007
Hovedveileder: Terje Rydland, IDI

Norges teknisk-naturvitenskapelige universitet
Institutt for datateknikk og informasjonsvitenskap

Forord

Denne avhandlingen er resultatet av et forskningsstudie i emnet ”IT3900 Masteroppgaven, Informatikkstudiet” ved NTNU fra høsten 2005 til våren 2007. Oppgaven ble definert i samarbeid med konsulentbedriften Kantega.

Kontaktperson hos Kantega har vært Ingunn Moen. Jeg vil takke henne for bra veiledning underveis i prosjektet og gode tilbakemeldinger på henvendelser under studiet.

Jeg vil også rette en stor takk til masteroppgavens veileder Terje Rydland. Takk for rettledning og konstruktiv kritikk.

Til slutt vil jeg takke den personen som har betydd mest for meg gjennom hele oppgavens utvikling. Anne Tronhus, kjæreste og samboer, takk for all støtte og motivasjon du har gitt meg.

Trondheim 22. Mai 2007

Øyvind Aarøe

Oppgavetekst

Brukbarhetstesting er en kjent evalueringsmetodene innen brukskvalitet (eng. usability). Den mest anvendte brukbarhetstestmetoden heter concurrent think aloud (CTA), som også blir omtalt som den tradisjonelle brukbarhetstestmetoden Under en brukbarhetstest skal testdeltagerne gjennomføre ett sett med oppgaver de får utdelt. Oppgavene skal løses mens de blir observert, filmet og i tillegg skal de verbalisere sine tanker (si høyt hva de tenker). Meningen med å verbalisere tanker er at testpersonellet skal forstå deltagernes tankegang. Hvis deltagerne glemmer å snakke høyt, vil de få en påminnelse fra testpersonellet. RTA er en alternativ, men mindre kjent metode. Denne metoden krever verken at testpersonene skal filmes eller at de skal verbalisere sine tanker under oppgaveløsningen. Verbaliseringen foregår først etter oppgaveløsningen ved å se på et opptak fra oppgaveløsningen (skjermopptak). Denne masteroppgaven forsker på om testsituasjonen har innvirkning på resultatene. Resultatene i denne sammenhengen er brukskvalitetsproblemer som er avdekket av de anvendte brukbarhetstestmetodene, CTA og RTA. Testsituasjonen kan være alt fra overvåkningseffekt, kunstige omgivelser, teknikker som brukes eller for eksempel personell som gjennomfører testen.

Sammendrag

CTA har vært utsatt for sterkt kritikk av eksperter. Dette er kritikk som spesielt fokuserer på metodens validitet og gjennomførelse. Denne masteroppgaven fokuserer på brukbarhetstester med et annet perspektiv; brukernes opplevelse av metoden. Oppgaven prøver finne ut om, og eventuelt i hvor stor grad, testsituasjonen påvirker testresultatene.

I dette studiet gjennomføres til sammen 20 brukbarhetstester. Testene ble gjennomført i to runder på et nettsted for ”Studentsamskipnaden i Trondheim". Studiet viser at RTA og CTA oppleves forskjellig av testbrukerne, og at opplevelsen av CTA har innvirkning på resultatene. CTA- testdeltagerne brukte lengre tid på å gjennomføre oppgavene og metoden avdekket i tillegg falske brukskvalitetsproblem. Alle testdeltagerne kom med flere negative tilbakemeldinger på CTA, mens det stort sett var positive tilbakemeldinger på RTA metoden. Den negative kritikken av CTA var gjennomgående temaer som; mislikte å bli filmet, dobbel kognitiv belasting, vanskelig å sette ord på tanker, prestasjonsangst, press til å snakke, dårlig tid til oppgaveløsning, vanskelig å holde verbaliseringen gående og tenker annerledes enn de snakker.

Innhold

Figurer

Tabeller

Kapittel 1 Introduksjon

1.1 Bakgrunn

Brukbarhetstesting har vært ett hett diskusjonstema i miljøer innen brukskvalitet siden de tidlige 90-årene. *Brukbarhetstesen* er den mest brukte evalueringsmetoden for å oppnå innsikt i hvordan mennesker samhandler med produkter eller brukergrensesnitt. Det finnes mye litteratur på hvordan teknikken skal nyttes, og det eksisterer forskjellige varianter av metoden. Brukbarhetstester utføres som regel i laboratorier, i felten eller i workshops. De mest typiske teknikkene som kombineres i en brukbarhetstest er: Think aloud, observasjon, videoopptak, automatisk logging av markør eller tastetrykk og spørreskjema. Det er en av disse teknikkene som har fått mye oppmerksomhet: *think aloud* også kalt TA. TA er den mest populære teknikken i MMI praktisering (Nielsen, Clemmensen & Yssing 2002)

CTA (concurrent think aloud) er en variant av TA og brukes under de fleste brukbarhetstester. CTA omtales også ofte som selve brukbarhetstestsmetoden (Nielsen 1992). Målet med CTA er å avdekke flest mulig problemer knyttet til brukergrensesnittet. Dette ønsker man å oppnå ved å få direkte tilgang til menneskers mentale prosesser.

CTA har blitt positivt omtalt hos enkelte innen fagmiljøet, som for eksempel Jacob Nielsen som har uttrykt følgende; ”Thinking aloud may be the single most valuable usability engineering method” (Nielsen 1993: 195)

Flere forskere og studier har påpekt svakheter ved CTA (Preece, Rogers & Sharp 2002, Van den Haak & De Jong 2003). Dette er svakheter som dobbel kognitiv belasting, stillhet, kunstige omgivelser og manglende retningslinjer for å gjennomføre testen. Retrospective think aloud (RTA) prøver å komme rundt CTAs svakheter. RTA er mindre utbredd enn CTA, men ble i 2006 erklært gyldig og pålitelig av Guan et al. (2006). CTA-metoden er i skrivende stund under utforskning angående dens validitet og reliabilitet (Ramey et al. 2006).

Det finnes mange guider på hvordan man skal utføre en brukbarhetstest, men mer forskning på området er etterspurt. ”Think aloud testing is a widely employed usability evaluation method, yet its use in practice is rarely studied” (Hornbæk & Nørgaard 2006: 209; Nielsen, Clemmensen og Yssing (2002) påpeker mangel av forskningslitteratur av brukerens anvendelse av teknikken. I denne oppgaven blir det fokusert på hvordan brukerne opplever å bli testet med den tradisjonelle CTA-metoden og sammenligning med brukernes opplevelse av RTA-metoden. Det blir også undersøkt om brukernes opplevelse av metoden kan ha innvirkning på resultatet av metodene.

1.1.1. Begrensninger

- Det vil kun bli forsket på brukbarhetstester av et nettsted. Resultatene kan vært annerledes ved testing på andre produkter.
- Alle testene er gjennomført i laboratorier.
- Teknikkene som brukes under CTA-gjennomgang er observasjon, videoopptak og concurrent think aloud (se kapittel 2)
- Teknikkene som brukes under RTA-gjennomgang er opptak av skjerm (under oppgaveløsning), observasjon, lydopptak og retrospectice think aloud
- Det vil kun bli utført tester med en deltager i hver gjennomføring

1.1.2. Forskningsspørsmål

Forskningsspørsmålet som blir stilt i denne masteroppgaven, er; ”i hvilken grad påvirkes resultatene av testsituasjonen av de to brukbarhetstestmetodene CTA og RTA?”. Testresultatene i denne sammenhengen er brukskvalitetsproblemer. Testsituasjonen kan være alt fra overvåkningseffekt, kunstige omgivelser, teknikker som brukes, eller for eksempel personell som gjennomfører testen. Forskningsspørsmålet er brutt ned til tre delspørsmål:

- Hvilke problem innen brukskvaliteten vil de to metodene CTA og RTA avdekke innen de begrensningene som er beskrevet ovenfor?

- Hvordan opplever testbrukerne metodene RTA og CTA?

- Vil opplevelsen av metodene ha innvirkning på resultatet?

1.1.3. Leserens guide

Oppgaven er bygd opp som følgende kapitler:

Kapittel 2: En innføring i brukbarhetstester

Først en liten forklaring av termer som blir brukt etterfulgt av en mer generell introduksjon til evaluering av brukergrensesnitt. Etter hvert fokuseres det i større grad mot brukbarhetstester. Det vil bli gitt en detaljert utreding om brukbarhetstester, etterfulgt av teorier og praktisering rundt TA-teknikkene/metodene CTA og RTA. Det vil også bli gitt en kort beskrivelse av alternative metoder

Kapittel 3: Forskningsdesign

Dette kapittelet omhandler valg av forskningsstrategi og vil kartlegge tilnærming av oppgaven. Her vil forskningsmetodene og analyse av dataene bli forklart i detalj.

Kapittel 4: Forstudie

Hensikten med forstudiet var å bli bedre kjent med fagområdet. Her ble det gjennomført åtte brukbarhetstester. Fire av testene ble gjennomført med CTA-metoden, og de resterende testene ble gjennomført med en metode kalt *cooperative usability testing* (CUT). Dette forstudiet var med på å forme forskningsspørsmålet.

Kapittel 5: En sammenligning mellom CTA og RTA

I dette kapittelet behandles ytterligere tolv brukbarhetstester. Dette er en rapportering av brukbarhetstestene som ble utført. Her konkretiseres planlegging og utførelse av brukbarhetstestene. Brukskvalitetsproblemene som dukket opp er rapportert og analysert. Til slutt i dette kapittelet fokuseres det på resultatforskjeller mellom RTA og CTA

Kapittel 6: Diskusjon

Diskusjonskapittelet inneholder en fremstilling av intervjuene basert på deltagernes opplevelse av metoden. Her diskuteres del 2 og 3 av forskningsspørsmålet.

Kapittel 7: Avslutning

Masteroppgaven avslutter med å fokusere på hva forskningen har resultert i, og viser til potensielle forskningsområder for framtiden.

Kapittel 2 Teori

2.1 En innføring i brukbarhetstesting

Brukbarhetstesting er en viktig og nyttig metode for å avdekke problemer innen brukskvalitet. Avhandlingen vil starte med å gi noen konkrete definisjoner og forklaringer på terminologier innen fagområdet, etterfulgt av beskrivelser av metoder og teknikker innen evaluering av brukskvalitet. Avhandlingen vil etter hvert fokusere i større grad på brukbarhetstesting.

2.1.1 Terminologi

Brukskvalitet

En kommer ikke utenom brukbarhetstesting uten å introdusere begrepet brukskvalitet. De fleste i dag kjenner til ordet *brukervennlighet*, og dets betydning. Begrepet kan være litt misvisende, siden de fleste av oss ikke trenger maskiner som er vennlige, men i stedet hjelper til med å få jobben gjort på best mulig måte. Profesjonelle innen design av brukergrensesnitt benytter blant annet andre termer som *CHI* (Computer Human Interaction), *HCI* (Human Computer Interaction), *UCD* (User-Centerd Design), *MMI* (Man-Machine Interface), *UID* (User Interface Design) eller *HF* (Human Factors). Men den mest veletablerte termen i internasjonal sammenheng er *usability*.

I 1994 ble Norsk Språkråd spurt av Den Norske Dataforening om hvilket norsk ord man skulle benytte for usability (Den Norske Dataforening, 1994). Etter å ha undersøkt begrepene i andre bransjer og andre nordiske land, anbefalte språkrådet begrepet brukskvalitet fremfor brukbarhet, brukervennlighet og anvendbarhet. Begrepet brukskvalitet var allerede et godt innarbeidet ord på norsk innen bl.a. møbeldesign og industridesign. Til tross for at brukskvalitet er en god oversettelse, blir den lite brukt i forhold til oversettelsene *brukervennlighet* og *brukbarhet* innen software. Denne avhandlingen følger Norsk Språkråds råd, og benytter konsekvent termen brukskvalitet. *Evaluering av brukskvaliteten* beskriver prosessene som forsøker å klarlegge brukskvaliteten (Stone et al. 2005).

Andre terminologier innen brukskvalitet

Brukbarhetstesting er en fellesbetegnelse på enkelte metoder som evaluerer brukskvaliteten. Man involverer brukerne for å teste brukergrensesnittet med mål for å finne problemnivået av brukskvaliteten (Dumas & Redish 1999, Nielsen 1993, Preece, Rogers & Sharp 2002). Engelske termer for brukbarhetstesting er *usability testing* eller *user testing*. I Norge brukes også den direkte oversatte termen *brukertesting*, men kan lett gi assosiasjoner om at det er brukeren som skal testes. Hensikten er å teste brukergrensesnittet, og ikke brukerne. Brukbarhetstesting er ikke en konkret metode med faste retningslinjer, men en generell metode som har mange varianter. *Concurrent think aloud* (CTA) er den mest kjente varianten, men det er uenighet om hvilke retningslinjer man skal følge. (Nielsen, Clemmensen & Yssing 2002). *Retrospective think aloud* (RTA) er en mindre kjent metode som har likheter med CTA, men skiller seg ut på vesentlige områder. CTA vil også omtales som *sanntids TA* og RTA omtales også som retrospektiv TA. Begge metodene beskrives i detalj dette kapittelet.

Think aloud protocol, think aloud technique eller *think aloud method,* som den også kalles, brukes for å samle data under brukbarhetstestingen ved at deltagerne verbaliserer sine tanker. Denne teknikken brukes på flere fagområder som psykologi, sosialvitenskap, produkt design og ikke minst i den brukersentrerte delen av systemutviklingsprosessene (Preece, Rogers & Sharp 2002).

Testdeltagere eller testbrukere er personer som deltar i en brukbarhetstest. De skal samhandle med produktet som skal testes og kan i enkelte brukbarhetstester inspisere brukergrensesnittet for å gi feedback angående brukskvaliteten.

Tilretteleggeren (facilitator) har som ansvar å guide testdeltageren gjennom brukbarhetstesten. Hvordan dette skal gjennomføres er avhengig av metode eller teknikk. En *observatør* observerer testdeltageren, og noterer testdeltagerens kommentarer eller problemer med brukskvaliteten. En observatør kan enten observere under selve oppgaveløsningen eller etterpå når man går igjennom videoopptak eller logg av oppgaveløsningen. Hvis det er bare en person som gjennomfører brukbarhetstester på deltagere, spiller denne personen både rollen som observatør og tilrettelegger.

2.1.2 Definisjon av brukskvalitet

I litteraturen finner vi at *International Organization of Standardization* (ISO) ofte refereres når man snakker om brukskvalitet og brukermedvirkning. ISO definer brukskvalitet som *"*…the extent to which a product can be used by **specified users** to achieve **specific goals** with effectiveness, efficiency and satisfaction in a **specified context** of use" (ISO 9241-11 1998), hvor *effectiveness, efficiency and satisfaction* betyr;

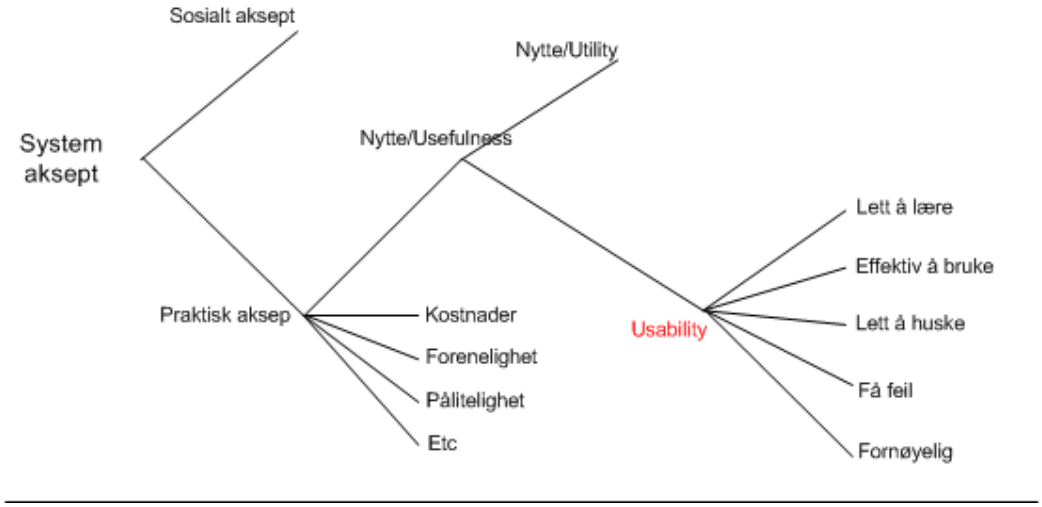
"Effectiveness: The accuracy and completeness with which specified users can achieve specified goals in particular environments.

Efficiency: The resources expended in relation to the accuracy and completeness of goals achieved.

Satisfaction: The comfort and acceptability of the work system to its users and other people affected by its use."

I boka "Usability Engineering" (Nielsen 1993) finnes en utvidet definisjon av brukskvalitet. Han understreker at brukskvalitet ikke er et enkelt endimensjonalt mål på et brukergrensesnitt, men at det omfavner flere aspekter ved et system. Figur 1.1 viser en oversikt av hvilke faktorer som påvirker brukernes opplevelse av et systemet, der faktorene bestemmer hvordan systemet vil aksepteres av brukerne. Brukere vil ha forskjellige verdier knyttet til de ulike faktorene. Basert på Jacob Nielsens fem kjente nøkkelfaktorer er brukskvalitet en kvantitativ og kvalitativ måling av designet til et brukergrensesnitt. De fem nøkkelfaktorer er:

Lett å lære, så brukere kan gå raskt fra å ikke kjenne systemet til å gjøre noe arbeid.
Effektivt, lar ekspertbrukeren oppnå en høy grad av produktivitet.
Lett å huske, så brukere med lav brukshyppighet kan returnere etter en periode med inaktivitet uten å måtte lære alt på nytt.
Relativt feilfritt og feiltolerant, slik at brukere ikke gjør mange feil, og at disse feilene ikke er katastrofale (og at man lett kan ta seg inn igjen)
Behagelig å bruke, tilfredsstiller brukerne subjektivt, slik at de liker å bruke systemet.



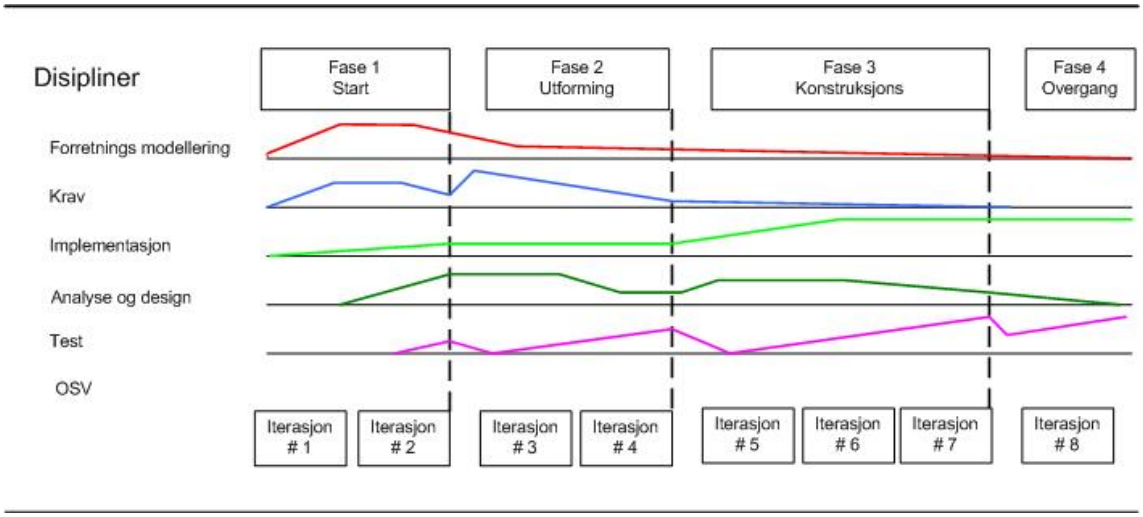
Figur 2.1: Modell av faktorer som påvirker brukernes aksept av systemet

Både definisjonen ISO-9241 og Nielsens fem kriterier definerer brukskvalitet, men med litt forskjellig vektlegging og perspektiv. Nielsen har en mer pragmatisk innfallsvinkel til brukskvalitet og har helt konkrete momenter som definerer brukskvalitet. ISO 9241-11 forklarer fordelene ved å måle brukskvalitet i termer av brukerutførelse og tilfredsstillelse, og den inkluderer retningslinjer på hvordan brukskvaliteten av ett produkt kan spesifiseres og evalueres. ISO 9241-11 dekker ikke systemutviklingsprosessen; menneskesentrerte designprosesser for interaktive systemer er beskrevet i ISO 13407 (ISO 13407 1999).

2.2 Evaluering av brukskvalitet

Evaluering av brukskvalitet er prosessene hvor en prøver å oppdage hvor enkelt et system er å bruke, om det er problemer som er knyttet til brukskvaliteten, og hvor disse problemene ligger. Det finnes mange metoder og teknikker for å evaluere brukskvalitet, og de fleste bedrifter (fra store til små) burde kunne finne en passende metode for sitt produkt, brukergruppe og budsjett. ”Usability engineering is not a one-shot affair where the user interface is fixed up before the release of a product. Rather, usability engineering is a set of activities that ideally take place throughout the lifecycle of the product…” (Nielsen 1993: 71). Evaluering av brukskvalitet skal altså helst gjennomføres iterativt, brukskvalitet bør være i fokus tidlig i en utviklingsfase, og aller helst skal man gjøre mest mulig før man startes på selve designet. Nielsen påpeker at avgjørelser som involverer design gir ringvirkninger for resten av utviklingen. Går det lang tid før man oppdager svakheter knyttet mot

brukskvaliteten, kan dette resultere i store og unødvendige kostnader. Ved å evaluere design i forskjellige faser under utviklingen, kan man i større grad forsikre seg om at produktet møter brukernes krav og ønsker. Dette gjennomføres iterativt i flere runder hvor man evaluerer designet og eventuelt kommer med forslag til redesign (Genov 2005, Nielsen 1993, ISO 13407 1999).”Iterative design, with its repeating cycle of design and testing, is the only validated methodology in existence that will consistently produce success results. If you don’t have user-testing as an integral part of your design process, you are going to throw buckets of money down the drain” (Tognazzi 2000). Rational Unified Process (RUP) er et rammeverkt til en iterative utviklingsprosess som er populær for utvikling av it systemer (utviklingsprosesser er stadig under utvikling, så RUP kan utvikles eller eventuelt bli mindre populær over tid). RUP er delt inn i fire faser, hvor hver fase inneholder ett sett med disipliner (se figur 1.2). Disiplinene er de samme for hver fase, men kan ha ulik påvirkning for hver av fasene. I en slik modell bør brukskvalitet integreres i integreres i hensiktsmessige disiplinene for hver fase. Hver fase består av flere iterasjoner eller sprinter.



Figur 2.2: Forenklet modell av RUP

2.2.1 Holdninger

En vanlig holdning blant designere i dag er at om designerene selv og deres kollegaer kan bruke programvaren og synes den er god, vil også andre synes det (Preece, Rogers & Sharp 2002). Men er utviklere forskjellige fra de som skal bruke systemet? Tognazzini skriver om Jungs typer i boken Tog on Interface (1991). Han skriver, med rot i undersøkelser gjort av Katharine Cook Briggs og Isabel Briggs Meyer, at brukere og utviklere har forskjellige psykologiske typer. Tognazzini påpeker at brukerne oftere reagerer og handler på bakgrunn av

det visuelle, mens utviklerne tenker oftere mer intuitivt og vant til å tenke mer abstrakt. Dette er selvsagt forskjellig fra person til person og forekommer i varierende grad. Poenget er at utviklerne må kjenne brukerne, og det er vanskelig å kjenne brukerne uten å la brukerne delta i utviklingsprosessen. Brukskvalitet skal innebære designprosesser som inkluderer mennesker (ISO 13407: 1999).

Mange foretrekker å utelukke evaluering fordi det vil forlenge utviklingstiden og fordi evaluering koster penger (Nielsen 1993). Problemet er at hvis man utelukker evalueringen kan man ikke være sikker på om programvaren er verken brukbar eller hva brukerne ønsker, og det hele kan bli ganske dyrt. Har man kommet langt i utviklingen før man oppdager en feil fra ett tidlig designvalg, kan det ende opp med å bli svært ressurskrevende å rette på feilen.

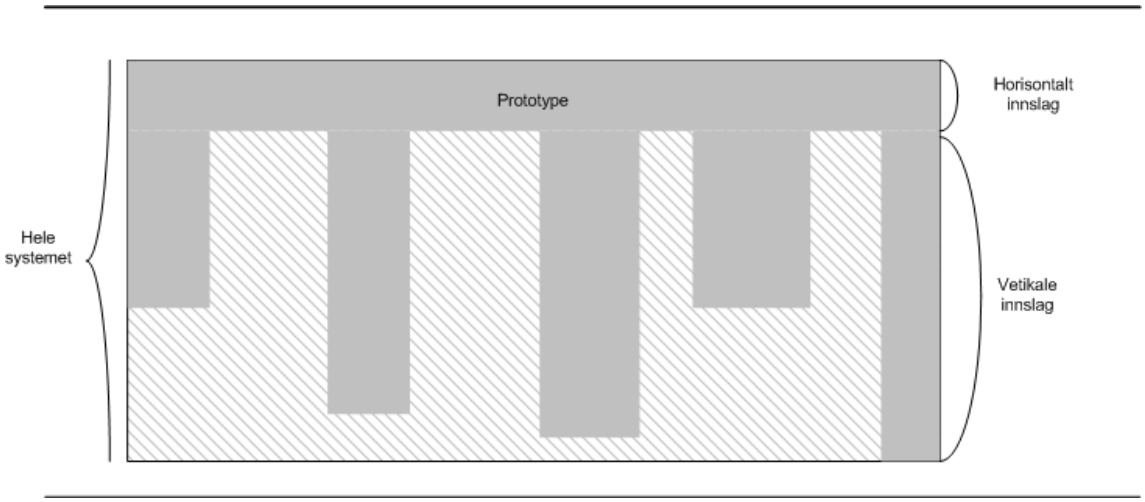
Vi har i dag flere typer evalueringsteknikker for å kartlegge brukskvaliteten. Preece, Rogers og Sharp (2002) definerer evaluering som prosessen av å systematisk samle data, som informerer oss om hva det er for en spesiell bruker eller gruppe av brukere, å bruke et produkt for en spesiell oppgave i bestemte omgivelser. Men for å evaluere brukergrensesnittet må man ha noe å evaluere. Dette kan være i form av tekst (for eksempel use-cases av brukergrensesnittet), prototyper eller det endelige produktet, og er avhengig av hvor man er i utviklingsprosessen.

2.2.2 Prototyper

Ved bruk av simuleringer, modeller, papirprototyper eller softwareprototyper, har man noe konkret visuelt å gå ut etter. Designavgjørelser blir mer eksplisitte, noe som muliggjør at medlemmene av designteamet kommuniserer med hverandre tidlig i prosessen. Dette tillater designere å utforske flere designkonsepter før man bestemmer seg for en. En prototype kan også fungere som en modell hvor man konkretiserer sine ideer til brukerne, og gjør det mulig å innlemme brukerfeedback av designet tidlig i utviklingen. Dette er med på å redusere risikoen slik at man forandre på produktet før man går videre i utviklingen (ISO 13407 1999):

I stedet for å starte med en fullskalaproduksjon basert på tidlige brukergrensesnittdesign, burde evalueringer av brukskvaliteten baseres på prototyper av det endelige systemet. Prototyper kan utvikles raskt og billig, og muliggjør evaluering av flere iterasjoner, slik at designet kan forandres inntil man får en bedre forståelse av brukergrensesnittet. Dette forbedrer kvaliteten og fullstendigheten av funksjonelle designspesifikasjoner.

I tillegg til fordelene nevnt over, er ideen bak å benytte prototyper å spare tid og kostnader. Ved å planlegge designet og i tillegg teste dette på brukerne, øker sannsynligheten på å utvikle noe kunden vil ha. Dette muliggjør involvering av brukere på ett tidlig steg i designfasen gjennom f eks statiske papirbaserte prototyper, og er et krav for å innfri ISO 13407 (ISO 13407 1999). En prototype er et forslag eller forsøk på en delvis speiling av det endelige systemet. Dette kan gjøres på flere måter (Nielsen 1993), men vi skiller mellom horisontale- og vertikale prototyper. En vertikal prototype tester en begrenset del av det endelige systemet, men går i dybden på realistiske omstendigheter med funksjoner som kan brukes for å løse virkelige oppgaver. Den horisontale prototypen inkluderer alle momenter av brukergrensesnittet, men her mangler underliggende funksjonalitet. Sistenevnte er en simulering av grensesnittet men man kan ikke utføre noe arbeid på dette. Ofte lages prototyper som en kombinasjon av disse.



Figur 2.3: En kombinasjon av vertikal og horisontal prototype

Det vil si at man reduserer antall momenter eller nivå på funksjonalitet til å passe ett scenario som bare er i stand til å simulere brukergrensesnittet så lenge den som testes følger en bestemt sti. Scenarier er også billige og enkle å bygge .

2.2.3 Forutsiende evaluering

Ikke alle evalueringsmetoder involverer brukeren. Noen ganger er det hensiktsmessig å la eksperter evaluere brukergrensesnittet. I *forutsiende* (engelsk: predictive) *evaluering* kan ekspertene bruke designprinsipper om hvordan brukere samhandler med brukergrensesnitt og

prinsipper rundt grensesnittet. De benytter også teoribaserte modeller (Preece, Rogers & Sharp 2002). Den store fordelen med forutsiende evaluering er at den er rask og relativt billig å gjennomføre. Ulempen er imidlertid at den ikke involverer brukerne. Dette betyr ikke at man bør utelukke forutsiende evaluering, men at denne metoden bør kombineres med andre metoder som involverer brukere.

2.3 Brukbarhetstesting

"The most basic advice with respect to interface evaluation is simply to do it, and especially to conduct some user testing" (Nielsen 1993: 102).

Brukbarhetsting er en anvendt form for eksperimentering som brukes av utviklere for å kunne teste om deres produkt (ferdig eller under utvikling) er brukbart for de tenkte brukerne, slik at de på best mulig måte skal kunne utføre sine oppgaver (Dumas & Redish 1999). Under en brukbarhetstest prøver testbrukerne å gjennomføre et sett med realistiske oppgaver på en prototyp eller eventuelt det endelige systemet. I løpet av en slik test kan man måle tiden det tar å gjennomføre oppgaver, antall feil, hvilke typer feil som oppstår, hvor brukerne står fast, valg av sti for å utføre oppgaven, hvordan de tolker grensesnittet, antall "steg" de bruker før målet er nådd, negative kommentarer eller for eksempel finne ut hvor tilfredstilt brukerne er. Det finnes flere varianter av brukbarhetstester som kan benyttes; CTA, RTA, *Constructive Interaction, Cooperative Usability Test* (CUT), *Quick and Dirty* og for eksempel Felttester.

2.3.1 Hvorfor brukbarhetsteste

Bruce Tognazzi (2000) har lagt ut en liste med hvorfor brukbarhetstesting er viktig;

Kapittel 3 Problemet er utbedret før produktet er levert, ikke etter. "It is far less expensive to have one person spend an hour revamping a light-weight prototype than for a room full of customer support people to spend two months fending off calls while your entire engineering team tries frantically to fix reported problems and get out release 1.1".

Kapittel 4 Teamet kan konsentrere seg om virkelige problem framfor innbilte. "This is perhaps the greatest money-saver. Developers know about most of the problems that will be found before the testing even begins. The problem is, they also know about a raft of other "problems" that aren't real. And they have no tool other than testing to tell the difference"

Kapittel 5 Utviklerne implementerer i stedet for å debattere. "We've all been to those project team meetings where perhaps ten $100/hr engineers, designers, and marketing people sit around and debate how users are likely to respond. That's $1000 an hour for uninformed

opinion. One usability professional, applying the scientific method, can have a real answer in two hours for a tiny fraction of that amount. Not only that, it will be the right answer."

Kapittel 6 Forhandlingstiden er betydelig redusert. "The iterative design process cycles perhaps 8 to 12 times in a properly-tested product before reaching a reasonable stasis. I have routinely performed that many iterations in a two to three week period. At the end, what may have started as a brave stab in the dark has been honed into a solid, compact, successful design. By having the design team be first off the blocks, you will usually have the major part of design ready to go before the implementers are ready for it. Where time to market is even shorter, designers are trained to work with the rest of the team, concentrating first on those areas that need to be implemented first"

"Finally, upon first release, your sales department has a rock-solid design they can go sell without having to pepper their pitches with how it is all going to actually work in release 1.1 or 2.0."

2.3.2 Antall brukere

Hvor mange testdeltagere trenger man for å gjennomføre en brukbarhetstest? Her er det ganske stor uenighet mellom forskerne, og det er også avhengig av hva slags type brukbarhetstest man skal gjennomføre. Noen peker på at man trenger bare fem testdeltagere for å avdekke 80-85 prosent av alle brukskvalitetsproblemer (Nielsen 2000), mens andre mener at fem personer er for få, og dekker bare 55 prosent av problemene (Faulkner 2003). Dumas og Redish (1999) nevner at det er vanlig at mellom fem til tolv personer blir testet, men på grunn av budsjett og tidsskjema blir færre testet. Dessuten finnes former for brukbarhetstester, som *Quick and dirty* (se kapittel 2.5.2), som sjelden involverer mer enn et par personer (Preece, Rogers & Sharp 2002). Det finnes også metoder hvor man skal teste to eller flere personer samtidig (for eksempel constructive interaction).

2.3.3 Testomgivelser

En brukbarhetstest kan gjennomføres i forskjellige omgivelser. Den vanligste varianten er å ta med brukerne i ett testlaboratorium (Preece, Rogers & Sharp 2002). Denne formen for testing blir ofte omtalt som *brukbarhetstesting*, men egentlig er den bare en *variant* av brukbarhetstesting. Det finnes mobilt utstyr for brukbarhetstesting, slik at man kan dra ut til brukeren for å teste i brukerens egne familiære omgivelser, eller at man følger brukeren

utendørs (Kaikkonen 2005). I laboratorier måler man utførelsen av oppgaver i mer kontrollerte omgivelser, hvor man enkelt kan observere og evaluere utførelsene. Målet med en brukbarhetstest, uavhengig av metode, er å finne ut om produktets brukskvalitet holder mål i henhold til hva som defineres som god brukskvalitet (Nielsen 1993).

Etter en brukbarhetstest, sitter man igjen med store mengder data. Disse dataene skal organiseres og analyseres. Før en test bør man sette seg inn i hva man faktisk vil finne ut, samt hvordan man skal samle inn dataene (Dumas & Redish 1999). Typiske data som samles inn kan være;

Kapittel 7 Start og slutt-tid
Kapittel 8 Antall steg som er gjennomført
Kapittel 9 Brukerens navigasjons sti
Kapittel 10 Valg av løsning
Kapittel 11 Negative eller positive kommentarer eller uttrykk i løpet av testen
Kapittel 12 Forslag til forbedring

Brukbarhetstesting har noen felles karakteristiske trekk med vitenskapelige eksperimenter. Begge måler for eksempel utførelse. Forskjellen er at brukbarhetstesting er en systematisk tilnærming for å evaluere brukerutførelsen, med den hensikten å forbedre designet av brukskvaliteten. Eksperimenter innen forskningsformål er å oppdage ny kunnskap. Et forskningsrettet eksperiment krever at forsøksprosedyrene skal være svært rigorøse og grundig dokumentert slik at resultatene kan benyttes med sikkerhet av andre forskere. Brukbarhetstesting bør også planlegges og gjennomføres med grundighet, men en må ta hensyn til restriksjoner i den virkelige verden og kompromiss må inngåes (Preece, Rogers & Sharp 2002). Resultatene kan sjelden repliserbare, men det bør være mulig å gjenta testen for så å finne lignende resultat. Forskningsrettede eksperimenter fokuserer også mer på validitet ved bruk av for eksempel statistiske tester. Validitet innebærer at vi virkelig måler det vi ønsker å måle.

2.3.4 Planleggingsfasen

Hvis man ikke planlegger en brukbarhetstest, vil man mest sannsynlig kaste bort mye tid og ressurser (dette gjelder ikke i samme grad enklere metoder som *quick and dirty*). Det er viktig å vite hva man vil teste, hvordan man vil teste det, velge ut passende oppgaver, osv. DECIDE

er et enkelt rammeverk for planlegging av en brukbarhetstest (Preece, Rogers & Sharp 2002) som er mye brukt. Dumash og Redish (1999) beskriver omtrentlig de samme stegene. I hovedsak skal disse punktene være dekket under planleggingsfasen;

Bestemme målene

Å teste store deler av produktet krever mye ressurser. Hvis man ikke fokuserer er det lett å gå glipp av viktig informasjon. Under en brukbarhetstest skjer det mye samtidig. Derfor skal man fokusere på hva som er de overordnede målene for evalueringen, hvem som har interesse i målene og hvorfor de har det. Eksempler på mål er å;

Kapittel 13Sammenligne grensesnitt og finn det beste
Kapittel 14Se hvordan teknologien forandrer brukerens arbeidspraksis
Kapittel 15Finne ut hvordan testbrukerne håndterer menyene
Kapittel 16Hvilke stier velger testbrukerne får å få svar på oppgavene?
Kapittel 17Se om en gitt metafor fungerer

Målene bør veilede evalueringen, så målene må settes på plass tidlig. Dumash og Redish skiller mellom mål og antagelser. Antagelser er hva vi tror kan være brukskvalitetsproblemer.

Forberede spørsmålene

Det å avgjøre hvilke oppgaver som skal benyttes i testen er et kritisk område. Ofte utvikles et sett med ferdigdefinerte oppgaver. Hver oppgave kan dekke ett eller flere mål. Nøye valg av oppgaver, forsterker ønsket fokus. Oppgavene kan utvikles med tanke på kvantitative målinger. Dette kan omhandle;

Kapittel 18hvor lang tid det tar å fullføre oppgavene
Kapittel 19hvor ofte og hvilken type feil som oppstår i de forskjellige oppgavene
Kapittel 20antall navigasjonsflytt til hjelpefunksjoner eller manualer
Kapittel 21antall personer som gjør den samme feilen

Oppgave bør være enkle og presise, slik at de ikke kan tolkes på ulike måter.
Løsningene bør variere fra enkle til mer komplekse. De første oppgavene bør være enkle, noe som ofte vil gi testdeltageren selvtillit. Hver oppgave varer som regel en plass mellom 1 til 20 minutter. Alt dette er avhengig av produkt som testes og mål.

Valg av testdeltagere

Hvilke testdeltagere bør man velge? En god start er å finne ut hvem som er målgruppe for systemet, slik at systemet testes av de som faktisk skal bruke det. Det er ofte hensiktsmessig å teste mennesker med ulike typer erfaring. Eksempel på dette kan være innføring av et datasystem i den kommunale sektoren, der man kan bruke testdeltagere med forskjellig nivå av datakunnskaper. Utdanning, alder, kjønn eller personlighetstyper er også faktorer man kan vurdere under utvelgelse av testdeltagere.

Omgivelser

Det er ønskelig å forhindre bråk og forstyrrelser under brukbarhetstester, og til dette finnes spesialdesignede laboratorier. Et laboratorium består ofte av to rom, et rom hvor testdeltageren skal løse ett sett med oppgaver, og et observasjonsrom hvor testdeltagerens handlinger blir observert og til en viss grad analysert. Testdeltageren blir ofte filmet under en brukbarhetstest, så mye av evalueringen og analyseringen foregår etter selve brukbarhetstesten. Rommene kan være spesialdesignet for å ligne testbrukernes vante omgivelser. Dette kan gjøres ved å flytte på vegger, innrede med passende møbler eller sette opp utstyr man finner igjen fra testbrukerens hverdag.

Planlegge gjennomføringen

Det finnes mange metoder for gjennomføring av en test, men det finnes også forskjellige teknikker innen hver metode. Man kan sette av tid på å la deltagere øve seg på å si høyt hva man tenker, slik at det blir lettere å forstå hva testdeltagerne tenker/tenkte under utførelsen av oppgavene. Man må avgjøre om testbrukerne skal snakke mens man løser oppgaver eller om man snakker etterpå. En kan gi ut et spørreskjema etter testen hvor brukeren evaluerer hvor godt han eller hun likte brukergrensesnittet, eller man kan intervjue. Hvis man vet hva man skal fokusere på, kan en lage et skjema som man krysser/fyller ut under gjennomkjøringen. Det finnes mange teknikker å velge blant, noen bedre enn andre.

Det etiske

En brukbarhetstest kan være belastende for testdeltagerne. Derfor er det viktig å informere om at det ikke er deltagerne vi tester, men systemet. Testdeltagerne kan hoppe til neste oppgave eller avslutte testen når de selv føler for det. Det er viktig å peke ut all overvåkning, slik som kameraer og speil. Man må også gjøre seg kjent med de lovene som gjelder overvåkning.

2.3.5 Pilottest

Etter selve planleggingsfasen er det en god idé å teste planen (gjelder de fleste brukbarhetstest metodene). Dette kan man gjøre ved å foreta en pilottest. Hensikten med en pilottest er å avdekke eventuelle problemer før man starter brukbarhetstestene for alvor. Også i løpet av dette studiet ble viktighetene av å foreta en pilottest erfart. ”The study revealed numerous practical problems that usability evaluators experience when testing. In 12 sessions (of 14) we observed examples of such problems or practical realities. These include system failures, users not showing up for a session, disturbing surroundings, and technical problems with recording devices” (Hornbæk & Nørgaard 2006: 214)

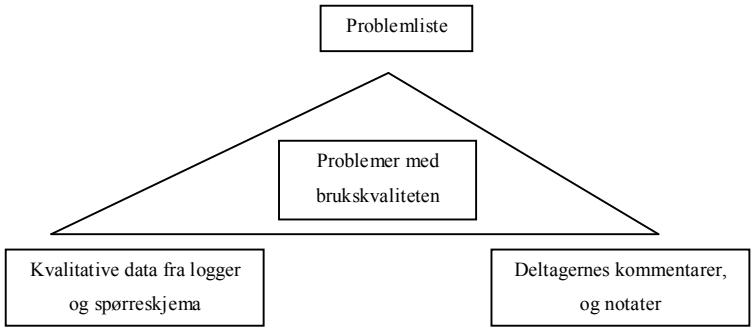
En pilottest legger til rette for å avdekke tekniske problemer, forstyrrende omgivelser eller for eksempel forståelsen av oppgavene. Tilretteleggeren og eventuelt observatøren får også sjansen til å forbedre seg selv.

2.3.6 Analyse av brukbarhetstest

Evalueringssesjonen forekommer mellom og etter brukbarhetstestene. En brukbarhetstest genererer store menger med verdifulle data som kan benyttes for videre analysering. Dataene kan ende opp i store mengder nedskrevet tekst; egne notater, andre observatørers notater, deltagerens notater eller spørreskjema som deles ut før eller etter selve testsesjonen. Enkelte forskere har påpekt viktigheten ved å foreta en strukturert analyse umiddelbart etter hver brukbarhetstest (Duman & Redish 1999, Hornbær & Nørgard 2006), fordi det er viktig å få skrevet ned ferske tanker og observasjoner på papiret. På denne måten de lett å hente igjen ved en senere anledning.

Med analyse menes forståelse av hva som var viktige observasjoner av brukbarhetstesten(e). Dette handler om å forstå brukerens oppførsel, årsak til observasjonene og tenke ut eventuelle designløsninger. Hornbæk og Nørgaard (2006) viser til eget forskningsstudium hvor strukturert analyse i brukbarhetstesting sjeldent blir foretatt rett etter sesjonene etter hvert som de oppstår. Utfordringen er å finne de *virkelige* problemene i all datamengden som samles inn. ”The real problems are the ones that are going to cause difficulties for users when they get product in their homes or offices to their work” (Dumas & Redish 1999: 310). Mange av problemene som dukker opp under brukbarhetstest er symptomer på større problemer. Dumas

og Redish har foreslått en rekke teknikker for å strukturere dataene. Triangulering er en av teknikkene der man ser de innsamlede dataene, og prøver å se etter sammenhenger.



Figur 2.4: Triangulering

Trianguleringsteknikken krever at man skal gå strategisk fram. Det første steget er å lage sammendragsskjema av utførelsesmålinger og subjektive målinger. En kan begynne med å sette opp utførelsesmålingene i tabeller som viser deltagernummer og oppgavenummer. Her kan man for hver oppgave vise forløpt tid, antall feil (av forskjellige typer) og eventuelt andre viktige målinger. Dette kan være et eksempel på en tabell som viser utførelsestidene på hver oppgave:

	TB #1	TB #2	TB #3	…	TB #N	Total	Gj.snitt
Oppg. 1	00.50	00.28	00.28	…		06.43	00.34
Oppg. 2	05.34	04.24	02.32	…		15.12	04.02
…	…	…	…	…			
Oppg. M							
Total	14.55	11.23	16.25				
Gj.snitt	01.02	01.12	00.56				

Figur 2.5: Tidsskjema for brukbarhetstesting

Ut i fra dette kan man regne ut total- og gjennomsnittstid per oppgave, samt total- og gjennomsnittstid per bruker. Ved å benytte regneark kan man beregne enkle statistikker som gir en god oversikt over tidsbruken. I tillegg lager man en liste over kommentarer fra testbrukerne. Dette kan være kommentarer fra under tester, i debrief-sesjonen eller fra samtale

etter endt test. Til slutt lages en liste med notater som testpersonellet noterte under testen. Mange bruker også skjema som fokuserer på type feil under selve testen.

Så er turen kommet for å se etter trender og overraskelser i dataene. Her kan man finne mye, og det er mange faktorer en kan ta i betrakting. For eksempel kan man sammenligne personer med ulik erfaring med produktet som testes. Man kan se om noen bruker lengre tid enn forventet. Man kan se etter mønster ved for eksempel å sammenligne de med og de uten dataerfaring. Man kan se etter "outsidere". Hvis man har for eksempel fem testbrukere og en av disse sliter med en oppgave kan denne testpersonen gjerne representere 20-30 prosent av alle framtidige brukere. Ved å se på videoopptak av skjerm, notater og kommentarer fra testbrukere, har man et bedre utgangspunkt for å evaluere dataene. Ser noe ut til å være et potensielt brukskvalitetsproblem, bør man fokusere på dette i neste iterasjon av brukbarhetstesten.

Undersøke dataene

Etter man har sett på dataene som nevnt over er tiden kommet for å utforske dataene for å se om de er konsistent med hypotesene man hadde før man startet utførelsen av testene. Med dette kan man se om:

Kapittel 22Hvorvidt antagelsene stemmer
Kapittel 23I hvilken grad brukskvalitetsmålene er oppnådd.
Kapittel 24Om de kvantitative målene som er satt for oppgavene er oppnådd.

Globale og lokale problem

Etter at problemene er avdekket, skal de kategoriseres. Dumas og Redish skiller også mellom globale og lokale problem. Lokale problem kan være;

Kapittel 25Manglende steg i prosedyren som forårsaker at de ikke kan fullføre oppgaven
Kapittel 26Undermenyene med lite passende navn
Kapittel 27Tekstfelt som ikke indikerer hvordan man skal skrive dataene. For eksempel når man skal skrive dato.

- dd.mm.åå
- dd.mm.åååå
- åå.mm.dd

Globale problem har ett større omfang, for eksempel;

Kapittel 28Det er vanskelig å navigere seg fram i menyene.
Kapittel 29Generell dårlig feedback når man har utført en oppgave
Kapittel 30Det er ingen forklaring på noen av tekstfeltene man skal skrive i. Da vet ikke brukerne hvilket format man skal skrive i.

Globale variabler er ofte mer alvorlige enn lokale.

Organisering av problemnivåene

Når man skal finne de problemene som har videst omfang kan problemene deles inn i disse nivåene:

Nivå 1: Problem som hindrer utførelse av en oppgave. Brukere velger konsekvent feil menyvalg, og vet ikke hvilke andre plasser man skal gå. Eller at testbrukerne gir opp etter de har prøvd å skrive ut.

Nivå 2: Problem som forårsåker signifikante forsinkelser og frustrasjon. For eksempel manglende feedback når en bruker har utført en oppgave, som resulterer i at man gjør oppgaven en gang til for å forsikre seg at den er gjort.

Nivå 3: Problemer som har en mindre effekt på brukskvaliteten. For eksempel ved å bruke det samme ordet på to ulike handlinger, noe som gjør at brukerne må stoppe opp og tenke over om de gjør det rette valget.

Nivå 4: Dette punktet er mer finurlig og peker ofte på en forbedring som kan bli lagt til i framtiden. Dette kan være forbedringsmuligheter som er blitt påpekt underveis eller etter testen.

2.4 Think aloud - TA

Evaluering av brukskvalitet involverer ofte testing av brukere som samhandler med datamaskiner, og det finnes et bredt spekter med teknikker og metoder som kan benyttes. Noen av de mest grunnleggende teknikkene er; think aloud (TA), observasjon, videoopptak, automatisk logging av markør eller tastetrykk, retningsledet samhandling, intervjuer og spørreskjema, og ofte blir disse teknikkene kombinerte. En teknikk utpeker seg - think aloud.

Think aloud brukes for innsamling av verbale protokoller under brukbarhetstestingen. Denne teknikken brukes på flere fagområder som psykologi, sosialvitenskap, produktdesign og ikke minst i den brukersentrerte delen av systemutviklingsprosessene (Preece, Rogers & Sharp 2000). Teknikken finnes i flere varianter, men to teknikker/metoder skiller seg ut; sanntids TA og retrospektiv TA, henholdsvis CTA og RTA. Det finnes også mange lignende varianter av disse metodene, men disse skiller klart på hvilket tidspunkt man skal snakke høyt. Ved gjennomføring av CTA er meningen at brukerne skal verbalisere hva de oppfatter mens de løser ett sett med oppgaver. Brukerne kan verbalisere hva de ser, hva de føler, hva de tenker og hva de gjør. Med RTA-metoden skal man løse oppgavene i fred og ro uten ay man behøver å snakke høyt. Verbaliseringen foregår etter oppgaveløsingen.

Think aloud- og talk aloud- protokoller benyttes ofte om hverandre i litteraturen, men Ericsson og Simon (1984) skiller mellom disse uttrykkene. Begge protokollene har likheter, men de skiller seg spesielt ut på ett område. Talk aloud protokollen beskriver testpersonenes handlinger uten at de gir forklaringer på hvorfor de gjør som de gjør.

Think aloud er en av de mest anvendte og populære teknikkene som benyttes under brukbarhetstesting, uavhengig om man gjennomfører felttester, laboratorietester eller andre metoder (Nielsen, Clemmensen & Yssing 2002). Det er lett å blande begrepene *teknikk* og *metode* i denne sammenhengen, og at begrepene brukes om hverandre i litteraturen gjør ikke saken enklere. Karl Duncker (1945) referert av Nielsen, Clemmensen og Yssing (2002) var den første som beskrev think aloud teknikken i sitt arbeid innen eksperimentell psykologi hvor han studerte produktiv tenking. Det refereres sjeldent til dette verket av MMI-forskere. I litteraturen er think aloud beskrevet under mange navn; *verbal reports*, *concurrent verbal*

protocols, retrospective verbal protocols, retrospective verbal protocols, after think aloud og *verbal protocols* (Boren Ramey 2000).

2.4.1 Concurrent Think Aloud - CTA

CTA metoden refereres ofte som selve *brukskvalitetsmetoden* og benyttes hovedsaklig til settinger innen laboratorier, men også workshops og felttesting (Nielsen, Clemmensen & Yssing 2002). Mange skriver at dette er en svært god og enkel teknikk, og at den avdekker de fleste problemer med brukskvaliteten. Jacob Nielsen har uttrykket følgende; ”Thinking aloud (CTA) may be the single most valuable usability engineering method” (Nielsen 1993: 195)

Metoden gir umiddelbare resultater og er regnet som kostnadseffektiv (Nielsen 1993). Teknikken kan også gjennomføres av de som ikke er spesialister på brukskvalitet. Det viktigste er at den antaes å gi tilgang til kognitive prosesser under brukernes engasjement med datamaskiner. CTA er allikevel kritisert av en del forskere; ”Verbal protocols place added strain on users, who are required to do things at once – the task itself and talk about their actions or what they are thinking about. Evidence from cognitive psychology, shows that humans are poor at maintaining divided attention for more than a few minutes. so you will need to think of ways to support users if you want to collect this kind of data. How can you prevent the long silence, which will not tell you much?” (Preece, Rogers & Sharp 1994: 319)

Forestill deg at du kjører bil gjennom en by, og du skal si alt hva du gjør og har tenkt å gjøre. Det er lite trafikk og du kjører en rute du kjenner godt fra før. Da du er halvveis blir du omdirigert pga veiarbeid, og må forandre til en rute som er dårlig skiltet og ukjent for deg. I tillegg øker trafikken. Du skal fortsette med å verbalisere dine tanker. Hva vil mest sannsynlig skje med den verbale protokollen?

I starten er det gode sjanser for at du ville behersket den verbale protokollen. Du kjenner ruten og du trenger ikke tenke så mye igjennom hva du gjør, så det går greit å snakke samtidig som man kjører (dette er selvsagt subjektivt). Problemene oppstår når en må begi seg på den ukjente ruta. Du vet nødvendigvis ikke hvor du er og hvor du skal, og samtidig må du konsentrere deg om trafikken. Antageligvis vil du stoppe med å prate fordi du må konsentrere deg om å kjøre. Hvis noen spør deg om ”hva tenker du nå?” kan det hende du fortsette å kommentere for en liten stund, men hvis påminnelsen forekommer mer enn ett par ganger vil

du antagelig føle deg forstyrret, spesielt hvis forstyrrelsene forårsaker at du gjør feil. Dette er ikke en uvanlig situasjon i brukertesting og andre hverdagslige situasjoner.

Stillheten er ett problem, og det samme kan sies om forstyrrelser. Hvis brukeren har gjort en feil eller sliter med å komme videre, skal tilretteleggeren la være å hjelpe vedkomne (Nielsen 1993, Ericsson & Simon 1984). Deltageren skal på egen hånd finne ut av sine problemer, slik at de som observerer skal kunne se hvordan testbrukerne tolker grensesnittet. Testbrukeren skal si hva han eller hun tenker, men i praksis blir de fleste stille i vanskelige situasjoner. Det kreves trening for å kunne verbalisere hva man tenker til enhver tid, og det er store forskjeller mellom individer. Vi kan som sagt be testbrukeren om å tenke høyt med hyppige mellomrom, men det kan slå igjennom som forstyrrende, i alle fall i lengden (Preece, Rogers & Sharp 1994; Boren & Ramey 2000). Det å verbalisere sine tanker, er en unaturlig prosess for de fleste. Mange føler seg også overvåket, evaluert og bedømt når de vet noen hører på eller at de vet det blir benyttet opptak. Dette kan gå på bekostning av troverdigheten til den praktiske bruken av think aloud som en evaluering av brukskvaliteten. Teknikken blir ofte omtalt som kunstig og begrenser en konstruktiv dialog hos brukerne (Frøkjer & Hornbæk 2005).

2.4.2 Retrospective Think Aloud - RTA

RTA metoden kalles også *post-task testing*, *retrospective report* og *think after*. Her arbeider testpersonene uten at de trenger å si høyt hva de tenker under oppgaveløsningen. Dette gjør at situasjonen ligner mer på en hverdagssituasjon hvor personen kan arbeide stille. Det vil allikevel ha et kunstig preg, pga overvåkning og bruk av laboratorium, men i mindre grad i forhold til CTA.

Men denne metoden har fått kritikk fordi den ikke gir direkte tilgang til testpersonenes tankeprosess (Ericsson & Simon 1984), og man kan lettere gå glipp av uttrykk som irritasjon, overraskelser og andre følelser. Det tar også lengre tid å gjennomføre en RTA-test enn en CTA gjennomgang, fordi man kjører gjennom hele oppgavesesjonen to ganger. Men som nevnt påståes det at CTA utsettes med en dobbel kognitiv belastning (Nielsen, Clemmensen & Yssing 2002, Boren & Ramey 2000), som kan ha en påvirkning på utførelsen. Det kan også føre til at man arbeider på en annen måte, enten at man yter raskere enn vanlig, eller at man blir hemmet av den doble kognitive belastingen. Alt dette kan resultere til at man ikke oppdager de potensielle brukskvalitetsproblemene. Etter testbrukeren har løst oppgavene i stillhet, er tiden inne for å benytte TA-teknikken, altså retrospektiv TA.

2.4.3 Stimuli Retrospective Think Aloud - RTA

Det er kjent innen hukommelsesforskning at vi gjenkjenner materiale mye lettere enn vi kan erindre det fra hukommelsen (Preece et al. 1994). Dette har også stor betydning for utvikling av grafiske brukergrensesnitt. Preece et al. skriver at ”The superiority that the phenomenon of recognition has over recall has obvious ramifications for interface design. Indeed, during the last decade there has been a shift towards designing interface where the amount of information users are required to recall has been reduced in favour of requiring them to recognize the information that is needed to perform a task:” (s 118). Med ”recognize” menes at en person husker noe som er gjort før fordi han gjenkjenner den grafiske representasjonen. Hjernen kjenner ofte igjen former umiddelbart, og derfor brukers ofte symboler som man lett kjenner igjen når man designer grafiske brukergrensesnitt. Men det er ikke bare i utvikling av grafiske brukergrensesnitt dette gjelder. Når testdeltagerne ved hjelp av videoopptak, datalogg -filer eller opptak av skjerm skal tenke høyt, er det enklere å verbalisere sine tanker når de ser hva de selv har gjort (se neste avsnitt). Når RTA utføres uten noen form for stimuli, kan dette gå på bekostning av at nøyaktigheten av de kommentarene som produseres. Et problem er at testpersonene kan produsere tanker de egentlig ikke har tenkt, og i feil rekkefølge (van der Haak & de Jong 2003, Guan et al. 2006).

2.4.4 Tidligere forskning

I den senere tid er etterspørsel på forskning på metoder innen brukbarhetsester blitt mer etterspurt (van den Haak & de Jong 2003, Guan et al. 2006). Litteratur rundt CTA og RTA har en tendens til å framstille metodene som like alternativ, ved at man enten kan velge det ene eller det andre uten å gå i dybden på forskjeller (Nielsen 1993). Så langt har få studer sammenlignet RTA og CTA. Hoc og Leplat (1983) undersøkte metodene CTA, RTA med stimuli og RTA uten stimuli for å sammenligne problemløsingsprosessene fra de ulike metodene. I den retrospektive varianten skulle testdeltagerne først gi en fremstilling av deres tankeprosess uten stimuli. Etterpå skulle de tenke høyt ved å se på en log fil fra alle stegene i prosessen. De konkluderte med at RTA uten stimuli burde unngåes, pga forvrengningene og mangler i protokollene. Men de påpeker at både retrospektiv og sanntid TA produserer like resultat. Det skal legges til at Hoc og Leplats forskning ikke ble brukt i en brukbarhetsteststssituasjon.

Van den Haak, De Jong og Jan Schellens (2003) gjennomførte et eksperiment der de sammenligner CTA og RTA metoden på en online biblioteks katalog. De kom også fram til at metodene avdekker like sett med resultater, men at problemene kommer i lyset på forskjellige måter. I RTA ble flere problemer avdekket av selve verbaliseringen, mens CTA metoden avdekket flere problemer med observasjon. De viser også til av kravet om å snakke høyt har negativ effekt på utførelsen under oppgaveløsningen.

Browers og Snyder (1990, refereses via Van den Haak & de Jong 2003) ga også et bidrag hvor de sammenligner CTA- og RTA- (med stimuli) metodene på brukbarhetstester. De fant i motsetning til dette studiet ingen signifikante forskjeller i henhold til utførelse og tid på oppgaveløsing, men RTA-metoden ga vesentlig færre verbaliseringer. Dette er de stikk motsatte funnene som denne oppgaven har kommer fram til. Men i likhet med denne oppgaven, påpeker de at verbaliseringen er av annen type enn sanntidsverbalisering, fordi det fokuseres mer på forklaringer enn prosedyrer. Ut i fra Browers og Snyders artikkel, finner man ingen rapporterte nr av feil eller type problemer som ble avdekket av deltagerne i de to TA-typene. Dette gjør at sistenevnte artikkel ikke blir tatt i betraktning videre her, siden misforståelser lett oppstår når man ikke har noe konkret å gå etter.

Noen mener at CTA er den beste metoden. Katalin (2000) referert i Nielsen, Clemmesen og Yssing (2002) ser ingen problemer med den doble kognitive belastningen. De bruker TA for å få tilgang til studentenes leseforståelse. ”The closest possible way to get to the cognitive processes of readers” (s1). Konteksten kan ha noe å si, og ikke alle aktiviteter krever like mye av mentale prosesser. For eksempel er det lite trolig at å løse enkle oppgaver krever like mye konsentrasjon som å løse vanskelige oppgaver. Det vil også være nærliggende å tro at interaksjonsapplikasjoner krever mer deltagelse enn typiske informasjonsapplikasjoner. Dette begrunnes med at teknikken legger på en kognitiv belastning og krever en kognitiv involvering fra brukeren som kan forstyrre de kognitive kravene fra interaksjonen eller oppgaven.

2.4.5 Verbale protokoller

Think Aloud brukes som sagt til å studere søkestrategier og navigeringsoppførsel av folk som leter etter detaljert informasjon, og kombineres ofte med andre teknikker slik som videoopptak eller intervju etter testen osv. Det vi ønsker er å få tilgang til testdeltagerens mentale prosess, hvor testsituasjonen ikke påvirker den mentale prosessen.

Blant de første som benyttet verbalisering var W. James og W. Wundt og andre som benyttet introspeksjon for å undersøke psykologiske påstander og teorier av våre mentale prosesser. Introspeksjon er en systematisk metode for selviakttakelse i psykologien, og var vanlig i eksperimentalpsykologiens fase på 1800 tallet. Ved introspeksjon forsøker trente testpersoner å rapportere sine mentale modeller direkte. Resultatene av disse forsøkene var vanskelige å gjenskape, og ble derfor i senere tid angrepet av atferdspsykologer som J.B. Watson som førte til introspeksjonens bortgang som en psykologisk forskningsteknikk (Boren & Ramey 2000). Etter hvert som den kognitive psykologien fikk fortgang i 1960 årene, kom også introspeksjon på banen igjen. Nisbett og Wilson (1977), referert av Ericsson og Simon (1984), gikk i 1977 ut med en omfattende og anstrengende kritikk mot bruken av slike verbale protokoller. I deres artikkel "Telling more than we know" argumenterte de med at mennesker ikke har direkte tilgang til sine mentale prosesser, og derfor kan man ikke presisere dette i noen form for rapporter.

2.4.6 Ericsson og Simons modell

Ericsson og Simon var på mange måter enige med Nisbett og Wilson, men de påpekte at kritikken var begrenset primært til spesifikke typer verbaliseringer. De hevder at enkelte typer (eller nivåer) av verbalisering på rettmessig vis kan bli betraktet som gyldig data, men bare hvis de ble vurdert som indikatorer av hvilken informasjon som ble innsamlet og i hvilken rekkefølge. Ericsson og Simons modell blir ofte referert som teoretisk forankring i regi av brukbarhetstesting.

Ericsson og Simon argumenterer med at verbale rapporter er nødvendige for å forstå menneskelige handlinger. En setning kan være en verbal realisering av en idé, og verb i en setning kan brukes til å identifisere forskjellige typer informasjon og forskjellige kognitive prosesser. De skiller mellom introspeksjon, retrospektive rapporter og kommunikasjon til eksperimentatoren på den ene siden, og verbalisering av nåværende tanker (tanker som reflekterer det som skjer nå) på den andre siden. Forskjellen på disse to har med korttidsminnet å gjøre. Alt vi vet, har på ett eller annet tidspunkt gått gjennom korttidsminnet.

Vi kan verbalisere hva vi oppfatter i selve prosessen der vi oppfatter, og vi kan verbalisere hva vi var bevisst på hvis vi blir spurt rett etter selve tankeprosessen. Dette er fordi tankene fremdeles ligger i korttidsminnet. Men hvis det er et tidsrom mellom oppfattelsen og når man

bes om å gjenkalle erindringene, vil vi produsere beskrivelser og forklaringer. Dette er ikke en rapport av våre øyeblikkelige tanker, fordi informasjonen fra korttidshukommelsen er borte. Retrospektive rapporter er også av interesse hvis de benyttes som dobbeltkontroll og hvis de er koblet mot informasjon fra korttidshukommelsen.

Skal man ha tak i de kognitive prosessene, vil bare verbale rapporter fra CTA si noe om informasjonen som benyttes under oppgaveløsningen (Ericsson og Simon 1984). Ved å identifisere og analysere kom de fram til tre nivåer for organisere verbaliseringene etter. Nivåene er sorterte etter grad av pålitelighet;

Nivå 1:

Verbalisering som ikke trenger å transformeres før de uttrykkes. En testdeltager som uttrykker sekvenser av tall mens man løser et matteproblem produserer nivå 1-data fordi tallene kan uttrykkes i samme form som de originalt ble kodet i korttidshukommelsen. Dette er den mest pålitelige typen av verbalisering, men er ofte vanskelig å oppnå. Å snakke høyt krever kognitive prosesser som opererer direkte på muntlig kodet informasjon.

Nivå 2:

I motsetning nivå 1 må her verbalisering transformeres før de uttrykkes gjennom en utførelse av oppgaver. Her snakker vi om bilder eller abstrakte konsepter som må transformeres til ord før de kan verbaliseres. Denne transformeringen er den eneste som danner overgang til kognitiv prosess mellom korttidshukommelsen og verbalisering. Dette framskaffer også pålitelige data ifølge Ericsson og Simon. Think aloud-oppgaver krever både muntlig kodet informasjon og andre former for tanker som ligger i korttidshukommelsen. Protokollene tar form som setninger som kan forståes som tanker uten innen kontekst av andre tanker.

Nivå 3:

Verbalisering som krever ytterligere kognitive prosesser utover det som kreves ved utførelse av oppgaver eller verbalisering. Eksempel på ytterligere kognitiv prosesser kan være filtreringsprosesser (for eksempel det å kunne uttrykke informasjon som er knyttet til ett tema), lage slutning om individets egen kognisjon, og informasjon som er hentet fra langtidshukommelsen etter anmodning fra forskerne. Det kan også være hvilken som helst innflytelse fra utsiden slik som kommentarer eller påminnelse fra forskerne. Innflytelsen snur følgende verbalisering til nivå 3 fordi den normale flyten av informasjon i

korttidshukommelsen igjennom oppgaven har blitt forandret. Nisbett og Wilson og Ericsson og Simon argumenterer mot bruk av dataene på dette nivået. RTA kommer under Nivå 3.
I tillegg til disse nivåene finnes verbaliseringer som ikke betraktes som data i denne modellen, på grunn av vanskeligheter med tolkning og analyse (s 223). Her snakker vi om følelser, osv.

Nivå 1 og 2 kan altså betraktes som pålitelige data, hvis de hentes inn på riktig måte.

2.4.7 Forsoning av teori og praksis

Boren og Ramey (2000) studerte hvordan praktikanter faktisk utførte think aloud på brukbarhetstester og diskuterte sine observasjoner med relasjoner til det klassiske verket til Ericsson og Simon (1984). Når man skal utvikle anvendelige metoder som skal identifisere mangler ved et system, handler det ikke bare om å forstå deltagernes kognisjon i form av verbalisering, men også hvordan de samhandling med systemet. Når man gjennomfører brukbarhetstester, oppleves en større variabilitet enn ved testapparater som benyttes til kognitiv forskning, slik som Ericsson og Simons modell. Produkter som utvikles inneholder flere valg og faktorer som er vanskeligere å kontrollere. Et eksempel på dette kan være noe så enkelt som en nettside under utvikling som har mange stier til samme mål. Lite forskning har blitt gjort på dette området (Boren & Ramey 2000)

Boren og Ramey har prøvd å følge Ericsson og Simons retningslinjer til å gjennomføre brukbarhetstester. De foreslår ett sett med implikasjoner basert på modellen:

Samle og analysere kun ”harde” verbale data

Prosedyrer som brukes til å samle muntlige data må motstå Nisbett og Wilsons kritikk av verbal data. Dette er beskrevet under nivå 1 og 2 tidligere.

Gi detaljerte innledende instruksjoner for Think aloud

Her bør man blant annet skille mellom forklaring og think aloud. Testdeltagerne skal motiveres til å snakke konstant som om de var alene i rommet uten å tenke på sammenhengen. Think aloud-teknikken skal praktiseres før selve testen. De skal også informeres at det blir gitt en påminnelse hvis de blir stille.

Ikke forstyrr

Etter at testen har begynt, skal man ikke forstyrre testdeltageren. Den eneste innblandingen skal være en påminnelse om at de må fortsette å snakke. Testpersonene påminnes alt fra 15 til 60 sekund, alt ettersom hvilket forskningsmål man er ute etter. Også i ISO 13407 (1999) legger de vekt på at man ikke skal bryte inn: ”In order to determine whether the overall objectives have been met, more formal evaluation should be conducted in realistic context, for example, without help or interruptions from the evaluator” (s. 7).

“We encountered evaluators asking questions that differed dramatically from how Ericsson and Simon, and in part also Boren and Ramey, suggest to interact with test participants” (Hornbæk & Nørgaard 2006: 214).

I publisert litteratur som refererer til det klassiske verket til Ericsson og Simon, er prinsippene ovenfor sjeldent artikulert. Boren og Ramey (2000) argumenter med at think aloud-tester innen brukbarhetstesting i virkeligheten er mindre rigorøse, og har mange motsetninger til Ericsson og Simons anbefalinger. Det er også ofte stor forskjell på prosedyrer som benyttes av forskjellige forskere, forskjellige introduksjoner til deltagere, forskjellige grad av innblanding under testen av forskerne og forskjellige forskningsmotivasjoner. ”If Ericsson and Simon were being consistently applied, this level of variation should not exist” (s. 263).

For å få en bedre forståelse av disse tydelige forskjellene, gjennomførte Boren og Ramey (2000) et litteraturstudium kombinert med feltobservasjoner, intervju og artefakt innsamlinger med mål om å utvide forståelsen av hvordan verbale protokoller ble samlet og analysert. Utfallet viste eksempler hvor det var klare forskjeller mellom Ericsson og Simons teorier og praksis.

Brukskvalitetskonsulenter ga sjelden think aloud instruksjoner på den foreskrevne måten

Flere av disse forklarte ikke forskjellen mellom think aloud og forklaringer. Bare en av brukskvalitetekspertene lot testbrukeren praktisere for think aloud, mens bare en annen ga instruksjoner for praksis (av totalt 25 eksperter innen brukskvalitet).

Brukskvalitetspraktikere ga sjelden påminnelser om å tenke høyt på den foreskrevne måten

Påminnelsene som ble gitt var alt fra korte til lange (Eks; ”keep talking” til ”yeah, even as you’re reading, just read that out loud so I can hear you better”), fra personlige til upersonlige og fra retningsgivende til ikke-retningsgivende. Under testene som ble foretatt hadde ikke spesialistene noe fast intervall mellom hver påminnelse som kunne variere fra fem til 21 sekunder. Mange spesialister lot være å komme med påminnelser. Av 23 *pauser* som overgikk 15 sekunder var det bare fire som kom med en påminnelse, mens resten tillot stillheten inntil testpersonene selv fortsatte å tenke høyt.

Brukskvalitetspraktikere griper inn på teoretiske motstridende måter

Her ble heller ikke Ericsson og Simons teorier fulgt. Det ble observert 125 forstyrrelser i løpet av de sju sesjonene, hvor bare 16 av de var påminnelser om å tenke høyt. Også i en undersøkelse av Hørnbæk og Nørgaard (2006) hvor de gjennomfører et forskningsstudie av 14 CTA tester med forskningsspørsmålet; ”what do usability evaluators do in practice?”, var ett av de seks observasjonene at upassende spørsmål ble stilt under testen.

Brukskvalitetspraktikanter fokuserer ikke på å samle inn verbaliseringer av ”hard” data

Testdeltagerne ble ofte spurt om ”hva liker du, eller hva liker du ikke”. Testdeltagerne fikk fritt spillerom til å uttrykke sine følelser, meninger og opplevelser under og etter testen. Verbaliseringer av denne typen blir ikke betraktet som valide data på grunn av vanskeligheter med tolkning og analyse.

Årsak til manglende overensstemmelse

Til tross den hyppige henvisningen til Ericsson og Simon i litteraturen, sier tilsynelatende ikke brukskvalitetspraktikanter seg enige i Ericsson og Simons modell som rettferdiggjør samlingen av verbal data. I tillegg til at teorien og praksis ikke synkroniserer er det heller ikke uniforme uoverensstemmelser mellom brukskvalitetseksperter (Boren & Ramey 2000).

"Usability needs to be simplified even more and even more actionable. There is a full research agenda here, and we better get started finding the answers, because it is already too late" (Nielsen 2005)

Jakob Nielsen foreslår at CTA-metodene må forenkles enda mer (2005) i stedet for den mer tradisjonelle CTA-metoden som baser seg i høyere grad på Ericsson og Simons modell, men ifølge Boren og Ramey er termene dårlig definert og har heller ingen samsvarende teoretisk støtte eller henvisninger/referanser til grunnlag for forenklingene og prosedyrene han har gjort.

Områder innen brukskvalitet mangler metodisk konsistens og teoretiske bevis (Boren & Ramey 2000). Deltagerne av en brukbarhetstest kan ikke gi en fullstendig sammenhengende logisk forklaring på hvorfor de gjør det de gjør. Hvis brukbarhetstesting skal bli betraktet som en disiplin, må metodene som benyttes være teoretisk forankret og systematisk anvendt. Hvis ikke, hvordan kan vi sammenligne eller gjenta studier og gå god for validiteten? Hvordan skal av praktiseringen av brukbarhetstesting læres bort til studenter eller framtidige praktikanter? Boren og Ramey uttrykker at de savner en metodisk konsistent, og legger fram tre alternativ;

- Slutte å samle verbale data og i stedet fokusere på observerbare utførelser.
- Starte en rigid anvendelse av Simon og Ericsson's teori?
- Utforske andre teorier

"Dersom en generell teori ikke kan brukes til å kaste lys over det som skjer i miljøet, eller det empiriske materialet peker i en annen retning enn det man skulle forvente ut fra teorien, er det kanskje grunn til å sette spørsmålstegn ved etablert teori og etablerte begreper" (Repstad 1998: 19)

2.5 Alternative testmetoder

Det finnes mange teknikker og metoder for brukbarhetstesting, samt kombinasjoner av disse som utgjør ”nye” modeller. Her omtales noen av disse.

2.5.1 Felttester

Felttester er en litt annen form for evaluering av brukskvaliteten. Fordelen med felttester er at man kan samle data fra de naturlige omgivelsene som brukerne arbeider i. Et eksempel; under en felttest observeres det at en testbruker ofte blir forstyrret av andre kollegaer, slik at han må forlate applikasjonen (eventuelt prototypen) han jobber med for å hjelpe de andre. Når testbrukeren skal fortsette sitt arbeide, har han glemt hva han holdt på med og må gjøre alt på nytt. Neste fase i designet kan være å implementere statusbeskjeder som gjør at testpersonen lett kan fortsette sitt arbeide. I et laboratorium ville man mest sannsynlig ikke avdekket denne feilen. Man prøver altså å teste systemet i omgivelser der testbrukerne føler seg mer komfortable, og man kan i større grad redusere den kunstige laboratoriesituasjonen.

Utviklingen av mobile videoopptakssystem har økt mye de siste årene, noe som har gjort det enklere å utføre enkelte felttester. Man kan feste små kamera til å ta opp ansiktsuttrykk, pc skjermer, mobiltelefoner og for eksempel tastatur. Da kan brukerne gå, stå stille, sitte eller hva de enn brukte da de skulle utføre en oppgave. Tilretteleggeren som gir oppgaver går like bak og gir nye oppgaver etter at den forrige er avsluttet. Testbrukerens utstyr består typisk av flere kameraer og en mikrofon, mens tilretteleggerens utstyr kan bestå av en LCD-monitor, videokamera, en trådløs video transceiver og et batteri. Kameraets hensikt er å ta opp brukerens omgivelser .

I den siste tiden har det pågått en diskusjon rundt laboratorietester. Det påpekes at tester i laboratorium skaper en kunstig atmosfære og kan resultere i forvrengte data (Kaikkonen et al. 2005), men allikevel forekommer felttester mye sjeldnere. Keikkonen nevner at det etterlyses mer forskning på dette, og at det trengs å gjennomføre flere tester med ulike applikasjonstyper for å validere og generalisere disse resultatene.

Men felttester har sine begrensinger. De er ofte tungvinte å organisere og sette opp. I tillegg er de dyre med tanke på utstyr og lån av arbeidskraft og krever mye tid. Som med mye annet, kommer det an på hva man er ute etter å teste.

2.5.2 Quick and dirty

Ikke alle metoder krever mye ressurser. En *Quick and dirty*-evaluering brukes når designere trenger feedback fra brukere eller konsulenter for å avgjøre om deres ideer passer med brukernes behov (Preece, Rogers & Shart 2002). Metoden kan brukes under hele utviklingsprosessen og kan brukes når kjapp feedback er viktigere enn godt dokumenterte funn. Quick and dirty kan for eksempel brukes til å avgjøre design ideer tidlig i utviklingsfasen og til spesifikk valg som å avgjøre om et ikon fungerer eller ikke. Datainnsamlingen fra brukere er som regel beskrivende og informativ, og blir overført til utviklingsprosessen i form av verbalisering, notater, anekdote eller for eksempel skisser. En annen type feedback kan komme fra konsulentene, som benytter sin kunnskap og erfaring. Denne metoden har blitt veldig populær innen webutvikling.

2.5.3 Ekspertevaluering

Predictive evaluation eller *ekspertevaluering* som vi ofte det kaller det her i landet, er en metode hvor man bruker ekspertenes kunnskap for å forutsi problemer med brukskvaliteten. Ekspertene kombinerer ofte sin erfaring med kunnskap innen heuristikk og retningslinjer for å evaluere et produkt, enten det er det virkelige produktet, eller en prototype. Denne metoden utelukker brukerinvolvering, som igjen gjør denne metoden relativt kjapp og billig. Men den har en ulempe. Designere blir noen ganger ledet på villspor på bakgrunn av funn fra denne type evaluering (Preece et at, 2002).

2.5.4 CUT- Cooparative Usability Testing

Erik Frøkjær og Kasper Hornbæk (2005) presenterte teknikken *cooperative usability testing* (CUT). Meningen med denne teknikken er at den skal komplettere testing av brukskvalitet med en sesjon som inkluderer en brukerstøttet tolkningssamhandling. CUT reduserer på kravene til think aloud testing, men legger i stedet til rette for at brukeren skal ha innflytelse på fortolkningen av hva man gjennomfører på testen. Formålet med CUT er å hjelpe til med identifiseringen av brukskvalitetsproblemer med bedre validitet enn de som blir funnet ved hjelp av think aloud, samtidig forbedre brukerens og testernes erfaringer i å delta i en brukbarhetstest.

Ideen er å knytte sammen brukere og de som evaluerer til en konstruktiv dialog med mål om å avdekke brukskvalitetsproblem. Dette utføres gjennom en interaksjonssesjon som er rettet mot brukeren, der man utfører relevante oppgaver med systemet. Etter denne sesjonen gjennomføres en tolkningssesjon som er basert opp mot et opptak av interaksjonssesjonen.

Samhandlingssesjon (Interaction session - IAS)

Denne delen kan for eksempel utføres som en slags think aloud test. I denne sesjonen er det brukeren som er initiativtaker. Testbrukeren kan be om hjelp. Forfatterne anbefaler å benytte to personer som evaluerer; en guide, og en som skriver logg. Prosedyrene av evalueringen og evaluatorens rolle skal forklares nøye til brukeren i starten av sesjonen, og gjøre brukeren klar over at det er han eller henne som er den primært aktive personen. Denne sesjonen skal ikke vare lengre enn 45 minutter. Det er også viktig at lyden klar og bildet er godt.

Tolkningssesjon (Interpretation session - IPS)

Denne sesjonen utføres i samarbeid mellom bruker og de som evaluerer. Her skal man prøve å identifisere og forstå de viktigste brukskvalitetsproblemene som dukket opp i den forrige sesjonen (IAS). For de som ikke er godt trent i å oppdage problemer med brukskvalitet, kan denne fasen underbygges med en solide teknikker for inspeksjon av brukskvaliteten. Forfatterne anbefaler en metode som heter MOT (som de selv har utviklet).

Denne fasen kan begynne rett etter IAS-sesjonen, men det anbefales å ta en kort pause først. Fasen skal ikke vare lengre enn 45 minutter. Det anbefales at guiden og loggeren bytter rolle fra IAS sesjonen, slik at loggeren fra IAS sesjonen kan bruke notatene sine gjennom IPS-sesjonen. Loggeren i IPS-fasen skal hjelpe til med navigasjonen av opptaket og ta notater fra diskusjonen mellom guide og testbrukeren. Loggeren skal også passe på at guiden ikke går for detaljert til verks på uviktige utfall fra IAS.

Når videoen evalueres, peker guiden ut i samarbeid med brukeren sekvenser der det har oppstått brukskvalitetsproblemer. Det er viktig å følge med om brukeren er enig i uthevelsen av problem. Er han ikke skal man gå videre. Frøkjær og Hornbæk anbefaler å spille gjennom hele IAS i IPS-sesjonen.

I forhold til CTA og RTA

CUT skiller seg fra CTA og RTA fordi den har en interaksjonssesjon der brukeren ikke trenger å snakke høyt. Ifølge Frøkjær og Hornbæk er det heller ikke nødvendig å bruke TA-teknikken. CUT teknikken legger opp i større grad til samarbeid mellom testdeltager og tilretteleggere. Brukeren kan også stille spørsmål og be om hjelp under oppgaveløsningen. IPS fasen kan minne noe om den retrospektive fasen under RTA-metoden, men det legger heller ikke her vekt på TA-teknikken her.

2.5.5 Constructive interaction

Constructive interaction eller *konstruktiv samhandling* (også kalt *covery learning*) er en variant av CTA-metoden. Denne metoden involverer to testdeltagere som skal samhandle med en prototyp eller det aktuelle produktet. Denne metoden prøver å dra fordelen av samarbeid og fange den naturlige samtalen som oppstår under samarbeid. Mange er vant til å snakke sammen når de skal løse et problem, og prater ofte mer enn de ville ha gjort under en CTA-gjennomgang. Metoden har sine negative sider, slik som at mennesker har sine egne strategier for å lære og å bruke datamaskiner. Dette kommer klart fram i studie av Benedikte S. Alsa, Janne J. Jensen, Mikael B. Skova (2005), hvor de prøver ut konstruktiv samhandling på barn. Men også hvem som samhandler med hverandre kan ha innvirkning på resultatene; "…the pairing of the children had impact on identification of usability problems as acquainted dyads identified more problems both in total and of the most severe than non-acquainted dyads and individual testers. Finally, the acquainted pairs reported that they had to put less effort into the testing than the think-aloud and non-acquainted children" (s. 9)

Konstruktiv interaksjon passer best for prosjekt hvor det er lett å skaffe brukere, og hvor det er en fordel at disse er forholdsvis billig å bruke. Denne metoden krever dobbelt så mange testbrukere som CTA- eller RTA-metoden.

Vi har nå sett på teorien som danner grunnlaget for denne oppgaven, samt gitt et overblikk over hvilke metoder og teknikker som brukes for å måle brukskvaliteten. Vi har sett at CTA-metoden har mottatt kritikk for sin TA-variant, og at det finnes andre metoder som prøver å omgå CTAs påståtte svakheter. I forstudiet fokuseres det i større grad på gjennomføringen og detaljene rundt CTA, som blir utført i samarbeid med konsulent bedriften Kantega. Dette forstudiet er ment å gi en bredere innføring rundt metodens styrke og svakheter. Etter forstudiet sammenlignes RTA og CTA etterfulgt av et diskusjonskapittel

Kapittel 3 Forskningsdesign

3.1 Tilnærming av oppgaven

Brukbarhetstesting inngår i fag som *menneske-maskin-interaksjon* og *design av grafiske brukergrensesnitt*. Gjennom utførelse av disse brukbarhetstestene gikk det fram at testene må kunne gjøres på en bedre måte. Metoden som undervises ved NTNU er en meget forenklet utgave av CTA. Under disse testene fungerte TA-teknikken svært dårlig, fordi deltagerne sa lite under testen.

Masteroppgaven startet med temaet; ”fallgruver innen brukbarhetstesting”. Som vi har sett i innføringskapittelet, er ikke brukbarhetstesting en konkret metode, men en generell betegnelse på for flere retninger innenfor brukbarhetstesting. Tema og problemstillingen ble endret i løpet av studiet.

For å oppnå dypere innsikt og for å bli bedre rustet til å stille et mer presist og relevant forskningsspørsmål ble ett forstudium gjennomført. Forstudiet bestod hovedsakelig av fordypning i fagbøker og artikler, samtale med domeneeksperter og gjennomføring av ni brukbarhetstester (hvorav en var pilottest). Fokuset i forstudiet var å se på de positive og negative aspektene ved CTA og sammenligne metoden med en alternative metoder

3.1.1 Oppgavens særpreg

Denne masteroppgaven involverer situasjoner fra den virkelige verden. Dette er ikke en analyse av et abstrakt matteproblem, men blant annet et forsøk på å forske på interaksjon mellom mennesker, og mellom mennesker og maskiner ved hjelp av kvalitative og kvantitative metoder. Det vil bli rettet et kritisk blikk mot den enkelte metode, slik som den tradisjonelle brukbarhetstesten (CTA) og den alternative metoden RTA. Det må understrekes at oppgaven har et nøytralt utgangspunkt. Laboratoriet blir brukt som scene, og brukbarhetstestingen er handlingen. De som testes vil i utgangspunktet tro dette er

laboratorieforsøk, mens min agenda er å studere de menneskelige aspektene ved brukbarhetstesten. Dette kan forsvares etisk (se kapittel 3.3.2)

3.2 Forskningsstrategi

I 1967 skrev den norske professoren Johan Galtung; ”De som i dag lærer sosiologi uten å kunne matematikk, vil om tjue år ikke kunne følge med i de sosiologiske fagtidsskrifter” (Galtung 1967 referert av Repstad 1998: 11)

Dette viste seg å være feil. Nå, nesten 40 år etter, er de teknisk-matematiske sidene ved slike artikler sjelden dominerende. Denne masteroppgaven vil ha ett fleksibelt perspektiv med hovedsakelig kvalitative forskningsmetoder, men også kvantitative metoder. Begrepene *fleksibel* og *kvalitativ* viser til viktige trekk innen slik design. Disse tilnærmingene benytter utvalgte metoder for å få tak i kvalitative data, men fleksibel design benytter også metoder for å få tak i kvantitative data. Ved bruk av termen fleksibel kreves det mindre forberedelser av spesifikasjoner (i forhold til ”fixed design” som kan være forskningsstrategier som eksperiment). Et trekk ved laboratorieforskning eller eksperimenter, er at man i større grad må kontrollerer tilstandene (Robson 2002).

Oppgaven oppstår, utvikles og brettes ut ettersom forskningen beveger seg fremover. Under denne prosessen kan man bruke en miks av både kvalitative og kvantitative metoder. ”Målet med forskning er ganske enkelt å avdekke hittil ukjent kunnskap. Det ukjente kan også komme i form av uventede sammenstillinger av etablert kunnskap” (Hartvigsen 1998: 24)

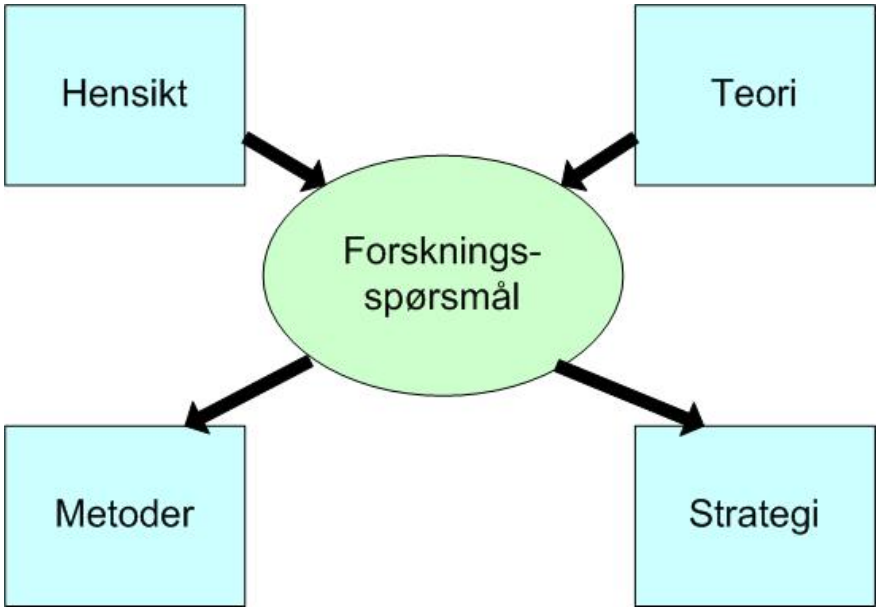
3.2.1 Valg av forskningsstrategi

For å finne en passende angrepsmetode til problemstillingen og ryggrad til oppgaven, ble forskningsstrategi innen fleksibel design studert. Med strategi menes den generelle brede orienteringen når man adresserer forskningsspørsmål. Det er en tidkrevende prosess, og enkelte strukturer kan oppleves som svært komplekse og vanskelige å få tak på (for eksempel grounded theory). Etter å ha studert forskjellige metoder og strategier, sto valget mellom *case studie* og *grounded theory*. For å være bedre rustet til å velge en adekvat strategi, ble veileder og spesialist innen *forskningsmetoder i informatikk* kontaktet. Det ble klart at det var unødvendig å gå i detaljer på en kompleks forskningsstruktur som groundet therory, fordi dette var for mye å kreve av en masteroppgave. ”The general principle is that the research

strategy or strategies, and the methods or techniques employed, must be appropriate for the questions you want to answer*"*(Robson 2002: 89)

På bakgrunn av studiets størrelse og type problemstilling, ble det enighet om å benytte et fleksibelt forskningsperspektiv, med vekt på kvalitative metoder. "Tre grep på gitaren kan være både til hygge og nytte i mange sammenhenger, men ytelsen bør ikke presenteres som symfonier av høy klasse."*(*Repstad 1998: 115)

Forskningsspørsmålet ble utviklet etter hvert i forskningens løp. Det var vanskelig å sette opp ett konkret forskningsspørsmål i starten, siden man bare ser symptomer på at brukbarhetstesten (CTA) ikke fungerer optimalt. For å sette merkelapper på problemenes lokasjon, benyttes en fleksibel tilnærnærming med åpent sinn i starten som snevrer seg inn etter hvert som studiet prosesserer. Robson (2002) har utviklet en modell som beskriver denne tilnærmingen;



Figur 3.1: Strategi for forskningsspørsmål

Denne modellen viser at hensikt og teori har påvirket forskningsspørsmålet, og dette bør være utgangspunkt for valg av strategi og metode. Avklaring av problemstillinger finner sted samtidig med at nye data samles inn og samtidig med at en bygger opp en analyse.

3.3 Forskningsmetoder

3.3.1 Kvantitative metoder

Under forstudiet ble det benyttet en spørreundersøkelse for å få et bredere perspektiv på hva man skal forske videre på. Intensjonen bak undersøkelsen var å på en forståelse på hvem testdeltagerne er (jobb, studier, alder og datakunnskaper), hvordan de opplevde å bli testet i en brukbarhetstest (enten ved hjelp av CTA eller CUT), hvordan de taklet å bli observert, og hvordan de håndterte TA-teknikken. Under denne prosessen fikk testdeltagerne et spørreskjema hvor de skulle evaluere opplevelse av testmetoden. Spørreskjemaet ga en pekepinn på hvordan testbrukerne opplevde situasjonen. Undersøkelsen ga et bedre utgangspunkt til å forbedre neste runde med tester.

3.3.2 Kvalitative metoder

Kvalitative metoder handler om å karakterisere (Repestad 1998). Vi kan si at kvalitative metoder går i dybden og ikke i bredden, og dette innebærer at vi studerer i helhet ett eller flere miljøer. Dette knytter da forsker og det miljø eller de personer som utforskes. Et annet kjennetegn er metodenes fleksibilitet. I motsetning til fast design (som er i mye større grad kvantitativ), kan man endre på spørsmålsstilingen underveis. Forandrer man på en kvantitativ spørreundersøkelse underveis, er dette vanligvis en grov synd. I et kvalitativt intervju kan svar på spørsmål resultere i nye oppfølgende spørsmål. Siden denne avhandlingen prøver å få innsikt i grunntrekk og særpreg under en brukertestsituasjon, er observasjon og kvalitative intervjuer gode alternativ. Kvalitative tilnærmingsmåter beskriver nyansert det som finnes og er mindre opptatt av hvor ofte det finnes. Denne masteroppgaven vil nok også ha kvantitative tilnærminger fordi det blir umulig å unngå mengdeangivelser og hyppighetsanslag av den kvantitative sorten. ”Syv av ti personer opplevde at..” og ”alle vi snakket med påpekte..” er utsagn som ofte vil dukke opp i mitt kvalitative studie. Kvalitative studier gir ofte ett godt grunnlag for å forstå konkrete, lokale utviklingsforløp (Repstad 1998:).

Observasjoner

Observasjoner er studier av mennesker for å se hvilke situasjoner de naturlig møtes i, og hvordan de pleier å oppføre seg i slike situasjoner (Repestad 1998). Normalt er observasjoner i forskningssammenheng at man studerer mennesker i bestemte kontekster, for eksempel på en arbeidsplass. En slik observasjon kan vare fra noen timer til flere år, og er avhenging av

hva man vil studere. I denne masteroppgaven blir observasjon benyttet for å studere menneskelig atferd under en brukbarhetstestsituasjon, noe som gjør at det er begrenset hvor lenge observasjonen skal vare. Observasjonene er nedtegnet observasjoner i form av notater som er grunnlaget for analysen, samt videoopptak med lyd og transkribering av hver brukbarhetstest.

Oppgavens observasjonsscene er litt spesiell siden det er en kunstig situasjoner studeres. Det må legges til at det ved observasjon skjer en tolkning fra forskerens side.

Skjult eller åpen observasjon

Ved skjult observasjon forteller en ikke aktørene, i dette tilfellet testdeltagerne, at en forsker. Ved åpen observasjon forteller man aktørene at man driver observasjon, men her trenger man nødvendigvis ikke gå detaljer innpå hvilke problemstillinger man er ute etter å belyse. Så hvor i bildet ligger denne masteroppgaven? Testbrukerne blir fortalt at de observeres mens de utfører ett sett med oppgave osv, men de får ikke vite at det er brukbarhetstestmetodene som studeres. Men dette kan forsvares.

Den nasjonale forskningsetiske komité for samfunnsvitenskap og humaniora (NESH) har som oppgave å opplyse om og gi råd om forskningsetiske spørsmål. ”De forskningsetiske retningslinjene har ikke samme funksjon som loven. Retningslinjene er et hjelpemiddel for forskerne selv” (NESH 2006: retningslinje 2). Noen av de etiske retningslinjene står skrevet i loven, og kan dog være straffbart å bryte. Denne avhandlingen tar uansett utgangspunkt i de retningslinjene som er foreslått. Når det gjelder hensyn til personer, er det lagt fram krav på å informere dem som forskes; ”De som er gjenstand for forskning, skal få all informasjon som er nødvendig for å danne seg en rimelig forståelse av forskningsfeltet, av følgene av å delta i forskningsprosjektet og av hensikten med forskningen. Det skal også informeres om hvem som betaler for forskningen” (NESH: utdrag retningslinje 8)

Altså, det skal informeres om;

- … at deltagelsen er frivillig
- … at det forskes på testmetoder, men ikke i detalj
- … at de blir tatt opp på video
- … at forskerne har taushetsplikt og at testdeltagerne forblir anonyme

- … at testpersonene kan avbryte når de vil
- … at all informasjon om personlige forholdt blir holdt konfidensielt
- … det skal ikke lagres noen opplysninger som kan identifisere enkeltpersoner

”Særlig aktsomhet må utvises når informasjon ikke kan gis før forskningen settes i gang, for eksempel dersom det ikke kan oppgis hva som er den egentlige hensikten med et eksperiment” (NESH 2006: utdrag retningslinje 8). Først etter at testdeltagerne er ferdig med oppgaveløsningen, vil de få detaljert informasjon om forskningens hensikt. Da vil de bli spurt om materialet fra testen kan brukes videre i forskning eller om det skal forkastes. Etter dette gjennomføres et kvalitativt intervju. Det er viktig at testdeltagerne ikke vet forskningens fulle hensikt fra starten av. Dette begrunnes med at vi vil at deltagerne skal oppføre seg som om dette var en vanlig brukbarhetstest. Hvis ikke er det en viss sjanse for at enkelte testdeltagere fokuserer mer på egen oppførsel. Det ville i så fall ha påvirket forskningseffekten.

Det opereres ikke med virkelige navn under denne forskningen for å vektlegge de respektive testpersonenes anonymitet. Siden denne avhandlingen ikke behandler personopplysninger (Personopplysningsloven 2000), faller dette prosjektet utenfor personsopplysningsloven.

Intervjuer

I løpet av studiet oppsto behov for å studere individers personlige erfaringer og oppfatninger. Kvalitative intervjuer ble utført for å få tak i testbrukernes tolkning av de begivenhetene denne avhandlingen går i dybden på. I tillegg ga intervjuene dypere innsyn enn spørreskjema og teoretisk fordypning, og de var med på å kartlegge brukbarhetstestsituasjonen. Dette var en tidkrevende prosess, men ga nærhet til feltet. Etter intervjuene ble fokus på forskningsspørsmålene mer konkretisert og endret noe fokus. Intervjuene tok sted etter brukbarhetstesten.

Siden oppgavens problemstilling involverer menneskelige aspekter, er uformelle kvalitative intervjuer en praktisk metode (Repstad 1998: 67). Et kvalitativt intervju går i dybden, og har ikke så fast struktur som kvantitative undersøkelser. Ved å erstatte intervjuet med for eksempel spørreskjema, kan gå glipp av menneskelige holdninger og erfaringer.

Intervjuguide

I denne forskningen er det benyttet en intervjuguide, men guiden ble ikke brukt slavisk. Det ble lagt vekt på å behandle alle punktene fra guiden, men spørsmålene ble ikke stilt i en fast rekkefølge. Poenget var å fokusere på det som ble sagt, og følge opp med oppfølgingsspørsmål. Denne type intervju kalles et semistrukturert intervju, og er mye brukt i fleksibelt design (Robson 2002).

Intervjuguiden som ble benyttet under andre runde av brukbarhetstestene ble utvidet etter hvert som nye fokus dukket opp under intervjuperioden. Noen av spørsmålene gjelder bestemte metoder, og dette er løst med at metodenavnet er skrevet med parentes til de gjeldende spørsmålene. Intervjuguiden kan leses i appendiks C.

Det ble forsøkt å formulere spørsmålene etter disse kriteriene;

- Unngå lange spørsmål.
- Unngå sammensatte spørsmål. Del heller sammensatte spørsmål opp i enkle spørsmål.
- Unngå sjargong. Bruk enkelt språk som testdeltagerne forstår.
- Unngå ledende spørsmål.
- Vær nøytral.

Det var vanskelig å forholde seg til kriteriene i de første intervjurundene, men etter hvert som man fikk praktisert intervjueteknikk ble intervjuene bedre.

Båndopptaker ble benyttet under intervjuene slik at man kan konsentrere seg i større grad om hva deltageren sier, siden en ikke trenger å skrive underveis. En kan fokusere i større grad på ikke-verbal oppførsel slik som kroppsspråk, tonefall og annen stemmebruk. I analysefasen kan det være en fordel å ha ordrett gjengivelse, spesielt når man skal transkribere resultatene. Intervjuene i dette studiet ble tatt opp med en mp3-spiller. Testpersonene ble informert angående bruk av båndopptaker, og opptakene ble slettet så snart studiet var avsluttet.

Etter datainnsamling av brukbarhetstester og kvalitative intervju ble dataene prosessert. For hver testdeltager ble en side med observasjoner notert på et ark. For eksempel hvem som var involvert, hvilke temaer ble dekket, relevansen i forhold til forskningsspørsmålet, forslag til

ny hypotese, implikasjoner for senere data innsamlinger. Også etter at selve intervjuet var avsluttet, ble nyttig informasjon gitt. Denne informasjonen ble notert og brukt videre i forskningen.

3.4 Analyse av de kvalitative dataene

En av farene ved kvalitativ datainnsamling er at man lett drukner i alle dataene. Det er lett at kvalitative data kumulerer. Materialet er ustrukturert og vanskelig å håndtere (Robson 2002). Derfor er det hensiktsmessig å lage en slagplan på hvordan dataene skal analyseres.

Allerede i løpet av forstudiet oppsto potensialet for å strukturere alle dataene som kom inn. Før påfølgende runde med brukbarhetstester, ble det lagt til rette for å håndtere større mengder med data. I løpet av datainnsamlingen, ble store datamengder redusert igjennom sammendrag, referater, koding, skrevne memoarer osv. Det ble lagt vekt på hva man skal ta vare på og hvordan det skal organiseres. Etter hvert ble mønster og regularitet oppdaget, og det ble enklere å organisere strukturer.

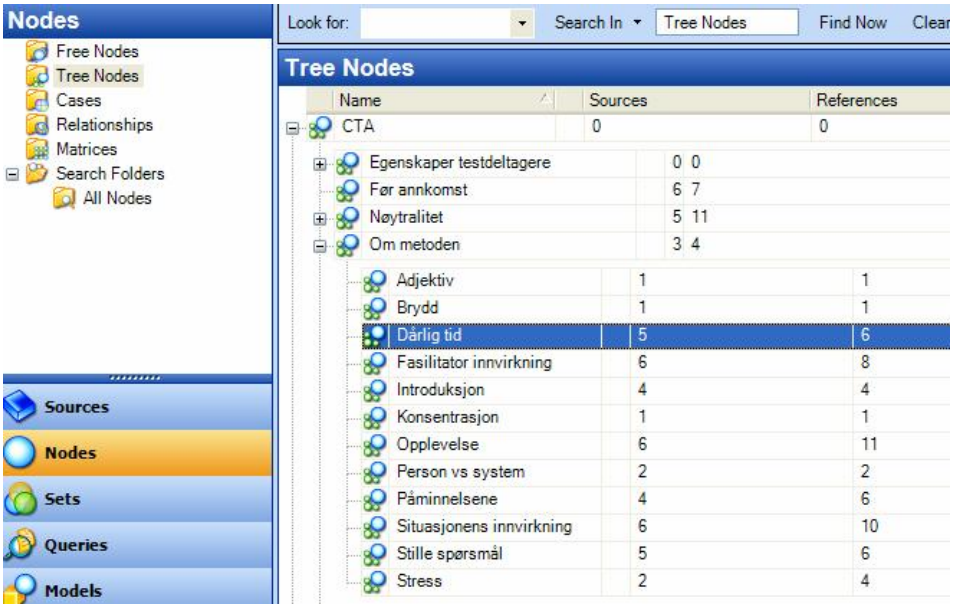
3.4.1 Koding av intervjuer

Analyse av kvalitative data blir ofte sett på som en vanskelig oppgave. Det er fordi analysen ikke er en mekanisk eller teknisk oppgave, men en mer dynamisk og kreativ resoneringsprosess (Robson 2002). Målet med å analysere kvalitative data er å bestemme kategorier, relasjoner og antagelser som informerer forskerens funn eller bidrar til å dra konklusjoner. Koder er merkelapper som brukes for å allokere ”enheter med mening”. En kode er et symbol som brukes for å klassifisere eller kategorisere en seksjon med tekst, som vanligvis tilknyttet er ”beholdere” av ord, fraser, setninger eller hele avsnitt. Beholderne kan være tilknyttede til en spesifikk setting. Koder kan være alt fra ”rett fram”-kategorier til mer komplekse kategorier, for eksempel en metafor. Ved å kode kan man oppdage relevante fenomen, samle eksempler av fenomenene og analysere fenomenene.

NVivo

Koding har spilt viktig rolle i oppgavens analyseringsfase. Både sammendrag, memoarer, transkriberte intervju og andre notater ble benyttet i kodefasen. Ved å dele opp de innsamlende dataene og velger ut kategorier, har likheter, forskjeller, mønster og strukturer oppstått. Kodene ble knyttet mot forskningsspørsmålene. I dette studiet ble programvaren NVivo fra *Qualitative research software for qualitative data analysis and research* (QSR

2007) kjøpt inn for å analysere de kvalitative dataene som kommer fra intervjuer, skrevne notater og andre dokumenter. NVivo er særlig egnet for store mengder data og gjør analysen mer effektiv og sikker (Robson 2002). Programmet fungerer på den måten at man koder de tekstene (også video hvis ønskelig) man har lastet inn.



Figur 3.2: Skjermbilde fra NVivo

Kapittel 4 Forstudie

4.1 Innledning

Etter en teoretisk fordypning ble det oppdaget forskjellige sider ved brukbarhetstesting hvor det manglet forskning; ”Er CTA en valid metode? Er RTA en valid metode? Hvordan opplever testdeltagerne testsituasjonen? Hvilke metoder avslører flest brukskvelitetsproblemer? Når skal man bruke hvilken metode?” Flere av disse temaene er heftig diskuterte, og parallelt med utviklingen med dette forstudiet erklærte Guan et al. (2006) RTA som en gyldig og pålitelig metode.

Utgangspunktet for dette studiet var å fokusere på de positive og negative aspektene ved CTA og sammenligne metoden med alternative metoder. Etter hvert som forskningen pågikk utviklet forskningsspørsmålet seg til å bli mer fokusert rundt andre oppdagelser. Hovedhensikten med forstudiet var å bli bedre kjent med brukbarhetstestsituasjonen for å oppnå en større innsikt til oppgaven. Det ble i den forbindelse opprettet kontakt med konsulentfirmaet Kantega som blant annet foretar brukbarhetstester i enkelte prosjekter. I samarbeid med Kantega ble det gjennomført åtte brukbarhetstester. Disse testene er utgangspunkt for forstudiet. Senere ble det gjennomført 13 ytterligere brukbarhetstester. ”Å gjennomføre en brukbarhetstest er et viktig aspekt i kvalitetssikringsprosessen fordi målet er å tidligst mulig å avdekke hva som generelt oppleves som problematisk, og om reelle brukere finner den informasjonen som ligger i løsningen. Resultatene fra en slik test gir gjerne innspill til videre utvikling og bekreftelse på om de prinsippene som er implementert så langt fungerer.” (Fra Kantegas brukbarhetstest mal)

4.2 Case

”SiT” er en forkortelse for ”Studentsamskipnaden i Trondheim”, og disse tilbyr en rekke tjenester for studenter. Kantega har i sitt samarbeid med SiT utviklet et nytt nettsted for studenter, samt et nytt intranett for de ansatte. Begge løsningene ble lansert i mars 2006.

Sit.no ble testet med CTA-metoden av fire studenter, og intranettet ble testet med CUT-metoden av fire personer. CUT ble valgt fordi den prøver å omgå de påståtte problemene til CTA-metoden (se kapittel 2.5.4). I ettertid ble det bestemt å droppe den mindre kjente CUT-metoden, fordi sammenligningrunnlaget var for dårlig. Resultatene fra CUT-metoden ble ikke brukt videre i oppgaven, både på grunn av sammenligningsgrunnlaget, men også fordi mye gikk galt under utførelsen av CUT.

SiT er en stor organisasjon som i skrivende stund består av 382 personer. SiT driver med alt fra forvaltning av boliger, barnehage, reisebyrå og kaféer til bokhandel.

4.3 CTA - forstudie

4.3.1 Litt om sit.no

Via nettstedet sit.no skal man gjøre informasjon relatert til Studentstudentsamskipnaden lett tilgjengelig for studentene. Her skal studenter finne det meste av informasjonen de behøver, og de skal kunne sende inn spørsmål til relevante kontaktpersoner. Det nye nettstedet har ett mer omfattende innhold enn det gamle. Løsningen har en todelt menystruktur, og siden lanseringsdatoen har SiT i samarbeid med Kantega jobbet med å legge inn informasjon.

Figur 4.1: Skjermbilde startsiden til sit.no

Informasjonsstrukturen er også i stor grad gjort om fra det forrige nettstedet. Etter ønske fra SiT.no skulle dette nettstedet testes for brukskvalitet ved hjelp av fire testdeltagere.

4.3.2 Planlegging

Planleggingen av testene ble utført i hovedsak av Kantega, SiT og skribenten av denne oppgaven. Planleggingen besto av møter, telefonsamtaler og e-postkommunikasjon. Det ble tidlig satt opp en liste med mulige problemområder som både SiT og Kantega ønsket å undersøke. Listen baserer seg på mulige problemer innen brukskvaliteten;

Testområde

- Under tab/fanen *bolig* kan vi f.eks. se på beboersidene. Hopping mellom boligtorget og sit.no kan være et problem.
- Vi kan se på om det som finnes under helse, rådgivning, barn og familie er logisk. Flere tema vil finnes under flere av disse kategoriene. Finner brukeren fram?
- Finn oppgaver som illustrerer bruk av eksterne linker i og utenfor meny.
- Ris og ros er noe som blir brukt, og som skal vise at SiT bryr seg. Kommer dette riktig fram?
- Oppgaver som ber brukeren finne konkrete ting, f.eks.:
    - Treningstidspunkt, treningslokale. Her er det veldig ulik bruk av timeplan, noe vi vil prøve å få fram i testen.
    - Leiligheter som passer brukerens familiesituasjon, f.eks. parleiligheter.
    - Bestille legetimer.
    - Finne storkiosken på Dragvoll.
    - Finne telefonnummeret til en person.
    - Finne dagens middag.
- Fokus på navigasjon.

Figur 4.2: Liste over potensielle problemområder

Formål

En selger kan ha andre formål enn en kjøper. Det kan være nærliggende å tro at en selger ønsker å tjene penger på å selge et produkt, men også at han ønsker å levere et produkt med høy brukskvalitet. Kunden vil også ha et produkt med høy brukskvalitet, men det skal kanskje aller helst være billig. Dette trenger ikke å stemme, men poenget er her å belyse alle formål. Kantega har utviklet en egen rapport som omhandler sit.no. ”Målet med å gjennomføre brukertester er generelt sett å så tidlig som muligavdekke hvor og på hvilke måter det planlagte grensesnittet gir reelle brukere problemer med å utføre arbeidsoppgavene sine og hva som generelt oppleves som problematisk”.

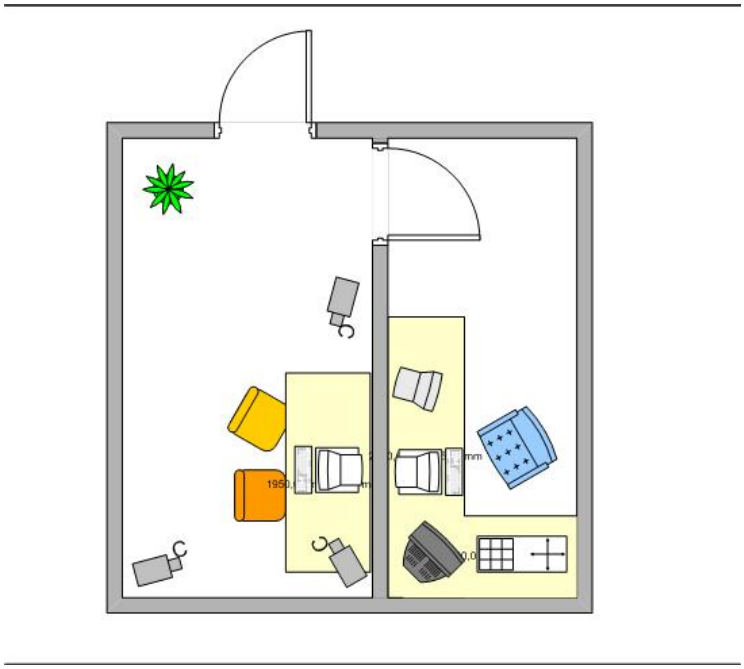
Videre utdypes at målet med testen er å se om brukerne finner den informasjonen de trenger og om de oppfatter nettstedet som nyttig. Kunden har også sine ønsker, noe Kantega må

forholde seg til. Målet med dette forstudiet er å evaluere selve testmetoden for å oppnå bedre forståelse av den praktiske gjennomføringen av CTA-metoden. Denne avhandlingen forsker på gjennomføringen av brukbarhetstestmetoder, mens Kantega fokuserer på de problemer som er knyttet opp mot brukskvaliteten. Alle tre parter har litt forskjellige formål, og dette hadde til en viss grad innvirkning på forstudiet. Dette involverer avgjørelser som at CUT og CTA ble gjennomført i ulike testomgivelser. Målgruppene var ulike (CTA-deltagerne besto hovedsaklig av studenter, mens CUT-deltagerne arbeidet i barnehage), og få testpersoner ble benyttet.

Målgruppe og omgivelser

Selv om CTA-deltagerne studerte ulike studieretninger på ulike steder, var de ganske like med hensyn på alder og internettbruk. De fleste studentene er unge og har gode kunnskaper med internett. Ingen av testbrukerne hadde brukt den nye utgaven av sit.no, men en av testbrukerne hadde brukt den tidligere versjonen for å lete etter bolig.

Testene ble gjennomført i et brukbarhetslaboratorium, bestående av to rom. Ett rom på ca 12 kvm og ett på 8 kvm. I det største rommet foregår selve brukbarhetstestingen. Dette rommet består tre kameraer, en mikrofon, en datamaskin, ett gruppebord og en skillevegg mellom datamaskinen og gruppebordet.



Figur 4.3: MMI-lab. Redigeringsrom til høyre og testdeltagerrom til venstre.

Det ene kameraet filmer testbruker forfra, det andre ovenfra og det tredje filmer bakfra. Det andre rommet er bygget for observasjon. Av utstyr finnes en TV, en datamaskin, headset, en redigeringsbenk, samt utstyr som får dette til å fungere sammen. Her kan man ta opp bilde og lyd, for så å brenne dette ut som DVD. Som kildemateriale for bilde kan det velges mellom opptak fra de tre kameravinklene og testbrukerens skjerm.

Oppgavene

Oppgavene ble utarbeidet i henhold til Dumas og Redish's retningslinjer (1999). Det ble i testen gitt 10-12 korte oppgaver. Sit.no er et informasjonsrikt nettsted, så testen måtte konsentrere seg om utvalgte deler av nettstedet. Det ble gjort et utvalg av oppgaver ut fra hva som ble ansett som viktig eller komplisert. Studentene ble bedt om å tenke seg at de var nye studenter i Trondheim; ute etter et sted å bo eller ute etter å finne informasjon om sitt studiested. Oppgavene ble notert på et ark som studentene fikk utlevert da testen startet. Deltagere som brukte kort tid på oppgaveløsningen fikk tilleggsoppgaver. Oppgavene var formulert som vist i figur 3.2;

Du har nettopp flyttet til Trondheim, og skal begynne å studere ved NTNU. Inntil videre bor du hos en venn Petter som anbefaler deg som ny student å benytte SiT.no for litt starthjelp.

1. Finn ut hva som er til middag i morgen der du studerer.

2. Kan du bestille time hos lege via nettstedet?

3. Kan du bestille time hos psykolog?

4. Du fikk lyst på sjokolade. Hvor ligger storkiosken på Dragvoll?

5. Åjsann, det ble visst litt mye sjokolade. Finn et sted du kan trene aerobic, på et tidspunkt som passer din timeplan.

6. Skriv ut timeplanen for det stedet du vil trene.

7. Du skal begynne å studere i Trondheim, og vil sjekke om SiT kan hjelpe deg med å finne leilighet. Du har samboer, men ingen barn. Finn en leilighet eller studentby du har lyst å søke om plass på. Sjekk om det er ledige plasser eller om det er lang kø der.

8. Du har oppdaget flere symptomer på stress. Du har liten tid til overs, og vil gjerne forhøre deg med en spesialist før du eventuelt bestiller time hos lege. Hvordan kan sit.no hjelpe deg?

9. Du vil gjerne gi tilbakemelding til SiT på hva du mener om det å finne fram på sit.no. Hvor kan du gjøre dette?

10. Hvor går du dersom du vil snakke med noen på SiT, og når er resepsjonen åpen?

Figur 4.4: Oppgaver

4.3.3 Gjennomføring

Testen ble gjennomført i samarbeid med prosjektlederen fra Kantega, og det ble vekslet på å ha rollen som guide og observatør. Testen ble gjennomført ved Norges teknisk-naturvitenskapelige universitet (NTNU) den 23. mai 2006. Testbrukerne var to jenter og to gutter. Brukerne testet løsningen en og en, og de fikk alle oppgavene utdelt på et ark. Under gjennomføringen av testen ble det gjort video- opptak av bruker og av brukers skjerm. Det ble også gjort lydopptak for å kunne høre kommentarer fra bruker. Det er laget en DVD med opptak fra testen, som brukes til evaluering. I forkant ble det utført en pilottest hvor selve testen ble evaluert. Meningen med pilottesten var å observere tidsforbruk og hvordan oppgavene fungerte, samt at det tekniske fungerer som det skal. Det ble gjort endringer i oppgaveformuleringen etter pilottesten.

CTA-metoden ble gjennomført etter Bruce Tognazzinis ti punkter for brukbarhetstesting. Dette er ikke en modell på linje med CTA, CUT eller RTA, men med mer konkrete retningslinjer på hvordan brukbarhetstester skal gjennomføres. De ti punktene for brukbarhetstester har etter hvert blitt sett på som en standard for hvordan man bør gjennomføre slike tester og er blant annet pensum i MMI relaterte fag ved NTNU.

- Det ble lagt vekt på høflighet. Det er mulig å få brukerne til å føle seg tryggere hvis testpanelet legger til rette for en hyggelig og avslappet atmosfære.

- Kameraer og mikrofoner ble pekt ut, og de fikk en rask innføring i hvordan brukbarhetstesten gjennomføres, inklusiv guiding på kontrollrommet.

- Det ble gitt klar beskjed om at det var systemet som skulle testes og ikke testbrukerne, og at det var verdifulle opplysninger å se hvor de hadde problemer.

- Testbrukerne kunne også avslutte når som helst.

- De kunne derimot ikke stille spørsmål under oppgaveløsingen, men de kunne skrive spørsmålene på ark, får så å stille de etter oppgaveløsingen.

- Det ble instruert og demonstrert hvordan man skulle snakke høyt, og poengtert flere ganger hvor viktig det var at testbrukerne snakket høyt under hele oppgaveløsningen om ikke bare hva de gjorde, men også hva de vurderte og hva de forventet hva ville skje, osv. Etter demonstrasjonen fikk de selv øve på teknikken. Dette pågikk inntil vi så at de behersket teknikken. Mobiltelefon ble bl.a. brukt som objekt, slik at de skulle si høyt hva de tenke mens de for eksempel skulle slå den på, sende melding eller ringe.

- Oppgavekonteksten ble forklart og nettstedet sit.no ble introdusert.

- De ble spurt om det var noe de lurte på før testen ble kjørt. Det ble notert ned hvilke problemer testpersonen hadde når han/hun utførte de forskjellige oppgavene, samt relevante spørsmål som ble avdekket underveis. Det ble ikke grepet inn når ting gikk galt, bortsett fra ved spesielle anledninger, for eksempel når testbrukeren føler seg ukomfortabel eller vil gi opp testen. I en slik situasjon ble brukeren gradvis hjulpet forbi problemet (det ligger mye verdifull informasjon i en slik situasjon).

- Testen ble avsluttet med en debrief-sesjon. Det ble spurt om konkrete sider ved designet som brukeren hadde problemer med når han/hun skulle gjøre de forskjellige oppgavene. De ble bedt om en subjektiv vurdering og eventuelle forslag til redesign.

4.3.4 Testresultater

Testresultatene er utarbeidet i samarbeid med kantega. Resultatene er noe endret i forhold til Kantegas egen rapport, fordi noen av kommentarene var irrelevante i forhold til dette studiet. Alle kommentarer er dog skrevet av representanter fra Kantega.

Skjermbilde/ Funksjon	Problembeskrivelse og andre observasjoner	Kommentar
SiT.no		
Dagens middag	Link til dagens middag ligger nederst i menyen og i en egen boks nederst på forsiden. Boksen ble aldri sett, og brukerne gikk innom kafèer først i menyen før de så dagens middag.	Dagens middag kunne vært flyttet opp hvis den er forventet å være noe av det mest brukte. Legg også til link på kafèer.
Boligsidene	Disse sidene er ikke så godt tilpasset nye studenter. Det virker som det er forventet at brukere skal kjenne til studentbyene. Det er mye tekst på midtsidene, noe som tar fokus vekk fra venstremenyen. Ledige boliger, søknadsprosessen osv. er godt gjemt under listen over studentbyer.	Listen med studentbyer bør inneholde kun studentbyer. Link til annen info forsvinner når det ligger under listen. Vurder om noe av teksten på midtsidene kan klippes, evt. gjøres mer oversiktlig.
Kniven/Forsiden	Kun en bruker brukte kniven. Da testene ble gjennomført var det veldig mye informasjon på forsiden, samt mange forskjellige farger og typer linker. ”Alt” er viktigst og krangler om oppmerksomheten. Det som var under kniven ble aldri brukt, og flere nevnte at dette ga et rotete preg på forsiden.	Vurder hvordan førstesiden kan ryddes. Her er mange hurtigmenyer, kniven er bare en av dem.
Timeplaner	Det var ikke så enkelt å finne timeplaner for treningsstedene - disse kunne f.eks. vært linket fra treningstilbud. Ingen kommenterte at timeplanene var ulike, men det var vanskelig å lage en oppgave som gjorde at de måtte lese flere timeplaner. Det er kanskje heller ikke en realistisk problemstilling.	Oppgaven ledet brukeren til først å finne hva som er interessant å trene, og så finne ut når.

Tunge midtsider.	Generelt var det en observasjon at testbrukerne fokuserte på midtdelen av en side, og brukte venstremenyen når de ikke fant det de skulle. Noen sider har for ”tunge” midtsider, blant annet legesentersiden har en høyremeny som overdøver den til venstre.	
Utskrift	Mange av testbrukerne lette etter ikon for å skrive ut sidene. Noen trodde også at header og menyer ville være med på utskriften hvis man valgte ”Fil→Skriv ut”. Mye usikkerhet.	
Trening	Treningssidene opplevdes som noe vanskelig å finne fram på. Brukerne var ikke sikre på om de skulle velge ”Treningstilbud”, ”Treningsanlegg” eller ”Timeplaner” for å finne ut hva de vil trene og hvor. I tillegg er det innført en høyremeny under timeplaner.	Det er ikke sikkert det skal så mye til for å strukturere disse sidene bedre.
Helse	Dette er også sider med mange navigasjonsmuligheter til samme informasjon, men få så de lilla boksene med linker til høyre i vinduet. Det ser ut til at menyen til venstre også drukner i all informasjonsmengde.	Vurder å forenkle antall veier inn til hvert sted, sånn at brukerene læres opp til å bruke venstremeny.
Kontaktinfo	Det var to oppgaver ang. kontakt: finn hjelp til internettproblemer, og finn resepsjon. Internettproblemer var vanskeligst å finne, selv om de som fant hjelp på boligsidene opplevde det som en helt logisk plassering.	

Tabell 4.1: Testresultater fra CTA gjennomgang

4.4 CUT - forstudie

CUT skiller seg ut fra den tradisjonelle testmetoden CTA. CUT består to sesjoner, en IAS-fase hvor testbrukeren skal samarbeide med tilrettelegger og en IPS-fase hvor tilrettelegger bytter rolle med observatør. Under IPS-fasen går observatøren gjennom IAS-fasen sammen med brukeren, hvor de snakker ut om oppgaveløsningen. For mer informasjon, les om ”CUT” i kapittel 2.5.4.

Denne runden med tester ble ikke like vellykket som testingen av nettstedet fordi det oppsto problemer underveis. Det tekniske utstyret ble testet dagen før, og alt fungerte som det skulle. I denne omgang ble kun ett kamera og en mikrofon benyttet. Etter evaluering av første testbruker, ble det sjekket om opptak og lyd fungerte som det skulle. Det virket tilsynelatende OK. Først da filene ble lagt over PC, ble det oppdaget at det var kameraets mikrofon som tok opp lyden, og ikke den eksterne mikrofonen. Hvorfor dette skjedde er usikkert. Dette førte til dårlig lydkvalitet pga kameraets avstand til testbrukerne, som gjorde det til tider umulig å høre hva som blir sagt. Utfallet av dette gjorde det lite hensiktsmessig å transkribere opptaket, siden det ville bli stort sett tipping og mangel av sitater. Et annet problem var at testbrukerne brukte mye lengre tid enn først antatt. Underveis måtte det anskaffes mer teip til kameraet fordi tiden ble feilberegnet. Det oppsto også problemer da opptaket skulle redigeres og overføres til DVD. Meningen var at innspillingen skulle kombineres med opptak av skjerm. Opptakene resulterte i ulik avspillingslengde, noe som gjorde det svært vanskelig å få synkronisert opptakene. Det hele endte med at begge opptakene ble benyttet separate. Alt dette (bortsett fra rekrutteringen) kunne vært unngått med en mer omfattende pilottest, ved å inkludere en testbruker og redigering av opptakene etterpå.

Gjennomføringen av CUT var ikke helt intakt i henhold til Horbæk og Frøkjærs anbefalte retningslinjer; “Our impressions from reviewing the IPS videos are that replay of the IAS in full length gives a more easygoing, yet very focused session compared to presenting only the important usability problems” (2005: 1386)

I utgangspunktet var meningen å bruke hele opptaket I IPS-fasen, men det var knappe tidsintervall mellom hver test (etter kundens ønsker). Derfor ble det fokusert på områder der

testbrukerne hadde brukskvalitetsproblemer. Dette er ikke stikk i strid med CUTs retningslinjer, men heller ikke anbefalt.

Et annet problem med valg av metode, er at denne metoden ikke er spesielt kjent. I denne forskningen ble det ikke funnet noen referanser til Hornbæk og Frøkjærs studie. Det kan også virke som om modellen fremdeles er på prøvestadiet; "Further work is needed to develop better support for the interaction and interpretation session of CUT" (s 1386). På grunnlag av alt dette ble det bestemt at resultatene fra CUT ikke skulle brukes videre i denne forskningen. Det ble derimot avdekket flere brukskvalitet problem som ble rettet på av Kantega.

4.4.1 Observasjon

Utførelsen av CUT ble på ingen måte bortkastet. Den ga innblikk i alternative måter å avdekke brukskvalitetsproblemer, og den poengterte hvor viktig det var med en godt gjennomført pilottest.

Vi observerte blant annet at det ble sagt svært lite under IAS-fasen, og det var lite som minnet om samarbeid. Dette kan ha årsak i at testbrukerne hadde liten kunnskap med data, og at de brukte mye tid på å orientere seg på intranettet, samt hvordan teknologien fungerer. Det var ellers lite respons under IPS-fasen.

4.5 Observasjoner fra forstudiet

Etter forstudiet sitter vi igjen med mange interessante observasjoner, og da spesielt fra CTA-metoden. Forstudiet inklusiv teorifordypningen har ført til at man på kloss hold har belyst problemer ved den mest brukte brukbarhetstest metoden.

4.5.1 Utført på flere måter

Til tross for at det på forhånd ble definert hvordan CTA-metodene skulle gjennomføres, fungerte ikke dette i praksis. Prosjektlederen fra Kantega gjennomførte en litt annerledes variant av CTA enn Bruce Tognazzini (1991) sine retningslinjer. Prosjektlederen åpnet for en bredere dialog mellom henne og testbrukeren i stedet for påminnelser som ”fortsett å snakk” og ”hva tenker du nå?”. Dette begrunner hun med; ”Det er fordi det føles mest naturlig. Som du har lagt merke til har jeg ikke fulgt teoriene strengt, det er jo ikke mulig. Når det er faktiske folk du har med å gjøre. Jeg prøver å ikke hjelpe dem, men tror situasjonen blir mindre ubehagelig om vi har en viss dialog” (Personlig kommunikasjon via epost). Boren og Ramey (2000) viser til publisert litteratur som refererer til det klassiske verket til Ericsson og Simon, men at disse prinsippene sjeldent er artikulerte. De påpeker at CTA-metoden gjennomføres mindre rigorøst enn i teorien, og at gjennomføringen har mange motsetninger til Ericsson og Simons anbefalinger. Dette fører til at think aloud læres bort i mange varianter, som er stikk i strid med teorien. Dette gjør det vanskelig å innestå for resultatenes validitet, samt vanskelig å lære bort en standard som nybegynnere skal praktisere. I skrivende stund forskes det på i hvilken grad CTA-metoden kan godtas som en valid metode (Guan et al. 2006).

4.5.2 Det å si høyt hva man tenker versus Stillhet

Etter å ha transkribert alle gjennomføringene av CTA ble det registrert perioder med mye stillhet hos alle testbrukere. Enkelte av testbrukerne mumlet relativt mye under testen, noe som gjorde det veldig vanskelig og til tider umulig å høre hva testbrukerne sa. Ved hjelp av en stoppeklokke ble prosenter av prating målt og notert. Meningen her er å kartlegge stillhet og å se om det er sammenheng mellom vanskelighetsgraden av oppgaver og stillhet. Disse tallene er ikke hundre prosent nøyaktige, men skal gi en god pekepinn. Tallene er oppgitt i prosent av den totale tiden det tok fra de startet på oppgaveløsingen til de var ferdige.

Testbruker	Prater aktivt	Mumling eller stillhet
#1	21 %	79 %
#2	26 %	73 %
#3	17 %	83 %
#4	22 %	78 %

Tabell 4.2: Aktiv prat i forhold til stillhet - CTA forstudie

Ut fra videoopptakene tyder mye på at testdeltagerne snakker mindre når de står fast. Dette er ingen ny oppdagelse (Preece et al. 1994), men er viktige målinger når man skal forsøke å få innsikt i deltagernes mentale prosesser.

4.6 Spørreundersøkelsen

Spørreundersøkelsen (se appendiks D) fokuserte på hvordan testbrukerne opplevde å bli testet. Kolonnene til venstre indikerer nr på testperson. # 1 betyr testperson nr 1.

I hvilken grad de opplevde de stress under selve oppgaveløsningen

	Høy	Middels	Liten	Ingen	Vet ikke
# 1			X		
# 2			X		
# 3			X		
# 4			X		

I hvilken grad det var vanskelig å verbalisere sine tanker

	Høy	Middels	Liten	Ingen	Vet ikke
# 1	X				
# 2		X			
# 3			X		
# 4			X		

I hvilken grad de følte seg komfortabel med å verbalisere sine tanker

	Høy	Middels	Liten	Ingen	Vet ikke
# 1	X				
# 2	X				
# 3		X			
# 4		X			

Hvor ofte de sluttet å snakke under oppgaveløsningen

	Ofte	Av og til	Sjelden	Aldri	Vet ikke
# 1		X			
# 2		X			
# 3		X			
# 4		X			

Hvor ofte de tenkte raskere enn du greide å verbalisere

	Ofte	Av og til	Sjelden	Aldri	Vet ikke
# 1		X			
# 2		X			
# 3		X			
# 4	X				

I hvilken grad de følte seg overvåket

	Høy	Middels	Liten	Ingen	Vet ikke
# 1			X		
# 2			X		
# 3		X			
# 4			X		

I hvilken grad oppfattet de situasjonen som kunstig

	Høy	Middels	Liten	Ingen	Vet ikke
# 1			X		
# 2	X				
# 3			X		
# 4			X		

I hvilken grad de følte seg komfortabel gjennom hele testen

	Høy	Middels	Liten	Ingen	Vet ikke
# 1		X			
# 2		X			
# 3			X		
# 4			X		

Tabell 4.3: Spørreundersøkelse

4.6.1 Det å verbalisere tanker

Flere av spørsmålene fokuserte på det å verbalisere sine tanker. En av testbrukerne syntes det var vanskelig å si høyt hva man tenker, to av testbrukerne syntes det var av middels grad vanskelig, mens sistemann følte det i liten grad. Altså følte alle testbrukerne en viss grad av vanskelighet med å verbalisere sine tanker. På spørsmålet i hvilken grad de følte seg komfortabelt med å snakke høyt, svarte to av testpersonene ”høy”, mens to svarte ”middels”. At man i middels grad føler seg komfortabel, kan bety at de til tider følte seg ukomfortabel. Det kan også bety at de følte noe fra eller til, eller at de hadde det bra men de har en høyere list for å oppnå begrepet *komfortabel.* Samtlige av testpersonene svarte at de av og til stopper å snakke under oppgaveløsningen. Når man ser på figur 3.6, skulle man heller tro at de kom til å svare sjeldent. Man kan spørre seg om hvorfor spørreundersøkelsen gir et litt annerledes bilde enn hva de kvantitative resultatene i tabell over sier. Når testbrukerne er stille mellom 73-83% av tiden, skulle man tro at de svarte ”sjeldent” framfor ”av og til”, men dette er også en definisjonssak på hva som menes med av og til. Uansett indikerer begge resultatene at de stopper opp med å snakke. Det siste spørsmålet som omhandler TA-teknikken var om de tenkte raskere enn de greide å verbaliserte sine tanker. En av testbrukerne svarte ”ofte”, mens de tre andre svarte ”av og til”. Det kan virke som om dette er med på å begrense informasjon vi får fra testbrukerne.

4.6.2 Komfort

Hvorfor er det viktig at testdeltagerne føler seg komfortabel under en brukbarhetstest? Er det ikke viktigere å få fram gode resultater framfor å skape hygge? Det er i alle fall viktig at det må legges til rette for at omstendighetene i minst mulig grad skal påvirke resultatene på en negativ måte. Det er nærliggende å tro at om testdeltagerne føler seg lite komfortabel så kan dette ha innvirkning på resultatene. Et eksempel er om testdeltageren er av nervøs og beskjeden karakter, og to brautende personer får testdeltageren til å føle seg ukomfortabel, kan det resultere i at testdeltageren sier mindre og mumler mer enn hvis han eller henne hadde følt seg komfortabel. Det kunne også ha resultert at testdeltageren hadde blitt så nervøs at han eller hun hadde gjort flere navigeringsfeil. I utgangspunktet er det lagt opp til at testbrukerne skal føle seg komfortabel. Dumas og Redish (1999) har skrevet et kapittel, *Caring for the Test participants,* om hvordan man skal ta seg av testdeltagerne. ”Making them comfortable and caring about them and for them throughout the test is part of the test team’s job” (s 274)

Sett fra et subjektivt synspunkt ble det lagt stor vekt på å få testdeltagerne til å føle seg komfortabel. Men ut fra testdeltagernes dom, følte to deltagere seg middels komfortabel, mens to følte seg i liten grad komfortabel. Når man ser på måten testen ble gjennomført på, ble både TOG- og Dumas og Redish’s retningslinjer fulgt opp.

4.6.3 Overvåking

Et annet spørsmål som ble stilt før testen var i hvilken grad overvåkningen hadde på resultatene. På spørsmål om i hvilken grad testbrukerne følte seg overvåket, svarte tre ”liten”, mens den siste person krysset av for ”middels”. Det kan her virke som om overvåkningseffekten har liten innvirkning.

4.6.4 Kunstig

På spørsmål om deltagerne oppfattet situasjonen som kunstig, svarte tre av fire ”i liten grad”. Et spørreskjema går ikke tilstrekkelig i dybden, så det er vanskelig å finne ut hva som ligger bak ”i liten grad”. Ingen av testdeltagerne svarte ”ikke i det hele tatt”, mens en person svarte ”i svært høy grad”. Ut fra disse resultatene, ble det bestemt å gå i dybden for å finne ut hva som ligger bak.

4.7 Videre fokus

Gjennomføringen av forstudiet var svært lærerikt. I ettertid er det lett å se viktigheten av en godt planlagt test, som bl.a. inkluderer en omfattende pilottest. Det skal sies at runden med CTA-testene nesten gikk smertefritt, men det ble lagt litt for liten vekt på TA-trening og påminnelser underveis.

Meningen med forstudiet var å bli bedre kjent med brukbarhetstestsituasjonen. Spørreundersøkelsen og observasjon avdekket problemområder som ga grunnlag for videre forskningen i masteroppgaven. Vi har sett at CTA utføres i ulike varianter og at dette har resultert i at flere har stilt spørsmålstegn ved metodens validitet. Man kan spørre seg om det er poeng i at metoden må være valid så lenge den avdekker brukskvalitetsproblemer, men hva hvis testsituasjonen påvirker resultatene? Vi skal ikke videre avgjøre om CTA er valid eller ikke, men vi belyser temaet. Det vi skal se nærmere på er om og eventuelt hvordan testsituasjonen påvirker resultatene. Vi har sett at testbrukerne er stille ca 80 % av tiden, og at de synes det er vanskelig å snakke høyt. Samtidig antydes det at de føler seg mindre komfortabel under en test. Dette følges opp i de påfølgende kapitlene.

CUT blir lagt på hylla pga vanskelig sammenligningsgrunnlag og blir erstattet med en annen metode. RTA er en metode for brukbarhetstesting som påstår å komme rundt problemene med dobbel kognitiv belastning. RTA har mange fellestrekk med CTA men noen vesentlige forskjeller er det. Dette blir behandlet i kapittel 6

Etter forstudiet fikk man en pekepinn på at diverse faktorer kan ha innvirkning på resultatet. Dette er tema som videre følges opp i påfølgende kapittel; Verbalisering av tanker, kognitiv belasting, overvåkningseffekt og kunstige situasjoner. I tillegg skal vi se hvilken metode som er best til å avdekke brukskvalitetsproblem

Kapittel 5 CTA versus RTA

5.1 Innledning

I forstudiet avdekket vi en rekke brukskvalitetsproblemer. Noen av disse ble forsøkt forbedret av Kantega, mens andre ble utelatt. Runde 2 av brukbarhetstesten fokuserer i stor grad på de samme brukskvalitetsområdene som i forstudiet, men her er det stilt strengere krav til gjennomføring av metode. Denne runden ble utført uten hjelp eller samarbeid fra andre, fordi en da slipper å ta hensyn til andre interessenters formål. Planleggingsfasen var langt mer omfattende, og det ble lagt større vekt på utvalget av testbrukerne. TA-teknikken ble grundig gjennomgått, og testbrukerne måtte praktisere før de startet på oppgaveløsningen. CTA-deltagerne øvde før oppgaveløsningen, mens RTA-deltagerne øvde seg før den retrospektive sesjonen. Etter hver brukbarhetstest ble deltagerne intervjuet.

Dette kapitelet fokuser det å avdekke brukskvalitetsproblemer, og ser på hvilke problem CTA og RTA avdekker innen brukskvalitet. Siden disse intervjuene er en viktig del av denne forskningen, ble det bestemt at det bør gjennomføres seks brukbarhetstester av hver metode. I følge Nielsens (2000), vil fem testpersoner kunne avsløre 80 – 85 prosent av problemene med brukskvaliteten. Første pilottest ble gjennomført 29. oktober 2006.

”A practical guide to usability testing” skrevet av Joseph S. Dumas og Janice C. Redish (1999) er blitt benyttet som mal i denne iterasjonen. Dette gjelder i hovedsak:

- Planlegging av neste runde med brukbarhetstester
- Definere mål
- Valg og formulering av oppgaver
- Hvordan man skal måle brukskvalitet
- Utførelse av pilottest

I runde 2 med brukbarhetstester av sit.no, ble tolv tester utført i perioden 20. oktober til 1. desember. Det ble nødvendig å kjøre en ekstra test fordi et av opptakene manglet lyd.

5.1.1 Planleggingen

Med utgangspunkt i sunn fornuft fra testområdene og gjennomføringen beskrevet i forstudiet, ble det satt opp en liste med konkrete mål for utførelsen av oppgavene. Siden dette i hovedsak er et informasjonsnettsted med liten funksjonalitet, er navigering og struktur veldig sentralt. Hvor ligger informasjonen? Hvor lett er det å finne den? Hvordan navigerer man seg fram? Hvor forståelig er informasjonen? Dette er spørsmål som er svært sentrale i denne brukbarhetstesten.

Pilottest

Under pilottesten ble det lagt vekt på å teste både teknisk utstyr, forskningsmetode, spørsmål og gjennomføring. Pilottesten besto av en full gjennomføring av en brukbarhetstest (utført 29.oktober 2006), etterfulgt av analyse og rapportering av resultatene. Dette inkluderer en analyse av brukskvalitetsproblem, transkribering av intervjuet, og analyse av transkriberingen. Pilottesten gikk som planlagt, og det var ingenting uvanlig å rapportere.

Testbrukerne

Siden testen blir gjennomført på et nettsted som er beregnet på studenter, er alle testbrukere studenter. Alle testdeltagerne som er med gjennomførte emnet *Datastøttet læring* høsten 2006. Studentene har forskjellig bakgrunn og kompetanse innen IT. Noen er IT-studenter, noen går lærerutdanningen, mens andre tar realfag som matematikk og fysikk. De fleste av testbrukerne er nye studenter. ”One of the most cardinal rules of usability testing is that people who work with the product in the usability test must be like the people who will actually use the product.” (Dumas & Redish, 1999: 120)

Når det gjelder bruk av Internett, har alle testbrukerne grunnerfaring. Det vil si alle at alle studentene skal være i stand til å finne fram til den informasjonen som er tilgjengelig på sit.no. Testdeltagerne hadde varierende erfaring med datamaskiner, men alle hadde det grunnleggende inntakt. Et par av testdeltagerne hadde i tillegg noe kunnskaper innen webutvikling. De fleste av testbrukerne var førsteårstudenter, mens resten hadde flere år på baken.

Dumas og Redish (1999) foreslår å dele testbrukerne opp i grupper. Hver gruppe deler felles karakteristikker, som for eksempel de som har erfaring med en applikasjon og de uten

erfaring. Siden et nettsted skal passe alle studenter, ble studenter fra ulike campuser valgt som deltagere. Alderen på deltagerene varierte fra ca 20 til 30 år.

5.1.2 Mål og antagelser

”Even with a simple product, so much happens so quickly in a usability test that if you have not thought about what to focus on, you may miss important events. For each usability test, therefore, you have to start by considering what you want to learn-that is, by defining specific goals and concerns.” (Dumas & Redish 1999)

På grunn av oversettelsesproblemer med å finne et godt norsk ord til det engelske uttrykket ”concerns”, brukes *antagelser* som erstatningsord. Antagelser betyr i denne sammenhengen områder vi antar eller er usikre på om kan ha brukskvalitetsproblemer, idéer vi vil teste ut, eller bare spørsmål vi vil ha svar på. Antagelser kan i noen tilfeller brytes ned ett nivå. Disse antagelsene ble satt opp:

1. Det å finne fram på SiT.no
    a. Hvor raskt setter nye brukere seg inn i strukturen?
    b. Vil det bli foretatt hyppige navigeringsfeil?
    c. Hvordan kombineres bruken av toppmenyen og høyremenyen?
    d. Hvor lang tid tar det å utføre en oppgave?
    e. Finner testbrukerne det de leter etter på de forventede plassene?
    f. Hvordan er forholdet mellom mengde tekst og oppdeling av undermenyer?

2. Relasjon mellom boligtorget og SiT.no
    a. Vil relasjonen mellom boligtorget og SiT.no være et problem?
    b. Kommer testbrukerne til å besøke boligtorget?
    c. Hvis a, vil brukerne bli forvirret av denne kombinasjonen?

3. Ris og ros
    a. Ris og ros skal vise at SiT bryr seg. Kommer det fram?

   b. Hvor synlig er ris og ros linken?
   c. Bruker testbrukerne linken i fanen øverst til høyre?
   d. Hvis c, når fikk de øye på den?

4. Trening og timeplaner
   a. Hvordan opplever testdeltagerne strukturen av timeplanene?
   b. Vil ulik bruk av timeplaner være forvirrende?

5. Nettrådgivning
   a. Vil testbrukerne benytte rådgivningslinken i toppmenyen?
   b. Hvis nei, hvor leter de?
   c. Er nettrådgivningen synlig plasser?
   d. Vil testbrukerne forstå denne tjenesten?

6. Kontakt
   a. Hvordan oppsøker testbrukerne kontaktinformasjon?
   b. Hvilken kontakt tjeneste benytter de?
   c. Er det lett å finne riktig kontaktinformasjon?
   d. Er det lett å få tak i de aktuelle personene som er ansatt i SiT?
   e. Hvordan finner testbrukerne fram til aktuelle ansatte?

7. Organisasjon
   a. Vil testbrukerne finne organisasjonskartet?
   b. Vil testbrukerne forstå organisasjonskartet?
   c. Hvordan tolker testbrukerne symbollinkene i organisasjonskartet?

Ut fra dette kan man lage oppgaver til testen, sette konkrete mål, sette opp en test og finne ut hva man skal måle.

Testoppgavene

De fleste av oppgavene fra forstudiet ble brukt igjen i denne iterasjonen, mens noen ble forkastet og erstattet med nye. Noen av oppgavetekstene er forandret, slik at de skal bli enklere å forstå. Med utgangspunkt fra forrige runde med brukbarhetstester og antagelsene satt opp i denne runden, ble disse testoppgavene satt opp:

1. Finn ut hva som er til middag i morgen der du studerer

2. Kan du bestille time hos lege via nettstedet?

3. Kan du bestille time hos psykolog?

4. Du fikk lyst på sjokolade. Hvor ligger storkiosken på Dragvoll?

5. Åjsann, det ble visst litt mye sjokolade. Finn et sted du kan trene aerobic, på et tidspunkt som passer din timeplan.

6. Skriv ut timeplanen for det stedet du vil trene.

7. Du trenger en plass å bo i Trondheim, og vil sjekke om SiT kan hjelpe deg med å finne leilighet. Du har samboer.

8. Finn en leilighet eller studentby du har lyst å søke på, og se om det er ledige plasser eller venteliste.

9. Du har oppdaget flere symptomer på stress. Du har liten tid til overs, og vil gjerne forhøre med en spesialist før du eventuelt bestiller time hos lege.

10. Finn et organisasjonskart over SiT

11. Skriv ut en oversikt over alle ansatte som hører til under avdelingen ”bolig”

12. Hvor kan du gi tilbakemelding til SiT?

13. Når er SiT’s ekspedisjon åpen?

Figur 5.1:Oppgaver runde 2

Mål

Med bekymringer og oppgaver som utgangspunkt er disse konkrete målene satt opp:

Oppg. #	Mål
Alle	Det skal være enkelt å navigere seg fram til tema med toppmenyen.
3, 6, 11	Disse oppgavene bør kunne løses innen 30 sekunder, fordi man enten kun skal skrive ut fra et skjermbilde man allerede er i (skriv ut timeplan), repetere en tidligere oppgave (bestill time hos psykolog) eller finne en enkel link som er synlig hele tiden (ris og ros).
1, 2, 4, 5, 8, 9, 10, 12	Dette er typiske navigasjonsoppgaver, hvor man kanskje må se litt i menyene og teksten for å finne fram. Målet er maks 1 minutt pr oppgave.
7	Denne oppgaven krever litt tid i forhold til de andre oppgavene. Her skal man i utgangspunktet finne en aktuell plass å bo før man leter i ventelisten. Det er å forvente at dette bør kunne gjøres på under 2 minutter, men må se ann hvor mye tid de bruker på å finne ønsket bolig.
8	Det skal være enkelt å finne fram til SiT’s nye tjeneste *nettrådgivning*.
Alle	Mål med tanke på antall navigasjonssteg. Testbrukerne skal kunne velge riktig toppkategori på første forsøk. Mest ideelt skal testbrukerne også finne riktig underkategori på første forsøk, enten ved å bruke høyre margen eller linkene midt på siden. Mer enn en navigasjonsfeil er ikke godt nok.

Tabell 5.1: Mål

Måling av brukskvalitet

I runde 2 ble det lagt mer vekt på hvordan man skal måle brukskvaliteten. Ved å benytte konkrete kvantitative målinger kan man i større grad fokusere på det man mener er viktig og interessant, og i tillegg får man nøyaktige resultat å forholde seg til. Målinger er kvantitative, og ut i fra listen med *mål, oppgaver* og *bekymringer*, ønsker vi verdiene:

- Tid per oppgave
- Antall navigeringsfeil
- Type navigeringsfeil
- Antall gjentatte feil (samme feil blir gjort flere ganger)
- Observasjoner av frustrasjon
- Alternative løsinger
- Hvilke oppgaver som ble hoppet over

For hver brukbarhetstest, ble resultatene ble notert i dette skjemaet:

Brukbarhetstest's målinger									
Deltager #	T=Feil valg av toppmen			A=Andre feil			O=Hoppet over		
Dato:	H=Feil valg i høyremen			L=Alternative løsninger			F=Frustrasjon		
Oppg.	Tid	T	H	A	F	L	O	Kommentar	Notater
#1 Dagens									
#2 Legetime									
#3 Psykolog									
#4 Storkiosk									
#5 Trening									
#6 Skriv ut									
#7 Bolig									
#8 Stress									
#9 Org. Kart									
#10 Ansatte									
#11 Ris og ros									
#12 Sentralbord									

Figur 5.2: Skjema for notering av brukskvalitetsproblemer

5.1.3 Resultater

Notater, videoopptak, måleskjema og kommentarer fra testbrukerne ble notert i et oppsummeringsskjema (se appendiks A og B). Ut fra dette skjemaet ble det bestemt hvilke problemer som ble erklært *brukskvalitetsproblemer* og *mulige brukskvalitetsproblemer*.

Globale og lokale problem

Noen problem er langt mer alvorlige enn andre. Vi har delt problemene opp i lokale og globale. Lokale problemer kan være at man hoppet over et steg i prosedyren som forårsaker at de ikke kan fullføre oppgaven eller at en av undermenyene har ett dårlig navn. Disse feilene er lokale og gjelder som oftest bare i ett skjermbilde. Globale variabler har ett større omfang, som er mer generelle problemer som gjelder større deler eller hele systemet. Globale variabler er ofte mer alvorlige enn lokale.

5.2 Resultater fra CTA

Tidstabellen (figur 4.5) fra CTA-metoden viser totaltid per oppgave. I snitt brukte hver testdeltager til sammen cirka et kvarter på selve oppgaveløsingen (av oppgaver som kan godkjennes), og i tillegg gikk tiden med på å lese gjennom oppgaver, tenke igjennom som egen besvarelse, venting på grunn av tregt nettverk og en liten tilbakerapporteringssesjon etter hver test. Blå farge indikerer at målene ikke ble nådd, mens svart betyr at oppgavene ikke ble fullført. Av totalt 70 gyldige besvarelser, greide nøyaktig halvparten (35 besvarelser) tidsmålet.

CTA	TB #1	TB #2	TB #3	TB #4	TB #5	TB #6
Oppg. 1	00:39	00:21	01:34	00:50	02:55	01:21
Oppg. 2	00:50	00:34	00:46	00:59	00:55	00:28
Oppg. 3	00:07	00:45	00:46	00:23	00:35	00:35
Oppg. 4	00:35	00:29	00:25	00:37	02:35	00:20
Oppg. 5	00:39	00:52	00:46	00:25	01:39	01:54
Oppg. 6	01:21	00:38	00:15	00:23	01:08	00:29
Oppg. 7	02:05	09:33	01:50	01:11	04:19	02:42
Oppg. 8	03:33	01:59	01:52	00:40	01:29	00:58
Oppg. 9	00:40	00:12	01:28	00:59	01:17	02:41
Oppg. 10	01:39	02:25	00:03	02:12	03:00	01:13
Oppg. 11	00:33	00:13	00:25	00:05	00:27	00:20
Oppg. 12	00:20	01:13	02:22	00:43	00:48	01:56

Ikke innen for målrammen

Oppgave ikke fullført

Tabell 5.2: Tidstabell over CTA oppgaveløsing

Målte problemer - CTA

Etter å ha delt opp i lokale og globale problem, er problemene sortert etter nivå av alvorlighetsgrad i stigende rekkefølge;

Problem:	**Skriv ut** (oppgave 6 og 10)
Omfang:	Global
Frekvens:	6 / 6
Utdyping:	Dette globale problemet oppstår når testdeltagerne skal skrive ut. *Skriv ut timeplan*: Alle testdeltagerne kommenterer at de savner en utskriftslink eller en link til en utskriftsvennlig versjons. *Skriv ut ansatt:*
Forslag til forbedring:	Innfør et ikon med teksten ”skriv ut” som er med konsekvent på hele nettsiden til SiT. Når man klikker på ikonet kan ”skriv ut”-vinduet dukke opp foran en utskriftsvennlig versjon. Se skjermbilde 1 i vedlegg 1.

Problem:	**Bolig** (oppg. 7)
Omfang:	Lokalt / Globalt
Frekvens:	4/6
Utdyping:	Testpersonene brukte lang tid på denne oppgaven (Gjennomsnittstid på 3 min 37 sek). Mye kaos. Når man kommer inn på boligtorget (et annet nettsted med annen oppbygging), blir mange av testdeltagerne forvirret. Ventelisten er heller ikke oppdatert. Noen av testdeltagerne nevner at de savner en direktelink fra ventelisten til den aktuelle boligen. Og når man trykker på ”til søknadsskjema” under den aktuelle boligen, kommer man inn på nytt nettsted man må sette seg inn i, for så å navigere seg fram til den rette boligen, velge type leilighet og så sjekke om det er ledig bolig. Enkelte av ”boligdetaljer og pris” linkene var også defekte.
Forslag til forbedring:	• Direktelink fra ventelisten til de aktuelle boligene • På hver av boligene bør det stå hva som er ledig • Når man klikker på søknadsskjema, burde man komme til selve søknadsskjemaet. Se skjermbilde 2 vedlegg 1.

Problem:	**Nettrådgivning** (Oppgave 8)
Omfang:	Lokalt / Globalt
Frekvens:	6 / 6
Utdyping:	Ingen brukte nettrådgivningstjenesten, til tross for at to av testdeltagerne var innom ”Rådvill? Spør oss”. Dette ble begrunnet med at de ikke visste om tjenesten, eller at den ikke var nok synlig. Alle testdeltagerne ville løst denne på en alternativ måte, bortsett fra en testbruker som ville hoppe over oppgaven.
Forslag til forbedring:	Nettrådgivningen burde gjøres mer synlig. Nettrådgivningen burde ikke åpnes i nytt vindu?

Problem 5:	**Resepsjon og åpningstider** (Oppgave 12)
Omfang:	Lokalt
Frekvens:	5/6
Utdyping:	Hovedproblemet var at de leita i undermenyene på ”kontakt oss” og flere testdeltagere bruker lang tid på å leite. I følge RTA metoden, hadde kun 1 av 6 problemer. En person foreslo at ”kontakt oss” burde vært en link på fanen øverst i høyre hjørne, altså sammen med *English, Ris* og *ros* og *Kart. En annen nevnte* at det burde vært en egen link som het ”kontaktinformasjon” under *kontakt oss*. En annen observasjon under testen: Under *bolig* finnes en link som heter *Åpningstider for resepsjonen,* men siden forklarer ikke hvor dette er.
Forslag til Forbedring:	Hva med en egen undermeny under *kontakt oss* som heter *Ekspedisjon og åpningstider,* eller *Sentralbord*? På denne siden kunne man også satt inn en oversikt over de forskjellige ekspedisjonen med kart.

Problem:	**Dagens middag** (Oppgave 1) CTA
Omfang:	Lokalt
Frekvens:	2 av 6
Utdyping:	Det kommer uttydelig fram hvor menyene serveres. Flere testdeltagere leita etter hva som ble servert. *"Det tok litt tid før jeg skjønte at dette var de to rettene som serveres"*
Forslag til Forbedring:	Hva med forklaring under tabellen på hvor dette serveres?

Problem:	**Ansatte**
Omfang:	Lokal
Frekvens:	4/6
Utdyping:	Bortsett fra en testdeltager, var det ingen som så ikonlinken. Etter oppgaveløsningen kommenterte de fleste det er vanskelig å se linken. Enten at den går for mye i ett med layouten slik at den blir skjult, og /eller at linken er for liten.
Forslag til forbedring:	Gjøre linken større, og vurder bruk av andre farger til selve linkene.

Tabell 5.3: Oppsummerte resultater fra CTA runder, samt forslag til forbedreing.

5.3 Resultater fra RTA

Tidstabellen fra RTA-metoden (se figur 4.6) viser totaltid og gjennomsnittstid i tillegg til tid per oppgave. I snitt brukte hver testdeltager til sammen ca ni minutter på selve oppgaveløsingen (av oppgaver som kan godkjennes). Tabellen nedenfor viser også fargekoder som indikerer overtid på oppgaveløsningen (blå) for å se om tidsmålene ble nådd, samt tider som ikke er med i utreging av gjennomsnittstiden pga ugyldig løsning (svart).

RTA	TB # 7	TB # 8	TB # 9	TB # 10	TB # 11	TB # 12
Oppg. 1	00:23	00:42	00:20	00:55	01:47	00:10
Oppg. 2	00:38	00:27	00:39	00:45	00:20	00:00
Oppg. 3	00:15	00:22	00:17	00:50	00:18	00:37
Oppg. 4	00:44	03:45	00:26	00:50	00:13	00:42
Oppg. 5	00:34	00:18	00:37	01:12	00:21	01:35
Oppg. 6	00:25	00:29	00:17	01:05	00:12	00:20
Oppg. 7	00:35	01:39	00:33	01:40	02:48	01:59
Oppg. 8	00:18	00:22	00:34	00:44	00:24	00:52
Oppg. 9	00:19	00:30	00:40	00:15	00:22	00:23
Oppg. 10	00:55	00:45	00:22	00:54	02:35	03:27
Oppg. 11	02:06	00:10	00:24	00:10	00:20	00:13
Oppg. 12	00:31	00:28	00:25	00:26	00:39	00:30

Ikke innenfor målrammen

Oppgave ikke fullført

Tabell 5.4: Tidstabell over RTA oppgaveløsing

Målte problemer - RTA

Hvis vi går igjennom RTA-metoden på samme måte som med CTA, får vi disse brukskvalitetsproblemene i organisert rekkefølge. De største problemene først;

Problem 1	**Leter etter *skriv ut*** (Oppgave 6 og 10)
Omfang:	Global
Frekvens:	6/6
Utdyping:	Testbrukerne er usikre på hvordan de skal skrive ut. De leter etter en utskriftslink eller ikon. Flere av testbrukerne sier de ikke vet hva som vil bli skrevet ut, slik som ramme, timeplan over to sider osv
Forslag til forbedring:	Innfør ett skriv ut ikon med teksten ”skriv ut” som er med konsekvent på hele nettsiden til SiT. Når man klikker på den kan ”skriv ut”-vinduet dukke opp foran en utskriftsvennlig versjon. Se skjermbilde 1 i vedlegg 1.

Problem: #1	**Finne bolig** (oppgave 6)
Omfang:	Global
Frekvens:	6/6
Utdyping:	Testdeltagerne har store problemer med å finne det de letet etter, samt å navigere seg mellom venteliste, aktuell bolig og boligtorget. Fire av seks testdeltagere var innom boligtorget en eller flere ganger for å prøve å få svar på det var ledig bolig. En person nevnte også at det burde vært linker for beskrivelser til hva de forskjellige boligtypene betydde.
Forslag til forbedring:	Direktelink fra ventelisten til de aktuelle boligene På hver av boligene bør det stå hva som er ledig. Når man klikker på søknadsskjema, burde man komme til selve søknadsskjemaet. Se bilde X.

Problem: #1	**Nettrådgivning** (Oppgave 8)
Omfang:	Lokal
Frekvens:	6/6
Utdyping:	Nettrådgivningen er bortgjemt. Testbrukerne vet ikke hva dette er, og fant derfor andre alternative måter å søke hjelp på. Det skaper også forvirring når undermenyene til nettrådgivning dukker opp i ett nytt vindu.
Forslag til forbedring:	Synliggjøring av tjenesten. En av testbrukerne mente at en ”lilla boks” ville gjøre nytten. Man kan også bruke hovedsiden til ”rådgivning” som reklame for tjenesten. ”Spør oss” under nettrådgivning kan med fordel flyttes på toppen av undermenyene.

Problem: #	**Dagens middag** (Oppgave 1)
Omfang:	Lokalt problem
Frekvens:	2 av 6
Utdyping:	To testbrukere skjønte ikke at *”Det tok litt tid før jeg skjønte at dette var de to rettene som serveres”*
Forslag til forbedring:	Hva med forklaring at dette serveres i alle Tellus kafeene under tabellen?

Problem:	**Ansatte** (oppgave 7)
Omfang:	Lokal
Frekvens:	4/6
Utdyping:	Fire av testdeltagerne fant ingen ”skriv ut”-knapp. En av deltagerne brukte veldig lang tid.
Forslag til forbedring:	Gjøre linken større, og vurder bruk av andre farger til selve linkene.

Tabell 5.5: Oppsummerte resultater fra CTA runder, samt forslag til forbedreing.

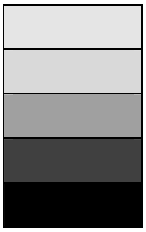
Ut fra brukskvalitetsproblemene beskrevet over er resultatene like, men under gjennomføringen av CTA dukket det opp et ekstra brukskvalitetsproblem. Fem av seks personer hadde problemer med å finne fram til resepsjonens åpningstider, mens ved RTA-metodene ikke hadde betydelige problemer. Dette kan i utgangspunktet være tilfeldig.

Skjemaet (figur 4.7) viser hvor lang tid hver bruker benytter på hver oppgave under metodene. Det kan være noen sekunder avvik i forhold til virkeligheten, fordi enkelte situasjoner er vanskelig å beregne eksakt tid på oppgave. Tabellen viser også gjennomsnittstid testbrukeren brukte på hver oppgave, samt gjennomsnittstid for å løse en bestemt oppgave. Enkelte greide ikke å løse alle oppgavene. Oppgavene som ikke ble godkjent ble ikke regnet med i totalen og gjennomsnittet. En del av oppgavene ble også løst på alternative måter. Her var det vanskelig å avgjøre om løsningene kunne godkjennes i forhold til gjennomsnittstid.

Ut fra tabellene kan man se at testbrukerne i RTA-metoden lot være å fullføre flere oppgaver enn i CTA. Med ”fullførte ikke oppgave” menes at testbrukerne prøvde seg på oppgaven, men hoppet over til neste uten å fullføre. Det er også store tidsforskjeller på oppgaveløsningen av RTA og CTA. RTA-testdeltagerne brukte ca 2/3 av tiden i forhold til CTA-testdeltagerne under oppgaveløsningen.

CTA	TB #1	TB #2	TB #3	TB #4	TB #5	TB #6		Tot.tid	Gj. snitt
Oppg. 1	00:39	00:21	01:34	00:50	02:55	01:21		**07:40**	**01:00**
Oppg. 2	01:10	00:34	00:46	01:33	00:55	00:28		**05:26**	**00:49**
Oppg. 3	00:07	00:45	00:46	00:23	00:35	00:35		**03:11**	**00:32**
Oppg. 4	00:35	00:29	00:25	00:37	02:35	00:20		**05:01**	**00:50**
Oppg. 5	00:39	00:52	00:46	00:25	01:39	01:54		**06:15**	**01:03**
Oppg. 6	01:21	00:38	00:15	00:23	01:08	00:29		**04:14**	**00:42**
Oppg.. 7	02:05	09:33	01:50	01:11	04:19	02:42		**21:40**	**03:37**
Oppg. 8	03:33	01:59	01:52	00:40	01:29	00:58		**07:03**	**01:46**
Oppg. 9	00:40	00:12	01:28	00:59	01:17	02:41		**07:17**	**01:13**
Oppg. 10	01:39	02:25	00:03	02:12	03:00	01:13		**08:20**	**01:40**
Oppg. 11	00:33	00:13	00:25	00:05	00:27	00:20		**02:03**	**00:21**
Oppg. 12	00:20	01:13	02:22	00:43	00:48	01:56		**07:22**	**01:14**
Totaltid	13:21	17:15	12:32	09:02	21:07	14:57		**01:25:32**	**14:42**
Gj. Snitt	**01:07**	**01:34**	**01:03**	**00:50**	**01:46**	**01:15**			

RTA	TB #7	TB #8	TB #9	TB #10	TB #11	TB #12		Totaltid	Gj. snitt
Oppg. 1	00:23	00:42	00:20	00:55	01:47	00:10		**04:17**	**00:43**
Oppg. 2	00:38	00:27	00:39	00:45	00:20	00:00		**02:04**	**00:31**
Oppg. 3	00:15	00:22	00:17	00:50	00:18	00:37		**02:39**	**00:27**
Oppg. 4	00:44	03:45	00:26	00:50	00:13	00:42		**06:40**	**01:07**
Oppg. 5	00:34	00:18	00:37	01:12	00:21	01:35		**04:37**	**00:46**
Oppg. 6	00:25	00:29	00:17	01:05	00:12	00:20		**02:48**	**00:28**
Oppg.. 7	00:35	01:39	00:33	01:40	02:48	01:59		**07:15**	**01:27**
Oppg. 8	00:18	00:22	00:34	00:44	00:24	00:52		**03:14**	**00:32**
Oppg. 9	00:19	00:30	00:40	00:15	00:22	00:23		**02:29**	**00:25**
Oppg. 10	00:55	00:45	00:22	00:54	02:35	03:27		**08:58**	**01:30**
Oppg. 11	02:06	00:10	00:24	00:10	00:20	00:13		**03:23**	**00:34**
Oppg. 12	00:31	00:28	00:25	00:26	00:39	00:30		**02:59**	**00:30**
Totaltid	**07:43**	**09:57**	**05:34**	**09:46**	**10:19**	**11:25**		**54:44**	**09:07**
Gj. Snitt	**00:39**	**00:50**	**00:28**	**00:49**	**00:52**	**00:57**			



Gal løsning (ikke godkjent)
Alternativ løsning (godkjent)
Hoppet over uten å prøve (ikke godkjent)
Skjønte ikke spørsmålet (ikke godkjent)
Fullfører ikke oppgave (godkjent/ ikke godkjent. Etter skjønn)

Tabell 5.6: Tidstabell over RTA og CTA med totaltid og gjennomsnittstid

Kapittel 6 Diskusjon

6.1 Sammenligning av testresultatene

Ved å oppsummere resultatene fra forrige kapittel kom vi fram til at;

- Resultatene er ganske like
- CTA avdekket ett brukskvalitetsproblem mer enn RTA
- RTA-testdeltagerne brukte ca 2/3 av tiden i forhold til CTA-testdeltagerne under oppgaveløsningen.
- RTA testdeltagerne hoppet over langt flere oppgaver enn CTA-deltagerne

6.1.1 CTA avdekket flere brukskvalitetsproblem

RTA-metoden identifiserte fem brukskvalitetsproblem, mens CTA identifiserte seks. CTA identifiserte alle problemene identifisert av RTA, og i tillegg identifiserte den et ekstra brukskvalitetsproblem. Betegnelsen *identifisere* kan være litt misvisende, fordi det er ikke gitt at de problemene som dukker opp er brukskvalitetsproblemer. Det kan være andre årsaker til at testbrukerne hadde problemer med å løse oppgavene, slik som *tidspress, nervøsitet* eller for eksempel *hvilken metode* man bruker. Fem av seks testdeltagere under CTA-metoden hadde problemer med å finne fram til resepsjonens åpningstider, mens ingen av RTA-deltagerne hadde problemer med denne oppgaven.

Tegnet # etterfulgt av et nummer indikerer nummeret på testdeltageren på samme måte som i forstudiet. Ved å se på den loggen fra testdeltagernes oppgaveløsning, kommer det fram at:

De som brukte **CTA-metoden**:

1 ville benyttet alternativ løsning ved å benytte et kontaktskjema for å spørre om åpningstidene.

2 Valgte feil toppmeny først, men fant fram litt senere.

3 Var innom feil toppmeny tre ganger. Valgte feil i høyremenyen to ganger og viste tegn til frustrasjon. Ville til slutt ringt en person i boligsektoren.

4 Valgte først feil i høyremenyen under ”Hva er SiT”, og endte til slutt med den alternative løsningen å ringe noen innenfor ”Enheter og bedrifter”.

5 Oppdaget at åpningstidene sto i bunnteksten. Denne løsningen ble godkjent fordi testbrukeren fikk svar på det han ønsket ved å kun bruke nettstedet.

6 Fant fram etter andre toppmeny - og andre høyremeny feilsteg. Testbrukeren brukte to minutter.

Altså er det ingen tvil i følge Dumas og Redish’s (1999) retningslinjer at dette er et brukskvalitetsproblem. Resultater fra RTA-loggen viser noe helt annet; ”Bortsett fra en person som løste denne oppgaven på en alternativ måte, som var like grei, må det konstateres at det var enkelt å finne fram til åpningstider og eksedisjon” (Se vedlegg 1).

Den samme oppgaven ble også benyttet under forstudiet (CTA-metoden). Videoopptaket kunne konstatere at tre av fire hadde problemer med å finne fram til kontaktinformasjonen. Dette er med på å forsterke teorien om at CTA-metoden avdekket et falskt brukskvalitetsproblem. I tillegg ble i ettertid fire personer fra samme målgruppe som de andre i runde 2 spurt om de kunne finne åpningstider til SiTs resepsjon. Det skal tilføyes at deltagerne hadde ulik erfaring med data og internett. Her ble det ikke brukt noen form for overvåkning, men tiden ble tatt for å se om de greide målene. Navigasjonen ble heller ikke registrert, men deltagerne fortalte hvordan de kom fram til svaret. En person brukte søkefunksjonen, mens de tre andre fant fram ved å gå via ”Hva er SiT” | ”Kontakt oss”. Tre testdeltagere påsto de greide dette uten omveier, mens en sa at han var innom ”Kontakt enheter og bedrifter” før han skjønte det lå på ”Kontakt oss”. Alle greide dette på under 40 sekunder. Det ble totalt registrert at åtte av ti CTA-deltagere hadde problemer med å finne fram til ”Ekspedisjon og åpningstider”, mens ingen av ti testdeltagere fra RTA-gruppen hadde problemer med samme oppgave.

6.1.2 Tidsforskjeller på oppgaveløsning

Det var også tidsforskjeller på gjennomføringen av CTA- og RTA-metoden. Tiden det tok å gjennomføre alle oppgavene under testene (Se kapittelet 5), er med CTA-metoden 1 time og 26 minutter, mens RTA metoden tok 55 minutter (det er foretatt en avrunding mot nærmeste minutt). Gjennomsnittstiden per oppgavegjennomgang ligger på 15 minutter med CTA metoden, mens ni minutter med RTA metoden. RTA metoden brukte i underkant av 2/3 av tiden til CTA. Det understrekes at dette er en sammenligning av tiden det tar å gjennomføre oppgavene.

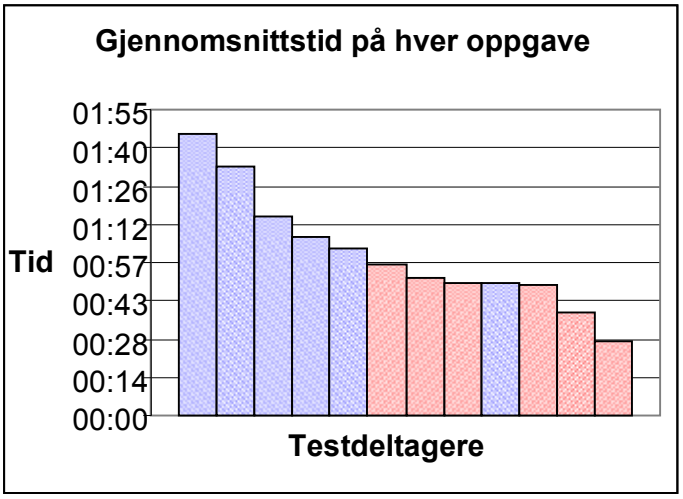
Det er ikke funnet andre studier som viser til dette resultatet, men det er gjort lite forskning på området (Van den Haak, de Jong 2003). Spørsmålet er *hvorfor* det er så stor tidsforskjell. Ser vi på gjennomsnittstiden per oppgave, ser resultatet slik ut:

CTA:	TB #1	TB #2	TB #3	TB #4	TB #5	TB #6
Gj. Snitt	**01:07**	**01:34**	**01:03**	**00:50**	**01:46**	**01:15**

Tabell 6.1: Utdrag fra tidsskjema fra CTA runden

RTA:	TB #7	TB #8	TB #9	TB #10	TB #11	TB #12
Gj. Snitt	**00:39**	**00:50**	**00:28**	**00:49**	**00:52**	**00:57**

Tabell 6.2: Utrag fra tidsskjema fra RTA runden



Figur 6.1: Graf over gjennomsnittstid pr oppgave for CTA og RTA runden

Fem av seks deltagere under CTA-metoden brukte over mellom ett og to minutter per oppgave, mens alle under RTA-metoden brukte mellom 28 og 57 sekunder. En start er å se på hovedforskjellen mellom metodene CTA og RTA. Som det ligger i navnet (sanntids og retrospektiv TA), er det en forskjell på når man snakker. Ved bruk av RTA snakker man først etter oppgaveløsningen. Dette er hovedforskjellen. I tillegg er det ingen krav i RTA at man skal overvåke og ta opp seansen, men vi skiller mellom RTA og stimuli RTA. I sistenevnte metode skal man kunne bruke en form for stimuli i oppsummeringssesjonen. Fordelen med stimuli er at man kan få hjelp til å huske. Dette har rot i psykologiens verden, hvor vi skiller mellom det *å huske* (eng. recall) versus det *å gjenkjenne* (eng. recognition). Vi kan gjenkjenne materiale langt enklere enn vi kan huske det fra minnet (Guidon 1988). I dette studiet ble stimuli benyttet under RTA-testen, på bakgrunn av Hoc og Leplats anbefaling (1983 via van den Haak & de Jong 2003). Det ble da ikke gjort videoopptak av testbrukerne, men fra skjerm.

Vi sitter da igjen med to mulige årsaker for tidsforskjellen;

- *Tilfeldigheter*. Man kan ikke utelukke tilfeldigheter. Man kan redusere sannsynligheten for tilfeldigheter ved å foreta flere brukbarhetstester, eller man kan finne bevis på sammenheng.

- *CTA metodens innvirkning på testsituasjonen.* Dette kan være en dobbel kognitiv belastning. Kan denne belastningen med å tenke og å snakke samtidig ha innvirkning? Eller har *overvåkningseffekten* noe å si på resultatet? Vil kameraene eller tilretteleggerens tilstedeværelse ha innvirknings på utførelsen? (Preece, Rogers & Sharp 2002, Nielsen, Clemmensen & Yssing 2002)

6.2 Intervju

Intervjuet er en viktig del av denne masteroppgaven, fordi man prøver å komme inn på de som testes, samt å kartlegge brukbarhetstestsituasjonen. Intervjuet ble gjennomført som et semistrukturert intervju, der det på forhånd ble definert hvilke områder man skulle gjennom i løpet av intervjuet. Semistrukturerte intervju gjør at man holder seg til konkrete områder, samtidlig som intervjuobjektene har spillerom til å gå i dybden på sine besvarelser (se kapittel

3.3.2) og intervjuguide i appendiks C. Intervjuene ble transkribert og senere analysert ved hjelp av softwareprogrammet NUD*IST som er utgitt av QSR (QSR 2007).

6.3 Opplevelse av metodene

6.3.1 Før ankomst

Før brukbarhetstestene ble deltagerne spurt hva de følte før testen. Poenget var å undersøke;

- Hvorvidt det var opplagte holdningsforskjeller på deltagergruppene
- Følelsesrelaterte endringer i løpet av testen.
- Redusere tilfeldige forskningsresultat

Under intervjuet ble deltagerne spurt hva de følte og hvilke holdninger de hadde før de kom til laboratoriet. Holdningene var ganske like i begge gruppene. Under kodingen av de transkriberte resultatene ble holdningene delt i to grupper; ”avslappet/likegyldig” og ”spent/nervøs”. I RTA-gruppen var fordelingen halvt om halvt. I CTA-gruppen var fire spent og to avslappet. Ved å lese kommentarene (se Appendiks A og B) fra transkriberingen, er det ikke store forskjeller.

6.3.2 Introduksjon til metodene

Det var heller ingen stor forskjell på introduksjonen som ble holdt både i CTA- og RTA-metoden. Forskjellen ligger i at deltagerne fikk ulik informasjon om;

- RTA: Det vil ikke bli brukt kamera under oppgaveløsningen mens skjermbevegelsene blir tatt opp. Et eget program som er installert på datamaskinen tar opp skjerm og musebevegelser.
- RTA: Under oppsummeringssesjonen vil en MP3-spiller med opptaksfunksjon bli brukt for å ta opp hva som blir sagt.
- RTA: deltagerne skal sitte alene uten tilrettelegger ved siden av.
- RTA: TA-teknikken blir praktisert etter oppgaveløsningen
- CTA: Det brukes kamera under oppgaveløsningen
- CTA: Tilrettelegger skal sitte ved siden av
- CTA: TA teknikken ble praktisert før oppgaveløsningen
- CTA: Kan ikke stille spørsmål relatert til oppgavene.

Stort sett oppfattet testdeltagerne fra begge metodene introduksjonen som grei, og det var stort sett positiv tilbakemelding. Testdeltagerne opplevde de fikk god oversikt over hva de skulle gjøre, og de fleste fikk oppklart det de lurte på. En testbruker kommenterte;

> *"Nei, det gikk greit fram til du begynte å fortelle om videogreiene da. Fram til da var jeg ganske rolig… da du fortalte om kameraene ble det litt ubehagelig."*
> (# 2)

6.3.3 Helhetsinntrykk

Under intervjudelen ble testdeltagerne stilt spørsmål rundt deres opplevelse av den metoden de var med på. Med metodens opplevelse menes gjennomføring av en test fra start til slutt. Ingen av testpersonene som var med på RTA-gjennomgangen kom med negative kommentarer.

> *"Jeg synes det var helt topp jeg*" (# 10)

De fleste skrev at de syntes det var greit eller at de ikke hadde noen problemer med å bli testet på denne måten. Et av argumentene for at de opplevde situasjonen som positiv var at de kunne sitte i ro og mak og at det var ingen som fulgte med på hva de gjorde. En kommenterte også at han satte stor pris på at han fikk servert kaffe (det ble tilbudt kaffe eller te under samtlige brukbarhetstester i runde 1 og 2). På spørsmål om de følte stress, svarte alle RTA-deltagerne *nei*. Alle testdeltagerne svarte også at de var fokuserte på oppgaveløsningen under RTA metoden;

> *"Jeg greide å konsentrere meg om det jeg skulle gjøre. Det var liksom bare meg, dataen og oppgavene"*
> (#11)

Det eksisterer lite forskningsmateriale om hvordan CTA-metoden oppleves av brukeren (Nielsen, Clemmensen & Yssing 2002), og dette er noe vi vil til bunns i fordi testbrukernes opplevelse *kan* ha innvirkning på resultatene. Da testbrukerne ble spurt om hvordan de opplevde å bli testet på denne måten, var det her, i motsetning responsen på RTA-metoden, mange negative tilbakemeldinger. Dette var alt fra at de følte metoden som ubehagelig, at de

følte seg stresset, at det var vanskelig å verbalisere sine tanker, prestasjonsangst, nervøsitet, anspenthet, lite spontanitet og at det hele opplevdes som en kunstig situasjon;

"Jeg opplevde dette som anstrengt. Og da spesielt da du så på hva jeg gjorde… og at jeg ble filma."
(# 2)

"Tja.. synes ikke det fungerte helt optimalt. Vanskelig å sette ord på sine tanker"
(# 3)

"Ja, det er greit egentlig. Men fordi om du sier jeg ikke skal ha noen presentasjonsangst, så føler du det. Å gud, nå synes han jeg er dum"
(# 4)

"*Jeg synes… hva skal jeg si. Jeg ble kanskje litt anspent. Jeg følte at alt jeg sa og gjorde blei registrert så det føltes som om det kanskje ikke ble så spontant"*
(# 5)

"Litt kunstig situasjon egentlig. Det er sånn… å sitte der liksom å..hei, hva du tenker. Det er ikke sånn man gjør til daglig. Man blir liksom litt stresset og man finner ikke ting enkelt. "Nå må jeg finne det, nå må jeg finne det"
(# 6)

Men opplevelsen av CTA-metoden skal ikke svartmales helt. Det kom også positive tilbakemeldinger som gikk på opplevelse av metoden;

"Jeg synes det var veldig greit. Det var ikke noe ubehagelig eller noe sånt. Det var greit å fortelle hva jeg gjorde. Jeg merka ikke at kameraene var på heller."
(# 1)

"Jeg synes det var greit. Det var ikke så skummelt som jeg hadde trodd. Jeg syntes ikke det var skummelt til å begynne med, men når jeg så kameraene og greier, da var

jeg redd for å få helt sånn jernteppe. Men det gikk greit altså. Det var litt interessant faktisk. Ser litt åssen jeg tenker selv da. Bevisst åssen man tenker kanskje"
(# 5)

Samtlige testdeltagere svarte at de en eller flere ganger var stresset under testen. Alt i alt virker det som testdeltagerne hadde mer eller mindre dårlige opplevelser med denne metoden. Noen mener kanskje at det viktigste er å avdekke brukskvalitetsproblemer, men hva om opplevelsen av metoden har innvirkning på resultatene? Vi har allerede sett at det er forskjeller på tiden deltagerne bruker på å fullføre oppgavene, men testresultatene er omtrent de samme. Når vi sier omtrent det samme blir det i dette tilfellet at CTA i forhold til RTA påpekte ett ekstra falskt brukskvalitetsproblem. I dette tilfellet er det enkelt å gjøre åpningstidene på resepsjonen mer synlige, men hva hvis det hadde vært en langt mer alvorlig og kostbar feil?

6.3.4 Tid

Tid var et tema som dukket opp under de første intervjuene, uten at det i utgangspunktet ble spurt om det. Samtlige testdeltagerne under RTA-gjennomføring beskriver at de i flere anledninger følte de hadde god tid, at de ikke følte noen form for tidspress, og at de tok den tiden de trengte. Den eneste kommentaren som pekte litt i den andre retningen var svaret på et oppfølgingsspørsmål om skjermopptaket hadde noe å si;

"Nei, men mer at du kanskje ventet, men det er jo jobben din… men det var jo også det at jeg kunne sitte der og ta den tiden jeg trengte. Det var bra det"
(# 7)

Under RTA-testene satt tilrettelegger i samme rom som testdeltagere, men personene ble skilt med en skillevegg. Testdeltagerne hadde ingen restriksjoner på når de kunne stille spørsmål, men ble bedt om å prøve før de spurte.

Testdeltagerne som var med på CTA-gjennomgangen hadde en annen opplevelse av tid. De følte ikke at de hadde god tid.

"Nei. Jeg ble som sagt litt stressa og ville prestere bra. Så gjør ikke saken bedre at du sitter ved siden av"
(#2)

"Det er som jeg sa til deg at du føler deg overvåket"
(# 4)

"Det blir jo litt sånn i en test. Kan liksom ikke svitsje innom alt annet som jeg ofte gjør. Du ser jo på hva jeg presterer så føler på tidspresset"
(# 5)

"Nei, det gjorde jeg ikke. Sånn som jeg pleier når jeg er fokusert… bare få det overstått. Ble stressa av å bli overvåka sånn"
(# 6)

Testdeltagere følte tidspress, prestasjonsangst, overvåkningseffekten og/eller at de jobbet annerledes enn de ellers ville gjort. Det ble lagt vekt på å få testdeltagerne til å føle seg komfortabel ved blant annet å si at de måtte ta den tiden de trenger, at de kan hoppe over oppgaver, avslutte når de ville og at vi testet systemet og ikke dem.

6.3.5 Tilretteleggeren

Under intervjuet, ble det stilt spørsmål til testdeltagerne om hvordan de oppfattet tilretteleggeren. Poenget med dette var å se om deltagerne på hver av gruppene oppfatter tilretteleggeren på samme måte. Hvis CTA-gruppene oppfattet tilretteleggeren som brautete eller uforutsigbar, kan dette virke skremmende på deltagerne. Dette kunne for eksempel resultert i prestasjonsangst og redusert ytelsen under oppgaveløsningen. Deltagerne ble gjort oppmerksomme på at tilretteleggeren spilte en rolle, og at dette på ingen måte ville bli tatt personlig. Samtlige RTA-deltagere hadde bare positive tilbakemeldinger om tilretteleggerens rolle.

Du var veldig behagelig. Det var ikke noe stress. Jeg kunne gjøre det i mitt tempo.
(# 8)

Deltagerne i CTA grupper følte tidspress av forskjellig grad, og to deltagere nevnte at tilretteleggeren hadde en form for innvirkning. Men på spørsmål om hvordan de opplevde tilretteleggerens rolle, svarte de med adjektiver som *avbalansert, rolig, behjelpelig* og *beroligende.* Allikevel ville samtlige av testdeltagerne aller helst, hvis de kunne velge, foretrukket å sitte alene under oppgaveløsingen.

> *"Du var jo hyggelig og alt sånn der, men det hadde som sagt vært bedre å ikke ha deg der. Men du var hyggelig"*
> (# 4)

Dette gikk også igjen under RTA-intervjuet, hvor de samme adjektivene over ble brukt om tilretteleggeren. De fleste av deltagerne ville foretrukket å sitte alene framfor å ha en tilrettelegger ved sin side. For å få klarhet i hvorfor deltagerne helst vil sitte alene, svarte deltagerne:

> *"Det at du ser på hva jeg gjør. En føler jo seg mer overvåket og at det er meg du tester"* (# 2)

> *"Det at du satt ved siden av meg og så hva jeg gjorde, gjorde at jeg følte meg …. ble kanskje litt satt ut. Jeg følte meg til tider dum"*
> (# 4)

> *"Jeg hadde forventa at jeg skulle sitte alene. At du kanskje skulle sitte litt unna meg. Jeg følte kanskje at du ble litt sånn nært da. Du satt lit oppi meg. Når du sitter på skolebenken og du skal løse oppgaver, så står læreren og henger over deg, det kan være litt… det kunne kanskje vært litt større avstand. Det hadde vært mer behagelig hvis du hadde sittet litt lengre unna"*
> (# 5)

> *Jeg følte meg litt overvåket. Det blir litt ekstra press av det også.*
> (# 6)

Testbruker #1 var den første som ble intervjuet, men under dette intervjuet ble det ikke stilt spørsmål om hvordan hun opplevde tilretteleggerens nærvær. Testdeltager #3 ga uttrykk for at han ikke hadde problemer med at tilretteleggeren satt ved siden av.

6.4 Testsituasjonens innvirkning

Testdeltagerne ble spurt om de ville utført oppgavene på en annen måte i "virkelige" omgivelser enn de gjorde under testen. Under en brukbarhetstest ønskes det at testbrukene skal løse oppgavene på en mest mulig realistisk måte, slik at de som evaluerer testen får reelle data å jobbe ut i fra. Ingen av RTA-deltagerne trodde testsituasjonen hadde innvirkning på resultatene.

> *"Jeg tenkte at hvis jeg skulle ha gjort en lignende oppgave, så ville jeg gjort noe av det samme… Jeg tror ikke det har noe å si at jeg sitter på en lab og at skjermen blir overvåket. Det tror jeg ikke"* (# 10)

Resultatene fra CTA-intervjurunden viser at fire testdeltagerne trodde testsituasjonen hadde en innvirkning på hvordan de arbeidet. Deltagerne vill gjort det annerledes om de hadde hatt bedre tid eller når de ikke blir overvåket.

> *"Du blir ikke nødvendigvis det samme som å sitte hjemme å finne informasjon da. Det blir en forskjell… det skal godt gjøres å unngå. Det blir ikke helt nøytralt rett og slett"* (# 3)

6.4.1 Grad av overvåkning

Alle testdeltagerne ble bedt om å svare på i hvilken grad de følte seg overvåket. Fem av seks testdeltagere i RTA gruppen følte svært liten eller ingen grad av overvåkning enda det ble brukt opptak av skjerm og båndopptager. Tre av deltagerne kommenterte at de glemte båndopptakeren omtrent med det samme, og at de ikke følte seg overvåket.

> *"Nei, det gjorde ingenting. Jeg tenkte ikke over den liksom. Du setter den på, så glemmer man at den er der"*
> (# 9)

Bare en av testpersonene følte seg litt overvåket. I introduksjonen fikk hun beskjed om at de ville bli benyttet *screen recording* under testen. Problemet var at hun ikke visste hva screen recording var, og at hun trodde hun ble filmet. På spørsmål om hun følte at kameraene gjorde henne nervøs, svarte hun;

> *"Litt. Jeg tok jo en test, og hvis jeg ble filma så hadde jo alle andre blitt filma, så det gjorde jeg ikke så mye. Jeg ble ikke noe spesielt nervøs. "*
> (# 8)

Etter dette ble *opptak av skjerm* benyttet i introduksjonen i stedet for *screen recording*.
Fire av testdeltagerne hadde den oppfatning at de ville følt seg mer overvåket med bruk av kamera, og at de foretrakk å bli testet uten kamera. Tre av testdeltagerne ga utrykk for at det hadde kanskje vært verre å bli filmet, men at det ikke var noe stort problem

Resultatene fra CTA-metoden viser seg fra en litt annen side enn resultatene fra RTA. Her benyttes kamera til å filme testdeltagerne, med tilretteleggeren ved siden av. Vi har allerede sett (se kapittel) at tilretteleggeren kan ha innvirkning ved at deltagerne føler seg overvåket eller at tilretteleggeren bryter intimgrensen. Tre av testdeltagerne uttrykket klart at de følte seg overvåket, mens to svarte konsekvent "nei" til at de følte noen grad av overvåking;

> *"Jeg følte meg ganske overvåket. Det at du sitter ved siden av meg og så hva jeg utførte, gjorde at jeg ble litt satt ut"* (# 4)

Den siste personen i CTA-gruppen hevdet han ikke følte seg overvåket, men uttalte likevel;

> *"Jeg følte at alt jeg sa og gjorde ble registrert, så det føltes som om det kanskje ikke ble så spontant"*(# 5)

Testbruker #1 og #3 følte seg ikke overvåket under oppgaveløsningen, og på tidsresultatene var disse blant de tre raskeste. Deltager #4 hadde en bedre gjennomsnittstid, til tross for at hun ikke satte pris på overvåkninga. De andre uttrykte at de ikke var særlig begeistret for å bli filmet.

Som man kan trekke ut fra observasjonene over, føler deltagerne i CTA-gruppen en større grad av overvåking enn deltagerne i RTA-gruppen. Det er ikke funnet noen detaljerte framgangsmåter hvordan man skal overvåke i RTA-metoden, men det finnes grove skisser på hvordan man skal gå fram (Se kapittel 2.4.2). Vi kan si at vi i dette studiet har gjennomført en *variant* av RTA, der det ble benyttet opptak av skjerm og lydopptak av debrief-sesjonen. En brukbarhetstest med CTA er detaljert beskrevet av både Tognazzini (1991), Nielsen (1993) og Dumash og Redish (1999), mens RTA-metoden mangler en slik detaljert beskrivelse. Man kan bruke både loggfiler og videoopptak av personene (Nielsen 1993) fra oppgaveløsningen. Opplevelsen av metodene kunne derfor vært annerledes med annen type overvåkning, og det er derfor viktig å understreke at vi her har kombinert RTA med opptak av skjerm og lysopptak.

Hvis vi ser bort fra personen som trodde hun ble filmet, ble det ikke registret noen som følte seg overvåket under RTA-gjennomføringen. Skjermopptak så ikke ut til å plage noen, og båndopptakeren ble glemt rett etter det ble spurt om tillatelse. Med tilretteleggeren ved siden av og tre kameraer som filmer deltagerne fra forskjellige sider, kan det likevel se ut til å ha en effekt på deltageren. Begge overvåkningsfaktorene ble ofte nevnt under intervjuet.

6.4.2 TA teknikkens innvirkning

En annen mulig innvirkning på resultatet kan være TA-teknikken. Det er imidlertid gjort lite forskning på brukerens refleksjoner ved bruken av denne teknikken (Nielsen, Clemmensen & Yssing 2002). Etter å ha selv utført lab-tester (stort sett på studenter) i regi av skole, har enkelte medstudenter påpekt at de tenker raskere enn de snakker og at det er umulig å gjengi akkurat hva man tenker. I tillegg har medstudenter oppfattet TA teknikken som et forstyrrende element under oppgaveløsningen.

Flere av RTA-deltagerne kommenterte at det var vanskelig å verbalisere sine tanker, spesielt i starten, mens andre RTA-deltagere mente det motsatte. Noen syntes det var uvant eller vanskelig, mens en kommenterte at det føltes helt naturlig. CTA-deltagerne beskrev teknikken med adjektiver som *vanskelig, uvant* eller *belastende*. Samtlige testdeltagere syntes denne teknikken var vanskelig å håndtere på et eller annen nivå.

Å holde det gående

Det viser seg at CTA-deltagerne synes det er vanskelig å holde det gående når de skal fortelle hva de tenker, noe som resulterte i redusert verbalisering. Problemet er at de glemte å snakke, tiltross for påminnelse fra tilrettelegger. Testdeltagerne mente at dette kunne skje når tankevirksomheten økte og de konsentrerte seg om oppgavene, noe som også kan bekreftes av videoopptaket av sesjonen. Det kan være at de leter etter noe, står fast, resonerer eller når de står foran valg. Ut fra hva de sier i intervjuet og hva vi ser på videoen, ser det ut til at konsentrasjon og verbalisering av tankene har en sammenheng.

> *"Ja, for da måtte jeg konsentrere mer. Da må jeg fokusere. Her har jeg ett problem som jeg må finne løsningen på. For å finne den raskes mulig, kan jeg ikke holde på med noe annet"*(# 3)

På samme måte som i forstudiet, ble tiden med aktiv snakking målt. Når testdeltagerne hadde problemer med å sette ord på sine tanker, ble de ofte stille. Andre ganger kunne det resultere i mye nøling, som førte til at det var helt umulig å forstå hva testdeltagerne sa.

> *"Av og til hadde jeg problemer med å sette ord på hva jeg tenkte. Som førte til at det ble litt nølende språk"*
> (# 5)

Tabellen for CTA ser slik ut:

Testbruker	Prater aktivt	Mumling eller stillhet
#1	19 %	81 %
#2	20 %	80 %
#3	20 %	80 %
#4	15 %	85 %
#5	17 %	83 %
#6	22 %	78 %

Tabell 6.3: Aktiv prat i forhold til stillhet - CTA

Til tross for at det ble lagt mer vekt på TA-trening i denne sesjonen, virker det ikke som om dette hadde betydning på hvor aktivt deltagerne snakket. I snitt var deltagerne stille 81 % av tiden under oppgaveløsingen. RTA-deltagerne hadde ikke like store problemer med å holde det gående, for her var det mer prat. I snitt var testdeltagerne stille 45 % av den hele tidsperioden. Det skal legges til at det var stor sprik mellom deltagerne, noe som kan skyldes personlighetsegenskaper eller holdning til testen.

Testbruker	Prater aktivt	Mumling eller stillhet
#7	82 %	18 %
#8	21 %	79 %
#9	57 %	43 %
#10	55 %	53 %
#11	93 %	7 %
#12	30 %	70 %

Tabell 6.4: Aktiv prat i forhold til stillhet – RTA

Tenker raskere

I likhet med egne mottatte tilbakemelding fra testbrukere, skriver Janni Nielsen (2002) følgende; ”Students complain that they think faster that they can speak” (s. 102). Et av de kjente problemene med TA-teknikken i CTA-metoden er at testdeltagerne ofte tenker raskere enn de greier å verbalisere. Denne undersøkelsen viser ingen unntak. Som vi kunne se ut fra forstudiet, svarte deltageren at de *av og til* (en svarte *ofte*) tenkte raskere enn de greide å verbalisere. Under intervjuene i runde 2 kunne fem av seks CTA-deltagere bekrefte de hadde det samme problemet som de i forstudiet, de følte at de tenkte for raskt til at de greide å si det de tenkte.

> *”Du vann ikke å registrere det man tenkte over… Tenkinga gikk så fort at du ikke vant å si det nesten”* (# 6)

RTA-metoden prøver å omgå disse problemene ved at testdeltagerne kunne stoppe opptaket om de har mer på hjertet, noe som ble positivt mottatt av deltagerne.

"Da jeg tenkte raskere enn jeg greide å snakke, kunne jeg bare trykke på pause"
(# 11)

En av testdeltagerne studerte MMI på mastergradsnivå ved NTNU. Han hadde selv utført en rekke brukbarhetstester og var svært begeistret for den retrospektive varianten av TA;

"... jeg kunne stoppe der og forklare hva jeg hadde gjort, og jeg tror det gir mye bedre tilbakemelding"
(# 7)

Ut fra dette ser vi at stoppknappen løser problemet med at testdeltagerne tanker raskere enn de greier å verbalisere. Det ble lagt vekt på å gi klare instrukser på hvordan deltagerne skulle verbalisere sine tanker, og de fikk de tid til å øve seg på TA-teknikken.

Tenker annerledes

Samtlige av RTA-testdeltagerne sa de tenkte på en eller annen måte annerledes enn de greier å uttrykke seg verbalt. Dette går ut på at de ikke greier å gjenfortelle nøyaktig hva de tenkte under testen, noe de begrunner med;

- Det var vanskelig å gjenfortelle presist hva man har tenkt
- De er ikke vant til å formulere seg helt likt etter tankene
- De er mer opptatt av å si hva de har gjort fremfor hva de hadde tenkt
- Ikke alt de tenker er like relevant (vi kan renske ut mye av det vi tenker)
- De tenker ikke over alt (mangler begrunnelse på hvorfor vi gjør som vi gjør)

"Du trenger kanskje ikke si "åh, hvor er den hen!". Du sier i stedet "Her tenkte jeg at det var der". Det blir litt forskjell"
(# 8)

Kommentarene fra CTA runden er ikke så ulik;

- Det er ikke så naturlig
- Vanskelig å få fram det man tenker
- De tenker og snakker annerledes

- De *hører* ikke sine egne tanker, og det blir derfor rart å verbalisere tankene
- Lettere å kommentere avgjørelser enn tanker
- Tenker ikke over hva de gjør, de *bare* gjør det uten begrunnelse

I utgangspunktet synes nok de fleste det er uvant å tenke samtidig som de må sette ord på hver tanke, til tross for at de fleste av oss har møtt personer som har kort vei fra tanke til munn. TA-teknikken kan også oppfattes som kunstig for enkelte av oss, siden vi i enkelte situasjoner utføres handlinger uten å tenke over *hvorfor*. Vi bare *gjør det*.

> *"Det er liksom ikke noe å tenke på. Det er bare sånn det er. Skal du begynne å liksom forklare hvorfor du gjør det, blir det liksom at du må forklare hvorfor du tar på deg ei jakke. Den ene arma først og så den andre"*
> (# 3)

Hukommelse og RTA

Testdeltagerne ble spurt om hvor mye de trodde de husket av det de hadde tenkt under oppgaveløsningen;

# 7	Ca 90 %
# 8	Ca 50 %
# 9	Ca 80 %
# 10	Ca 80–90 %
# 11	Ca 70–80 %
# 12	Ca 90 %

Tabell 6.5: Hukommelse fra RTA runden

Dette gir ett snitt på 78 %, men dette må selvfølgelig tas med en klype salt. Dette beviser ingenting, men det er interessant å se hva deltagerne selv tror. Det hadde også vært interessant å studere øyebevegelse her for å sammenligne resultatene, hvor kan man studere testdeltagernes øyebevegelser og i hvilken rekkefølge de så. Dette kunnet da vært sammenlignet med hva deltagerne sa under TA-sesjonen.

Verbalisering av tanker

I etterpåklokskapens lys er det lett å se at en burde vært mer konkret da det ble stilt spørsmål om deltagernes evne til å verbalisere sine tanker. Når testdeltagerne ble spurt om de greide å verbalisere det de tenkte, ser man det var rom for tolkning av spørsmålet. Hva menes egentlig med *det de tenkte?* Noen tolket det som gjenfortelling av *hva de gjorde* og *hvorfor de tok de valgene de gjorde*, mens andre tolket det som *all salgs tanker som svirret i hodet* på det gitte tidspunktet. På spørsmål om de greide å verbalisere sine tanker, sa en av testdeltagerne;

> *"Jeg så i hvert fall selv hva jeg hadde tenkt da jeg så opptaket" (# 10)*

Mens en annen sa;

> *"Nei, absolutt ikke. Det tror jeg blir vanskelig. Men jeg tror jeg husket de vesentlige tingene som var problemet*" (# 11)

Med utgangspunkt i det som står over skulle en tro de tippet svært ulikt i hvor mye de hukset, men de tippet omtrent samme. Testdeltager # 12 kommenterte også at han ikke husket alt han tenkte. Han var overrasket over at han brukte så lang tid på hvert skjermbilde, og syntes at det var vanskelig å verbalisere sine tanker når det gikk så lang tid og han ikke husket helt hva han hadde tenkt. Allikevel trodde deltageren at han husket mye fra framgangsmåten. Testdeltager # 8 hadde minst tro på at hun husket så mye;

> *"Når du skal si hva du tenkte, greier du ikke si akkurat hva du tenkte, fordi du ikke kommer på det"*

Tenkte noe, sa noe annet

Fem av seks CTA-testdeltagerne poengterte at de til tider sa noe annet enn det de egentlig tenkte. En av testpersonene ville gjerne stille seg mer positiv til nettstedet enn hun egentlig var. Hun lot være å kommentere de negative opplevelsene siden hun trodde dette var tilretteleggerens verk, mens hun forsterket det positive. En annen testdeltager sa at han la vekt på å verbalisere noen tanker, mens andre lot han være å fortelle om. Flere av testdeltagerne følte også press for å si noe;

"Kanskje mer at jeg følte jeg måtte si noe hele tiden. Da sa jeg hva jeg hadde tenkt men det blir kanskje mer ord enn det jeg tenkte"
(# 2)

"Det ble litt påtatt. Som jeg sa tenker jeg ikke hele tiden. Det blir litt sånn at en sier mer hva en driver med i forhold til det som foregår i hodet"
(# 3)

"Du føler at du er nødt til å si noe hele tiden. Nesten sånn at man sitter og finner på ting, bare for å ha noe å si. Jeg synes jeg egentlig at jeg snakket mer om det jeg gjorde, enn det jeg tenkte."
(# 6)

6.5 Retrospektive utsagn

I vitenskapelig forskning og blant brukskvalitetseksperter refereres det ofte til Ericsson og Simons kjente verk (se kapittelet 2). Ericsson og Simon (1984) mener vi kan verbalisere hva vi oppfatter i selve prosessen hvor vi oppfatter. Det som er verbalisert utgjør en protokoll. Ikke alt fra protokollen kan ifølge Ericsson og Simon regnes som pålitelige data. Derfor deler de inn protokollen i tre nivåer. I følge modellen kan ikke nivå 3, inkludert retrospektiv TA, erklæres som valide data.

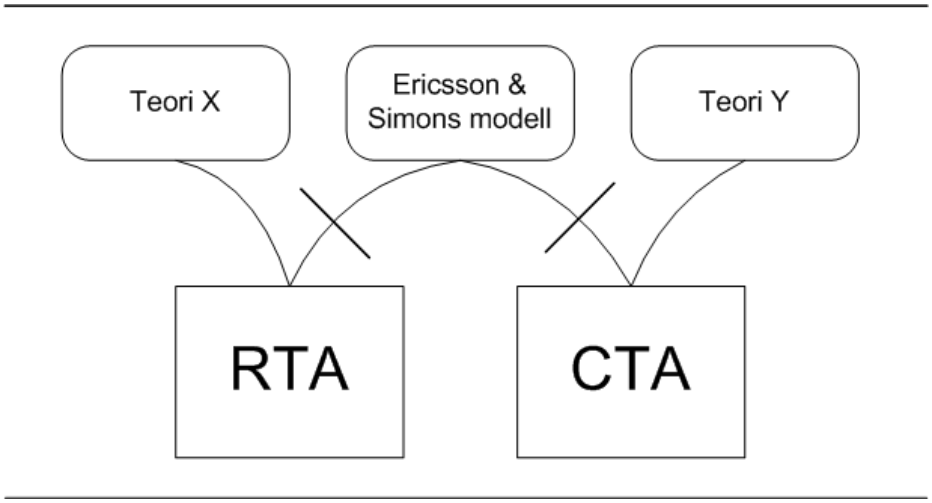
Nivå 3 går kort repetert ut på verbaliseringer som krever ytterligere kognitive prosesser utover det som kreves ved utførelse av oppgaver eller verbalisering. Eksempel på ytterligere kognitive prosesser kan være filtreringsprosesser (for eksempel det å kun uttrykke informasjon som er knyttet til et tema), lage slutning om deltagerens (opprinnelig individ) egen kognisjon, og informasjon som er hentet fra langtidshukommelsen etter anmodning fra tilrettelegger (opprinnelig forskerne). Det kan også være hvilken som helst innflytelse fra utsiden slik som kommentarer eller påminnelse fra tilrettelegger (opprinnelig forsker). Innflytelsen snur følgende verbalisering til nivå 3 fordi den normale flyten av informasjon i korttidshukommelsen igjennom oppgaven har blitt forandret.

Ericsson og Simons betraktet som nevnt *forklarende utsagn,* i dette tilfellet den retrospektive sesjonen, som upålitelige fordi de kunne forvrenge rapporten av hva testpersonen gjorde og i hvilken rekkefølge. Boren og Ramey (2000) har en annen oppfatning, og mener forklarende informasjon kan indikere hvordan informasjon ble prosessert og oppklarer hvilke spesifike strategier mennesker bruker for å fullføre oppgaver. ”This study also shows that the logic inference and strategy explanation information in people’s verbalization also provide valid information about users’ task performance… Usability evaluators can use this information to assess whether a product or interface is successful in supporting users in doing the tasks it is designed for and to identify what parts of the design negatively affect user’s behavior.” (s 1261)

Ønsket til Ericsson og Simon var å gjeninnsette verbale data i vitenskapelig forskning, og derfor utvikle en analyseprotokoll (*protocol of analysis)* hvor all *støy* var fjernet. Dette førte til veldig rigide restriksjoner. De mente at introspeksjon (som er verbalisering av nylig tenkte

tanker) ville forsterke forskningen til oppgaverettede kognitive prosesser som ledet til et informasjonsprosesserings paradigme.

Ericsson og Simon behandler ikke alle deler av de problemer og situasjoner man vanligvis møter i en brukbarhetssituasjon. Dette kan være slik som at datamaskiner henger seg opp, programvarelus og uferdige system. Alt dette forstyrrer monologen. Så da er spørsmålet er om det er rettferdig å bruke Ericsson og Simons modell som bakgrunnsteori for RTA eller CTA for den saks skyld. Ericsson og Simons modell gir en begrensende forståelse av mennesker som verbaliserende individer som handler i isolasjon uten kontekst, uten sanser, uten følelser og ser på mennesker som en slags informasjonsprosess. Derfor er det vanskelig å overføre Ericsson og Simons modell over på brukbarhetstesting (se kapittel 2) slik mange forskere tidligere har gjort. En brukbarhetstest er oppgave- og systemfokusert framfor kognisjonsfokusert. Målet med å innsette verbale data i vitenskapelig forskning er å introdusere utvikling av protokoller for analyse hvor all støy er fjernet, men dette fører til ekstremt strenge innskrenkning.



Figur 6.2: Spørsmålstegn ved Ericsson & Simons modell

Guan et al. (2006) ga ut en artikkel hvor de konkluderer med at RTA er en valid metode. Deres mål var å kartlegge RTA-metodens validitet, evaluere innvirkning av oppgavekompleksiteten på validiteten av RTA og karakterisere hvilken annen informasjon som framkommer under RTA. Under dette forskningseksperimentet ble det benyttet data som fanget øyebevegelser, slik at de kunne sammenligne øyebevegelsene med hva som ble sagt i

den retrospektive fasen. RTA ble betraktet valid på bakgrunn av resultatet. Øyebevegelsene samsvarte i stor grad, mer enn 80 prosent, med sekvensen av hva som ble sagt i den retrospektive sesjonen. Dette stemmer i stor grad med hva testdeltagerne under denne RTA gjennomføringen tippet.

Kapittel 7 Avslutning

Ericsson og Simon argumenter (1984) med at verbale protokoller er nødvendige for å forstå menneskelige handlinger. For å lage gode brukergrensesnitt, er det også nødvendig å forstå menneskelige handlinger. Det finnes uttallige eksempler på dårlige brukergrensesnitt, og mange av grunnene er at designerne/utvikleren ikke har samhandlet godt nok med de fremtidige brukerne. I hverdagen brukes kognitive modeller som å tenke, huske, lære, dagdrømme, se, lese og snakke. En kombinasjon av de kognitive prosessene å snakke og å tenke vil forhåpentligvis være inngangsporten til en stor del av testpersonenes mentale modell av prototypen (eller et ferdigstilt system). Når man løser problemer, planlegger, resonerer eller tar avgjørelser er alle disse kognitive modellene nært knyttet til det å tenke. Dette inkluderer tanker som ”Hva skal jeg gjøre nå? Hvilke muligheter har jeg? Hva er konsekvensene ved denne avgjørelsen? Siden jeg fikk opp informasjon X ved å trykke på knapp Y tidligere, vil jeg kanskje få informasjon A ved å trykke på knapp B?”. Derfor er det viktig at systemer som brukere skal samhandle med, baseres på konseptuelle modeller som gjør det enkelt for brukerne å lære og effektivt å bruke (Donald Norman 1988). Derfor vil vi at testbrukerne skal si høyt hva de tenker, slik at vi får tilgang til de mentale prosessene. Vi har nå sett at dette ikke er en enkel oppgave, verken for CTA- eller RTA-deltagerne. Testdeltagerne er ikke individer uten tanker og følelser, og derfor bør betraktes fra et psykologisk perspektiv. Få, hvis noen i det hele tatt, greier å åpne dørene til ens mentale prosesser ved å verbalisere sine tanker, men vi kan sette døren på gløtt. Mennesket er ikke bare en entitet som framskaffer informasjonsprosesser. Det å tenke er mer enn hva vi eksplisitt kan uttrykke i ord, men ord har vist seg i denne sammenhengen å være ett nyttig verktøy for bruk i brukbarhetstester.

Vi har vist at både RTA- og CTA-metoden har avdekket brukskvalitetsproblemer, og har forhåpentligvis bidratt til å øke produktets kvalitet. Resultatene fra studiet er oppsummert i følgende tabell:

CTA	RTA
• Avdekket et falskt brukskvalitetsproblem • Testdeltagerne brukte lengre tid enn RTA deltagerne på oppgavene • Negative tilbakemeldinger på spørsmål om opplevelse av metoden • Deltagerne misliker å bli filmet • Dobbel kognitiv belastning • Vanskelig å sette ord på tanker • Prestasjonsangst • Føler de må si noe • Deltagerne føler stress og har dårlig tid • Vanskelig å holde TA-teknikken gående • Påminnelser bare midlertidig effekt • Sa noe, tenkte noe annet • Tenker raskere enn man greier å verbalisere • Testsituasjonen har innvirkning	• Brukte kortere tid enn CTA deltagerne • Hoppet over flere oppgaver • Positive tilbakemeldinger på metoden • Foretrekker å sitte uten tilrettelegger • Følte ingen stress og hadde god tid • Liten eller ingen grad av overvåkning • Vanskelig / uvant å verbalisere tanker • Husker mye fra oppgaveløsningen • Kan ikke huske alt man tenker • Kunne stoppe opptaket. • Kunne konsentrere seg i større gras om oppgavene • Erklært valid av Ramey et al.(2006) • Testsituasjonen hadde ingen eller liten innvirkning • Noen savnet det å stille spørsmål til tilretteleggeren

Tabell 7.1: Oppsummering

Kort oppsummert fra intervjuene kan vi si at det var mange negative tilbakemeldinger fra CTA-deltagerne, og det var langt mer negativ enn positiv respons. Responsen fra RTA-deltagerne var stort sett positive. Skal man velge metode ut fra testbrukernes opplevelse, kan man trygt si at RTA kommer bedre ut enn CTA-metoden.

7.1 Sammenheng mellom brukskvalitet og opplevelse

Kan vi si at det er en sammenheng mellom hvordan deltagerne opplevde metoden og resultatene fra brukbarhetstestene? CTA-deltagerne brukte lengre tid på oppgaveløsningen, og i tillegg ser vi at CTA-metoden avdekket ett brukskvalitetsproblem mer enn RTA. CTA-deltagerne brukte lengre tid.

Metodene hadde i utgangspunktet de samme forutsetningene. Det var lik oppfatning av tilrettelegger, omgivelsene var de samme (alle testene ble utført på samme lab), oppgavene var lik for alle, og alle testdeltagerne tilhørte målgruppen. Likevel opplevde de representative testdeltagerne metodene ulikt. CTA-deltagerne sa selv at de trodde metoden hadde innvikning på resultatene i form av (C)TA teknikken, tilrettelegger ved siden av, kameraovervåkning, tidspress og stress. Vi har tidligere sett at tidspresset blant annet kunne henge sammen med overvåkningsfølelsen. Testdeltagerne nevnte både kamera og tilrettelegger som en negativ opplevelse av metoden.

Det er nærliggende å tro at dette har sterk tilknytning til tidsforskjellene mellom metoden og at disse faktorene var med på å framstille et falskt brukskvalitetsproblem. Ut fra resultater og intervjuer er det en klar sammenheng mellom overvåkningseffekten og resultatene fra brukbarhetstestene. Men overvåkningseffekten er ikke alene om å påvirke resultatene.

7.1.1 Verbalisering av tanker

Testdeltagerne sier selv det er en sammenheng mellom oppgavens kompleksitet og verbalisering av tanker, noe vi også ser ut fra resultatene. Preece, Rogers & Sharp (2002) kommenterer at deltagerne antageligvis synes det er vanskelig å snakke når oppgavene blir mer krevende. Snakker man mer, går det ut over konsentrasjonen. Når vi sliter, snakker vi mindre. Og vice versa. Nielsen, Clemmensen og Yssing (2002) mener dette kommer av at man må fokusere på flere kognitive prosessene samtidig, noe som resulterer at prosessene konkurrerer og svekker hverandre.

RTA omgår den doble kognitive belastningen ved å prate etter oppgaveløsningen. Da kan man i første omgang konsentrere seg om oppgaveløsningen, etterfulgt av en oppsummeringssesjon der deltagerne fokuserer på verbaliseringen. Deltagerne synes det er enklere å holde det gående når man kan fokusere på en oppgave om gangen, og i tillegg kan

de trykke pause når de tenkte raskere enn de greide å verbalisere. Tabell 6.3 og tabell 6.4 viser at RTA-deltagere prater mer med tilhørende TA-teknikk enn CTA-deltagerne, som er med på å styrke Nielsen, Clemmensen og Yssings teori (2002) om at CTA-deltagere utsettes for en dobbel kognitiv belastning.

Noen av deltagerne fra CTA-gruppen følte press til å si noe hele tiden, fordi de fikk instrukser om hele tiden om å verbalisere sine tanker og hyppige påminnelser om å verbalisere sine tanker når de ikke sa noe. Vi er igjen inne på overvåkningseffekten hvor deltagerne føler de må prestere noe siden tilretteleggeren sitter ved siden av og maser mens kameraet ruller. Påminnelseseffekten hadde bare midlertidig effekt, siden deltagerne fikk problemer med den doble kognitive belastingen. Ericsson og Simons modell ble også kritisert for deres påminnelse "keep talking" (Boren & Ramey 2000). Det påpekes at modellen alltid vil ha en svakhet med at man skal snakke og tenke samtidlig. Det går kjapt når man tenker tanker som allerede er kodet verbalt, aktivering av gamle tanker er noe tregere og generering av nye tanker er virkelig treg. Ericsson og Simon mener derfor er det viktig å roe ned tankeprosessen eller å øke verbaliseringen. De foreslår av den grunn keep talking. Nielsen, Clemmensen og Yssing. (2002) angriper Boren og Ramey for deres kritikk av keep talking. Bakgrunnen for Nielsen, Clemmensen og Yssings kritikk er at Broren og Ramey fokuserer på kommunikasjonen mellom tilrettelegger og testperson, mens dette faller utenfor Ericsson og Simons fokus. De mener at Ericsson og Simon verken var interesserte i den tekniske eller praktiske bruken i hvordan rådata blir generalisert, men de var opptatt av analysen av protokollene. "The keep talking solution is explained as a means to circumvent the delay in time as verbalizations lack behind thinking. But how do we know that it is a question of speed and not a question of cognitive interference?" (s 106)

Ut fra dette kan det virke som det blir feil å referere til Ericsson og Simons TA-modell i RTA-metoden, og at vi kan få fram verdifulle data uten disse strenge restreksjonene. Boren og Ramey kom fram til følgende tre alternativer tilnærminger for å forsone teori med praksis:

1. Starte en rigid anvendelse av Simon og Ericsson's teori?
2. Slutte å samle verbale data og i stedet fokusere på observerbare utførelser og målinger.
3. Utforske andre teorier

Vi har allerede diskutert det første alternativet og kommet fram til at en slik tilnærming ikke passer i en brukbarhetstest. Mange av brukskvalitetsproblemene som oppsto under testene, ble avdekket etter samtale med testbrukerne. Det andre alternativet krever metoder og teknikker som ingen så langt har funnet opp, og muligens heller aldri vil bli funnet opp. Problemer som blir avdekket ved observasjon og kvantitative målinger, kan ha mange årsaker som først kommer fram ved TA- teknikken, intervjuer eller andre verbale teknikker/metoder. Hvis vi slutter å samle inn verbale data gir vi slipp på en svært verdifull informasjonskilde (Nielsen 1983). Vi står igjen med alternativ nummer tre.

7.1.2 Oppgavens bidrag

Masteroppgaven har i hovedsak bidratt til å kaste lys på testmetodens innvirkning på testresultatene. Denne forskningen svekker CTA-metodens posisjon slik den er anbefalt gjennomført av kjente brukskvalitetseksperter (Nielsen 1993; Dumas & Redish 1999; Bruce Tognazzini 1991). Forskere som Nielsen (Janni), Mia Nørgaard, Erik Frøkjær, Ted Boren, Zhiwei Guan og Judith Ramey.har, som vi har sett, påpekt aspekter ved CTA metoden som kan ha innvirkning på resultatene, men uten å vise til brukernes opplevelse av metoden. Vi har i denne oppgaven belyst dette, og hvorvidt opplevelsen påvirker testresultatene. Oppgaven konkluderer med at CTA-metodens testsituasjon påvirker testresultatene.

Videre har masteroppgaven gitt et nytt bidrag til sammenligning av RTA versus CTA, noe det finnes lite av fra før (se kapittel 2.5.4) Vi ser at testresultatene mellom metodene er nesten helt identiske. Den eneste forskjellen er at CTA-metoden avdekket et falskt brukskvalitetsproblem, et falskt problem som i teorien kunne vært langt dyrere å rette på en de vi avdekket her.

Til tross for at RTA-deltagerne stort sett bare hadde positive tilbakemeldinger på metoden, syntes deltagerne TA-teknikken var vanskelig. Det finnes også alternative metoder hvor testbrukerne skal i større grad samarbeide med tilretteleggeren eller en annen testperson, slik som CUT eller konstruktiv samhandling (se kapittel 2). Denne avhandlingen leverer en grundig evaluering av RTA, og gir grunnlag for å sammenligne metoden med andre metoder. Hvorvidt RTA-metoden har innvirkning på resultatene er vanskelig å si. Vi har vist at RTA-deltagerne synes det er vanskelig å beherske TA-teknikken. Dette temaet krever ytterligere forskning.

7.2 Videre forskning

RTA- deltagerne prater mer under debrief sesjonen enn hva CTA- deltagerne gjør under oppgaveløsningen. Allikevel er sesjonen preget av mye stillhet, men dette varierer voldsomt fra deltagere til deltager. Metoder som CUT og konstruktiv samhandling legger til rette for mer samhandling mellom deltagerne (konstruktiv samhandling, se kapittel 2.5.4) og mellom tilrettelegger og deltager (CUT, se kapittel 2.5.4). En detaljert sammenligning av RTAs debrief sesjon med overnevnte teknikker er et interessant tema for videre forskning.

Denne masteroppgaven viser at CTA metodens testsituasjon påvirker testresultatene, men det er ikke funnet tegn til at RTA har denne innvirkningen. Det betyr ikke at RTA metoden avdekker alle og kun reelle brukskvalitetsproblem. Fagområder innen brukbarhetstesting mangler fremdeles mye forskning, men vi er på god veg.

Referanseliste

- Alsa, B. S. og J. J. Jensen, M. B. Skova. 2005. *Comparison of Think Aloud and Constructive Interaction in Usability Testing with Children*. Proceeding of the 2005 Conference on interaction Design and Children, June, s 9- 16

- Boren, M. T. og J. Ramey. 2000. *Thinking aloud: Reconciling Theory and Practice*. Professional Communication, IEEE Transactions on. Vol. 43, Issue 3, September, s 261- 278

- Den Norske Dataforening. 1994. *Usability = Brukskvalitet*. Tilgjengelig: [http://dataforeningen.no/?module=Articles;action=Article.publicShow;ID=2745] (21 november. 2006)

- Dumas J. S. og J. C. Redish. 1999. *A Practical Guide to Usability Testing*. Exeter: Intellect

- Duncker, K. 1945. *On Problem-solving, in Dashiell*. Psychological Monographs, The American Psychological Association. Vol. 58, s 101-114.

- Ericsson, K. A., & H. A. Simon. 1984. *Protocol Analysis: verbal reports as data*. Cambridge: MIT Press

- Faulkner, L. 2003. *Beyond the five-user assumption: Benefits of Increased Sample Sizes in Usability Testing*. Behavior Research Methods, Instruments, & Computers. Vol. 35, Issue 3, s 379- 383

- Frøkjær, E. og K. Hornbæk. 2005. *Cooperative Usability Testing: Complementing Usability Tests With User-supported Interpretation Sessions*. CHI '05 ACM Press, s 1383- 1386

- Galtung, J. 1967. *Theory and Methods of Social Research*. Oslo: Universitetsforlaget.

- Genov, A. 2005. *Iterative usability testing as continuous feedback: A control system perspective*. Journal of Usability Studiess, Vol. 1, s 18- 27

- Guan, Z., S. Lee, E. Cuddihy og J. Ramey. 2006. *The Validity of the Stimulated Retrospective Think-aloud Method as Measured by Eye Tracking*. CHI '06 ACM Press, s 1253- 1262

- Guindon, R. 1988. *Cognitive Science and its Applications for Human-Computer interaction*. Hillsdale: Lawrence Erlbaum Associates, Inc

- Hartvigsen, G.1998. *Forskerhåndboken*. Kristiansand: Høyskoleforlaget.

- Hoc, J. M. og J. Leplat. 1983. *Avaluation of Different Modalities of Verbalization in Sorting Task*. International Journal of Man-Machine Studies. Vol. 18, s 283- 306.

- Hornbæk, K. og M. Nørgaard. 2006. *What do usability evaluators do in practice? : An Explorative Study of Think-aloud Testing*. DIS '06 ACM Press, s 209- 218

- ISO 13407. 1999 (Ikrafttredelse 2004-04-14.). *Human-centred Design Processes for Interactive Systems*. Tilgjengelig: [http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=21197] (2. Oktober 2006)

- ISO 9241-11. 1998. Guidance on Usability. Ikrafttredelse 2003-05-19. Tilgjengelig: [[http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=16883&ICS1=13&ICS2=180&ICS3] (3. Oktober 2006)

- Kaikkonen, A., A. Kekäläinen, M. Cankar, T. Kallio, og A. Kankainen. 2005. *Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing*. Journal of Usability Studies. Vol. 1, Issue 1, s. 4- 17

- NESH. Den nasjonale forskningsetiske komité for samfunnsvitenskap og humanoria. 2006. *Forskningsetiske retningslinjer for samfunnsvitenskap, humaniora, juss og teologi*. Tilgjengelig: [http://www.etikkom.no/HvaGjorVi/Publikasjoner] (12. februar 2006)

- Nielsen, (Jacob). 1993. *Usability Engineering*. New Jersey: Academic Press.

- Nielsen, (Jacob). 2000. *Why You Only Need to Test With 5 Users*. UseIT.com. Tilgjengelig via [http://www.useit.com/alertbox/20000319.html] (6. februar 2006)

- Nielsen, (Jacob). 2005. *Usability for the masses*. Boston: Academic Press

- Nielsen, (Janni)., T. Clemmensen, & C. Yssing. 2002. *Getting Access to what goes on in People's Heads: Reflections on the Think-aloud Technique*. NordiCHI '02 ACM Press. Vol. 31, s 101- 110

- Nisbett, R. E. og Wilson, T. D. 1977. *Telling More Than We Can Know: Verbal Reports on Mental Processes*. Psychological Review. Vol. 84, issue 3, s. 231- 259

- Personopplysningsloven. 2000. *LOV-2000-04-14-31*. Ikrafttredelse 2001-01-01 Tilgjengelig via: [http://www.lovdata.no/all/nl-20000414-031.html] (3. mars 2006)

- Norman, D.A. 1988. *The Psychology of Everyday Things*. New York : Basic Books

- Preece, J., Y. Rogers, H. Sharp, D. Benyon, S. Holland og T. Carey. 1994. *Human Computer Interaction*. Wokingham : Addison-Wesley

- Preece, J., Y. Rogers & H. Sharp. 2002. *Interaction Design*. New York : Wiley.

- QSR. 2007. Qualitative Research Software for Qualitative Data Analysis and Research. *Programvare: NVIVO 7*. Tilgjengelig via: [http://www.qsrinternational.com/products/productoverview/NVivo_7.htm] (2. september 2006)

- Ramey, J., T. Boren, E. Cuddihy, J. Dumas, Z. Guan, M. J. van den Haak og M. D. de Jong. 2006. *Does Think Aloud Work? How do we know?* CHI '06 ACM Press, April, s 45- 48

- Repstad, P. 1998. *Mellom nærhet og distanse*. Universitetsforlaget AS

- Robson, C. 2002. *Real World Research: a Resource for Social Scientists and Practitioner*. Oxford : Blackwell

- Stone, D., C. Jarrett, M. Woodroffe og S. Minocha. 2005. *User Interface Design Evaluation*. Amsterdam: Elsevier

- Tognazzini, B. 1991. *Tog on Interface*. Mass: Addison-Wesley

- Tognazzini, B. 2000. *If They Don't Test, Don't Hire Them*. Tilgjengelig: [http://www.asktog.com/columns/037TestOrElse.html] (10. august 2006)

- Van den Haak, M. J., M. D. T. de Jong. 2003. *Exploring two Methods of Usability Testing: Concurrent versus Retrospective Think-aloud Protocols*. IEEE International. September, s 285- 287

- Van den Haak, M. J., M. De Jong, P. Jan Schellens. 2003b. *Retrospective vs. Concurrent Think-aloud Protocols: testing the Usability of an Online Library Catalogue*. Behaviour & Information Technology. Vol. 22, Issue 5, s 339- 351

Vedlegg

Appendiks A: Sammendrag fra RTA sesjonen

Dette vedlegget er sammendragene fra RTA sesjonen. Hver brukbarhetstest ble transkribert og denne oppsummeringen inneholder kommentarer fra testdeltagerne brukte pr oppgave, notarer fra testen og øvrige observasjoner.

1. Finn ut hva som er til middag i morgen der du studerer

Testbruker #11 gikk til dagens middag med det samme, men skjønner ikke at menyen gjelder alle kafeene hvor man serverer middag.

> *"Det tok litt tid før jeg skjønte at dette var de to rettene som serveres"*

Derfor benyttet testpersonen mye tid på å lete under de forskjellige kategoriene, før han skjønte sammenhengen. Testbruker #8 hadde også litt av det samme problemet;

> *"Jeg sjekket hva som var dagens middag, så sjekke om den kafeen som hadde fantes på Gløshaugen"*

Hun skjønte at dette var to menyer som serves på de ulike kafeene, men måtte undersøke om Gløshaugen hadde en slik kafé.

2. Kan man bestille time hos lege via nettstedet?

Her ser vi at fire av seks fullførte oppgaven med god tidsmargin. To av testpersonene løste ikke oppgaven på tilfredsstillende måte. Den ene gikk via helsestasjon for studenter. Dette er en ordning hvor studentene møter opp uten å bestille time. Dermed endte kandidaten å svare "nei" på spørsmålet om man kan bestille time hos lege. Den andre kandidaten hadde hoppet litt fram og tilbake i oppgavene hun løste, så hun overså rett og slett oppgave 2. Men det er for tynt grunnlag å si at denne er en kandidat for forbedring, da testbruker #10 ikke helt visste forskjellen på helsestasjon og legesenter, og testbruker #12 hoppet over oppgaven.

3. Kan du bestille time hos psykolog?

En av testpersonene så linken ”bestill time” fra undermenyen ”psykisk hjelp”, og derfor erklærte oppgaven som løst. Dette er ikke direkte feil, men vi ønsket at personen skulle finne ut hvordan de bestilte time. En av testbrukerne brukte 7 sekunder over estimert mål. Denne virker ok.

4. Du fikk lyst på sjokolade. Hvor ligger storkiosken på Dragvoll?

Her har vi en outsider. Testbruker #8 bruker lang tid på å løse denne oppgaven. Feilen som ble gjort, var at vedkomne leter under ”kart”, men etter hvert finner hun fram. Også en testbruker kommenterte at det var uklart hva den oransje prikkene betydde. I teksten står det;

”Velkommen til oss på Dragvoll, vi er den **oransje prikken i bygg 5** (post og storkiosk)”

mens symbolforklaringen sier;

Det skal ikke mye forandring til for å få dette klarere fram. Det er også mye informasjon i kartet som er unødvendig i forhold til kartets beliggenhet. Det kunne for eksempel bare vært et kart over Dragvoll med en prikk som viser hvor kiosken ligger. Et alternativ er en flash animasjon med hvor man kan velge hva man vil se. Dette kartet kunne man også eventuelt brukt under ”kart”. Dette kan være en typisk lokal feil, som ikke prioriteres på høyeste nivå, men den må dokumenteres.

6. Skriv ut timeplanen for det stedet du vil trene.

Tilsynelatende ser denne grei ut, i forhold tiden testbrukerne brukte. Men kommentarene sier noe helt annet;

"Det burde være en print funksjon. Jeg har dårlige opplevelser med å skrive ut bra en browser. Da får utskriften ramme. Når trykker jeg bare på print og håpet på det beste. Jeg synes det er litt dårlig tilbakemelding" #7

*"Jeg så etter om det var en utskriftsversjon" (*ser for seg at hele margen kommer ut)#10

"Jeg så ingen utskriftsversjon. jeg ser for meg hele siden kommer ut" #9

"Usikker på om jeg har skrevet ut på riktig måte, eller om det er noen enklere måte å få ut selve timeplanen. Nå vil jeg få ut hele skjermbildet med ramme og sider…antagelig er sidene fordels på to steder. Jeg leter etter en link som heter skriv ut, som skriver ut akkurat det vi er på utkikk etter" #11

Leter etter en link hvor man skriver ut
#12

Testbruker#8 skriver ikke ut, fordi hun overså oppgaven.
Dette er en klar kandidat til forbedring.

Problem: #1	Skriv ut
Scope:	Global
Frekvens:	6/6
Utdyping:	Testbrukerne er usikre på hvordan de skal skrive ut. De leter etter en utskriftslink eller ikon. Flere av testbrukerne sier de ikke vet hva som vil bli skrevet ut, slik som ramme, timeplan over to sider osv
Forslag til forbedring:	Innfør ett skriv ut ikon som skal brukes konsekvent på hele nettstedet

7. Du trenger en plass å bo i Trondheim, og vil sjekke om SiT kan hjelpe deg med å finne leilighet. Du har samboer men ingen barn. Finn en leilighet eller studentby du har lyst å søke på, og se om det er ledige plasser eller venteliste.

Testbruker #11 brukte mest tid på denne oppgaven, og ved å se på videoopptaket ser vi at denne personen gikk rett på "ventelisten", og prøver etterpå å finne en fristende plass å bo. Under flere studentbyer valgte han "Boligdetaljer og pris", men noen av linkene til boligtorget var defekte.

Testbruker #12 savner linker direkte fra "ventelisten" Slik det er nå må man sjekke hver enkelt plass, uten å gå gå bort fra ventelisten. Ventelisten er heller ikke oppdatert til enhver tid, noe som virker forvirrende på studentene Denne tabellen viser forskjellen i hvilken informasjon de legger ut på nettstedet;

SiT.no			***Boligtorget***	
Studentboliger	*Type*	*Antall Ledig*	*Antall ledig*	*Korrekt informasjon*
Nedre Singsaker	Hybel	4	3	Nei
Steinan Studentby	Dublett	1	1	Ja
Berg studentby	Hybel	3	4	Nei
Karinelund	Hybel	2	0	Nei
Jakobsliveien 55	Parbolig	1	1	Ja

Testbruker #10 bestemmer seg for en parleilighet, men fullfører ikke oppgaven med å se om den er ledig eller ikke. Han begrunner dette med;
"Siden den dukket opp, regnet jeg med den var ledig"

Testbruker #7 syntes det var vanskeligere å forstå hva de forskjellige boligtypene var; "Her er det oversikt over forskjellige ting som er ledig, men jeg har for eksempel ikke peiling på hva en dublett er. Man burde kunne klikka på dette og fått informasjon på forskjellen.

Problem: #2	Finne bolig
Scope:	Global
Frekvens:	6/6
Utdyping:	Testdeltagerne har store problemer med å finne det de leitet etter, samt å navigere seg mellom venteliste, aktuell bolig og boligtorget. 4 av 6 testdeltagere var innom boligtorget en eller flere ganger for å prøve å få svar på det var ledig bolig. En person nevnte også at det burde vært linker til hva de forskjellige boligtypene betydde.
Forslag til forbedring:	Direktelink fra ventelisten til de aktuelle boligene På hver av boligene bør det stå hva som er ledig. Når man klikker på søknadsskjema, burde man komme til selve søknadsskjemaet. Se bilde X.

8. Du har oppdaget flere symptomer på stress. Du har liten tid til overs, og vil gjerne forhøre med en spesialist før du eventuelt bestiller time hos lege. Hvilke tjenester tilbyr sit.no?

Dette er interessant. Alle greier det under tidsmålet, men alle har en alternativ løsning. Alle ville sendt e post. #12 mener at man kunne satt opp en rød boks som reklamerer for nettrådgivningstjenesten.

Kunne spørsmålet vært stilt annerledes?

”9 Ble forvirret over at det åpnet seg et nytt vindu med nettrådgivning. Dette er en aktuell kandidat for neste runde.

Problem: #3	Nettrådgivning
Scope:	Lokal
Frekvens:	6/6
Utdyping:	Nettrådgivningen er bortgjemt. Testbrukerne vet ikke hva dette er, og fant derfor andre alternative måter å søke hjelp på. Det skaper også forvirring når submenyene til nettrådgivning dukker opp i ett nytt vindu.
Forslag til forbedring:	Synliggjøring av tjenesten. En av testbrukerne mente at en ”lilla boks” ville gjøre nytten. Man kan også bruke hovedsiden til ”rådgivning” som reklame for tjenesten. ”Spør oss” under nettrådgivning kan med fordel flyttes på toppen av undermenyene.

9. Finn et organisasjonskart over SiT

To av testbrukerne var innom ”kart” først, men disse var ikke helt sikre på hva et organisasjonskart var. Ellers intet negativt rapportert. Denne ser ikke ut til å være en kandidat.

10. Skriv ut en oversikt over alle ansatte som hører til under avdelingen ”bolig”

#7 Leter etter en printknapp, men finner ikke. Så ser han over en gang til, men finner den fremdeles ikke.
#8 og #9 skrev ikke ut. #8 leitet etter en måte å skrive den ut, men så ingen utskriftslink. Hun trodde ikke at dette var så nøye å skrive ut, så lenge hun lenge hun fant ansatte. Men hun poengterte etter testet at hun ville valgt skriv ut fra menyen.
#10 Brukte lang tid på denne oppgaven. Hun leita

11. Hvor kan du gi tilbakemelding til SiT?

Bortsett fra testdeltager #7, fullførte alle denne oppgaven under tidskravet. Når det gjelder #7, er han en utvekslingsstudent fra Nederland. Han kunne snakke, skrive og lese norsk uten problemer, men utrykket ”ris og ros” skjønte han ikke med det samme,

så hans besvarelse vil ikke telle i denne sammenhengen. ”Ris og ros” funksjonen er utelatt som forbedringskandidat.

12. Finn resepsjon og åpningstider

Bortsett fra en person som løste denne oppgaven på en alternativ måte, som var like grei, må det konstateres at det var enkelt å finne fram til åpningstider og ekspedisjon.

#12 ”det er litt gjemt og jeg var litt konsentrert om undermenyen” og poengterer at ”kontakt oss” også kanskje burde vært en egen undermenylink, men også at det da kanskje ville vært litt mye dobbelt opp. Denne testdeltageren synes også at menyen burde inneholdt ”om sit” i stedet for ”hva er sit”.

Om oppgaven

”Oppgavene var litt enkle, siden ordene i oppgavene er akkurat de samme ordne på siden”

Problem: #4	Oppgave 1: Dagens middag
Scope:	Lokalt problem
Frekvens:	2 av 6
Utdyping:	Testbrukerne (2 stk) skjønte ikke at *”Det tok litt tid før jeg skjønte at dette var de to rettene som serveres”*
Forslag til forbedring:	Hva med forklaring at dette serveres i alle Tellus kafeene under tabellen?

Appendiks B: Sammendrag fra CTA sesjonen

Dette vedlegget er sammendragene fra CTA sesjonen. Hver brukbarhetstest ble transkribert og denne oppsummeringen inneholder kommentarer fra testdeltagerne brukte pr oppgave, notarer fra testen og øvrige observasjoner.

1. Finn ut hva som er til middag i morgen der du studerer

Testbruker #3 brukte 1 minutt og 34 sekunder. Her ser vi at testbrukeren først går helt riktig, men tror ikke at hun har løst oppgaven. Da testbrukeren ser dette bildet, skjønner hun ikke at dette serveres på alle kantinene ved NTNU.

Dagens middag

Dato		Husmann	Østen
Mandag	27.11	Blomkålgrateng med skinke	Tex Mex gryte
Tirsdag	28.11	Juletallerken	Orientalsk svinenakke
Onsdag	29.11	Pasta bolgonese	Thailanske kjøttboller
Torsdag	30.11	Potet bakt sei med salat	Indisk karrygryte
Fredag	1.12	Kafeens egen godbit	Fredagsbuffet

Testbrukeren leter videre under ”kafeer” i stedet for dagens middag. Testbruker #5 har akkurat det samme problemet, og kommenterer at det burde linkes videre til dagens middag fra kafeer. [Testbruker #6]

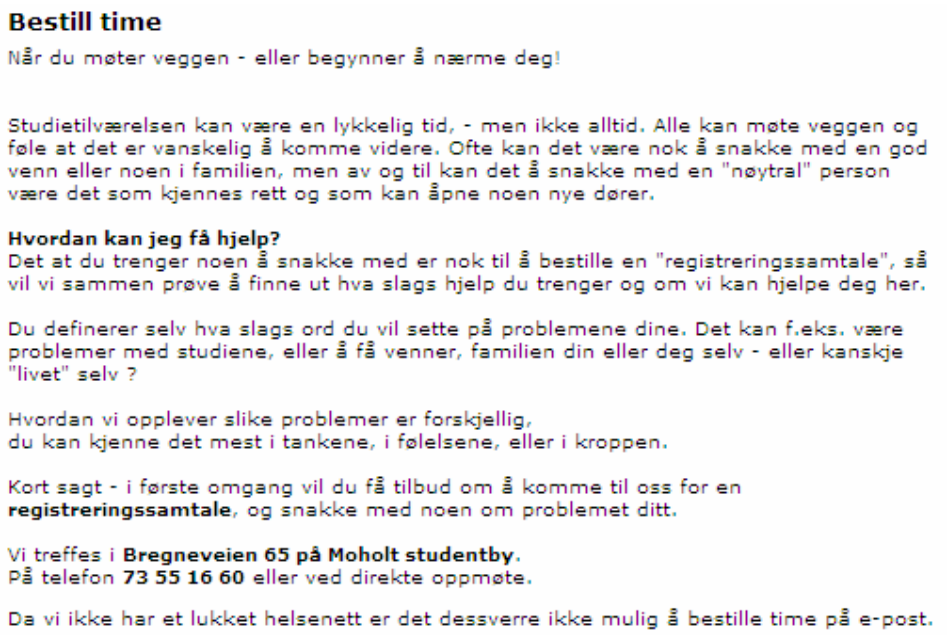
Slik dagens middag er organisert nå, forutsettes det at brukerne forstår at det er felles middag uavhengig av kantine. Problemet er bare at alle ikke vet dette.
Tre navigasjonsfeil innen høyremenyen fordelt på tre personer.

2. Kan man bestille time hos lege via nettstedet?

Alle deltagerne greide målet. Gjennomsnitt tiden ligger på 39 sekunder, og vi kan fastslå ut fra denne brukbarhetstesten at denne ikke er en kandidat til forbedring
Ingen navigasjonsfeil

3. Kan du bestille time hos psykolog?

Denne oppgaven har lik framgangsmåte som oppgave 2, bare man skal velge psykisk hjelp i stedet for legesenter. To av testbrukerne ligger 5 sekunder over målet, mens to bruker 15 sekunder over målet. Det ble ikke notert noe spesielt på disse løsningene, verken i tilretteleggerens notater, eller på målingsskjemaet. Ved å se på videoene, kan man konstatere at det var mye tekst på siden, og man måtte lese litt får og finne ut om hvordan man kunne bestille time hos psykolog. Ingen navigasjonsfeil.



Bestill time

Når du møter veggen - eller begynner å nærme deg!

Studietilværelsen kan være en lykkelig tid, - men ikke alltid. Alle kan møte veggen og føle at det er vanskelig å komme videre. Ofte kan det være nok å snakke med en god venn eller noen i familien, men av og til kan det å snakke med en "nøytral" person være det som kjennes rett og som kan åpne noen nye dører.

Hvordan kan jeg få hjelp?
Det at du trenger noen å snakke med er nok til å bestille en "registreringssamtale", så vil vi sammen prøve å finne ut hva slags hjelp du trenger og om vi kan hjelpe deg her.

Du definerer selv hva slags ord du vil sette på problemene dine. Det kan f.eks. være problemer med studiene, eller å få venner, familien din eller deg selv - eller kanskje "livet" selv ?

Hvordan vi opplever slike problemer er forskjellig,
du kan kjenne det mest i tankene, i følelsene, eller i kroppen.

Kort sagt - i første omgang vil du få tilbud om å komme til oss for en **registreringssamtale**, og snakke med noen om problemet ditt.

Vi treffes i **Bregneveien 65 på Moholt studentby**.
På telefon **73 55 16 60** eller ved direkte oppmøte.

Da vi ikke har et lukket helsenett er det dessverre ikke mulig å bestille time på e-post.

Alle testdeltagerne navigerte seg raskt fram til denne siden, så man kan bare stille spørsmålstegn ved teksten. Dette er ikke et typisk brukskvalitetsproblem, siden testbrukerne vet de er på riktig sted, men her kan man forandre på teksten om man ønsker. Dette er ikke en kandidat for forbedring, men nevnes for utviklerfirmaet.

4. Du fikk lyst på sjokolade. Hvor ligger storkiosken på Dragvoll?

Testbruker #5 brukte 2 minutter og 35 sekunder på denne oppgaven, og dette er en typisk outsider. Outsidere er ikke ufarlige, fordi de kan representere en stor prosent av alle framtidige brukere av nettstedet. Når vi ser på videoopptak og notater, kommer det fram at vedkomne leter på feil plass. I stedet for å lete under toppmenyen til ”mat og drikke”, velger han å lete etter de ulike avdelingene til sit”. [MER]
Hvorvidt dette er et brukskvalitetsproblem eller ikke, er vanskelig å si. Denne er en kandidat for neste iterasjon med brukbarhetstesting. En navigasjonsfeil av toppmeny.

5. Åjsann, det ble visst litt mye sjokolade. Finn et sted du kan trene aerobic, på et tidspunkt som passer din timeplan.

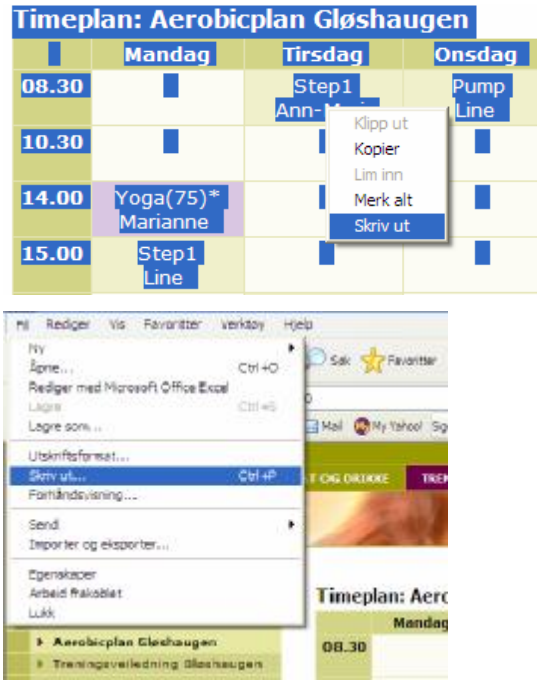
En person ser ikke treningstabben i første omgang, og leter etter under de forskjellige avdelingene til SiT. En annen valgte riktig toppfane, men roter litt rundt i høyremenyen før han finner riktig timeplan. Det er disse to som bruker over estimert tid. Denne er også en kandidat for neste runde, men det er vanskelig å peke på eksakt hva som gjør at et par stykker sliter på denne oppgaven. Resultatene fra forstudiet sier:

> ”Det var ikke så enkelt å finne timeplaner for treningsstedene, de kunne f.eks. vært linket fra treningstilbud. Ingen kommenterte at timeplanene var ulike, men det var vanskelig å lage en oppgave som gjorde at de måtte lese flere timeplaner. Det er kanskje heller ikke en realistisk problemstilling.”

6. Skriv ut timeplanen for det stedet du vil trene.

Denne er veldig inntresann. Samtlige av testbrukerne sier at de savner en utskiftsknapp eller en utskriftslink;
”Jeg fant ingen utskriftsknapp…”
”Jeg tenkte at det var en skriv ut knapp…”
”[Testbruker #3]”
”Så ingen skriv ut link…”
”Er det noe utskriftsvennlig format her da…”
”Det skulle kanskje vært en utskriftsknapp…”

Dette er en superkandidat som forbedringsobjekt. Alle savner en utskriftslink eller link som ikke er der. Likevel greier alle testbrukerne å skrive ut timeplanen ved å bruke browseren. Tre av testbrukerne merket timetabellen, [sjekk opp] får så å høyreklikke og velge ”skriv ut”. De andre valgte bare ”skriv ut” i browser menyen, og trodde at ”alt rundt” (marg, toppmeny etc.) ble skrevet ut. En var også i tvil om hele tabellen kom til å være med på utskriften.



7. Finn en par- leilighet du har lyst å søke på, og se om det er ledige plasser eller venteliste

Gjennomsnittstiden på denne oppgaven er 3 minutter og 37 sekunder. Dette er langt over hva som ble forventet. Man kan kanskje tro at dette er en oppgave det er lett å bruke tid på å finne en passe plass å bo, men dette var ikke tilfelle i noen av oppgaveløsningene. Alle testdeltagerne utenom testdeltager #4 var innom boligtorget minst en gang. Målet var at testbrukerne skulle løse denne oppgaven uten å gå via boligtorget. En av testdeltagerne hopper fram og tilbake mellom sit og boligtorget 5 ganger. [Mer] Et av hovedprisnippene innen design er kontinuitet, og når man hopper fra ett nettsted til et annet, er det lett å bli forvirret. Det var også lett å se at flere av testdeltagerne hadde problemer med å navigere seg fram på boligtorget, og mange fortsatte å lete, selv om svaret sto rett foran nesa på dem. Denne er også en sterk kandidat til forbedringsprosessen.

Problem:	**Bolig** (Oppg. 7)
Scope:	Lokalt / Globalt
Frekvens:	4/6
Utdyping:	Testpersonene brukte lang tid på denne oppgaven (Gj. snittstid på 3 min 37 sek). Mye kaos. Når man kommer inn på boligtorget (et annet nettsted med annen oppbygging), blir mange av testdeltagerne forvirret. Ventelisten er heller ikke oppdatert. Noen av testdeltagerne nevner at de savner en direktelink fra ventelisten til den aktuelle boligen. Og når man trykker på "til søknadsskjema" under den aktuelle boligen, kommer man inn på nytt nettsted man må sette seg inn i, for så å navigere seg fram til den rette boligen, velge type leilighet og så sjekke om det er ledig bolig. Enkelte av "boligdetaljer og pris" linkene var også defekte, men i skrivende stund ser det problemet ut til å være løst
Forslag til forbedring:	• Direktelink fra ventelisten til de aktuelle boligene • På hver av boligene bør det stå hva som er ledig • Når man klikker på søknadsskjema, burde man komme til selve søknadsskjemaet. Se skjermbilde 2 vedlegg 1.

8. Du har oppdaget flere symptomer på stress. Du har liten tid til overs, og vil gjerne forhøre deg med en spesialist før du eventuelt bestiller time hos lege. Hvordan kan sit.no hjelpe deg?

Testbruker #5 løste oppgaven ved å kontakte noen som arbeidet i helsesektoren, noe som ikke kan godtas som løsnings, fordi meningen er å benytte SiT' s nettrådgivningstjeneste. Testbruker #2 hoppet over denne oppgaven, av ukjente årsaker. Den eneste som var innen undermenyen "spør oss" var testbruker #3 og #6, men de skjønte ikke hva denne tjenesten gikk ut på, og endte opp med;

"Jeg ville ha ringt" og "jeg sender mail"

Dette er første siden man kommer til når man velger nettrådgivning. Målet var at studentene skulle benytte seg av ”nettrådgivning”, og på denne siden skriver SiT; ”vi tilbyr nå også gratis nettrådgivning”. Meningen er at studentene skal benytte linken i høyremenyen som heter ”Rådvill? Spør på nett”, for så å velge ”spør oss!” for å stille konkrete spørsmål innen diverse kategorier. Flere av testdeltagerne var innom ”Rådvill? Spør oss”, og to av de var også innen ”Spør oss”. Problemet var bare at ingen helt skjønte hva dette var.

”Nå vet jeg ikke helt hva jeg er kommet inn på”

Ifølge CTA metoden er denne absolutt en kandidat. Den må gjøres mer synlig.

Problem:	**Nettrådgivning** (Oppgave 8)

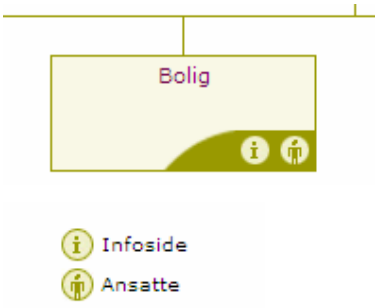
Scope:	Lokalt / Globalt
Frekvens:	6 / 6
Utdyping:	Ingen brukte nettrådgivningstjenesten, til tross for at 2 av testdeltagerne var innom ”rådvill? Spør oss”. Dette ble begrunnet med at de ikke visste om tjenesten, eller at den ikke var nok synlig. Alle testdeltagerne ville løst denne på en alternativ måte, bortsett fra en testbruker som ville hoppet over oppgaven.
Forslag til forbedring:	Nettrådgivningen burde gjøres mer synlig. Nettrådgivningen burde ikke åpnes i nytt vindu?

9. Finn et organisasjonskart over SiT

Ut i fra timetabellen kan man lett tro at dette er en klar kandidat til forbedringsprosessen. Men den er ikke det. Testbruker #3 er den første som sliter, og personen innrømmer at han/hun ikke vet hva et organisasjonskart er. Testperson #5 vet hva et organisasjonskart er, men brukte bare lang tid på oppgaven. Han gjorde ingen navigasjonsfeil, og mente at den lå på en logisk plass. Når det gjelder siste testbruker, forteller han først etter testen at han heller ikke visste hva et organisasjonskart er.

10. Skriv ut en oversikt over alle ansatte som hører til under avdelingen ”bolig”

Fire av testbrukerne sa det vanskelig å se linken. En av testdeltagerne ser:



Men allikevel ser han ikke ikonet i boligblokkene.
Alt peker på at denne linken blir for lite framhevet.

Problem:	**Skriv ut** (oppgave 6 og 10)
Scope:	Global
Frekvens:	6 / 6 og 4/6
Utdyping:	Dette globale problemet oppstår når testdeltagerne skal skrive ut. *Skriv ut timeplan*: Alle testdeltagerne kommenterer at de savner en utskriftslink / utskriftsvennligversjons link *Skriv ut ansatt:*
Forslag til forbedring:	Innfør ett skriv ut ikon med teksten ”skriv ut” som er med konsekvent på hele nettsiden til SiT. Når man klikker på den kan, ”skriv ut” vinduet dukke opp foran en utskriftsvennlig versjon. Se skjermbilde 1 i vedlegg 1.

Problem:	**Ansatte**
Scope:	Lokal
Frekvens:	4/6
Utdyping:	Bortsett fra en testdeltager, var det ingen som så ikonlinken. Etter oppgaveløsningen kommenterte de fleste det er vanskelig å se linken. Enten at den går for mye i ett med layouten slik at den blir skjult, og /eller at linken er for liten.
Forslag til forbedring:	Gjøre linken større, og vurder bruk av andre farger til selve linkene.

11. Hvor kan du gi tilbakemelding til SiT?

En testbruker brukte tre sekunder over tidsmålet, ellers kom alle under. Intet negativt ble registrert på testbruker #1, så denne er utelukket som brukskvalitetsproblem (Også på grunnlag av at det ikke ble notert noen negative kommentarer fra andre deltagere eller i tilretteleggerens notater.)

12. Når er SiT's ekspedisjon på Gløshaugen åpen?

#1 ville sent forespørsel om når de har åpnet.

#2 greide det fint

#3 Er inne på "kontakt oss" og leter i undermenyen. Finner det i bunnmenyen på "kontaktskjema", men leter videre i "kontakt enheter og bedrifter". Er inne på kontakt oss igjen. Men ser bare innom undermenyene. Ser også innen rådgiving. Så innom bolig. "Spørs om det gjelder for alt, eller bare bolig rett og slett". Ville tilslutt ringt noen på resepsjonen på Moholt 13.29- 17.41

#4 Ville kontaktet enheter og bedrifter.

#5 27.44- 28.40 Ville brukt bunnlinjen

#6 Går via resepsjon. Men så til hva er SiT, og kontakt oss. "burde kanskje hatt en kontakt oss i toppmenyen".

Appendiks C: Intervjuguide

GUIDE

Takk for at du var villig til å stille opp på denne brukbarhetstesten, noe som betyr mye for min masteroppgave. Først av alt vil jeg garantere deg at du vil forbli helt anonym. Ditt navn og andre personalia som kan avsløre din identitet vil ikke bli lagret på noe vis. Jeg hadde satt stor pris på om jeg kunne ta opptak av intervjuet, slik at jeg lettere kan konsentrere meg om selve intervjue uten å måtte skrive. Er det ok? Opptakene vil bli slettet så snart oppgaven er ferdigskrevet.

Om testbrukeren:
Studie / arbeidssitasjon
Dataerfaring
Internetterfaring

Før oppgaveløsningen
Forventinger til brukbarhetstesten
Følelser før testen
Opplevelse av introduksjon

Overvåkning
Grad av overvåkning
Opptak av skjerm (RTA og delvis CTA)
Båndopptaker (RTA)
Tilretteleggeren ved siden av (CTA)
Kamera (CTA)

Oppgaveløsningen
God/Dårlig tid?
Avslappet/stresset?
Grad av konsentrasjon?

Ville stille spørsmål? (CTA)

Var det noe du lurte på under testen, som du lot være å spørre om? (RTA)

Ville prestere gode resultater?

Redd for å fornærme? (CTA)

Debrief

God/Dårlig tid? (RTA)

Avslappet/stresset?

Tilfredsstillende å kunne hjelpe?

Brydd av sine prestasjoner? (RTA)

Redd for å fornærme? (RTA)

Stoppe opptak (RTA)

TA

Opplevelse av teknikk

Husket de alt de tenkte? (RTA)

Greide du å verbalisere dine tanker? Hvordan opplevde du det?

Holdte du det gående?

Hvordan var det å sette ord på situasjoner der du sliter (CTA)/ hvor de hadde slitt (RTA)?

Synkronisering av tenke versus snakke (hurtighet)?

Sa du noe du ikke tenkte?

Metoden

Testmetodens innvirkning?

Opplevelse av testmetode?

Hvordan opplevelse tilretteleggeren?

Takk så mye for at du hjelper meg med dette, og at du har stilt din tid til disposisjon. Til slutt vil jeg spørre deg om det er noe du vil tilføye som ikke er blitt dekket under dette intervjuet?

Appendiks D: Spørreskjema

Evaluering av testmetoden

Hvor gammel er du?

15-20 21-25 26-30 31-35 Over 36

Hvor arbeider du?

Hva arbeider du med?

Hvordan vil du karakterisere dine datakunnskaper?

Ingen Liten Middels God Vet ikke

Hvordan opplevde du å si verbalisere egne tanker?

Vanskelig litt vanskelig Ok lett Vet ikke

Stoppet du ofte med å snakke?

Ja Av og til Nei Vet ikke

Var det vanskelig å uttrykke/sette ord på problemer under evaluering av video-opptaket?

Ja Av og til Nei Vet ikke

Hvordan opplevde du å samarbeide under oppgaveløsningen?

Vanskelig litt vanskelig Ok lett Vet ikke

I hvilken grad følte du deg overvåket under testen?

I høy grad I middels grad I liten grad Ingen grad

Vet ikke

I hvilken grad var du stresset under selve oppgaveløsningen?

I høy grad I middels grad I liten grad Ingen grad

Vet ikke

I hvilken grad følte du deg komfortabel under testen?

I høy grad I middels grad I liten grad Ingen grad

Vet ikke