NTNU
Det skapende universitet

Brukbarhetstesting av en mobil butikkløsning i laboratorium og felten

En sammenligning av tre metoder

Kenneth Devik

Master i informatikk
Oppgaven levert: Juni 2009
Hovedveileder: Dag Svanæs, IDI

Norges teknisk-naturvitenskapelige universitet
Institutt for datateknikk og informasjonsvitenskap

Sammendrag

I dette studiet har vi sammenlignet resultatene fra tre metoder for å brukbarhetsteste en mobil butikkløsning. De tre metodene er en felttest utført i en dagligvarebutikk, en fullskalatest i laboratorium med realistisk kontekst og en test i laboratorium uten realistisk kontekst. Det var fem testpersoner i hver metode, og testpersonene i felten skilte seg ut som en egen brukergruppe med signifikant høyere alder og dårligere datakunnskaper. Dette kan til en viss grad ha påvirket validiteten til funnene.

Oppdagelsen av antall brukbarhetsproblemer har vist seg å være nesten lik mellom de tre metodene. Det er heller ingen stor forskjell på alvorlighetsgraden til problemene som ble oppdaget i de ulike metodene. Ingen av metodene klarte å avdekke mer enn 59% av de oppdagede brukbarhetsproblemene. Det viste seg at alle problemene i teorien kunne blitt oppdaget i alle tre metodene, og at ingen problemer dermed var knyttet direkte til konteksten i en metode. Etter en gjennomgang med butikkløsningens systemleverandør viste det seg at det eneste kritiske problemet som bare ble oppdaget i felten ikke ble prioritert å rette av dem. Dermed kunne alle kritiske prioriterte problemer bli oppdaget i laboratoriemetodene. Av de prioriterte problemene ble bare 3 av 7 kritiske brukbarhetsproblemer oppdaget i felten.

Den lave realismen i laboratoriet uten realistisk kontekst førte til oppdagelsen av et *falskt positivt problem*. Det var knyttet til at testpersonene ikke skjønte at de kunne benytte varelappene som lå på bordet under hele brukbarhetstesten. Det falske positive problemet ga innvirkning på både effektiviteten og tilfredsheten. Tidsbruken var lik mellom de forskjellige metodene hvis vi trekker fra den ekstra tiden det falske positive problemet medførte at testpersonene brukte i laboratoriet uten realistisk kontekst. Tilfredsheten var best i fullskalametoden, mens den var en del lavere for de andre metodene. Grunnen til det skyldes trolig brukergruppen i felten og det falske positive problemet i laboratoriet uten realistisk kontekst. På grunnlag av de to momentene blir det derfor vanskelig å sammenligne tilfredsheten.

Studien har vist at felttesten ikke gir noen ekstra verdi og at de utfordringene med å teste i felten ikke er verdt strevet. Vi mener at en enkel laboratoriemetode med begrenset gjenskapning av realistisk kontekst, er tilstrekkelig for å avdekke alle brukbarhetsproblemer knyttet til mobile butikkløsninger. Vi anbefaler bruken av ekte varer i denne metoden for å unngå falske positive problemer. Det er ikke nødvendig å benytte statister som spiller kunder. Under felttesten ble de butikkansatte avbrutt tolv ganger av kunder, men de var tydeligvis vant til det, og avbrytelsene førte ikke til problemer vedrørende gjennomføringen av oppgavene.

Forord

Denne masteroppgaven i informatikk med studieretningen systemarbeid og menneske-maskin interaksjon, er laget ved institutt for datateknikk og informasjonsvitenskap, ved Norges Teknisk-Naturvitenskaplig Universistet i Trondheim.

Jeg vil takke veileder professor Dag Svanæs for råd og innspill det siste året. Alt utstyr som ble brukt i felttestingen var lånt av Norsk Senter for Elektronisk Pasientjournal(NSEP). Brukbarhetstestingen i laboratorium ble gjort på NSEP, og jeg vil takke overingeniør Terje Røsand for hjelpen under testingen i laboratoriet og fotograferingen av illustrasjonsbilder.

Uten Lindbak som støttespiller og pådriver hadde aldri denne studien blitt realisert. Jeg vil takke Trond Klevstuen for initiativet med å definere en masteroppgave fra Lindbak. Jeg vil også rette en takk til Tormod Fjeldskår og Asle Mæhre fra Lindbak for vurderingen av brukbarhets-problemene fra et systemleverandørperspektiv.

Til gjennomføring av brukbarhetstestene vil jeg rette en stor takk til alle 15 personene som stilte opp, og en spesiell takk til daglig leder Bjørg Jørstad ved Coop Mega i Steinkjer som gjorde felttestingen mulig.

Lars Flem har vært en god støttespiller til å vurdere testoppgavene og brukbarhetsproblemene som oppstod. Jeg vil også takke samboer Magnhild Hoås for korrekturlesingen og for å ha holdt ut det siste semesteret sammen med en stresset masterstudent.

Til slutt vil jeg rette en takk for det fine samholdet blant medstudentene på arbeidsplassen Gribb i IT-vest. Alle har vært med å gjøre masterstudiet til to fantastiske år. Gribb har bestått av følgende personer: Amund Mortensen, Aleksander Grande, Glenn Ruben Bakke, Bjarne Muri, Mats Ringstad, Anders Thunem, Øystein Solberg, Hans Petter Eide, Håvard Holvik, Ole Georg Pettersen, Lars Flem og Håvard Tronhus.

Kenneth Devik

Innholdsfortegnelse

1 Innledning 1

1.1 Problemstilling 2

1.2 Disposisjon 3

1.3 Begrepsavklaringer 4

2 MMI Teori 7

2.1 Menneske-maskin interaksjon 7

2.2 Brukbarhet 8

2.2.1 Definisjon av brukbarhet, IEC/ISO 9241-11 9

2.3 Brukbarhets heuristikker 9

2.4 Brukergrupper 11

2.5 Brukbarhetstesting av mobile enheter 12

2.6 Brukbarhetstesting i forskjellig kontekst 13

2.6.1 Brukbarhetstesting i felten 13

2.6.2 Brukbarhetstesting i laboratorium 14

3 Handel og håndterminalløsningen 17

3.1 Forskning på brukbarhet innen handelsbransjen 17

3.2 Testproduktet 18

3.2.1 Håndterminalen 18

3.2.2 Lindbak POS Mobile 19

4 Evalueringsmetoder 23

4.1 Empiriske metoder 23

4.1.1 Observasjon 23

4.1.2 Intervju 24

4.1.3 Spørreskjema 24

4.1.4 Brukbarhetstesting 25

4.1.5 Logging 29

4.2 Analytisk evalueringsmetoder av brukbarhet 30

4.2.1 Inspeksjon (review) 30

4.2.2 Konsistenssjekk 30

4.2.3 Guidelines sjekk 30

4.2.4 Heuristisk evaluering 31

4.2.5 Kognitiv gjennomgang 31
4.3 Valg av metode 31
4.3.1 Kvantitative- og kvalitative metoder 31
4.3.2 Brukbarhetstest 32
4.3.3 Spørreskjema 32
4.3.4 Intervju 33
4.3.5 Observasjon 33
4.3.6 Logging 33
5 Forskningsmetoder og tidligere metodesammenligninger 35
5.1 Tidligere metodesammenligninger 35
5.1.1 Is it woth the hassle? Exploring the added value of evaluation the usability of context-aware mobile systems in the field 35
5.1.2 It's worth the hassle! The added value of evaluating the usability of mobile systems in the field 36
5.1.3 Usability testing of mobile applications: A comparison between laboratory and field testing. 37
5.1.4 Usability testing of mobile devices: A comparison of three approaches 37
5.1.5 Usability evaluation for mobile device: A comparison of laboratory and field tests 38
5.1.6 Creating realistic laboratory setting: Comparative studies of three think-aloud usability evaluations of a mobile system 39
5.2 Resultatoppsummering av tidligere metodesammenligninger 39
5.2.1 Problemfordeling 39
5.2.2 Alvorlighetsgraden til brukbarhetsproblemene 40
5.2.3 Overlappende problemer 41
5.3 Forskningsmetoder for metodesammenligning 42
5.3.1 Anvendbarhet 43
5.3.2 Effektivitet 43
5.3.3 Brukertilfredshet 43
5.3.4 Mental belastning 44
5.3.5 Brukeradferd 44
5.3.6 Ressursbruk 44
5.3.7 Oppsummering 44
5.4 Valg av sammenligningsmetode 45

5.5 Evalueringssammenligninger 45
6 Gjennomføring 47
6.1 Rekruttering 47
6.1.1 Anskaffelse av testbutikk 47
6.1.2 Rekruttering av testpersoner til laboratorietestene 47
6.2 Oppgaver 48
6.2.1 Oppgavelaging 48
6.2.2 Utvelgelse av oppgaver 48
6.2.3 Problemer med oppgavene 49
6.3 Gjennomføring av hver enkelt test 49
6.3.1 Opplæringen av den enkelte testperson 50
6.3.2 Pretest spørreskjema 50
6.3.3 Brukbarhetstesten 50
6.3.4 Intervjuet etter endt brukbarhetstest 50
6.4 Brukbarhetstest i felten 50
6.4.1 Oppsett av lokasjon 51
6.4.2 Gjennomføring av brukbarhetstester 52
6.4.3 Problemer og utfordringer 53
6.4.4 Observasjon av daglig bruk av håndterminaler 56
6.5 Fullskala brukbarhetstest i laboratorium 56
6.5.1 Oppsett av lokasjon 57
6.5.2 Pilottest 59
6.5.3 Problemer og utfordringer 59
6.6 Desktop brukbarhetstest i laboratorium 59
6.6.1 Oppsett 59
6.6.2 Problemer og utfordringer 61
6.7 Bearbeiding av data 62
7 Resultater 63
7.1 Testpersoner 63
7.1.1 Testpersonene i felttesten i butikk 63
7.1.2 Testpersonene i fullskala brukbarhetslaboratorium 64
7.1.3 Testpersonene i desktoptest i brukbarhetslaboratorium 65
7.1.4 Oppsummering av testpersonene 66

7.2 Interaksjonen med håndterminalen 67
7.3 Anvendbarhet 68
7.3.1 Feltresultater 69
7.3.2 Fullskalaresultater 70
7.3.3 Desktopresultater 71
7.3.4 Fullførte oppgaver og sammenbrudd 71
7.4 Effektivitet 73
7.5 Tilfredshet 74
7.6 Brukbarhetsproblemer 76
7.6.1 Gradering av alvorlighet 76
7.6.2 Kategorisering av problemene 76
7.6.3 Tabell med brukbarhetsproblemer 77
7.6.4 Oversikt som viser hvor de forskjellige feilene oppstod 85
7.6.5 Oversikt som viser de forskjellige problemtypene 88
8 Analyse 91
8.1 Analysering av brukbarhetsproblemer 91
8.1.1 Modal dialogboks var forvirrende 91
8.1.2 Språkproblemer laget utfordringer i felten 93
8.1.3 Metodeproblem grunnet lav realisme 94
8.1.4 Systemfeil fremprovosert av raske testpersoner 96
8.2 Sammenbrudd 97
8.3 Prioriterte brukbarhetsproblemer 97
8.4 Unike problem i metodene 99
8.4.1 Unike problemer i feltmetoden 99
8.4.2 Unike problemer i fullskalametoden 100
8.4.3 Unike problemer i desktopmetoden 100
8.4.4 Problemer i snittet mellom to metoder 101
8.4.5 Oppsummering 102
8.5 Tilfredsheten blant testpersonene 102
9 Diskusjon 105
9.1 Resultatkvalitet 105
9.1.1 Validitet 105
9.1.2 Reliabilitet 106

9.2 False positive resultater 107
9.3 Sammenlignet mot tidligere forskning 108
9.3.1 Anvendbarhet 108
9.3.2 Effektivitet 109
9.3.3 Tilfredshet 110
9.3.4 Hvordan påvirkes brukeren av konteksten i felten? 111
9.3.5 Hvordan påvirkes brukeren av konteksten i laboratoriet? 112
9.4 Problemstilling, forskningsspørsmål og hypoteser 113
9.4.1 Forskningsspørsmål og hypoteser 113
9.4.2 Problemstilling 116
9.5 Metodediskusjon 117
9.5.1 Observasjon 117
9.5.2 Intervju 117
9.5.3 Brukbarhetstesting 117
9.5.4 Spørreundersøkelse 117
9.5.5 Logging 117
9.6 Anbefalinger for brukbarhetstesting i handelsbransjen 118
9.7 Hva ville jeg gjort annerledes hvis jeg skulle gjort det om igjen? 119
10 Konklusjon 121
10.1 Konklusjon 121
10.2 Videre arbeid 122
11 Referanser 123
Vedlegg I. Oppgaver 127
Vedlegg II. Rekrutteringsplakat mot dagligvarebutikker 129
Vedlegg III. Rekrutteringsplakat 131
Vedlegg IV. Pretest spørreskjema 133
Vedlegg V. Sjekkliste for gjennomføring av brukbarhetstest 135
Vedlegg VI. Opplæring 137
Vedlegg VII. Intervjuveiledning 139

Figurliste

Figur 1 - En modell som viser attributtene for system akseptabilitet. Brukbarhetstesting er med andre ord ikke alt vi skal ta hensyn til (Nielsen 1993:25)..... 8

Figur 2 - Håndterminalen Symbol MC3090 som blir benyttet under testingen. 18

Figur 3 - Bak på håndterminalen er styluspennen festet 19

Figur 4 - Skjermbilde som viser ABC-tastaturet på skjermen. 19

Figur 5 - Skjermbilder fra Lindbak POS Mobile som viser menyflyten og sentrale funksjoner 21

Figur 6 - En graf som viser hvor mange brukbarhetsproblemer som avdekkes ved å teste x antall personer. Hentet fra Nielsen (2000) 27

Figur 7 - Minikamera for å filme interaksjonen med den mobile enheten. Hentet fra Kjeldskov og Skov et al. (2004)..... 35

Figur 8 - Viser antall problemer som ble oppdaget i de tre forskjellige metodene. Emulator = desktopemulator, document camera = desktopmobil, wireless camera = feltmobil. Fra Betiol og de Abreu Cybis (2005). 38

Figur 9 – Problemenes alvorlighetsgrad i kapittel 5.1.1..... 40

Figur 10 – Problemenes alvorlighetsgrad i kapittel 5.1.2..... 40

Figur 11 – Problemenes alvorlighetsgrad i kapittel 5.1.3..... 41

Figur 12 – Problemenes alvorlighetsgrad i kapittel 5.1.5..... 41

Figur 13 – Problemenes alvorlighetsgrad i kapittel 5.1.6..... 41

Figur 14 - Linjediagrammet viser hvor mange problemer som er funnet i alle metodene oppover, og antall testpersoner bortover. OBS, eksperimentelt diagram med lav validitet. 42

Figur 15 - Viser hvilke metoder som blir brukt for å sammenligne resultater 42

Figur 16 - Oppsummering av alle fire CUE. Hentet fra Molich og Dumas (2006) 46

Figur 17 - Bakrommet på lageret der alle håndterminalene til butikken er plassert til høyre, mens jeg fikk sette opp min test håndterminal på skrivebordet til venstre..... 51

Figur 18 - Skisse som viser dagligvarebutikken i felten. De oransje sirklene indikerer hvor de ulike oppgavene ble løst. 52

Figur 19 - Illustrasjonsbilde fra feltbutikken som viser hvordan brukbarhetstestingen ble gjennomført..... 53

Figur 20 - Skjermbilde fra den synkroniserte videostrømmen med loggen nederst og skjermopptaket oppe i høyre hjørne. 55

Figur 21 - Skisse over oppsettet for fullskalatest i laboratorium 57

Figur 22 - Illustrasjonsbilde som viser en butikkansatt som selger en vare med håndterminalen til en kunde. Til venstre i bilde er lagerdøren, mellom personene står det en kasse og til høyre har vi varehyllene. Fra venstre står Kenneth Devik og Terje Røsand. 58

Figur 23 - Kontrollrommet i brukbarhetslaboratoriet på NSEP. På bilde, Overingenør Terje Røsand. 58

Figur 24 - Skisse over desktoptestoppsett i brukbarhetslab. 60

Figur 25 - Skjermbilde fra videostreamen under en desktoptest. 60

Figur 26 - Illustrasjon av desktopmetoden. Nederst på bilde er lappene som skal forestille varer, mens på den andre siden er sleden håndterminalen dokkes i. 61

Figur 27 - Kakediagram med interaksjonstypebruk i fullskala laboratorium 67

Figur 28 - Kakediagram med interaksjonstypebruk i felt 67

Figur 29 - Kakediagram med interaksjontypebruk i desktoptest 67

Figur 30 - Viser hvor mange som fullførte de forskjellige oppgavene i de ulike metodene. 72

Figur 31 - Stolpediagram som viser hvor lang tid i sekunder testpersonene i de tre metodene brukte i snitt på oppgavene, hvor minimum 3 av 5 har klart oppgaven. 73

Figur 32 - En skala fra 0-100 som viser graden av tilfredshet. 0 er dårligst og 100 er best. 74

Figur 33 - Tavle med post-it-lapper over hvor problemene oppstod. Tavlen ble benyttet i innledende fase av problemanalyseringen. 86

Figur 34 - Venndiagram som viser hvor problemene oppstod 87

Figur 35 - Venndiagram som viser hvor de forskjellige problemtypene oppstod 88

Figur 36 - Skjermbilde av en popup-dialog som ingen hadde noen problemer med. 91

Figur 37 - Skjermbilde av en popup-dialog som flere testpersoner slet med å lukke. 91

Figur 38 - UAC i Windows Vista 92

Figur 39 - Skjermbilde som viser synkroniseringsfeilmeldingen TP får når en er offline og synkroniseringsknappen trykkes 93

Figur 40 - Lappene som ble brukt som varer og skapte en del forvirring i desktoptesten. 94

Figur 41 - Illustrerer desktopmetoden fra testpersonen sin synsvinkel. 95

Figur 42 - Viser hva slags problemer Lindbak anser som viktig å fikse. De uviktige er markert som svake bobler. 98

Figur 43 - Viser alle tre metodene. Fra venstre er felt-, fullskala- og desktopmetoden. 116

Figur 44 - Anbefalt testoppsett med tilstrekkelig realisme. 119

Tabelliste

Tabell 1 - Oversikt over de kvantitative- og kvalitative forskningsmetodene som blir benyttet i dette studiet 32

Tabell 2 - Antall brukbarhetsproblemer som ble oppdaget i de forskjellige metodene. Viser også antall overlappende problemer mellom metoder, og antall unike problemer for hver metode. . 40

Tabell 3 - Tabell med bakgrunnsinformasjon om testpersonene i felttesten 64

Tabell 4 - Tabell med bakgrunnsinformasjon om testpersonene i fullskalatesten 65

Tabell 5 - Tabell med bakgrunnsinformasjon om testpersonene i desktoptesten 65

Tabell 6 – Anvendbarhetsresultater fra feltmetoden. 69

Tabell 7 – Anvendbarhetsresultater fra fullskalametoden. 70

Tabell 8 – Anvendbarhetsresultater fra desktopmetoden. 71

Tabell 9 - Gir en oversikt over hvilke testpersoner som hadde sammenbrudd på de forskjellige oppgavene. 72

Tabell 10 - Tabell som viser tidsbruker i sekunder for hver oppgave per testperson. De tidene som er markert med rød tekst illustrerer sammenbrudd 74

Tabell 11 - SUS resultatet for hver enkelt testperson 75

Tabell 12 - SUS tabell med gjennomsnittet pr spørsmål i de ulike kontekstene og totaloppsummeringsverdier 75

Tabell 13 - Brukbarhetsproblemene som ble oppdaget 85

Tabell 14 - Tabell som viser hvor mange hvor mange problemer og alvorlighetsgraden til disse i de forskjellige kontekstene 87

Tabell 15 - Viser hvor mange av de forskjellige problemtypene som oppstod i de forskjellige kontekstene 89

Tabell 16 – Kategorisering av språkproblemene 93

Tabell 17 - Viser i hvor stor grad testperson 11 ikke benyttet varelappene og ikke valgte å skanne 94

Tabell 18 - Viser i hvor stor grad testperson 12 ikke benyttet varelappene og ikke valgte å skanne 94

Tabell 19 - Viser i hvor stor grad testperson 13 ikke benyttet varelappene og ikke valgte å skanne 95

Tabell 20 - Viser i hvor stor grad testperson 14 ikke benyttet varelappene og ikke valgte å skanne 95

Tabell 21 - Viser i hvor stor grad testperson 15 ikke benyttet varelappene og ikke valgte å skanne 95

Tabell 22 - Antall brukbarhetsproblemer som ble oppdaget i de forskjellige metodene. Denne tabellen stammer fra tabell 2 og er utvidet med resultatene fra dette studiet. Se tabell 2 for en forklaring på stjernene. 108

1 Innledning

I 1981 ble Xerox Star med sitt grafiske brukergrensesnitt sluppet på markedet. Xerox Star var det første systemet i verden som ble brukbarhetstestet under utviklingen. De satset bevisst på brukervennlighet for å selge mest mulig datamaskiner. Siden den gang har feltet menneske-maskin interaksjon (MMI) vokst seg stort og blitt et viktig område innen informatikk. Brukbarhetstesting har tradisjonelt sett foregått foran en stasjonær datamaskin i et laboratorium som en enkel 'desktoptest'. Når de mobile enhetene gjorde sitt inntog på markedet, var det bekymringer for at en desktoptest i laboratoriet ikke var tilstrekkelig for å simulere mobile enheters kontekstuelle omgivelser. De innså også at datainnsamling i en naturlig setting ville være veldig vanskelig (Kjeldskov, Skov et al. 2004).

I det siste tiåret har det blitt gjort en del studier som sammenligner brukbarheten til mobile enheter i felten og i laboratorium. Resultatene fra disse studiene er litt sprikende i forhold hvilken ekstra verdi en felttest gir, og om det er verdt strevet å gjennomføre en felttest. Vi skal i dette studiet se nærmere på det ved å brukbarhetsteste en mobil butikkløsning i felten og i et laboratorium både med og uten simulert kontekst. En slik sammenligning av tre metoder er det få som har gjort tidligere. Det har vært mest vanlig å bare benytte en laboratoriemetode til å sammenligne resultatene mot en felttest. Vi skiller de to metodene i laboratoriet ved å kalle den med simulert kontekst og flere realistiske momenter fra felten for fullskalatest. Den andre metoden uten simulert kontekst er enkel med lite realisme og kalles i det videre for desktoptest.

Håndterminalløsningen vi skal brukbarhetsteste i denne studien er Lindbak POS Mobile. Det er et program til bruk av butikkansatte i handelsbransjen, og kan foreta operasjoner som for eksempel varetelling, varebestilling og varemottak. I motsetning til mange andre system som blir brukbarhetstestet er Lindbak POS Mobile et ferdig produkt og ikke en prototype. Innen handelsbransjen er det foretatt få brukbarhetsstudier, og butikkens kontekst er derfor ikke særlig utprøvd.

Bakgrunnen for dette studiet er en deltidsjobb undertegnede har hatt hos Lindbak. Han har hatt hovedansvaret for uviklingen av håndterminalløsningen Lindbak POS Mobile. I utgangspunktet ble det formet en masteroppgave hvor en pilotbutikk for Lindbak POS Mobile skulle observeres, og funnene skulle sammenlignes mot en regissert brukbarhetstest i felten. Tiden gikk uten at pilotbutikken ble satt i gang, og derfor ble dagens oppgave en realitet i slutten av 2008.

1.1 Problemstilling

Vi ønsker å se nærmere på hva slags resultater de tre metodene gir. Det er ofte vanskelig og tidkrevende å gjennomføre realistiske studier. Hvis det viser seg at en felttest avdekker noe som ikke kan identifiseres i et laboratorium kan det være helt nødvendig å brukbarhetsteste i felten. Hvis ikke vil vi se på hvilken grad av realisme som er tilstrekkelig for å kunne avdekke alle problemer i et laboratorium.

Problemstillingen vi stiller oss i dette studiet er: Hvilken av de tre metodene er best egnet til evaluering av mobile butikkløsninger? Til å underbygge denne problemstillingen har vi laget noen forskningsspørsmål med hypotese:

A) Blir samme problemer og fenomener funnet i alle tre metodene?
En mer realistisk kontekst vil trolig kunne bidra til oppdagelsen av noen problemer som ikke er mulig å finne i metodene med mindre realisme.

B) Er alvorlighetsgraden til problemene forskjellige i de tre metodene?
De to mer realistiske metodene vil kunne avdekke flere alvorlige problemer på grunn av en mer realistisk kontekst, mens desktopmetoden vil finne flere kosmetiske problemer.

C) Vil det være forskjell på problemtypene som oppdages i de forskjellige metodene?
Desktopmetoden vil i større grad generere brukbarhetsproblemer knyttet til design og interaksjon med systemet.

D) Er det forskjeller på fullføringsgraden av oppgavene i de forskjellige metodene?
Mer realisme i form av større kontekst vil gjøre oppgavene vanskeligere. Løsningen på oppgavene vil ikke nødvendigvis stå rett foran nesen på testpersonene, noe den i praksis vil gjøre i desktopmetoden.

E) Vil interaksjonen med håndterminalen være forskjellig ved bruk av de tre metodene?
Det er mer naturlig å benytte fingrene til å trykke på den trykkfølsomme skjermen i felten. De som sitter på en stol og blir testet i desktopmetoden vil på samme måte bruke pennen mer aktivt.

F) Hvilke fordeler og ulemper har de tre metodene sett opp imot hverandre?
Det er vanskeligere gjennomføre en felttest i forhold til laboratorietest.

G) Er det verdt det ekstra strevet å gjennomføre en felttest?
Sett opp imot tidligere forskning finnes det mye varierende resultat. Det kan derfor virke som at mange faktorer i stor grad kan påvirke resultatene. Dette kan for eksempel være selve systemet, testpersonene, kontekstene og scenarioene. Derfor er det usikkert hvordan det vil fortone seg med en metodesammenligning innen handelsbransjen.

1.2 Disposisjon

Oppbygningen av denne masteroppgaven blir presentert her, og vi forteller litt om hva som kan forventes under de ulike kapitlene.

2 - MMI teori
Vi vil presentere relevant brukbarhetsteori i dette kapitlet. Fokuset vil være på brukbarhets-testing av mobile enheter og de ulike kontekstene de kan brukbarhetstestes i. Metodeteori blir nevnt i kapittel 4 og 5.

3 - Handel og håndterminalløsningen
I den første delen går vi inn på tidligere brukbarhetsstudier gjort innen handelsbransjen. I del to vil vi beskrive håndterminalløsningen som skal benyttes i dette studiet, slik at det skal være mulig å forstå de forskjellige brukbarhetsproblemene som blir avdekket.

4 - Evalueringsmetoder
Her ser vi på tradisjonelle empiriske evalueringsmetoder som intervju, spørreskjema, brukbarhetstesting osv. Vi har delt kapittelet inn i en teoridel som presenterer evalueringsmetodene og en del der vi snakker om hvordan og hvorfor vi benytter de ulike evalueringsmetodene.

5 - Forskningsmetoder
Først vil vi oppsummere seks forskningsstudier som benytter mobile enheter og sammenligner resultatene fra brukbarhetstesting i felt og laboratorium. Disse studiene vil vi benytte for å redegjøre for hvordan resultatene fra brukbarhetstesting i ulik kontekst sammenlignes. Vi vil si litt om hva slags sammenligningsmetoder vi velger å bruke i dette studiet. Til slutt ser vi på hvorfor evalueringer på tvers av ulike forskningsteam kan være problematisk.

6 - Gjennomføring
Vi forteller om den praktiske gjennomføringen av studiet i dette kapittelet. Vi sier litt om utfordringene med å rekruttere feltbutikk og testpersoner. Hvordan oppgavene som testpersonene skulle løse ble laget. Oppsettet av de forskjellige kontekstene sier vi noe om og utfordringene knyttet til dem.

7 - Resultater

I resultatkapitlet oppsummerer vi de ulike testpersonene i studiet og fremstiller bearbeidingen av data fra brukbarhetstestingen.

8 - Analyse

Under analysen går vi nærmere inn på noen interessante brukbarhetsproblemer som dukket opp. Vi ser litt på hvor de ulike problemene dukket opp, og vi ser på hva slags problemer vi oppdaget som systemleverandøren prioriterer å fikse.

9 - Diskusjon

Det første vi gjør i diskusjon er å se på kvaliteten til resultatene våre i henhold til validitet og reliabilitet. Vi sammenligner også våre resultater mot tidligere metodesammenligninger. Problemstillingen og alle forskningsspørsmålene vil bli besvart, og vi kommer med en anbefaling om hvordan mobile enheter for handelsbransjen bør brukbarhetstestes.

1.3 Begrepsavklaringer

I oppgaven benyttes en del ord og begreper som ikke nødvendigvis er allmennkunnskap, og disse forklares her:

Begrep	Forklaring
POS	Engelsk forkortelse for Point Of Sale, og kan oversettes med et kasseapparat.
Lindbak POS	Moderløsningen som Lindbak POS Mobile jobber mot, og er en fullblods kasseapparatløsning.
Lindbak POS Mobile	En mobil kasseløsning og selve programmet som blir benyttet i testingen. Det kjører på mobile enheter i det videre kalt håndterminaler.
Bong	Fagbegrep i Lindbak POS Mobile (kapittel 3.2.2).
Slede	Dokkingstasjonen håndterminalen settes i når databasen skal synkroniseres. Se kapittel 3.2.1 og figur 2.
SMS-tastatur	Fysiske knapper fra 1 til 9 som kan benytte for å skrive bokstaver, ála tastaturet vi har på mobiltelefoner.
ABC-tastatur	Et fult qwerty-tastatur på den trykkfølsomme skjermen.
TP	Forkortelse for testperson, altså en av deltakerne som ble brukbarhetstestet.
NSEP	Norsk Senter for ElektroniskPasientjournal ved St. Olavs Hospital, NTNU.
MMI	Menneske-maskin interaksjon (kapittel 2.1).
SUS	System Usability Scale. Spørreskjema for å måle tilfredshet. (Brooke 1996).

Begrep	Forklaring
CIF	Common Industry Format (kapittel 4.1.4).
CUE	Comparative usability evaluation (kapittel 5.5).
Fullskala-test/metode	En brukbarhetstest i laboratorium med mye realisme skaffet fra felten (kapittel 2.6.2).
Desktop-test/metode	En enkel brukbarhetstest i laboratorium uten noe særlig realisme (kapittel 2.6.2).
Desktopmobil-test/metode	Et begrep for en metode som blir benyttet av Betiol og de Abreu Cybis (2005). Den går ut på at de har en enkel brukbarhetstest i laboratorium hvor de benytter en mobil (kapittel 5.1.4).
Desktopemulator-test/metode	Et begrep for en metode som blir benyttet av Betiol og de Abreu Cybis (2005). Den går ut på bruken av en mobilemulator som kjører på en datamaskin(kapittel 5.1.4).

2 MMI Teori

I dette menneske-maskin interaksjon(MMI) teorikapitlet vil relevant brukbarhetsteori bli presentert. Det vil bli gått nærmere inn på målbare kritterier for brukbarhet som ISO-definisjon, og hvordan ulike brukergrupper kan påvirke en brukbarhetstest. Utfordringene vi har med tanke på brukbarheten til mobile enheter vil bli nevnt, og de forskjellige kontekstene vi skal brukbarhetsteste i blir forklart. Brukbarhetstesting som metode blir nevnt i dette kapitlet, men blir utdypet nærmere i evalueringsmetoder under delkapittel 4.1.4.

2.1 Menneske-maskin interaksjon

Menneske-maskin interaksjon(MMI) er studien som omhandler brukernes samhandling med datamaskiner. Fokuset til MMI er å forbedre kvaliteten på interaksjonen mellom bruker og datamaskin. MMI er det egentlig en disiplin på tvers av vitenskapsgrener. Det kommer fra spesialfelt som angår flere disipliner, hver med forskjellig vekt: informatikk (design av applikasjoner og konstruksjon av brukergrensesnitt), psykologi (teorier om kognitive prosesser og emirisk analyse av brukeradferd), sosiologi og antropologi (interaksjon mellom teknologi, arbeid og organisasjon). (Hewett, Baecker et al. 1996)

> *Human-computer interaction is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them.* (Hewett, Baecker et al. 1996)

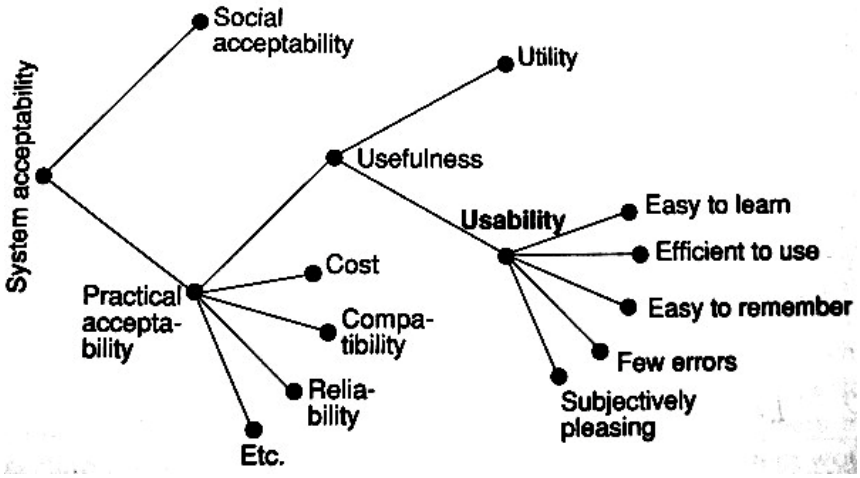
Det er flere personligheter innen MMI miljøet som har definert generelle prinsipper for en MMI tilnærming i design av programvare. Disse prinsippene er både teknologi og enhetsuavhengig, og kan derfor benyttes på datamaskiner uavhengig av operativsystem eller for eksempel til å designe programvare for mobile enheter (Blum and Khakzar 2007). Under kommer en liste med forskjellige designprinsipper innen MMI:

- Simpson (1985): ”*Design of user-friendly programs for small computers*”
- Shneiderman (1992): ”*Designing the user interface strategies for effective human-computer interaction*”
- Dumas (1988): ”*Designing user interfaces for software*”
- Nielsen (2001): “Ten usability heuristics”
- Norman (2002): “*The design of everyday things*”
- Chang (2002): “*Gestalt theory in visual screen design*”

Begrepet brukbarhet er sentralt innen MMI, og kort fortalt handler om hvor brukbart et datasystem er. Brukbarhet vil bli forklart nærmere i kapittel 2.2. Til å forbedre brukbarheten til et system finnes det flere fremgangsmåter. Den mest kjente fremgangsmåten er en brukbarhetstest og er den metoden som benyttes i dette studiet. Det finnes også andre evalueringsmetoder for brukbarhet, som for eksempel heuristisk evaluering, kognitiv gjennomgang, konsistenssjekk og guidelines sjekk. Alle disse evalueringsmetodene vil bli presentert nærmere i kapittel 1.

2.2 Brukbarhet

Brukbarhet(usability) innen informatikk handler om nytteverdien og hvor enkelt et system er for brukeren. Det er mange som har definert brukbarhet på sin måte som Nielsen, Shneiderman og Molich med flere. Felles for dem alle er at de definerer brukbarhet ganske likt. Figur 1 viser en graf fra Nielsen (1993) som plasserer brukbarhet i en enkel system akseptabilitets modell. Under usefulness i grafen er utility og usability. Utility går ut på om funksjonaliteten til et system kan gjøre det som brukeren trenger, mens usability på den andre siden sier noe om hvor brukervennlig funksjonaliteten er.



Figur 1 - En modell som viser attributtene for system akseptabilitet. Brukbarhetstesting er med andre ord ikke alt vi skal ta hensyn til (Nielsen 1993:25)

Punktene under brukbarhet i figur 1 til Nielsen (1993:24-26) oppsummeres slik med hans definisjon:

- **Lærbarhet**: Systemet bør være lett å lære slik at brukeren raskt kan bli produktiv.
- **Effektivitet**: Systemet bør være effektivt å bruke, slik at når brukeren har lært seg systemet, kan brukeren bli enda mer produktiv.

- **Huskbarhet**: Systemet bør være enkelt å huske, slik at en vanlig bruker kan begynne å bruke systemet igjen etter en lengre tids pause uten å måtte lære seg alt på nytt igjen.
- **Problemer**: Systemet bør ha en lav feilrate, sånn at brukerne gjør få feil når de bruker systemet. Hvis de gjør en feil, bør systemet gjøre det lett å komme seg ut av feilen. Katastrofefeil må ikke opptre.
- **Tilfredshet**: Systemet bør være behagelig å bruke, slik at brukerne er subjektivt tilfreds når de bruker det og at de liker det.

2.2.1 Definisjon av brukbarhet, IEC/ISO 9241-11

For å få en veletablert definisjon av brukbarhet har brukbarhet blitt standardisert i ISO9241-11. Brukbarhet defineres av ISO9241-11 på følgende måte:

> *The extent of which a product can be used by specified users to achieve specified goals within effectiveness, efficiency and satisfaction in a specified context of use. (ISO 1997)*

I tillegg til å definere brukbarhet, forklarer også ISO9241-11 hvordan vi finner informasjonen som er nødvendig å ta hensyn til når vi evaluerer brukbarheten til et system med tanke på brukerytelsen og brukertilfredsheten. ISO9241-11 gir også retningslinjer for hvordan brukbarheten til et produkt kan bli spesifiser og evaluert. Nedenfor vil de tre evalueringsmålene for brukbarhet oppsummeres:

- **Anvendbarhet** (effectiveness): Sier noe om nytteverdien og hvor anvendbart et system er, slik at brukeren kan oppnå sine mål.
- **Effektivitet** (efficiency): Sier noe om hvor mye ressurser en bruker trenger for å oppnå sine mål.
- **Tilfredshet** (satisfaction): Sier noe om brukerens subjektive reaksjoner og holdninger til bruken av systemet.

Anvendbarhet måles vanligvis ved å se hvor mange prosent en bruker klarer å løse av en gitt oppgave, telle antall problemer, telle antall ganger brukeren har fått hjelp og telle antall ganger brukeren har brukt dokumentasjonen til hjelp. *Effektiviteten* måles gjerne ved å ta tiden en bruker trenger på å fullføre en oppgave. Brukerens *tilfredshet* blir vanligvis målt ved hjelp av et spørreskjema som måler lettheten til systemet og nytteverdien. (Theofanos 2006)

2.3 Brukbarhets heuristikker

Jacob Nielsen og Rolf Molich definerte heuristikkene for første gang i 1990 og Nielsen endret endret disse litt i 2001 til den listen nedenfor. Listen er basert på en gjennomgang av 249 brukbarhetsproblemer som er gruppert sammen etter hva slags type problem det er snakk om.

Brukbarhetsheuristikkene kan brukes til å gjennomføre en heuristisk evaluering. En slik heuristisk evaluering er en evalueringmetode på linje med brukbarhetstesting innen feltet menneske-maskin interaksjon. Mer om heuristisk evaluering og andre evalueringsmetoder innen brukbarhet står under i kapittel 4.2.

1. **Synliggjør statusen til systemet**
   Systemet bør alltid holde brukeren informert om hva som skjer gjennom passende tilbakemelding innen rimelig tid.
2. **Systemet bør tilsvare den virkelige verden**
   Systemet bør snakke brukeren sitt språk med ord, fraser og konsepter som er kjent for brukeren i stedet for systemorienterte termer. Følg konversasjoner fra den virkelige verden og få informasjonen til å fremstå naturlig og logisk ordnet.
3. **Brukerkontroll og frihet**
   Brukere velger ofte systemfunksjoner ved en feil og vil trenge en klart markert "nødutgang" for å komme seg ut av uønsket tilstand, uten å gå gjennom unødvendig mange steg. Støtte funksjoner som å angre og gjøre om igjen.
4. **Konsistens og standarder**
   Brukere bør ikke trenge å lure på om forskjellige ord, situasjoner eller handlinger betyr det samme. Følg plattformens konversasjon.
5. **Forebygg feil**
   Det som er bedre enn gode feilmeldinger er et omhyggelig designet system som forebygger feil fra å opptre. Enten eliminere feilutsatte tilstander eller presenter brukeren med en bekreftelsesdialog før vedkommende skal til å gjøre en slik handling.
6. **Gjenkjenn heller enn å tenke tilbake(recall)**
   Minimer det brukeren må huske ved å lage synlige objekter, handlinger og opsjoner. Brukeren bør ikke trenge å huske informasjon fra en dialog til en annen. Instruksjoner for å bruke systemet bør være synlige og enkelt gjenopprettelig når det er hensiktsmessig.
7. **Fleksibilitet og effektiv bruk**
   Akseleratorer(for eksempel tastatursnarveier) som ikke kan ses av en nybegynner kan ofte øke farten på interaksjonen for en ekspertbruker sånn at systemet imøtekommer både erfarne og uerfarne brukere.
8. **Estetisk og minimalistisk design**
   Dialoger bør ikke inneholde informasjon som er irrelevant eller sjeldent brukt. Hver ekstra informasjon i dialogen konkurrerer med den relevante informasjonen og reduserer deres relative synlighet.
9. **Hjelp brukeren til å kjenne igjen, diagnostisere og gjenvinne kontrollen etter feil**
   Feilmeldinger bør uttrykkes i enkelt språk (ingen koder), presist indikere problemet og strukturert foreslå en løsning.

10. **Hjelp og dokumentasjon**
    Selv om det er bedre hvis et system kan brukes uten dokumentasjon, kan det være nødvendig å tilby hjelp og dokumentasjon. All slik informasjon bør være enkel å søke i, fokusert på brukerens oppgaver og liste opp konkrete steg som kan utføres og må ikke være for stor.

2.4 Brukergrupper

Når et nytt system skal designes foretar vi en analyse for å sikre systemets nytteverdi, som er en viktig del av brukbarhet. Det er viktig å ha grundig kjennskap til hvem som skal bruke et system når vi designer det. En slik analyse gjør vi ved å finne ut hvem brukerne til systemet er. Disse brukerne blir gjerne delt inn i grupper på bakgrunn av deres karakteristikker, og kalles brukergrupper. Hver brukergruppe beskrives kort, og vi lager gjerne en personas som er en brukerkarakteristikk for brukergruppen. En personas er en realistisk beskrivelse av en fiktiv person i brukergruppen. (Molich 2003:40-47)

> *For det meste er der flere målgrupper med ret forskjellige interesser og forudsætninger. I nogle tilfælde har målgrupperne modstridende forudsætninger, for eksempel nybegyndere og øvede brugere, eller amatører og professionelle brugere.* (Molich 2003)

Det er mange faktorer som kan være med å forme de ulike brukergruppene vi har for et system. Nedenfor vil en rekke slike faktorer fra en forelesningsfoil i faget TDT4180 (2006) ved NTNU bli presentert:

Personlig historie:
- Alder, kjønn, holdninger til typen av produkt, ”hendthet”, læringsstil, holdning til teknologi

Utdanningshistorie:
- Høyest oppnådde grad, Fag

Erfaring med IKT:
- (hvor lenge, hvor ofte, hvilken type operativsystem)

Produkterfaring:
- Tid brukt på produktet, frekvens, hvilken type oppgaver, hvilken type (bruker de "ditt" produkt)?

Jobbhistorie:
- Nåtidig og tidligere jobb, ansvar for hva, hvor lenge i nåtidig stilling etc.

Ved gjennomføring av brukbarhetstester på et system bør vi helst teste alle definerte brukergrupper for systemet. I praksis blir det som regel ikke testet flere brukergrupper enn de primære målgruppene. Det kan i mange tilfeller være ressurser til kun å teste en brukergruppe.

> *Testlederen bør sikre, at han får et repræsentativt udvalg af typiske testdeltagere, hva angår alder, erfaring og motivation.* (Molich 2003)

Enhver brukergruppe bør bestå av minst fire testpersoner sier Molich (2003), men han anbefaler seks. Færre testpersoner enn fire mener han at kan føre til at viktige problemer blir oversett. Mer om antall testpersoner står under kapittel 4.1.4.

Forskjellige brukergrupper vil trolig finne mange like problemer, men det er også mulig at disse brukergruppene finner noen unike problemer som kan relateres direkte til brukergruppen. Dersom det er forskjellige brukergrupper i dette studiet, kan det potensielt utgjøre en fare med tanke på sammenligning av resultatene fra de tre metodene.

2.5 Brukbarhetstesting av mobile enheter

Mobile enheter skiller seg fra stasjonære datamaskiner på flere måter. En mobil enhet har økt mobilitet, ustabilt nettverk, begrenset minne og prosessorkraft, mindre skjermstørrelse og vanskeligere input-mekanismer. (Weiss 2002)

> *Mobile systems are typically used in highly dynamic contexts. Moreover, their use often involves several people distributed in the user's physical surroundings*. (Kjeldskov and Stage 2004)

Nettopp på grunn av at mobile enheter gjerne benyttes i dynamiske kontekster, kan det virke tiltalende og uunnværlig å brukbarhetsteste mobile systemer i felten sier Kjeldskov et al. (2004) på bakgrunn av tidligere forskning. Det er en del utfordringer knyttet til testing i felten, og disse utfordringene kan reduseres signifikant i et brukbarhetslaboratorium. Mer om forskjellig type kontekst vi kan brukbarhetsteste mobile enheter i omtales nærmere i neste kapittel.

En studie utført av Kjeldskov og Graham (2003) viser at 71 % av alle brukbarhetsevalueringer av mobile enheter blir gjort i et laboratorium, og få av disse brukbarhetslaboratoriene involverer spesielle teknikker for å møte utfordringene ved å teste mobile systemer.

> *A product can have significantly different levels of usability when used in different contexts.(ISO 1997)*

Til å bedre brukbarhetstestingen av mobile system har Kjeldskov et al. (2004) utført en studie hvor de tester ut seks nye teknikker for å evaluere brukbarheten i laboratorium. Målet har vært å finne teknikker som får det kontrollerte laboratoriemiljøet til å være mest mulig likt virkeligheten. De har fokusert på at brukeren går og forflytter seg mens mobile enheter benyttes(andre relevante faktorer som de ikke har tatt hensyn til er fysiske, sosiale og temporal kontekst for mobil bruk). Til å kunne sammenligne de seks teknikkene mot noe, gjennomførte de en feltstudie. Resultatene fra de seks ulike teknikkene viste at alle hadde likheter med felttesten, men at ingen var helt sammenlignbare med metoden i feltevalueringen. Flest brukbarhetsproblemer ble avslørt ved hjelp av den enkleste teknikken, som innebar at testpersonen satt ved et bord. Mange av problemene som ble oppdaget av den enkleste teknikken var kosmetiske.

2.6 Brukbarhetstesting i forskjellig kontekst

Brukbarhetstesting i forskjellig kontekst er gjerne et spørsmål om hva slags grad av realisme vi trenger for å finne de reelle brukbarhetsproblemene. Det kan være mange faktorer som spiller inn for hva slags kontekst det bør testes i for et gitt mobilt system. Nettopp det vises i gjennomgangen av tidligere metodesammenligninger i kapittel 5.1. Noen sier at vi trenger å teste i en så realistisk kontekst som mulig, mens andre sier at det ikke er verdt strevet fordi det beriker resultatet lite.

Innen brukbarhetstesting skiller vi tradisjonelt mellom felt- og laboratoriekontekst. En feltstudie karakteriseres ved at testingen skjer i "*den virkelige verden*". Det er en del utfordringer knyttet til feltstudier med tanke på at vi har mindre kontroll på variablene. Datainnsamlingen kan også være vanskeligere. Fordelen er den økte realismen og muligheten for å studere komplekse situerte interaksjoner og prosesser. (Kjeldskov and Graham 2003)

I kontrast til feltstudier, kjennetegnes laboratoriestudier ved at de skjer i et kontrollert miljø. Det må nevnes at vi trenger ikke et dedikert brukbarhetslaboratorium. En kan gjennomføre "*lab*"-eksperimenter i andre kontrollerte miljøer som for eksempel ett møterom. Fordeler med å brukbarhetsteste i et laboratorium er at fasilitetene for å samle inn data er gode. Ulemper med denne metoden er mindre grad av realisme. (Kjeldskov and Graham 2003)

Denne oppgavens fokus er å sammenligne resultatene mellom brukbarhetstesting i tre forskjellig kontekster, hvorav den ene er i felten mens de to andre er i laboratorium. I tidligere sammenligninger skilles det ikke alltid like godt mellom fullskala- og desktoptester. Mange av laboratorietestene som sammenlignes mot felten er egentlig desktoptester. Nedenfor vil de tre kontekstene det skal sammenlignes i bli presentert grundigere.

2.6.1 Brukbarhetstesting i felten

Ved å gjennomføre en brukbarhetstest i felten får vi testet det mobile systemet i en reell setting. Feltstudier blir vanligvis gjennomført for å finne ut hvordan et produkt blir brukt naturlig i sitt hverdagslige miljø. Realismen i felten kan være vanskelige å gjenskape noen annen plass. Det finnes en del utfordringer knyttet til brukbarhetstesting i felten, og tre av disse har Kjeldskov og Stage (2004) oppsummert i listen nedenfor:

1. Det kan være komplisert å skape realistiske studier som fanger nøkkelsituasjoner i dynamiske kontekster som ofte involverer flere folk distribuert i brukerens fysiske omgivelser.
2. Det er langt fra trivielt å anvende etablerte evalueringsteknikker som observasjon og høyttenkning når en evaluering blir utført i en felt setting.
3. Feltevalueringer kompliserer datainnsamlingen og begrenser kontrollen siden brukeren beveger seg i et fysisk miljø med ett antall ukjente variabler som potensielt kan påvirke testoppsettet.

2.6.2 Brukbarhetstesting i laboratorium

De tre utfordringene i felten kan reduseres signifikant ved brukbarhetstesting i et laboratorium. Det er ingen problemer knyttet til å samle inn data i et laboratorium, og det er heller ikke noen problemer med å benytte etablerte evalueringsteknikker. Ulempen er at en laboratoriesetting mangler realisme. Det kan også være vanskeligere å gjøre laboratoriet mer realistisk når vi tester mobile enheter enn ved testing av en stasjonær datamaskin, fordi en mobil enhet kan brukes veldig mobilt og dynamisk, og de fysiske omgivelsene kan være vanskelig å gjenskape. (Kjeldskov and Stage 2004)

> *The laboratory environment is a peaceful space, where a test user can concentrate on the given task.* (Kallio and Kaikkonen 2005)

I kontrast til en felttest er brukbarhetsforskere og praktikere bekymret for at evalueringer i laboratoriet ikke klarer å simulere konteksten til mobile enheter. Faktorer som bråk, multitasking, bevegelse, avbrytelser etc. kan være med å påvirke blant annet effektiviteten i felten. De faktorene er i mindre grad eller ikke eksisterende i laboratorietester. (Kallio and Kaikkonen 2005)

Fullskala brukbarhetstest i laboratorium

En fullskala brukbarhetstest, eller simulert brukbarhetstest som det også kalles i litteraturen, kan benyttes for å gjøre laboratoriet mest mulig lik en reell setting. Vi prøver med andre ord å kombinere det beste fra brukbarhetstesting i felten og laboratoriet slik at vi får best mulige resultater.

> *It is widely acknowledged in the usability community that much can be gained from bringing aspects from the field into the lab.* (Nielsen 1998)

I en fullskalakontekst tas fysisk kontekst inn fra felten for å bøte på den manglende realismen, og vi lager dermed en mer realistisk arbeidssetting. Det kan i noen tilfeller være vanskelig å gjenskape relasjonen mellom systemet og aktivitetene i de fysiske omgivelsene, og systemer for veldig spesifikke domener kan også være utfordrende (Kjeldskov and Skov 2003).

Ett annet aspekt som bør overveies er hvor realistisk laboratoriet skal gjøres i en fullskalametode. Dahl, Alsos et al. (2009) sier at det er ingen "*one size fits all*" i sykehussammenheng, fordi det beror på situasjonen. De følger "*just enough*" prinsippet med tanke på hvor realistisk en setting skal være. De mener at brukbarhetseksperter og domeneeksperter sammen bør hente inn elementer fra den virkelige verden som ekspertene betrakter som aspekter som trolig vil påvirke oppfattelsen av brukbarheten til systemet. (Dahl, Alsos et al. 2009).

Desktop brukbarhetstest i laboratorium

Desktoptest er en enkel og billig metode for brukbarhetstesting. Der bruker vi ikke noe tid eller penger på å gjenskape realismen fra den virkelige verden. Testpersonene trenger ikke å forholde seg til noe annet enn å tenke høyt og gjennomføre en gitt oppgave.

3 Handel og håndterminalløsningen

Handelsbransjen setter pris på verktøy som er med på å forenkle og effektivisere hverdagen. Håndterminaler med strekkodelesere er et slikt verktøy og kan benyttes til for eksempel varebestilling, varemottak og svinnregistrering. De butikkansatte kjennetegnes gjerne med varierende bakgrunn, datakunnskaper og alder. Miljøet de jobber i består av mange visuelle artefakter, jevnlige avbrytelser og en del støy. På grunn av alle disse faktorene bør mobile butikkløsninger lages med hensyn til brukbarhet. I dette kapittelet vil det bli gått nærmere inn på tidligere brukbarhetsforsking innen butikkmiljøet og løsningen som testes i dette studiet vil bli presentert.

3.1 Forskning på brukbarhet innen handelsbransjen

Det har blitt foretatt lite brukbarhetstesting på mobile løsninger innen handelsbransjen. Vi har ikke funnet noen som har utført en lignende brukbarhetstesting med fokus på de butikkansatte. Den eneste vitenskaplige artikkelen som har brukbarhetstestet i en butikk er Newcomb et al. (2003), men de har fokuset på kunden i butikken. En annen artikkel fokuserer på bruk av håndterminaler i handelsbransjen er Blum og Khakzar (2007). De har laget 17 designretningslinjer for utvikling av programvare for håndterminaler. Det finnes derimot en god del artikler rundt allestedsnærværende (pervasive/ubiquitous) datamaskiner innen handelsbransjen, men det blir litt utenfor rekkevidden til denne studien.

I artikkelen ”Mobile computing in the retail area” av Newcomb et al. (2003) brukbarhetstester de en prototype for elektronisk vareliste i en dagligvarebutikk. Newcomb et al. (2003) lurer blant annet på hvordan kunden vil holde og bruke håndterminalen samtidig som de må bruke hendene når de skal plukke varer under handlingen. Når kunden er i en butikk sier de at vedkommende leter mye med øynene etter varer og plukker de opp. Selve butikkmiljøet har overveldende mye informasjon i form av visuell, følbar, hørbar og lukt sier de.

De testet fem personer i butikken og ba de tenke høyt mens de utførte oppgavene. Mange av disse personene var overrasket over hvor fort de handlet ved bruk av den elektroniske varelisten. Personene merket seg også at de ble mer fokusert på å skaffe varene på listen, enn å se seg rundt i butikken. Det hendte at de la fra seg håndterminalen når de ikke hadde bruk for den. Det hyppigste ønsket for personene var en dokkingstasjon for håndterminalen på handlevognen, for de mislikte å holde og bære på håndterminalene hele tiden mens de handlet. Alt bråket i en butikk førte også til at personene ikke la merke til et lydsignal håndterminalen ga fra seg.

Newcomb et al. (2003) sier at det var forskjellige utfordringer knyttet til brukbarhetstestingen i butikken. De kunne ikke benytte videokamera og var helt avhengig av lydopptak, visuelle observasjoner og feltnotater.

> *"Unlike a typical usability test, which is performed at a stationary desktop PC, actions on the interface were not easily monitored"* (Newcomb, Pashley et al. 2003).

3.2 Testproduktet

Håndterminalløsingen som ble benyttet i denne studien er Lindbak POS Mobile fra Lindbak retail systems. Produktet er funksjonelt komplett og er i pilotstadiet. Nettopp derfor er det velegnet for brukbarhetstesting fordi ingen er kjent med produktet fra før av. Dette kapittelet vil gi en innføring i hvordan Lindbak POS Mobile virker slik at det vil være mulig å forstå de fleste brukbarhetsproblemene som blir avdekket og andre aspekter.

3.2.1 Håndterminalen

Det ble benyttet en håndterminal av typen Symbol MC3090 i denne studien. Den har en trykkfølsom skjerm som kan betjenes av en finger eller en styluspenn festet på baksiden (se figur 3). På toppen har den et strekkodeleserhode som kan vris til sidene og aktiveres ved hjelp av en gul knapp rett nedenfor skjermen eller ved bruk av to knapper på begge sider av håndterminalen (se figur 2). Knappene består av tall fra 1-9, piltaster, backspace, space, enter og noen andre funksjoner. Ved aktivering av en oransje "alpha"-knapp blir tallknappene et SMS tastatur.

Figur 2 - Håndterminalen Symbol MC3090 som blir benyttet under testingen.

Håndterminalen kjører operativsystemet Microsoft Windows CE 5.0. Plastdingsen den står i på figur 2 blir i det videre kalt for en slede som håndterminalen dokkes i. Når den er dokket vil håndterminalen lade og det blir opprettet en kobling til datamaskinen sleden er koblet til via programmet Microsoft ActiveSync. Symbol MC3090 har også innebygd støtte for de trådløse teknologiene bluetooth og WLAN.

Det er to måter å skrive tekst på håndterminalen. Den ene blir kalt ABC-tastatur og er et på-skjermstastatur (figur 4). Den andre metoden går ut på å bruke et mobillignende tastatur ved hjelp av tallknappene 1-9 som vi ser i figur 2, og metoden blir i denne studien kalt SMS-tastatur. Et problem med SMS-tastaturet er at aplha-knappen må trykkes ned for å aktivere bokstavskriving, og det deaktiveres ved enda et trykk. Tilstanden til aplha-knappen vises ikke. Ved skriving av bokstaver med SMS-tastaturet, vises ikke det en skriver før x antall trykk på et tall er ferdigtrykket.

3.2.2 Lindbak POS Mobile

Lindbak POS Mobile er produktet som ble testet i denne studien og det kjører på håndterminalen i figur 2. Det er utviklet ved hjelp av Microsoft .NET Compact Framework 3.5. I stedet for å benytte medfølgende standard windows-komponenter er det laget et eget GUI-bibliotek med komponenter som ser bedre ut enn standardkomponentene.

Figur 3 - Bak på håndterminalen er styluspennen festet

Utviklingen av produktet har blitt gjort av to studenter som sommer- og deltidsjobb hvor undertegnede var med helt fra starten i 2007. Vi har ikke hatt noen bestemt kunde å forholde oss til. Det er ledelsen hos Lindbak som har vært interessenter på alle vis. Produktet er ment brukt hos flere av Lindbak sine butikkdatakunder. Lindbak har ikke tradisjon for å brukbarhetsteste produkter under utvikling, og derfor har heller ikke Lindbak POS Mobile blitt brukbarhetstestet tidligere.

Produktet har støtte for følgende funksjonalitet: Varetelling, varebestilling, varemottak, svinnregistrering, intern overføring, internt forbruk, etikettbestilling, priskontroll, prisendring, lageroppslag og varesalg. Lindbak POS Mobile benytter begrepet *bong* om alle disse funksjonene. Har du foretatt en varetelling, så blir det opprettet en varetellingsbong. Ved en fullført varebestilling blir det laget en varebestillingsbong. Alle disse bongene blir liggende i bongkøen (bongkø finnes under administrasjonsmenyen i figur 5), og overføres under synkronisering.

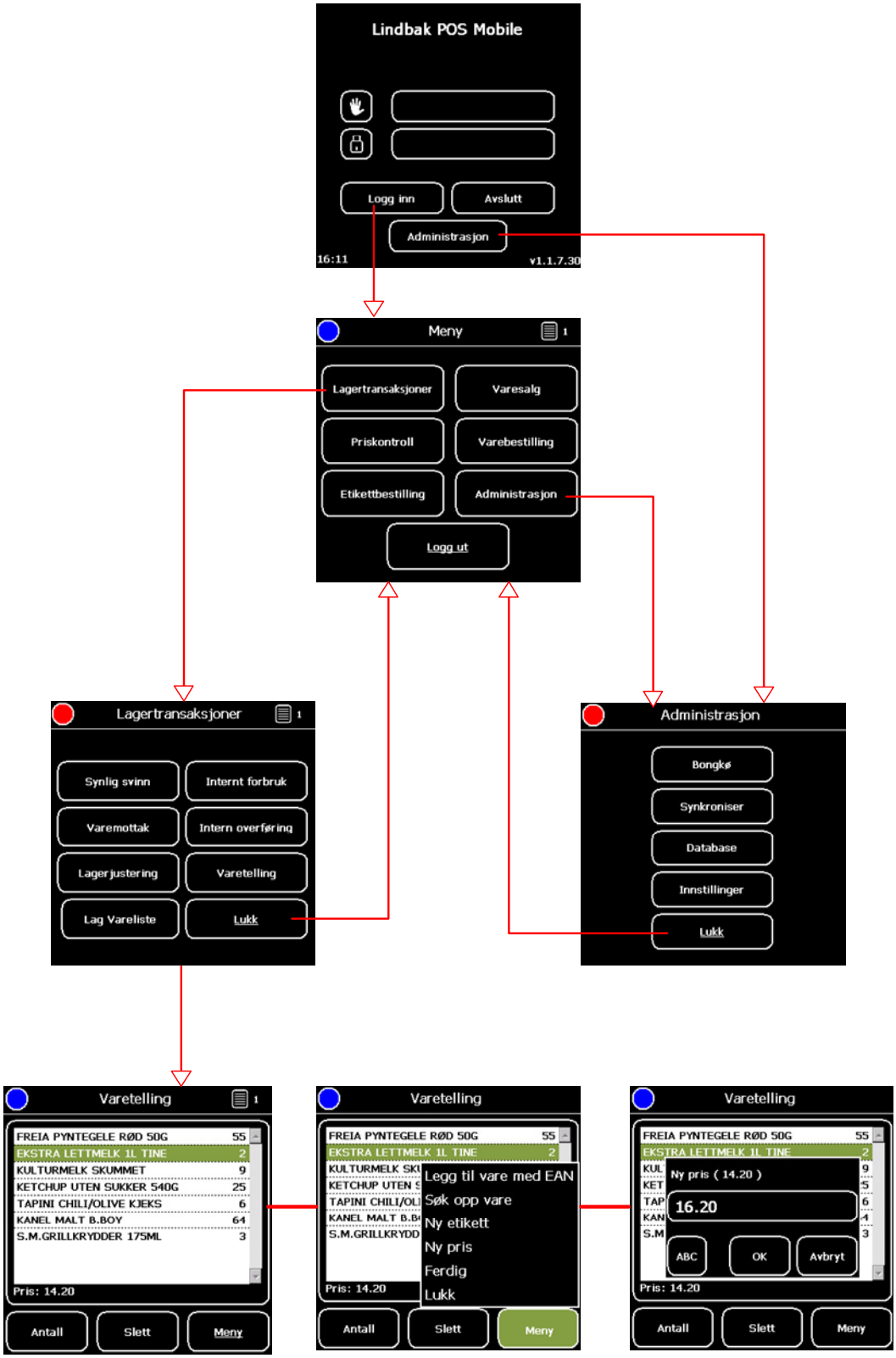
Tradisjonelt har slike håndterminalprodukter i butikk blitt brukt som en batch-terminal. Brukeren har tatt med seg håndterminalen i butikken og for eksempel foretatt en varetelling. Når vedkommende er ferdig har håndterminalen blitt satt i sleden (se figur 2) på butikkens bakrom. Varetellingen og eventuelt andre jobber som har blitt foretatt blir overført som en batch-jobb til bakromspcen sleden er tilkoblet. Lindbak POS Mobile støtter denne typen batchsynkronisering, men i tillegg kan Lindbak POS Mobile utnytte et trådløst nettverk i butikken. Ved bruk av trådløst nettverk vil alle jobber den butikkansatte fullfører bli automatisk overført i bakgrunnen. I tillegg muliggjør det trådløse nettverket funksjonalitet som krever direkte kommunikasjon med bakromspcen som for eksempel et varesalg der prisene må være riktige og totalsummen beregnes korrekt.

Figur 4 - Skjermbilde som viser ABC-tastaturet på skjermen.

Lindbak POS Mobile har tre tilstander: online, offline og dokket. Disse programtilstandene avgjør hvilken funksjonalitet som er tilgjengelig. Gjeldende tilstand til programmet vises øverst i venstre hjørne ved hjelp av en sirkel som skifter farge etter hvilken tilstand programmet er i (se figur 5). Nedenfor vil de tre tilstandene bli kort forklart

- **Dokket** (blå)
  Håndterminalen står nede i sleden som angitt på figur 2 og sirkelen lyser blått. Når håndterminalen blir satt ned i sleden startes det automatisk en synkroniseringsjobb som oppdaterer den lokale databasen til håndterminalen og overfører bonger som ligger i kø.
- **Online** (grønn)
  Håndterminalen er fri fra sleden og er tilkoblet et trådløst nettverk. Alle funksjoner kan utføres når håndterminalen er online og ferdige bonger blir automatisk overført i bakgrunnen.
- **Offline** (rød)
  Håndterminalen jobber frakoblet og databasen kan ikke synkroniseres. Noen funksjoner kan ikke benyttes og alle bonger blir liggende i kø inntil håndterminalen dokkes eller den kommer online.

På side 21 blir det vist en del ulike skjermbilder av produktet (figur 5). De øverste skjermbildene i den figuren viser navigeringen i programmet. Innloggingen er det første vi kommer til når programmet startes. Når en bruker logger seg inn åpnes hovedmenyen som har tittelen 'meny'. I denne hovedmenyen har vedkommende tilgang til et par funksjoner, mens 'lagertransaksjoner'-knappen og 'administrasjons'-knappen fører til undermenyene vist på skjermbildene nedenfor. Helt nederst på figur 5 blir det vist tre skjermbilder av varetellingsfunksjonen. Alle de andre funksjonene bortsett fra varesalg og priskontroll er bygd opp på akkurat samme måte og ser helt like ut sett bort fra tittelen. Det første varetellingsbilde er normaltilstanden og viser listen med de telte varene og vi kan bruke strekkodeleseren til å skanne flere varer. I varetellingsbilde to vises menyen i funksjonen. Fra der kan vi legge til varer ved hjelp av EAN-kode, søke opp varer, angi ny pris for valgt vare, bestille ny vareetikett for valgt vare, sette varetellingsbongen som 'ferdig' etter endt varetelling eller velge "lukk" for å avbryte og gå ut. I det siste varetellingsbilde vises et eksempel på en dialogboks som dukker opp når noe data må skrives inn.



Figur 5 - Skjermbilder fra Lindbak POS Mobile som viser menyflyten og sentrale funksjoner

4 Evalueringsmetoder

Dette kapitelet vil presentere en del evalueringsmetoder innen brukbarhet som er relevant for gjennomføringen av brukbarhetstester i denne studien. Valget av de ulike metodene vil bli begrunnet i kapittel 4.3. Der vil det blant annet bli diskutert litt rundt valget av kvantitative- og kvalitative metoder, og hvordan de forskjellige empiriske metodene vil bli brukt.

4.1 Empiriske metoder

Empiriske metoder er vitenskapelige undersøkelser av virkeligheten, og er kunnskap som er bygd på erfaring. I dette kapitlet vil vi gå nærmere inn på teorien bak de ulike empiriske metodene vi skal benytte i denne studien.

4.1.1 Observasjon

Observasjon går ut på å se og legge merke til noe. Det brukes innen forskingen som en datainnsamlingsmetode for å finne ut hva personer faktisk gjør. Det kan nemlig være avvik mellom hva en person sier han gjør hvis vedkommende blir intervjuet, og hva personen ville gjort ved observasjon. (Oates 2006:202-211)

Det skilles mellom to typer observasjon, skjult og åpenlys observasjon. Disse typene har begge sine fordeler og ulemper. Fordelen med skjult observasjon er at de som blir observert blir ikke forstyrret på noen måte og oppfører seg naturlig. Ulempene er at en må passe på og ikke bli avslørt og forskeren kan selvfølgelig ikke spørre personene om noe han lurer på. Det er også en etisk del å passe på når de observerte ikke har gitt tillatelse til å bli observert. Skjult observasjon kan være akseptert på offentlige områder, fordi der vet alle at de kan bli sett på av andre. I åpenlys observasjon derimot er fordelen at forskningen blir mer etisk ved at personene har gitt tillatelse til å bli observert. Det negative er ”Hawthorn effekten” som går ut på at mennesker kan endre adferd når de vet de blir observert. (Oates 2006:202-211)

Under en observasjon kan forskeren ha forskjellig grad av deltakelse. Forskeren kan være en ren observerer som aldri tar del i det som skjer. På motsatt side kan forskeren være en komplett deltaker, der han for eksempel blir medlem av en gruppe han ønsker å forske på. Dersom forskeren ikke kan skli inn som komplett deltaker på grunn av alder, kjønn eller andre ting, kan han velge å følge etter folk i stedet. (Oates 2006:202-211)

4.1.2 Intervju

Intervju er en spesiell form for konversasjon mellom mennesker, der vanligvis en person har til hensikt å skaffe informasjon fra andre. Det er en metode som passer bra hvis en forsker vil grave litt dypere for å finne detaljert informasjon, og vil kunne utforske følelsene/ sinnsstemningen til den som intervjues.

Det finnes tre hovedtyper intervju, og vi må velge den typen som passer best til de forskjellige situasjonene. Oates (2006:187-188) definerer tre intervjutyper;

- ***Strukturert intervju***: Her har du forhåndsdefinerte standardspørsmål som du spør hvert enkelt intervjuobjekt i en bestemt rekkefølge. Det er viktig at intervjueren ikke blir lokket med til å starte en samtale når intervjuobjektet svarer, for da kan intervjuobjektet få innsikt i intervjueren sitt syn og det kan medføre at svarene blir påvirket.
- ***Semistrukturert intervju***: På samme måte som i et strukturert intervju, har du definert en liste med temaer og spørsmål du vil spørre om, men du er villig til å endre på spørsmålsrekkefølger etter hvordan konversasjonen flyter framover og du kan stille ekstraspørsmål dersom intervjuobjektet støter innom noe interessant du ikke har forberedt spørsmål om.
- ***Ustrukturert intervju***: Et ustrukturert intervju er mer løsere og forskeren har mindre kontroll. Intervjueren introduserer gjerne et team og lar intervjuobjektet få snakke fritt rundt temaet.

  It has been shown that people respond differently on how they perceive the person asking the questions, that is, the data generated can depend on the perceived role and identity of the researcher. (Oates 2006:188)

Kjønn, alder, etnisitet, status og dialekt, kan være faktorer som kan påvirke hva slags informasjon et intervjuobjekt er villig til å dele. Derfor skal vi opptre så profesjonelt som mulig og forholde seg nøytral. Hvis ikke kan det føre til at vi får ukorrekte eller uønskede data. Dataene samles inn ved hjelp av feltnotater tatt underveis i intervjuet, eller med tekniske hjelpmidler som lydopptaker eller videoopptaker.

4.1.3 Spørreskjema

Spørreskjema er et forhåndsdefinert skjema med spørsmål som en forsker kan analysere resultatene av og se etter mønster og foreta generaliseringer av synspunktene til forskjellige grupper. Når en skal lage et spørreskjema må vi være påpasselig for hvordan det blir laget og strukturert, slik at det genererer gyldige- og pålitelige data. Det er også viktig at spørsmålene skrives på en slik måte at alle svarende tolker spørsmålene likt. (Oates 2006:219-221)

Oates (2006:223) definerer følgende typer spørreskjema:

- Ja/nei svar
- Kvantitative spørsmål
- Enig/uenig utsagn
- Grad av enig eller uenig ved bruk av en Likert skala

The System Usability Scale (SUS)

Et enkelt, pålitelig, robust og populært spørreskjema for subjektiv måling av brukertilfredshet innen brukbarhet er "*the System Usability Scale*" (SUS). Spørreskjemaet består av en Likert skala, der det angis om en er sterkt enig eller sterkt uenig i et utsagn på en skala fra 1-5. Det er totalt ti utsagn å ta stilling til, og de finnes i tabell 12 på side 75. Dette spørreskjemaet skal fylles ut etter gjennomført test og før debrifingen med intervju. (Brooke 1996)

4.1.4 Brukbarhetstesting

Brukbarhetstesting er en metode som har som mål å forbedre brukbarheten til et system. Deltakerne i en brukbarhetstest representerer ekte brukere og gjør reelle oppgaver mens vedkommende blir observert og gjerne filmet. Dataene vi sitter igjen med blir analysert, og vi oppdager problemer med brukergrensesnittet og kommer med forslag til hvordan problemene kan løses. (Dumas and Redish 1999:22-25)

En brukbarhetstest starter med at testpersonen får informasjon testen og nødvendig opplæring. Selve testen gjennomføres ved at en testperson får en oppgave som skal løses uten mulighet til å kunne spørre om hjelp. Når testpersonen har fullført brukbarhetstesten, er det normalt med en debrifing hvor vi snakker om eventuelle problemer som dukket opp, og andre forhåndsdefinerte spørsmål. Videre i dette kapitlet vil vi presentere en del aspekter rundt brukbarhetstesting.

Common Industry Format

Common industry format(CIF) er en ISO standardisert metode(ISO/IEC 25062 Software engineering – Software Product Quality Requirements and Evaluation) for rapportering av resultater fra en brukbarhetstest. Den er basert på definisjonen av brukbarhet ISO 9241-11. Formatet ble laget på bakgrunn av at det eksisterer mange måter å rapportere brukbarhetstestresultater på. Ett felles format gjør det enklere i ettertid å kunne verifisere validiteten, og det fører til mindre feiltolkning av resultatet, enn om alle rapporterer sine resultat på forskjellige måter. (Theofanos 2005)

> *The CIF does not tell you what to do; it tells you how to report on what you did*
> (Theofanos 2005)

Som sitatet sier vil CIF bare hjelpe deg med hvordan du skal rapportere resultatene. CIF antar at "*best practice*" prinsippet er benyttet når det gjelder gjennomføringen av brukbarhetstester. Theofanos (2005) oppsummerer hvilke seksjoner en CIF rapport inneholder:

- En beskrivelse av produktet
- Målene med testen
- Testpersonene
- Oppgavene disse testpersonene utførte
- Oppbyggingen av testen
- Metoden eller prosessen for hvordan testen ble gjennomført
- Datainnsamlingsmetoder og brukbarhetsmålene
- De numeriske resultatene.

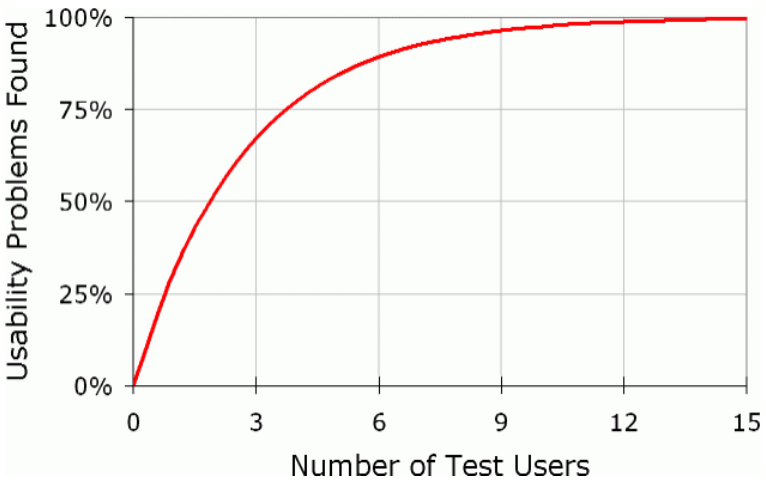
Testpersonene må representere reelle brukere

Det er viktig for validiteten til en brukbarhetstest at testpersonene representerer reelle brukere. Til å illustrere denne viktigheten kan vi se på et eksempel; Det vil være vanskelig for en bilmekaniker å brukbarhetsteste et produkt for sykepleiere eller motsatt. Det sier seg selv at det vil oppstå problemer med at bilmekanikeren ikke klarer å sette seg inn i rollen og oppgavene til en sykepleier, og han vil også slite med fagbegreper. For at en brukbarhetstest skal være gyldig, må testpersonene være en del av målgruppen for produktet. Vanligvis finner vi aktuelle kandidater ved å lage en brukerprofil eller en personas. (Jacko and Sears 2003:1131)

> *If the participant in the usability test do not represent the real users, you are not seeing what will happen when the product gets to the real users* (Dumas and Redish 1999:23)

Hvor mange testpersoner behøves

En av grunnene til at brukbarhetstesting har blitt så populært er evnen til å finne brukbarhetsproblemer med få testpersoner. Dette ser vi fort ved å observere flere tester med samme oppgaver. Det skal ikke mange testene til før vi ser mange av de samme problemene går igjen. En plass mellom fem og åtte testpersoner er tilstrekkelig for å avdekke de fleste problemer. Det vil mest sannsynlig være unødvendig å teste med flere, fordi oppdagelsen av nye problemer flater ut i grafen i figur 6. (Jacko and Sears 2003:1140)



Figur 6 - En graf som viser hvor mange brukbarhetsproblemer som avdekkes ved å teste x antall personer. Hentet fra Nielsen (2000)

Nielsen (2000) har gjennomført en studie og funnet ut at det beste resultatet får vi ved å teste fem personer. Ut fra grafen til Nielsen i figur 6 leser vi ut at 85 % av alle brukbarhetsproblemene er oppdaget etter fem brukbarhetstester. For å oppdage så og si alle brukbarhetsproblemene mener Nielsen vi trenger å gjennomføre 15 brukertester.

Det har kommet noe kritikk av Nielsen sin antagelse. Faulkner (2003) har gjennomført en studie der 60 personer ble brukbarhetstestet. Hun viser risikoen ved at et utvalg av fem personer bare fant 55 % av alle brukbarhetsproblem, mens et annet utvalg med fem personer fant hele 99 % av alle problemer. Faulkner (2003) understreker at det kommer veldig an på, om det holder å teste med fem personer. Målgruppen for et produkt har en god del å si med tanke på antall personer som bør testes. Når det er et produkt for allmennheten som skal testes, bør det testes med mange brukere med forskjellig erfaring.

Et annet aspekt som også er relevant når det kommer til brukbarhetstesting er balansen mellom tid, penger og informasjonen vi sitter igjen med. Tid og penger er dessverre det som vanligvis bestemmer hvor mange som skal testes, men i mange tilfeller er det ikke nødvendig å bruke masse ressurser på å finne absolutt alle brukbarhetsproblemene. De viktigste og mest sentrale problemene oppdages som regel med bare et par testpersoner. (Dumas and Redish 1999:127-129)

Høyttenkning

Høyttenkning går ut på at testpersonen tenker høyt når oppgavene løses, og dette hjelper den som skal analysere testen. Når en testperson observeres finner vi ut hva vedkommende gjør, men observasjonen sier ingenting om hvorfor testpersonen handler som vedkommende gjør. Ved å bruke høyttenkning finner vi ut hvorfor vedkommende handler akkurat på den måten. Det gir en innsikt i personens tankeprosess, og kan hjelpe til med å finne elementer i brukergrensesnittet som kan føre til misforståelser, og som derfor bør redesignes. (Nielsen 1993:195-198)

> *It is surprising how much of this analysis is dependent on the think aloud protocol. We depend on what participants say to help us understand the problems. (Jacko and Sears 2003:1133)*

Tradisjonelt sett har høyttenkning blitt brukt som en metode innen kognitiv psykologi, men har blitt adoptert av MMI-miljøet. Styrken med denne metoden er alle kvalitative data vi sitter igjen med. Ebling (2000) har utført en studie som viser at høyttenkning avdekket 1/3 av alle alvorlige problemer og 2/3 av alle mindre alvorlige problemer. Svakheten med dette studiet er at resultatene stammer fra bare én enkelt brukbarhetsstudie, men det gir en pekepinne og er med på å understreke viktigheten av høyttenkning under brukbarhetstester. Ebling (2000) poengterer også viktigheten av å transkribere ned verbale data, hvis ikke er det stor sjanse for å gå glipp av kritiske kommentarer.

Det ligger ikke i menneskets natur å tenke høyt, derfor er det nødvendig å gi tilstrekkelig med opplæring i metoden. De beste høyttenkerne snakker som en foss og deler hele tankerekken deres, mens de dårligste sier nesten ingenting eller bare mumler. Folk flest er nok en mellomting av dette og det kan hende testpersonene må minnes på å tenke høyt. (Dumas and Redish 1999:278-281)

Etikk

I tillegg til det generelle etiske ansvaret forskere eller profesjonelle har innen brukskvalitet, er det også noen ekstra aspekter å passe på innen brukbarhetstesting. Ved bruk av høyttenkning sier Molich (2003:161) at vi må ta hensyn for å unngå situasjoner som er ubehagelig for testpersonen. Dumas og Fox har definert følgende tre viktige punkter om etikk ved brukbarhetstesting i Jacko og Sears (2003:1144-1145):

- **Informert samtykke**: Går ut på at testpersonene får informasjon om deres rettigheter før de starter en brukbarhetstest. Det sier noe om retten de har til å trekke seg når som helst uten straff. De bør få beskjed om eventuelle risikoer ved å delta og om hvordan dataene vil bli behandlet(spesielt lyd og film). Dumas og Fox sier at testsesjoner som blir utført uten slik samtykke er uetiske.
- **Konfidensialitet til dataene**: Det er ansvaret til testlederen å passe på at testpersonene sine navn ikke brukt noen plass, selv ikke på videokassettene. Testeren bør være forsiktig ved bruk og distribusjon av høydepunktvideo. Dersom det utvikles noe i et selskap og selskapets ansatte er testpersoner bør det ikke benyttes video med høydepunkt. Den som blir testet skal få beskjed på forhånd om at sjefen til den ansatte ikke får se videoen for å slippe det presset.
- **Balanser ubehag med formål**: Forskeren må tilse at formålet med testingen kan rettferdiggjøre det emosjonelle stresset testpersonene blir utsatt for. Det er normalt at testpersonene får beskjed om at det er produktet som skal testes og ikke de. Til tross for dette skylder mange av testpersonene på seg selv når de sliter eller gjør noe feil. Dumas og Fox mener at stresset testpersonene blir utsatt for kan rettferdiggjøres med balanseprinsippet om at fremtidige brukere

blir spart stress ved å finne årsaken i testen. De sier at strevingen til et par testpersoner kan gi motivasjon for utviklerne til å fikse problemene, og på den måten forhindre de problemene hos fremtidige brukere.

Sjekkliste for gjennomføring av brukbarhetstester

Tognazzini (1992:83-86) har laget en sjekkliste som er fornuftig å følge ved gjennomføring av brukbarhetstester:

1. Introduser deg selv
2. Beskriv hensikten med testen
3. Fortell deltakerne at de kan avbryte når de vil
4. Beskriv utstyret i rommet og begrensningene til prototypen
5. Lær bort hvordan en tenker høyt
6. Forklar at du ikke kan tilby hjelp under testen
7. Beskriv oppgaven og introduser produktet
8. Spør om det er noe de lurer på og kjør testen
9. Avslutt testen med å la brukeren uttale seg før du samler evnt. løse tråder
10. Bruk resultatene

Pilottest

En pilottest er en fullstendig testgjennomføring av en brukbarhetstest. Grunnen til at pilottester gjennomføres er for å feilsøke alt utstyr, programvare, materiell og prosedyrene før vi begynner å brukbarhetsteste med dem vi skal teste. Pilottesten bør gjennomføres nøyaktig som en full brukbarhetstest og pilottestpersonen bør representere en reell bruker. En slik pilottest er også god trening for brukbarhetsekspertene som skal gjennomføre alle testene. (Dumas and Redish 1999:264-269)

Etter endt pilottest skal vi kritisk vurdere oppgavene, testforløpet og testlederen sin rolle. Dersom pilottesten avdekker for mange feil, eller hvis testoppsettet endres radikalt bør det vurderes om det skal gjennomføres en ny pilottest. En pilottest vil for eksempel avsløre typiske feil som: Første oppgave er for stor. Testdeltakerne misforstår oppgaveteksten. Irrelevante oppgaver. Relevante oppgaver mangler. For mange oppgaver. (Molich 2003:148)

4.1.5 Logging

Logging involverer at en datamaskin samler inn detaljert data og statistikk over bruken. Metoden er spesielt nyttig fordi den viser akkurat hvordan brukeren har utført arbeidet, og det er lett å samle inn data automatisk fra et stort antall brukere. Logging blir gjerne utført ved å koble seg mot lavnivådeler i en datamaskin som tastatur og mus, eller ved å modifisere det konkrete programmet vi vil logge bruken til. Den siste måten er å foretrekke fordi den gjør det enklere å logge det som er interessant uten å analysere lavnivådata. Vanligvis blir logging brukt som en metode for å samle informasjon om bruken i felten etter at et system er sluppet ut på markedet,

men logging kan også bli brukt som en supplementmetode under brukbarhetstesting for å kunne samle mer detaljert data. (Nielsen 1993:216-220)

4.2 Analytisk evalueringsmetoder av brukbarhet

I tillegg til brukbarhetstesting, eksisterer det også et par metoder som ikke involverer eksterne brukere. Den som evaluerer er gjerne en domeneekspert eller brukergrensesnittekspert. Disse metodene kan gjerne kalles inspeksjoner og består av ekspertevaluering. En ekspertevaluering kan bestå av inspeksjon, konsistenssjekk, guidelines sjekk, heuristisk evaluering og kognitiv gjennomgang. Alle disse evalueringsmetodene er billige og enkle å gjennomføre. Det finnes også flere inspeksjonsmetoder som pluralistisk gjennomgang, formell brukbarhetsinspeksjoner og standard sjekk med mer, men de vil ikke bli utdypet noe nærmere her. (Nielsen 1994)

4.2.1 Inspeksjon (review)

Eksperten innehar rollen som både testleder og testperson ved testingen av brukbarhet. Han går gjennom hele systemet og noterer ned hver gang eksperten ser ut til å ha oppdaget et problem. Inspeksjonen bygger på sunn fornuft og egne erfaringer fra tidligere brukbarhetstesting. Faren med gjennomføring av inspeksjoner er at det er en upålitelig metode, fordi den bygger på meninger. Resultatene har en verdi dersom en annen ekspert kan bekrefte samme funn. (Molich 2003)

4.2.2 Konsistenssjekk

Konsistensen til system kan bli sjekket av en eller flere eksperter ved å passe på at alt i brukergrensesnittet er konsistent (gjøres likt overalt). I en konsistenssjekk passer vi på at GUI elementer benytter samme type knapper, font, menystiler, farge utforming osv… Vi kan også passe på at språkbruken er lik overalt og at like ord brukes til å forklare samme funksjonalitet. Det er også andre elementer vi kan verifisere som konsistensen til interaksjonsstil, visuell kommunikasjon, konsistens i forhold til operativsystem med mer. (TDT4180 2006)

4.2.3 Guidelines sjekk

En guideline sjekk går ut på at en eller flere eksperter evaluerer et system etter en sjekkliste med guidelines. Det finnes mange forskjellige guidelines. Microsoft og Apple har definert guidelines for designing av programmer på deres plattform. Det eksisterer guidelines for mobile enheter, programvare for barn og så spesifikt som design guidelines for brukergrensesnitt i en handelskontekst som Blum og Khakzar (2007) har laget.

4.2.4 Heuristisk evaluering

I en heuristisk evaluering benyttes et mindre antall evaluerere som undersøker et brukergrensesnitt og bedømmer om det overholder brukbarhetsprinsippene i heuristikkene (heuristikkene er gjengitt i kapittel 2.3). Normalt sett holder det å bruke 3-5 personer til å evaluere. Selve evalueringen blir gjort individuelt ved at hver person inspiserer brukergrense-snittet flere ganger. De forskjellige elementene i grensesnittet blir sammenlignet opp imot de ti heuristikkene. Etter endt evaluering skal vi sitte igjen med en liste med brukbarhetsproblemer, der hvert problem refererer til en heuristikk. (Nielsen)

4.2.5 Kognitiv gjennomgang

En kognitiv gjennomgang går ut på at en eller flere eksperter setter seg inn i brukerens rolle og simulerer brukerens problemløsningsprosess. Gjennomgangen skjer på bakgrunn av vanlige og kritiske brukeroppgaver (Molich 2003:166). Vi går igjennom konkrete oppgaver som en vanlig bruker ville ha gjort og stiller seg følgende spørsmål; Vil brukeren skjønne hva som må gjøres, se hvordan vi kan gjøre det og forstå fra tilbakemeldingen om en handling var korrekt eller ei. Sammenlignet med en heuristisk gjennomgang, fokuserer en kognitiv gjennomgang mer på identifisering av høynivå brukerproblemer. Metoden kan være nyttig til bruk på systemer som innebærer komplekse operasjonen for å fullføre gitte oppgaver. (Sharp, Rogers et al. 2007:592, 702-703)

4.3 Valg av metode

Dette kapittelet vil gi en oversikt over de metodene som er benyttet i dette studiet. Vi vil gå inn på skillet mellom kvantitative og kvalitative metoder i studiet. Det redegjøres hvorfor de forskjellige metodene er valgt, og hvordan de blir brukt opp imot hovedmålet med studiet, nemlig metodesammenligningen.

4.3.1 Kvantitative- og kvalitative metoder

De to hovedkategoriene forskningsmetoder grupperes i er kvantitative- og kvalitative metoder. Kvantitative data er basert på numeriske data som vanligvis er generert gjennom forskjellig type eksperimenter eller spørreundersøkelser. Analysen av kvantitative data går ut på å se etter mønster ved for eksempel å bruke tabeller, diagrammer eller grafer og konkludere deretter. Kvalitative data derimot tar for seg alle data som ikke er numeriske. Eksempler på slike data er bilder, lyder, ord og så videre. Disse datatypene samles gjerne inn ved hjelp av case studier, action research eller etnografi. Analysen av kvalitative data går ut på å få dataene ned i et fornuftig format(for eksempel tekst). Kategoriser dataene og se etter mønster som vi igjen kan gå dypere ned i. (Oates 2006:245-280)

I dette studiet vil det bli benyttet både kvantitative- og kvalitative forskningsmetoder. Hovedfokuset i denne rapporten vil være på kvantitative metoder ettersom at oppgaven går ut på å sammenligne tre metoder for brukbarhetstesting av mobile enheter. Definisjon for brukbarhet som nevnt i kapittel 2.2.1 er anvendbarhet, effektivitet og tilfredshet, og disse måles alle ved hjelp av kvantitative metoder. Selv om vi måler brukbarhet ved hjelp av kvantitative metoder er den kvalitative høyttenkningen under brukertestingen verdifull i dette studiet. Den gir innsikt i hvorfor brukerne gjør som de gjør og det er med på å avdekke flere feil (Ebling and John 2000). I tabell 1 er de forskjellige forskningsmetodene som vil bli brukt delt inn om de er en kvantitativ eller kvalitativ metode.

Kvantitative metoder	Kvalitative metoder
Brukbarhetstest	Brukbarhetstest
Spørreskjema	Intervju
SUS	Observasjon

Tabell 1 - Oversikt over de kvantitative- og kvalitative forskningsmetodene som blir benyttet i dette studiet

4.3.2 Brukbarhetstest

Evalueringsmetoden brukbarhetstest er sentral i dette studiet, ettersom at det er den metoden som blir sammenlignet i tre ulike kontekster. Det som er viktig å ta hensyn til i dette studiet er å holde oppsettet i alle tre kontekstene mest mulig likt. Alle oppgavene er holdt like, det er laget en sjekkliste for informasjon og opplæring før teststart, og en intervjuguide til å debrife testpersonene etter test. En mer detaljert gjennomgang av testoppsettet blir presentert i kapittel 6.2 og 6.3 under gjennomføring. Videre har råd gitt i CIF blitt brukt under testingen og noen resultater blir presentert etter eksempeltabeller anbefalt av CIF.

4.3.3 Spørreskjema

Det ble laget et pretest spørreskjema i den hensikt å hente inn mer detaljert bakgrunnsinformasjon som kan gjøre det enklere å tolke dataene fra testen. Den andre grunnen som Dumas og Redish (1999:209) sier er å bekrefte kvalifikasjonen for testpersonene og gå mer i dybden på hva slags type butikkerfaring de har. Fordelen med å lage spørreskjema er at testpersonen alltid blir spurt om det samme spørsmålet og det fører til at vi ikke glemmer å spørre et spørsmål. (Dumas and Redish 1999:208-212)

Etter endt brukbarhetstest sitter gjerne testpersonen igjen med et inntrykk av produktets brukbarhet og vi kan bruke et posttest spørreskjema for å samle inn de inntrykkene (Dumas and Redish 1999:211-212). Et slikt spørreskjema er med på å dekke tilfredshetspunktet i ISO9241-11. Hornbæk og Law (2007) sier at et standardisert spørreskjema måler vanligvis tilfredshet mer pålitelig enn et hjemmelaget spørreskjema.

I denne studien blir System Usability Scale(SUS) benyttet. Det skal utfylles av hver enkelt testperson rett etter endt brukbarhetstest og før intervjuet (Brooke 1996). Grunnen til at det er viktig å gjøre det før intervjuet er fordi testpersonens inntrykk av produktet kan bli påvirket for eksempel av gjennomgangen av problemene testpersonen hadde underveis.

4.3.4 Intervju

Molich (2003:54-57) sier at vi på forhånd bør utarbeide en sjekkliste for gjennomføringen av intervjuet etter endt brukbarhetstest. Molich poengterer at vi må ikke la en slik sjekkliste styre intervjuet, fordi han sier det er viktig at den som blir intervjuet føler vedkommende har mulighet til å komme med egne meninger og utspill. Denne sjekklisten er et hjelpemiddel for å passe på at vi husker å spørre testpersonen om alt viktig å spørre om.

På bakgrunn av Molichs (2003) anbefalinger valgte jeg derfor å benytte meg av et semistrukturert intervju. På forhånd laget jeg en liten intervjuveiledning som ligger i vedlegg VII, og det består av konkrete spørsmål og tema som jeg ville spørre om. Under brukertestingen i laboratoriet hadde jeg mulighet til å notere ned ting som testpersonen gjorde og som jeg kunne spørre ut om i intervjuet. I felttestingen ble dette derimot litt vanskeligere og jeg måtte prøve å huske mest mulig av interessante situasjoner selv om det er vanskelig.

4.3.5 Observasjon

Det vil være fornuftig å gjennomføre en åpen observasjon i felten. En slik observasjon vil kunne gi innsikt i om deres daglige erfaring er med på å påvirke resultatene vi får etter brukbarhetstestingen. Planen er å gjennomføre en slik observasjon med en av testpersonene i felten.

4.3.6 Logging

Den opprinnelige problemstillingen i denne masteroppgaven gikk ut på å sammenligne bruken av POS Mobile i en pilotbutikk mot å sammenligne resultatet med en arrangert brukbarhetstest i felten. For å kunne samle detaljert data av bruken i pilotbutikken ble det derfor laget utvidet logging av programmet med fokus på å fange opp alt av interaksjon om brukerhendelser.

Selv om problemstillingen ble endret, ble loggingen i programmet tatt vare på og kvalitetssikret før gjennomføring av brukbarhetstestene. Som Nielsen (1993:216-220) sier kan logging brukes for å samle mer detaljert data, men det vil også fungere som en sikkerhet med tanke på om skjermopptakingen under brukbarhetstestingen skjærer seg.

5 Forskningsmetoder og tidligere metodesammenligninger

Dette kapittelet vil gi en oppsummering av tidligere forskning som har sammenlignet forskjellige metoder for brukbarhetstesting av mobile enheter. Denne oppsummeringen vil danne grunnlaget for en gjennomgang av hvilke forskningsmetoder vi vanligvis benytter for å foreta en metodesammenligning. Deretter vil det bli diskutert hva slags forskningsmetoder som vil bli brukt i denne studien. Til slutt vil vi si noe om sammenligninger på tvers av team og hvorfor det er problematisk å sammenligne ulike studier.

5.1 Tidligere metodesammenligninger

Noen av de viktigste vitenskaplige artiklene som omhandler metodesammenligninger av mobile enheter i forskjellig kontekst vil bli oppsummert under dette kapittelet. En bør være klar over at en del slike sammenligningsartikler ikke skiller like godt på laboratoriesimulering med eller uten kontekst. De kaller det gjerne bare brukbarhetstest i laboratorium, men i praksis kan det være fullskala eller desktop brukbarhetslaboratorium. Til å skille disse litt mer konsekvent i oppsummeringen nedenfor blir ordene felt, fullskala eller desktop brukt entydig for å tydeliggjøre hva slags setting det er snakk om.

Felles for alle sammenligningene er at det benyttes flere moderatorer til å gå gjennom resultatene. Det er ikke unormalt å ha egne team med moderatorer som evaluerer resultatene i de enkelte metodene. I de ulike metodesammenligningene blir det benyttet vanlige mennesker i noen sammenligninger, og fagfolk innen for et domene i andre.

5.1.1 Is it woth the hassle? Exploring the added value of evaluation the usability of context-aware mobile systems in the field

Kjeldskov og Skov et al. (2004) utforsket den ekstra verdien en brukbarhetstest av mobile systemer har i felten i forhold til i et fullskala laboratorium. Hver av de to evalueringene involverte 6 sykepleiere som skulle teste en løsning for elektronisk pasientjournal. Laboratoriet ble satt opp som en fullskala brukbarhetstest med realistiske pasientrom. Til å samle inn data i laboratoriet benyttet de et minikamera(se figur 7) festet på den mobile enheten for å fange interaksjonen i tillegg til faste kamera i taket. Samme minikamera ble benyttet i felttesten i tillegg til at en person gikk etter testpersonen og filmet.

Figur 7 - Minikamera for å filme interaksjonen med den mobile enheten. Hentet fra Kjeldskov og Skov et al. (2004).

Resultatet fra studien viser at de fant flere brukbarhetsproblemer i laboratoriet enn i felten (36 mot 23). Det ble funnet nesten like mange kritiske feil, men en betydelig større del alvorlige og kosmetiske feil i laboratoriet. Testpersonene i laboratoriet fant i snitt 18,8 brukbarhets-problemer hver mot 11,8 i felten. Til å gjennomføre de to evalueringene brukte de 34 timer i laboratoriet mot 65 timer i felten.

De konkluderer med at den ekstra verdien felttestingen gir er overraskende liten. I laboratoriet klarte de å finne alle brukbarhetsproblemene som ble oppdaget i felten bortsett fra en. Den ene kritiske feilen som bare ble oppdaget i felten sier de er på grunn av manglende realisme i laboratoriet. På bakgrunn av studiet sier de at vi bør kanskje ikke bruke verdifull tid i felten hvis vi har mulighet til å sette opp en realisk fullskala brukbarhetslab. De brukte nesten dobbelt så mange timer på feltevalueringen. Studien viser også at mangelen på kontroll i felten gjør det utfordrende å passe på at alle aspekter ved systemet blir testet.

5.1.2 It's worth the hassle! The added value of evaluating the usability of mobile systems in the field

Nielsen og Overgaard et al. (2006) har foretatt en studie som sammenligner resultatene fra en desktoptest i brukbarhetslaboratorium og en felttest. Til å samle inn data benyttet de systemlogging i tillegg til det samme utstyret som Kjeldskov og Skov et al. (2004) benyttet. Testpersonene ble bedt om å tenke høyt under brukbarhetstesten. I hver enkelt metode ble det benyttet 7 testpersoner som var i alderen 16-36 år gamle.

Det ble funnet signifikant flere brukbarhetsproblemer under felttesten med 60 mot laboratorietestens 48 brukbarhetsproblemer. Felttesten identifiserte tre kritiske problemer mer enn i laboratoriet, og hele ni flere kosmetiske brukbarhetsproblemer. 42% av problemene ble funnet i begge metodene. Når over halvparten av problemene bare identifiseres i en av metodene, tyder dette på at det kan være fornuftig å benytte flere metoder mener forfatteren. Desto mer alvorlig et brukbarhetsproblem er, jo større er sjansen for at problemet blir funnet i begge metodene. Effektiviteten mellom de to metodene er ganske jevn, bortsett på en oppgave der testpersonene i felten brukte signifikant lengre tid. Den oppgaven skyldtes at de i felten måtte gjøre litt ekstraarbeid i forbindelse med oppgaven. Brukertilfredsheten blir målt og indikerer at testpersonene synes det samme om systemet uavhengig av evalueringsmetode. Brukbarhetsproblemene er kategorisert etter hva slags type problem det er snakk om.

Konklusjonen er at felttesten fant signifikant flere brukbarhetsproblemer. Det var bare felttesten som avslørte problemer relatert til kognitiv belastning og interaksjonsmåte. Dette sier de at indikerer en mer realistisk kontekst gir samlet sett mer gyldig brukbarhetsinformasjon om systemet.

5.1.3 Usability testing of mobile applications: A comparison between laboratory and field testing

Kallio og Kaikkonen (2005) sammenlignet resultatene av brukbarhetstestingen med en mobiltelefon i en desktoptest og en felttest. De testet hele 20 personer i hver metode. Testpersonene var mellom 22 og 35 år gamle med en lik fordeling mellom begge kjønn. Høyttenkning ble benyttet og de brukte et minikamera til å fange interaksjonen. Produktet de testet ut var en mobil designet for vanlige forbrukere. Til å gjennomføre brukertestene ble det benyttet flere moderatorer for å unngå effekten vi kan få av å bruke bare en.

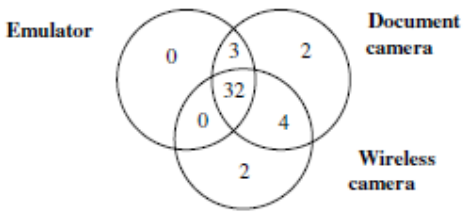
Resultatmessig fokuserer de på 22 brukbarhetsproblem som opptrådte hos minst to personer (46 distinkte brukbarhetsproblemer og bruksobservasjoner), og alle problemene ble oppdaget i begge metodene. Av disse er ni kategorisert som kritisk og det ble funnet et kritisk problem mer i felten, men i snitt var ikke feltproblemene alvorligere enn i laboratoriet. 10 alvorlige brukbarhetsproblemer ble funnet og tre mindre alvorlige problemer. I bare 3 av de 22 brukbarhetsproblemene var det statistisk signifikant forskjell mellom antall opptredener i laboratorium og felt. Tiden de brukte på oppgavene ble målt, og det var ingen signifikant forskjell mellom metodene. Det var heller ingen signifikant forskjell på hvor mange oppgaver de klarte. Mental belastning kunne bare oppdages i felten ved at personene ble veldig oppslukt mens de stirret på skjermen og kanskje gikk til siden.

De konkluderer med at studien ikke fant noen forskjell i antall problemer og alvorlighetsgrad i de forskjellige metodene. Ved gjennomføring av en brukergrensesnittevaluering mener de at felttesting ikke gir noen særlig ekstra verdi og ikke er verdt det med tanke på den ekstra tidsbruken og de ekstra kostnadene.

5.1.4 Usability testing of mobile devices: A comparison of three approaches

Betiol og de Abreu Cybis (2005) har brukbarhetstet et mobilt system ved hjelp av tre metoder: desktoplab med bruk av en datamaskin som kjørte en mobilemulator(desktopemulator), desktoplab med bruk av en mobiltelefon(desktopmobil) og en felttest med en mobilfelefon koblet til et minikamera(feltmobil). Totalt ble 36 personer mellom 21-40 år testet, med tolv personer i hver enkelt metode. Mobiltelefonen i laboratoriet ble låst fast i en posisjon som muliggjorde filming av interaksjonen mot den. Testpersonen ble aldri filmet, men lyden ble tatt opp.

Det viste seg at testpersonene med desktopmobil ga opp i større grad enn både feltmobil og desktop-emulatoren. Effektiviteten på noen oppgaver var også lavere for desktopmobil enn for de andre. Brukertilfredsheten målt ved hjelp av SUS var på 58 % for desktopmobil, 73 % for desktopemulator og 72 % i feltmobil. I figur 8 vises hvor mange brukbarhetsproblemer som ble oppdaget i de ulike metodene. Figuren viser at 74 % av problemene ble oppdaget av alle metodene. Desktopmobil fant flest problemer med sine 95 %. Hvis vi sammenligner den enkleste metoden(desktopemulator) mot den mest realistiske(feltmobil) ser vi at den enkleste metoden finner bare tre problemer færre. Betiol og de Abreu Cybis (2005) sier at av alle de åtte problemene som ikke ble oppdaget i metoden desktopemulator, var det bare ett av de problemene som ikke kunne ha skjedd i emulatoren fordi det var relatert til mobiltelefonen i bruk(på grunn av en liten forskjell i hvordan programmet kjørte på mobiltelefonen og på emulatoren).



Figur 8 - Viser antall problemer som ble oppdaget i de tre forskjellige metodene. Emulator = desktopemulator, document camera = desktopmobil, wireless camera = feltmobil. Fra Betiol og de Abreu Cybis (2005).

De foretar en kost-nytte analyse og finner ut at desktopmobil gir best ROI med 11.38 mot desktopemulator og feltmobil som gir henholdsvis 7,49 og 8,79. De mener at testpersonene var mindre tilfreds med desktopmobil fordi mobilen måtte stå fast i et stativ og det gjorde interaksjonen litt unormal. Testpersonene i felten ble ikke forstyrret av et stressende og støyete miljø.

5.1.5 Usability evaluation for mobile device: A comparison of laboratory and field tests

Duh et al. (2006) har brukbarhetstestet en mobiltelefon på ti personer i en desktoptest og ti personer i en felttest. Testpersonene tenkte høyt under testingen og de benyttet blant annet et minikamera til å fange interaksjonen. Et pretest spørreskjema ble brukt for å sjekke forskjeller i demografisk informasjon, men det ble ikke funnet noen signifikant forskjell. De har valgt å fokusere på brukbarhetsproblemene, tidsbruken og brukertilfredsheten.

Resultatet av testingen viser at det ble funnet signifikant flere brukbarhetsproblemer i felttesten. Av alle brukbarhetsproblemer ble 171 funnet i felttesten og 92 i laboratoriet. Det ble faktisk funnet 52 flere kritiske problem i felten enn i laboratoriet (64 i felt og 12 i laboratorium). Felttesterne brukte også signifikant lengre tid på løsingen av oppgavene. De sammenlignet også hvor positiv/negativ adferden var til testpersonene. Feltpersonene oppførte seg mindre positiv og mer negativ enn labtesterne. Adferden mener de kan være med på å påvirke resultatene fra brukbarhetstesten med hvor mange problemer som ble funnet og hvor lang tid de brukte.

Konklusjonen er at det ble funnet mange flere brukbarhetsproblemer i felttesten og mange av disse var kritiske. Noen av problemene var relatert til hvordan mobiltelefonen ble brukt i felten

og ville ikke vært mulig å oppdage i laboratoriet. Grunnen til disse forskjellene mener forfatterne er eksterne faktorer i miljøet som støy, kjørende tog, privatliv på et folksomt sted, mental og fysisk belastning som har påvirket testpersonene.

5.1.6 Creating realistic laboratory setting: Comparative studies of three think-aloud usability evaluations of a mobile system

I denne studien bringer Kjeldskov og Skov (2003) opp spørsmålet om det skal lage et realistisk fullskala-brukbarhetslaboratorium ved testing av mobile enheter. De benytter seg av tre forskjellige evalueringsmetoder: desktoplab med fagfolk, desktoplab uten fagfolk og fullskalalab med fagfolk. For hver av metodene ble det brukt tre lag med to personer. Disse skulle manøvrere et konteinerskip fra havn.

Resultatene fra testingen viste at desktoplab metoden uten fagfolk fant 64 % av alle brukbarhetsproblemene. Med fagfolk i desktoplab fant de 69% av alle problemene, mens 62 % ble funnet i fullskalalab med fagfolk. 12 % av alle problemene ble funnet i alle tre metodene. Det viste seg at alle tre metodene var gode til å identifisere store deler av de kritiske og alvorlige problemene, men bare fullskalametoden fant alle kritiske problemene.

Konklusjonen er at de ikke finner noen store forskjeller med tanke på antall brukbarhetsproblemer som ble oppdaget i de tre metodene. Det ser ut til at uerfarne personer finner like mange problemer som fagfolk. Realismen i miljøet hadde liten eller ingen påvirkning på antall problemer, men det virker som at fullskalalaboratoriet fant flere unike problemer.

5.2 Resultatoppsummering av tidligere metodesammenligninger

Resultatoppsummeringen vil bruke dataene fra tidligere metodesammenligninger til å enklere kunne sammenligne de kvantitative dataene. De oppsummerte resultatene vil være med å danne grunnlag for å sammenligne resultatene fra dette studiet opp mot tidligere metodesammenligninger. En bør være klar over at det er mange ulike faktorer i studiene som gjør at de ulike metodesammenligningene ikke ukritisk kan sammenlignes direkte ut fra dataene i dette kapitlet.

5.2.1 Problemfordeling

I tabell 2 ser vi hvor mange og hvor stor prosentandel av brukbarhetsproblemene som blir oppdaget ved hjelp av de ulike metodene. Den viser også hvor mange av problemene som er overlappende, det vil si antall problemer som ble oppdaget i alle metodene. På samme måte vises det i den nest siste kolonnen hvor mange problemer som bare blir oppdaget i en metode.

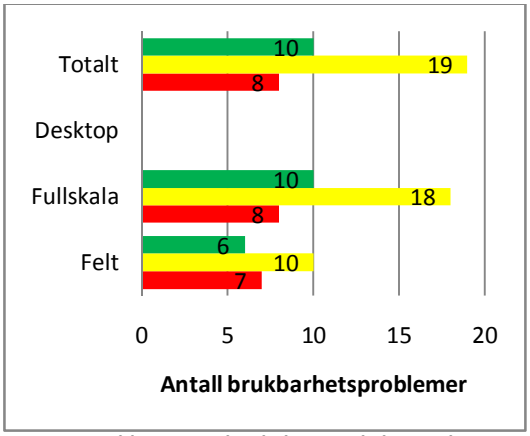
Kapittel	Felt	fullskala	Desktop	Overlappende problemer	Unike metode-problem	Totalt
5.1.1	23 (62%)	36 (97%)		22 (59%)	15 (41%)	37
5.1.2	60 (79%)		48 (63%)	32 (42%)	44 (58%)	76
5.1.3*	22 (100%)		22 (100%)	22 (100%)	0 (0%)	22
5.1.4**	38 (88%)	41 (95%)	35 (81%)	32 (74%)	4 (9%)	43
5.1.5	171		92			
5.1.6***		36 (64%)	40 (75%)	37 (70%)	16 (30%)	53

*** Studiet fokuserer bare på de problemene som opptrådte minst to ganger. Det ble egentlig funnet 46, men de sier ikke noe mer enn det.**
**** Det passer ikke helt å sette de to laboratoriemetodene som ble brukt i båsene fullskala og desktop. Resultatene i desktop stammer fra en test med mobilemulator, mens i fullskala benyttes en ekte mobiltelefon i et "desktopmiljø".**
***** De gjennomfører test på to brukergrupper i desktopmetoden, og det er testen med fagfolk som står under desktop i denne tabellen.**

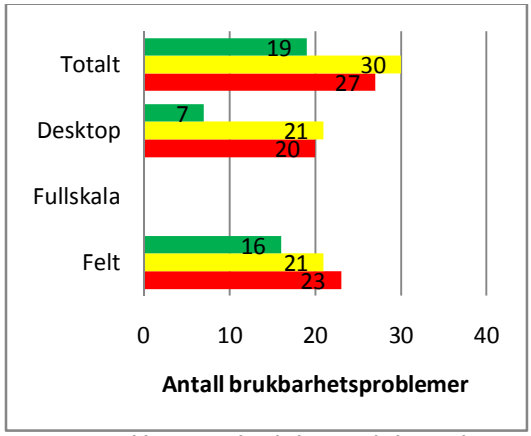
Tabell 2 - Antall brukbarhetsproblemer som ble oppdaget i de forskjellige metodene. Viser også antall overlappende problemer mellom metoder, og antall unike problemer for hver metode.

5.2.2 Alvorlighetsgraden til brukbarhetsproblemene

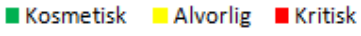
Alvorlighetsgraden til problemene blir presentert i figur 9-13 på neste side. Stolpene i figurene har fargekoder etter alvorlighetsgraden til problemene som ble oppdaget i de forskjellige metodene. Grønn betyr kosmetisk problem, gul betyr alvorlig problem og rød betyr kritisk problem (hvis fargeblind, er den øverste stolpen kosmetisk og den nederste kritisk). Artikkelen i kapittel 5.1.4 er ikke med på neste siden fordi den ikke gir tilstrekkelig med data for å kunne generere et stolpediagram.

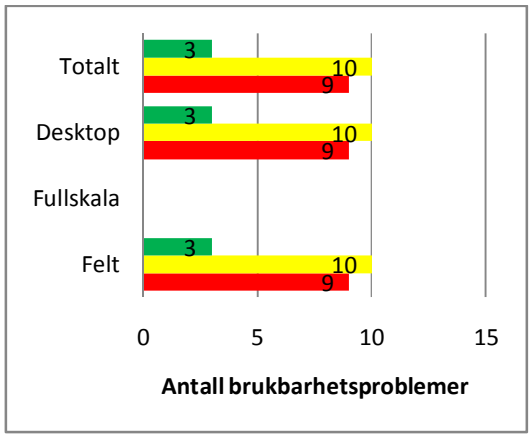


Figur 9 – Problemenes alvorlighetsgrad i kapittel 5.1.1

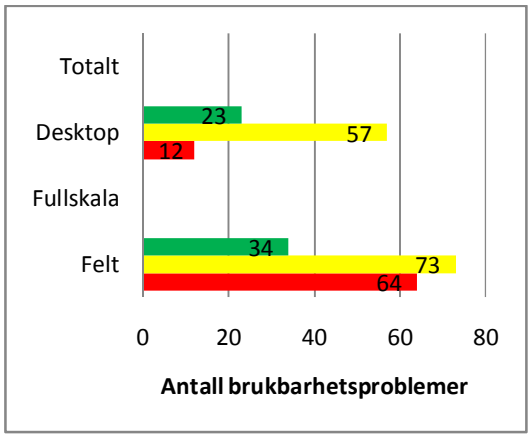


Figur 10 – Problemenes alvorlighetsgrad i kapittel 5.1.2

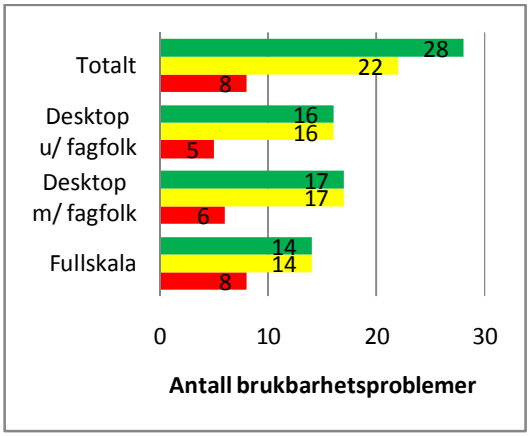




Figur 11 – Problemenes alvorlighetsgrad i kapittel 5.1.3



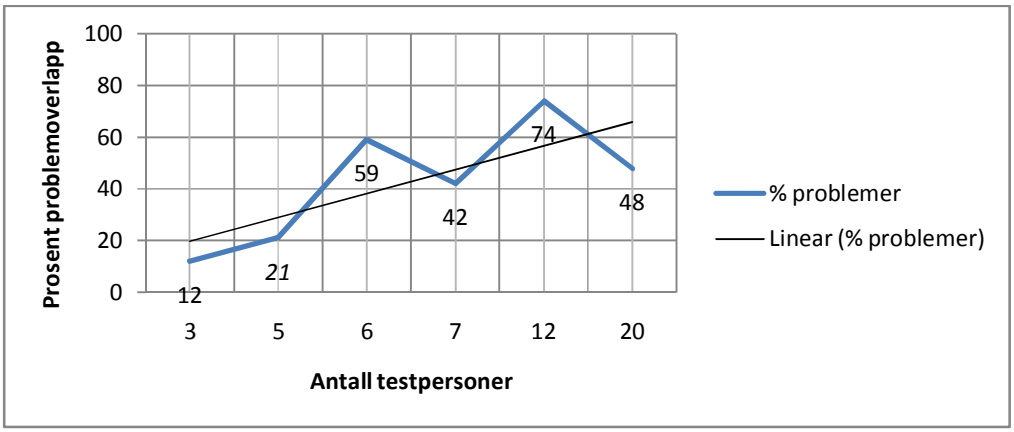
Figur 12 – Problemenes alvorlighetsgrad i kapittel 5.1.5



Figur 13 – Problemenes alvorlighetsgrad i kapittel 5.1.6

5.2.3 Overlappende problemer

Det er tilsynelatende en sammenheng mellom antall personer vi tester i hver metode og hvor stor del av problemene som overlapper mellom flere metoder. I figur 14 vises et diagram som bekrefter en trendlinje som viser at sjansen er større for å finne overlappende problemer, hvis vi tester på mange personer. Diagrammet er i og for seg interessant, men vi må være klar over at gyldigheten er lav ettersom at det er mange ulike faktorer i de tidligere metodesammenligningene som dataene stammer fra. Studiet som benytter fem personer i figur 14 er dette studiet.



Figur 14 - Linjediagrammet viser hvor mange problemer som er funnet i alle metodene oppover, og antall testpersoner bortover. OBS, eksperimentelt diagram med lav validitet.

5.3 Forskningsmetoder for metodesammenligning

Det finnes ingen veldefinerte veiledninger for hvordan resultatene fra brukbarhetstesting av forskjellige kontekster skal sammenlignes. Derfor må vi se på hvordan det har blitt gjort tidligere. Vi bør generere samme type data slik at disse kan sammenlignes som tidligere resultat.

I figur 15 vises en tabell med forskjellige forskningsmetoder de ulike metodesammenligningene baserer seg på. Kapittelkolonnene i figur 15 refererer til de tidligere metodesammenlignningene i kapittel 5.1. De forskjellige måtene metodene blir sammenlignet på blir oppsummert i underkapitlene til dette kapitlet.

Kapittel 5.1.X ->	1	2	3	4	5	6
0 Anvendbarhet (effectiveness)	X	X*	X	X*	X	X
5.3.2 Effektivitet (efficiency)		X	X	X		
5.3.3 Tilfredshet (satisfaction)		X		X		
5.3.4 Mental belasting (workload)		X		X		
5.3.5 Brukeradferd (user behavior)					X	
5.3.6 Ressursbruk	X		(X)			

*** Har målt fullføringsprosenten på oppgavene i tillegg til antall brukbarhetsproblemer**

Figur 15 - Viser hvilke metoder som blir brukt for å sammenligne resultater

5.3.1 Anvendbarhet

Anvendbarhet (effectiveness) kan presenteres på flere måter ifølge CIF, og to av dem blir brukt blant de forskjellige metodesammenligningene: opptelling av antall brukbarhetsproblem og hvor mange oppgaver testpersonene har klart å løse.

Den helt klart mest brukte måten å sammenligne forskjellige brukbarhetsmetoder er ved sammenligning av brukbarhetsproblemene. Alle tidligere metodesammenligninger har registrert brukbarhetsproblemer og sammenlignet de mellom metodene. To(2 og 4 i figur 15) har i tillegg valgt å si noe om fullføringsgraden på oppgavene for testpersonene. figur 15 viser at alle tidligere metodesammenligninger omtalt i kapittel 5.1 har sammenlignet anvendbarhet.

Til å finne dataene og analysere seg fra til brukbarhetsproblemene er det ikke unormalt at det benyttes et team til å evaluere felttesten og et annet team til å evaluere laboratorietesten. Når disse har funnet sine brukbarhetsproblem og klassifisert alvorlighetsgraden blir listene slått sammen. Den sammenslåtte listen blir dannet etter diskusjon og konsensus er oppnådd blant teamene.

Alle de evaluerte metodesammenligningene har benyttet en trenivågradering av brukbarhets-problemene. Alvorlighetsgradene til Molich (kritisk, alvorlig og kosmetisk) er mye brukt og blir omtalt i kapittel 7.6.1.

Kompletterningsgraden derimot viser hvor mange oppgaver testpersonene klarte å løse. Det presenteres for å vise om det er noen spesielle oppgaver som skiller seg ut i de forskjellige metodene. Resultatene blir gjerne presentert med antall personer som fullfører de angitte oppgavene og visualiseres for eksempel ved hjelp av et stolpediagram. Ut fra resultatene bør vi se etter signifikante forskjeller mellom de ulike metodene og annet som skiller seg ut på en interessant måte.

5.3.2 Effektivitet

Effektiviteten (efficiency) blir målt ved å se hvor lang tid testpersonene bruker på hver oppgave. Tidene blir brukt til å se om det er noen signifikant forskjell mellom tidsbruken ved bruk av ulike metoder. Alle som måler effektiviteten i kapittel 5.1 benytter stolpediagram for å vise gjennomsnittstiden til testpersonene i de forskjellige metodene bruker per oppgave. Det hender seg også at standardavviket blir presentert i samme stolpediagram.

5.3.3 Brukertilfredshet

Brukerens tilfredshet (satisfaction) blir målt for å se om det er noen signifikant forskjell mellom tilfredsheten til testpersonene i de ulike metodene. Brukertilfredsheten blir målt rett etter endt brukertest og det blir brukt egenlaget eller velkjent spørreskjema som for eksempel SUS og QUIS.

5.3.4 Mental belastning

NASA-TLX blir brukt i to av de tidligere metodesammenligningene for å måle mental belastning (workload). Bakgrunnen for å gjennomføre en slik test er for å finne ut om en mer realistisk settings i en felttest vil øke arbeidslasten signifikant i forhold til laboratorietester. NASA-TLX test kan ha noe for seg å gjennomføre dersom systemet vi skal brukbarhetsteste er et system som skal brukes i en særdeles stresset situasjon. Eksempler på slike situasjoner kan være systemer som en astronaut skal benytte i romferjen eller systemer for en jagerpilot i kamp.

En studie av Kjeldskov og Stage (2004) viser at den mentale belastningen øker signifikant når en person må løse en oppgave i varierende gåfart eller ved spasering hvor kursen endres flere ganger i forhold til å sitte i ro.

5.3.5 Brukeradferd

Duh et al. (2006) er den eneste som har målt brukeradferden når de har løst oppgavene. De har registrert om testpersonen har hatt en positiv, negativ eller nøytral adferd under gjennomføringen og mener at adferden kan relateres til brukbarhetsproblemer. En positiv adferd kan være at personen smiler og ser avslappet ut, men en negativ adferd er det motsatte. De fant signifikant forskjell på adferden i studiet til Duh et al. (2006), der feltpersonene var mer negativ enn laboratoriepersonene.

5.3.6 Ressursbruk

De fleste sier at det tar mer tid å gjennomføre en felttest kontra en laboratorietest. Dersom vi bruker mye mer tid på en felttest og verdien vi får ut av den ikke er god nok kan det hende at det ikke er verdt det ekstra strevet. Kjeldskov et al. (2004) brukte totalt 34 timer på planlegging, oppsett, testgjennomføring og evaluering i laboratorium. Til sammenligning brukte de 65 timer i felten.

Grunnen til at Kallio og Kaikkonen (2005) står i parentes i figur 15 er fordi de sier at de beregner dobbelt så lang tid på gjennomføringen av hver enkelt brukbarhetstest i felten i motsetning til laboratorietestene. Det skyldes måten de gjennomførte felttestene på og det trenger ikke nødvendigvis være sånn i andre felttester. De har ikke laget noe regnskap over det totale timeforbruket i de forskjellige metodene som Kjeldskov et al. (2004).

5.3.7 Oppsummering

Sammenligning av anvendbarhet med analysering av brukbarhetsproblemene er den metoden som går igjen blant alle i figur 15. To av forskningsartiklene som sammenlignes i figur 15 benytter brukbarhetsproblemene som eneste grunnlag for metodesammenligningen (hvis vi ser bort ifra ressursbruk, som ikke har noe med brukbarheten å gjøre). Måling av kognitiv last blir

gjort i to av sammenligningene, og kan være unødvendig å måle dersom testpersonen ikke skal stresstestes i konteksten.

Måten brukeradferd måles i Duh et al. (2006) er kanskje unødvendig, tidkrevende og noe unøyaktig. Det er meget trolig at resultatene fra brukeradferdsmålingene vil sammenfalle med resultatene fra et klassisk tilfredshetsspørreskjema. Det virker sannsynlig at en person med negativ adferd hos Duh et al. (2006) også vil få lavere score på en SUS test.

> *"Experience from empirical research shows that these measures can not be reduced to one single measure, and that all three should be reported when usability tests are done."* (Svanæs, Das et al. 2008) etter (Hornbæk and Law 2007)

Oppsummert kan vi konkludere med at en metodesammenligning bør som et minimum analysere anvendbarhet i form av oppdagede brukbarhetsproblemer. Det anbefales å måle de to andre målbare punktene i definisjon av brukbarhet ISO9241-11 også, nemlig effektivitet og tilfredshet.

5.4 Valg av sammenligningsmetode

I denne studien vil alle kriteriene som definerer brukbarhet i ISO9241-11 bli målt, presentert og analysert. Flere anbefalinger fra CIF om hva som bør måles og hvordan det bør presenteres vil bli tatt med i resultatkapittel 1. Den kognitive lasten for testpersonene blir ikke målt. Grunnen til det er at NASA-TLX testen er beregnet for kritisk programvare som for eksempel romfartsprogramvare. Den mobile butikkløsningen vi tester skal ikke brukes i en stressende omgivelse der det ikke handler om liv og død. Det ble gjort forsøk på å måle ressursbruken på samme måte som Kjeldskov et al. (2004), men det ble vanskelig å følge opp på grunn av hyppig skift av arbeidsfokus. Nielsen et al. (2006) sier at en del studier som sammenligner forskjellige metoder ikke nødvendigvis benytter samme type teknikkene for å samle data. I dette studiet vil samme type teknikker(høyttenkning, oppgaver, opplæring etc.) bli brukt i alle tre kontekstene og alt tilstrebes å bli gjort så likt som mulig.

5.5 Evalueringssammenligninger

I 1998 gjennomførte Rolf Molich første studie i en rekke der han sammenligner resultatene fra flere forskjellige team som har brukbarhetstestet det samme produktet (disse studiene kalles *comparative usability evaluation* og forkortes CUE). Totalt har det blitt gjennomført fire CUE-studier av Molich som vist i figur 16, men andre har også utført lignende studier inspirert av CUE-1. Disse evalueringssammenligningene er med på å sette denne studien i et perspektiv, og viser hvorfor det er vanskelig å sammenligne resultater på tvers av studier.

Formålet har vært å fylle gapet med manglende informasjon om hvordan praktiserende brukbarhets profesjonelle faktisk evaluerer og presenterer resultatene fra testing. Noen av målene med CUE er interessant for denne studien. På tvers av studiene ser de blant annet på om evalueringsresultatene er reproduserbare, vurderingen av alvorlighetsgrad og om det er viktige kvalitetsforskjeller mellom resultatene de forskjellige teamene har oppnådd. (Molich and Dumas 2006)

CUE-studiene viser at forskjellen mellom resultatene fra teamene er mye større en vi kanskje hadde regnet med. Hvis vi går nærmere inn på den siste CUE-studien (CUE-4), ble det benyttet 17 team med en lett blanding mellom brukbarhetstesting og ekspertevaluering som evalueringsmetoder. Teamene fikk i oppdrag å evaluere en relativt liten webside. Totalt ble det oppdaget 340 brukbarhetsproblemer. Av disse 340 ble hele 205 (60 %) brukbarhetsproblemer ikke funnet av andre enn bare enkeltteam som vi ser i figur 16. Av de 205 problemene ble 21 kategorisert som kritiske og 40 som alvorlige. Ingen av problemene ble funnet hos alle teamene. Den største overlappen var to problemer som ble funnet blant 15 av de 17 teamene. Den gjennomsnittlige overlapprosenten var 11,5 %. (Molich and Dumas 2006)

Study	CUE-1	CUE-2	CUE-3	CUE-4
Total number of participating teams	4	9	12	17
US teams	2	5	0	14
European teams (excluding Denmark)	2	1	0	3
Danish teams	0	3	12	0
Usability testing	4	9	0	9
Expert reviews	0	0	12	8
Evaluated system	Windows calendar program	Hotmail.com	Avis.com	Hotelpenn.com
Number of issues reported by single teams only	128 of 141 (91%)	232 of 310 (75%)	135 of 220 (61%)	205 of 340 (60%)
Conducted	March 1998	December 1998	August 2001	March 2003
References	Molich *et al.* 1998	Molich *et al.* 2004	Molich 2003, Hertzum *et al.* 2002	Molich 2003

CUE = Comparative Usability Evaluation.

Figur 16 - Oppsummering av alle fire CUE. Hentet fra Molich og Dumas (2006)

> *Usability testing is not the 'high quality gold standard' against which all other methods should be measured. Our study shows that usability testing – just like any other method – overlooks problems, even critical problems.* (Molich and Dumas 2006)

Oppsummert kan vi konkludere med at ulike team finner forskjellige problemer og kategoriserer alvorlighetsgraden forskjellig. Molich og Dumas (2006) mener at vi bør fokusere mindre på å finne 'alle' brukbarhetsproblemer. Selv med et relativt enkelt produkt kan det oppdages hundrevis av brukbarhetsproblemer, hvis flere brukbarhetseksperter evaluerer produktet uavhengig sier de.

6 Gjennomføring

Vi skal først se nærmere på rekrutteringsprosessen, oppgavelagingen og hvilke rutiner som ble laget for at hver test skulle gjennomføres mest mulig likt for å opprettholde god reliabilitet. Deretter vil vi gå inn på oppsettingen av lokasjonene til de tre metodene og utfordringene som dukket opp. Selve gjennomføringen av felttesten gikk over tre dager i uke 4 og begge laboratorietestene ble utført over fem dager i uke 5.

6.1 Rekruttering

Rekrutteringen kan vi dele i anskaffelse av butikk til felttesten og rekruttering av testpersoner til laboratoriemetodene. Vi vil si litt om utfordringene ved begge disse rekrutteringene under dette kapitlet.

6.1.1 Anskaffelse av testbutikk

Til å gjennomføre felttesten var det ønskelig å anskaffe en butikk som ville gå med på å la oss benytte butikklokalet til å sette opp nødvendig utstyr. Det var også ønskelig at butikken ville la oss benytte deres ansatte som testpersoner i arbeidstiden, uten at butikken fikk noen kompensasjon for det.

Det viste seg vanskelig å skaffe en butikk på disse premissene. Derfor ble det benyttet en god kontakt for å skaffe tilgang til en dagligvarebutikk i Steinkjer. Den butikken gikk med å på la meg benytte deres ansatte.

6.1.2 Rekruttering av testpersoner til laboratorietestene

Det ble satt som et krav at testpersonene måtte ha butikkerfaring for at de skulle kunne delta i brukbarhetstestingen. Kravet ble satt for at testpersonene skulle være reelle brukere, og mest mulig like brukergruppe som i felten er ideelt for å kunne sammenligne metodene mot hverandre. Som kompensasjon for å få folk til å verve seg som testpersoner, ble det lokket med 250kr for en times arbeid.

I første rekrutteringsrunde ble fire dagligvarebutikker i området rundt brukbarhetslaboratoriet besøk. Det ble lagt igjen rekrutteringsplakater (se vedlegg II), men ingen hadde tatt kontakt etter en uke og dermed ble plan B satt i gang. Plan B og andre rekrutteringsrunde gikk ut på å skaffe studenter med butikkerfaring, og plakaten i vedlegg III ble benyttet på NTNU Gløshaugen.

En uke etter at studentrekrutteringen ble satt i gang, hadde 16 personer tatt kontakt og ønsket å være testperson i studiet. Blant disse personene var det deriblant en person fra den første

rekrutteringsrunden. I og med at så mange tok kontakt, ble det mulig å sortere ut tre personer som hadde litt for lite butikkerfaring (f.eks. en person som hadde erfaring billettluka på den lokale kinoen).

6.2 Oppgaver

Her vil prosessen med å lage de forskjellige oppgavene som ble brukt under brukbarhetstesten bli forklart. Det ble lagt et par hensyn til grunn for oppgavene som ble laget og i hvilken rekkefølge de skulle utføres. Oppgaver som krevde forståelse av konteksten ble vektlagt, for å se om realismen spiller noen rolle. Til slutt vil vi nevne et problem med en av oppgavene som ikke ble avdekket før det var for sent.

6.2.1 Oppgavelaging

Opprinnelig ble det laget 20 oppgaver og som en liten pilottest ble de testet på masterstudenten Lars Flem (spesialiserer seg innen MMI og har sommerjobberfaring fra Netlife Research). Vi diskuterte oppgavene, og fant et par oppgaver som kanskje var litt unødvendige, og kuttet ned til 15 oppgaver. Etterpå ble oppgavene gjennomgått med veileder som ga råd om å redusere antall tester litt slik at brukbarhetstestene ikke overskrider totalt én time ved gjennomføring. Dette resulterte i en nedgang til elleve oppgaver.

Oppgaveteksten i de tre forskjellige metodene ble holdt like. Den eneste forandringen mellom metodene var matvarene som ble benyttet, og det ble gjort på grunn av praktiske årsaker. Oppgavene som ble brukt ligger i vedlegg I. Testpersonene brukte i snitt totalt 40 minutter på gjennomføringen av oppgavene.

De to første oppgavene ble laget veldig enkle. Dette er som Molich (2003:143) sier et psykologisk triks for å avstresse testpersonen. Testpersonen får en med dette en hurtig en mestringsfølelse og føler ubevisst at ”*dette var egentlig ikke så ille*”. Figur 30 viser at alle testpersonene klarte de to første oppgavene.

6.2.2 Utvelgelse av oppgaver

Oppgavene ble laget for å være mest mulig realistisk, og det var testpersonene enige i på spørsmål under intervjuet etter brukbarhetstesten. Måten oppgavene ble bygd opp og rekkefølgen de skulle gjennomføres i ble satt opp for å teste tre hovedmomenter:

- Teste mest mulig sentral funksjonalitet i Lindbak POS Mobile.
- Se om brukerne forstår hvordan programmet fungerer i de tre tilstandene offline, online og dokket.

- Teste om brukerne forstår hvordan konteksten skal brukes når løsningen ikke er i syne og de må gå en plass for å løse oppgaven.

De første syv oppgavene skulle gjennomføres uten bruk av online-tilstanden (tilstandene styres sentralt og kan ikke settes direkte i programmet), og det er slik de fleste i handelsbransjen bruker håndterminaler i dag. Fra og med oppgave 8 benyttes online-tilstanden aktivt og testpersonene må forholde seg deretter. Rekkefølgen er også laget for å være litt praktisk i felten, slik at vi slipper å springe mellom kontor og butikk hele tiden. På figur 18 kan vi se en skisse av butikken i felten og boblene indikerer hvor de ulike oppgavene ble løst. Oppgavene ble enten gjennomført på kontoret eller i butikken. Samme skille er forsøkt gjenskapt i fullskalametoden, hvor vi på figur 21 har kombinert kontor og lager, og har butikkarealet adskilt.

De tre første oppgavene ble gjennomført på kontoret, mens oppgave 4 tvang testpersonen til å forflytte seg til butikken for å telle varer. Oppgave 5 og 6 ble også gjennomført i butikken, og testpersonen var i butikken når oppgave 7 ble overgitt. Oppgave 7 gikk ut på at alle fullførte bonger skulle sendes over til bakromssystemet, og det gjøres ved å sette håndterminalen i sleden som står på kontoret. Oppgave 8 og 9 ble utført på kontoret og oppgave 9 gikk ut på å foreta et varemottak og sørge for at varemottaksbongen ble overført. Hovedmomentet i oppgave 9 var å se om testpersonene la merke til at bongen ble sendt over automatisk i bakgrunnen på grunn av online-tilstanden (i tidligere oppgaver ble bongene køet opp). Oppgave 10 tvang testpersonen til å hente noen varer i butikken, og oppgave 11 ble gjennomført på kontoret.

6.2.3 Problemer med oppgavene

Nielsen (1993:174-175) sier vi gjerne lager oppgavene vanskeligere for testpersonene enn hva vi hadde forventet på forhånd. Det vises ganske tydelig at ut fra figur 30 at oppgave 8 var alt for vanskelig når bare 3 av 15 klarte oppgaven. Dette kunne vært avdekket ved å gjennomføre en pilottest på en eller flere personer.

6.3 Gjennomføring av hver enkelt test

Til gjennomføringen av brukbarhetstestene ble det laget en egen sjekkliste(se vedlegg V) basert på de ti punktene til Tognazzini (1992) som ble omtalt under kapittel 4.1.4. Sjekklisten viser stegene i hele testen. Dumas og Redish (1999:250-261) sier at hver rolle skal ha en egen sjekkliste. I vårt tilfelle innehadde en person alle rollene og det ble derfor en global sjekkliste. I dette kapitlet vil vi går nærmere inn på et par av de sentrale stegene i sjekklisten.

6.3.1 Opplæringen av den enkelte testperson

Det ble gitt grunnleggende opplæring i bruk av håndterminalen og nødvendig informasjon om sentrale aspekter i Lindbak POS Mobile til hver enkelt testperson før brukbarhetstesten kunne starte.

For at alle testpersonene skulle stille på mest mulig likt grunnlag fra vår side, ble det laget en egen sjekkliste som gikk på opplæring, og den ligger i vedlegg VI. Det ble gitt opplæring i bruk av håndterminalen slik at de som har brukt samme type håndterminal tidligere ikke skulle ha noen fordel. Det viste seg at de som hadde brukt sammen håndterminal ikke visste hvordan all grunnleggende funksjonalitet fungerte(SMS- og ABC-tastatur).

6.3.2 Pretest spørreskjema

Før brukbarhetstesten besvarte testpersonene skjemaet i vedlegg IV. Det skjemaet ble brukt for å innhente mer informasjon om den enkelte. Vi spurte blant annet om hva slags oppgaver de gjorde i dagligvarejobben, og litt rundt datakunnskaper. Hensikten med skjemaet var å se om de forskjellige bakgrunnene kunne forklare hvorfor testpersonene handlet som de gjorde under testen. En annen grunn er for å få bekreftet at vi har fått riktige testpersoner. (Dumas and Redish 1999:209-210)

6.3.3 Brukbarhetstesten

Like før selve brukbarhetstesten ble testpersonene trent i å tenke høyt. Til å demonstrere det benyttet vi oss av Lindbak POS (kasseapparat) som kjørte på den bærbare datamaskinen (bakromspcen). Selve brukbarhetstesten ble gjennomført på vanlig måte. Etter at testpersonene hadde besvart SUS-skjemaet, gikk vi gjennom oppgavene som testpersonene slet med eller ikke fikk til.

6.3.4 Intervjuet etter endt brukbarhetstest

Etter gjennomført brukbarhetstest ble det foretatt et intervju. Til støtte under veis i intervjuet ble det benyttet en intervjuveiledning vedlagt i vedlegg VII. I tillegg til de forhåndsdefinerte spørsmålene, ble det også spurt om hendelser som dukket opp i den enkelte test.

6.4 Brukbarhetstest i felten

Brukbarhetstestingen i felten ble gjennomført på daværende Coop Mega (nedlagt til fordel for Obs! Hypermarked) i kjøpesenteret Amfi Steinkjer. Brukbarhetstestingen i felten ble gjennomført på tre dager i uke 4 2009. I dette kapitlet vil vi gå nærmere inn på hvordan lokasjonen i felten ble satt opp og gjennomført. Til slutt vil vi oppsummere problemene og

utfordringene knyttet til felttestingen. Det ble også gjennomført en observasjon av hvordan de butikkansatte benytter håndterminalene til daglig.

6.4.1 Oppsett av lokasjon

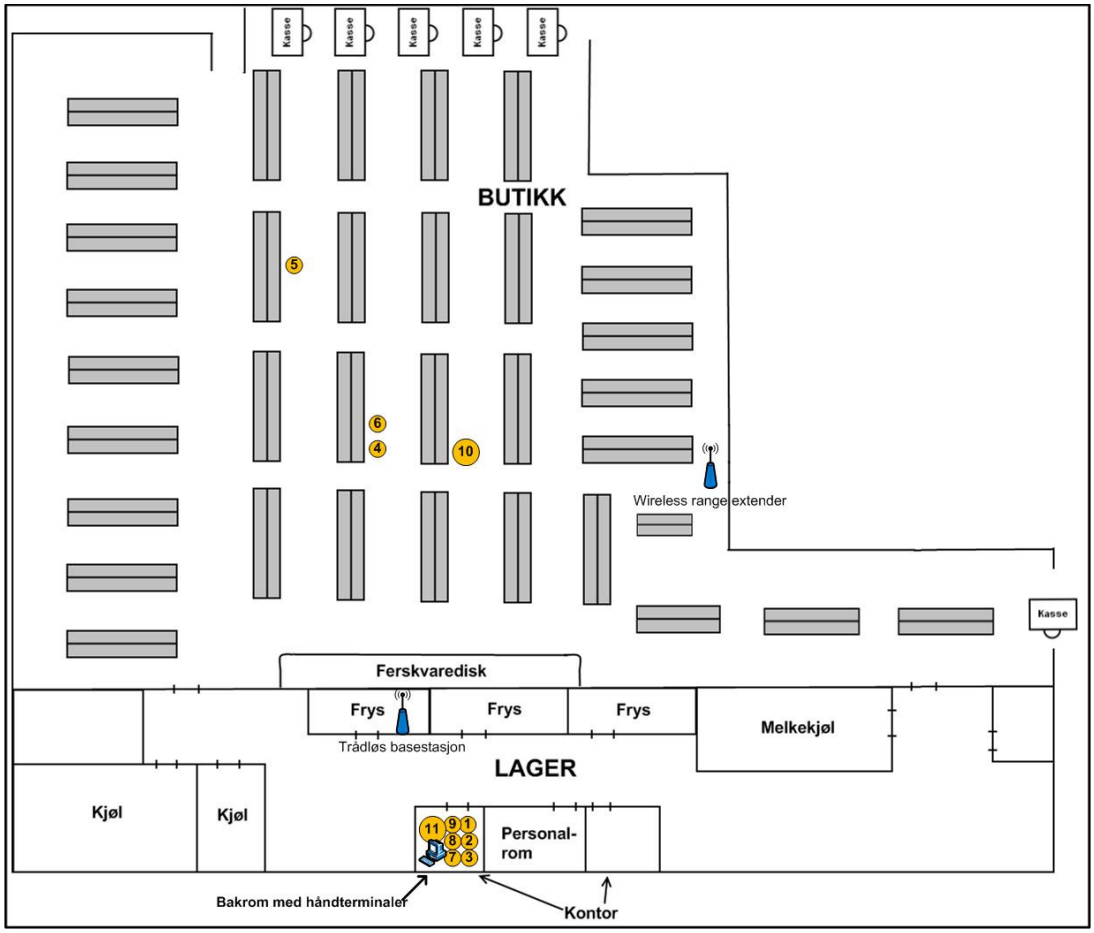
Den første dagen gikk med på å sette opp utstyret og teste det på stedet. Jeg fikk disponere et eget skrivebord inne på kontoret hvor de ansatte til dagelig administrerer butikkens håndterminaler (se figur 18). Rommet er avbildet på figur 17, og til høyre vises butikkens håndterminaler og bakromspc, mens til venstre er mitt testoppsett med håndterminal og pc fra Lindbak. Dette bakrommet ble naturlig nok hovedbasen hvor informasjon og opplæring ble gitt og stedet hvor intervjuene ble gjennomført etter endt brukbarhetstest.

Figur 17 - Bakrommet på lageret der alle håndterminalene til butikken er plassert til høyre, mens jeg fikk sette opp min test håndterminal på skrivebordet til venstre.

Det ble satt opp en trådløs basestasjon sentralt i butikken som både håndterminalen (illustrert med ett blått antenneikon på toppen av det ene fryserommet i figur 18) og bakromspcen ble koblet til. Butikkområdet er ganske stort og vi var på forhånd usikker på om en basestasjon ville dekke hele butikken, derfor hadde vi med en D-Link Wireless Range Extender som ble satt opp i butikkområdet (se figur 18) og ble tilkoblet hovedbasestasjonen. Disse to trådløse enhetene ble plassert strategisk i butikken for å oppnå best mulig dekningsgrad overalt.

På den bærbare datamaskinen til venstre på figur 17 inne på et kontor på lageret (se figur 18) ble SOTI Pocket Controller 6 installert og brukt til å ta opp skjermbilde på håndterminalen. Pocket Controller fungerer ved at hovedprogramvaren ligger på en pc, mens det installeres en liten klient på håndterminalen som hovedprogramvaren kobler seg til. Det var helt nødvendig at det trådløse nettverket holdt seg stabilt for at den bærbare datamaskinen skulle være i stand til

å ta opp skjermbilde fra håndterminalen. Dette ble testet grundig ved å gå rundt overalt i butikken for å sjekke at forbindelsen holdt seg stabil. Under uttestingen av opptaksfunksjonen til Pocket Controller ble det oppdaget et kritisk problem. Det oppstod nemlig ett nettverkshikk når håndterminalen hoppet fra den ene basestasjonen og til den trådløse range extenderen eller motsatt. Nettverkshikket førte til at videostrømmen mellom den bærbare datamaskinen og håndterminalen ble avbrutt. Løsningen ble å kutte ut den trådløse range extenderen, og bare beholde en basestasjon sentralt plassert i butikken Det viste seg å fungere utmerket under uttestingen. Det trådløse nettverket ble også brukt av onlinefunksjonaliteten til Lindbak POS Mobile.



Figur 18 - Skisse som viser dagligvarebutikken i felten. De oransje sirklene indikerer hvor de ulike oppgavene ble løst.

6.4.2 Gjennomføring av brukbarhetstester

Ingen pilottest ble utført i butikken på grunn av mangel på kandidater som kunne være pilottester. Jeg gikk derfor selv gjennom alle oppgavene to ganger på samme måte som en testperson ville gjort for å avdekke feil i oppgavene og systemene. Dumas og Redish (1999:264-269) anbefaler alltid å utføre en pilottest med mindre testlederen er veldig erfaren og skal teste

et kjent produkt. Jeg har litt erfaring med å gjennomføre brukertester i forbindelse med fagene MMI(TDT4180) og design av grafiske brukergrensesnitt(IT3402), men kan det kvalifiserer ikke til å kalle meg selv erfaren. Kjennskap til produktet derimot har jeg mye av ettersom at jeg har utviklet det, og det burde derfor gå greit å starte rett på brukbarhetstestingen av personer uten pilottest. Til tross for at det ikke er optimalt.

Gjennomføringen av brukbarhetstestene ble gjort på dag to og tre, med henholdsvis to tester på dag to og tre tester på dag tre. Før jeg startet testingen på hver av testdagene gikk jeg gjennom og testet alle oppgavene for å tilse at alle systemene kjørte korrekt. Under kjøringen av brukbarhetstestene gikk jeg etter testpersonene med et videokamera for å filme deres bruk av håndterminalen og reaksjoner. Det ble benyttet en trådløs mikrofon som testpersonene hadde på seg under testen for å fange opp kvalitativ høyttenking. Svanæs et al. (2008) understreker viktigheten av god verbal lydkvalitet og anbefaler bruk av trådløs mikrofon til testpersonene. Videokameraet ble også brukt til å ta opp intervjuet etter testen.

Figur 19 - Illustrasjonsbilde fra feltbutikken som viser hvordan brukbarhetstestingen ble gjennomført

6.4.3 Problemer og utfordringer

Erfaringen fra felttesten viser en del problemer og utfordringer knyttet til en felttest. Miljøet har vi lite kontroll på og det er en effekt av å gjennomføre i en helt realistisk kontekst. I dette kapitlet skiller vi mellom disse med praktiske utfordringer og tekniske problemer.

De praktiske utfordringene gikk blant annet ut på at hele gjennomføringen var et enmannsshow. Det var ganske vanskelig å se og filme hva testpersonen trykket på og hva vedkommende gjorde på håndterminalen uten at jeg hang over skulderen og det kunne jeg selvfølgelig ikke gjøre. Newcomb og Pashley et al. (2003) som også har brukbarhetstestet i en dagligvarebutikk poengterer at det er vanskelig å overvåke en testperson sin interaksjon med grensesnittet på håndterminalen.

Jeg kunne bare glemme å ta feltnotater mens jeg filmet, og det medførte at jeg ikke fikk notert interessante funn underveis som jeg kunne spørre ut om etter testen. Mennesket er ikke god til å huske slike ting og når en i tillegg har nok med å passe på at gjennomføringen går på skinner sier det seg selv at en del oppfølgingsspørsmål kan gå tapt.

En annen praktisk utfordring var at jeg ikke hadde noe eget rom til disposisjon for å gjennomføre intervjuene. Bakromskontoret på figur 17 ble brukt som intervjurom. Dette rommet ble brukt til mange oppgaver i den daglige driften og jeg kunne derfor ikke stenge døren. Det gikk folk inn og ut under brukbarhetstestingen og intervjuet. Til tross for det gikk gjennomføringen av intervjuene der relativt greit. Jeg merket ikke at de holdt tilbake noe, men det er selvfølgelig ikke optimalt. Det er vanskelig å få det 100 % optimalt i felten, og det er gjerne en av utfordringene med å bedrive felttester.

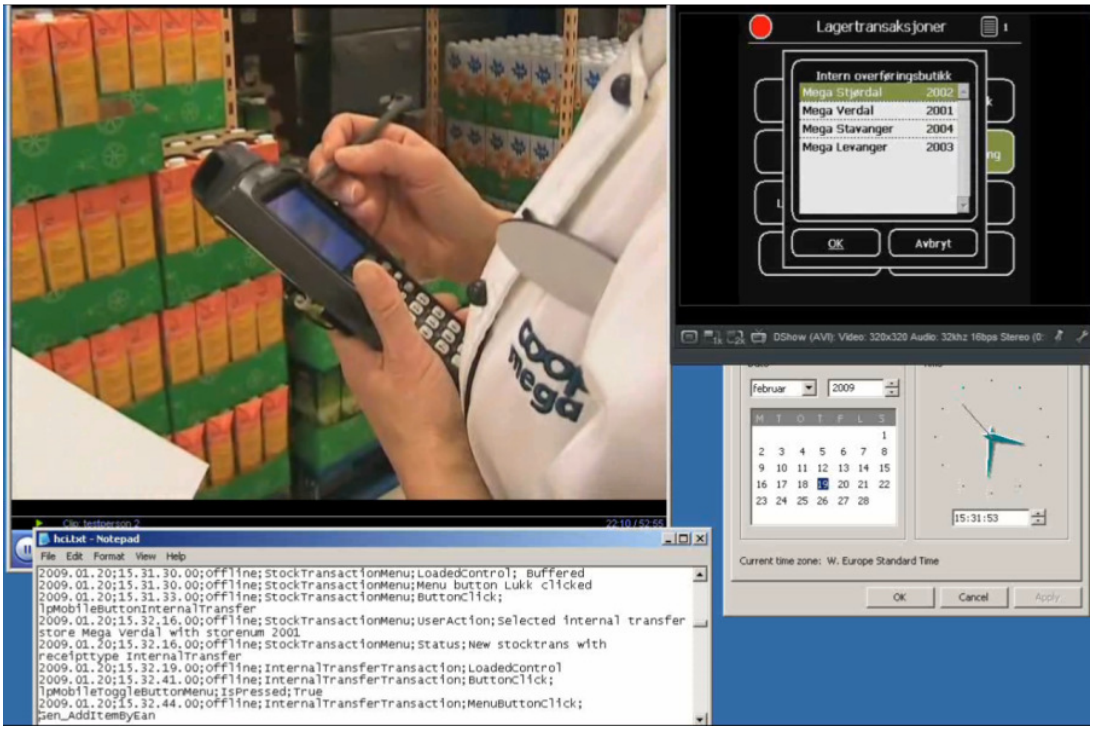
Høyttenkning for testpersonene i butikken var også en liten utfordring. Testpersonene var ikke helt bekvem med hele situasjonen. Jeg måtte minne de på å tenke høyt en del ganger. Det virket som at de syntes det var litt ubehagelig. Jeg ønsket derfor ikke å mase for mye, slik at situasjonen ble mer ubehagelig. Høyttenkningen foregikk tross alt blant kunder og kollegaer.

Det var en del ulike tekniske problemer som dukket opp med forskjellig alvorlighetsgrad, men totalt sett kan felttestingen oppsummeres som vellykket. Det første problemet som dukket opp under intervjuet av testperson to. Jeg hadde glemt å sette videokameraet til ladning etter endt brukbarhetstest. Konsekvensen av det var at batteriet gikk tom cirka fem minutter ut i intervjuet. Dette oppdaget jeg dessverre ikke underveis, og jeg tok heller ikke notater fordi jeg tenkte at det holdt i massevis å transkribere opptaket i etterkant. Problemet ble oppdaget rett etter at intervjuet var ferdig, og jeg endret da tidsplanen litt slik at videokameraet fikk mer tid på å lade mellom testene, og at jeg passet også på å sette det på ladning før intervjuet startet.

Det andre tekniske problemet viste seg å være den trådløse mikrofonen. Under brukbarhetstestingen av testperson fire var det tidvis masse skurring og det var umulig å høre hva som ble sagt i ettertid. Dette kan potensielt ha ført til at noen brukbarhetsproblemer ikke ble oppdaget, ettersom mange av brukbarhetsproblemene oppdages ved høyttenkning (Ebling and John 2000). Det var også en del lydtap under intervjuet med testperson fem. Alt opptak etter ti minutter av intervjuet var ubrukelig. På opptakene hørtes det ut som en kontaktfeil, og den trådløse mikrofonen indikerte aldri at batteriet var dårlig. Jeg hadde testet mikrofonen på forhånd, men da merket jeg ikke noe til kontaktproblemet. Det skal sies at mikrofonen fungerte i cirka 90 % av tiden.

Hos testperson fire oppstod det et problem med programvaren Lindbak POS Mobile, og det førte til at vi ikke fikk testet oppgave 10 og 11 på vedkommende. Dumas og Redish (1999:281-286) påpeker at vi ikke må glemme testpersonen når det oppstår et problem med programvaren, og at testpersonen bør bli satt til side og forklart hva som skjer. Jeg passet på å forklare situasjonen der og da mens jeg brukt et par minutter på å konstatere at det var ingenting jeg kunne gjøre. Testpersonen hadde gjort riktig og det så jeg selv under utførelsen. Problemet skyldtes en konfigurasjonsfeil og ble fikset før neste person skulle testes.

Skjermopptaket fra håndterminalen med Pocket Controller pro 6 fungerte utmerket bortsett fra når jeg brukbarhetstestet med testperson to. Etter 30min ut i testen skjedde det et eller annet med det trådløse nettverket som stoppet opptaket. Dette medførte at vi mangler skjermopptak fra de siste 15 minuttene av testen, og det var tilnærmet umulig å se hva testpersonen faktisk gjorde med det håndholdte videokameraet. Heldigvis hadde jeg på forhånd laget detaljert logging i programmet som nevnt i kapittel 4.3.6, og led derfor ingen datatap av den grunn. Jeg restartet den trådløse ruteren rett før de tre gjenværende testene som et forebyggende tiltak.



Figur 20 - Skjermbilde fra den synkroniserte videostrømmen med loggen nederst og skjermopptaket oppe i høyre hjørne.

Det har tidligere blitt nevnt at det var vanskelig å filme all interaksjon med håndterminalen. Det gjorde ikke saken noe bedre at det viste seg å være vanskelig å lage en videofil (se figur 20 som viser et skjermbilde fra en sammensatt feltfilm), der skjermopptaket og tapen fra DV kameraet ble synkronisert. Det viste seg nemlig at skjermopptaket fra SOTI Pocket Controller gikk en del raskere enn DV filmen. Det gikk med mye tid på å finne ut nøyaktig hvor mye fortere skjermopptaket gikk, slik at hastigheten kunne settes ned. Resultatet ble ikke helt perfekt, men

det ble godt nok til at det var mulig å analysere den sammensatte filmen. Skjermbilde av den sammensatte feltvideoen i figur 20 stammer fra testperson to og er like før skjermopptaket stoppet, og vi var helt avhengig av loggen som vises nederst i skjermbildet. I de andre felttestene trengte vi ikke loggen, og de sammensatte feltvideoene består bare av videoen fra DV kameraet og skjermopptaket.

6.4.4 Observasjon av daglig bruk av håndterminaler

For å kunne få en større innsikt i hvordan håndterminaler brukes til dagelig valgte jeg å observere hvordan disse brukes i felten. På en måte er testpersonene i felt ganske homogene med tanke på at de benytter samme håndterminal og følger samme rutiner i daglig bruk. Nettopp derfor kan det være til hjelp å se hvordan de benytter håndterminalen for kanskje å kunne forstå hvorfor visse ting gjøres på en spesiell måte under brukbarhetstestingen i felten.

> *One advantage of observing users doing their own tasks is that one often finds that they use the software in unexpected ways that one would (by definition) not have sought to test in a planned laboratory experiment*. (Nielsen 1993:208)

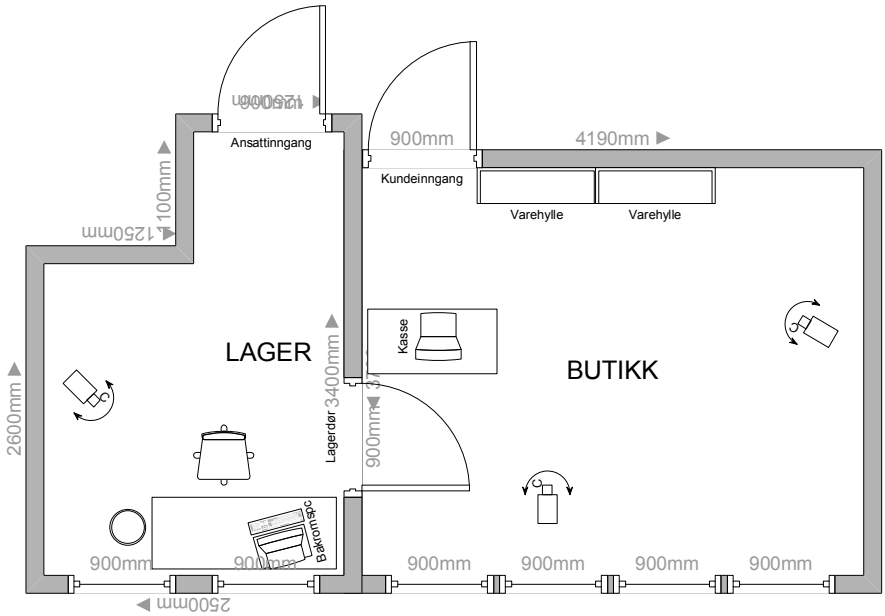
Jeg benyttet meg av åpen observasjon sammen med en av testpersonene jeg hadde fått god kontakt med i løpet av brukbarhetstestingen. Nettopp det gjorde det lett for meg å kunne stille spørsmål underveis mens jeg observerte om hvorfor vedkommende handlet som hun gjorde. Jeg fikk blant annet vite at de skannet stort sett bare skannet hylleetikettene og ikke varene. Når de hadde fullført en operasjon gikk de tilbake til bakrommet og satte håndterminalen i sleden.

6.5 Fullskala brukbarhetstest i laboratorium

I dette kapitlet blir alle elementer rundt gjennomføringen av fullskalametoden omtalt. Vi vil gå gjennom oppsetningen av laboratoriet og si noe om hvorfor vi har valgt å innføre de ulike elementene av realisme. Før vi begynte å teste på de rekrutterte personene ble det gjennomført en pilottest, uten at den oppdaget noen problemer med testoppsettet. I motsetning til feltmetoden, var det ingen problemer av betydning knyttet til gjennomføring av en fullskala brukbarhetstest i laboratorium.

6.5.1 Oppsett av lokasjon

Fullskalatesten ble gjennomført i brukbarhetslaboratoriet på NSEP(Norsk senter for elektronisk pasientjournal) som er et topp moderne laboratorium i brukbarhetssammenheng. Laboratoriet har skinner i taket med veggmoduler som kan flyttes slik at testrommet kan bygges opp slik vi vil. Til å ta opp høyttenkningen til testpersonene benyttes det trådløse mikrofoner på den enkelte person. I takplatene har vi tre videokamera (se figur 21), som er plassert for å filme all interaksjon. Kameraene blir styrt fra kontrollrommet (se figur 23).



Figur 21 - Skisse over oppsettet for fullskalatest i laboratorium

Oppdelingen med butikkareal og lagerareal er vektlagt som viktig i oppbyggingen av en realistisk butikk i brukbarhetslaboratoriet som vi ser i figur 21. Denne oppdelingen ligner på butikken i felten, og den vil kunne hjelpe testpersonene med å leve seg inn i rollen som butikkansatt. Det ble satt opp to varehyller med ekte varer, og det ble laget etiketter på hyllene for hver av varene(se figur 22). På figur 22 vises kasseapparatet som ble satt opp. Det kjører Lindbak POS programvaren og den ble som nevnt brukt til å lære testpersonene å tenke høyt. Ingen av oppgavene går ut på at kassen skal benyttes, og derfor er den egentlig bare en kulisse. Bakromspcen som er avbildet på figur 21 er bare en kulisse den også som ikke ble bruk i forbindelse med brukbarhetstestingen.

Figur 22 - Illustrasjonsbilde som viser en butikkansatt som selger en vare med håndterminalen til en kunde. Til venstre i bilde er lagerdøren, mellom personene står det en kasse og til høyre har vi varehyllene. Fra venstre står Kenneth Devik og Terje Røsand.

Under gjennomføringen av de enkelte testene satt vi på kontrollrommet (se figur 23) og observerte testpersonene. De to hovedgrunnene til det var å for å kunne se skjermopptaket fra håndterminalen direkte og kunne derfor ta mer årsaksberettigete notater. Den andre grunnen var for å slippe og stresse testpersonene unødvendig med nærvær i testlokalet. Til å styre kameraene fikk vi hjelp av Terje Røsand.

Figur 23 - Kontrollrommet i brukbarhetslaboratoriet på NSEP. På bilde, Overingenør Terje Røsand.

6.5.2 Pilottest

Selv om systemmiljøet rundt håndterminalen og rutinene fikk testet seg i felten ble det gjennomført en pilottest i brukbarhetslaboratoriet dagen før de rekrutterte personene ankom. En doktorgradsstipendiat som til daglig sitter på NSEP ble brukt som pilottester. Hun fikk samme type opplæring som de butikkansatte hadde fått i feltbutikken og fikk prøve seg på de samme oppgavene.

Det viste seg fort at hun ikke var en kvalifisert testperson, fordi hun slet mye med mange av fagbegrepene som ble benyttet i oppgaveteksten og i programmet. Hun slet også med å skanne varer. Dette understreker bare det Dumas og Redish (1999:135-141) sier om at testpersonene må representere reelle brukere. Bortsett fra at pilottestpersonen ikke var representativ ble det ikke oppdaget noen andre problemer.

6.5.3 Problemer og utfordringer

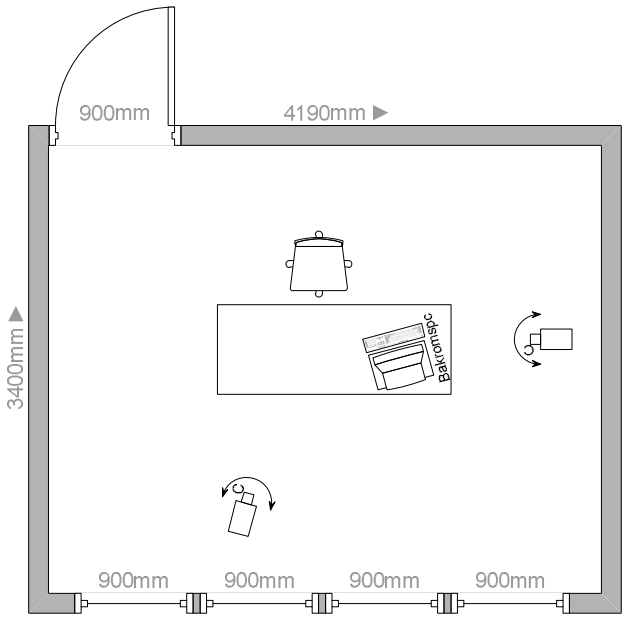
Det tekniske utstyret fungerte utmerket. Det eneste problemet vi hadde ved gjennomføringen av brukertestene var at testpersonene noen ganger hadde en tendens til å stille seg med ryggen til alle kamera slik at vi ikke fikk filmet interaksjonen med håndterminalen. Ved et par tilfeller måtte vi be testpersonen flytte seg litt slik at kameraet fikk filme bedre. Etter tre tester i brukbarhetslaboratoriet flyttet vi det ene kameraet litt for å unngå dette problemet.

6.6 Desktop brukbarhetstest i laboratorium

I den enkleste og minst realistiske metoden ble alt av testoppsett gjort billigst mulig. Vi vil gå nærmere inn på hvordan laboratoriet ble satt opp i desktopmetoden, og se litt på hva slags problemer metoden medførte.

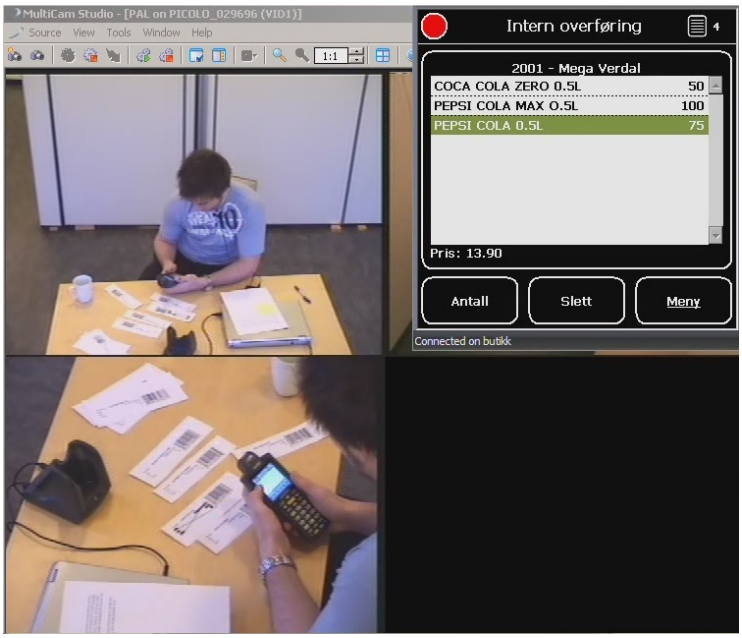
6.6.1 Oppsett

I desktopmetoden ble alt i laboratoriet strippet ned til den enkleste og billigste måten å gjennomføre en brukbarhetstest på. Vi benyttet fortsatt alt teknisk tilgjengelig utstyr i brukbarhetslaboratoriet på NSEP som kamera i taket og trådløs mikrofon. Det ble benyttet bare ett rom til testen som vist på figur 24, der testpersonen satt på en stol inntil ett skrivebord under hele testen. Til å fange opp ansiktsreaksjoner ble det satt opp et kamera rett foran skrivebordet og kameraet på siden av skrivebordet ble brukt til å fange opp interaksjonen med håndterminalen. På figur 26 vises et bilde som illustrerer metoden.



Figur 24 - Skisse over desktoptestoppsett i brukbarhetslab

I stedet for ekte varer som ble brukt både i felt og fullskalametodene, ble det laget papirlapper med nødvendig vareinformasjon. Informasjonen som stod på en slik lapp var navnet på varen, EAN-koden, en skannbar strekkode og hvor mange eksemplarer av varen som er i butikken. Eksempler på slike varelapper er avbildet i figur 40 på side 94. På figur 25 kan vi se varelappene ligge utover bordet under en test.



Figur 25 - Skjermbilde fra videostreamen under en desktoptest.

Det var en liten utfordring å gjenskape flere av elementene som trengtes for å kunne fullføre de ulike oppgavene liten realisme. Til å illustrere hvor mange varer som fantes, ble det skrevet på et antall på varelappen. Varelappen med utkladdet informasjon rett ved skannerhode på toppen av håndterminalen i det nederste bilde på figur 25 skal forestille en manglende hylleetikett. Varemottaket ble illustrert ved at testpersonene fikk tre varelapper limt fast på selve oppgavearket. Til tross for veldig enkle løsninger, kan bare tre problemer knyttes til manglende realisme, og det skal vi se nærmere på i resultatkapitlet.

Figur 26 - Illustrasjon av desktopmetoden. Nederst på bilde er lappene som skal forestille varer, mens på den andre siden er sleden håndterminalen dokkes i.

6.6.2 Problemer og utfordringer

Desktoptestingen forløp uten noen problemer av noe slag. Det eneste vi kan pirke på er at opptaket av den ene testpersonen ikke ble startet før litt ut i oppgave 3 og det var en liten menneskelig glipp. Det ble funnet bare ett kosmetisk brukbarhetsproblem blant de 14 andre testpersonene når de løste oppgave 1 og 2. Sjansen er derfor ganske liten for at vi har gått glipp av noe. Loggen som programmet genererer ble brukt til å bekrefte at testpersonen klarte å løse de to oppgavene. Tidsmålingen for gjennomføringen av de tre første oppgavene mangler på grunn av dette. Det ble vurdert å se på tiden ut i fra loggen, men den sier ingenting om når testpersonen starter en oppgave og sier seg ferdig med den.

Det var litt uheldig bare å ha kamera i taket. På figur 25 vises ett skjermbilde fra videoen som ble laget ved en gjennomføring. Den viser at oppsettet fungerte relativt bra, men at det tidvis var vanskelig å få med ansiktet til testpersonen. De hadde en tendens til å ha håndterminalen nær egen kropp og se ned på den. Det medfører at det er vanskelig å legge merke til ansiktsuttrykk.

6.7 Bearbeiding av data

Råfilmene fra alle testene ble sett gjennom totalt tre ganger. Den første gangen ble benyttet til å transkribere ned hver eneste detalj. Ut fra de transkriberte dataene ble det laget en liste med alle problemer og interessante hendelser. På bakgrunn av den listen var det tydelig at vi manglet noen observasjoner, og vi gikk derfor gjennom filmene enda en gang for å se saker vi visste skulle være der men som ikke hadde blitt med i første omgang. Vi merket oss også et 20-talls nye problemer som var litt vanskeligere å oppdage, men uansett viktige funn. Den tredje gjennomgangen ble brukt til å ta tiden testpersonene brukte på oppgavene.

7 Resultater

Gjennomføringen av 15 brukbarhetstester i tre metoder gir mye data, og de blir strukturert og presentert i dette resultatkapitlet. Vi vil først av alt oppsummere personkarakteristikkene til testpersonene som deltok i testene. Deretter vil vi se på interaksjonen med håndterminalen, før vi går nærmere inn på resultatene knyttet til anvendbarhet, effektivitet og tilfredshet. Gjennomgangen av brukbarhetsproblemene er skilt ut i et eget kapittel, og der ser vi hvor de ulike problemene har oppstått. Det blir også redegjort for alvorligheten til problemene og hva slags type problemer det er snakk om.

7.1 Testpersoner

Resultatene fra spørreskjemaet i vedlegg IV lager en basis for kartleggingen av testpersonene og hva slags bakgrunnserfaring de sitter inne med. På bakgrunn av svarene fra det spørreskjemaet, er det laget en liten oppsummering av hva vi mener er interessant blant testpersonene i de forskjellige metodene.

7.1.1 Testpersonene i felttesten i butikk

Det var en overvekt med kvinner i alderen 40-50 år som ble testet i felt (gjennomsnittsalder på 39.2, med standardavvik på 9.04). Alle hadde varebestilling som en del av sin hovedoppgave, og i den forbindelse benyttet de en håndterminal. Deres datakunnskaper angir de til å være litt under middels (2.8). Det er en yngre mann på 25 (TP3) som skiller seg ut med over middels datakunnskaper og lavere alder.

Felttestpersoner	**TP 1**	**TP 2**	**TP 3**	**TP 4**	**TP 5**
Kjønn	Kvinne	Kvinne	Mann	Kvinne	Kvinne
Alder	40	50	25	39	42
Varebestilling	X	X	X	X	X
Varepåfylling	X		X	X	
Rydding	X			X	
Kasse		X			
Leverandøroppfølging			X		
Benytter håndterminal i jobben?	Ja	Ja	Ja	Ja	Ja
Benytter du data regelmessig?	Ja	Nei	Ja	Ja	Nei
Datakunnskaper på en skala fra 1-5 (5 er best)	3	2	4	3	2

Felttestpersoner	TP 1	TP 2	TP 3	TP 4	TP 5
Benytter du mindre dingser?	Nei	Nei	Ja	Nei	Ja
Har du mobil?	Ja	Ja	Ja	Ja	Ja

Tabell 3 - Tabell med bakgrunnsinformasjon om testpersonene i felttesten

7.1.2 Testpersonene i fullskala brukbarhetslaboratorium

Blant testpersonene i fullskalatesten er det en lett blanding mellom kjønnene og alle er i 20-årene (22.2 år i gjennomsnitt, standardavvik på 2.77). Alle bortsett fra en benytter håndterminal i sin jobb og alle har bestilt varer med den. Det er to personer som ikke er studenter. TP9 er en av de som ikke er student, og faktisk den eneste blant de som ble forsøkt rekruttert direkte fra dagligvarebutikker. En annen sak som kan være verdt å merke seg blant disse fem er at TP9 og TP10 har jobbet i en Coop dagligvarebutikk, og det betyr at de har erfaring med akkurat samme håndterminal og program som de fra feltmetoden. Datakunnskapene til testpersonene mener de selv at ligger over middels.

Fullskalatestpersoner	TP 6	TP 7	TP 8	TP 9	TP 10
Kjønn	Kvinne	Mann	Kvinne	Mann	Mann
Alder	21	20	22	27	21
Student	Ja	Ja	Nei	Nei	Ja
Studie	Nanotek.	Energi og miljø	(ferdigutd. sykepleier)		Elektronikk
Årstrinn	2	2			1
Butikkerfaring	Byggmaker, bartender	Mix, Rimi	Narvesen, Shell	Coop Prix	Coop Mega
Varebestilling	X		X	X	X
Varepåfylling	X	X	X	X	X
Varetelling	X	X			
Kasse	X	X	X	X	X
Benytter håndterminal i jobben?	Ja	Nei	Ja	Ja	Ja
Benytter du data regelmessig?	Ja	Nei	Ja	Ja	Ja
Datakunnskaper på en skala fra 1-5 (5 bes er best)	5	5	4	4	4

Fullskalatestpersoner	TP 6	TP 7	TP 8	TP 9	TP 10
Benytter du mindre dingser?	Ja	Ja	Nei	Ja	Ja
Har du mobil?	Ja	Ja	Ja	Ja	Ja

Tabell 4 - Tabell med bakgrunnsinformasjon om testpersonene i fullskalatesten

7.1.3 Testpersonene i desktoptest i brukbarhetslaboratorium

Alle testpersonene i desktopmetoden var bare unge menn tidlig i 20-årene med gjennomsnitt på 22.2 år (standardavvik på 1.64) og samtlige studerer. Tilfeldig nok er det bare TP 11 som har benyttet en håndterminal i forbindelse med butikkjobben, og det var bare til å foreta noen svinnregistreringer. Alle har hovedsakelig sittet i kassen, og det er mest trolig derfor såpass få har benyttet en håndterminal i jobbsammenheng. Datakunnskapen er i følge dem selv veldig god, og det skyldes blant annet at tre av personene studerer IKT (informatikk- og kommunikasjonsteknologistudier).

Desktoptestpersoner	TP 11	TP 12	TP 13	TP 14	TP 15
Kjønn	Mann	Mann	Mann	Mann	Mann
Alder	21	21	25	22	22
Student	Ja	Ja	Ja	Ja	Ja
Studie	Elektronikk	Kom. Tek.	Informatikk	Kom. Tek.	Energi og miljø
Årstrinn	2	2	5	2	2
Butikkerfaring	Ultra	Platebutikk	Coop OBS	Kiwi	Ica Maxi
Varebestilling		X			
Varepåfylling				X	
Varetelling		X			
Kasse	X	X	X	X	X
Benytter håndterminal i jobben?	Ja	Nei	Nei	Nei	Nei
Benytter du data regelmessig?	Ja	Ja	Ja	Ja	Ja
Datakunnskaper på en skala fra 1-5 (5 er best)	4	5	5	5	4
Benytter du mindre dingser?	Ja	Ja	Ja	Ja	Ja
Har du mobil?	Ja	Ja	Ja	Ja	Ja

Tabell 5 - Tabell med bakgrunnsinformasjon om testpersonene i desktoptesten

7.1.4 Oppsummering av testpersonene

Alle testpersonene hadde tilstrekkelig relevant butikkerfaring til å skjønne oppgavene og forstå de fleste fagbegreper som ble benyttet. Viktigheten av at testpersonene må ha butikkerfaringen kom fram ved gjennomføring av pilottest i laboratorium med en person uten butikkerfaring, som nevnt i kapittel 6.5.2.

En del av testpersonene var litt usikre på hva noen av fagbegrepene betydde i forhold til deres erfaring. Til tross for det skjønte de det etter hvert. Dette ble påpekt både under og etter testen, og kan oppsummeres med et sitat fra TP12 angående fagspråk:

> *Det jeg ofte lurte på var, det her kalles for det. Er det samme som vi kalte for noe annet, på en måte?* (TP12)

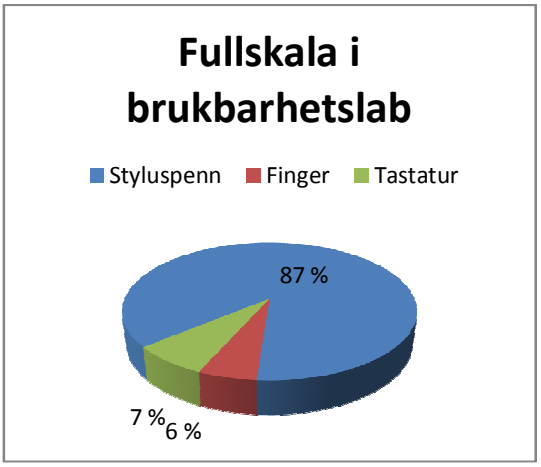
Det er noen ulikheter mellom testpersonene i de forskjellige metodene som er av interesse. Dersom vi ser på gjennomsnittsalderen til personene i fullskala- og desktopmetoden er begge 22.2 år, mens den i feltmetoden er 39.2 år. Det er statistisk signifikant forskjell i alderen mellom personene i feltmetoden og både fullskala- ($t=4.02$, $p < 0.02$) og desktopmetoden ($t=4.13$, $p < 0.02$). Mellom fullskala- og desktopmetodene er det ingen statistisk forskjell i alder.

En annen forskjell som er interessant å se på er testpersonene sin vurdering av egne datakunnskaper på en skala fra 1-5 (hvorav 1 er dårligst og 5 er best). Gjennomsnittsverdien for feltmetoden er 2.8 med standardavvik på 0.84, fullskalametoden har snitt på 4.4 med standardavvik på 0.55 og desktopmetoden har 4.6 i gjennomsnitt med 0.55 i standardavvik. Også her er det statistisk signifikant forskjell mellom metodene felt sett i mot fullskala ($t=3.57$, $p < 0.01$) og desktop ($t=4.02$, $p < 0.01$), mens det er ingen signifikant forskjell mellom fullskala og desktop ($t=0.58$, $p > 0.5$).

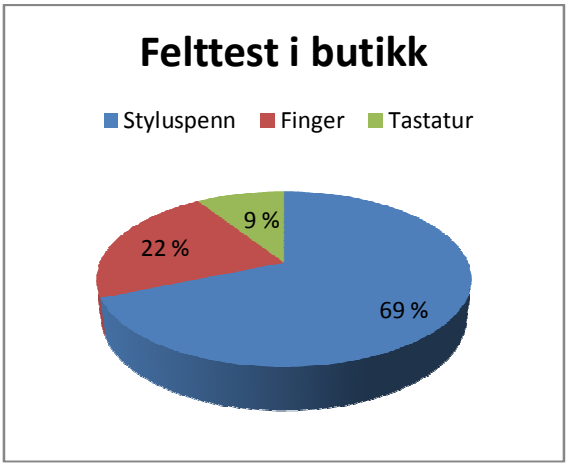
TP3 skiller seg litt ut blant testpersonene i felten. Alderen til personen er 25 år og er 14 yngre enn snittalderen på 39 år. Den nest yngste er faktisk 39 år. Dette er en stor grunn til at standardavviket på alderen til testpersonene i felten er på 9. Det andre TP 3 skiller seg ut på blant feltpersonene er datakunnskapen. Han har angitt 4 og det er med på å dra opp standardavviket med 0.3 prosentpoeng. Med andre ord er TP3 ganske lik profilen til testpersonene i fullskala- og desktoptesten.

7.2 Interaksjonen med håndterminalen

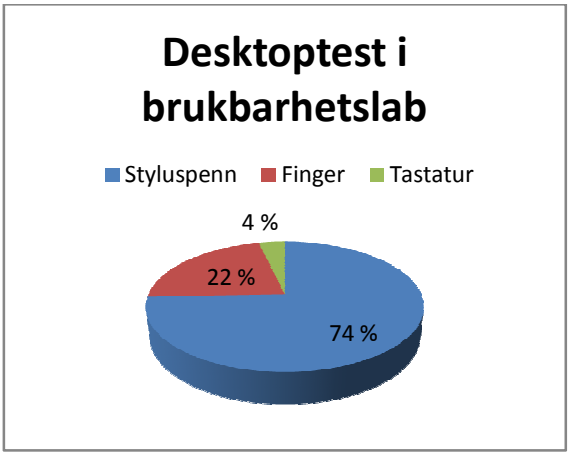
Under gjennomgangen av videodataene ble det notert ned hva slags interaksjonsmetode personene benyttet. De tre interaksjonstypene håndterminalen støtter er; syluspenn (til vanlig festet på baksiden av enheten, se figur 3), selve fingeren direkte på skjermen eller ved å bruke tastaturet på håndterminalen (med tastatur menes det de fysiske hardwareknappene og ikke på-skjermstastaturet). Det ble notert ned hva slags interaksjonstype hver testperson benyttet mest i de enkelte oppgavene. Bruken av de forskjellige interaksjonstypene vises prosentvis i kakediagrammene i figur 27-29.



Figur 27 - Kakediagram med interaksjonstypebruk i fullskala laboratorium



Figur 28 - Kakediagram med interaksjonstypebruk i felt



Figur 29 - Kakediagram med interaksjontypebruk i desktoptest

Ut fra figurene ser vi at styluspennen er mest brukt i alle metodene. Dette til tross for at de fleste GUI-komponentene er laget store for å kunne bruke fingrene. Statistikken viser at det er ingen signifikant forskjell på bruken av interaksjonstypene i de ulike metodene. Den største forskjellen er bruken av styluspenn mellom felt- og fullskala metodene ($p > 0.3$).

Seks av testpersonene benyttet konsekvent bare styluspennen i alle oppgavene. Flere av testpersonene som benyttet forskjellige interaksjonstyper kommenterte underveis at de foretrakk å benytte styluspennen. En situasjon som er verdt å merke seg er at TP12 benyttet fingeren helt fram til oppgave 8. I den oppgaven måtte han foreta et varesøk. Han åpnet ABC-tastaturet og kommenterte at tastene var veldig små, og tok derfor opp styluspennen. Fra og med den oppgaven benyttet TP12 styluspenn på resten av oppgavene. Denne hendelsen er bare tydelig hos TP12, og kan derfor ikke ses på som en trend. Grunnen til det er at de andre benyttet styluspennen i mye større grad fra starten av.

I forbindelse med interaksjon mot håndterminalen, bør det nevnes at den har en god passform i hånden. Når testpersonene skulle skanne en vare, hadde de håndterminalen i den ene hånden og varen i den andre. Ingen opplevde noen problemer med å skanne varene. Skannerknappene er lett tilgjengelig når håndterminalen opereres med en hånd, men det er vanskelig å trykke på skjermen ved bruk av en hånd. Alle testpersonene benyttet begge hender når de trykket på skjermen.

I felten gikk alle testpersonene bortsett fra TP5 og trykket på håndterminalen når de skulle gå en lengre distanse, som for eksempel fra lagret til butikkområdet eller motsatt. Dette ble ikke oppdaget i fullskala laboratoriet. De trykket hovedsakelig seg gjennom menyene for å gjøre seg kjent og bekrefte at de hadde funnet fram til riktig funksjon. De hadde ikke noen problemer med interaksjonen selv om de gikk, og det skyldtes nok store knapper. Dette står i kontrast til Brewster (2002) som sier at brukbarheten ble signifikant redusert når testpersonene gikk. Grunnen til at vi ikke oppdaget noe lignende var fordi testpersonene våre ikke løste oppgavene når de gikk, de bare utforsket menyene og funksjonene.

7.3 Anvendbarhet

Det ble nevnt i kapittel 2.2.1 at anvendbarhet kan måles ved flere forskjellige måter. I dette kapitlet vil alle disse målene bli presentert. Vi vil se på hvor mange prosent av oppgavene de enkelte testpersonene har fullført og antall sammenbrudd, vise antall brukbarhetsproblemer hver enkelt har oppdaget og se hvor mange ganger hver enkelt har fått hjelp (assist). Selv om brukbarhetsproblemene hører hjemme under dette kapitlet, er de i stedet skilt som et eget kapittel 7.6 fordi gjennomgangen av brukbarhetsproblemene er stor og grundig.

Common industry format(CIF) (ISO 2006) gir noen veiledninger for hvilke data som bør presenteres i en brukbarhetstestrapport og konkrete eksempler på hvordan noen av disse dataene kan presenteres. Nedenfor i tabell 6-8 oppsummeres noen resultater fra hver enkelt metode.

Tabell 6-8 viser prosentvis hvor mye hver enkelt klarte å fullføre på alle oppgavene. Det er viktig å ha klart for seg at fullføringsprosenten sier noe om hvor korrekt løsingen av oppgavene har vært. Det har for eksempel blitt trukket 10 % for hver detaljfeil en person har gjort, og mer for

grovere feil. Fullføringsprosenten sier ikke direkte noe om hvor mye funksjonalitet en testperson har testet ut, selv om det henger litt sammen.

CIF sier at fullføringsraten skal deles opp mellom hvor mange prosent hver enkelt har fullført både med- og uten hjelp. Det har vært bevisst fra vår side å være restriktiv med hjelping. Vi har derfor valgt å gi mindre hint, og ikke indirekte gi de løsningene. Dersom en testperson har vært på jordet, har vi for eksempel bedt vedkommende om å lese oppgaveteksten en gang til eller lignende som hjelp.

I tabellene i underkapitlene for hver metode er den tiden testpersonene brukte på alle oppgavene. Tiden er oppgitt som effektiv tid i minutter. Mindre pauser mellom oppgavene, tiden feltpersonene brukte på å gå lengre distanser og tiden testpersonene brukte på å lese oppgavetekstene er ekskludert. Problemkolonnen sier hvor mange unike brukbarhetsproblemer den enkelte testperson oppdaget under hele brukbarhetstesten. Den siste kolonnen tar for seg hvor mange ganger testpersonene har fått hjelp underveis.

7.3.1 Feltresultater

Testperson	Uten hjelp (% fullført)	Med hjelp (% fullført)	Oppgavetid (min)	Problemer	Hjelp
1	62	22	32,7	19	5
2	35	18	37,4	18	4
3	89	0	23,6	7	0
4*	70	0	23,8	10	0
5	45	0	31,8	13	2
Snitt	60,2	8	29,9	13,4	2,2
Standard-avvik	21	11	6,0		
Min	35	0	23,6	7	0
Maks	89	22	37,4	19	5

*** Pga programvarefeil kunne ikke TP4 løse oppgave 10 og 11.**
Tabell 6 – Anvendbarhetsresultater fra feltmetoden.

Det som skiller seg ut blant testpersonene i felt som vist i tabell 6 er at de i gjennomsnitt har en del lavere fullføringsgrad enn testpersonene i de andre metodene med 60% uten hjelp og 68% inkludert hjelp. Hvis vi isolert sett ser på fullføringsgraden uten hjelp blant fullskala- og desktoptesterne i tabell 7 og 8 er den laveste fullføringsgraden TP15 med 78%. Alle felttesterne bortsett fra TP3 ligger godt under 78%. Fullføringsgraden er ganske sprikende, og det medfører også et betydelig større standardavvik enn i de andre metodene.

Under gjennomføringen av felttesten stilte testpersonene gjerne mange åpne spørsmål underveis. Dette til tross for at de på forhånd hadde fått beskjed om at vi ikke kunne hjelpe dem. Bakgrunnen for det har nok litt med at jeg var i umiddelbar nærhet i felten, i motsetning til i laboratoriet. De fleste spørsmålene besvarte jeg ikke. Grunnen til at gjennomsnittet for hjelpingen er litt høyere i felten, er rett og slett fordi testpersonene der slet mer med å fullføre oppgavene.

En annen viktig bemerkelse under testingen i felten, er at mens testpersonene løste oppgavene ble de avbrudd av både kunder og medansatte. De ble avbrutt hele tolv ganger og samtlige fem testpersoner ble avbrutt minst en gang. Avbryterne la rett og slett ikke merke til at testpersonen ble filmet. Avbruddene var alt fra kjappe spørsmål som tok fem sekunder å besvare til lengre spørsmål som tok opptil tre minutter. De fleste spørsmålene gikk ut på at en kunde lurte på hvor en bestemt vare befant seg. Det medførte ofte at testpersonen gikk til varelokasjonen for å hjelpe kunden.

Til tross for at testpersonene ble avbrutt underveis, påvirket ikke det gjennomføringen av oppgavene. Etter å ha hjulpet en kunde fortsatte de med oppgaven de holdt på, med uten å lese seg opp igjen på hva oppgaven gikk ut på. Alle situasjonene rundt avbrytelsene har blitt analysert grundig, og det var bare en gang at det skjedde noe som kan skyldes avbrytelsen. Etter at TP5 hadde hjulpet en kunde skannet hun feil vare (feil type appelsinjuice). Strengt talt er det bare en fillesak. Alle taklet avbrytelsene forbausende bra, i forhold til det vi kanskje skulle tro. Grunnen til det er trolig at de er vant til å bli avbrutt mange ganger daglig mens de jobber ute i butikken.

7.3.2 Fullskalaresultater

Testperson	Uten hjelp (% fullført)	Med hjelp (% fullført)	Oppgavetid (min)	Problemer	Hjelp
6	86	0	34,9	13	1
7	98	0	29,7	12	0
8	81	0	29,8	15	1
9	81	9	26,9	12	1
10	89	9	24,0	8	1
Snitt	86,9	3,6	29,1	12	0,8
Standard-avvik	7,1	5,0	4		
Min	80,9	0,0	24	8	0
Maks	98,2	9,1	35	15	1

Tabell 7 – Anvendbarhetsresultater fra fullskalametoden.

Testpersonene i fullskalatesten er de som har størst fullføringsgrad av oppgavene. Det kan være en medvirkende årsak til at de oppdaget minst problemer i gjennomsnitt. Tabell 7 viser at fullføringsgraden uten hjelp er på hele 87 % og 91 % inkludert hjelp. Standardavviket på de forskjellige verdiene er også relativt lav i forhold til resultatene fra felttesten. Det viser at det ikke er all verden til forskjeller på fullføringsgraden mellom testpersonene i denne testen.

7.3.3 Desktopresultater

Testperson	Uten hjelp (% fullført)	Med hjelp (% fullført)	Oppgavetid (min)	Problemer	Hjelp
11	82	0	39,7	16	0
12	95	0	24,5*	14	0
13	88	0	47,0	17	0
14	84	0	32,0	19	0
15	78	9	47,7	14	1
Snitt	85,5	1,8	38,2	16	0,2
Standard-avvik	7	4	10		
Min	78	0	24	14	0
Maks	95	9	48	19	1

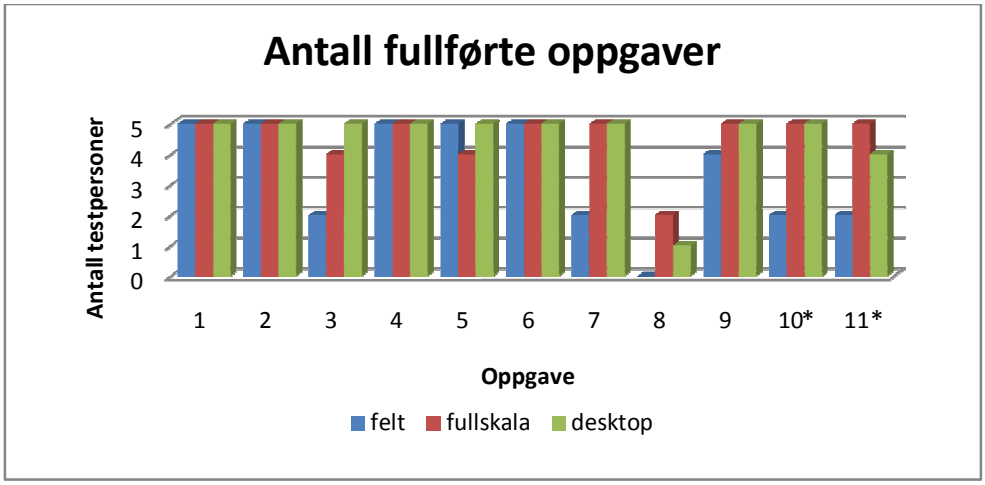
*** Mangler tiden for oppgave 1-3 og tiden er derfor noen minutt mindre enn den ellers ville vært.**
Tabell 8 – Anvendbarhetsresultater fra desktopmetoden.

Fullføringsgraden til testpersonene i desktoptesten viser en fullføringsgrad som er rett under fullskalatesten. De har en fullføringsgrad på 86 % uten hjelp og 87 % inkludert hjelp i følge tabell 8. Det ble gitt hjelp bare en gang blant desktoptesterne, og det er relativt lavere enn både i fullskalatesten og felttesten. Oppgavetiden er i gjennomsnitt nesten ti minutter lengre enn både fullskala og felt, med et standardavvik på 10. Denne tidsforskjellen vil bli gått nærmere inn på i nesten kapittel. Det bør også nevnes at de har også oppdaget mest problemer med 16 i snitt.

7.3.4 Fullførte oppgaver og sammenbrudd

En oppgave blir definert som fullført når testpersonen mener at de er ferdig med en oppgave og når personen har gjort en troverdig innsats. Oppgaven trenger med andre ord ikke være løst 100% etter boka. Mindre detaljfeil blir godtatt så lenge intensjonen har vært korrekt. Det ble tolket som ett sammenbrudd dersom en testperson satte seg fast og ikke fikk til en oppgave, eller når personen slet lenge uten framgang og oppgaven ble avbrutt.

I figur 30 blir antall fullførte oppgaver illustrert ved hjelp av et stolpediagram. En må være klar over at oppgave 10 og 11 ikke ble testet av to personer i felten. Den ene testpersonen hadde fått nok og ønsket ikke å løse de to siste oppgavene, mens den andre testpersonen ikke kunne fullføre oppgave 10 og 11 på grunn av en konfigurasjonsfeil på håndterminalen. I tabell 9 kan vi se hvem som hadde sammenbrudd på de ulike oppgavene.



*** TP2 ønsket ikke å løse de to siste oppgavene og TP4 kunne ikke fullføre de to siste oppgavene på grunn av en konfigurasjonsfeil på håndterminalen.**
Figur 30 - Viser hvor mange som fullførte de forskjellige oppgavene i de ulike metodene.

Figur 30 viser at oppgave 8 skiller seg veldig ut i stolpediagrammet. Det var bare 3 personer som klarte oppgaven, mens de 12 andre hadde sammenbrudd. Det skyldes ikke realisme eller grove brukbarhetsproblemer knyttet til programmet, men heller at oppgaven var for vanskelig. Oppgaven gikk ut på at de skulle benytte en funksjon som lå gjemt bak en annen funksjon, uten at disse hadde en åpenbar kobling. Det er derfor høvelig å anta at det bare er tilfeldig at de tre som fullførte oppgaven, klarte den. Den burde blitt silt bort før brukbarhetstestingen startet, og derfor blir den ikke vektlagt i dette studiet.

Oppgave 3 og 7 skiller seg ut med mange flere personer fra felten som ikke klarte å fullføre disse oppgavene. Årsaken til det vil bli analysert nærmere i kapittel 8.2.

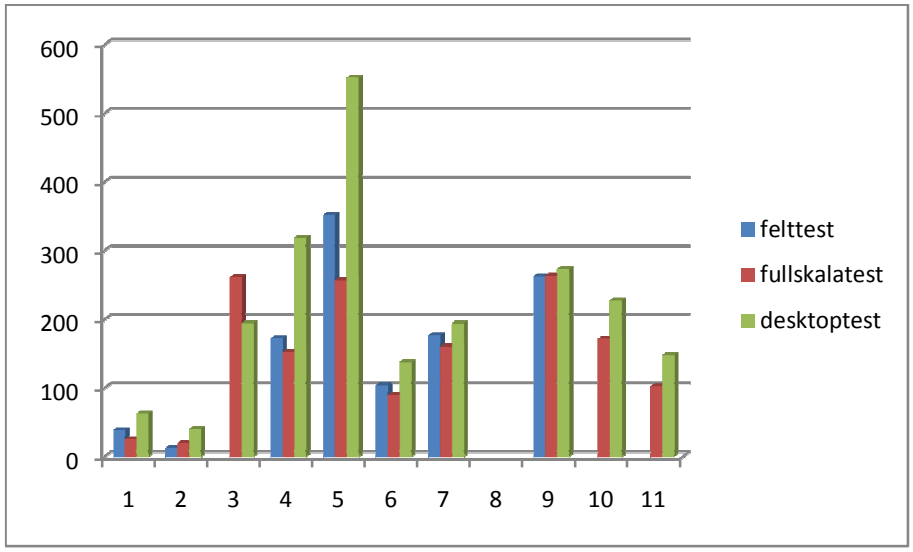
Oppgaver	1	2	3	4	5	6	7	8	9	10	11
Testperson	0	0	1, 2, 4, 5, 8	0	6	0	1, 4, 5	1, 2, 3, 4, 5, 6, 8, 9, 11, 13, 14, 15	5	5	5,11

Tabell 9 - Gir en oversikt over hvilke testpersoner som hadde sammenbrudd på de forskjellige oppgavene.

7.4 Effektivitet

Effektiviteten ble målt ved å ta tiden på hvor lang tid hver enkelt brukte på å løse de forskjellige oppgavene. Tiden ble målt ved å se gjennom alle videoopptakene. Stoppeklokken ble startet i det personene var ferdig med å lese oppgaveteksten og vendte blikket mot håndterminalen. I felten medførte noen av oppgavene at testpersonene måtte gå fra kontoret på lageret til en plass i butikken eller motsatt. Tiden ble stoppet når de måtte gå en lengre distanse. Det ble gjort for at tidsmålingene skal være sammenlignbare. Stoppeklokken ble også stoppet når de fikk spørsmål fra kunder eller kollegaer underveis i brukbarhetstesten.

I figur 31 vises gjennomsnittstiden for testpersonene i de ulike metodene. Tiden blir bare presentert der minimum 3 av 5 har gjennomført oppgaven. Det som skiller seg kraftig ut er hvor mye lengre tid personene i desktopmetoden har brukt på oppgave 4 og spesielt oppgave 5 i forhold til de andre to metodene. På oppgave 5 bruker desktoptesterne signifikant lengre tid enn fullskalatesterne ($t=3.07$, $p < 0.05$). Dette er den eneste statistisk signifikante tidsforskjellen mellom de ulike metodene. Det kan nevnes at tidsforskjellen på desktoptesterne og felttesterne er nesten signifikant ($t=2.03$, $p < 0.1$) i oppgave 5.



Figur 31 - Stolpediagram som viser hvor lang tid i sekunder testpersonene i de tre metodene brukte i snitt på oppgavene, hvor minimum 3 av 5 har klart oppgaven.

Grunnen til den signifikante tidsforskjellen i oppgave 4 skyldes den forenklede konteksten i desktopmetoden. Hele fire av fem testpersoner i desktopmetoden skjønte ikke at de til enhver tid kunne benytte seg av varelappene som lå på bordet (se figur 26). Dette førte til at de søkte opp varene i stedet for å skanne dem. Det står mer om dette spesielle *falske positive* metodeproblemet under analysen i kapittel 8.1.3. Dette problemet er også skyld i at personene i desktopmetoden brukte noe lengre tid i oppgave 4 enn de andre. Vi kan på bakgrunn av det *falske positive* metodeproblemet si at den signifikante tidsforskjellen ikke er reell. Derfor kan vi oppsummere med at det ikke er noen signifikante forskjell på tidsbruken mellom metodene.

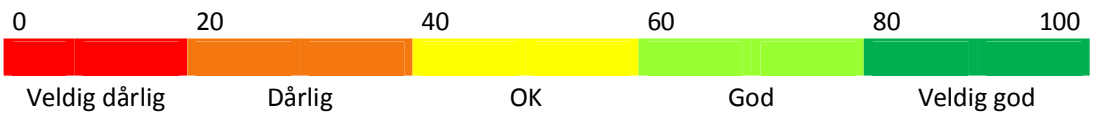
Gjennomsnittstiden for alle testpersonene i en metode fra tabell 6-8 og figur 31 viser at tidene til felt og fullskala er forbausende like. Tiden for felt er 29,9 minutter i snitt per person, mens den i fullskala er 29,1. Desktop skiller seg derimot ut med 38,2 minutter, som er nesten 10 minutter lengre enn de andre. Deler av grunnen til dette er det *falsk positive* metodeproblemet. I figur 31 vises det at desktoptesterne i snitt brukte litt lengre tid enn de andre metodene på alle oppgavene. Samtlige personer i desktoptesten svarte at de ikke klarte å leve seg inn i rollen som butikkansatt. Den forenklede konteksten gjorde det kanskje vanskeligere å forstå oppgavene og hvordan konteksten fungerer relatert til dem. Nedenfor i tabell 10 vises tiden som testpersonene brukte på de enkelte oppgavene i sekunder.

	Oppgaver ->	1	2	3	4	5	6	7	8	9	10	11
Felttest	Testperson 1	23	12	*227*	172	401	69	*176*	*166*	251	312	154
	Testperson 2	71	11	*306*	*393*	451	149	116	*194*	302	*104*	*145*
	Testperson 3	20	7	121	89	219	82	227	*166*	195	206	85
	Testperson 4	39	14	*143*	120	333	79	167	*151*	382		
	Testperson 5	38	19	*292*	89	357	141	198	*134*	*183*	*311*	*144*
Fullskalatest	Testperson 6	25	60	280	265	*280*	200	100	*295*	310	200	80
	Testperson 7	23	10	214	100	360	27	145	276	*345*	168	113
	Testperson 8	23	6	*417*	95	219	108	178	*271*	227	128	116
	Testperson 9	30	9	202	87	264	72	138	*281*	234	202	94
	Testperson 10	24	12	194	212	159	41	241	87	201	159	109
Desktoptest	Testperson 11	97	31	181	161	617	88	158	*305*	324	178	*242*
	Testperson 12				108	384	65	247	163	245	139	118
	Testperson 13	32	79	191	309	862	145	184	*284*	279	278	174
	Testperson 14	74	24	227	361	*369*	57	229	*271*		213	92
	Testperson 15	48	28	178	656	*527*	331	151	*256*	246	329	114

Tabell 10 - Tabell som viser tidsbruker i sekunder for hver oppgave per testperson. De tidene som er markert med rød tekst illustrerer sammenbrudd

7.5 Tilfredshet

Brukertilfredsheten ble målt ved hjelp av en SUS test. Den testen genererer et prosentvis tall for hver person mellom 0-100 (SUS score). Det tallet sier noe om hvor høy tilfredsheten er for de enkelte personene. Desto høyere tallet er, desto mer tilfreds er personene med systemet som ble testet. I figur 32 er det laget en skala som hjelper til med å vise graden av tilfredshet. (Brooke 1996)



Figur 32 - En skala fra 0-100 som viser graden av tilfredshet. 0 er dårligst og 100 er best.

I tabell 11 presenteres SUS score for hver enkelt testperson. Snittverdiene for de tre metodene er samlet opp i tabell 12. Ut ifra den tabellen kan vi lese at fullskalatesterne er mest tilfreds med en snittverdi på 84,5. Felt- og desktoptesterne er litt mindre tilfreds med sine henholdsvis snittverdi på 72,5 og 69,5. Standardavviket for de to siste gruppene er også merkbart høyere enn fullskalatesterne. Hvis resultatene blir satt opp i mot figur 32 er tilfredsheten i gjennomsnitt god eller veldig god. Dette er gode resultat for et system i brukbarhetssammenheng og viser at testpersonene jevnt over er godt fornøyd med systemet.

Metode	**Felttest**					**Fullskalatest**					**Desktoptest**				
Testperson	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SUS score	73	58	88	83	63	78	88	85	93	80	55	78	85	53	78

Tabell 11 - SUS resultatet for hver enkelt testperson

Til tross for at det er en betydelig forskjell mellom resultatet til fullskalatesten og de to andre, så er ikke denne forskjellen signifikant. Forskjellen mellom de med størst avstand, fullskala og desktop er nesten signifikant vel og merke (t=2.11, p = 0.08). Mellom felt og fullskala er forskjellen litt mindre og kan også kalles nesten signifikant (t=1.9, p = 0.11).

	Spørsmål	**Felttest**	**Fullskalatest**	**Desktoptest**
1	Jeg kunne tenke meg å bruke dette systemet ofte	4,4	4,2	3,8
2	Jeg synes systemet var unødvendig komplisert	2	1,2	1,8
3	Jeg synes systemet var lett å bruke	3,6	3,8	3,6
4	Jeg tror jeg vil måtte trenge hjelp fra en person med teknisk kunnskap for å kunne bruke dette systemet	2,4	1,4	2,6
5	Jeg syntes at de forskjellige delene av systemet hang godt sammen	4,4	4	3,8
6	Jeg syntes det var for mye inkonsistens i systemet (Det virket "ulogisk")	2	1	2
7	Jeg vil anta at folk flest kan lære seg dette systemet veldig raskt	4,6	4,6	4,2
8	Jeg synes systemet var veldig vanskelig å bruke	1,8	1	1,8
9	Jeg følte meg sikker da jeg brukte systemet	2,6	3,6	3,2
10	Jeg trenger å lære meg mye før jeg kan komme i gang med å bruke dette systemet på egen hånd	2,4	1,8	2,6
	Gjennomsnittsverdi	**72,5**	**84,5**	**69,5**
	Standardavvik	13	6	15
	Minste verdi	57,5	77,5	52,5
	Største verdi	87,5	92,5	85

Tabell 12 - SUS tabell med gjennomsnittet pr spørsmål i de ulike kontekstene og totaloppsummeringsverdier

7.6 Brukbarhetsproblemer

Brukbarhetsproblemene er som nevnt tidligere en del av anvendbarhet, og i dette kapitlet vil vi gå nærmere inn på brukbarhetsproblemene som ble oppdaget. Alvorligheten til alle brukbarhetsproblemer blir vurdert og de får en alvorlighetsgradering. Brukbarhetsproblemene blir også kategorisert etter hva slags type problem det er snakk om. Alle brukbarhetsproblemene er lagt inn i tabell 13 med en forklaring på problemet og skal fungere som et oppslagsverk. Det kan nemlig være nyttig å gå inn og se på detaljene når de ulike problemene analyseres og diskuteres senere. Til slutt vises det en oversikt ved bruk av venndiagram som illustrerer hvor de ulike brukbarhetsproblemene har oppstått etter både alvorlighetsgrad og problemtype.

7.6.1 Gradering av alvorlighet

Alle brukbarhetsproblemene har fått en alvorlighetsgrad som er kritisk, alvorlig eller kosmetisk. Disse alvorlighetsgradene stammer fra Molich (2003:27,28,154-157) og blir benyttet i en del andre vitenskaplige artikler som Kjeldskov et al. (2004), Nielsen et al. (2006), Kjeldskov og Stage(2004), Duh et al. (2006) og Kjeldskov og Skov (2003). Felles for disse artiklene er at alle har foretatt en brukbarhetsstudie med mobile enheter der de sammenligner minst to metoder. Grunnen til at vi også benytter samme alvorlighetsgradering er fordi det gjør resultatene enklere å sammenligne mot andres resultater.

Kritisk
Brukbarhetsproblemer blir vurdert som kritiske hvis testpersonen ikke får fullført en ønsket operasjon og dermed ikke får løst oppgaven. Dersom en testperson gjør en forretningskritisk feil, vil det også bli gradert som kritisk.

Alvorlig
Dersom et problem vurderes som alvorlig vil det si at personen klarer oppgaven, men at personen sliter med problemet og bruker en del ekstra tid enn de ellers ville ha gjort.

Kosmetisk
Mindre problem som egentlig ikke lager noen større utfordringer for brukeren enn at de bruker litt mer tid på oppgaven.

7.6.2 Kategorisering av problemene

En ikke helt uvanlig praksis er å knytte alle brukbarhetsproblemene som oppdages opp mot en av Nielsens (2001) ti heuristikker. Av praktiske årsaker har vi i denne studien valgt å korte ned på antall kategorier. Nedenfor presenteres de ulike kategoriene og vi sier hva slags heuristikk de kan knyttes mot.

Designproblemer
Det ble oppdaget flere problemer som skyldes designet. Alle de grafiske komponentene som er brukt i programmet er spesialdesignet, og dette kan være med på at brukerne ikke kjenner igjen velkjente standardkomponenter. Denne kategorien omfavner punktene 2. "*Match between system and the real world*" og 4. "*Consistency and standards*" i Nielsens reviderte Heuristikker.

Manglende tilbakemelding
Generelt var det dårlig med tilbakemelding når programmet jobbet, og i enkelte situasjoner var det også dårlig tilbakemelding på handlingene brukerne gjorde. Mangelen på tilbakemelding kan kobles mot heuristikk 1. "*Visibility of system status*".

Språkproblem
Enkelte plasser ble det brukt noen tekniske ord og enkelte engelske ord. Noen setninger var også rett og slett dårlig oppbygd. Slike problemer går under punkt 2. "*Match between system and the real world*" som sier at systemet bør snakke brukerens språk og bruke kjente konsepter.

Metodefeil
Dette er en kategori for oppsamling av handlinger som helt tydelig var et resultat av konteksten de ble utført i, og er ikke feil i selve systemet. Disse problemene er gjerne knyttet til en spesiell metode, og er derfor ikke reelle problemer som kan oppdages overalt.

Systemfeil
Systemfeil er rett og slett en god gammel bug som avdekkes i programmet under testingen.

7.6.3 Tabell med brukbarhetsproblemer

Brukbarhetsproblemene som ble oppdaget ligger i tabell 13. Alle problemene har fått sitt unike nummer og vil bli referert til etter det senere i studien. Resultatene under kolonnene felt, fullskala lab og desktop lab viser hvor mange testpersoner i de forskjellige metodene som oppdaget de enkelte problemene. Den nest siste kolonnen sier noe om hvor det var mulig å oppdage de enkelte problemene. Hvis vi for eksempel ser på problem 1 i tabell 13 står det "X2", og det betyr at dette problemet i all hovedsak bare var mulig å observere i fullskalatesten og desktoptesten. Problemet var altså vanskelig å oppdage i felten på grunn av utfordringene knyttet til å fange interaksjonen som nevnt under kapittel 6.4.3. Den siste kolonnen i tabellen sier hva slags testpersoner som oppdaget de forskjellige problemene. Opprinnelig bestod listen av 88 problemer, men etter fjerning av duplikater og problemer som ikke var problemer ble listen krympet ned til 56. Det er bakgrunnen for at problemnumrene strekker seg fra 1 til 88 med en del hull.

I de tidligere metodesammenligningene som ble presentert i kapittel 5.1, har det vært flere personer inne for å registrere problemer og analysere de. Det er derfor en svakhet at denne masteroppgaven stort sett er et one-man-show. Til å bøte litt på det har jeg sammen med en Lars Flem (som nevnt i kapittel 6.2.1) gått gjennom alle registrerte problemer. Vi har diskutert

problemene til vi ble enige om hvilke registrerte problemer som er faktiske problemer, og avgjorde alvorlighetsgraden og problemtypen til de.

Nr	Problembeskrivelse	Felt	Fullskala lab	Desktop lab	Alvorlig-hetsgrad	Type problem	Felt(1) Fullskala(x) desktop(2)	Testperson
1	Fokuset i en tekstboks forsvinner ikke når en trykker utenfor tekstboksen			1	kosmetisk	Design-problem	**X2**	14
2	Prøver å bruke den gule skannerknappen som enterknapp når de bruker piltastene. Problemet er at den gule knappen er i midten med piltastene rundt og på en mobiltelefon ville tasten i midten vanligvis være <enter>		1	2	kosmetisk	Design-problem	**X2**	6, 11, 12
4	Legger ikke merke til eller forstår ikke at pipingen betyr at en bong overføres ved synkroniseringsjobben som kjøres når håndterminalen står i sleden	2		3	kosmetisk	Manglende tilbakemelding	**1X2**	2, 3, 11, 13, 15
5	Legger ikke merke til bongkø ikonet oppe til høyre som indikerer antall bonger som ligger i kø, klar for å bli sendt	5	1	3	Alvorlig	Manglende tilbakemelding	**1X2**	1, 2, 3, 4, 5, 8, 11, 12, 14
6	Leser ikke hele feilmeldingen som kommer på synkroniseringssiden når en er offline og trykker på "synkroniser"-knappen. De fikk da følgende feilmelding og leste bare til online uten å legge merke til resten før etter flere min. "Må være online eller docked for å benytte denne funksjonen"		3		Alvorlig	Språk-problem	**1X2**	7, 8, 9
7	Forstår ikke hvorfor gammel pris står i en lagertransaksjonsliste etter at de har satt ny pris. De får heller ikke noen tilbakemelding når prisen endres		3	5	Alvorlig	Manglende tilbakemelding	**1X2**	6, 7, 9, 11, 12, 13, 14, 15
8	Synes full synkronisering tar litt lang tid og noen tror programmet har klikket på grunn av lite tilbakemelding på at håndterminalen jobber	1	2	3	Alvorlig	Manglende tilbakemelding	**1X2**	3, 6, 7, 12, 13, 15

Nr	Problembeskrivelse	Felt	Fullskala lab	Desktop lab	Alvorlig-hetsgrad	Type problem	Felt(1) Fullskala(x) desktop(2)	Testperson
13	Legger inn en vare uten å bruke strekkoden. TP søker opp varen i stedet eller skriver inn EAN-koden			4	Alvorlig	Metodefeil	**1X2**	11, 13, 14, 15
14	Forstår ikke at en varelinje må valgt under varesalg for å få aktivert linjemenyknappen og funksjonaliteten under den	1		1	Alvorlig	Design-problem	**1X2**	5, 11
15	Åpner EAN/varesøk-dialogen når vedkommende skal skanne en vare og velger avbryt på dialogen når vedkommende er ferdig med å skanne. Dialogboksen må ikke være åpen for å kunne skanne.	3		3	Alvorlig	Design-problem	**1X2**	1, 4, 5, 11, 12, 14
16	Usikker på om ny etikett har blitt bestilt i intern overføringen, og derfor prøver de å bestille samme etiketten enda en gang før de går videre		2	1	Alvorlig	Manglende tilbakemelding	**1X2**	7, 8, 11
17	Glemmer å angi antall for en vare i varetellingen			2	kosmetisk	Metodefeil	**1X2**	12, 13
19	Alphaknappen er i en tilstand slik at brukeren skriver bokstaver når vedkommende vil ha tall eller motsatt. Tilstanden indikeres ikke på noen måte, så TP må passe på å skru av og på alphaknappen	3	5	4	alvorlig	Manglende tilbakemelding	**1X2**	1, 2, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
20	Forstår ikke hva "docked" betyr i offlinefeilmeldingen TP får i synkroniseringssiden når håndterminalen ikke er dokka	3			Alvorlig	Språk-problem	**1X2**	1, 4, 5
21	Forstår ikke hva "synkroniser" betyr og får dermed ikke løst oppgave 3	1			Alvorlig	Språk-problem	**1X2**	2
22	Legger ikke merke til at en eller flere priser på hylleetiketten er feil i forhold til hva som står på håndterminalen	2	1		Alvorlig	Metodefeil	**1X2**	1, 2, 10

Nr	Problembeskrivelse	Felt	Fullskala lab	Desktop lab	Alvorlig-hetsgrad	Type problem	Felt(1) Fullskala(x) desktop(2)	Testperson
23	Klarer ikke å skrive med SMS tastatur. Tastaturet er litt vanskeligere enn på en mobiltelefon fordi de ser ikke bokstaven de holder på å skrive før etter at de har trykket en numpad-tast x antall ganger	3			Alvorlig	Design-problem	**1X2**	1, 2, 5
24	Forstår ikke at det er en modal dialogboks som er oppe og må avbrytes. TP prøver å trykke på knapper som er utenfor dialogen	3	1		Alvorlig	Design-problem	**X2**	1, 2, 5, 6
25	Setter for eksempel ikke varetellingen til ”ferdig” før håndterminalen puttes ned i sleden for å synkronisere over bongen. Dette medfører at den aktuelle varetellingen ikke blir overført, fordi den ikke er satt til ”ferdig”	1	1		kritisk	Manglende tilbakemelding	**1X2**	2, 10
26	Trykker OK i dialoger hvor det forventes inputtekst og får rød tekstboks fordi de ikke skjønner det er en tekstboks i dialogen.	5	1	1	kosmetisk	Manglende tilbakemelding	**1X2**	1, 2, 3, 4, 5, 9, 12
28	Trykker flere ganger unødvendig fordi det ikke indikeres at håndterminalen jobber med en operasjon	2			Alvorlig	Manglende tilbakemelding	**X2**	2, 4
30	TP tror programmet har låst seg fordi TP har glemt av at skjermen på håndterminalen er trykkfølsom. Problemet er altså at en dialog mangler støtte for navigering med tastatur		1		kritisk	Systemfeil	**1X2**	6
32	Legger ikke merke til pipet etter endt varemottak. Pipet skal indikere at varemottaket har blitt sendt i bakgrunnen. I tillegg blir telleren i bongkøikonet oppe i høyre hjørne telt ned.	2	2	3	Alvorlig	Manglende tilbakemelding	**1X2**	2, 3, 6, 8, 11, 14, 15

Nr	Problembeskrivelse	Felt	Fullskala lab	Desktop lab	Alvorlig-hetsgrad	Type problem	Felt(1) Fullskala(x) desktop(2)	Testperson
37	Søker etter vare men får ikke treff pga dårlig implementert søkefunksjon	1	1	2	kritisk	Systemfeil	**1X2**	4, 7, 13, 15
45	Når TP skulle oppdatere varedatabasen i oppgave 3 slet de med å finne fram til "synkroniser"-knappen og gikk gjerne via "database"-knappen som også er under synkroniser	4	5	3	Alvorlig	Språk-problem	**1X2**	1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14
46	Sliter med ABC tastatur pga veldig små taster		1	2	Alvorlig	Design-problem	**1X2**	8, 11, 14
47	I oppgave 8 gikk TP inn på "vareliste" under lagertransaksjonsmenyen og anga "rogaland" som varelistenavn. Vareliste er et veldig misvisende navn fordi det egentlig er bare en telleliste. Det er ikke mulig å få oversikt over alle varer i databasen.		3	5	Alvorlig	Språk-problem	**1X2**	6, 8, 9, 11, 12, 13, 14, 15
48	Prøver å skanne en vare i lagertransaksjonsmenyen uten at det skal være mulig. Manglende indikasjon på når en kan skanne og ikke.			3	kosmetisk	Design-problem	**X2**	11, 14, 15
49	Skanner tre ganger raskt i et varesalg men programmet registerer bare to skanninger selv om håndterminalen piper tre ganger.			2	kritisk	Systemfeil	**1X2**	12, 13
53	Dårlig indikasjon på at varesøksfunksjonen jobber. TP klager/trykker flere ganger på OK	1	2	3	kosmetisk	Manglende tilbakemelding	**1X2**	1, 8, 9, 13, 14, 15
54	Glemmer å bestille ny etikett	1		3	Alvorlig	Metodefeil	**1X2**	2, 13, 14, 15

Nr	Problembeskrivelse	Felt	Fullskala lab	Desktop lab	Alvorlig-hetsgrad	Type problem	Felt(1) Fullskala(x) desktop(2)	Testperson
57	Ved varesalg godkjennes alle operasjoner asynkront av en ekstern tjener og i ett tilfelle sier TP feil totalsum til kunden pga at TP var rask og sa totalsum før alle operasjonene var godkjent.			1	Alvorlig	Design-problem	**1X2**	13
60	TP har foretatt et søk som ikke ga ønsket resultat og trykker på søkeknappen i resultatvisningen. Det dukker da opp en ny dialog hvor en kan skrive ny søkespørring over resultatvisningen. TP tror da at det nye søket blir gjort blant resultatet fra forrige søk pga den overlappende dialogen, uten at det er tilfelle.			1	kosmetisk	Design-problem	**1X2**	15
61	TP har vært innom varelistesiden i forbindelse med oppgave 3 og 8 uten at de noen gang skulle dit. Vareliste er rett og slett et misvisende navn	5	4	5	kosmetisk	Språk-problem	**1X2**	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15
64	Forstår ikke tastaturknappen "ABC" og får derfor heller ikke til å skrive inn noe tekst. (TP benyttet heller aldri SMS tastaturmuligheten).	2			kritisk	Design-problem	**1X2**	2, 5
65	Tømmer databasen under database når TP skal prøve å oppgradere varedatabasen i oppgave 3.	1	2		alvorlig	Språk-problem		2, 8, 9
67	I dialogen hvor TP angir rabatt legger ikke TP merke til at det er en tekstboks rett under hvor en angir hva slags rabattype vedkommende ønsker. TP trykker derfor på OK-knappen og får rød bakgrunnsfarge i tekstboksen som indikerer at noe må skrives der. Tekstmarkøren burde blinket i tekstboksen for å hjelpe til med å indikere at det er en tekstboks.	1	4	2	kosmetisk	Design-problem	**1x2**	1, 6, 8, 9, 10, 12, 14

Nr	Problembeskrivelse	Felt	Fullskala lab	Desktop lab	Alvorlig-hetsgrad	Type problem	Felt(1) Fullskala(x) desktop(2)	Testperson
68	I listen hvor alle varene ligger holder TP på å redigere antallet på en av varelinjene prøver TP å trykke pil opp for å komme til varelinjen over, men det fungerer ikke.		1		kosmetisk	Design-problem	**x2**	7
69	TP er inne på bongkøsiden og prøver å gjenopprette en bong ved å dobbelklikke på ønsket bong i listen.		1		kosmetisk	Design-problem	**x2**	7
70	Får programmet til å krasje ved å trykke ok(som brukes til å legge en vare) i resultatlisten til varesøk når det ikke er noen varer i listen.		1		kritisk	Systemfeil	**1x2**	7
71	Trykker på varelinjen i varesalg mens den fortsatt er disabled og venter på godkjenning.		1		kosmetisk	Design-problem	**x2**	7
72	Prøver å gå tilbake til en funksjon(for eksempel varetellingen i oppgave 6) for å finne noe TP har gjort tidligere uten at vedkommende finner det.	2	4	2	alvorlig	Design-problem	**1x2**	3, 5, 6, 8, 9, 10, 13, 15
74	Sliter med å skrive med SMS tastatur.	1	1		alvorlig	Design-problem	**1x2**	2, 8
76	Under rollespillet i oppgave 11 hører ikke TP etter når jeg vil ha prosentvis rabatt. TP sier jeg skal få 10kroner i avslag og velger rabattype pris i stedet for beløp og dermed blir prisen på varen feil. TP legger ikke merke til at han gjør denne feilen.		1		alvorlig	Språk-problem	**x**	9

Nr	Problembeskrivelse	Felt	Fullskala lab	Desktop lab	Alvorlig-hetsgrad	Type problem	Felt(1) Fullskala(x) desktop(2)	Testperson
77	TP går inn på en funksjon og registerer noen varer. TP finner ut at dette blir feil og går ut ved å sette bongen til ferdig. Dette medfører at denne ugyldige bongen blir lagt inn i systemet og vil medføre feil. Navngivingen kunne vært bedre for å unngå dette.			2	kritisk	Språk-problem	**1x2**	11, 14
78	Benytter "send bonger"-knappen i bongkø som egentlig skulle vært deaktivert når tilstanden er offline.	2	3	2	alvorlig	Systemfeil	**1x2**	1, 4, 6, 8, 10, 12, 13
79	Sliter med å få bort popup menyen(ala startmeny i Windows) som hver funksjon(varetelling, varesalg osv…) har.TP trykker lukk som fører til at funksjonen lukkes (og TP må dermed gå inn igjen hvis det ikke er noen elementer i listen). En kan heller ikke trykke utenfor popup menyen for å få den bort.	2		2	alvorlig	Design-problem	**1x2**	1, 4, 13, 14
80	Backspaceknappen på ABC tastatur fungerer ikke.			1	alvorlig	Systemfeil	**1x2**	14
81	Finner ikke "ø" på SMS/ABC tastatur. Dette kunne bare bli funnet i desktoptesten pga at det ble brukt en vare med "ø" i navnet i den metoden.			2	alvorlig	Design-problem	**2**	13, 14
82	Synkroniseringen startes ikke automatisk dersom TP har en dialog oppe på synkroniseringssiden.			1	alvorlig	Systemfeil	**1x2**	15
83	Trykker på antallknapp selv om antalltekstboks er åpen for å skrive inn antallet.	2			kosmetisk	Design-problem	**1x2**	1, 5

Nr	Problembeskrivelse	Felt	Fullskala lab	Desktop lab	Alvorlig-hetsgrad	Type problem	Felt(1) Fullskala(x) desktop(2)	Testperson
84	TP er litt usikker på om status "ferdig" for en bong i bongkøen betyr om den er sendt eller ikke.	1			alvorlig	Språk-problem	**1x2**	1
85	En oppgave blir ikke gjort som planlagt og vedkommende får beskjeden "Får ikke kontakt med web service". TP gir uttrykk for at vedkommende ikke vet hva en web service er.	1	1		alvorlig	Språk-problem	**1x2**	1, 10
86	Velger en vare i varesøkresultatet og skal legge den til listen ved å trykke søk i stedet for OK. Navngivingen kunne vært bedre og søkeknappen kunne vært fjernet.	1			kosmetisk	Språk-problem	**1x2**	1
87	Prøver å foreta et varesøk i EAN-dialogboksen og skjønner ikke hvorfor bakgrunnen i tekstboksen blir rød(som indikerer feil input) når vedkommende trykker OK.	1			alvorlig	Design-problem	**1x2**	1
88	Klarer ikke å angi antall fordi ALPHA-knappen er aktivert og alle tastetrykk blir bokstaver i stedet for tall.	1			alvorlig	Design-problem	**1x2**	2

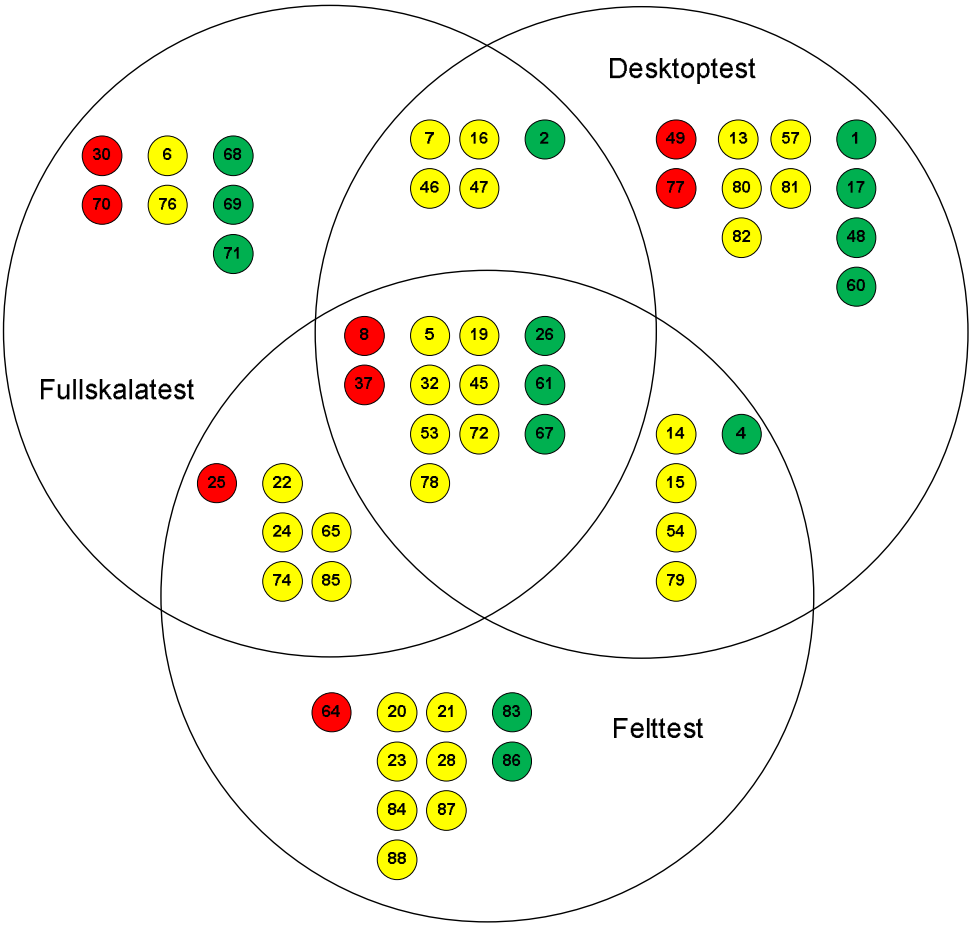
Tabell 13 - Brukbarhetsproblemene som ble oppdaget

7.6.4 Oversikt som viser hvor de forskjellige feilene oppstod

Tidligere forskning mellom forskjellige metoder har i all hovedsak bare sammenlignet to metoder: Fullskalatest i laboratorium mot en felttest eller en desktoptest i laboratorium mot en felttest. Resultatene i slike eksperimenter har vanligvis blitt presentert i tabellform. I denne oppgaven sammenlignes tre forskjellig metoder og det vanskeliggjør helhetsbilde. Rett etter vurderingen av alvorlighetsgraden til problemene ble det tegnet opp tre overlappende sirkler på en tavle (se figur 33). Inne i de forskjellige regionene ble det festet røde, gule og grønne post-it-lapper som symboliserer alvorlighetsgraden for hvert av problemene.

Figur 33 - Tavle med post-it-lapper over hvor problemene oppstod. Tavlen ble benyttet i innledende fase av problemanalyseringen.

På neste side og figur 34 ser vi den reviderte utgaven i digital form. De røde boblene symboliserer kritiske problem, de gule symboliserer alvorlige problem og de grønne symboliserer kosmetiske problem. Boblenes er gruppert etter alvorlighetsgraden i de forskjellige regionene med de kritiske til venstre og de kosmetiske til høyre. I midten av hver boble står det ett tall som refererer til problemets unike nummer, og for en grundig analyse av en problemboble kan vi slå det opp i tabell 13.



Figur 34 - Venndiagram som viser hvor problemene oppstod

I tabell 14 er figur 34 oppsummert i henhold til hvor mange problemer av forskjellig alvorlighetsgrad som oppstod innenfor de forskjellige metodene. Ut ifra tabell 14 ser vi at det er oppdaget så og si like mange kritisk problemer i alle metodene, men at felttesten avdekker litt flere alvorlige problemer enn de andre. Ingen av disse forskjellene er store, og det er langt ifra noen signifikant forskjell på alvorlighetsgradene sett opp imot de forskjellige metodene.

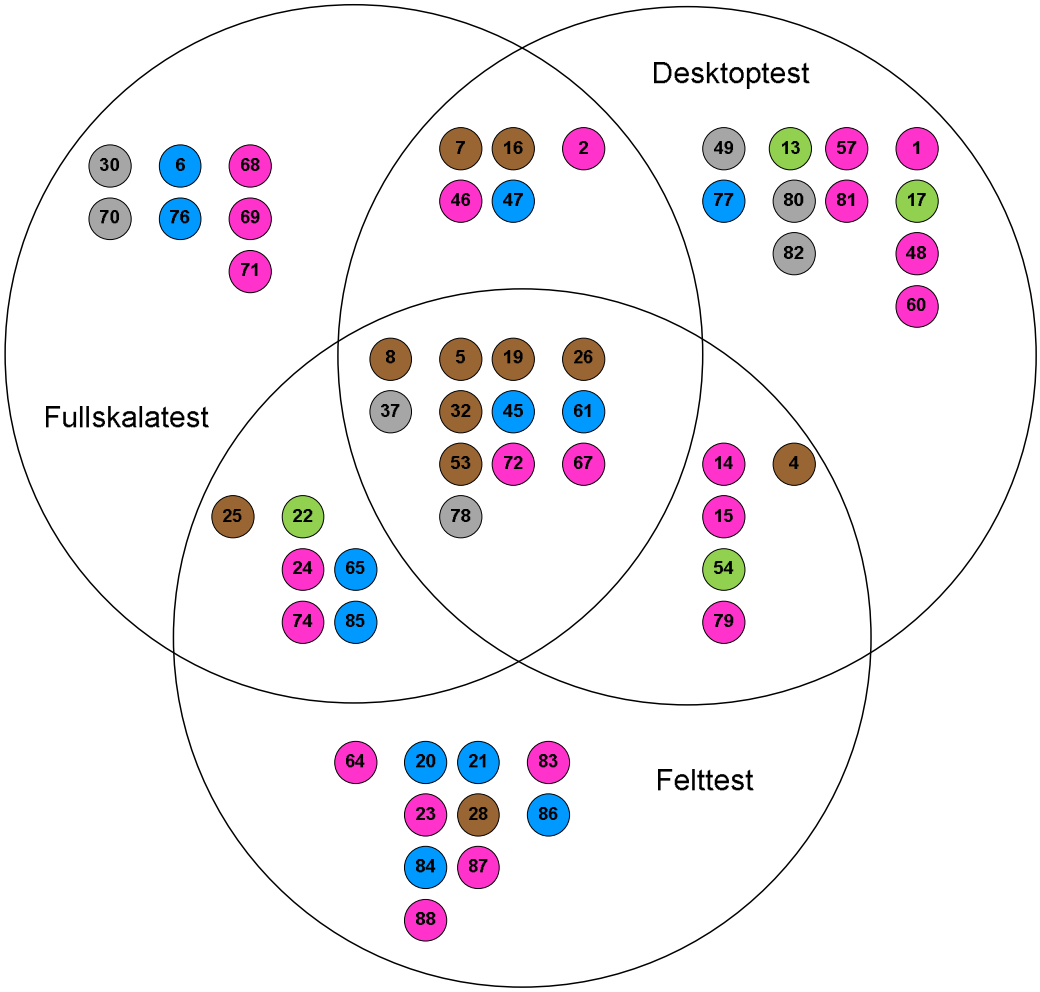
Ut ifra figur 34 ser vi at hele 28 av de 56 problemene bare har opptred i en av metodene. Disse unike metodeproblemene vil bli analysert grundig i kapittel 8.4. Det vil være interessant å se om noen av disse problemene skyldes metoden.

Alvorlighetsgrad	Felt	Fullskala	Desktop	Totalt
Kritisk	4 (50%)	5 (63%)	4 (50%)	8
Alvorlig	23 (68%)	18 (53%)	20 (59%)	34
Kosmetisk	6 (43%)	7 (50%)	9 (64%)	14
SUM	33 (59%)	30 (54%)	33 (59%)	56

Tabell 14 - Tabell som viser hvor mange hvor mange problemer og alvorlighetsgraden til disse i de forskjellige kontekstene.

7.6.5 Oversikt som viser de forskjellige problemtypene

Til å illustrere hvor de ulike typer problem opptrer har vi også valgt å benytte et venndiagram. Bakgrunnen for det er at en slik visuell fremstilling vil gjøre enklere å få et oversiktbilde på hva slags problemer som oppstår hvor. Fargene i figur 35 betyr følgende: Rosa betyr designproblem, brun betyr manglende tilbakemelding, blå betyr språkproblem, grønn betyr metodefeil og grå betyr systemfeil.



Figur 35 - Venndiagram som viser hvor de forskjellige problemtypene oppstod

I tabell 15 oppsummeres figur 35 med henhold til hvor mange av hver problemtype som oppstår i de forskjellige metodene. Antall språkproblemer skiller seg ut i felt og fullskala med henholdsvis 8 og 7 problemer, i motsetning til desktop sine 4. Det ble oppdaget minst dobbelt så mange systemfeil i begge laboratoriemetodene i forhold til felten. Manglende tilbakemelding er helt klart i følge figur 35 den problemtypen som var lettest å oppdage i alle metodene.

Problemtype	Felttest	Fullskalatest	Desktoptest	Totalt
Designproblem	12 (55%)	9 (41%)	12 (55%)	22
Tilbakemelding	9 (82%)	9 (82%)	9 (82%)	11
Språkproblem	8 (67%)	7 (58%)	4 (33%)	12
Systemfeil	2 (29%)	4 (57%)	5 (71%)	7
Metodefeil	2 (50%)	1 (25%)	3 (75%)	4
SUM	33 (59%)	30 (54%)	33 (59%)	56

Tabell 15 - Viser hvor mange av de forskjellige problemtypene som oppstod i de forskjellige kontekstene

8 Analyse

Fra resultatkapitlet ble det presentert en del interessante data, og de mest sentrale aspektene vil bli analysert i dette kapitlet. Først vil vi analysere noen brukbarhetsproblemer som kan knyttes til en metode eller brukergruppe. Vi vil se nærmere på årsaken til at de forskjellige sammenbruddene fant sted. Tabellen med brukbarhetsproblemer er diskutert sammen med systemleverandøren og de har kommet fram til en prioritert liste med brukbarhetsproblemer. Vi skal se nærmere på de problemene som bare oppstod i en metode og analysere årsakene til det, før vi til slutt vil se grundigere på tilfredsheten til testpersonene i de ulike metodene.

8.1 Analysering av brukbarhetsproblemer

Under analysen av brukbarhetsproblemer vil det bli gått i dybden på noen av de mest interessante problemene. Alle de analyserte problemene vil være med å danne et grunnlag for senere kapittel.

8.1.1 Modal dialogboks var forvirrende

Problem 24 i tabell 13 er et spesielt tilfelle av et designproblem som i de fleste tilfeller ikke var noe problem. Ingen av testpersonene hadde noen særlige problemer med å forstå at det var en modal dialogboks oppe som er avbildet i figur 36. Samme type dialogboks i figur 37 skapte derimot problemer blant fire av testpersonene (tre i felt og en i fullskala). Dialogboksene ser helt like ut i formen, men det som var med på å lure testpersonene var den store lukk-knappen som vises rett nedenfor dialogboksen i figur 37.

Figur 36 - Skjermbilde av en popup-dialog som ingen hadde noen problemer med.

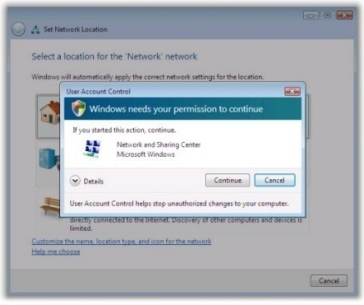
Figur 37 - Skjermbilde av en popup-dialog som flere testpersoner slet med å lukke.

Nielsens heuristikker (2001) sier at brukerne ofte velger feil systemfunksjon og trenger derfor en tydelig markert nødutgang. I dette tilfellet er lukk-knappen en lurenødutgang, og den var spesielt forvirrende blant to personer i felten. De slet i opptil ett minutt før de prøvde å trykke på avbrytknappen. Dersom dialogboksen på figur 37 hadde vært plassert litt lenger ned, slik at den tydelig hadde ligget over lukk-knappen, ville trolig ikke den modale dialogboksen vært noe problem å skjønne.

Normann (2002) sier at en handling skal alltid gi en tilbakemelding. Dersom lukk-knappen i figur 37 trykkes vil vi faktisk få en lavfrekvent pipelyd som skal varsle om at brukeren ikke kan trykke der. En av grunnene til at personene i felten brukte ekstra tid på skjønne hva som måtte gjøres, kan faktisk skyldes at pipet drunket i støyen fra omgivelsene. Støyen i en dagligvarebutikk var også årsaken til at en lydmelding ikke ble hørt i Newcomb et al. (2003).

Det er også interessant å legge merke til at dette problemet aldri oppstod i desktoptesten og det er meget mulig at den høye datakunnskapen der var en medvirkende årsak. Vi ser et lite mønster i at datakunnskapsnivået kan være en medvirkende årsak til dette problemet. De to personene som slet mest fra felttesten svarer på pretest spørreskjemaet at de ikke benytter datamaskiner regelmessig. De vurderer begge sine datakunnskaper til en toer på en skala fra 1-5 der 5 er best. Den ene fra fullskalatesten som også gjorde samme feil benytter en datamaskin regelmissig og vurderer sine datakunnskaper til en femmer, og trenger bare 2-3 sekunder på å skjønne at avbryt-knappen må trykkes.

Det viser seg altså at dette er et problem som dukker opp uavhengig av kunnskapsnivå, men alvorligheten av problemet er større desto dårligere datakunnskap de har. Potensielt kan dette føre til en situasjon der brukeren tror at systemet har låst seg, eller ikke fungerer som det skal.



Figur 38 - UAC i Windows Vista

En god løsning på problemet ville vært å grå ut alt som ligger bak dialogboksen på samme måte som når User Account Control (UAC) i Windows Vista dukker opp og krever din oppmerksomhet (se figur 38). En annen løsning kan være å gi mer visuell tilbakemelding for å følge Normann (2002) sitt utsagn om at handlinger alltid skal gi tilbakemelding ved å la rammen til dialogboksen blinke i noen sekunder når det trykkes utenfor.

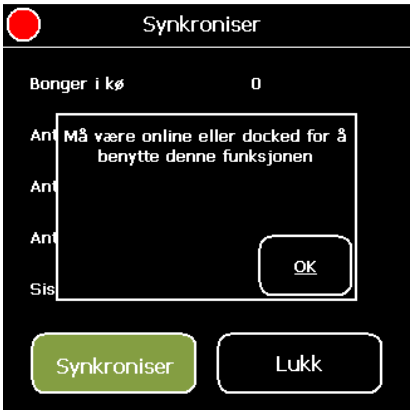
8.1.2 Språkproblemer laget utfordringer i felten

Alle språkproblemene som dukket opp kan deles inn i kategoriene teknisk- og butikkfagspråk, misvisende navn eller uheldige setningsformuleringer. I tabell 16 vises antall tilfeller av de forskjellige typer språkproblemene.

Type språkproblemer	Antall tilfeller	Problem nr
Teknisk språk	2	20,21
Butikk språk	1	33
Misvisende navn	3	45,47,61
Uheldig setningsformulering	1	6

Tabell 16 – Kategorisering av språkproblemene

Mest interessant er de språkproblemene som går under teknisk og butikkspråkgruppen i tabell 16. Det tekniske problemet nummer 20 ble oppdaget i oppgave 3 og delvis i oppgave 7. Begge oppgavene går ut på at de skal synkronisere håndterminalen, og løsningen er å sette håndterminalen i sleden. Det vi ofte oppdaget var at de gikk gjennom menyene og fant fram til synkroniseringsvinduet og trykket på synkroniseringsknappen i figur 39. De får da opp en feilmelding som sier "*Må være online eller docked for å benytte denne funksjonen*" når de er i tilstanden offline. Tre av testpersonene fra felten ga uttrykk for under eller etter testen at "*docked*" skjønte de ikke hva betydde, og derfor klarte de heller ikke å gjennomføre oppgavene. Dette er et brudd på Nielsens (2001) heuristikk "*Match between system and the real world*", som sier at systemet bør snakke brukerens språk ved bruk av ord, begreper og en framstillingsmåte som brukeren forstår.



Figur 39 - Skjermbilde som viser synkroniseringsfeilmeldingen TP får når en er offline og synkroniseringsknappen trykkes

Den eldste testpersonen er 50 år og slet også med oppgave tre(problem nr 21). Problemet hennes var at hun ikke visste hva "*synkroniser*" betydde, og var derfor aldri innom skjermbilde på figur 39. Det er interessant å se at problemene med teknisk språk bare er knyttet til de som er 40 år eller eldre og med under middels datakunnskaper. Språkproblemer er altså årsaken til at det var færre i felten som klarte oppgave 3 og 7 som vi ser på figur 30, og ikke den realistiske konteksten som vi kanskje ville trodd i utgangspunktet. I fullskalametoden skjønte alle hva de måtte gjøre når de leste feilmeldingen i figur 39, mens testpersonene i desktopmetoden bare satt håndterminalen rett i sleden.

8.1.3 Metodeproblem grunnet lav realisme

Totalt ble det registrert fire problemer som mest trolig skyldes den brukte metoden og er ikke direkte problemer i systemet. Det er spesielt problem 13 i tabell 13 som skiller seg ut. Hele fire av fem testpersoner i desktopmetoden slet med å skjønne at de til enhver tid kunne benytte varelappene som er avbildet på figur 40. Det er hovedsakelig oppgave 4, 5, 6 og 10 som baserer seg mye på bruk av varer som utmerket seg.

I tabell 17-21 presenteres det i prosent hvor stor grad av oppgavene som ble løst uten å benytte varelappene og i hvor stor grad de benyttet varesøk eller tastet EAN-koden i stedet for å skanne. Den beste måten å løse oppgavene på er å benytte varelappene og skanne de. En slik løsning vil i følge tabellene gi 0 % på ”uten varelapp”-kolonnen og 0 % på ”varesøk/EAN”-kolonnen. For testpersonene som deltok i felt og fullskalatesten var resultatet 0 % under begge disse kolonnene. 0 % er altså best.

Figur 40 - Lappene som ble brukt som varer og skapte en del forvirring i desktoptesten.

Oppgave	Uten varelapp	Varesøk/EAN
4	0 %	0 %
5	0 %	*57 %*
6	0 %	*100 %*
10	0 %	0 %

Tabell 17 - Viser i hvor stor grad testperson 11 ikke benyttet varelappene og ikke valgte å skanne

Oppgave	Uten varelapp	Varesøk/EAN
4	0 %	0 %
5	0 %	0 %
6	0 %	0 %
10	0 %	0 %

Tabell 18 - Viser i hvor stor grad testperson 12 ikke benyttet varelappene og ikke valgte å skanne

Oppgave	Uten varelapp	Varesøk/EAN
4	0 %	0 %
5	83 %	83 %
6	100 %	100 %
10	66 %	66 %

Tabell 19 - Viser i hvor stor grad testperson 13 ikke benyttet varelappene og ikke valgte å skanne

Oppgave	Uten varelapp	Varesøk/EAN
4	33 %	100 %
5	0 %	0 %
6	0 %	0 %
10	0 %	0 %

Tabell 20 - Viser i hvor stor grad testperson 14 ikke benyttet varelappene og ikke valgte å skanne

Oppgave	Uten varelapp	Varesøk/EAN
4	71 %	71 %
5	0 %	0 %
6	100 %	100 %
10	60 %	60 %

Tabell 21 - Viser i hvor stor grad testperson 15 ikke benyttet varelappene og ikke valgte å skanne

Det vi ser i tabellene er at spesielt testperson 13 og 15 i liten grad benytter varelappene. Det som går igjen er for eksempel at testperson 13 ikke tenker på at han kan benytte varelappene i oppgave 5 før mot slutten av oppgaven. Først da begynner han å lete blant varelappene og skjønner at han kunne brukt dem. På neste oppgave(6) har han igjen helt glemt at han kan benytte varelappene, før han i oppgave 10 kommer på at han kan benytte varelappene etter at han er litt over halvveis ferdig med oppgaven. Akkurat samme mønster går igjen hos testperson 15. På figur 41 ser vi et illustrasjonsbilde fra testpersonenes synsvinkel. Til høyre i bilde ser vi alle varelappene.

Figur 41 - Illustrerer desktopmetoden fra testpersonen sin synsvinkel.

En annen sak som fremkommer av tabellene er at testperson 11 og 14 i all hovedsak benytter varelappene, men det er ikke alltid at de skanner dem selv om strekkoden på varelappene kommer ganske tydelig fram på varelappene(se figur 40). Vi ser i tabell 17 at testperson 11 alltid benytter varelappene, men at vedkommende ikke alltid skanner varene i oppgave 5 og 6. Han benytter varesøk ett par ganger men legger inn flest varer ved å taste inn EAN-kode. Etter testen ble han konfrontert med hvorfor han ikke skannet varene i stedet og svarte følgende: *"Jeg burde jo egentlig ha skannet varene, men jeg vet ikke helt hvor jeg ikke gjorde det, kanskje fikk jeg for meg at det ikke var mulig, men jeg vet ikke om jeg testa det*". Testperson 13 og 15 svarte på samme spørsmål at det rett og slett glemte at de hadde varer liggende på bordet. Det virker som at *"ut av syne ut av sinn*" effekten slo til ved at settingen ble for urealistisk til at de klarte å leve seg inn i rollen som butikkansatt og forstå konteksten.

Isolert sett er ikke problem 13 kritisk, for det ville aldri skjedd i en ekte butikk. Problemet har faktisk ført til en del ringvirkninger som blant annet utstrakt bruk av varesøk, og det fremprovoserte faktisk et par ekstra problemer (37, 60, 80 og 81) som nødvendigvis ikke ville blitt funnet ellers. Det var knotet å skrive tekst på håndterminalen og varesøket var dårlig implementert. All varesøkingen har gitt tydelig utslag på grafen i figur 31 der desktoptesterne i snitt bruker betydelig mer tid på oppgave 4 og 5 enn testpersonene i de andre metodene. Problemene rundt varesøkingen er trolig en medvirkende årsak til at brukertilfredsheten til testpersonene i desktopmetoden er lavest av alle i følge tabell 12. Problem 13 er ikke et reelt problem, men et falskt positivt problem som har skapt ringvirkninger på resultatene og blir diskutert i kapittel 9.2.

8.1.4 Systemfeil fremprovosert av raske testpersoner

Det kritiske problemet nr 49 er litt interessant selv om den bare ble funnet av to personer under desktopmetoden. Kort fortalt går feilen ut på at testpersonene skannet en vare raskt tre ganger. Varesalgsdelen av programmet registrerte derimot bare to av skanningene, selv om håndterminalen pep bekreftende for tre skanninger. Testperson 13 poengterer også alvorligheten av dette i intervjuet:

> *Leseren var grei, den godtok å skanne, også hører du tre pip men har bare registrert to. Tenk deg at du holder på med varetelling og hører pip hele tiden. Du har vært der i seks timer og er dritt lei, så kanskje sjekker du ikke håndterminalen for hver eneste gang, og da får du mye svinn i varetellinga da.* (TP12)

8.2 Sammenbrudd

Testpersonene i felten slet litt på oppgave 3 og 7 ser vi på figur 30. Fullføringsgraden på oppgave 7 er nesten signifikant forskjellig mellom felt og de andre metodene ($t= 2.45$, $p < 0.1$). I disse to oppgavene skulle de oppdatere varedatabasen eller sende over noen ferdige bonger. Løsningen på begge disse oppgavene var å sette håndterminalen ned i sleden på bakrommet.

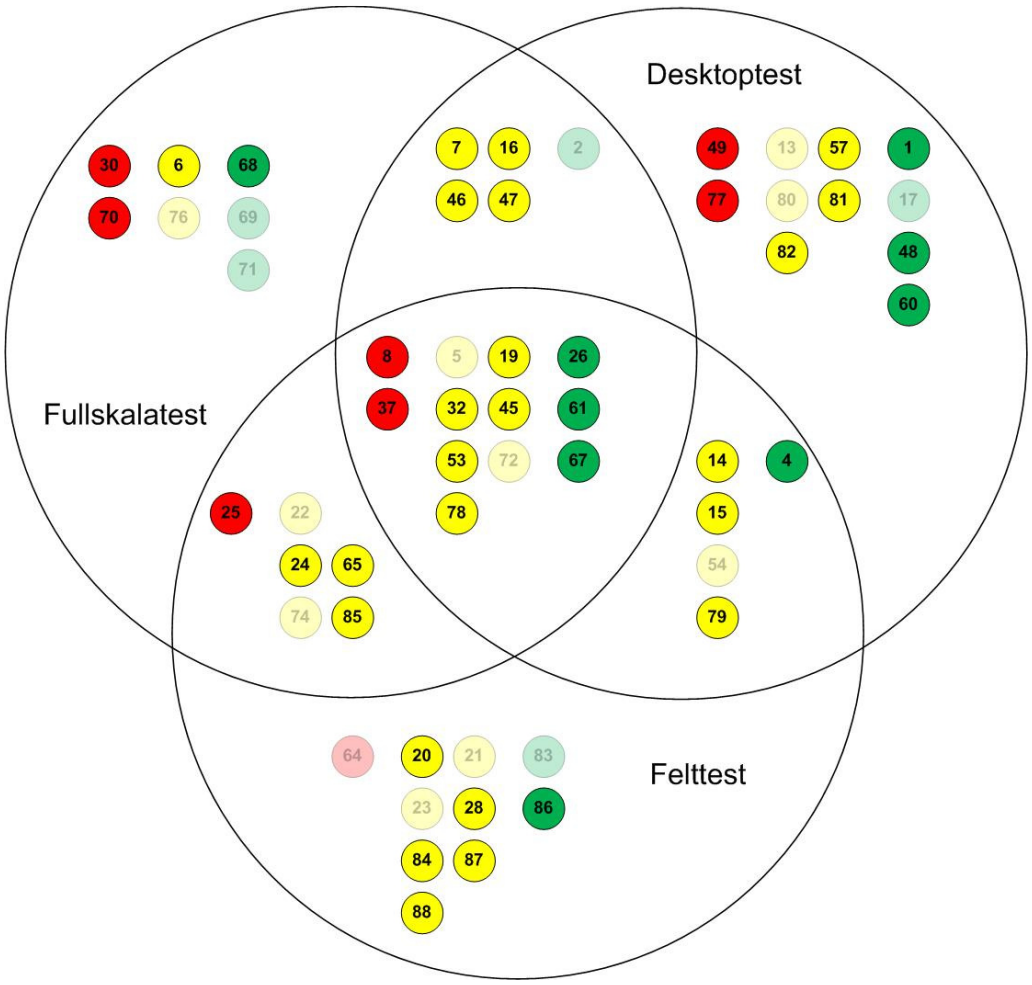
Oppgavene ble konstruert slik at sleden ikke var synlig når testpersonene fikk disse to oppgaver i felt- og fullskalatesten. Dette ble gjort for at løsningen ikke skulle ligge rett foran dem. Umiddelbart kan vi tro at forskjellen skyldes den realistiske konteksten i felten. Det kan derimot tilbakevises av to grunner. For det første ble akkurat samme utfordring gjenskapt i fullskala laboratoriet uten at samme trend kan påvises der. Den andre og viktigste grunnen til disse sammenbruddene skyldes språkproblemene som ble nevnt i kapittel 8.1.2.

I programmet ble det benyttet noen tekniske/engelske ord som tre av fem i felten ikke forstod. De var alle på rett spor, og fikk en feilmelding som forklarte hva som måtte gjøres. Problemet var dårlig ordbruk, og det var årsaken til at de ikke skjønte hva som måtte gjøres. Språkproblemene dette gjelder er registrert som egne brukbarhetsproblemer med nummer 20, 21 og 85 i tabell 13. De to første av disse problemene ble bare oppdaget i felten(se figur 34) og nummer 21 er hovedårsaken til at de ikke klarte disse oppgavene.

8.3 Prioriterte brukbarhetsproblemer

Den tette kontakten mellom undertegnede og systemleverandørene til produktet har blitt utnyttet etter endt empiriinnsamling. Bakgrunnen var at vi ønsket å se hva slags brukbarhets-problemer Lindbak ville fikse av alle oppdagede problemer. Det har ikke blitt gjort noe lignende i andre metodesammenligningsstudier.

Brukbarhetsproblemene i tabell 13 ble diskutert sammen med en utvikler på Lindbak POS Mobile prosjektet. Vi gikk systematisk gjennom hele tabellen og avgjorde om de enkelte problemene var noe som burde fikses. Listen med disse prioriterte problemene ble deretter gjennomgått i et møte med prosjektets ledere hos Lindbak. Resultatet fra det møte var mindre endringer i listen og en del forslag til hvordan flere av problemene kunne løses.



Figur 42 - Viser hva slags problemer Lindbak anser som viktig å fikse. De uviktige er markert som svake bobler.

I figur 42 illustreres de problemene som ikke ble ansett som nødvendig å fikse ved bruk av svake bobler. Hele 40 av de totalt 56 (71 %) registrerte problemene ville Lindbak ha fikset. Blant de som ikke skal fikses er alle metodefeilene. Problem 5, 21, 64, 69 og 72 blir det ikke gjort noe med, fordi de anses som en del av noe brukeren bare må lære seg og være inneforstått med før bruk. Andre problemer som 2, 23, 74, 80 derimot er reelle, men kan ikke løses av Lindbak fordi det er problemer knyttet til tredjeparts hardware på håndterminalen. De tre første av disse problemene omhandler måten de fysiske knappene på håndterminalen er satt opp og fungerer, mens den siste går på en feil i håndterminalens innebygde tastatur.

Ut ifra figur 42 ser vi at opptreden av ikke-prioriterte problemer er ganske jevnt blant de ulike metodene med 7 i desktoptesten, 8 i fullskalatesten og 9 i felttesten. Det er interessant å legge merke til at det eneste kritiske problemet som bare blir oppdaget i felten ikke blir ansett som et problem. Dette medfører at alle prioriterte problemer blir oppdaget ved hjelp av de to laboratoriemetodene, og at under halvparten(3 av 7) av de prioriterte kritiske problemene blir oppdaget i felten.

8.4 Unike problem i metodene

Vi tar utgangspunkt i figur 35 som viser et venndiagram over hvor de ulike problemtypene oppstod. Der er det interessant å gå inn på problemene som bare ble avdekken i en av metodene. Det kan hende at noen av disse problemene er unike for den metoden, og kan derfor kanskje ikke gjenskapes noen annen plass. Andre problemer kan derimot bare være en tilfeldighet at de har blitt oppdaget i en metode, og kunne i teorien blitt oppdaget i flere metoder.

8.4.1 Unike problemer i feltmetoden

Det ble oppdaget 10 unike problemer i felten. Disse problemene er i all hovedsak design og språkproblemer. Halvparten av de 10 problemene som ble oppdaget i felten viser seg å være på bakgrunn av at felttesterne skiller seg forskjellig ut som brukergruppe.

Det ble stilt et krav om at alle 15 testpersonene skulle ha butikkerfaring på forhånd, og det hadde de. Faktisk hadde ingen problemer med butikkreltaterte begreper eller arbeidsoppgaver. Alder og datakunnskaper ser derimot ut til å kunne spille inn. I kapittel 7.1.4 så vi at både alder og datakunnskaper er signifikant forskjellig mellom felten og laboratoriemetodene. I felten er alderen høyere og datakunnskapene lavere, og dette fører til at de i felten skiller seg ut som en egen brukergruppe.

To av de fem problemene som skyldes at felttesterne er en ulik brukergruppe, er språkproblemene med nummer 20 og 21, som går på bruk av engelsk/tekniske begrep (omtalt tidligere i kapittel 8.1.2). De tre siste problemene går under kategorien designproblem, og er relatert til å skrive inn bokstaver eller tall på håndterminalen. Det første av disse tre designproblemene er nr 23, og går ut på at tre av personene ikke klarte å skrive inn tekst ved hjelp av SMS-tastaturet på håndterminalen (se kapittel 3.2.1 for hvordan SMS-tastaturet og den tilhørende alpha-knappen fungerer). Det andre problemet er nr 64 som går ut på at to personer ikke skjønner at en knapp med teksten ”ABC” på vil åpne et tastatur på skjermen(se figur 36). Dette førte til at de to personene aldri klarte å skrive inn noen bokstaver, og det blir derfor karakterisert som et kritisk problem. Det siste problemet er nr 88, og går ut på at alpha-knappen er innstilt på SMS-tastatur. På grunn av det får ikke TP2 til å skrive inn antall varer i en tekstboks, fordi hun bare skriver bokstaver i stedet for tall.

De fem andre problemene(28, 83, 84, 86 og 87) som ikke skyldes brukergruppen, kunne like gjerne blitt oppdaget i de andre metodene. Det er ingenting ved disse problemene som skyldes kontekst og realisme.

8.4.2 Unike problemer i fullskalametoden

I fullskalametoden ble det funnet bare 7 unike problemer, og det er minst av alle. De to kritiske problemene som blir oppdaget kategoriseres som systemfeil. Problem 30 ble oppdaget på grunn av TP6 sin bakgrunn fra Byggmakker. Der hadde hun benyttet en håndterminal som ikke hadde trykkfølsom skjerm. Det førte til at hun automatisk glemte at hun kunne trykke på skjermen, og det til tross for at dette ble behøvelig forklart og demonstrert under opplæringen. De fire første oppgavene klarte hun fint ved hjelp av tastaturet. I oppgave 5 ble hun stående fast i en dialog med manglende muligheter for å navigere ved hjelp av tastaturet. Hun var overbevist om at programmet hadde klikket, og ønsket å restarte håndterminalen. Dette er for så vidt grunnen til at hun er den eneste som ikke fullfører oppgave 5 (se figur 30), og skyldes testpersonen sin bakgrunnserfaring. Det andre kritiske problemet (70) er rett og slett en bug som fikk programmet til å krasje og ble oppdaget ved en tilfeldighet.

Det er tydelig at de to kritiske problemene ikke ble oppdaget på grunn av konteksten i fullskalatesten. De andre design- og språkproblemene som bare ble oppdaget i fullskalatesten kunne like gjerne blitt oppdaget i den forenklede desktoptesten eller i felten.

8.4.3 Unike problemer i desktopmetoden

Mangelen på realisme og enkelheten i desktopmetoden kan, og har ført til oppdagelsen av fem brukbarhetsproblemer (13, 17, 77, 49 og 57). To av disse problemene blir også kategorisert som metodefeil og skyldes manglende realisme. Det ene er problem 13 som går ut på at fire av testpersonene glemte at de kunne skanne strekkoden på varelappene, og har blitt omtalt grundig i kapittel 8.1.3. Det andre er metodeproblem 17, og går ut på at to personer glemte å angi antallet som stod på varelappen. De klarte med andre ord ikke å visualisere varene mentalt. Ingen av disse problemene blir sett på som reelle problem.

Det tredje problemet som også kan skyldes realismen er problem 77, og går ut på at to personer går inn på en funksjon for å registrere noe varer. Midt i registreringen finner personene ut at de er inne på feil funksjon, og går ut ved å sette bongen til ”ferdig”. Dette medfører at de lager en ferdig bong som blir lagt inn i systemet, og vil medføre en feil i butikkdatabasen. Det kan være en tilfeldighet at bare to personer i desktoptesten gjør akkurat denne feilen. Vi kan derfor lure på om manglende realisme fører til at de ikke innser problemet og alvorligheten dette problemet kan medføre. Til sammenligning gikk også TP5 i felten inn på feil funksjon, og registrerte noen varer under gjennomføringen av oppgave 6. Da hun oppdaget at hun burde ha benyttet en annen funksjon, var hun påpasselig med å slette hver enkelt varelinje manuelt inne på den gale funksjon før hun valgte ”lukk” og gikk til rett funksjon.

To andre brukbarhetsproblemer kan skyldes enkelheten i desktoptesten, men kan også være bare en tilfeldighet. De problemene ble oppdaget på bakgrunn av rask bruk av håndterminalen, og en medvirkende årsak kan være at de satt foran et bord når de løste oppgavene. Problem 77

ble oppdaget av to personer og er gradert som kritisk og kommer av rask skanning. Testpersonene skannet tre varer raskt og fikk tre bekreftende pip fra håndterminalen, men programmet klarte bare å registrere to varer (omtalt i detalj tidligere under kapittel 8.1.4). Det andre problemet (57) ble oppdaget av TP13 og gikk ut på at han skannet fire varer relativt raskt under varesalg i oppgave 11. I et varesalg kreves det at hver enkelt linje godkjennes av en server, og disse kallene skjer asynkront og tar noe tid. Testpersonen trodde at alt var i boks etter endt skanning og sa at kunden skulle betale summen til halvparten av varene som var godkjent av serveren. Denne summen var følgelig noe lavere enn det endelige beløpet.

Et annet problem kunne bare oppdages i desktoptesten pga oppgaveteksten. I oppgave 9 skulle de foreta et varemottak. Der manglet en av varene strekkode for at de skulle bli tvunget til å skrive inn EAN nummeret til varen. En annen vare manglet både strekkode og EAN nummer for at de skulle benytte varesøk. Av praktiske årsaker ble det brukt litt forskjellige varer i de ulike metodene. Den varen som manglet bare strekkoden i desktoptesten var en pakke med risengrøt. To av testpersonene la ikke merke til EAN nummeret, og prøvde dermed å søke opp varen. Problem 81 som da ble oppdaget, var at verken SMS-tastaturet eller ABC tastaturet på skjermen hadde bokstaven ’ø’ (’æ’ og ’å’ eksisterte). De andre fire problemene i desktoptesten kan ikke sies å være spesielt knyttet til denne metoden. Problemene kunne like gjerne dukket opp i andre metoder.

8.4.4 Problemer i snittet mellom to metoder

Dersom vi ser på snittet mellom fullskala- og felttesten, finnes det to problemer som har nær tilknytning til den ulike brukergruppen i felten. Problem 24 går ut på at tre personer i felten og en i fullskala sliter med å forstå en modal dialogboks (problemet er forklart i kapittel 8.1.1). Overvekten med antall feltpersoner som slet med dette kan skyldes mindre datakunnskaper og erfaring med ulike brukergrensesnitt-komponenter. Det andre problemet er nr 74, og går ut på at en i felt og en i fullskala sliter med å skrive bokstaver ved hjelp av SMS-tastaturet (må ikke forveksles med de som ikke klarte å benytte SMS-tastatur i problem 23).

Et annet problem i samme snitt er nr 22. Det problemet er knyttet til metoden, fordi totalt tre personer ikke la merke til at prisen på en vare var feil og burde blitt endret. Det siste problemet som er verdt å merke seg her er det kritiske problemet nr 25. Problemet ble oppdaget av en i felt og en i fullskala. Når de var ferdige med for eksempel en varetelling og skulle overføre den, satte de håndterminalen direkte i sleden. De glemte å sette varetellingen til ”ferdig”, og derfor ble den ikke overført selv om det tilsynelatende kunne se slik ut på grunn av at det startes en full synkronisering ved å dokke håndterminalen i sleden. Testpersonen fra fullskalatesten hadde erfaring fra håndterminaler i en Megabutikk. Det er derfor rimelig å anta at dette problemet oppstod på grunn av håndterminalrutinene knyttet til håndterminalløsningen Mega benytter, fordi dette stemmer overens med min observasjon av daligbruk i felten.

I snittet mellom felt og desktop er det bare et metodeproblem (54) som er interessant. Det går ut på at en person i felten ikke legger merke til at en vare mangler hylleetikett. En manglende hylleetikett ble i desktopmetoden illustrert ved å la en varelapp være utkrysset, og det var tre testpersoner i desktopmetoden som ikke koblet det mot teksten i oppgave 5.

Mellom snittet i fullskala og desktop skiller tilbakemeldingsproblem 7 seg ut, ved at alle desktoptesterne og tre av fullskalatesterne brukte opp mot ett minutt ekstra på oppgave 5 på grunn av dette problemet. Problemet oppstod når testpersonene skulle endre pris på en vare. De angir ny pris, og ser at prisen på varen i listen ikke endrer seg. Endringen blir lagt til i bongkøen og illustreres svak ved at en teller telles opp oppe i høyre hjørne (se lagertransaksjonsmeny-skjermbilde i figur 5). Akkurat hvorfor ingen i felten savnet tilbake-melding på denne handlingen vet vi ikke. Det kan ha noe med hva de er vant til med dagens COOP-håndterminalløsning.

8.4.5 Oppsummering

Oppsummert kan vi si at det var vanskelig å gjenskape konteksten fra felt og fullskala i desktopmetoden. Testpersonene slet unødvendig med varelappene, de forstod ikke illustreringen av hylleetikett og forholdet til hvor mange eksemplarer det eksisterte av en vare har skapt litt problemer. Til tross for dette ble det ikke oppdaget noen problemer i de mer realistiske metodene, som ikke kunne vært oppdaget i den forenklede desktopmetoden.

Det kommer også fram at personene i felttesten avviker en del fra de som ble testet i laboratoriet. Det er derfor ikke tilfeldig at det ble oppdaget dobbelt så mange språkproblemer i felt som i desktop med tanke på ulik brukergruppe.

Noen av problemene som ble oppdaget gir antydninger til at vi ikke bør ignorere bakgrunnserfaringene testpersonene har fra butikkarbeid. Det virker som at noen tar med seg erfaringer fra ulike håndterminalløsninger de har benyttet tidligere. Vi bør oppmerksom på dette ved tilfeldig rekruttering av personer med butikkerfaring.

8.5 Tilfredsheten blant testpersonene

I kapittel 7.5 ble det vist at det er nesten signifikant forskjell når det gjelder tilfredsheten mellom fullskalametoden og både felt- og desktopmetoden. Denne betydelige forskjellen er verdt å se litt nærmere på. Det kan være flere årsaker til denne forskjellen og derfor være vanskelig å konkludere.

Hvis vi først ser på resultatene fra de mest realistiske metodene felt og fullskala, er det flere potensielle grunner til tilfredsheten er større i fullskalametoden. Forskjellen på brukergruppene i de to metodene med tanke på alder og datakunnskaper kan bidra på flere måter. Vi ser at TP3 i felten egentlig tilfører samme brukergruppe som de i laboratoriet, og hans resultater på SUS er

88, noe som faktisk er over gjennomsnittet til de i fullskalametoden. De med lavest tilfredshet var også blant de som klarte minst oppgaver og hadde negativ adferd til egne datakunnskaper underveis.

Feltpersonene slet mer med oppgavene og det vises i tabell 6. De klarte bare å fullføre 68 % av alle oppgavene og hadde flest sammenbrudd av alle. På den andre siden klarte personene i fullskalatesten å fullføre hele 91 % av oppgavene(mest av alle). Fullskalametoden oppdaget også minst brukbarhetsproblemer, med bare 12 i gjennomsnitt. Det er derfor rimelig å anta forskjellen på brukergruppene har hatt en betydning for den lavere tilfredsheten vi ser blant testpersonene i felten.

Det er usikkert om den mentale belastningen kan ha noe å si og det ble heller ikke målt. Tidligere metodesammenligninger viser at det ikke er noen statistisk forskjell på den mentale belastningen i felten og i laboratorium (Betiol and de Abreu Cybis 2005; Nielsen, Overgaard et al. 2006). Kjeldskov og Stage (2004) sier derimot i en studie at økt fysisk aktivitet fører til signifikant økt subjektiv mental belastning.

Litt overraskende er kanskje resultatene mellom fullskala og desktopmetodene. Testpersonene i disse to metodene tilhører samme brukergruppe, og det gjør grunnlaget for sammenligning bedre. Desktoptesterne har en fullføringsprosent på 87 % og den er nesten identisk med fullskalametoden. Derimot avslørte desktopmetoden signifikant flere brukbarhetsproblemer med 16 i gjennomsnitt mot fullskalametoden sine 12 ($t= 2.69$, $p < 0.05$). I gjennomsnitt brukte desktoptesterne 10 minutt mer enn fullskalatesterne ($t= 1.89$, $p = 0.11$), og det skyldtes at de slet mer med oppgavene på grunn av mangel på realisme (hovedårsaken er varelappene som omtalt i kapittel 8.1.3).

Det førte til at testpersonene i desktopmetoden søkte opp varene i større grad. Implementasjon av varesøket på håndterminalen var dårlig, og det førte til at testpersonene i desktopmetoden brukte lengre tid og ble tydelig oppgitt. Den dårlige implementasjonen gjorde at testpersonene ofte fikk dårlig eller ingen treff overhode (problem 37). Flere av testpersonene i desktopmetoden valgte å søke opp Tine ekstra lettmelk ved å skrive ”Tine ekstra lettmelk” eller ”Tine lettmelk”, men de fikk null treff. Grunnen er at varenavnet for Tine ekstra lettmelk er skrevet på en annen måte i varedatabasen (ekstra lettmelk 1L Tine).

På intervjuet etter brukbarhetstesten svarte to i desktopmetoden at de ikke klarte å leve seg inn i rollen som butikkansatt, mens alle testpersonene i fullskalametoden sa de klarte det. Desktoptesterne klaget også i større grad over at håndterminalen var treg. Totalt sett kan det være forståelig at personene i desktopmetoden var mindre tilfreds på bakgrunn av lav realisme som førte til at de måtte søke opp mange varer.

9 Diskusjon

Det første vi vil gjøre under diskusjon er å redegjøre for kvaliteten på resultatene i dette studiet. Deretter vil vi si noe om det spennende funnet av falske positive brukbarhetsproblemet, før vi sammenligner våre resultater mot de tidligere metodesammenligningene som ble presentert i kapittel 5.1. De ulike evalueringsmetodene som ble brukt i dette studiet vil også bli diskutert i henhold til nytteverdien. Forskningsspørsmålene og hypotesene som ble stilt i innledningen vil bli besvart og vi vil komme med anbefalinger for brukbarhetstesting innen handelsbransjen. Til slutt blir det diskutert kort hva som kunne vært gjort annerledes i dette studiet.

9.1 Resultatkvalitet

Kvaliteten til resultatene i dette studiet vil bli diskutert i dette kapitlet. Vi si noe om både validiteten og reliabiliteten til studien. Reliabiliteten deler vi opp i intern reliabilitet mellom metodene i dette studiet, og reliabiliteten mellom ulike studier.

9.1.1 Validitet

Validiteten sier noe om i hvilken grad vi kan trekke gyldige konklusjoner om det som skal undersøkes i en studie ut ifra resultatene (Braut). Vanligvis skiller vi mellom ytre og indre validitet. Ytre validitet sier noe om resultatene fra en studie med et mindre utvalg mennesker, kan gjøre seg generaliserbare til mange andre mennesker (Braut). Indre validitet eksisterer hvis vi kan begrunne at for eksempel hypotesen A er årsaken til handlingen B (Oates 2006:293).

Hovedfokuset i dette studiet har vært å sammenligne tre forskjellig metoder med ulik grad av realistisk kontekst. Problemet med denne sammenligningen er validiteten til resultatene ikke er helt på topp. Det har tidligere blitt bekreftet at testpersonene i feltmetoden tilhører en annen brukergruppe, enn testpersonene i de to laboratoriemetodene. Dette gjør at en sammenligning vil ha dårlig ytre validitet. I tillegg er fem testpersoner i minste laget dersom vi skal ha noe statistisk tyngde bak resultatene, vi burde hatt minst åtte testpersoner.

Ett annet aspekt som truer validiteten er datainnsamlingen i felten. Det var nesten umulig å fange alle aspekter ved interaksjonen uten å benytte et minikamera som vist i figur 7. Tabell 13 viser at 18 % av de oppdagede brukbarhetsproblemene var tilnærmet umulig å oppdage i felten. Et annet moment som også truer validiteten er at 10 % av lyden mangler i felttesten. Det kan ha ført til at noen brukbarhetsproblemer ikke har blitt oppdaget. Ebling og John (2000) understreker viktigheten av høyttenking og hvor mange brukbarhetsproblemer det avslører.

Senere i kapittel 9.6 vil vi komme med en anbefaling rundt hva som er tilstrekkelig med realisme for å brukbarhetsteste mobile butikkløsninger. Vi anser vår anbefaling som generaliserbar til

andre lignende miljø med en statisk kontekst hvor brukerne ikke skal forholde seg direkte til mennesker.

9.1.2 Reliabilitet

Reliabiliteten til resultatene handler om studiet kan bli reprodusert og gi like resultat. Når en skal reprodusere et resultat må mest mulig identiske omstendigheter gjenskapes og det bør benyttes samme brukergruppe og testmetodikk. (Oates 2006:293)

Reliabiliteten mellom de tre metodene

Hvis vi ser på reliabiliteten til dette studiet kan den sies å være god. Grunnen til det er at det har vært en person som har gjennomført alle testene. Det ble benyttet sjekklister under opplæringen av den enkelte testperson. Oppgavene de gjennomførte var de samme og testpersonene ble behandlet likt under selve brukbarhetstestingen. Det hendte seg et par ganger at jeg gjorde en personlig feil når jeg skulle spille kunde i rollespillet på oppgave 11. Jeg skulle si at jeg ønsket avslag på en vare jeg ønsket og kjøpe, men sa i stedet at jeg ønsket rabatt. Knappen de skulle benytte heter nemlig rabatt. Denne feilen så ikke ut til å påvirke resultatene noe fordi rabattknappen var litt gjemt, og alle som fant fram til menyen hvor den var klarte å løse oppgaven uavhengig om jeg sa rabatt eller avslag.

Et annet moment som kan ha påvirket testpersonene noe er intervjuet. Etter å ha sett gjennom videofilmene la jeg merke til at jeg i noen sammenhenger var for dårlig lytter. Det hendte seg at jeg i noen tilfeller ble overengasjert på grunn av svarene noen av testpersonene ga, og lot dem nesten ikke fullføre svaret fordi jeg ville skyte inn med noe. Oppfølgingsspørmålene kunne være litt ulik mellom testperson ettersom at vi benyttet semistrukturert intervju. Disse momentene har dog hatt lite å si i metodesammenlignings øyemed, fordi dataene fra intervjuene er i liten grad vektlagt.

Reliabiliteten i forhold til testteam

Hvis vi ser på dette studiet i en større sammenheng, er trolig reliabiliteten noe svakere. I kapittel 5.5 ble et par studieevalueringer sammenlignet. Resultatene fra CUE-1,2,3 og 4 sier i klartekst at det er vanskelig å reprodusere samme resultat i forskjellige studier med forskjellig testteam. Til tross for at vi har prøvd å gjøre det under like omstendigheter. Det er spesielt oppdagelsen av brukbarhetsproblemer og alvorlighetsgraden til disse som er vanskelig å reprodusere. Dette er et problem når det kommer til evalueringsmetoder for brukbarhet. Derfor kan det sees på som en svakhet at bare en student har lett etter brukbarhetsproblemer i dette studiet, og alvorlighetsgraden ble diskutert sammen med en annen MMI student.

9.2 False positive resultater

Innen brukbarhetstesting er det et stort fokus på problemer som vi ikke finner(falske negative problemer). Det er gjort mye forskning på hvor mange testpersoner vi trenger for å avdekke størst mulig andel problemer. Vi er i mindre grad opptatt av falske positive resultater. Falske positive problemer (også kalt falske alarmer i brukbarhetslitteraturen) er problemer som tilsynelatende kan se ut til å være et brukbarhetsproblem, men som egentlig ikke er det (Gray and Salzman 1998).

> *Artificial based evaluations e.g. think-aloud protocols in laboratory evaluations or heuristic evaluations may generate false positive problems that are not really problems in everyday use.(Kjeldskov, Skov et al. 2004) etter (Molich 2003)*

Falske positive resultater kan oppstå i alle evalueringsmetoder innen brukbarhet, der gjennomføringen blir konstruert. Graden av hvor mange falske positive resultater de ulike evalueringsmetodene genererer er forskjellig. I følge Molich og Dumas (2006) har det utviklet seg en konsensus om at ekspertevalueringer genererer flere falske positive brukbarhets-problemer enn empirisk brukbarhetstesting.

Mye av forskningen som omtaler falske positive problemer innen menneske-maskin interaksjon ser på forskjellige evalueringsmetoder(UEM) opp mot hverandre. De ulike problemene som blir oppdaget i de forskjellige metodene blir analysert opp imot hverandre. Et tilsynelatende problem i en metode kan på bakgrunn av en annen metode bli tolket som en falsk positiv. (Gray and Salzman 1998)

> *Problems of interpretation arise when the number of problems identified by one UEM is compared to those identified by another. When different techniques identify different problems, do the differences represent misses for one UEM or false alarms for the other?* (Gray and Salzman 1998)

Falske positive problemer som oppstår på grunn av manglende realisme i konteksten er det derimot ikke forsket noe på. Ingen av de seks sammenligningsartiklene nevner falske positive problemer i forbindelse med manglende realisme. Et Google scholar søk gir heller ingen relevante treff. Det er nettopp i den kategorien varelappproblemet(13) hører hjemme. Desktoptesting kan altså i større grad fremprovosere falske positive resultater enn de andre metodene i denne studien. Nettopp det kan være god nok grunn til at vi bør benytte oss av litt mer realistisk kontekst for å unngå slike problemer.

9.3 Sammenlignet mot tidligere forskning

Resultatene fra tidligere forskning er ganske varierende. Det er gjerne forskjeller på hva de tester, hvordan de tester, hvem som tester og hvor de tester det. Dette kan være ganske avgjørende faktorer for resultatene vi får.

I underkapitlene skal vi se nærmere på våre resultater sett opp mot tidligere forskning. Vi går kategorisk til verks ved å diskutere anvendbarhet, effektivitet og tilfredshet hver for seg. Det er ikke alle tidligere studier som har resultater innen effektivitet og tilfredshet, og disse blir derfor ikke nevnt.

9.3.1 Anvendbarhet

I tabell 22 ser vi det er ulik fordeling av problemer i de forskjellige metodene i ulike studiene. På grunn av mange forskjellige faktorer i de studiene er det vanskelig å sammenligne dataene i tabellen mot hverandre direkte. Tabell 22 understreker bare at det er forskjell på hvilke metoder som oppdager mest problemer i studiene. Nettopp derfor er det vanskelig å komme med generelle konklusjoner om at en av metodene er bedre enn andre, ettersom dette er lite generaliserbart. Videre i anvendbarhetskapitlet skal vi se nærmere på våre resultater sett mot de andre studienes resultater rundt oppdagelsen av brukbarhetsproblemer og litt om hvorfor de har fått slike resultater.

Kapittel	Felt	fullskala	Desktop	Overlappende problemer	Unike metode-problem	Totalt
5.1.1	23 (62%)	36 (97%)		22 (59%)	15 (41%)	37
5.1.2	60 (79%)		48 (63%)	32 (42%)	44 (58%)	76
5.1.3*	22 (100%)		22 (100%)	22 (100%)	0 (0%)	22
5.1.4**	38 (88%)	41 (95%)	35 (81%)	32 (74%)	4 (9%)	43
5.1.5	171		92			
5.1.6***		36 (64%)	40 (75%)	37 (70%)	16 (30%)	53
Dette studiet	*33 (59%)*	*30 (54%)*	*33 (59%)*	*12 (21%)*	*28 (56%)*	*56*

Tabell 22 - Antall brukbarhetsproblemer som ble oppdaget i de forskjellige metodene. Denne tabellen stammer fra tabell 2 og er utvidet med resultatene fra dette studiet. Se tabell 2 for en forklaring på stjernene.

I studie til Kjeldskov et al. (2004)(kapittel 5.1.1) ser vi at fullskalametoden avdekket alle problemer bortsett fra ett. Det ene problemet som Kjeldskov et al. (2004) bare oppdaget i felten var på grunn av manglende realisme i laboratoriet. Sett i mot vår studie er det likheter i at så å si alle problemer kunne i teorien blitt oppdaget i laboratoriet, men vår studie avdekket i alle metodene bare i underkant av 60 % av alle brukbarhetsproblemer.

Fordelingen av antall problemer mellom felt og laboratorium er helt motsatt i Nielsen et al. (2006)(kapittel 5.1.2) dersom vi sammenligner mot Kjeldskov et al. (2004). Nielsen et al. (2006) målte mental belastning og oppdaget problemer knyttet til det. De sier at feltmetoden var den

eneste som klarte å avdekke problemer knyttet til mental belastning og interaksjonsstil. Hvor mange problemer det er snakk om sier de ikke nøyaktig, bare at de to problemtypene i tillegg til seks andre problemtyper stod for 27 av de totalt 76 problemene. Det kunne vært interessant å vite om de to problemtypene har ført til signifikant flere problemer i felten. I vår studie ble ikke mental belastning målt og det er derfor litt usikkert om det ville ført til oppdagelse av flere problemer.

Kallio og Kaikkonen (2005)(kapittel 5.1.3) har bare valgt å fokusere på de problemene som minimum to personer oppdaget. Det førte til at alle de problemene ble oppdaget både i felten og i laboratoriet. På grunn av det fremhever de tre problemer der det er signifikant forskjell på problemopptreden i felten mot laboratoriet. De har også et større statistisk grunnlag for å kunne måle slik signifikans med 20 testpersoner i hver metode. Det er litt for få personer i vår studie til å kunne gjøre det samme, men tre problemer (19, 45 og 61) skiller seg ut ved at nesten alle 15 har oppdaget de.

Duh et al. (2006) (kapittel 5.1.5) oppdaget nesten dobbelt så mange problemer i felten. Figur 12 viser at en hele 64 av problemene i felten var kritiske, mens de fant bare 12 i laboratoriet. Dette står i sterk kontrast til vår studie som viser minimal forskjell. Duh et al. (2006) sier at noen av problemene som ble oppdaget var relatert til bruken i felten, og kunne ikke blitt oppdaget i laboratoriet. Vår studie viser derimot at ingen problemer er direkte relatert til feltkonteksten. Duh et al. (2006) mener at forskjellene mellom metodene skyldes eksterne faktorer som støy, kjørende tog, manglende privatliv og den ekstra mentale belastningen i felten. Disse faktorene har tilsynelatende ingen innvirkning på felttesterne i vår studie. Det eneste kan være at støyen i butikkmiljøet kan ha noe å si. Problem 32 går ut på at et pip var eneste tilbakemelding, og det var bare to personer i felten som ikke la merke til lyden. Det ble heller ikke hørt av to i fullskala og tre i desktopmetoden.

Kjeldskov og Skov (2003) (kapittel 5.1.6) har en jevn fordeling av problemoppdagelser blant de tre type testene de utførte, og det er likt som resultatene i vår studie. De mener på samme måte som oss at realismen i miljøet hadde liten eller ingen påvirkning på antall brukbarhetsproblemer. Kjeldskov og Skov (2003) testet ut både vanlige folk og fagfolk i desktopmetoden, og fant ut at vanlige folk er tilsynelatende like gode til å finne problemer som fagfolk. Når det gjelder alvorlighetsgraden på problemene, ser vi i figur 13 at alle kritiske brukbarhetsproblemer blir oppdaget i fullskalametoden, mens desktopmetoden finner litt færre.

9.3.2 Effektivitet

Det er bare tre av de tidligere metodesammenligningene som har målt effektiviteten. Nielsen et al. (2006) og Kallio et al. (2005) sine studier viser at det ikke er noen forskjell på tidsbruken i felt og laboratorium. Dette til tross for at i Nielsen et al. (2006) sin studie må testpersonene i felten forflytte seg fra A til B i en by ved bruk av t-bane, mens i laboratoriet satt testpersonene på en stol og løste oppgavene. I Kallio et al. (2005) måtte testpersonene i felten gjøre noe som de i

laboratoriet ikke trengte på en oppgave, og brukte derfor signifikant lengre tid på akkurat den oppgaven. Resultatene fra disse to studiene ligner på tidsbruken i vår studie. Testpersonene i de ulike metodene i vår studie brukte nesten like mye tid hvis vi ser på figur 31. Eneste unntaket er i oppgave 4 og 5, der testpersonene i desktopmetoden bruker lengre tid på grunn av det falskt positive problemet knyttet til manglende realisme.

I studiet til Betiol og de Abreu Cybis (2005) skiller desktopmobil-metoden (begrepsforklaring i kapittel 5.1.4) seg ut ved at testpersonene i den metoden brukte signifikant lengre tid på 4 av 7 oppgaver. Dette til tross for at ingen av metodene krevde at de brukte konteksten. Alle testpersonene satt på en stol både i felt og i laboratoriet uten å måtte forholde seg til noe annet enn mobiltelefonen (eller emulatoren i "desktopemulator"). De oppgir ingen annen grunn enn at det skyldes selve metodeoppsettet. Det virker litt underlig ettersom at de benytter samme type mobiltelefon i desktopmobil og i felten. Eneste forskjellen hvis vi ser bort fra miljøet rundt, er at mobiltelefonen i desktopmobil måtte stå fast i et stativ for at det skulle være enklere å filme interaksjonen.

9.3.3 Tilfredshet

Når det gjelder tilfredsheten viser Nielsen et al. (2006) sin sammenligning at brukerne synes det samme uavhengig av evalueringsmetode. Derimot merket Duh et al. (2006) en helt klart mer negativ adferd i felten enn i laboratoriet. De mener at det skyldes at testpersonene i felten både brukte lengre til på å løse oppgavene, og at de fant nesten dobbelt så mange brukbarhetsproblemer. Hos Betiol og de Abreu Cybis (2005) er det motsatt av Duh et al. (2006). Der er testpersonene minst tilfredse med metoden desktopmobil (58 % SUS), som er en laboratoriemetode. Den andre laboratoriemetoden desktopemulator gir 73 % på SUS-skjemaet og feltmetoden gir 72 %. De sier ingenting om hvorfor, men vi skal ikke se bort ifra at det har en sammenheng med at de brukte mye lengre tid på 4 av 7 oppgaver i desktopmobil metoden. Dette kan sammenlignes med vår studie og den lavere tilfredsheten i desktopmetoden sett mot fullskalameotden. Testpersonene i desktopmetoden vår brukte i gjennomsnitt 10 minutter mer enn de andre metodene på grunn av det falske positive problemet med varelappene(se kapittel 8.1.3 og 9.2).

Det virker som at det er en sammenheng mellom tidsbruk/oppdagede problemer og tilfredshet. Det kan tyde på at testpersonene ikke blir mindre tilfredse på grunn av konteksten i felten. Målinger av mental belastning i Nielsen et al. (2006) og Betiol et al. (2005) viser at det ikke er noen forskjell i den mentale belastningen mellom felt og laboratorium.

9.3.4 Hvordan påvirkes brukeren av konteksten i felten?

Duh et al. (2006) mener at konteksten i felten har påvirket testpersonene i stor grad og har vært med på å avdekke mange flere problemer og testpersonene har hatt en negativ adferd. Felttesten ble gjennomført på et tog, og Duh et al. (2006) sier at eksterne faktorer som støy, manglende privatliv og den ekstra mentale belastningen i felten er årsaken til resultatet.

> *Although there is no unified ways to explain how these factors would affect everyone, they cannot be ignored, especially for the purposes of conducting accurate usability evaluation.* (Duh, Tan et al. 2006)

En lignende studie er fortatt av Kallio og Kaikkonen (2005). De har også testet med mobiltelefon, og testpersonene i felten har måttet reise med t-banen til et kjøpesenter og gått rundt der mens de løste oppgavene. De fant to interessante funn som er knyttet til konteksten i felten. Det første viste at når testpersonene fikk en kognitivt belastende oppgave, så oppførte de seg forskjellig fra en mindre kognitivt belastende oppgave. Når testpersonene fikk en oppgave om en funksjon de ikke var kjent med på mobiltelefonen, hendte det at de bare stirret på skjermen mens de gikk. Det var bare så vidt de unngikk å gå på folk. Det hendte seg også at de gikk til siden på en roligere plass hvor de kunne fullføre oppgaven. Dette gir innsikt i hvor vanskelig oppgavene er. I vår studie var det vanskeligere å avdekke denne observasjonen i felten. Testpersonene våre sto i ro når de skulle finne riktig funksjon, for så å gå til riktig varehylle og skanne varene.

Det andre interessante funnet i Kallio og Kaikkonen (2005) var knyttet til samme type eksterne faktorer som Duh et al. (2006) mener at skyldtes avdekkingen av flere problemer. I motsetning til Duh et al. (2006) sier Kallio og Kaikkonen (2005) at eksterne faktorer og potensielle avbrudd så ikke ut til å plage testpersonene under testen. I felten konsentrerte testpersonene seg hardt om oppgavene, og la i mindre grad merke til omgivelsene.

> *It could be interpreted that users were performing the tasks inside a bubble, which in fact is normal behavior in public places in Helsinki. People are concentrating on their own activities and ignoring what the others do unless the activity is noisy or threatening.* (Kallio and Kaikkonen 2005)

Det andre funnet til Kallio og Kaikkonen (2005) stemmer godt overens med vår studie. Testpersonene våre var også konsentrerte, og det viste seg også at de ikke ble forstyrret av kundespørsmål underveis i testingen. Funnet til Kallio og Kaikkonen (2005) og vår studie kan bekreftes av Betiol og de Abreu Cybis (2005), der de heller ikke fant noen forskjell i felten. Testpersonene i Betiol og de Abreu Cybis (2005) var mindre mobile enn de øvrige studiene, for der ble alle oppgavene løst ved at testpersonen satt.

> *Although the external environment was busy and noisy, in the field the user was not as susceptible to outside interference as expected.* (Betiol and de Abreu Cybis 2005)

9.3.5 Hvordan påvirkes brukeren av konteksten i laboratoriet?

Det er ikke mange som nevner noe om hvordan brukeren påvirkes av laboratoriekonteksten. Ett hederlig unntak er Kallio og Kaikkonen (2005). Der klaget tretten testpersoner i laboratoriet over lang nedlastningstid, mens ni i felten klaget over det samme. Ordene som ble brukt i laboratoriet viste mer frustrasjon enn i felten.

Samme fenomen er også gjeldende i vår studie. To av oppgavene medførte at håndterminalen skulle synkroniseres, og det tar cirka ett minutt. Det var tre i desktopmetoden, to i fullskalametoden og en i feltmetoden som klaget over at synkroniseringen tok lang tid. Språket til de i desktopmetoden viser mer frustrasjon enn de andre på samme måte som i Kallio og Kaikkonen (2005).

I felten sier Kallio og Kaikkonen (2005) at testpersonene gjerne brukte denne ventetiden til å sjekke egen mobil, se på andre mennesker og andre kontekstuelle ting. I laboratoriet og spesielt i desktopmetoden har testpersonene bare oppgaven og håndterminalen å forholde seg til. Det skaper et sterkt fokus på håndterminalen og kan være grunnen til at de er mer utålmodig. Nedenfor er frustrasjonsutsagnene oppsummert:

Desktop
TP14*: "Synkroniseringen tar fryktelig lang tid"*
TP13: *"Ser ikke ut til at det skjer noe. Hadde nesten lyst til å avbryte fordi det ikke skjedde noe"*
TP12: *"Synkroniseringen tar kjempelang tid"*

Fullskala
TP7: *"Fort går det ikke. Lurer nesten på om den har klikka"*
TP6: *"Synes det tar litt lang tid å oppdatere varene"*

Felt
TP3: *"Dette tok da sin tid det også, som det bruker å gjøre"*

9.4 Problemstilling, forskningsspørsmål og hypoteser

I innledningen presenterte vi en problemstilling med underliggende forskningsspørsmål. I dette kapittelet vil begge disse delene bli besvart. Vi starter med å besvare forskningsspørsmålene før vi svarer på problemstillingen til slutt.

9.4.1 Forskningsspørsmål og hypoteser

Her vil forskningsspørsmålene og deres hypotese bli besvart, og besvarelsen vil være med å danne grunnlaget for svaret på problemstillingen og den endelige konklusjonen.

A) Blir samme problemer og fenomener funnet i alle tre metodene?

> *En mer realistisk kontekst vil bidra til oppdagelsen av noen problemer som ikke er mulig å finne i metodene med mindre realisme.*

Figur 35 viser at bare 12 av 56 (21 %) brukbarhetsproblemer blir funnet i alle metodene, og halvparten av disse problemene er relatert til manglende tilbakemelding. Alle problemer som skyldes manglende tilbakemelding bortsett fra ett blir oppdaget i flere enn en metode. Denne problemtypen er helt klart den som blir funnet i flest metoder.

Hele 28 av 56 (50 %) av brukbarhetsproblemene blir funnet i bare en av metodene. Analyseringen av de unike problemene som ble oppdaget i felten (kapittel 8.4.1), viser at ingen av disse problemene kan relateres til butikkmiljøet. Dermed kan vi konkludere med at hypotesen kan avkreftes, fordi mer realistisk kontekst oppdaget ingen unike problemer knyttet til felten. Det er mulig at flere testpersoner ville ført oppdagelsen av flere problemer som ville blitt oppdaget i mer enn én metode, slik trenden i figur 14 antydninger.

B) Er alvorlighetsgraden til problemene forskjellige i de tre metodene?

> *De to mer realistiske metodene vil kunne avdekke flere alvorlige problemer på grunn av en mer realistisk kontekst, mens desktopmetoden vil finne flere kosmetiske problemer.*

Antall brukbarhetsproblemer med ulik alvorlighetsgrad mellom de forskjellige metodene har vist seg å være veldig like. Tabell 14 viser tydelig at det er bare mindre forskjeller mellom metodene, og det er langt ifra noen signifikant forskjell mellom dem.

Vi kan se antydninger til at hypotesen kan ha noe for seg ved å se nærmere på tabell 14. Feltmetoden har avdekket 23 alvorlige problemer mot desktopmetoden sine 20, og desktopmetoden har avdekket 9 kosmetiske problemer mot feltmetoden sine 6. Derimot har desktopmetoden avdekket et kritisk problem mer. Uansett er dette for små forskjeller til at vi kan dra noen konklusjon ut ifra det.

C) Vil det være forskjell på problemtypene som oppdages i de forskjellige metodene?

> *Desktopmetoden vil i større grad generere brukbarhetsproblemer knyttet til design og interaksjon med systemet.*

Det ble oppdaget betydelig flere språkproblemer i feltmetoden (8 av 12, tabell 15), og grunnen til det skyldtes at felttesterne tilhører en annen brukergruppe. Nettopp på grunn av ulik brukergruppe, kan vi ikke si at dette ville vært resultatet dersom alle hadde tilhørt samme brukergruppe. Hvis vi ser på de andre problemtypene er det bare små forskjeller mellom de ulike metodene.

I forhold til hypotesen ble det i gjennomsnitt funnet totalt 11 designproblemer med et standardavvik på 1,7. Med andre ord er det minimal forskjell mellom de ulike metodene. Det ble faktisk funnet like mange i felt og desktopmetodene med 12 problemer, mot fullskalametoden sine 9.

D) Er det forskjeller på fullføringsgraden av oppgavene i de forskjellige metodene?

> *Mer realisme i form av større kontekst vil gjøre oppgavene vanskeligere. Løsningen på oppgavene vil ikke nødvendigvis stå rett foran nesen på testpersonene, noe den i praksis vil gjøre i desktopmetoden.*

Hvis vi ser på fullføringsgraden til tabell 6-8 ser vi ingen forskjell mellom fullskala og desktopmetodene. Derimot er forskjellen mellom de to mot feltmetoden begge nesten signifikant ($p < 0.1$). Denne forskjellen skyldes også i stor grad ulik brukergruppe og dermed er validiteten for sammenligning dårlig. Det er derfor sannsynlig at forskjellen ville vært mindre dersom brukergruppene hadde vært like, men dette bør avklares i en senere studie.

E) Vil interaksjonen med håndterminalen være forskjellig ved bruk av de tre metodene?

> *Det er mer naturlig å benytte fingrene til å trykke på den trykkfølsomme skjermen i felten. De som sitter på en stol og blir testet i desktopmetoden vil på samme måte bruke pennen mer aktivt.*

Hypotesen stammer fra Kjeldskov et al. (2004) der felttesterne benyttet fingeren, mens de i laboratoriet valgte styluspennen. Dette viste seg ikke å stemme i denne studien. Styluspennen ble brukt mest, og mange uttrykte verbalt at de foretrakk å benytte den.

F) Hvilke fordeler og ulemper har de tre metodene sett opp imot hverandre?

> *Det er vanskeligere å gjennomføre en felttest i forhold til laboratorietest.*

Hypotesen stemte ganske bra og det var også som forventet. All litteratur som omtaler brukbarhetstesting i felten sier at det er vanskeligere enn å brukbarhetsteste i laboratorium.

Fordelen med felttesting er at konteksten er helt realistisk, men dette studiet har vi vist at vi fant ikke noe ekstra ved å teste i felten. Derfor kan det ikke sies å ha vært en fordel i dette studiet.

Ulempene ved å brukbarhetsteste i felten er det derimot flere av. De to siste av punktene til Kjeldskov og Stage (2004), som er omtalt i kapittel 2.6.1 viste seg å være gjeldende. Det var ikke like enkelt for testpersonene i felten å benytte høyttenkning, og det viste seg å være mye vanskeligere å få samlet inn data. Til tross for skjermopptak, ble ikke alle trykk registrert. Dersom en person prøvde å trykke på noe som ikke var en knapp eller lignende så ble ikke det registrert. Til å kompensere for det var tanken at jeg med DV-kameraet skulle prøve å få filmet mest mulig av interaksjonen fra avstand, men når det viste seg at det var umulig å få begge videostrømmene synkronisert helt riktig i ettertid hjalp det lite. Konsekvensen av dette er at noen problemer som ble oppdaget i laboratoriet, kunne være vanskelig å fange opp i felten (se ”1X2” i kapittel 7.6.3).

I fullskalametoden var det mye enklere å gjennomføre brukbarhetstestingen. Fordelene der er at vi har full kontroll over miljøet. Testpersonene sa de i stor grad skjønte konteksten, og de klarte å sette seg inn i rollen som butikkansatt. Ulempene er at settingen kan bli litt for liten, slik at den indirekte gir hint om hva testpersonen skal gjøre i de forskjellige oppgavene. Det er også mulig at den ’perfekte’ settingen kan ha bidratt til at testpersonene får en unaturlig stor tilfredshet og fullføringsrate.

Den enkleste metoden av alle var desktopmetoden, og den avdekket like mange brukbarhetsproblemer som de andre metodene. En annen interessant oppdagelse ved denne metoden var at testpersonene var raske i handlingene sine og avdekket ett kritisk og ett alvorlig brukbarhetsproblem på grunn av ’stresstestingen’ de indirekte gjorde. Ulempen var at testpersonene ikke klarte å sette seg godt nok inn i situasjonen og konteksten. Det førte til oppdagelsen av et falskt positivt brukbarhetsproblem som hadde innvirkning på effektiviteten og tilfredsheten.

G) Er det verdt det ekstra strevet å gjennomføre en felttest?

> *Sett opp imot tidligere forskning finnes det mye varierende resultat. Det kan derfor virke som at mange faktorer i stor grad kan påvirke resultatene. Dette kan for eksempel være selve systemet, testpersonene, kontekstene og scenarioene. Derfor er det usikkert hvordan det vil fortone seg med en metodesammenligning innen handelsbransjen.*

Til tross for manglende validitet på resultatene som danner grunnlaget for sammenligningen mellom metodene, mener vi at det er mulig å si noe om dette spørsmålet. Dette studiet har vist at det tilsynelatende ikke er verdt strevet å gjennomføre brukbarhetstester i felten. Alle brukbarhetsproblemene som ble funnet i felten kunne blitt oppdaget ved hjelp av enklere metoder. Derfor er den ekstra verdien av en felttest veldig liten.

Det viste seg også at andre forstyrrelser vi bare hadde i felten som kundeavbrytelser hadde liten innvirkning og testpersonene hadde ingen problemer med den statiske konteksten. Hvis vi også legger til grunn at datainnsamlingen var mye vanskeligere i felten, sier det seg selv at det ekstra strevet ikke er verdt det.

9.4.2 Problemstilling

Innledningsvis i denne studien stilte vi oss en problemstilling om hvilken av de tre metodene avbildet i figur 43 som er best egnet til evaluering av mobile butikkløsninger. Vi skal forsøke å svare på det spørsmålet på bakgrunn av blant annet forskningsspørsmålene.

Figur 43 - Viser alle tre metodene. Fra venstre er felt-, fullskala- og desktopmetoden.

Det viste seg at ingen av metodene skilte seg ut på noen spesiell måte. Det var ingen av de reelle brukbarhetsproblemene vi oppdaget som kunne knyttes til en bestemt metode, til tross for at det var stor spredning i hvor de ulike problemene ble oppdaget. Tidsbruken var ganske like i metodene hvis vi ser bort fra det falske positive problemet. Brukertilfredsheten var litt forskjellig mellom metodene og kan skyldes forskjellige årsaker. Derfor er det vanskelig å si noe helt sikkert om forskjellene i brukertilfredsheten.

På bakgrunn av at den realistiske konteksten i felten ga ingen ekstra verdi, og med tanke på ulempene i felten vil vi anbefale at laboratorietesting er tilstrekkelig når det gjelder mobile butikkløsninger. Det viste seg at en kontekstsimulering i fullskalametoden heller ikke avdekket noen unike kontekstuelle problemer, og derfor kan vi konkludere med at den enkleste metoden faktisk gir bra resultater i denne studien. Svakheten med desktopmetoden var genereringen av falske positive problemer, og derfor kan det være fordelaktig å benytte seg av en metode midt i mellomting fullskala og desktop. Denne mellomtingen vil bli presentert nærmere i kapittel 9.6, hvor vi kommer med anbefalinger til brukbarhetstesting av mobile butikkløsninger. Den metoden vi anbefaler der er en mellomting mellom fullskala og desktop, og er den metoden vi tror er best egnet til å evaluere mobile butikkløsninger.

9.5 Metodediskusjon

I metodediskusjon skal vi se nærmere på nytteverdien til de forskjellige evalueringsmetodene som ble brukt i dette studiet. Vi vil si litt om hva som var lurt og hva som ikke var så lurt.

9.5.1 Observasjon

Observasjonen av hvordan de butikkansatte til daglig brukte håndterminalen i felten har vært nyttig. Med bakgrunn i observasjonen kunne vi relatere brukbarhetsproblem 25 til håndterminalrutinene til COOP.

9.5.2 Intervju

Intervjuet etter brukbarhetstesten fungerte greit som en arena for å spørre ut om handlinger de hadde gjort under testen, og andre interessante synspunkt. Bortsett fra det har ikke resultatene fra intervjuene blitt brukt i så stor grad, men det har helt klart vært en bra evalueringsmetode for å supplere resultatene fra brukbarhetstestingen. I etterkant oppdaget vi også visse tema det hadde vært interessant å spurt testpersonene om.

9.5.3 Brukbarhetstesting

Brukbarhetstestingen gikk fint, men det var et par ekstra utfordringer i felten som nevnt i kapittel 6.4.3. Det hadde vært en stor fordel med et minikamera som vi ser i figur 7. Minikameraet ville hjulpet godt for å fange interaksjonen når testpersonene trykket på håndterminalen. I dette studiet var det vanskelig å filme interaksjonen i felten, og et slikt minikamera ville også vært nyttig i laboratoriet.

9.5.4 Spørreundersøkelse

Både pretest spørreskjemaet og SUS-spørreskjemaet kom godt med. Uten pretest-spørreskjemaet kunne vi ikke klart å definere de ulike brukergruppene, og konkludert med at de forskjellige brukergruppene forringet resultatgrunnlaget for å sammenligne metodene. Ut i fra SUS-spørreskjemaet kunne vi lese ut at tilfredsheten var lavere i desktopmetoden enn i fullskalametoden, og at det trolig skyldes det falske positive problemet.

9.5.5 Logging

Det var bra vi la inn logging i programmet med tanke på problemet med skjermopptaket til TP2 i felten som nevnt i kapittel 6.4.3. Uten det ville vi ikke visst hva TP2 foretok på håndterminalen det siste kvarteret av brukbarhetstesten. Hvis vi ser bort fra den hendelsen, ble ikke loggen brukt

til noe. Dersom vi benytter minikamera som vist på figur 7, eller er trygg på at skjermopptaket fungerer 100 % er det helt unødvendig å bruke tid og krefter på å lage et loggesystem. Det gikk med mange timer på å lage fullstendig interaksjonslogging i programmet.

9.6 Anbefalinger for brukbarhetstesting i handelsbransjen

Dette kapitlet vil oppsummere noen anbefalinger rundt brukbarhetstesting i handelsbransjen på bakgrunn av dette studiet. Vi vil si noe om graden av realisme som er nødvendig for å finne relevante brukbarhetsproblemer. Vi kommer til å se på hva slags elementer som er viktige og hva som er uviktige momenter. Til slutt vil vi komme med en oppsummering over hva det holder å ta tak i når det skal brukbarhetstestes mobile enheter for handelsbransjen.

De butikkansatte i felten jobber i stor grad selvstendig når de benytter en håndterminal. Under testingen ble det notert ned tolv avbrytelser av kunder og andre butikkansatte. I snitt ble testpersonene avbrutt en gang cirka hvert tiende minutt når de befant seg i butikkarealet. Ingen av disse avbrytelsene hadde noen nevneverdig innflytelse på oppgaveløsingen de holdt på med. Grunnen til at avbrytelsene ikke hadde noen effekt på de butikkansatte er trolig fordi de er så pass vant til å bli avbrutt til dagelig. Det er derfor rimelig å anta at statister som avbryter en butikkansatt under testing i et brukbarhetslaboratorium vil ha ingen effekt på resultatene.

En annen sentral faktor for brukbarhetstestingen er konteksten. Hvis vi ser på tabell 13 og analysen i kapittel 8.4, ser vi at det er få brukbarhetsproblemer som har sitt utspring i konteksten. Faktisk er ingen av problemene som ble oppdaget i felt- og fullskalametoden knyttet direkte til den aktuelle metodekonteksten. Derfor kan vi konkludere med at alle tre metodene i teorien kunne avdekket samtlige problemer.

Ulike brukergrupper i de forskjellige metodene har vært med å undergrave validiteten til sammenligningen i dette studiet. Hell i uhell, kan vi på bakgrunn av det si noe om viktigheten av å brukbarhetsteste med ulike brukergrupper. Det viste seg å være signifikant forskjell på datakunnskapene mellom testpersonene i felten mot de i laboratoriet. Brukergruppen med minst datakunnskap slet i større grad med å skrive inn tekst og hadde problemer med noe teknisk språk. Derfor anbefaler vi at det testes med personer fra like brukergrupper.

Vi bør være bevisst på hva slags butikkløsninger de vi brukbarhetstester har erfaring med. Minst to av problemene som ble avdekket stammer fra at testpersonene handler på bakgrunn av tidligere erfaring med slike håndterminaler. Vi må være klar over at det kan dukke opp problemer på bakgrunn av erfaring til den enkelte testperson. Et eksempel på dette er problem 25. Det problemet skyldes at testpersonene fra COOP ikke er vant til å kunne gjøre flere jobber før de synkroniserer. Når en jobb var ferdig gikk de for å synkronisere den, og det var på grunn av rutiner fra COOP håndterminaler. Et annet eksempel er problem 30, og det problemet kommer av at testperson tre har erfaring med håndterminaler uten trykkfølsom skjerm fra Byggmakker. Det var også flere testpersoner fra laboratoriet som kommenterte at de var litt

usikre på hva noen av fagbegrepene betydde i forhold til deres erfaring. Dette ga ingen problemer under gjennomføringen av oppgavene, fordi de etter hvert skjønte hva begrepene betydde.

Figur 44 - Anbefalt testoppsett med tilstrekkelig realisme.

Konklusjonen er at strevet med en felttest er unødvendig når vi skal brukbarhetsteste en håndterminalløsning til bruk i handelsbransjen. Det vil holde lenge å teste i et laboratorium. I et laboratorium bør det benyttes en mellomting mellom fullskala- og desktopmetoden. En enkel desktopmetode med ekte varer og hylleetiketter liggende foran varene vil trolig være nok realisme og er illustrert i figur 44. Hvis det skal benyttes en slede kan det være en fordel å benytte to bord i rommet, der sleden står på det ene og varene på det andre som vist i figur 44. Dette bør gjøres for å unngå at 'løsningen' står rett foran ansiktet på vedkommende, slik tilfellet var i desktopmetoden. I tillegg bør det benyttes testpersoner fra flere enn en brukergruppe.

9.7 Hva ville jeg gjort annerledes hvis jeg skulle gjort det om igjen?

Dersom jeg skulle gjort dette studiet om igjen er det et par momenter jeg ville gjort annerledes. Det viktigste momentet er at jeg ville satt flere krav til hva slags testpersoner vi skulle rekruttere i tillegg til kravet om butikkerfaring. Jeg burde passet på at alder og datakunnskapene var mer lik, slik at validiteten ved sammenligning av resultatene fra de forskjellige metodene hadde blitt bedre. Det er enkelt å si hva slags personer vi vil ha. Rekrutteringsarbeidet vi gjorde i dette studiet viste at det var bare en person (TP9) som meldte seg frivillig etter en rekrutteringsrunde blant fire dagligvarebutikker i Trondheim. Det hadde også vært en fordel med flere personer i de ulike metodene. Det holder å teste fem personer i en utviklingsprosess, men i en vitenskapelig

sammenligning av metoder bør vi benytte minimum åtte personer i hver metode for å ha nok statistisk grunnlag.

Underveis i testingen og i etterkant har jeg ergret meg over at jeg ikke stilte et par spørsmål som det hadde vært interessant å få et svar på. På pretest-spørreskjemaet kunne jeg tenkt meg å stilt et spørsmål om skriving av meldinger på mobiltelefon, ettersom at flere testpersoner i felten slet med å skrive på SMS-tastatur. I etterkant burde jeg blant annet spurt om hva slags interaksjonsmetode testpersonene foretrakk (finger, styluspenn eller tastatur), hva de foretrakk best av SMS eller ABC-tastatur. Det kunne også vært interessant og målt kognitiv last for å se det er forskjeller på belastningen i de forskjellige metodene og om det kan være med på å avsløre noen flere problemer i felten som i studiet til Nielsen et al. (2006).

10 Konklusjon

Dette kapittelet vil gi en konklusjon for denne studien. Til slutt sier vi litt om hva det bør arbeides videre med innen dette temaet.

10.1 Konklusjon

I denne studien har vi brukbarhetstestet en mobil butikkløsning i felten og i laboratorium, hvor vi har sammenlignet tre forskjellige metoder. Målet har vært å se på hva slags resultater vi får fra de ulike metodene, å finne ut hva slags metode som har fungert best. Det har tidligere blitt gjort en del studier som sammenligner resultatene fra felt og en type laboratorietest, men ikke mange har testet både fullskala og desktopmetoder. Det har heller ikke blitt gjort noen slike sammenligninger av systemer for bruk i handelsbransjen.

Grunnlaget for sammenligningen i dette studiet har dessverre et par validitetssvakheter. For det første er fem testpersoner i hver metode for få til å kunne ha tilstrekkelig statistisk tyngde bak dataene. Den andre svakheten går ut på at testpersonene i felten skiller seg ut som en egen brukergruppe, med signifikant høyere alder og lavere datakunnskaper, og det har ført til litt forskjellig problemoppdagelse.

Vi ser at oppdagelsen av brukbarhetsproblemer i de ulike metodene i tabell 14 er jevn mellom alle metodene og det er ingen som skiller seg ut, verken på antall eller alvorlighetsgrad. Hvis vi også ser på hvilke problemer programmets systemleverandør ønsker å løse i figur 42, er det fortsatt en jevn fordeling mellom metodene. Det som er interessant å merke seg er at systemleverandørene ikke ønsker å fikse det eneste kritiske problemet som er unikt for feltmetoden. Det vil si at laboratoriemetodene fant alle prioriterte kritiske problemer, og feltmetoden fant bare 3 av 7.

I desktopmetoden fant vi et problem som er falskt positivt, og det problemet oppstod på grunn av lav realisme. Testpersonene klarte ikke å forstå at de kunne benytte varelappene med strekkode under hele testen (se figur 40). Det førte til at de benyttet et dårlig implementert varesøk, i stedet for å skanne varelappene. Ringvirkningene fra dette førte til at de brukte mye lengre tid på to oppgaver enn de som skannet varene i de andre metodene, og ble av den grunn mindre tilfreds. Testpersonene i desktopmetoden bruker tilsynelatende signifikant lengre tid på alle oppgavene, men det skyldes dette falske positive problemet. Derfor kan vi konkludere med at det er ingen forskjell i tidsbruken mellom metodene.

Det oppstod flere sammenbrudd i felten på oppgaver som spilte mye på konteksten, og det kunne tilsynelatende virke som at det var årsaken. Det viste seg derimot at problemet lå i brukergruppen. Testpersonene i felten oppdaget flere språkproblemer på grunn av teknisk språk

og bruk av engelske ord. I figur 35 ser vi at de i felten oppdaget signifikant flere språkproblemer enn de andre, og to av disse problemene var årsaken til flere sammenbrudd på oppgave 3 og 7.

Hele 5 av de 10 unike problemene som bare oppstod i felten skyldes brukergruppen. To av problemene var de nevnte språkproblemene, mens de tre siste var relatert til å skrive inn tekst på håndterminalen. Det viste seg at alle problemene som ble oppdaget i felten også kunne blitt oppdaget i laboratoriet. På samme måte kunne alle problemene som ble funnet i laboratoriet bli funnet i felten. Ingen av problemene som ble oppdaget i denne studien var direkte knyttet til en bestemt metode, hvis vi ser bort fra det falske positive problemet.

Vi konkluderer derfor med at utfordringene og ekstraarbeidet en felttest medfører er unødvendig når vi skal brukbarhetsteste en mobil butikkløsning. Vår studie viste at de butikkansatte er vant til kundeavbrytelser, og at det ikke forstyrret dem i utførelsen av oppgavene. Vi fant like mange problemer i desktopmetoden som i feltmetoden, og ettersom at ingen problemer er unike for sin metode, kan vi anbefale å bruke en enklere metode. For å unngå problemer med for lav realisme, vil vi anbefale en mellomting mellom desktop- og fullskalametoden, hvor vi benytter ekte varer i desktopmetoden som illustrert i figur 44.

10.2 Videre arbeid

Det kunne vært interessant å gjennomføre denne studien med flere personer. For at sammenligningsgrunnlaget skal være bedre, bør det benyttes personer som tilhører samme brukergruppe. I en slik studie bør den anbefalte desktopmetoden med ekte varer prøves ut, i tillegg til de tre metodene som ble testet i denne studien for å se om det er hold i anbefalingen.

11 Referanser

Betiol, A. and W. de Abreu Cybis (2005). "Usability testing of mobile devices: A comparison of three approaches." LECTURE NOTES IN COMPUTER SCIENCE **3585**: 470.

Blum, R. and K. Khakzar (2007). "Design Guidelines for PDA User Interfaces in the Context of Retail Sales Support." LECTURE NOTES IN COMPUTER SCIENCE **4551**: 226.

Braut, G. S. "Store norske leksikon, snl.no." Retrieved 17. May, 2009, from http://www.snl.no/.sml_artikkel/validitet.

Brewster, S. (2002). "Overcoming the lack of screen space on mobile computers." Personal and Ubiquitous Computing **6**(3): 188-205.

Brooke, J. (1996). "SUS-A quick and dirty usability scale." Usability Evaluation in Industry: 189-194.

Chang, D., L. Dooley, et al. (2002). Gestalt theory in visual screen design: a new look at an old subject, Australian Computer Society, Inc. Darlinghurst, Australia, Australia.

Dahl, Y., O. Alsos, et al. (2009). "Evaluating Mobile Usability: The Role of Fidelity in Full-Scale Laboratory Simulations with Mobile ICT for Hospitals." J.A. Jacko (Ed.): Human-Computer Interaction: 232-241.

Duh, H., G. Tan, et al. (2006). Usability evaluation for mobile device: a comparison of laboratory and field tests, ACM New York, NY, USA.

Dumas, J. (1988). Designing user interfaces for software, Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

Dumas, J. and J. Redish (1999). A practical guide to usability testing, Intellect Books.

Ebling, M. and B. John (2000). "On the Contributions of Different Empirical Data in Usability Testing."

Faulkner, L. (2003). "Beyond the five-user assumption: Benefits of increased sample sizes in usability testing." Behavior Research Methods, Instruments, & Computers **35**(3): 379-383.

Gray, W. and M. Salzman (1998). "Damaged merchandise? A review of experiments that compare usability evaluation methods." Human-Computer Interaction **13**(3): 203-261.

Hewett, T., R. Baecker, et al. (1996). "ACM SIGCHI curricula for human-computer interaction." Retrieved April **1**: 2004.

Hornbæk, K. and E. Law (2007). Meta-analysis of correlations among usability measures, ACM New York, NY, USA.

ISO (1997). "9241-11." Ergonomic requirements for office work with visual display terminals (VDT's). Part **11**.

ISO (2006). "ISO/IEC, ISO/IEC 25062." Software Engineering - Software product Quality Requirements and
Evaluation (SQuaRE)-Common Industry Format (CIF) for Usability Test Reports. **ISO/IEC 25062:2006**.

Jacko, J. and A. Sears (2003). The human-computer interaction handbook: fundamentals, evolving technologies, and emerging applications, Lawrence Erlbaum Associates.

Kallio, T. and A. Kaikkonen (2005). "Usability testing of mobile applications: A comparison between laboratory and field testing." Journal of Usability Studies **1**: 4-16.

Kjeldskov, J. and C. Graham (2003). "A Review of Mobile HCI Research Methods." LECTURE NOTES IN COMPUTER SCIENCE: 317-335.

Kjeldskov, J. and M. Skov (2003). Creating realistic laboratory settings: comparative studies of three think-aloud usability evaluations of a mobile system.

Kjeldskov, J., M. Skov, et al. (2004). "Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field." LECTURE NOTES IN COMPUTER SCIENCE: 61-73.

Kjeldskov, J. and J. Stage (2004). "New techniques for usability evaluation of mobile systems." International Journal of Human-Computer Studies **60**(5-6): 599-620.

Molich, R. (2003). "Brugervenligt webdesign, 2. udgave." Ingeniøren/Bøger, side **23**.

Molich, R. and J. Dumas (2006). "Comparative usability evaluation (CUE-4)." Behaviour & Information Technology(1): 1-19.

Newcomb, E., T. Pashley, et al. (2003). Mobile computing in the retail arena, ACM New York, NY, USA.

Nielsen, C. (1998). Testing in the Field. Proceedings of the Third Asia Pacific Computer Human Interaction Conference (APCHI 98), California.

Nielsen, C., M. Overgaard, et al. (2006). It's worth the hassle!: the added value of evaluating the usability of mobile systems in the field, ACM Press New York, NY, USA.

Nielsen, J. "How to Conduct a Heuristic Evaluation." Retrieved 3. May, 2009, from http://www.useit.com/papers/heuristic/heuristic_evaluation.html.

Nielsen, J. (1993). Usability engineering, Morgan Kaufmann.

Nielsen, J. (1994). Usability inspection methods, ACM New York, NY, USA.

Nielsen, J. (2000). "Why You Only Need to Test With 5 Users." Retrieved 31. March, 2009, from http://www.useit.com/alertbox/20000319.html.

Nielsen, J. (2001). "Ten usability heuristics." Retrieved October. Retrieved 23. March, 2009, from http://www.useit.com/papers/heuristic/heuristic_list.html.

Norman, D. and B. Collyer (2002). The design of everyday things, Basic Books New York.

Norman, D. A. (2002). The design of everyday things. [New York], Basic Books.

Oates, B. J. (2006). "Researching Information Systems and Computing."

Sharp, H., Y. Rogers, et al. (2007). Interaction design: Beyond human computer interaction, John Wiley & Sons.

Shneiderman, B. (1992). Designing the User Interface: Strategies for Effective Human-computer Interaction, Addison Wesley Publishing Company.

Simpson, H. (1985). Design of user-friendly programs for small computers, McGraw-Hill Companies.

Svanæs, D., O. Alsos, et al. (2008). "Usability testing of mobile ICT for clinical settings: Methodological and practical challenges." International Journal of Medical Informatics.

Svanæs, D., A. Das, et al. (2008). "The Contextual Nature of Usability and its Relevance to Medical Informatics." Stud Health Technol Inform **136**: 541-546.

TDT4180 (2006). "Menneske-maskin interaksjon." Forelesningsfoil.

Theofanos, M. (2005). "Common Industry Format approved as international standard." interactions **12**(5): 46-47.

Theofanos, M. (2006). "A practical guide to the CIF: usability measurements." interactions **13**(6): 34-37.

Tognazzini, B. (1992). TOG on Interface, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Weiss, S. (2002). Handheld usability, ACM New York, NY, USA.

Vedlegg I. Oppgaver

Oppgave 1
Start programmet Lindbak POS Mobile

Oppgave 2
Du skal logge deg inn med operatør 8 og passord 88.

Oppgave 3
Før du begynner å bruke håndterminalen, bør du passe på at varedatabasen er oppdatert og ikke eldre enn 5 dager.

Oppgave 4
Det er januar og på tide å telle opp butikkbeholdningen. Du har fått i oppgave å telle alle pakkene med ferdig pizzabunn fra Toro. Fullfør operasjonen når alt er telt og registrert.

Oppgave 5
Coop Mega i Verdal har gått tom for diverse juice og trenger desperat mer juice. Din butikk har gått med på å hjelpe dem. En butikkansatt fra Verdal har allerede vært innom tidligere i dag og hentet juice. I hastverket rakk han ikke å snakke med personalet i Steinkjer, men han la igjen en postitlapp med hva han tok med seg.

Registrer dette på håndterminalen. Dersom noen av prisene er ukorrekt må du passe på å korrigere disse. Hvis noen av varene mangler hylleetikett du må bestille nye.

50stk. X-tra appelsinjuice 1.5L
100stk. Røra eplejuice 1.5L
75stk. Coop frokostjuice 1.5L

Oppgave 6
Mens du rydder og fjerner unødvendig papp i varehyllen, oppdager du plutselig at 5 poser med ”PIZZABUNN FIN TORO 216G” har falt bak varehyllen. Disse ble derfor ikke med i den varetellingen du foretok i oppgave 4. Sørg for at varetellingen blir korrekt.

Oppgave 7

Du har registrert et par bonger mens du har vært offline. Sørg for at disse bongene blir sendt over til bakromssystemet.

Oppgave 8

En kunde kommer inn og vil ha Rogaland avis, men din butikk fører naturligvis ikke den avisen. Kunden spør om du kan finne nærmeste Megabutikk som har avisen inne som en del av butikkbeholdningen.

Oppgave 9

Bring express har kommet med en pakke med varer. Registrer disse og overfør de til bakromssystemet.

Oppgave 10

Coop Mega ønsker flere bedriftskunder og har i den forbindelse gjort det mulig for bedrifter å sende inn en mail med alle varene de ønsker, slik at disse står klar til avhenting dagen etter.

Din jobb er å klargjøre denne forespørselen slik at varene er klar til avhenting. Du skal registrere varene på håndterminalen slik at ordren kan betales i en hvilket som helst kasse når kunden kommer for å hente varene. Husk å merk ordren med ID slik at det ikke blir noen forvekslinger.

Fra: Fagerheim barnehage
Sendt: 19. januar 2009 08:53
Til: steinkjer@mega.coop.no
Emne: Vareklargjøring

Hei

Vi på Fagerheim barnehage ønsker følgende varer klargjort til avhenting i morgen

Bjørn havregryn små 750g
Indisk ris boil in bag
2 pakker Toro risengrøt 245g
10 pakker Toro tomatsuppe

Vår avtale id er: mx23

Hilsen
Gerd
Fagerheim BH

Oppgave 11

Rollespill med kunde.

Vedlegg II. Rekrutteringsplakat mot dagligvarebutikker

Lei av dataprogrammer som ikke er brukervennlige?

Bli med som testperson i et brukervennlighetsstudie og få 250Kr rett i lomma for 1 times arbeid

NÅR: Uke 5, (27-30. Januar). Ønsket time avtales

HVOR: Norsk senter for elektronisk pasientjournal (NSEP) ved innkjørslen til St. Olavs Hospital (se baksiden for veibeskrivelse)

HVEM: Personer med erfaring fra butikkarbeid er eneste kriterie. Du trenger ingen erfaring fra bruken av håndterminaler. Det mobile kassedataprogrammet som blir benyttet er Lindbak POS Mobile(se bilde) og er ikke sluppet ut på markedet enda.

Kort om studiet

Kenneth Devik er masterstudent ved NTNU i Trondheim, han skriver oppgave om brukervennlighetstesting av mobile enheter i forskjellige miljø skoleåret 08/09.

Hvordan skal brukertesten gjennomføres?

Selve testen skal utføres i et miljø som ligner mest mulig på en vanlig butikk. Du skal prøve å utføre et par scenarioer med håndterminalen. Hvert scenario er kort og enkelt beskrevet. Eksempler på slike scenarioer er at du skal bestille flere varer, eller for eksempel foreta en varetelling.

Påmelding

Kontakt Kenneth Devik snarest dersom du ønsker å være testperson. Oppgi når og hvilke av de angitte dagene du kan delta, så avtaler vi et nærmere tidspunkt senere.

Telefon: 92 60 63 40
Epost og MSN: kenneth@devik.net
Facebook: Kenneth Devik

Ta kontakt dersom du har spørsmål eller bare er nysgjerrig

Vedlegg III. Rekrutteringsplakat

Har du butikkerfaring?

Bli med som testperson i et brukervennlighetsstudie og

få 250Kr for 1 times arbeid

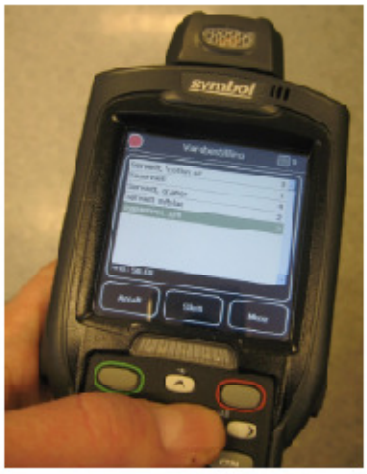
NÅR: Torsdag 29. eller fredag 30. Januar. (evt. 2-4 feb.) Time må avtales.

HVOR: Norsk senter for elektronisk pasientjournal (NSEP) ved innkjørslen til St. Olavs Hospital.

HVEM: Personer med erfaring fra butikkarbeid er eneste kriterie. Du trenger ingen erfaring fra bruken av håndterminaler. Det mobile kassedataprogrammet som blir benyttet er Lindbak POS Mobile(se bilde) og er ikke sluppet ut på markedet enda.

Kort om studiet

Kenneth Devik er masterstudent ved NTNU i Trondheim, han skriver oppgave om brukervennlighetstesting av mobile enheter i forskjellige miljø skoleåret 08/09.



Hvordan skal brukertesten gjennomføres?

Selve testen skal utføres i et miljø som ligner mest mulig på en vanlig butikk. Du skal prøve å utføre et par oppgaver med håndterminalen. Hver oppgave er kort og enkelt beskrevet. Eksempler på slike oppgaver er at du skal bestille flere varer, eller foreta en varetelling.

Påmelding

Kontakt Kenneth Devik snarest dersom du ønsker å være testperson. Oppgi når og hvilke av de angitte dagene du kan delta, så avtaler vi et nærmere tidspunkt senere.

Telefon: 92 60 65 40 Epost: kennethj@stud.ntnu.no

Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no
Brukertest: kennethj@stud.ntnu.no

Vedlegg IV. Pretest spørreskjema

1. Alder?
2. Kjønn?
3. Er du student?
    a. Hva studerer du?
    b. Hvilket årstrinn er du på?
4. Hva slags butikkrelaterte jobber har du hatt og hva slags arbeidsoppgaver har utførte du der? (stikkord)
5. Benytter du en håndterminal i forbindelse med butikkjobben? JA ☐ NEI ☐
6. Bruker du en datamaskin regelmessig? JA ☐ NEI ☐
7. Hvordan er dine datakunnskaper sammenlignet med en gjennomsnittlig normann?

Mindre god				Meget god
1	2	3	4	5
☐	☐	☐	☐	☐

8. Benytter du andre mindre enheter, som MP3-spiller, iPod eller lignende privat?
   JA ☐ NEI ☐
9. Har du mobiltelefon? JA ☐ NEI ☐

Vedlegg V. Sjekkliste for gjennomføring av brukbarhetstest

1. Ønske velkommen og introdusere deg selv
2. Beskriv hensikten med testen
3. Fortell detakerene at de kan avbryte når de vil
4. Beskriv utstyret som blir benyttet
5. Lær bort hvordan en tenker høyt ved hjelp av praktiske eksempler i Lindbak POS
6. Få testpersonen til å besvare spørreskjemaet om bakgrunn.
7. Gi opplæring i bruk av håndterminalen og nødvendig informasjon om testprogrammet. (Se vedlegg VI)
8. Sette testpersonen inn i rollen som butikkansatt og forklaresettingen med bruk av butikk og lagerområdet.
9. Forklare at jeg ikke kan tilby hjelp under testen
10. Spørre om det er noe de lurer på før testen startes
11. Gjennomfør selve brukbarhetstesten.
12. Få testpersonen til å besvare SUS skjemaet
13. Intervju. (Se vedlegg VII)
14. Takke for at testpersonen stilte opp.

Vedlegg VI. Opplæring

Håndterminalen

- Vis hvordan strekkodeleser fungerer
- Forklart at håndterminalen har touchscreen, og at de kan benytte finger eller styluspenn
- Vis hvordan de benytter SMS-tastaturet og ABC-tastaturet som benytter seg av full tastatur på touchskjermen.
- Forklart at håndterminalen har innebygd trådløst nettverk, for å kunne kommunisere med bakromspcen.

Bakrompc/slede

- Sleden brukes til å lade håndterminalen og til å koble den mot bakrompc
- Bakrompc er koblet til det trådløse nettverket, for å kunne kommunisere med håndterminalen trådløst
- Ingen av oppgavene vil kreve at du skal gå inn på bakromspc.

Lindbak POS Mobile

- Har 3 tilstander. Dokket, online og offline. Programmet vil kunne oppføre seg forskjellig i de 3 tilstandene. F.eks. er det noen funksjoner som krever at de er på nett.
- Støtten for bruk av bestillingsnummer er ikke ferdig, dvs at hylleetiketten ikke kan skannes.

Feltrelatert

- Hele oppsettet kjører seperat, uten kobling mot driftssystemet. Det er derfor ikke mulig å gjøre noe galt.
- Håndterminalen inneholder en kopi av varedatabasen til en Megabutikk.

Vedlegg VII. Intervjuveiledning

1. Hva synes du om realismen i...
    a. Settingen
    b. Oppgavene
2. Jeg skal gjennomføre akkurat de samme oppgavene på testpersoner i et rom som er satt opp til å ligne mest mulig på en butikk. Tror du jeg vil dårligere, like gode eller bedre resultater der kontra testingen vi nå har gjort i butikken?
3. Lett å bruke
4. Hvor lang tid tror du det tar for en helt person å lære seg å bruke håndterminalen?
5. Hvoran er løsningen sammenlignet med dagens rutiner
    a. Annet grensesnitt
    b. Online funksjonaliteten
6. Vil denne løsningen medføre endring av rutiner?
7. Er denne løsningen god nok til å settes I drift?
8. Er det noe du savner i denne løsningen?
9. Har du noen forslag til ny funksjonalitet, eller endring av eksisterende funksjonalitet?
10. Eventuelt en gjennomgang av oppgavene der testpersonen mistolket eller opplevde problemer med oppgaven.
11. Annet