

Osallistujien valinta

18

Jenni Anttonen

Osallistavat käytettävyyden arviointimenetelmät ovat sellaisia arviointimenetelmiä, joihin osallistuu käyttäjiä. Näille menetelmille on yhteistä, että niissä on päätettävä, keitä osallistujiksi valitaan ja kuinka monta osallistujaa tarvitaan. Osallistujien valinnassa keskeisin tekijä on osallistujien edustavuus. Osallistajat ovat edustavia, jos he ovat sovelluksen todellisia käyttäjiä tai jos he ovat sovelluksen käytön kannalta merkityksellisten ominaisuuksien suhteen mahdollisimman samanlaisia kuin todelliset käyttäjät. Edustavien osallistujien löytämiseksi tuotteen käyttäjäpopulaatio on ensin jaettava käyttäjäryhmiin. Käyttäjäryhmien muodostamiseksi on erotettava, mitkä käyttäjien ominaisuudet ovat merkityksellisiä sovelluksen käytön kannalta. Seuraavaksi keskeiset käyttäjäryhmiä kuvaavat ominaisuudet muunnetaan mitattavissa oleviksi kvantitatiivisiksi kriteereiksi, joiden perusteella osallistujia voidaan valita. Toinen tärkeä asia osallistujien valinnassa on otoskoon määrittäminen. Tarvittavaan otoskoon vaikuttavat ennen kaikkea, tehdäänkö formatiivista vai summatiivista käytettävyyden arviointia, montako käyttäjäryhmää testataan ja ollaanko kiinnostuneita tilastollisesta merkitsevyydestä tai tulosten yleistettävyydestä koko populaatioon. Jos sovelluksen käytettävyyttä testataan vain yhden ainoan kerran, otoskoon on oltava suurempi kuin yksittäisessä testissä, joka kuuluu usean iteraatiokierroksen muodostamaan testisarjaan. Lisäksi otoskoon suuruuteen vaikuttavat käytettävä menetelmä ja testin järjestäjän taidot.

18.1. Johdanto

Käytettävyyden arviointimenetelmät voidaan jakaa osallistaviin ja ei-osallistaviin sillä perusteella, onko niissä mukana käyttäjiä vai ei. *Osallistavia käytettävyyden arviointimenetelmiä* ovat muun muassa käytettävyydesti, haastattelu, kyselytutkimus, fokusryhmät, osallistava ryhmäläpikäynti ja tilannetutkimus. Kaikkiin osallistaviin käytettävyyden arviointimenetelmiin liittyy kysymys, kuinka osallistajat tulisi valita ja kuinka monta osallistujaa tarvitaan luotettavien tuloksien saamiseksi. Suurin osa kirjallisuudesta käsittelee osallistujien valitsemista käytettävyydestiin, mutta tässä luvussa esitettävät osallistujien valinnan pääperiaatteet ovat sovellettavissa muihinkin osallistaviin käytettävyyden arviointimenetelmiin.

Osallistujien edustavuus (representativeness) on keskeisin tekijä osallistujien valinnassa. Edustavuus liittyy keskeisesti *ulkoiseen validiteettiin* eli tulosten yleistettävyyteen otoksesta koko populaatioon: pätevätkö tietyssä tilanteessa, tietyssä hetkenä ja tietyillä käyttäjillä saadut tulokset koko käyttäjäpopulaatioon (Holleran, 1991; Robson, 1994)? Osallistajat ovat edustavia, jos he ovat sovelluksen todellisia käyttäjiä tai jos he ovat sovelluksen käyttöä selittävien ominaisuuksiensa suhteen mahdollisimman samanlaisia kuin todelliset käyttäjät (Dumas & Redish, 1993; Holleran, 1991; Rubin, 1994). Tulosten yleistettävyyden ja ulkoinen validiteetti riippuvat oleellisesti siitä, kuinka hyvin osallistajat edustavat niitä käyttäjäryhmiä, jolle tuote on suunnattu (Holleran, 1991). Jos osallistajat eivät ole edustavia, suuretkaan otoskoot eivät tuo haluttua tietoa. Käytettävyydestissä ulkoiseen validiteettiin liittyvät

osallistujien edustavuuden lisäksi keskeisesti myös testitehtävien todenmukaisuus ja testiympäristön luonnollisuus (Koskinen, luku 13).

Otoskoko vaikuttaa tulosten reliabiliteettiin eli luotettavuuteen sekä yleistettävyyteen otoksesta koko populaatioon siten, että luottamus tulosten oikeellisuuteen paranee otoskoon kasvaessa (Holleran, 1991). Yhden tai kahden osallistujan käyttäytymisen perusteella tehdyt havainnot eivät ole luotettavasti yleistettävissä koko populaatioon. Toisaalta, jos löydetään vakava ongelma, jonka yksi tai kaksi osallistujaa kohtaa, todennäköisesti myös joku tuotteen tulevista käyttäjistä tulee kohtaamaan saman ongelman. Tämän vuoksi on hyvä erottaa, että käytettävyyssongelmien löytäminen ja tulosten yleistettävyys ovat osin erilliset tavoitteet.

Käytettävyyden arvioinnilla voi siis olla erilaisia tavoitteita. Käytännön käytettävyytyö on yleensä *formatiivista käytettävyyden arviointia*, joka on luonteeltaan kvalitatiivista ja diagno-soivaa. Tavoitteena on löytää sovelluksesta käytettävyyssongelmia ja siten parantaa tuotteen käytettävyyttä mahdollisimman paljon taloudellisten ja ajallisten resurssien puitteissa (Nielsen, 1993). Formatiivisen arvioinnin tuloksia ei yleensä haluta yleistää vaan tavoitteena on nimenomaan käytettävyyden parantaminen. Tällöin otoskoon osalta ollaan kiinnostuneita lähinnä siitä, voidaanko tietyn osallistujamäärän perusteella luottaa siihen, ettei vakavia käytettävyyssongelmia jäänyt huomaamatta. Kaikkien käytettävyyssongelmien löytäminen ei silloinkaan ole realistinen tavoite. *Summatiivisessa käytettävyyden arvioinnissa* tutkitaan käyttöliittymän käytettävyyttä ja laatua kokonaisuutena. Siinä verrataan usein erilaisten suunnitteluvaihtoehtojen tai samankaltaisten sovelluksien keskinäistä paremmuutta, koska absoluuttista käytettävyyttä ei voida mitata (Nielsen, 1993). Jotta eri sovellusten käytettävyyttä voitaisiin vertailla, tehdään kvantitatiivisia mittauksia esimerkiksi rekisteröimällä suoritusajoja, virheiden esiintymistiheyttä, hiiren napsautuksia tai vaikka käyttäjän katsepolkuja. Tällöin ollaan usein kiinnostuneita tulosten yleistettävyydestä koko populaatioon, mikä asettaa otoskoolle erilaiset vaatimukset kuin formatiivisessa arvioinnissa. Summatiivinen käytettävyyden arviointi on yleisempää tieteellisessä käytettävyytutkimuksessa kuin käytännön tuotekehityksessä. Tieteellisen tutkimuksen tavoitteet liittyvät usein tutkimushypo-teesien validoimiseen, jolloin tarvitaan summatiivista käytettävyyden arvioimista. Osallistujien edustavuus on tärkeää sekä formatiivisessa että summatiivisessa käytettävyyden arvioinnissa.

Tässä luvussa kerrotaan ensin, miten osallistujien valita tehdään. Sen jälkeen käsitellään otoksen kokoa ja lopuksi perehdytään kevennetyn käytettävyyden arvioinnin yhteydessä esitettyyn pienen otoskoon sääntöön.

18.2. Keitä osallistujiksi valitaan?

Käytettävyytestauksen peruseräite on, että testiin osallistuvien käyttäjien täytyy olla samanlaisia kuin tuotteen todelliset käyttäjät (Dumas & Redish, 1993; Holleran, 1991; Nielsen, 1993; Rubin, 1994). Edustavat osallistajat ovat mieluiten sovelluksen todellisia käyttäjiä. Aina ei kuitenkaan ole mahdollista testata todellisia käyttäjiä, ja silloin osallistajat valitaan siten, että he ovat sovelluksen käyttöä selittävien ominaisuuksien suhteen mahdollisimman samanlaisia kuin todelliset käyttäjät. Osallistujien edustavuuden takaamiseksi osallistajat pitää valita heidän taustansa ja osaamisensa sekä muiden sellaisten ominaisuuksien perusteella, jotka selittävät ja mahdollisesti ennustavat sovelluksen todellista käyttöä. Jotta voidaan tietää, mitkä ominaisuudet selittävät parhaiten sovelluksen erilaisia käyttötapoja, käyttäjiä analysoidaan ja heidän joukostaan tunnistetaan keskeiset käyttäjäryhmät.

Seuraavissa alakohdissa kerrotaan, miten käyttäjäryhmiä muodostetaan ja kuinka osallistajat valitaan muodostetuista käyttäjäryhmistä. Tarkastelun ulkopuolelle on rajattu esimerkiksi *osallistujien rekrytointi* eli kuinka ja mistä edustavia osallistujia voi hankkia. Lisätietoa osallistujien rekrytoinnista saa esimerkiksi Rubinin (1994) kirjan luvusta ”Selecting and acquiring participants” tai Dumasin ja Redishin (1993) luvusta ”Recruiting participants”.

18.2.1. Käyttäjryhmien muodostaminen

Osallistujien valinnassa ensimmäinen vaihe on käyttäjäprofiilien ja käyttäjryhmien muodostaminen. Käyttäjäprofiilit ja tuotteen keskeiset käyttötavat pitäisi olla määriteltyinä jo aikaisemmissa, lähinnä määrittelyvaiheen, dokumentaatioissa, mutta viimeistään ne on määriteltävä nyt (Dumas & Redish, 1993). Dokumenttien lisäksi tietoa käyttäjäprofiilien muodostamisen pohjaksi voi löytää tuotteen edellisen version käyttäjiltä, tehtäväänalyyseistä, markkinointitutkimuksista, benchmarking-tutkimuksista tai tuotekehityksestä ja markkinoinnista vastaavilta tuotepäälliköiltä (Rosenbaum, 1987; Rubin, 1994). Rosenbaum (1987) muistuttaa, että markkinointiosasto on usein enemmän tekemisissä asiakasprofiilien kanssa kuin suunnittelijat, ja siksi heidän tietämystään kannattaa hyödyntää esimerkiksi haastatteleamalla heitä. Markkinointiosaston kanssa keskusteltaessa on muistettava selvittää ero asiakkaan ja todellisen käyttäjän välillä. Kun asiakkaana on yksityisen asiakkaan sijaan kokonainen organisaatio, ostopäätöksen tekevä henkilö ei useinkaan ole sovelluksen todellinen käyttäjä eikä välttämättä edes tietoinen loppukäyttäjien tarpeista (Rubin, 1994; Rosenbaum, 1987).

Rosenbaum huomauttaa myös, että on syytä selvittää tarkat määritelmät keskeisille termeille (esimerkiksi mitä noviisi tarkoittaa), koska yrityksen muulle henkilökunnalle riittää varsin yleisen tason kuvaus käyttäjistä. Tärkeitä kysymyksiä ovat muun muassa ”mitä”, ”miten”, ”miksi”, ”milloin” ja ”missä”, esimerkiksi ”kuinka paljon tyypillisillä käyttäjillä on tietoteknistä kokemusta” (Rosenbaum, 1987).

Kun käyttäjäprofiilit on muodostettu, käyttäjäpopulaatio jaetaan niiden perusteella erillisiin käyttäjryhmiin. *Käyttäjryhmä* muodostuu sellaisista käyttäjistä, jotka ovat samanlaisia sovelluksen käytön kannalta merkityksellisten ominaisuuksiensa suhteen (Dumas & Redish, 1993). Käyttäjryhmien muodostamiseksi onkin erotettava, mitkä ominaisuudet selittävät parhaiten käyttäjien välisiä eroja käyttäytymisessä (Holleran, 1991; Rosenbaum, 1987). Tärkeimpiä käyttäjiä kuvaavia ominaisuuksia ovat ne, jotka ovat yhteisiä koko käyttäjäpopulaatiolle, sekä ne, jotka erottelevat käyttäjäpopulaatiota käyttäjryhmiin (Dumas & Redish, 1993). Käyttäjien demografiset ominaisuudet kuten sukupuoli, koulutus ja ikä eivät välttämättä ole käyttäytymisen selittäjinä yhtä oleellisia kuin esimerkiksi erot käyttökokemuksessa (Dumas & Redish, 1993). Sen sijaan testattaessa esimerkiksi lapsille (Höysniemi, luku 17) tai vanhuksille suunnattuja sovelluksia demografiset ominaisuudet, kuten ikä, ovat hyvinkin keskeisiä.

On tärkeää ymmärtää, että keskeiset osallistujia kuvaavat ominaisuudet ovat erilaisia eri sovelluksille. Esimerkiksi liikuntarajoitteisia käyttäjiä valittaessa voi olla merkityksellistä kysyä, mitä apuvälineteknologioita he käyttävät (Henry et al., 2001). Yleisiä käyttäjien keskeisiä ominaisuuksia ovat Rosenbaumin (1987) ja Rubinin (1994) mukaan

- yleinen tietotekninen kokemus (käyttöaika, käyttöiheets, käyttöjärjestelmä jne.),
- sovelluksen ja vastaavien sovelluksien tuntemus (käyttöaika, käyttöiheets, onko käytänyt saman yrityksen tai brandin tuotteita),
- sovelluksella suoritettavat tehtävät ja tehtävien käyttöiheets,
- sovelluksen aihepiirin tuntemus (domain knowledge),
- asennoituminen teknologiaan, tietokoneisiin, testattavan sovelluksen tyyppiin sovelluksiin ja sovelluksen aihepiiriin,
- työkokemus (tehtävänimikkeet, vastualueet, palvelusaika, kouluttautuminen),
- oppimistyyli (lukijat vs. kuuntelijat ja ”luen ensin, teen sitten”, ”teen ensin, luen sitten” tai ”opin tekemällä”),
- sovellukseen käytön kannalta oleelliset taidot ja ominaisuudet (konekirjoitustaito, luku-taito, kätsisyys jne.), ja
- demografiset ominaisuudet (ikä, sukupuoli, koulutustaso, koulutusala jne.).

Tietotekninen kokemus on keskeinen sovelluksen käyttöä selittävä tekijä, koska noviisien ja eksperttien tarpeet ovat erilaiset. Kaikki käyttäjät ovat ensin noviiseja ja suurin osa heistä oppii taitavammiksi käyttökokemusten karttuessa, mutta kuitenkin vain pieni osa edistyy ekspertiksi asti. Näin ollen suurin osa käyttäjistä on taidoiltaan keskitasoisia käyttäjiä (Barnum, 2002b). Hackos ja Redish (1998) ovat jakaneet tietoteknisen kokemuksen neljään

ryhmään: noviisit (novices), edistyneet aloittelijat (advanced beginners), pätevät käyttäjät (competent performers) ja ekspertit (expert performers). Noviisikäyttäjät ovat keskeinen käyttäjäryhmä erityisesti www-sovelluksissa, koska näiden kohdekäyttäjryhmä on hyvin laaja eikä kaikilla potentiaalisilla käyttäjillä ole paljon tietoteknistä kokemusta (Nielsen, 1997; Nielsen, 2000a).

Dumas ja Redish (1993) selventävät käyttäjäryhmien muodostamista esimerkillä lakitoimiston laskutusjärjestelmästä, jonka käyttäjäryhmät voidaan jakaa ensin ammattinimikkeen perusteella asianajajiin ja lainopillisiin sihteereihin (taulukko 1). Nämä kaksi ammattiryhmää muodostavat järjestelmän tulevien käyttäjien joukon, ja ammattiryhmällä voidaan olettaa olevan merkitystä siinä, millaisia tehtäviä sovelluksella suoritetaan. Tämän lisäksi tunnistetaan kaksi muuta käyttäjiä oleellisesti kuvaavaa ominaisuutta: yleinen tietotekninen kokemus ja laskutuskokemus, joiden oletetaan vaikuttavan sovelluksen käyttöön ja uuden järjestelmän hyväksyntään. Näin ollen asianajajien ja sihteerien ryhmien sisällä syntyy alaryhmiä tietoteknisen kokemuksen ja sovelluksen aihepiirin tuntemuksen eli tässä tapauksessa laskutuskokemuksen perusteella. Muodostamalla käyttäjäryhmiä keskeisten käyttäjiä erottelevien ominaisuuksien perusteella saadaan helposti suuri määrä käyttäjäryhmiä.

Taulukko 1: Lakitoimiston laskutusjärjestelmän keskeiset käyttäjäryhmät (suomennettu Dumas & Redish, 1993).

Asianajajat, joilla on	Sihteerit, joilla on
<ul style="list-style-type: none"> • paljon tietoteknistä kokemusta • paljon laskutuskokemusta 	<ul style="list-style-type: none"> • paljon tietoteknistä kokemusta • paljon laskutuskokemusta
<ul style="list-style-type: none"> • paljon tietoteknistä kokemusta • vähän laskutuskokemusta 	<ul style="list-style-type: none"> • paljon tietoteknistä kokemusta • vähän laskutuskokemusta
<ul style="list-style-type: none"> • vähän tietoteknistä kokemusta • paljon laskutuskokemusta 	<ul style="list-style-type: none"> • vähän tietoteknistä kokemusta • paljon laskutuskokemusta
<ul style="list-style-type: none"> • vähän tietoteknistä kokemusta • vähän laskutuskokemusta 	<ul style="list-style-type: none"> • vähän tietoteknistä kokemusta • vähän laskutuskokemusta

Dillon ja Watson (1996) kritisoivat käytettävyytutkimuksessa käytettäviä osallistujien yksilöllisiä eroja kuvaavia ominaisuuksia. Käyttäjäryhmiä erotellaan lähinnä sovelluksella suoritettavien tehtävien ja kokemukseen pohjautuvien erotteluiden sekä demografisten ominaisuuksien avulla. Dillonin ja Watsonin (1996) mukaan nämä ominaisuudet ovat kuitenkin riittämättömiä selittämään käyttäjien toimintaa. Yksi merkki näiden ominaisuuksien selityskyvyn riittämättömyydestä on, että käyttäjien reagoimista teknologiaan ei ole kyetty ennustamaan näiden ominaisuuksien perusteella. Heidän ehdotuksensa on, että käytettävyytutkimuksessa otettaisiin oppia psykologian piirissä tehdystä yksilöllisten erojen tutkimuksesta, jolla on jo pitkät perinteet.

18.2.2. Keskeisten ominaisuuksien operationalisointi

Osallistujien valintaa varten käyttäjäryhmien keskeiset ominaisuudet täytyy *operationalisoida* eli muuntaa mitattavissa oleviksi kvantitatiivisiksi kriteereiksi, joiden avulla osallistujia voidaan valita (Rosenbaum, 1987; Rubin, 1994). Keskeiset ominaisuudet määritellään tarkemmin, esimerkiksi (Dumas & Redish, 1993)

- aloittelija tietokoneen käytössä = 0–3 kuukautta käyttökokemusta,
- keskitasoinen tietokoneen käytössä = 3 kuukautta – vuosi käyttökokemusta, ja
- kokenut tietokoneen käytössä = yli vuosi käyttökokemusta.

Tällaisten kriteerien määrittäminen on tärkeää, jotta saadaan objektiivisempia ja luotettavampia vastauksia verrattuna siihen, että käyttäjiltä kysyttäisiin, oletko tietotekniseltä osaamistasoltasi tai tietyn sovelluksen käytössä noviisi, keskitasoinen vai ekspertti. Ilmaukset kuten noviisi ja ekspertti ovat hyvin subjektiivisia ja suhteellisia, ja siksi käyttäjien itsearviot ilman tällaisia absoluuttisia kriteerejä eivät ole luotettavia (Rubin, 1994).

Käyttökokemuksen pituuden sijasta kokemustasoa voi määritellä myös esimerkiksi käyttötiheyden eli käytön toistuvuuden perusteella. Esimerkiksi tietyn sovelluksen tai tietyn toiminnon eksperttitason osaamisen kriteerinä voi olla yli vuosi käyttökokemusta ja / tai tietyn sovelluksen päivittäinen käyttötiheys. On muistettava myös erottaa tietotekninen kokemus ja aihepiirin kokemus toisistaan (esimerkiksi taulukossa 1 laskutuskokemus on aihepiirin kokemusta). Dumas ja Redishin (1993) mukaan kvantitatiivisten kriteereiden tulisi olla mielekkäitä testin tavoitteiden kannalta ja niiden tulisi perustua esimerkiksi haastatteluilla ja havainnoinnilla saatuun tietoon. Esimerkiksi haastatteluiden perusteella voidaan saada selville, että alle kolme kuukautta tietokonetta käyttäneet ihmiset yleensä vielä opettelevat tietokoneen käyttöä, mutta sen jälkeen alkavat vähitellen hallita sitä paremmin. Tällöin kolme kuukautta voidaan asettaa rajaksi noviisin ja keskitasoisen käyttäjän välillä. Käytännössä kriteerien empiirinen selvittäminen on resurssien puitteissa usein mahdotonta ja sen sijaan on turvaututtava käytettävyyssiantuntijan arvioon.

Dumas ja Redish (1993) neuvovat kiinnittämään huomiota osallistujien jakaumaan alaryhmän sisällä. Tulokset voivat olla erilaiset, jos edellisen määrittelyn mukaisessa kokeneiden ryhmässä kaikkien osallistujien tietokoneen käyttöaika on vuodesta puoleentoista vuoteen verrattuna siihen, että se vaihtelee kymmenen ja viidentoista vuoden välillä. Lisäksi esimerkiksi käyttötiheyden perusteella tehokäyttäjäksi määrittyvä henkilö ei välttämättä edusta alaryhmän tyypillistä käyttäjää, jolloin tällainen käyttäjä voi olla perusteltua tiputtaa pois. Jotta tällaiset ääritapaukset saataisiin karsittua pois, käyttäjäryhmästä valittaville osallistujille voi asettaa minimi- ja maksimiehdon. Esimerkiksi valittaessa osallistujia käyttötiheyden perusteella saatetaan haluta rajata pois käyttäjät, jotka käyttävät testattavaa sovellusta tai testattavan kaltaisia sovelluksia useita tunteja päivittäin. Tällaiset mahdollisesti hyvin eksperttitasoiset käyttäjät eivät välttämättä ole tutkimuksen kannalta mielenkiintoisia (jossakin tutkimuksessa taas juuri he saattavat olla hyvin mielenkiintoisia riippuen tutkimuksen tavoitteista ja keskeisistä käyttäjäryhmistä).

Kun käyttäjäryhmille on määritelty kvantitatiiviset kriteerit, niistä voidaan laatia kyselylomake, jolla osallistujia voidaan *seuloa* (*screen*). Näin perusteellinen seulominen ei yleensä ole mahdollista käytettävyydestin yhteydessä, koska se vaatii paljon resursseja ja osallistujia on vaikea hankkia. Kuitenkin esimerkiksi valittaessa osallistujia fokusryhmiin seulontaan saatetaan kiinnittää enemmän huomiota. Lisäksi osallistujien edustavuuden varmistaminen on erityisen tärkeää silloin, kun otoskoko on hyvin pieni (Nielsen, 1993).

18.2.3. Osallistujien valinta käyttäjäryhmistä

Kun sovelluksen käyttäjäpopulaatiosta on tunnistettu käyttäjäryhmiä, tätä jakoa käytetään pohjana osallistujien valitsemisessa. Käytettävyyssiantuntijoiden kesken näyttää olevan hieman kaksijakoisuutta siinä, miten osallistujat tulisi valita näistä ryhmistä. Osa on sitä mieltä, että kaikista ryhmistä valitaan edustajia, jotta saadaan kattava otos. Toiset taas neuvovat, että keskitytään muutamaan keskeisimpään ryhmään, joista valitaan useampi osallistuja. Esimerkiksi Rubinin (1994, s. 126) mukaan on tärkeää saada edustava otos koko käyttäjäpopulaatiosta, jolloin kaikki keskeisistä ominaisuuksista muodostetun matriisin (esimerkiksi taulukko 1) solut tulisi olla edustettuna käytettävyydestissä. Sen sijaan Dumas ja Redishin (1993) mukaan käytettävyydestissä on ajan ja kustannusten rajoissa yleensä mahdollista testata kahdesta neljään alaryhmää. Testiin on silloin valittava käyttäjäpopulaation tyypillisimmät ryhmät tai ne ryhmät, jotka ovat testin tavoitteiden kannalta oleellisiä. Jos testin tavoitteena on esimerkiksi parantaa sovelluksen opittavuutta ja ensikäytön helppoutta, testiin on silloin järkevää valita noviisikäyttäjiä. Dumas ja Redish (1993) suosittelevat, että jokaisesta alaryhmästä otetaan testiin vähintään 3–5 käyttäjää, jotta ryhmän sisäinen vaihtelu ei vaikuttaisi tuloksiin. Osallistujien välistä yksilöllistä vaihtelua on kuitenkin mahdotonta hallita tämänkokoisilla otoksilla ja ilman tilastollisia menetelmiä. Tämän vuoksi testin jälkeen esimerkiksi haastattelulla tai kyselylomakkeella kerätty tieto on tärkeää, jotta yksilöllisten erojen vaikutuksia tuloksiin voisi ymmärtää ja analysoida.

Jos pääkäyttäjäryhminä ovat tietoteknisen kokemuksen perusteella aloittelijat, keskitasoiset ja kokeneet eikä ole mahdollista testata kaikkia ryhmiä, Dumas ja Redish (1993) neuvovat

jättämään keskitasoisten ryhmän pois. Heidän perustelunaan on, että jos noviisit ja ekspertit kohtaavat saman ongelman, niin silloin keskitasoisetkin todennäköisesti kohtaisivat sen. Holleran (1991) toteaa, että käytettävyydestiin valitaan usein noviiseja, jolloin pitkäaikainen käytettävyys (long-term usability) jää tutkimatta. Dumas ja Redish (1993) ehdottavat, että jos käytettävyydestiin valitaan vain noviisikäyttäjiä, eksperttikäyttäjillä mahdollisesti vastaan tulevia ongelmia voidaan selvittää rinnalla tehtävän asiantuntija-arvion avulla.

Rubin (1994) antaa omaan kokemukseensa pohjautuvan neuvon, että osallistujaksi kannattaa aina valita ainakin yksi sellainen tuotteen potentiaalinen käyttäjä, jolla on heikoimmat tietotekniset edellytykset käyttää tuotetta (least competent user), vaikka tämä ei kuuluisikaan kohdekäyttäjien enemmistöön. Vaikka tietotekniset edellytykset olisivatkin heikot, käyttäjän ei tarvitse muilta taidoiltaan olla aloittelija. Hän voi olla aihepiirin tuntemuksen osalta keskitasoinen tai ekspertti; pääasia on, että hänellä ei ole juurikaan tietoteknistä kokemusta. Tällaisten käyttäjien avulla saadaan Rubinin mukaan paljon hyödyllistä tietoa siitä huolimatta, että he eivät edusta tyypillisintä käyttäjää. Etenkin jos tämä käyttäjä selviytyy ohjelman käytöstä, voidaan olettaa, että silloin todelliset tietotekniikkaa osaavammat käyttäjät selviytyvät hyvin. Toisaalta, jos vähemmän kokeneella käyttäjällä on ongelmia, se ei tarkoita, että sovellus on epäonnistunut ja todelliset käyttäjät eivät osaisi käyttää sitä. Sen sijaan vähemmän kokeneen käyttäjän toiminnan kautta voidaan saada tietoa perustavanlaatuisista käytettävyysongelmista.

Eräs otoksiin liittyvä ongelma on vain ”parhaiden” testaaminen, erityisesti silloin kun ei itse voi päättää keitä osallistujiksi valitaan ja millä perusteella valinta tehdään vaan esimerkiksi esimies tai opettaja valitsee osallistujat (Höysniemi, luku 17; Nielsen, 1993; Rubin, 1994). Esimies saattaa ajatella, että testiin osallistuminen on palkinto, joka myönnetään hyvin tehdystä työstä tai hän saattaa haluta antaa hyvän kuvan yrityksestä ja valitsee testiin siksi parhaita työntekijöitään. Tällainen ”paras” työntekijä voi olla tietoteknisesti aloittelevakin käyttäjä, mutta yhteistä ”parhaille” on suorittaminen, yritteliäisyys ja esimerkiksi sovelluksen ongelmien vähätteleminen. Tällöin otos saattaa vääristyä eivätkä tulokset anna todellista kuvaa todellisten käyttäjien toiminnasta. Valinnan tekeväälle esimiehelle tai opettajalle kannattaakin selittää, että testin mielekkyyden kannalta on tärkeää, että osallistujat muodostavat mahdollisimman kattavan otoksen. Otoksen edustavuudelle tuottaa ongelmia myös vapaaehtoiset osallistujat: he saattavat esimerkiksi olla muita motivoituneempia, uteliaampia ja kiinnostuneempia uudesta teknologiasta, eivätkä he siten välttämättä edusta tyypillisimpiä käyttäjiä (Thomas & Kellogg, 1989).

18.2.4. Käytettävyyden arviointiin soveltuvia otantamenetelmiä

Käytettävyyttä arvioitaessa ei yleensä testata koko käyttäjäpopulaatiota vaan otosta tästä populaatiosta. Käytettävyyden arvioimiseen soveltuvien otantamenetelmien peruseräotteet on hyvä tuntea, jotta voi ymmärtää ja arvioida otantamenetelmän vaikutuksia tuloksiin. Otantamenetelmät voidaan jakaa karkeasti satunnaisotantoihin ja ei-satunnaisotantoihin. *Satunnaisotannassa (probability sample; myös random sample, todennäköisyysotanta)* jokaisen osallistujan valituksi tuleminen todennäköisyys tiedetään ja se on kaikille populaation jäsenille sama (Robson, 1994). Jos tätä todennäköisyyttä ei tiedetä, kuten yleensä käytettävyytustutkimuksessa, kyseessä on *ei-satunnaisotanta (non-probability sample; myös non-random sample, ei-todennäköisyysotanta)*. Sen sijaan esimerkiksi laajoissa puhelinhaastattelussa ja kyselyissä voi olla mahdollista käyttää satunnaisotantaa. Seuraavaksi esitellään tarkemmin kiintiöotanta, dimensionaalinen otanta ja mukavuusotanta, jotka ovat käytettävyyden testaamiseen soveltuvia ei-satunnaisia otantamenetelmiä.

Kiintiöotannassa (quota sampling) osallistujat valitaan siten, että kaikki populaation alaryhmät ovat edustettuina otoksessa samassa suhteessa kuin ne esiintyvät populaatiossa (Robson, 1994). Tätä varten jokaiselle alaryhmälle on määriteltävä kiintiö. Kiintiön sisällä osallistujat valitaan usein mukavuusotannan mukaisesti. Käytettävyydestä tarkastuksessa tämä tapahtuu käyttäjäryhmien ominaisuuksien prosenttijakaumien avulla. Jos koko käyttäjäpopulaatiosta 20 % on noviisikäyttäjiä, 60 % keskitasoisia käyttäjiä ja 20 % eksperttejä, niin samojen suhteiden tulisi päteä myös otoksessa. Tällaiset todellisten käyttäjien jakaumat eivät

kuitenkaan yleensä ole tiedossa eikä niiden selvittäminen ole resurssien puitteissa useinkaan järkevää, joten on käytettävä arvioita.

Dimensionaalissa otannassa (dimensional sampling) käytettävyydestin tavoitteiden kannalta merkityksellisistä ominaisuuksista muodostetaan matriisi, ja jokaisesta solusta valitaan vähintään yksi osallistuja (Robson, 1994). Edellä kuvattu matriisi lakitoimiston laskutusjärjestelmän käyttäjäryhmistä (taulukko 1) on esimerkki tällaisesta matriisista dimensionaalisen otannan pohjaksi. Kiintiötanta ja dimensionaalinen otanta pyrkivät huomioimaan sitä, että otos edustaisi populaatiota tilastollisesti lähentyen siten satunnaisotantaa. Kumpikin otantamenetelmä sisältää silti riskejä otannan vääristymiseen. Vääristyneen otannan riskiä voi kuitenkin minimoida osallistujien valinnan huolellisella suunnittelulla (Robson, 1994).

Mukavuusotanta (convenience sampling), suomennettu joskus myös tarkoituksenmukaiseksi otannaksi) on vähän resursseja vaativa otantamenetelmä, jossa osallistujiksi valitaan helpoiten saatavilla olevat osallistajat (Robson, 1994). Tällaisen otannan perusteella ei voida tehdä tilastollisia yleistyksiä otoksesta koko populaatioon eikä se tuota edustavia tuloksia. Mukavuusotanta soveltuu kuitenkin erittäin hyvin pilottitestaukseen ja esitutkimukseen. Käytettävyydestissä on osallistujina usein esimerkiksi opiskelijoita, koska heitä on helppo saada osallistumaan testiin. He eivät välttämättä edusta mitään keskeistä käyttäjäryhmää, ja silloin heidän valitsemisensa osallistujiksi ei ole perusteltua (Rosenbaum, 1987). Sama koskee yhtiön työntekijöiden käyttämistä osallistujina, jos sovellus on tarkoitettu yhtiön ulkopuoliseen käyttöön. Vaikka työntekijät olisivat edustavia kaikkien muiden ominaisuuksiensa suhteen, he eroavat todellisesta käyttäjäpopulaatiosta siinä, että he työskentelevät tuotetta valmistavassa yhtiössä (Dumas & Redish, 1993; Rubin, 1994). Tällöin esimerkiksi yrityksen kulttuuri ja kieli sekä mahdollisesti muut yrityksen valmistamat samankaltaiset tuotteet ovat heille tuttuja, toisin kuin ulkopuoliselle käyttäjälle.

Käytettävyyden testaamiseen soveltuvia otantamenetelmiä ovat lisäksi (Robson, 1994)

- *ääritapausotanta (extreme case samples)*, jossa ääritapaus sisällytetään otokseen, koska sen arvellaan tuovan erityisen paljon tietoa tutkittavasta ilmiöstä;
- *homogeeninen otanta*, joka kattaa tietyn muuttujan, esimerkiksi tietoteknisen kokemuksen, vain hyvin suppealta alueelta;
- *heterogeeninen otanta*, jossa varmistetaan, että osallistajat edustavat keskeisiä piirteitä laajasti, ja
- *harkintaotanta (purposive sampling)*, joka perustuu tutkijan arvioon osallistujien tyypillisyydestä tai kiinnostavuudesta. Esimerkiksi grounded theory -lähestymistavassa ensimmäisten osallistujien perusteella muodostuva teoria ohjaa seuraavien osallistujien valintaa. Tällaisen otannan perusteella ei voida tehdä yleistyksiä otoksesta populaatioon. Harkintaotantaa käytetään tyypillisesti tapaustutkimuksissa.

Ääritapausotantaa voi soveltaa käytettävyydestissä esimerkiksi sisällyttämällä otokseen vähintään yksi sellainen mahdollinen tuotteen käyttäjä, jolla on heikoimmat käyttöedellytykset tuotteen käyttöön (katso kohta 18.2.3). Tällä tavoin voidaan saada sellaista hyödyllistä tietoa, jota ei saataisi vain kokeneempia käyttäjiä testaamalla. Otos on homogeeninen esimerkiksi tietoteknisen kokemuksen suhteen silloin, jos keskitytään tutkimaan vaikkapa ainoastaan vasta aloittelevia käyttäjiä. Tällöin valintakriteerejä määritettäessä voidaan asettaa minimi- ja maksimiehto osallistujien jollekin ominaisuudelle. Voidaan esimerkiksi valita vain osallistujia, joiden testattavan sovelluksen käyttökokemus on 0–½ vuotta, jolloin otos kattaa tietoteknisen kokeneisuuden vain hyvin suppealta alueelta. Esimerkki heterogeenisesta otoksesta on tutkimus, johon tietoisesti valitaan osallistujia kattavasti esimerkiksi tietoteknisen kokemuksen koko vaihtelualalta aloittelijasta kokeneimpiin.

Joskus voi olla mahdollista sisällyttää testiin koko käyttäjäpopulaatio, esimerkiksi testattaessa yrityksen sisäiseen käyttöön tulevaa sovellusta, jota tulee käyttämään kymmenen yrityksen työntekijää. Tällöin testin ulkopuolelle jäävät ainoastaan sovelluksen mahdolliset *tulevat* käyttäjät, kuten uudet työntekijät tai eri osastojen työntekijät (Dumas & Redish, 1993). Tämän huomioon ottamiseksi saattaa olla hyödyllistä valita osallistujaksi esimerkiksi yksi toisen osaston työntekijä tai täysin ulkopuolinen henkilö, jolle sovelluksen aihepiiri on vieraampi, edustamaan testissä mahdollista uutta työntekijää.

18.2.5. Osallistujien valintaan liittyviä ongelmia

Edustavan otoksen saaminen ei käytännössä ole aina mahdollista (Holleran, 1991). Käyttäjäryhmiä voi olla vaikea tunnistaa, etenkin jos sovellus on suunnattu hyvin laajalle käyttäjäjoukolle. Lisäksi edustavia osallistujia voi olla hyvin vaikea, ellei jopa mahdoton hankkia, jolloin on tyydyttävä kompromissiin otoksen edustavuuden suhteen. Tällöinkin tavoitteena on olosuhteiden ja resurssien puitteissa maksimoida otoksen edustavuutta populaatioon nähden. Esimerkiksi sisäisten työntekijöiden käyttö voi olla liikesalaisuuden takia välttämätöntä, jos edes salassapitosopimus ulkopuolisen osallistujan kanssa ei tule kysymykseen. Yrityksen sisäiset työntekijät eroavat siinä keskeisessä sovelluksen käyttöön vaikuttavassa ominaisuudessa, että he eivät ole yrityksen ulkopuolisia käyttäjiä. Tällöinkin yrityksen sisäisistä työntekijöistä tulee valita sellaisia osallistujia, joiden muut ominaisuudet täsmäävät kohde-ryhmän käyttäjien ominaisuuksien kanssa mahdollisimman hyvin.

Koska osallistujia on vaikea saada, ja huolellinen osallistujien valinta vaatii paljon resursseja, käytännössä osallistujien jako käyttäjäryhmiin tapahtuu usein myös jälkikäteen. Jako ryhmiin voidaan tehdä joko ennen testiä tai testin jälkeen kyselylomakkeella tai haastattelemalla saatujen tietojen perusteella. Esimerkki tällaisesta kyselylomakkeesta löytyy liitteestä 1. Nämä tiedot ovat tärkeitä myös siksi että tuloksia voidaan tulkita yksilöllisiä eroja selittävien ominaisuuksien perusteella, ja jotta keskiarvosta selkeästi poikkeavia niin sanottuja vieraita havaintoja (outlier) voidaan selittää.

Tuloksia analysoitaessa on aina syytä pohtia, miten osallistujien ja todellisten käyttäjien välinen eroavaisuus saattaisi vaikuttaa tuloksiin ja millaisissa tilanteissa tutkimuksen osallistujilla saadut tulokset eivät ehkä päde todellisten käyttäjien todellisissa käyttötilanteissa. Thomasin ja Kelloggin (1989) mukaan laboratoriotestin tuloksia on yleistettävä hyvin monella eri tavalla, jotta voidaan arvioida sovelluksen todellista käytettävyyttä. Tulokset on yleistettävä otoksesta koko populaatioon, käytetyistä testitehtävistä kaikkiin sovelluksella suoritettaviin tehtäviin sekä käytettävyydestilanteesta käyttäjän todelliseen työkontekstiin. Osallistujiin liittyvät ongelmat tulosten ulkoisessa validiteetissa liittyvät ensinnäkin osallistujien välisiin yksilöllisiin eroihin ja toiseksi laboratorio-olosuhteiden ja todellisen käyttötilanteen väliseen eroon käyttäjän motivaatiossa. Osallistujien valinnalla voidaan vaikuttaa edellä mainittuun ongelmaan. Thomasin ja Kelloggin (1989) artikkeli on suositeltavaa luettavaa, jos haluaa lisätietoa ulkoiseen validiteettiin liittyvistä kysymyksistä käytettävyydestissä.

18.3. Osallistujien lukumäärä

Tähän mennessä on tarkasteltu, millaisia osallistujia käytettävyytutkimukseen tulisi valita. Osallistujien valinta ja otoskoko ovat molemmat tärkeitä osatekijöitä luotettavien tulosten saamiseksi, sillä suurelta osin osallistujamäärät eivät auta, jos osallistujat eivät edusta tyypillisiä käyttäjiä. Toisaalta, vaikka osallistujat olisivatkin tyypillisiä, heidän välisensä yksilölliset erot saattavat vääristää tuloksia, jos otoskoko ei ole tarpeeksi suuri. Osallistujien lukumäärälle ei voi antaa muuta yksinkertaista ohjetta kuin että se riippuu tilanteesta. Sopivan otoskoon arvioimiseen vaikuttaa ennen kaikkea käytettävä menetelmä, onko testaus tyypiltään formaatiivista vai summatiivista, ja ollaanko kiinnostuneita tilastollisesta merkitsevyydestä tai tulosten yleistettävyydestä koko populaatioon (Dumas & Redish, 1993; Nielsen, 1994).

Paineet otoskoon minimoimiseen tulevat siitä, että testaus on kallista ja aikaa vievää, ja siitä, että edustavia osallistujia on usein vaikea saada (Holleran, 1991). Otoskoko onkin kompromissi rajallisten resurssien, kustannustehokkuuden ja testauksen perusteellisuuden välillä. Testauksen perusteellisuus ei riipu ainoastaan otoskoon riittävyyydestä vaan se on yhteydessä myös testin kattavuuteen. Vain muutaman osallistujan testillä voidaan saada syvällisempää tietoa kuin jakamalla sama aika useamman osallistujan kesken.

Seuraavissa alakohdissa tarkastellaan ensin keinoja sopivan otoskoon arvioimiseksi osallistavissa käytettävyyden arviointimenetelmissä ja seuraavaksi tarkemmin formaatiivisissa ja summatiivisissa käytettävyydestissä. Lopuksi esitellään kevennettyyn käytettävyyden

arviointiin keskeisesti liittyvä ”pienen otoskoon periaate”, jonka mukaan viisi käyttäjää on iteratiiviseen suunnitteluun liittyvässä käytettävyydestestauksessa riittävä osallistujamäärä, ja sen osakseen saamaa kritiikkiä.

18.3.1. Osallistujien määrä osallistavissa arviointimenetelmissä

Nielsen (1993) on antanut viitteellisiä arvioita eri menetelmien vaatimille osallistujamäärille (taulukko 2). Taulukon perusteella eri menetelmiä voidaan vertailla niissä tarvittavan osallistujamäärän perusteella. Nielsenin mukaan menetelmiä, joissa selvitetään pienimmillä osallistujamäärillä, ovat ääneenajattelu, havainnointi ja haastattelu. Jos osallistujia on käytettävissä yli kymmenen, voidaan harkita fokusryhmiä ja summatiivista käytettävyydestä, jossa suoriutumista mitataan esimerkiksi rekisteröimällä suoritusnopeuksia ja virheiden esiintymistiheyksiä. Menetelmiä, jotka vaativat hyvin paljon osallistujia ovat kyselytutkimus, todellisen käytön rekisteröinti ja käytettävyyden arviointi käyttäjien palautteen perusteella. Taulukon osallistujamääriä tarkasteltaessa on huomattava, että Nielsen on *kevennetyn käytettävyyden arvioinnin (discount usability)* puolesta puhuja. Kevennetyssä käytettävyyden arvioinnissa pyritään minimoimaan käytettävyyden arvioimisen kustannukset. Vastaavanlainen taulukko menetelmien vaatimista osallistujamääräarvioista on myös tämän raportin johdantoluvussa.

Taulukko 2: Suuntaa-antavia osallistujamääriä eri menetelmillä (muokattu ja suomennettu Nielsen, 1993).

Menetelmä	Osallistujia
Heuristinen arviointi	- (asiantuntijoiden tekemä)
Ääneenajattelu	3–5
Havainnointi	3 tai enemmän
Haastattelu	5
Suoriutumisen mittarit	vähintään 10
Fokusryhmät	6–9 / ryhmä
Käytön rekisteröinti (logging actual use)	vähintään 20
Kyselytutkimus	vähintään 30
Käyttäjien palaute	satoja

Kuten kohdassa 18.2.3. tuli ilmi, osallistujien määrä riippuu kiinnostavien, testiin mukaan otettavien käyttäjäryhmien määrästä, koska jokaisesta käyttäjäryhmästä tulisi testata vähintään 3–5 käyttäjää. Jos siis testataan useaa eri käyttäjäryhmää, otoskoko on suurempi kuin vain yhtä ryhmää testattaessa. Käytettävyydestin lisäksi osallistujia valitaan tyypillisesti eri käyttäjäryhmistä myös esimerkiksi fokusryhmissä. Tällöin osallistujien valinnassa on pohdittava myös ryhmädynamiikkaan liittyviä kysymyksiä, kuten muissakin menetelmissä, joissa on monta yhtäaikaista osallistujaa (Parviainen, luku 4).

18.3.2. Osallistujien määrän arvioiminen formatiivisessa käytettävyydestestissä

Formatiivisen käytettävyydestin tarkoituksena on löytää tuotteesta käytettävyyso ongelmia ja sitä kautta parantaa tuotteen käytettävyyttä. Formatiivisessa käytettävyydestestissä selvitetään suhteellisen pienillä osallistujamäärillä, koska testin tuloksia ei ole tarkoitus yleistää. Kuitenkin on tärkeää, että testin otos edustaa kattavasti niitä käyttäjäryhmiä, joita testissä tutkitaan, jotta erilaisten käyttäjien kohtaamat käytettävyyso ngelmat paljastuisivat.

Osallistujien määrään vaikuttaa keskeisesti, testataanko tuotetta vain yhden kerran vai monta kertaa iteratiivisen prosessin eri vaiheissa (Nielsen, 1994; Rubin, 1994; Virzi, 1992). Käytettävyydestestissä, joka on osa iteratiivista prosessia, voidaan käyttää pienempiä otoskokoja kuin yksittäisessä testissä. Koska seuraavalla iteraatiokierroksella tehdään uusi testi, saavat ensim-

mäisellä kerralla mahdollisesti löytämättä jääneet ongelmat uuden mahdollisuuden tulla löydettyksi. Vaikka yhden osatestin osallistujamäärä onkin pieni, testausprosessin kokonais-osallistujamäärä on usein suurempi kuin tilanteessa, jossa sovelluksen käytettävyyttä arvioidaan vain yhdellä yksittäisellä testillä. Prototyyppi myös elää koko ajan, joten ensimmäistä versiota ei kannata testata liian perusteellisesti (Nielsen, 1994).

Jos sovelluksen käytettävyyttä testataan vain yhden kerran, otoskokoa ei kannata määritellä tarkasti etukäteen (Nielsen, 1994; Rubin, 1994; Virzi, 1992). Testausta jatketaan niin kauan kunnes viimeisimmät osallistujat eivät enää kohtaa juurikaan uusia ongelmia vaan ainoastaan samat ongelmat ilmenevät toistuvasti. Ennen testiä täytyy määritellä, kuinka paljon uusia ongelmia osallistujien tulee kohdata, jotta testausta voidaan pitää vielä kannattavana. Esimerkiksi sovelluksissa, joissa käytettävyysongelmat ovat katastrofaalisia, testauksessa pyritään löytämään kaikki mahdolliset ongelmat. Tällöin testaus saattaa olla vielä perusteltua, vaikka jokainen osallistuja ei välttämättä löytäisikään uusia ongelmia.

Nielsenin (1994) mukaan käytettävyydestin otoskoko vaikuttavat arvioijan taidot ja kokemus, iteraatioiden määrä sekä sovelluksen ja testin pohjalta tehtävän päätöksen taloudellinen tai muu vaikutus. Kokeneet arvioijat huomaavat ongelmia tehokkaammin kuin aloittelijat, mikä heijastuu myös otoskoon ja löytyneiden ongelmien suhteeseen (Nielsen, 1994; Perälä, luku 19). Osallistujien määrään vaikuttaa myös testattavan sovelluksen laajuus. Sopivaa, kustannustehokasta osallistujamäärää *www*-sivuston testaamiseen on tutkittu erityisen paljon. Spoolin ja Schroederin (2001) tutkimuksen mukaan *www*-sivustojen arvioinnissa tarvitaan suurempia osallistujamääriä kuin esimerkiksi yksinkertaisten Windows-sovellusten arvioinnissa. Heidän testaamallaan *www*-sivustolla vielä viidestoistakin osallistujaa kohtasi vakavan ongelman. Spoolin ja Schroederin tutkimusta on kuitenkin kritisoitu heidän käyttämänsä testitehtävän laajuudesta. He pyysivät osallistujia ostamaan haluamansa tuotteen, esimerkiksi CD-levyn, verkkokaupasta. Testin osallistujat saattoivat käydä neljällä eri sivustolla ja siten heidän navigointipolkunsa vaihtelivat hyvinkin radikaalisti. Käytettävyyssiantuntijoiden keskuudessa ei toistaiseksi ole yksimielisyyttä siitä, vaatiiko *www*-sivustojen testaus enemmän osallistujia kuin muut sovellukset (Barnum, 2002a; Barnum et al., 2003).

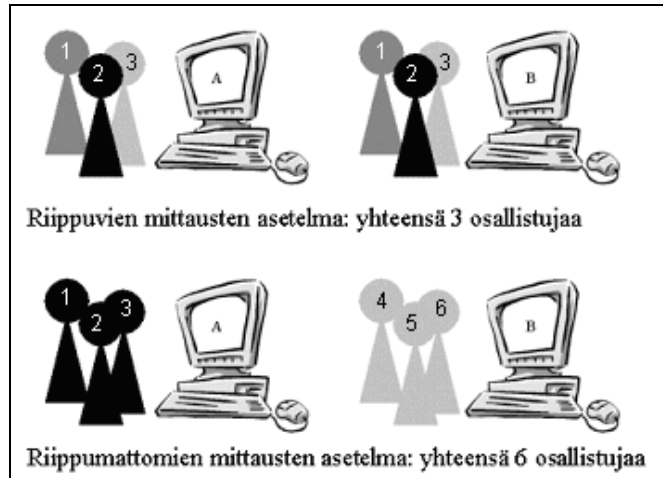
18.3.3. Otoksoon arvioiminen summatiivisessa käytettävyydestissä

Summatiivisen käytettävyydestin tarkoituksena on validoida käytettävyyteen liittyviä oletuksia, esimerkiksi todistaa empiirisesti, että tekstinsyöttö onnistuu uudella vuorovaikutustekniikalla huomattavasti nopeammin kuin perinteisellä tavalla. Koska absoluuttista käytettävyyttä ei voida mitata, summatiivisessa käytettävyydestissä yleensä verrataan kahden tai useamman sovelluksen tai suunnitteluvaihtoehdon käytettävyyttä, esimerkiksi uutta vuorovaikutustekniikkaa saatetaan verrata vastaavaan yleisesti käytettyyn tekniikkaan. Summatiivista testausta tehdään pääasiassa tieteellisessä tutkimuksessa, jolloin ollaan tyyppillisesti kiinnostuneita havaittujen erojen tilastollisesta merkitsevyydestä ja tulosten yleistettävyydestä. Käytännön käytettävyysoyössä ei yleensä ole tarvetta summatiivisille testeille.

Käytettävyydestissä osallistujien väliset yksilölliset erot saattavat olla hyvinkin suuria, mikä asettaa vaatimuksia testin *reliabiliteetille eli luotettavuudelle* (Nielsen, 1993). Reliabiliteetti tarkoittaa sitä, että menetelmällä saadut tulokset eivät ole sattumanvaraisia ja ne ovat toistettavissa. Tämän vuoksi summatiiviseen käytettävyydestiin tarvitaan riittävän suuri otoskoko, jopa useita kymmeniä osallistujia jokaisessa ryhmässä (Dumas, 2003). Käyttäjäpopulaation kirjavuus siis vaikuttaa otoskoko: mitä enemmän populaatiossa on vaihtelua, sitä suurempi otos tarvitaan (Robson, 1994). Nielsen (1993) on tehnyt yleiskatsauksen yksilöiden välisestä vaihtelusta eri tyyppisiä kvantitatiivisia muuttujia mitattaessa 36 julkaistun käytettävyyttutkimuksen perusteella. Tämän katsauksen mukaan virheiden esiintymistiheydessä on suurempaa vaihtelua kuin opittavuudessa ja eksperttien suorituksissa. Virheiden esiintymistiheyttä mittaavissa tutkimuksissa on siis yleisesti ottaen tarvittu suurempi otoskoko kuin opittavuutta tai eksperttien suoriutumista mittaavissa tutkimuksissa. Yksilöiden välisten erojen aiheuttama hajonta kuitenkin vaihtelee huomattavasti eri tutkimuksissa, joten mitään yleisiä johtopäätöksiä Nielsenin tuloksista ei voi tehdä.

Lisäksi otoskokoon vaikuttaa, käytetäänkö vertailevassa käytettävyydestissä *riippuvien mittausten asetelmaa (within-subjects design)* vai *riippumattomien mittausten asetelmaa (between-subjects design)*. Kuva 1 havainnollistaa sovellusten A ja B käytettävyyden vertaamista riippuvien ja riippumattomien mittausten asetelmilla. Riippuvien mittausten asetelmassa kaikki osallistujat käyttävät sekä sovellusta A että B. Riippumattomien mittausten asetelmassa taas puolet osallistujista käyttää vain sovellusta A ja puolet vain sovellusta B.

Riippuvien mittausten asetelmassa vältetään ryhmien erilaisuudesta johtuvat ongelmat. Tällöin validiteettiä uhkaa kuitenkin suoritusjärjestyksestä johtuva oppiminen. Jos osallistuja tekee ensin tehtävät sovelluksella A ja sen jälkeen samat tehtävät sovelluksella B, sovelluksella A saadut kokemukset vaikuttavat sovelluksella B suoritettavaan osioon. Tämän vuoksi riippuvien mittausten asetelmaa käytettäessä suoritusjärjestys on *tasapainotettava (counterbalancing)* eli puolet osallistujista käyttää ensin sovellusta A ja sitten B ja loput käyttää ensin sovellusta B ja sitten A. Riippuvien mittausten asetelmassa tilastollisten



Kuva 1: Riippuvien ja riippumattomien mittausten asetelmat.

erojen esiin saamiseksi tarvitaan vähemmän osallistujia kuin riippumattomien mittausten asetelmassa, koska edellisen asetelman tilastollinen voimakkuus on suurempi kuin jälkimmäisen. Tämä johtuu siitä, että riippuvien mittausten asetelmassa ei tarvitse huomioida ryhmien välistä satunnaisvaihtelua (Dumas, 2003). Riippuvia mittauksia käytettäessä ongelmana saattaa olla, että yksittäisen testin kesto on melko pitkä, koska sama osallistuja käyttää kumpaakin sovellusta.

Riippumattomien mittausten asetelmassa haasteena on, että molempien ryhmien osallistujien tulisi olla samankaltaisia niiden ominaisuuksien suhteen, joiden voidaan olettaa vaikuttavan testissä tutkittaviin asioihin. Jos esimerkiksi toisessa ryhmässä on yksi hyvin kompetentti osallistuja, tämä saattaa vaikuttaa merkittävästi vertailun tuloksiin etenkin jos osallistujamäärä on pieni (Dumas, 2003). Riippumattomien mittausten asetelmassa tarvitaan enemmän osallistujia kuin riippuvien mittausten asetelmassa (kuva 1).

18.3.4. Pienen otoskoon sääntö

Viime aikoina käytettävyydsiantuntijoiden keskuudessa on keskusteltu paljon siitä, onko viisi osallistujaa riittävä määrä käytettävyydestissä. *Pienen otoskoon säännön (small sample size)* mukaan viisi käyttäjää on riittävä määrä. Sääntö perustuu Virzin, Nielsenin ja Lewisin tutkimuksille, joissa on matemaattisen mallin avulla osoitettu, että 80 % käytettävyysongelmista löytyy viidellä testikäyttäjällä (Lewis, 1994; Nielsen, 1994; Nielsen, 2000b; Nielsen & Landauer, 1993; Virzi, 1992). Motivaationa pienelle otoskoolle on käytettävyydestäuksen kustannusten vähentäminen ja kustannustehokkuuden optimoiminen. Ajatuksena on, että vähän testausta on parempi kuin ei testausta ollenkaan, ja jos testeihin kuluu vähemmän resursseja, niin niitä tehdään enemmän (Nielsen, 1994). Aluksi on syytä painottaa, että pienen otoskoon sääntö soveltuu iteratiiviseen formatiiviseen käytettävyydestäukseen (Nielsen, 1994). Matemaattista mallia soveltavien tutkimusten (mm. Virzi, 1992) mukaan

- 80 % käytettävyysongelmista löytyy viidellä osallistujalla,
- seuraavat osallistujat tuovat yhä vähemmän uutta informaatiota, ja
- vakavimmat käytettävyysongelmat löytyvät todennäköisesti ensimmäisten osallistujien aikana.

Virzi (1992), Lewis (1994) ja Nielsen (1994) ovat käyttäneet todennäköisyysteoriaan pohjau-

tuva kaavaa käytettävyydestissä tarvittavan osallistujamäärän arvioimiseen. Virzin ja Nielsenin tutkimuksissa menetelmänä oli käytettävyydesti, jonka aikana ajateltiin ääneen.

$$1 - (1 - p)^n \quad (1)$$

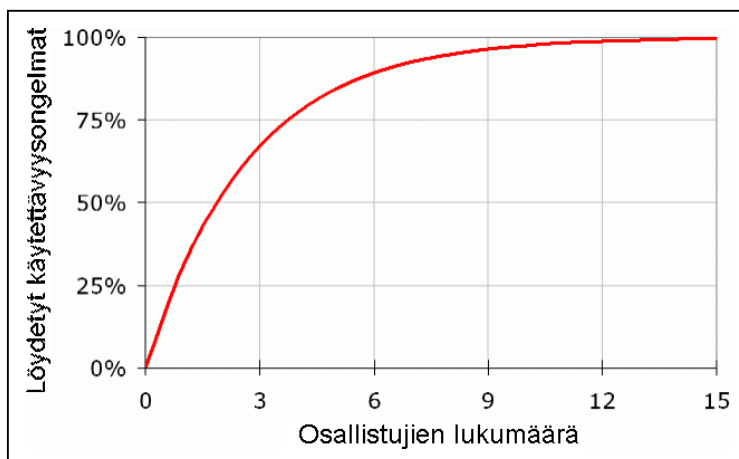
Kaava (1) ilmaisee, kuinka suuri osuus kaikista löydettyistä ongelmista löydetään tietyllä otoskokoilla. Kaavoissa p on käytettävyysongelman löytymisen todennäköisyys ja n on tarvittavien osallistujien määrä eli otoskoko. Kaavassa (1) 100 prosenttia tarkoittaa käytettävyydestissä kaikkien osallistujien yhteensä löytämien ongelmien lukumäärää (Virzi, 1992). 100 prosenttia ei siis tarkoita kaikkia sovelluksessa olevia ongelmia, vaan kaikilla käyttäjillä yhteensä löydettyjen ongelmien kokonaismäärää. Nielsenin (1994) mukaan kaikkien sovelluksessa olevien ongelmien määrää voidaan arvioida löydettyjen ongelmien lukumäärän perusteella kaavalla 2.

$$\text{kaikki ongelmat} = \frac{n : \text{llä osallistujalla löydetty ongelmat}}{1 - (1 - p)^n} \quad (2)$$

80 % käytettävyysongelmista löytyy viidellä osallistujalla. Kaavan (1) käyttö edellyttää, että käytettävyysongelmien löytymisen todennäköisyys (p) tiedetään. Virzi ja Nielsen ovat tutkineet ongelmien löytymistä erityyppisillä graafisilla sovelluksilla ja tulosten perusteella he ovat määrittäneet, että käytettävyysongelmien löytymisen todennäköisyys on vastaavan tyyppisissä sovelluksissa 30–40 prosenttia. Väitteen ”80 % käytettävyysongelmista löytyy viidellä osallistujalla” taustalla on siis oletus, että käytettävyysongelmat löytyvät tietyllä tunnetulla todennäköisyydellä. On aivan oikein päätellä, että viidellä osallistujalla löytyy noin 80 % käytettävyysongelmista, jos ongelmien löytymisen todennäköisyys vaihtelee välillä 0,32 ja 0,42 kuten Virzin (1992) testeissä. Ongelmana on se, että käytettävyysongelmien löytymisen todennäköisyys vaihtelee suurestikin eikä sitä siksi voida etukäteen määrittää. Esimerkiksi jos ongelmien löytymisen todennäköisyys on 10 %, silloin löytääkseen noin 80 % ongelmista täytyykin testata 15 osallistujaa. Vastaava laskutoimitus on siis seuraava

$$1 - (1 - p)^n = 1 - (1 - 0,10)^{15} = 1 - 0,90^{15} = 1 - 0,205 = 0,79.$$

Seuraavat osallistujat tuovat yhä vähemmän uutta informaatiota. Virzin (1992) toinen tutkimustulos, vähenevän tuoton laki (*law of diminishing returns*), on yleinen otoskokojen kasvattamiseen liittyvä ominaisuus: hyötysuhde tulosten tarkkuuden paranemisessa yleensä pienenee kun otoskokoa kasvatetaan (Robson, 1994). Vähenevän tuoton lain mukaisesti ensimmäinen osallistuja tuo eniten informaatiota, seuraava hieman vähemmän, kunnes lopulta viimeisimmät osallistujat eivät tuo enää paljonkaan uutta informaatiota (kuva 2). On selvää, että paras kustannus-hyöty -suhde saavutetaan ensimmäisillä osallistujilla, esimerkiksi Nielsenin ja Landauerin (1993) mukaan 3,2 osallistujalla. Tätä ei kuitenkaan pidä tulkita niin, että noin 3 osallistujaa olisi riittävä osallistujamäärä. Tulos kertoo vain, että kolmannen osallistujan jälkeen kustannus-hyöty -suhde alkaa laskea.



Kuva 2: Osotukseen ja löytyneiden käytettävyysongelmien suhde (Nielsen, 2000b).

Vakavimmat ongelmat löytyvät ensimmäisillä käyttäjillä. Virzin (1992) mukaan ongelmien yleisyys ja ongelmien vakavuus korreloivat keskenään. Jos siis ensimmäiset osallistujat löytävät keskimäärin eniten uusia ongelmia, niin silloin vakavimmat ongelmat löytyvät ensimmäisten osallistujien aikana. Lewisin (1994) sekä Spoolin ja Schroederin (2001) tulosten mukaan ongelmien vakavuus ja yleisyys eivät kuitenkaan korreloi keskenään.

Tarvitaan siis lisää tutkimusta, jotta voitaisiin tehdä varmempia johtopäätöksiä ongelmien vakavuuden ja yleisyyden suhteesta.

Pienen otoskoon sääntö on pitkään ollut osallistujien määrää ohjaavana normina. Se on kuitenkin otettu liian yleiseen käyttöön ja sen soveltaminen on virheellisesti yleistetty koskemaan kaikentyypisiä ja -laajuisia sovelluksia. Sääntöä on alettu kyseenalaistaa, koska se soveltuu vain tietynlaisiin sovelluksiin ja tilanteisiin, joissa ongelmien löytymisen todennäköisyys on noin 30 prosenttia. Tällainen *a priori* malli ei ota huomioon tietyn sovelluksen, tiettyjen testaaajien, tiettyjen testikäyttäjien, tietyn käyttökontekstin ja käytettyjen tekniikoiden sekä testitehtävien vaikutuksia, jotka ovat käytettävyydestissä merkittäviä (Molich et al., 1999; Spool & Schroeder, 2001; Thomas & Kellogg, 1989; Perälä, luku 19).

Pienen otoskoon säännön pohjana olevassa kaavassa oletetaan, että kaikki ongelmat löytyvät samalla todennäköisyydellä (Woolrych & Cockton, 2001). Kuitenkin ongelmien ilmenemisessä on suurtakin vaihtelua ja lisäksi osallistujien välillä yksilöllisiä eroja siinä, kuinka monta ongelmaa he kohtaavat. Tämä tulee esille myös Virzin (1992) tuloksista: tutkimuksessa yksittäinen osallistuja kohtasi 15–71 % kaikista löydetystä ongelmista ja yksittäisen ongelman kohtasi 5–95 % osallistujista. Osallistujien yksilöllisten erojen lisäksi ongelmien löytymisen todennäköisyyteen vaikuttavat testattava sovellus ja testitehtävät (Woolrych & Cockton, 2001). Sovelluksen koko vaikuttaa myös oleellisesti, koska laajaa sovellusta testattaessa testitehtävät eivät kata läheskään kaikkia toimintoja ja etenkin www-sovelluksissa kaikkia vaihtoehtoisia toimintaketjuja. Woolrych ja Cockton (2001) muistuttavat, että pelkkä ongelmien laskeminen ei palvele käytettävyyden arvioimista vaan on huomioitava myös ongelmien yleisyys ja vakavuus, jotta ongelmia voidaan ymmärtää ja priorisoida niiden korjaustarpeita.

18.4. Lopuksi

Käytettävyydestutkimuksessa tutkitaan, ”kuinka hyvin tietyt käyttäjät pystyvät käyttämään tuotetta tuloksellisesti, tehokkaasti ja miellyttävästi tiettyjen tavoitteiden saavuttamiseksi tietyssä käyttökontekstissa” (ISO 9241-11, 1998). ISO-standardin mukaan käytettävyys koskee siis tiettyjä käyttäjiä, tiettyjä tehtäviä ja tiettyä käyttökontekstia. Osallistujien valinnan tavoitteena on muodostaa edustava otos valitsemalla juuri niitä ”tiettyjä” käyttäjiä, joille tuote on suunnattu.

Käytettävyyden kontekstisidonnaisuudesta johtuen käytettävyydestin otoskoon suuruutta on vaikea arvioida etukäteen, koska käytettävyysongelmien löytymisen todennäköisyys riippuu tietyistä käyttäjistä, tehtävistä ja käyttökontekstista. Jos sovelluksen käytettävyyttä testataan vain yhden kerran, osallistujamäärää ei kannata kiinnittää etukäteen vaan jatkaa testausta resurssien puitteissa niin kauan, kun merkityksellistä uutta informaatiota tuntuu löytyvän (Virzi, 1992). Kun ongelmat alkavat toistaa itseään, testaamisen jatkaminen ei enää ole kustannustehokasta. On myös hyväksyttävä, että yksittäisellä käytettävyydestillä ei voida selvittää kaikkia ongelmia vaan ainoastaan osa niistä (Barnum, 2002a).

Käytettävyydestutkimuksella on erilaisia tavoitteita. Formatiivisen käytettävyydestin tavoitteena on havaita käytettävyyso ongelmia kun taas summatiivisessa käytettävyydestissä vertaillaan eri suunnitteluvaihtoehtoja ja ollaan kiinnostuneita tilastollisesti merkitsevistä eroista. Summatiivinen testaus vaatii suurempia osallistujamääriä kuin formatiivinen testaus. Joidenkin osallistavien menetelmien kuten esimerkiksi fokus-ryhmien ja tilannetutkimuksen tavoitteena ei niinkään ole käytettävyyso ongelmien löytäminen vaan käytettävyyden suunnittelu ja arviointi tuotekehityksen alkuvaiheessa.

Huolellisen osallistujien valinnan lisäksi osallistujilta on hyvä kerätä tarkempaa tietoa heidän ominaisuuksistaan ennen testiä tai testin jälkeen, jotta havaintoja voidaan tulkita tarkemmin (Dumas & Redish, 1993). Testien suorittamisen ja tulosten analysoinnin jälkeen tulisi pohtia, oliko valittu otos soveltuva ja miten valittu otos mahdollisesti vaikuttaa tuloksiin. Sen lisäksi, että osallistujien valintatavan ja määrän vaikutuksia tuloksiin arvioidaan, osallistujien valintaperusteet pitäisi myös raportoida ja perustella (Mirel, 1990).



Jenni Anttonen, fil.yo. Aloitin tietojenkäsittelytieteiden opiskelun Tampereen yliopistossa vuonna 1999. Olen opintojeni loppusuoralla, valmistun maisteriksi vuorovaikutteisesta teknologiasta vuonna 2004. Olen työskennellyt TAUCHI-yksikössä kesästä 2003 alkaen Emotions, Sociality, and Computing nimisessä tutkimusryhmässä. Tutkimukseni kohteena ovat käyttäjän emotioneihin liittyvät fysiologiset muutokset ja niiden havaitseminen.

Ohjaaja: Päivi Majaranta

Opponentit: Henna Heikkilä ja Katri Kosonen

Lähteet

Barnum, C.M. (2002a) The “magic number 5”: Is it enough for web testing? *Proceedings of the 1st European UPA Conference (EUPA 2002)*, September 2002, 45–52.

Barnum, C.M. (2002b) *Usability Testing and Research*. New York: Longman.

Barnum, C., Bevan, N., Cockton, G., Nielsen, J., Spool, J. & Wixon, D. (2003) The “magic number 5”: Is it enough for web testing? *Extended abstracts Human Factors in Computing Systems (CHI 2003)*, 698–699.

Dillon, A. & Watson, C. (1996) User analysis in HCI – the historical lessons learned from individual differences research. *International Journal of Human-Computer Studies*, 45(6), 619–637.

Dumas, J.S. (2003) User-based evaluations. In Jacko, J.A. & Sears, A. (eds.) *The Human-Computer Interaction Handbook—Fundamentals, Evolving Technologies and Emerging Applications*. Lawrence Erlbaum Associates, Inc, 1093–1117.

Dumas, J.S. & Redish, J.C. (1993) *A Practical Guide to Usability Testing*. Norwood N.J., Ablex.

Hackos, J.T. & Redish, J.C. (1998) *User and Task Analysis for Interface Design*. New York, Wiley.

Henry, S.L., Law, C. & Barnickle, K. (2001) Adapting the design process to address more customers in more situations. *Proc. of the Usability Professionals' Association (UPA 2001)*. <http://www.uiaccess.com/upa2001a.html> (21.6.2004)

Holleran, P. (1991) A methodological note on pitfalls in usability testing. *Behaviour & Information Technology*, 10(5), 345–357.

ISO 9241-11. (1998) Ergonomic requirements for office work with visual display terminals (VCTs) - Part 11: Guidance on usability.

Lewis, J.R. (1994) Sample sizes for usability studies: Additional considerations. *Human Factors*, 36(2), 368–378.

Mirel, B. (1990) Usability and hardcopy manuals: evaluating research designs and methods. *ACM SIGDOC Asterisk Journal of Computer Documentation*, 14(4), 69–77.

Molich, R., Thomsen, A. D., Karyukina, B., Schmidt, L., Ede, M., van Oel, W., & Arcuri, M. (1999) Comparative evaluation of usability tests. *Extended abstracts CHI 1999*, 83–84.

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

Nielsen, J. (1994) Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, 41 (3), September 1994, 385–397.

- Nielsen, J. (1997) Tech-support tales: Internet hard to use for novice users. Jakob Nielsen's Alertbox, April 1, 1997 <http://www.useit.com/alertbox/9704a.html> (21.6.2004)
- Nielsen, J. (2000a) Novice vs. expert users. Jakob Nielsen's Alertbox, February 6, 2000. <http://www.useit.com/alertbox/20000206.html> (21.6.2004)
- Nielsen, J. (2000b) Why you only need to test with five users. Jakob Nielsen's Alertbox, March 19, 2000. <http://www.useit.com/alertbox/20000319.html> (21.6.2004)
- Nielsen, J. & Landauer, T. K. (1993) A mathematical model of the finding of usability problems. *Proc. of Human Factors in Computing Systems (INTERCHI'93)*, ACM Press, 206–213.
- Robson, C. (1994) *Real World Research – A resource for social scientists and practioner-researchers*. Oxford: Blackwell.
- Rosenbaum, S. (1987) Selecting the appropriate subjects: subject selection for documentation usability testing. *Proceedings of the 1987 IEEE International Professional Communication Conference*, 135–142.
- Rubin, J. (1994) *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. New York: John Wiley & Sons.
- Spool, J. & Schroeder, W. (2001) Testing web sites: Five users is nowhere near enough. *Extended abstracts Human Factors in Computing Systems (CHI 2001)*, 285–286.
- Thomas, J.C. & Kellogg, W.A. (1989) Minimizing ecological gaps in interface design. *IEEE Software* 6 (1), 1989, 78–86.
- Virzi, R.A. (1992) Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34 (4), 457–468.
- Woolrych, A. & Cockton, G. (2001) Why and when five test users aren't enough. In Blandford, A., Vanderdonck, J. & Gray, P. (Eds.) *People and Computers XV Joint Proceedings of HCI 2001 and IHM 2001 (IHM-HCI2001)*, Springer-Verlag, Vol. 2, 105–108.

Kuvien lähteet

Kuva 2: Käännetty suomeksi. Alkuperäinen kuva löytyy lähteestä: Nielsen, J. (2000) Why you only need to test with five users. Jakob Nielsen's Alertbox, March 19, 2000. <http://www.useit.com/alertbox/20000319.html> (21.6.2004)

Liite 1. Taustatietolomake.

TESTIKÄYTTÄJÄN TAUSTATIEDOT

Nimi (etunimi riittää): _____

Ikä: 18-30 31-40 41-50 51-
 Sukupuoli: nainen mies
 Kätisyys: oikea vasen

Seuraavaksi toivon, että vastaisit seuraaviin tietotekniseen kokemukseesi liittyviin kysymyksiin.

1. Kuinka kauan olet käyttänyt tietokoneita?

- Alle vuoden
- 1-2 vuotta
- 3-5 vuotta
- 5-10 vuotta
- Yli 10 vuotta

2. Kuinka usein käytät verkkosivuja (www-sivuja)?

- 6-7 päivänä viikossa
- 3-5 päivänä viikossa
- 1-2 päivänä viikossa
- Harvemmin
- En koskaan

3. Mitä Internet-selainta käytät pääasiassa (voit rastittaa myös usean vaihtoehdon)?

- Microsoft Internet Explorer
- Netscape Navigator
- Muu selain, mikä? _____

4. Millaisia www-sivustoja ja -palveluita lähinnä käytät? Aseta itsellesi tärkeä verkkosivujen käyttö järjestykseen (1 = tärkein, 6 = vähiten tärkeä)

- ___ Julkisten palveluiden käyttö (esim. kunnan ja KELAn palvelut)
- ___ Kaupallisten palveluiden käyttö (esim. pankki, VR; myös tiedonhaku näistä)
- ___ Työhön ja opiskeluun liittyvä käyttö (esim. yliopiston tai työpaikan sivut)
- ___ Harrastuksiin liittyvä käyttö (esim. harrastajien omat sivut, tiedonetsintä, ...)
- ___ Viihdekäyttö (pelit, aikuisviihde)
- ___ Muu käyttö, millainen? _____

5. Kuinka usein haet tietoa Internetin hakupalveluiden avulla?

Työssä tai opiskelussa

Vapaa-aikana

- | | |
|--|--|
| <input type="checkbox"/> Päivittäin | <input type="checkbox"/> Päivittäin |
| <input type="checkbox"/> Muutaman kerran viikossa | <input type="checkbox"/> Muutaman kerran viikossa |
| <input type="checkbox"/> Muutaman kerran kuukaudessa | <input type="checkbox"/> Muutaman kerran kuukaudessa |
| <input type="checkbox"/> Harvemmin | <input type="checkbox"/> Harvemmin |
| <input type="checkbox"/> En koskaan | <input type="checkbox"/> En koskaan |

6. Kirjoita alla oleville riveille useimmin käyttämäsi Internet-hakupalvelut.

Työssä tai opiskelussa

Vapaa-aikana

Kiitos!