

READER

Rich Internet Application

Interaction Design

Les 7 – Usability testing

INHOUDSOPGAVE

INLEIDING	1
Hoofdstuk 1 Wat is Usability Testing?	2
§1.1 Gedeelte uit 'Don't make me Think'	2
§1.2 Gedeelte uit 'A Practical Guide to Usability Testing'	5
§1.3 Samenvattend 'Wat is Usability Testing'?	7
Hoofdstuk 2 Enkele basisbegrippen	10
Hoofdstuk 3 Het voorbereiden van de test	13
§3.1 Bepalen doelstellingen.....	13
§3.2 Bepalen gebruikers/deelnemers	14
§3.3 Kiezen van taken	16
§3.4 Opstellen scenario's.....	19
§3.5 Opstellen vragenlijsten.....	21
§3.6 Bepalen wat je gaat meten met bijbehorende norm	25
Hoofdstuk 4 Afname test en ordenen gegevens	27
§4.1 Afname van de test.....	27
§4.2 Ordenen en verwerken van de gegevens	28
§4.3 Presenteren van de gegevens	30
§4.4 Enkele maatstaven	33
§4.5 Gebruik van Excel.....	36
Hoofdstuk 5 Analyseren van gegevens	38
§5.1 Triangulatie	38
§5.2 Analyseren kwantitatieve gegevens.....	39
§5.3 Analyseren kwalitatieve (ordinale) gegevens.....	40
§5.4 Bepalen van de omvang en de ernst van de problemen	42
§5.5 Rapporteren	44
Bijlage	45
Why You Only Need to Test With 5 Users	45

INLEIDING

Deze syllabus bevat de kerninformatie die nodig is voor het uitvoeren van een Usability Test. Op de markt zijn er veel (Amerikaanse) boeken te krijgen die uitgebreid ingaan op dit onderwerp. Bij sommige boeken wordt er ook aandacht besteed aan het voortraject: gebruikersanalyse, taakanalyse en prototyping. Andere boeken richten zich volledig op het testgebeuren. Mocht je meer informatie willen hebben, dieper op de materie willen inzoomen (en dan met name op het testgebeuren, niet op analysegebied) dan kan ik de volgende boeken aanbevelen:

1. A Practical Guide in Usability Testing van Joseph S. Dumas en Janice C.Redish
2. Usability Testing and Research van Carol N. Barum

Bij het samenstellen van deze syllabus is o.a. gebruik gemaakt van deze 2 boeken.

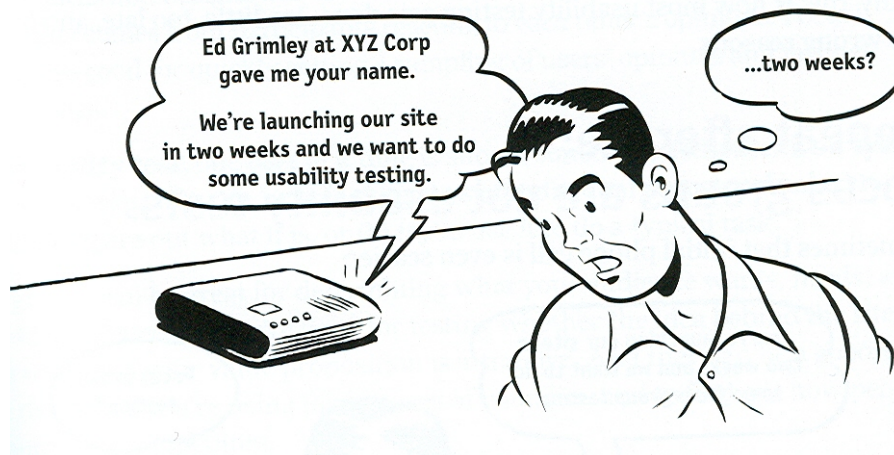
Hoofdstuk 1 Wat is Usability Testing?

In dit hoofdstuk gaan we in op wat Usability Testing nu precies inhoudt en wat het niet is. Ook behandelen we een aantal basisbegrippen vanuit de statistiek die nodig zijn om verder te kunnen gaan.

Om een idee te krijgen wat moet worden verstaan onder Usability Testing volgen in §1.1 en §1.2 twee stukken overgenomen uit respectievelijk 'Don't make me Think', 2nd edition van Steve Krug (blz 131-135) en 'A Practical Guide to Usability Testing' van J.S. Dumas en J.C. Redish (blz 22 t/m 25).

§1.1 Gedeelte uit 'Don't make me Think'

About once a month, I get one of these phone calls:



As soon as I hear "launching in two weeks" (or even "two months") and "usability testing" in the same sentence, I start to get that old fireman-headed-into-the-burning-chemical-factory feeling, because I have a pretty good idea of what's going on.

If it's two weeks, then it's almost certainly a request for a disaster check. The launch is fast approaching and everyone's getting nervous, and someone finally says, "Maybe we better do some usability testing."

If it's two months, then odds are that what they want is to settle some ongoing internal debates – usually about something very specific like color schemes. Opinion around the office is split between two different designs; some people like the sexy one, some like the elegant one. Finally someone with enough clout to authorize the expense gets tired of the arguing and says, "All right, let's get some testing done to settle this."

And while usability testing will sometimes settle these arguments, the main thing it usually ends up doing is revealing that the things they were arguing about aren't all that important. People often test to decide which color drapes are best, only to learn that they forgot to put windows in the room. For instance, they might discover that it doesn't make much difference whether you go with the horizontal navigation bar or the vertical menu if nobody understands the value proposition of your site.

Sadly, this is how most usability testing gets done: too little, too late, and for all the wrong reasons.

Repeat after me: Focus groups are not usability tests.

Sometimes that initial phone call is even scarier:



When the last-minute request is for a focus group, it's usually a sign that the request originated in Marketing. When Web sites are being designed, the folks in Marketing often feel like they don't have much clout. Even though they're the ones who spend the most time trying to figure out who the site's audience is and what they want, the designers and developers are the ones with most of the hands-on control over how the site actually gets put together.

As the launch date approaches, the Marketing people may feel that their only hope of sanity prevailing is to appeal to a higher authority: research. And the kind of research they know is focus groups.

I often have to work very hard to make clients understand that what they need is usability testing, not focus groups. Here's the difference in a nutshell:

- In a **focus group**, a small group of people (usually 5 to 8) sit around a table and react to ideas and designs that are shown to them. It's a group process, and much of its value comes from participants reacting to each other's opinions. Focus groups are good for quickly getting a sampling of users' opinions and feelings about things.
- In a **usability test**, one user at a time is shown something (whether it's a Web site, a prototype of a site, or some sketches of individual pages) and asked to either (a) figure out what it is, or (b) try to use it to do a typical task.

Focus groups can be great for determining what your audience wants, needs, and likes – in the abstract. They're good for testing whether the idea behind the site makes sense and your value proposition is attractive. And they can be a good way to test the names you're using for features of your site, and to find out how people feel about your competitors.

But they're *not* good for learning about whether your site works and how to improve it.

The kinds of things you can learn from focus groups are the things you need to learn early on, *before* you begin designing the site. Focus groups are for EARLY in the process. You can even run them late in the process if you want to do a reality check and fine-tune your message, but *don't* mistake them for usability testing. They *won't* tell you whether people can actually use your site.

Several true things about testing

Here are the main things I know about testing:

- **If you want a great site, you've got to test.** After you've worked on a site for even a few weeks, you can't see it freshly anymore. You know too much. The only way to find out if it really works is to test it.
Testing reminds you that not everyone thinks the way you do, knows what you know, uses the Web the way you do.

I used to say that the best way to think about testing was that it was like travel: a broadening experience. It reminds you how different – and the same – people are, and gives you a fresh perspective on things.

But I finally realized that testing is really more like having friends visiting from out of town. Inevitably, as you make the tourist rounds with them, you see things about your home town that you usually don't notice because you're so used to them. And at the same time, you realize that a lot of things that you take for granted aren't obvious to everybody.

- **Testing one user is 100 percent better than testing none.** Testing always works, and even the worst test with the wrong user will show you important things you can do to improve your site. I make a point of always doing a live user test at my workshops so that people can see that it's very easy to do and it always produces an abundance of valuable insights. I ask for a volunteer and have him try to perform a task on a site belonging to one of the other attendees. These tests last less than ten minutes, but the person whose site is being tested usually scribbles several pages of notes. And they always ask if they can have the recording of the test to show to their team back home. (One person told me that after his team saw the recording, they made one change to their site which they later calculated had resulted in \$100,000 in savings.)
- **Testing one user early in the project is better than testing 50 near the end.** Most people assume that testing needs to be a big deal. But if you make it into a big deal, you won't do it early enough or often enough to get the most out of it. A simple test early – while you still have time to use what you learn from it – is almost always more valuable than a sophisticated test later.

Part of the conventional wisdom about Web development is that it's very easy to go in and make changes. The truth is, it turns out that it's not that easy to make changes to a site once it's in use. Some percentage of users will resist almost any kind of change, and even apparently simple changes often turn out to have far-reaching effects, so anything you can keep from building wrong in the first place is gravy.

- **The importance of recruiting representative users is overrated.** It's good to do your testing with people who are like the people who will use your site, but it's much more important to test early and often. My motto – as you'll see – is "Recruit loosely, and grade on a curve."
- **The point of testing is not to prove or disprove something. It's to inform your judgment.** People like to think, for instance, that they can use testing to prove whether navigation system "a" is better than navigation system "b", but you can't. No one has the resources to set up the kind of controlled experiment you'd need. What testing *can* do is provide you with invaluable input which, taken together with your experience, professional judgment, and common sense, will make it easier for you to choose wisely – and with greater confidence – between "a" and "b."

- **Testing is an iterative process.** Testing isn't something you do once. You make something, test it, fix it, and test it again.
- **Nothing beats a live audience reaction.** One reason why the Marx Brothers' movies are so wonderful is that before they started filming they would go on tour on the vaudeville circuit and perform scenes from the movie, doing five shows a day, improvising constantly and noting which lines got the best laughs. Even after they'd settled on a line, Groucho would insist on trying slight variations to see if it could be improved.

§1.2 Gedeelte uit 'A Practical Guide to Usability Testing'

What is Usability Testing

While there can be wide variations in where and how you conduct a usability test, every usability test shares these five characteristics:

1. The primary goal is to improve the usability of a product. For each test, you also have more specific goals and concerns that you articulate when planning the test.
2. The participants represent real users.
3. The participants do real tasks.
4. You observe and record what participants do and say.
5. You analyze the data, diagnose the real problems, and recommend changes to fix those problems.

The Goal is to Improve the Usability of a Product

The primary goal of a usability test is to improve the usability of the product that is being tested. Another goal, as we will discuss in detail later, is to improve the *process* by which products are designed and developed, so that you avoid having the same problems again in other products.

This characteristic distinguishes a usability test from a research study, in which the goal is to investigate the existence of some phenomenon. Although the same facility might be used for both, they have different purposes.

This characteristic also distinguishes a usability test from a quality assurance or function test, which has a goal of assessing whether the product works according to its specifications.

Within the general goal of improving the product, you will have more specific goals and concerns that differ from one test to another.

- You might be particularly concerned about how easy it is for users to navigate through the menus. You could test that concern before coding the product, by creating an interactive prototype of the menus, or by giving users paper versions of each screen.
- You might be particularly concerned about whether the interface that you have developed for novice users will also be easy for and acceptable to experienced users.
- For one test, you might be concerned about how easily the customer representatives who do installations will be able to install the product. For another test, you might be concerned about how easily the client's nontechnical staff will be able to operate and maintain the product.

These more specific goals and concerns help determine which users are appropriate participants for each test and which tasks are appropriate to have them do during the test.

The Participants Represent Real Users

The people who come to test the product must be members of the group of people who now use or who will use the product. A test that uses programmers when the product is intended for legal secretaries is not a usability test.

The quality assurance people who conduct function tests may also find usability problems, and the problems they find should not be ignored, but they are not conducting a usability test. They are not real users – unless it is a product about function testing. They are acting more like expert reviewers.

If the participants are more experienced than actual users, you may miss problems that will cause the product to fail in the marketplace. If the participants are less experienced than actual users, you may be led to make changes that aren't improvements for the real users.

If the participants 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.

The Participants Do Real Tasks

The tasks that you have users do in the test must be ones that they will do with the product on their jobs or in their homes. This means that you have to understand users' jobs and the tasks for which this product is relevant.

In many usability tests, particularly of functionally rich and complex software products, you can only test some of the many tasks that users will be able to do with the product. In addition to being realistic and relevant for users, the tasks that you include in a test should relate to your goals and concerns and have a high probability of uncovering a usability problem.

Observe and Record What the Participants Do and Say

In a usability test, you usually have several people come, one at a time, to work with the product. You observe the participant, recording both performance and comments. You also ask the participant for opinions about the product. A usability test includes both times when participants are doing tasks with the product and times when they are filling out questionnaires about the product. Observing and recording individual participant's behaviours distinguishes a usability test from focus groups, surveys, and beta testing.

A typical focus group is a discussion among 8 to 10 real users, led by a professional moderator. Focus groups provide information about users' opinions, attitudes, preferences, and their self-report about their performance, but focus groups do not usually let you see how users actually behave with the product.

Surveys, by telephone or mail, let you collect information about users' opinions, attitudes, preferences, and their self-report of behaviour, but you cannot use a survey to observe and record what users actually do with a product.

A typical beta test (field test, clinical trial, user acceptance test) is an early release of a product to a few users. A beta test has ecological validity, that is, real people are using the product in real environments to do real tasks. However, beta testing seldom yields any useful information about usability. Most companies have found beta testing to be too little, too unsystematic, and *much too late* to be the primary test of usability.

Analyze the Data, Diagnose the Real Problems, and Recommend Changes to Fix Those Problems

Collecting the data is necessary, but not sufficient, for a usability test. After the test itself, you still need to analyze the data. You consider the quantitative and qualitative data from the participants together with your own observations and users' comments. You use all of that to diagnose and document the product's usability problems and to recommend solutions to those problems.

"Tabulating and Analyzing Data," is not a trivial task. Usability testing is distinguished from beta testing by both the quality and quantity of data that you have. The data are systematic, comparable across the participants that you saw, and very rich.

The Results Are Used to Change the Product - and the Process

We would also add another point. It may not be part of the definition of the usability test itself, as the previous five points were, but it is crucial, nonetheless.

A usability test is not successful if it is used only to mark off a milestone on the development schedule. A usability test is successful only if it helps to improve the product that was tested and the process by which it was developed.

Someone must use the results of the usability test.

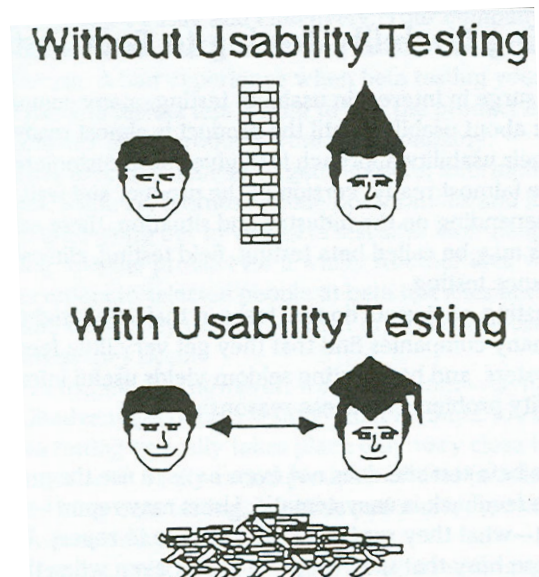


Figure 1 Usability testing can break down the wall between those who create the product and those who use it.

§1.3 Samenvattend 'Wat is Usability Testing'?

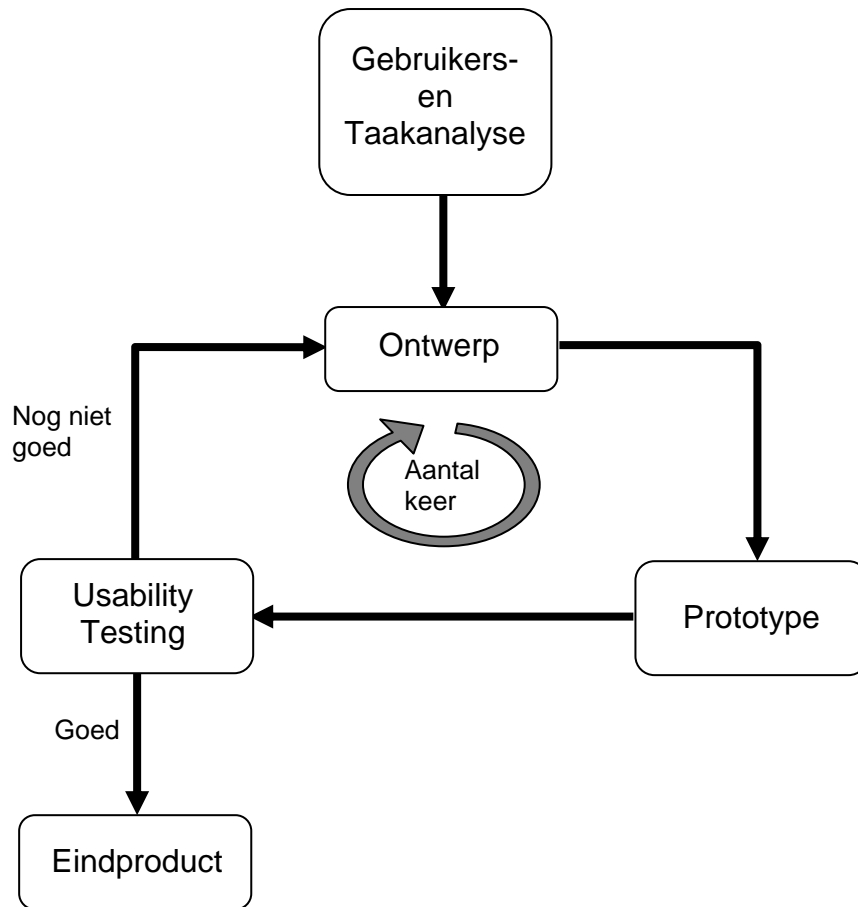
In de vorige 2 paragrafen is op verschillende wijzen getracht duidelijk te maken wat Usability Testing is en inhoudt. Het gedeelte uit 'Don't make me think' is vrij intuïtief geschreven. Het 2^e gedeelte is wat preciezer van aard. Duidelijk wordt wel uit beide stukken dat Usability Testing iets anders is dan een onderzoek, beta-testing, een functionaliteitstest of een focusgroep.

Ook wordt duidelijk dat Usability Testing noodzakelijk is bij de ontwikkeling (of aanpassing) van een product (ongeacht het soort product). Verder is het het beste als Usability Testing niet iets eenmaligs is binnen het ontwikkelproces, maar dat het onderdeel uitmaakt van een iteratief proces: Usability Testing komt niet (alleen) aan het eind van het ontwikkelproces, maar vindt geregeld plaats. Schematisch is een en ander weergegeven in figuur 1.2.

Dit schema is met een kleine aanpassing ook van toepassing op het verbeteren van een bestaand product. De start van de cyclus ligt dan niet bij het ontwerp, maar bij de Usability Test, de zogenaamde 0-meting.

Aangezien een Usability Test dus niet enkel aan het eind komt van een ontwikkelingsproces, voer je een Usability Test niet noodzakelijk uit op een afgerond product. Het is heel goed mogelijk (en zinvol) om een Usability Test uit te voeren met een prototype. Dit kan zowel een papieren prototype zijn als een highfidelity prototype. Ook kun je een Usability Test uitvoeren op een (klein) deel van het product om bijvoorbeeld een bepaald usability issue te behandelen.

Figuur 1.2 *Plaats van Usability Testing in het gehele ontwerp en ontwikkeltraject van een nieuw product.*



Kenmerken van een Usability Test

Een Usability Test heeft de volgende 5 kenmerken (volgens Dumas en Redish):

1. Het belangrijkste doel bij elke test is natuurlijk het verbeteren van de usability van een product. Dit is echter een ruim begrip en verschilt per product en test. Wat zijn nu voor deze test de specifieke doelstellingen? Deze geef je aan bij het opzetten van de test.
2. De deelnemers aan de test vertegenwoordigen de echte gebruikers
3. De deelnemers aan de test doen realistische taken.
4. Het waarnemen en vastleggen van wat de deelnemers aan de test doen en zeggen.
5. Het analyseren van de data, vaststellen van de werkelijke problemen en het doen van voorstellen voor aanpassingen om deze problemen te verhelpen.

Het laatste punt 5 is feitelijk waar alles omdraait en waarom de Usability Test werd gedaan. Dit is dus het belangrijkste (maar ook lastigste) onderdeel van de hele test.

Behalve dat elke Usability Test een vijftal kenmerken heeft, is een Usability Test opgebouwd uit een aantal achter elkaar te doorlopen fases. De hierboven genoemde 5 kenmerken vind je daarin terug. De verschillende fases binnen een Usability Test worden hieronder verder toegelicht.

Fasering binnen een Usability Test

Usability Testing is opgebouwd uit een aantal achtereenvolgens te doorlopen fasen. Deze fasen volgen na de gebruikersanalyse en de taakanalyse. We mogen aannemen dat gebruikers en de taken van de gebruikers tussentijds niet veranderen. Deze analyses hoeven daarom maar één keer te worden uitgevoerd gedurende het gehele ontwerp en ontwikkeltraject van een product. We spreken over 'product' omdat het zowel kan gaan om een internetsite als om andere software. De fasen waaruit Usability Testing bestaat, zijn:

1. Plannen en voorbereiden test

Deze fase omvat onder meer het vaststellen van de te benaderen groep gebruikers, het bepalen van de steekproefomvang, het kiezen van de gebruikers voor de test en het maken van afspraken daarmee. Daarnaast omvat deze fase ook het bepalen van de taken die je wilt laten uitvoeren en het opstellen van de vragenlijsten en alvast nadenken over de te hanteren analysemethoden.

2. Afname van de test

Deze fase omvat het uitvoeren van de test: observeren van de gekozen gebruikers bij het uitvoeren van de opgelegde taken en hen interviewen.

3. Verwerking van de gegevens

Deze fase omvat het ordenen van de verzamelde gegevens.

4. Analyse van de gegevens

In deze fase worden de vergaarde kwantitatieve en kwalitatieve gegevens geanalyseerd. Beide soorten gegevens zijn van belang. De aandacht moet dus niet alleen uitgaan naar de kwantitatieve gegevens, integendeel. De analyse van de kwalitatieve gegevens is weliswaar lastiger, maar levert erg waardevolle informatie op.

De analyse van kwantitatieve gegevens kan variëren van tabellen en grafieken en het interpreteren daarvan (ook niet altijd eenvoudig) tot allerlei geavanceerde analyses (deze vallen buiten het bestek van dit semester).

5. Rapportage/presentatie resultaat analyse

Deze laatste fase omvat de interpretatie van de resultaten van de analyses, het trekken van conclusies en het doen van aanbevelingen. Dit kan zowel schriftelijk als mondeling gebeuren.

Zoals hiervoor bij het noemen van de kenmerken van een Usability Test al is opgemerkt, komen de 5 kenmerken terug in bovengenoemde fasering. Zo omvat fase 1 de kenmerken 1 t/m 3 van een Usability Test. Zijn fase 2 & 3 samen kenmerk 4 en fase 4 & 5 samen kenmerk 5.

De indeling van de volgende hoofdstukken is o.a. gebaseerd op bovenstaande fasering.

We beginnen echter eerst met een aantal basisbegrippen (hoofdstuk 2). In hoofdstuk 3 komt fase 1 aan bod. In hoofdstuk 4 de fases 2 en 3. In hoofdstuk 5 tenslotte komen fase 4 en 5 aan bod.

Hoofdstuk 2 Enkele basisbegrippen

Voor een goed onderzoek is het belangrijk om duidelijk af te bakenen welke personen of zaken *wel* en welke *niet* bij het onderzoek betrokken zullen worden. De groep personen (of dingen) waarop het onderzoek betrekking heeft, heet een **doelpopulatie**. Meestal wordt slechts een gedeelte van de doelpopulatie onderzocht. We hebben dan te maken met een **steekproef**.

Tijdens de afname van de Usability Test probeer je via vragen en/of observaties informatie over/van je deelnemers aan de test te krijgen. Bijvoorbeeld de mate van computerervaring van een persoon of zijn/haar mening over bepaalde functies en hun hanteerbaarheid. Dit zijn beide kenmerken die variëren van persoon tot persoon. Om zo'n kenmerk te kunnen gebruiken in je onderzoek (analyse van de testgegevens) moet het eerst zoals dat heet *operationeel* gemaakt worden. Dit betekent dat je het kenmerk moet zien als een **variabele** die bepaalde waarden aan kan nemen. Je moet beschrijven welke waarden die variabele kan aannemen en je moet vastleggen hoe je op éénduidige wijze aan elke deelnemer aan de test een waarde van deze variabele kunt toekennen. Deze toekenning wordt meten genoemd en de toegekende waarde heet **meetwaarde, waarnemingsuitkomst** of **score**.

Soorten variabelen

Variabelen kunnen op meerdere manieren worden onderverdeeld. Belangrijk is de onderverdeling van variabelen naar hun meetniveau. Het meetniveau van een variabele bepaalt mede welke statistische analyses bij die variabele geschikt zijn.

Alle mogelijke waarnemingsuitkomsten van een variabele vormen samen een *schaal*. Een schaal kan zowel uit getallen als uit woorden bestaan, afhankelijk van het type variabele.

Gaat het om zaken zoals lengte, gewicht of inkomen, dan hebben we te maken met een

kwantitatieve variabele. In dat geval zijn de mogelijke waarnemingsuitkomsten altijd getallen.

Gaat het echter om bijvoorbeeld meningen, geslacht of merknamen, dan hebben we te maken met een **kwalitatieve variabele**. De mogelijke waarnemingsuitkomsten zijn dan woorden óf getallen. Zo zijn bij de variabele 'geslacht' de meetwaarden 'man' en 'vrouw', maar je kunt ook afspreken dat mannen code 1 krijgen en vrouwen code 2. Aan deze getallen kun je echter geen betekenis hechten in de vorm van 2x zoveel en dergelijke, het zijn slechts labels. Het feit dat waarnemingsuitkomsten uit getallen bestaat, zegt dus nog niets over het type variabele en wat je ermee kunt.

Hiervoor is al vermeld dat er in eerste instantie 2 type variabelen zijn: kwalitatieve variabelen en kwantitatieve variabelen. Binnen beide groepen kan nog een nader en zinvol onderscheid gemaakt worden. Deze nadere uitsplitsing vatten we hieronder kort samen:

Kwalitatieve variabele van nominaal niveau

Bij een nominale variabele is de volgorde waarin de verschillende uitkomsten genoemd worden geheel willekeurig. Voorbeelden van nominale variabelen zijn: geslacht, automerken of de opleidingen binnen ICA. De eventuele getallen die aan de antwoordmogelijkheden gekoppeld worden, dienen alleen als afkorting. Het zijn naambordjes voor de verschillende mogelijkheden. Dit is gemakkelijk bij het invoeren in de computer, maar verder heeft het getal als zodanig geen functie.

Kwalitatieve variabele van ordinaal niveau

Bij een ordinale variabele is er sprake van een logische volgorde bij de verschillende antwoordmogelijkheden. Voorbeelden van ordinale variabelen zijn: mening of Michelin-sterren. De eventuele getalcodes die aan de antwoordmogelijkheden gekoppeld worden weerspiegelen de logische volgorde die bij de antwoordmogelijkheden aanwezig is. Begrippen als 'meer of minder' of 'beter of slechter' bij de antwoorden komen overeen met het groter of kleiner zijn van de getallen. Verder is er geen betekenis toe te kennen aan deze getallen en kan er dus ook niet mee gerekend worden.

Kwantitatieve variabele

De waarden voor deze variabele fungeren als echte getallen. Niet alleen groter en kleiner, maar ook optellen en aftrekken hebben hun gebruikelijke betekenis. Een kwantitatieve variabele meet zaken als hoe veel, hoe groot en hoe lang.

Bij kwantitatieve variabele maken we onderscheid tussen **discrete** en **continue** variabelen. Men spreekt van een **discrete** variabele wanneer de mogelijke waarden van die variabele op een rijtje gezet kunnen worden (óf er zijn eindig veel mogelijkheden óf de mogelijkheden kunnen worden genummerd met de natuurlijke getallen). Voorbeelden zijn: aantal kinderen, aantal keren meespelen in de lotto voor je de hoofdprijs wint.

Daarnaast zijn er variabelen die een heel interval op de getallenlijn als mogelijke waarden hebben. Dit is het geval bij grootheden zoals lengte, massa en tijdsduur. Hoewel de waargenomen waarden meestal worden afgerond op hele centimeters, grammen, minuten of desnoods op honderdste seconden, zoals in de topsport, is het voor de theorie eenvoudiger te doen alsof de bewuste variabele in principe alle waarden op de getallenlijn (of een zeker interval daarvan) kan aannemen. Dergelijke variabelen noemt men **continu**.

Voorbeeld 1

Indeling naar geslacht: 1 = man; 2 = vrouw.

Dit is een kwalitatieve variabele van nominaal niveau.

Voorbeeld 2

"UserInterfaceDesign is een belangrijk vak in de opleiding CS".

Met deze uitspraak ben ik het (omcirkel uw keuze):

1	2	3	4	5
Geheel oneens	Oneens	Noch oneens noch eens	Eens	Geheel eens

Een hogere score betekent: meer eens met de uitspraak. De ordening tussen de getallen is dus betekenisvol voor de variabele. Het is dus een kwalitatieve variabele van ordinaal niveau. Tussen de opeenvolgende getallen 1, 2, 3, 4 en 5 bestaat in dit voorbeeld steeds een gelijk verschil van 1.

Mogen we hieraan betekenis hechten? Bestaan er tussen de waarderingen "geheel oneens" "geheel eens" ook gelijke afstanden?

Nee, dat hoeft niet! Tussen "noch oneens, noch eens" en "eens" kan psychologisch een veel bredere kloof liggen, dan tussen "eens" en "geheel eens". Als een meting op ordinaal niveau plaatsvindt, mogen uit de grootte van de verschillen tussen de scores géén conclusies worden getrokken. De praktijk leert echter dat dit veelvuldig gebeurt.

Wanneer je in een enquête een vraag stelt, dan kun je een antwoord vragen op verschillende meetniveaus. Een antwoord op nominaal niveau is het minst informatief, maar de vraag is wel erg eenvoudig te beantwoorden door de respondent. Wil je een antwoord op ordinaal niveau, dan zit daar meer informatie in, maar het antwoord vergt meer denkwerk van de respondent. Tenslotte zal een antwoord op kwantitatief niveau zowel het meest informatief zijn als het moeilijkste te geven.

Kortom: Hoe hoger het meetniveau hoe meer informatie daar in zit, maar des te moeilijker door de respondent te beantwoorden.

Voorbeeld 3

Bij een onderzoek onder bedrijven in de informaticabranche stel je vragen over het al dan niet geabonneerd zijn op een zeker vakblad. Bij het antwoord 'ja' vraag je door. We bekijken 3 manieren.

1. Zijn er buiten uzelf nog anderen die dit blad lezen?

- Ja
- Nee

2. Hoeveel personen buiten uzelf lezen dit blad?

- verder niemand
- 1-2 anderen
- 3-5 anderen
- 6 of meer anderen

3. Hoeveel personen buiten uzelf lezen dit blad?
nog anderen

Bij manier 1 is het antwoord op nominaal niveau. Bij manier 2 is dat op ordinaal niveau en bij manier 3 is het een kwantitatieve discrete variabele.

Welke vraagstelling je gebruikt hangt af van het doel dat je voor ogen hebt. Maar wees je er wel van bewust dat het uitmaakt welke antwoordmogelijkheden je aanbiedt.

Hoofdstuk 3 Het voorbereiden van de test

In dit hoofdstuk werken we de eerste fase van een Usability Test verder uit. Zoals al beschreven in hoofdstuk 1 omvat deze fase onder meer het vaststellen van de te benaderen groep gebruikers, het bepalen van de steekproefomvang, het kiezen van de gebruikers voor de test en het maken van afspraken daarmee. Daarnaast omvat deze fase ook het bepalen van de taken die je wilt laten uitvoeren en het opstellen van de vragenlijsten en alvast nadenken over de te hanteren analysemethoden. Uit deze opsomming volgt dat ook het plannen en voorbereiden van de test weer onder te verdelen is in een aantal fases. Deze fases zijn:

1. Bepalen doelstellingen.
2. Bepalen Gebruikers/deelnemers.
3. Kiezen van taken.
4. Opstellen scenario's.
5. Opstellen interviewvragen.
6. Bepalen wat je gaat meten met bijbehorende norm.

Duidelijk zal worden dat deze fases met elkaar samenhangen. Keuzes die je gemaakt hebt in een eerdere fase beïnvloeden en/of beperken keuzes in latere fases. Als je al deze fases doorlopen hebt, heb je voldoende materiaal om de testen te gaan afnemen.

§3.1 Bepalen doelstellingen

De meeste producten zijn zo complex dat het niet mogelijk is om bij elke groep gebruikers alle aandachtspunten in een keer te testen. Zelfs al bij een eenvoudig product kan er tijdens een test zoveel tegelijk en snel gebeuren dat belangrijke gebeurtenissen je ontgaan als je van te voren niet hebt nagedacht waar je op gaat letten. Bij elke Usability Test moet je beginnen met na te denken over wat je wilt bereiken met de test, wat zijn je specifieke aandachtspunten? Uit de aandachtspunten volgen dan de specifieke doelstellingen voor de Usability Test. Wat zijn eigenlijk doelstellingen en aandachtspunten?

Aandachtspunt = is meestal geformuleerd als een vraag.
Bijvoorbeeld: 'Zijn gebruikers in staat om het correcte icoontje snel te vinden?'

Doelstelling = is meestal geformuleerd als een beschrijving.
Bijvoorbeeld: 'Gebruikers zijn in staat om het correcte icoontje binnen 30 seconden te vinden, met maximaal 1 fout'.

De aandachtspunten bepalen de punten die door de Usability Test achterhaald moeten worden. Wat wil je te weten komen met de test? De doelstellingen zijn een nadere uitwerking daarvan. Zij stellen de criteria voor het bepalen van de gebruiksvriendelijkheid vast en daar kijk je later bij de analyses weer op terug.

Zonder het vaststellen van doelstellingen en aandachtspunten kan het gebeuren dat je een prachtige test ontwikkeld hebt, die het sleutelprobleem van de ontwikkelaars van het product totaal negeert. Een andere mogelijkheid is dat je in de situatie zit dat het opzetten van de test geheel vast komt te zitten omdat niemand het met elkaar eens kan worden. Dit alles enkel en alleen omdat de testdoelstellingen nooit waren toevertrouwd aan papier.

Bij het bepalen van de aandachtspunten (of bij het vaststellen van de criteria voor de doelstellingen) zul je ook merken dat je (soms) onderscheid wil maken in type gebruiker. Bijvoorbeeld: ervaren en onervaren gebruikers.

Het formuleren van duidelijke en beknopte doelstellingen is makkelijker gezegd dan gedaan. Hieronder volgen 2 voorbeelden van niet afgebakende en vage doelstellingen.

Voorbeeld 1

- I. Is het huidige product bruikbaar?
- II. Is het product klaar om op de markt gebracht te worden of moet er nog aan gewerkt worden?

Het probleem met deze 2 doelstellingen is **niet** dat ze onzin zijn, maar ze zijn incompleet en erg vaag. Wat is 'bruikbaar' en wat is 'klaar'? Hoe meet je dit?

Elke Usability Test die hierop is gebaseerd zal tot de conclusie komen dat het product 'bruikbaar' is; de gebruikers konden er toch mee werken? Dit is dus zonde van de tijd en het geld.

Als de doelstellingen duidelijk zijn, dan zijn de volgende beslissingen die genomen moeten worden zoals, wie gaan we ondervragen en observeren, welke taken laten we ze uitvoeren, wat willen we verder nog meten, niet zo moeilijk meer om te nemen.

§3.2 Bepalen gebruikers/deelnemers

Een belangrijk punt bij Usability Testing is dat de groep gebruikers die deelneemt aan de test (= steekproef) overeenkomt met de werkelijke gebruikers (= doelpopulatie) van het product. Bij Usability Testing krijg je een goed beeld van de doelpopulatie door goed te kijken naar de gebruikersanalyse. Het is daarom ook erg belangrijk voor het verdere verloop van de test dat de gebruikersanalyse (zeer) zorgvuldig is opgesteld. De testresultaten zijn namelijk alleen geldig als de personen die betrokken worden bij de test ook daadwerkelijk typische eindgebruikers zijn van het product (of daar zo dicht mogelijk tegen aan zitten). Worden namelijk de 'verkeerde' mensen bij de test betrokken dan maakt het niet uit hoeveel energie en tijd je steekt in de rest van de test voorbereiding en de resultaatverwerking. De resultaten zijn dan discutabel en dus van minder waarde.

De doelpopulatie wordt eventueel nog ingeperkt door de eerder opgestelde doelstellingen en/of aandachtspunten. Als daar gericht wordt op bijvoorbeeld beginnende gebruikers, zal dit de gebruikersgroep waaruit gekozen kan worden kleiner maken.

Wanneer eenmaal duidelijk is wat de doelpopulatie is en dus uit welke groep mensen gekozen kan worden, bestaat de volgende stap uit het kiezen van personen die daadwerkelijk deel gaan nemen aan de Usability Test. De volgende vraag is dan: behandelen we alle gebruikers gelijk of maken we onderscheid in gebruiker (bijvoorbeeld veel of weinig ervaring). Kortom: moeten de gebruikers ingedeeld worden in subgroepen?

Subgroepen

Bij veel producten zullen de eindgebruikers onder te verdelen zijn in een aantal subgroepen, waarbinnen de eindgebruikers een aantal gemeenschappelijke karakteristieken hebben. Deze subgroepen, vaak kunnen die geïdentificeerd worden door dezelfde beroepen of functienamen, zullen het product op verschillende manieren gebruiken en voor verschillende doeleinden. Bij het samenstellen van een steekproef voor de test is het dan zaak dat je uit elke subgroep een aantal mensen opneemt.

In het opdelen in subgroepen kun je erg ver gaan. Elke subgroep erbij betekent wel dat er meer personen geobserveerd moeten worden. Dit aantal neemt exponentieel toe met het aantal kenmerken waarin je onderscheid maakt en het aantal niveaus binnen een kenmerk. Een voorbeeld.

Voorbeeld 2

Stel je gaat een systeem testen dat gebruikt wordt bij een bank. Dit systeem wordt gebruikt door zowel baliemedewerkers als kantoormedewerkers. Aangezien de werkzaamheden van deze twee categorie medewerkers verschillend zijn en ze het product op verschillende manieren gebruiken, wil je medewerkers van beide groepen opnemen in je test. Op dit moment is er dus sprake van 2 subgroepen (zie figuur 3.1).

Figuur 3.1 Indeling in subgroepen enkel gebaseerd op functie (het getal tussen haakjes geeft het subgroepnummer aan).

Baliemedewerker	Kantoormedewerker
(1)	(2)

Je hebt verder het vermoeden dat de mate van ervaring in het omgaan met computers van invloed is op het gemak waarmee met het systeem gewerkt kan worden. De mate van ervaring kun je bijvoorbeeld opdelen in 2 niveaus: weinig ervaring en veel ervaring (je hebt dan al 4 subgroepen, zie figuur 3.2). Een indeling in 3 of meer niveaus is natuurlijk ook mogelijk (dan heb je al 6 of nog meer subgroepen).

Figuur 3.2 Indeling in subgroepen gebaseerd op functie en computerervaring

Categorie	Baliemedewerker	Kantoormedewerker
Weinig ervaring*	(1)	(2)
Veel ervaring*	(3)	(4)

- * Het is wel zaak dat je heel duidelijk aangeeft wat je verstaat onder 'weinig ervaring' en 'veel ervaring'. Deze termen worden vaak zo kritiekloos gebruikt, dat iedereen die te maken heeft met een product hier een verschillend idee bij heeft. Je moet een dergelijke vraag ook niet stellen aan gebruikers. De een is erg bescheiden en de ander juist niet. Maak gebruik van een objectieve meting: iemand met weinig ervaring is bijvoorbeeld iemand die minder dan 6 maanden met een bepaald product werkt. Daar kun je uiteraard wel naar vragen. Dat is een eenduidige vraag.

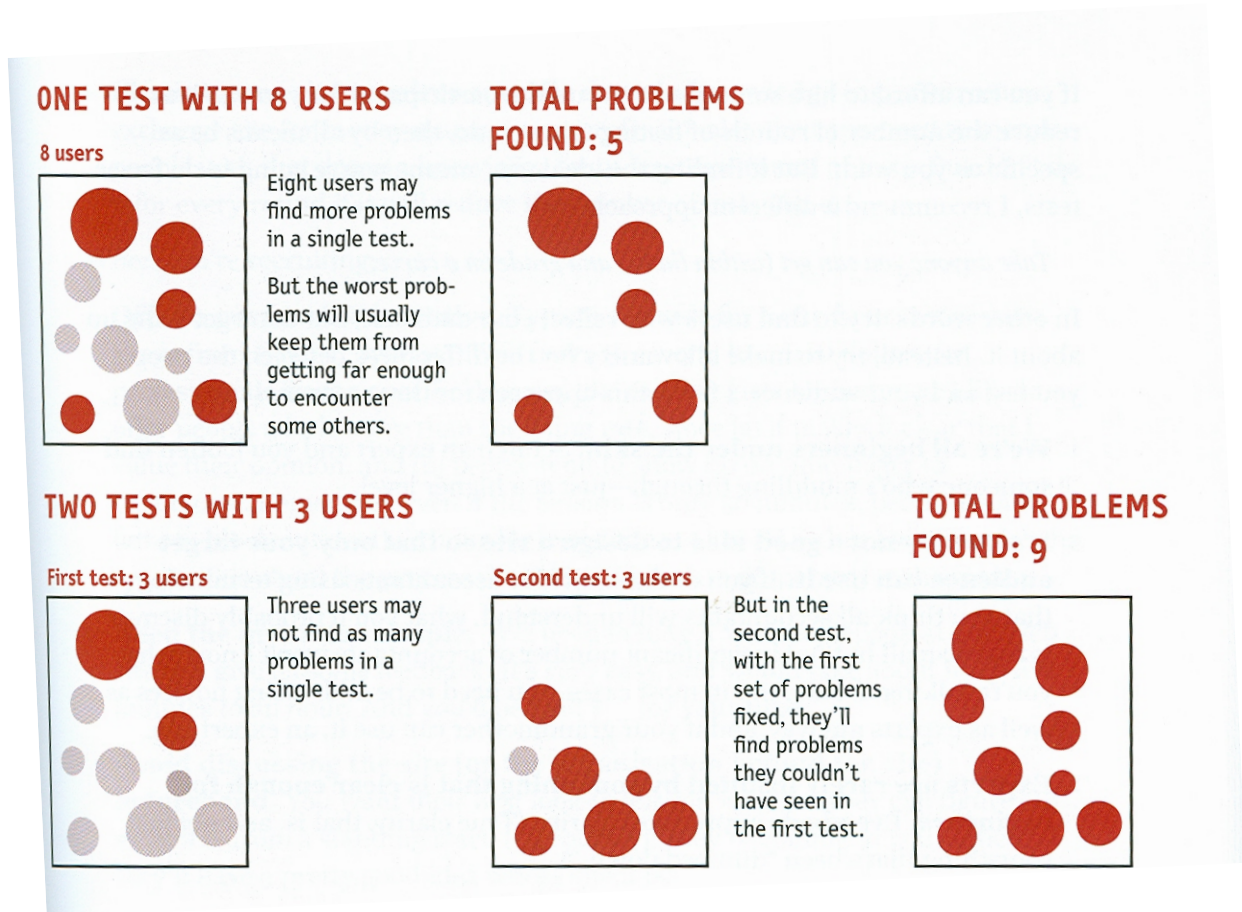
Nu eenmaal duidelijk is wat de doelpopulatie is en of er wel of geen onderscheid gemaakt moet worden in subgroepen is de volgende vraag: hoeveel gebruikers moet je mee laten doen aan de Usability Test?

Steekproefomvang

Dit aantal hangt van verschillende factoren af, waaronder de volgende:

- De mate van betrouwbaarheid van de resultaten die je wilt bereiken.
- De aanwezigheid van het type deelnemer dat je wenst.
- De duur van een testsessie.
- De tijd die je beschikbaar hebt om de test voor te bereiden.

Als je statistisch bruikbare resultaten wilt hebben, dan zul je voldoende deelnemers moeten hebben om bruikbare analyses er op los te kunnen laten en de resultaten te kunnen generaliseren naar de specifieke doelpopulatie. Wil je echter enkel proberen om zo veel mogelijk gebruiksproblemen boven tafel te krijgen in de kortst mogelijk tijd, dan heb je aan minder deelnemers genoeg. Een Usability Test is namelijk geen marktonderzoek. Het gaat er bij een Usability Test om, om de meest serieuze problemen die gebruikers met het product hebben te achterhalen. Uit verschillende onderzoeken is gebleken dat bijna 50% van alle grote usability problemen gevonden werden met 3 deelnemers aan de test, 80% van deze problemen met 4-5 deelnemers en 90% met 10 deelnemers (zie artikel 'Why You Only Need to Test With 5 Users' van Jakob Nielsen in de bijlage). Toename van het aantal deelnemers levert dus steeds minder kans op (extra) nieuwe informatie. Om meer problemen te vinden, kun je beter meerdere keren in het ontwikkeltraject een Usability Test uitvoeren met minder mensen per test. Zie hiervoor onderstaande tekening (overgenomen uit 'Don't make me Think').



De hierboven genoemde aantallen zijn wel steeds aantallen per subgroep. Hoe meer subgroepen je maakt, hoe meer deelnemers je dus nodig hebt. Uit bovenstaande volgt ook dat je het best zo rond de 4-5 deelnemers in elke subgroep moet hebben om ervoor te zorgen dat je de problemen die er zijn ook te zien krijgt. Drie deelnemers is echt het minimum.

Dit betekent dat bij een beperkt aantal deelnemers je de indeling in subgroepen zult moeten beperken. Je moet dan goed afwegen welke indeling in subgroepen van zwaarwegender belang zijn dan andere indelingen. Het klakkeloos bij elkaar voegen van subgroepen of in het geheel geen indeling meer maken is wel erg kort door de bocht.

§3.3 Kiezen van taken

In hoofdstuk 1, bij de kenmerken van een Usability Test, werd al opgemerkt dat het om realistische taken moet gaan. Uit de gemaakte taakanalyse in een eerdere fase zijn alle taken te vinden die de gebruikers uitvoeren met het product. Dit zijn vermoedelijk te veel taken om allemaal op te nemen in de Usability Test (anders zou de test te veel tijd en geld kosten). Hoe bepaal je nu welke taken je opneemt in de Usability Test? Hiervoor verwijzen we naar 2 stukken overgenomen uit respectievelijk 'Usability Testing and Research' van Carol M. Braun (blz 160-161) en 'A Practical Guide to Usability Testing' van J.S. Dumas en J.C. Redish (blz 160 t/m 163).

Beide stukken geven een goed idee hoe het best een deel van de veelheid aan taken gekozen kan worden. Belangrijk zijn in ieder geval dat je taken neemt die horen bij de eerder gestelde doelstellingen (zie §3.1) én dat je (ook) taken uitzoekt waarbij je usability problemen verwacht.

§3.3.1 Gedeelte uit 'Usability Testing and Research'

Selecting the Tasks to Test

The list of tasks you want to test typically far exceeds the time and budget you have for testing. So, the choice of tasks to test should be based on your operationalized goals (as discussed earlier in this chapter) and the needs of the users you have identified. Other considerations that might affect the tasks are the amount of time you have for testing, coupled with the number of participants you can use in the test. If, for instance, you want to test two subgroups of three users each and you have two days for testing, you may decide that you can test three users per day for two hours per user, total time, including pre-test and post-test questionnaires. If you have more time, you can test more users or test the same number of users performing more tasks. We have found that users tire after an hour of testing, so we design our tests typically to last one hour with a total time for each participant of one and one half hours.

With a general understanding of the amount of time you want to allow for each user, you can establish priorities for the tasks you want each user to do. These tasks may be the same for all users, or you may want to begin with the same tasks for all users and then add more advanced tasks for the more proficient users. To maximize the amount of time you have with each user, it's a good idea to prepare some optional tasks for any users who complete the tasks in less time than anticipated.

The biggest mistake some people make in creating the task list is to try to learn the answer to every question anyone has about the product. Instead, you should simplify the tasks to keep the test focused on specific goals. Because a usability test is typically exploratory, you can't expect to "validate" all phases or features of the product. As well, you want to avoid the "might as well" syndrome, which goes something like this: "While you're at it, you might as well ask users to. . ."

Tasks are frequently determined by using the following criteria:

- First impressions (look and feel of the product)
- First tasks (so important in fixing in users' minds whether they will consider the product easy or difficult to use)
- Tasks most frequently performed
- Critical tasks (even if performed less frequently)
- Specific problem areas (typically identified by sponsor or heuristic evaluation)
- New tasks added to a product or changes made from an earlier version of the product (including changes made after an earlier usability test of a prototype)

Organizing Tasks

Once you've created a list of the tasks you want to test, you need to arrange them in some order. The best way to begin testing is with a short and simple task for the following reasons:

- You want to make the user feel at ease by beginning with a simple task.
- Because many tests ask the user to think out loud, the test administrator will be able to remind the user to think out loud after the completion of the first task.
- A short first task provides the team with the opportunity to ask questions and also to make any adjustments needed to the equipment or in response to the user's needs.

§3.3.2 Gedeelte uit 'A Practical Guide to Usability Testing'

Selecting Tasks

Usability testing is a sampling process. You cannot test every possible task users can do with a product. What tasks, then, should you sample?

- Tasks that probe potential usability problems
- Tasks suggested from your concerns and experience
- Tasks derived from other criteria
- Tasks that users will do with the product

Tasks That Probe Potential Usability Problems

The first and most important criterion for selecting tasks is to use tasks that probe the potential usability problems with the product.

As with any testing procedure, the more problems you find in the limited time available, the more successful your test will be (Myers, 1979). The designers of the product may be surprised at this goal. They may view usability testing as a way to verify the usability of the product, that is, how easy it is to learn and use. You may need to remind them that the goal of quality assurance testing of software is to find the "bugs" before the product is released. The quality assurance process *assumes* that any complex software program has bugs in it. A good software testing procedure finds more bugs than a poor one. The goal of a usability test is similar: develop a procedure that will find the serious usability problems. Consequently, you look for tasks that will probe areas of potential usability problems.

Tasks Suggested From Concerns and Experience

Another source to use to identify tasks is the list of usability concerns you develop with designers. We have discussed this list in "Defining Your Goals and Objectives." The people who develop a product always have some ideas about where there are potential problems. They know what parts of the product were difficult to design and where they disagreed about the best approach.

Designers are often correct in knowing the kinds of *major* problems the test should probe. For example, if they made a major decision to select an interaction style such as touch, they may want you to look at the types of problems that users frequently have with touch-sensitive devices, such as touching outside of the touch pad.

While designers have useful ideas about where users will have problems, they often ignore the more basic problems users might have with the product. They may anticipate that users will have difficulty configuring software, but not that users will have difficulty starting a system and putting in a diskette.

There is a famous videotape made during a usability test at Digital Equipment Corporation. The tape shows how two naive users do not understand what a floppy diskette is and how to insert it into a disk drive. The designers had not anticipated that users would not know how to work with diskettes. This finding made it possible for designers to improve the design of the hardware so that users who have no experience with diskettes are less likely to have problems starting to install the product.

Besides probing problems that designers identify and the basic problems users often have with complex interfaces, there may be other potential problems. Your experience as a tester is invaluable here. In a very short time, you will see the common problems in the products you test. For example, if you know that users frequently have problems manipulating windows in applications that sit within a multitasking user interface, you will want to include some tasks that require participants to open, close, move, and resize windows.

Your list of potential problems may grow longer than you could possibly test. Do not worry about the length of the list at this point. You will shorten it later.

Once you have at least a preliminary list of problems, you need to begin to create tasks that will probe each problem.

Tasks Derived From Other Criteria

In addition to selecting tasks because they relate to usability problems and concerns and goals, you can use other criteria for selecting tasks for a usability test. Good choices are tasks that are difficult to recover from if done wrong.

Tasks That Users Will Do With the Product

All of the tasks you have selected thus far will be tasks that users will do with the product you are testing. There are other tasks that users can do with a product over and above those that relate to usability problems, concerns, or goals. For example:

- new or modified
- critical to the operation of the product
- frequently done
- done under pressure

If the development team has conducted a task analysis, you can extract the tasks from it. Otherwise, it is best to have a meeting with the developers, and with users as well, to create a list of the tasks.

Eliminating Tasks from the list

As you look for ways to eliminate tasks so that participants can complete the test on time, consider:

- The objective of each task and whether you can achieve more than one objective with it. For example, an objective such as having participants navigate three levels down a menu hierarchy can usually be combined with an objective for performing some function that is on the third level of the hierarchy.
- Whether a task that uses expensive resources is worth testing. For example, if you need an extra signal generator to simulate a complex signal to an oscilloscope, consider if the objective of that task is critical to the evaluation of the product
- Whether time-consuming tasks are more important than the two or three tasks you could include instead

Putting the Tasks in Order

The final step in creating the list of tasks is putting them in the order you will have the participants attempt them. There are two important points to consider here:

1. The tasks should flow in the natural order in which users will do them. For example, users will create a mail message before they edit and send it.
2. Tasks that are important to the evaluation of the usability of the product should come early in the test rather than near the end. It is likely that some participants will not finish all the tasks. So, if you leave important tasks until late in the test, you will collect less data on them.

§3.4 Opstellen scenario's

Wat is een scenario bij Usability Testing?

De ervaring heeft geleerd dat niet altijd duidelijk is wat onder een scenario wordt verstaan. De term scenario kom je ook in andere contexten tegen en heeft dan een heel andere betekenis. Een andere term die soms ook wordt gebruikt bij Usability Testing voor hetzelfde is 'schrijven van scripts'.

Een **scenario** is een beschrijving van een specifieke taak die je de deelnemers aan de test wilt laten doen. Scenario's beschrijven de taak op zo'n manier dat het kunstmatige karakter van de test wat wordt weggehaald.

Elke taak die je gekozen hebt (zie vorige paragraaf) moet worden vertaald in een scenario.

Hieronder volgen 2 voorbeelden van scenario's.

Voorbeeld 3

Je hebt net een telefoon met antwoordapparaat gekocht. De doos staat op de tafel. Haal het product eruit en zet het in elkaar zodat je telefoongesprekken kunt ontvangen en voeren.

Voorbeeld 4

Je kunt nummers in het geheugen van de telefoon opslaan en vervolgens deze nummers oproepen zonder dat je alle cijfers elke keer moet intoetsen. Het nummer van je beste vriend is 0212-4357658. Sla dit nummer op in het geheugen van de telefoon.

Wanneer is een scenario goed?

Een goed scenario is **kort**. Tijd is kostbaar. Je wilt niet dat de gebruikers onnodig veel tijd kwijt zijn met lezen. Bovendien lezen mensen niet allemaal even snel, dit beïnvloedt dan de data als er lange stukken tekst gelezen moeten worden.

Een goed scenario geeft de gebruiker **voldoende informatie om de taak te kunnen uitvoeren**. Om bepaalde taken te kunnen doen, hebben de gebruikers normaal gesproken zelf de informatie. Deze gegevens moet je hen nu geven. Als je het over laat aan de gebruiker om zelf iets te bedenken, zal dat bij sommige meer tijd kosten dan bij anderen. Dan kloppen de data dus niet meer.

Een goed scenario **gebruikt termen van de gebruiker en niet van het product**. Het hele punt bij Usability Testing is om te kijken of een gebruiker dit product zelfstandig kan gebruiken (zonder test-situatie) en wat er dan gebeurt. De gebruiker heeft dan alleen de beschikking over zijn eigen terminologie en zal die moeten vertalen naar die van het product. Als in een scenario de terminologie van het product wordt voorgezegd, dan mis je de test van die vertaalslag, waar het nu net omdraait.

Een goed scenario is **ondubbelzinnig**. Het moet duidelijk zijn wat je verwacht van de gebruiker, zodat daar geen misverstand over bestaat. Ook moet duidelijk zijn wanneer de taak afgelopen is. Bij de opdracht 'schrijf een antwoord' is het niet duidelijk of alleen het antwoord geschreven moet worden of dat het ook verstuurd moet worden.

Valkuilen bij het omzetten van taken in scenario's

Het maken van een lijst met taken is niet zo ingewikkeld. Deze vertalen in goede scenario's is complexer. Hiervoor zijn al 4 punten aangegeven waar een goed scenario aan moet voldoen. Ik wil jullie nog eens extra wijzen op 2 belangrijke valkuilen:

1. Het schrijven van een zogenaamde 'knoppencursus'.

Concentreer je op het *proces* en niet op de verschillende stappen die genomen moeten worden. Je wilt weten of de gebruiker het product zelfstandig kan gebruiken. Dit doe je niet door voor te zeggen op welke knoppen wanneer gedrukt moet worden.

2. Het gebruiken van producttermen.

De reden hiervoor is hierboven al aangegeven.

Deze 2 valkuilen lijken misschien open deuren. In de opdrachten volgen een aantal voorbeelden van eerdere uitvoeringen van een Usability Test door groepen studenten. Wat is er niet goed aan?

§3.5 Opstellen vragenlijsten

Voor de afname van de test wil je graag wat meer van de gebruiker weten (bijvoorbeeld ervaring met computers). Ook tijdens de test zijn er momenten dat je graag iets wilt vragen aan de deelnemers over uiteenlopende onderwerpen. Meestal zal dat zijn hoe ze iets ervaren of waarom ze iets op een bepaalde manier doen: dus **kwalitatieve** metingen. Het is zaak dat je de vragen die je wilt stellen op schrift zet. Dit vanwege 2 redenen: elke deelnemer stel je exact dezelfde vraag én je vergeet niet om de vraag te stellen. Hieronder volgen wat voorbeelden. Het gaat hierbij slechts om voorbeelden om je wat ideeën te geven waar je aan kunt denken. Dit wil absoluut niet zeggen dat al deze zaken bij elke Usability Test gemeten moeten worden. Dit hangt o.a. af van het product, de gekozen taak én de doelstellingen.

Voorbeelden van kwalitatieve metingen

- Waardering over het gemakkelijk aan te leren zijn van het product
- Waardering over het gebruiksgemak
- Waardering over de moeilijkheidsgraad van een specifieke taak
- Waardering over bruikbaarheid van de helpfunctie
- Waardering over het gemakkelijk kunnen vinden van informatie in de handleiding

- Waarom ze een taak op een bepaalde manier doen
- Oordeel over een soortgelijk product van een concurrent
- Vragen naar hun voorkeuren van werken en waarom
- Of ze dit product willen aanschaffen? (en waarom)
- Tegen welke prijs?

-

Gedurende de Usability Test zijn er drie momenten waarop je de deelnemers vragen zou *kunnen* willen stellen: voordat begonnen wordt met de taken, na elke/een taak en aan het eind van alle taken.

Er zijn dus drie mogelijke vragenlijsten:

- Pretest de vragen van deze lijst zijn bedoeld om informatie te krijgen over de achtergrond van de deelnemer (afgenomen voordat begonnen wordt met de taken).
- Posttask de vragen van deze lijst zijn bedoeld om waarderingen en oordelen te krijgen over een bepaalde taak (afgenomen na beëindigen van de taak, *kunnen* per taak verschillen).
- Posttest de vragen van deze lijst zijn bedoeld waarderingen en oordelen te krijgen over de gehele test en/of het product (afgenomen na beëindigen van alle taken).

Het is dus niet noodzakelijk om bij elke Usability Test altijd deze drie vragenlijsten op te nemen. De keuze hiervan hangt af van de diversiteit in de doelpopulatie (moet je een indeling in subgroepen maken?), de soort taken en product en de gestelde doelstellingen. Aan elke test zitten voor- en nadelen. Je moet deze keuze wel onderbouwen. Op de volgende bladzijden volgt een gedeelte uit het boek 'A Practical Guide to Usability Testing'. In dit gedeelte worden de drie soorten vragenlijsten uitgebreider besproken: wat is de meerwaarde en welke vragen neem ik wel en niet op.

We gaan hier verder niet in op het stellen van goede interviewvragen en bijbehorende valkuilen. Dat komt aan de orde bij vaardigheden. Het belangrijkste punt dat je moet onthouden bij het opstellen van vragen voor de posttask en de posttest is, dat deze vragen neutraal moeten zijn. Voorkom dat je gewenste antwoorden gaat krijgen. Ook zijn meerkeuze vragen, voornamelijk bij vragen naar een waardering, te verkiezen boven open vragen. Zorg ook voor consistentie in de antwoorden. Dus niet bij de ene vraag de antwoorden laten lopen van 'zeer oneens' tot 'zeer eens' en bij de volgende vraag net andersom. Aan het eind van de paragraaf (blz. 23) staan een paar voorbeelden van vragen.

Gedeelte uit het boek 'A Practical Guide to Usability Testing'.

Pretest Questionnaire

The purpose of a pretest questionnaire is usually to (a) gather background information to help you interpret the data from the test, and (b) to verify the qualifications of the test participant in cases when you have not already done so. Even when you have qualified the participants by having them complete a questionnaire during recruitment, there may be additional information you'll need to know.

For example, if you are testing the user interface to an electronic mail application, you may want to know how much experience the participant has using microcomputer software. People who use microcomputers frequently are familiar with common function key mappings, such as pressing the ESC key to move back to the previous screen. Conversely, people with little microcomputer experience may not know these mappings and, therefore, perform more poorly using electronic mail. When you conduct your analysis, you may, therefore, find wide differences in performance that are related to microcomputer experience, and these differences may highlight weaknesses in the interface for inexperienced microcomputer users. Therefore, you should ask about microcomputer experience on a pretest questionnaire for a microcomputer product.

Every question you include should have a purpose. Ask yourself, "What will I do with the information I get from this question?" and "How will I use this information in the test report?" If the question will help you diagnose problems with the product, it is a candidate to be in the questionnaire. If not, eliminate it. Most of our pretest questionnaires are about one page or less. What follows are examples of the types of questions we most frequently ask. We have not listed the alternative answer options as they vary with each test.

- What is your job title?
- How would you describe yourself? ("programmer," "computer operator," etc.)
- How long have you been doing this kind of work?
- How long have you worked with this product?
- How long have you been using personal computers?
- How often do you use a personal computer?
- Which of the following software products have you used?
- What features of - do you use most often?

Posttask Questionnaire

In usability tests that try to simulate having the user working alone with the product, you will normally wait until participants have completed all of the tasks or reached the time limit before you interview them. However, there are tests in which you want to get a reaction immediately after a particular task or a scenario, or after every task or scenario was completed.

- You might want to obtain an immediate reaction to the participant's experience during an important task. For example, you may have a task you have designed to see if users can find information in the manual or to see if they can navigate through an online help system. In these cases, you can get an immediate rating by the participant of how easy it was to use the manual or help system by including a posttask questionnaire that participants fill out as soon as they have completed the task.
- You might want to measure changes in perceptions over time. For example, different types of software often have major barriers to usability that appear at different levels of experience with them. Barriers to using communications software often occur early in use, while barriers to using a database management system may not occur until the user is faced with understanding concepts such as what a "view" of a database is. You can track users' changing perceptions of ease of learning by asking for ratings *during* the test.

In almost all cases, you will also want to cover these issues with the posttest questionnaire and interview.

Consequently, when you have a posttask questionnaire, make it short, usually one page or five to six questions at most, with a few questions and room for the participant to make a comment. Having a

short set of questions can actually be an advantage when the test team is new to testing. They may need a minute or two after each task to finish writing notes or typing into the data log. When you have the participant wait until these activities are completed, there is an awkward "dead" time for the participant. The posttask questions give the participant something to do during this time.

Posttest Questionnaire

After participants have completed the tasks, you have one final opportunity to gather data. After spending time using the product, participants have had an opportunity to gain some perspective about their impressions of its usability. The posttest questionnaire provides you with a vehicle to gather those impressions.

In our tests, we begin the debriefing by clarifying any ambiguities that may have occurred during the test. Frequently, events happen at a fast pace, and you need to clarify something that the participant did or said before you both forget about it. Then you present the posttest questionnaire. The debriefer should leave the room while the participant works on the questions. Participants pay more attention to the questions when they are left alone. After several hours of work both the test team and the participant are anxious to finish. It is almost impossible for the debriefer to avoid squirming and subtly hurrying the participant when the debriefer stays with participants while they fill out the questionnaire. After the participant is finished with the questionnaire, the debriefer goes over it with the participant to clarify the questionnaire responses.

In most posttest questionnaires there are usually two types of questions:

1. General questions that could apply to any product
2. Specific questions that apply only to the tested product

General Questions.

In our posttest questionnaire, we ask participants to answer questions such as:

- How do you rate the overall ease of use or difficulty of the product?
- How easy or difficult was it to find information in the manual?
- What do you like least about the product?
- What do you like best about the product?
- What one thing would you tell the designers to change?
- Would you use this product if your company bought it?
- Would you recommend that your company buy this product?

As with all questionnaires, structure as many as you can into formats such as ratings or YES/NO questions.

Specific Questions

In many cases, you will have usability concerns about the design of the product you are testing. When you know that you will want to focus on these concerns during the debriefing, put as many as you can in the form of structured questions. Even when you know that you will be covering a topic during the debriefing, put structured questions about it in the questionnaire. It is much easier to tabulate the answers to questions than it is to piece together the debriefer's notes from all the test participants.

Voorbeeld 5

Voorbeelden van posttask-vragen:

1. Het volbrengen van de taak was:

1	2	3	4	5
Zeer makkelijk	Makkelijk	Niet makkelijk of moeilijk	Moeilijk	Zeer moeilijk

2. Het gewenste menu-item vinden was:

1	2	3	4	5
Zeer makkelijk	Makkelijk	Niet makkelijk of moeilijk	Moeilijk	Zeer moeilijk

Voorbeeld 6

Voorbeelden van posttest-vragen:

1. Wat beviel u het meest aan het product?

.....
.....
.....

2. Noem één ding wat de ontwerpers volgens u moeten veranderen aan het product?

.....
.....
.....

3. Zou u mensen aanraden dit product te gebruiken?

Ja / Nee

Waarom.....
.....
.....

§3.6 Bepalen wat je gaat meten met bijbehorende norm

Bij een Usability Test verzamel je zowel **kwantitatieve** metingen als **kwalitatieve** metingen.

Bij **kwalitatieve** metingen moet je denken aan meningen van de gebruiker over allerlei verschillende aspecten waarnaar je vraagt (die stel je in de posttask en/of posttest zie vorige paragraaf), maar zeer zeker ook aan lichaamshouding zoals uitingen van frustratie, opluchting, etc en spontane opmerkingen (dit kun je bevorderen door de deelnemers te vragen hardop te denken).

Bij **kwantitatieve** metingen moet je denken aan de tijd die iemand nodig heeft voor het volbrengen van een taak, het aantal fouten dat iemand maakt, hoe vaak iemand dezelfde fout(en) maakt.

Hieronder volgen wat voorbeelden. Het gaat hierbij slechts om voorbeelden om je wat ideeën te geven waar je aan kunt denken. Dit wil absoluut niet zeggen dat al deze zaken bij elke Usability Test gemeten moeten worden. Dit hangt o.a. af van het product, de gekozen taak én de doelstellingen.

Voorbeelden van kwantitatieve metingen

- Tijd die nodig is om een taak te volbrengen
- Tijd besteedt aan het navigeren in menu's
- Tijd besteedt aan online help
- Tijd die nodig is om informatie te vinden in de gebruiksaanwijzing
- Tijd besteedt aan het lezen van de gebruiksaanwijzing
- Tijd die nodig was om fouten te herstellen

- Aantal verkeerde menukeuzes
- Aantal verkeerde keuzes in dialoogboxen
- Aantal verkeerde iconkeuzes
- Aantal verkeerde functietoets keuzes
- Aantal andere fouten
- Aantal herhaalde fouten

- Aantal schermen van online help waarnaar gekeken is
- Aantal herhaalde oproepen van hetzelfde helpscherm

-

Het is discutabel of het zinvol is om toetsaanslagen bij te houden. Dit is erg veel werk en levert eigenlijk weinig op. Zinvoller is het om foute keuzes bij te houden. Ook als het gebruik van bijvoorbeeld functietoetsen een essentieel onderdeel vormen van het gebruik van het product.

Voorbeelden van kwalitatieve metingen

- (zie hiervoor vorige paragraaf)

- Uitingen van frustratie (zuchten, handen in het haar.....).
- De moed opgeven.
- Enthousiasme....

- Spontane opmerkingen zoals: "Ik ben nu de draad helemaal kwijt"., "Dit was gemakkelijk", "Op dit moment zou ik normaal gesproken de technische dienst bellen.",

-

Wat ga je nu meten?

Aan de hand van de gestelde doelstellingen bepaal je welke kwantitatieve zaken je (per taak) gaat meten. Dit is voor de meeste taken hetzelfde, maar dat is niet per definitie zo. Je kunt dus gerust bij een aantal taken andere dingen meten.

Het hebben van een logging-programma is niet persé noodzakelijk. Onderstaand schema is ook heel goed bruikbaar. De gekozen items (zoals M, L etc) en het aantal zijn weer afhankelijk van je eigen keuze

Usability Test van							
Deelnemer:		Datum:			Naam notulist:		
M = fouten in menukeuze		A = Andere fouten			H = vraag helpdesk		
L =		O =			F =		
Taak	Tijd	M	L	A	O	H	F
Taak 1	<i>Start:</i>						
Korte omschr.	<i>Stop:</i>						
.....						

Het gaat bij het afnemen van een Usability Test echter niet alleen om bovengenoemde kwantitatieve gegevens. Dat is de ene kant van de medaille. Daarmee signaleer je dat er *ergens* problemen zijn. De bedoeling is dat er vervolgens verbeteringen aan het product plaatsvinden, om dit in de toekomst te voorkomen. Je moet dan wel weten *wat* het probleem is. Hier heb je kwalitatieve gegevens voor nodig. Die kun je o.a. achterhalen door de deelnemers tijdens het uitvoeren van de test goed te observeren en daar aantekeningen van te maken. Daarbij valt te denken aan:

- De dingen die ze zeggen als ze hardop denken.
- Spontane opmerkingen.
- De dingen die ze doen (of juist niet doen).
- Lichaamstaal en gezichtsuitdrukkingen.
- Geluiden die ze maken: zuchten, kreunen of zelfs schreeuwen, verbazing etc.

Hiervoor kun je een soort gelijk schema gebruiken als hiervoor:

Usability Test van		
Deelnemer:	Datum:	Naam notulist:
Taak	Opmerkingen gebruiker	Observaties
Taak 1		
Korte omschr.		
.....		

Het vaststellen van normen bij de kwantitatieve metingen

Het is belangrijk dat je voor elke taak afzonderlijk aangeeft wat je criteria zijn: wanneer is een tijd (aantal fout etc) nog acceptabel. Met andere woorden aan welke *norm* moet er voldaan zijn om tevreden te zijn? Vaak is dit te halen uit de doelstellingen.

Ook is het verstandig (in verband met de volgende taak en ter voorkoming van teveel frustratie) om voor elke taak een *maximale tijd* vast te stellen. Lukt het een deelnemer niet om de taak in de gestelde tijd te volbrengen, dan wordt er gestopt en verdergegaan met de volgende taak. Let op: de normen en maximale tijden verschillen per taak!

Hoofdstuk 4 Afname test en ordenen gegevens

In dit hoofdstuk werken we de tweede en derde fase van een Usability Test verder uit. Zoals al beschreven in hoofdstuk 1 omvatten deze fasen het uitvoeren van de test, dus het observeren van de gekozen gebruikers bij het uitvoeren van de opgelegde taken en hen interviewen, en het ordenen van de verzamelde gegevens.

§4.1 Afname van de test

In het vorige hoofdstuk is uitgebreid stil gestaan bij het opzetten van de Usability Test. Alle materialen zijn klaar, de deelnemers aan de test zijn uitgenodigd en de afspraken staan. De afname van de test kan beginnen.

Samenstelling testteam

Bij de afname van de test is het aan te bevelen om met meerdere personen de test bij te wonen. Aan de andere kant moet de groep ook weer niet te groot zijn. In dat geval voelt de deelnemer aan de test zich misschien erg overvallen. Er moet in ieder geval iemand zijn die de deelnemer instrueert en op zijn gemak stelt. Verder zal er iemand de tijd bij moeten houden en andere zaken die gemeten gaan worden. Ook is er iemand nodig die let op lichaamshouding en non-verbale signalen en dit ook noteert en ook gemaakte opmerkingen door de deelnemer opschrijft.

Begin van de test

Voordat begonnen wordt met het afnemen van een eventuele pretest of de test zelf, zal de deelnemer op zijn gemak moeten worden gesteld en hem/haar verteld moeten worden wat er gaat gebeuren en wat de bedoeling is van een Usability Test. Ook kunnen eventuele misverstanden over de aard van de Usability Test opgehelderd worden. Dit blijkt bijvoorbeeld uit opmerkingen die de deelnemer maakt.

Het feit dat de deelnemer opdrachten krijgt en er over de schouder meegekeken wordt, kan de deelnemer het gevoel geven dat hij getest wordt. Dit zal een druk op hem/haar leggen. Dit kan grotendeels ondervangen worden door uit te leggen dat de deelnemer niet getest wordt, maar dat het om een test van het product gaat. De deelnemer kan geen fouten maken. Als iets niet goed gaat, ligt dat aan het product.

De volgende punten komen over het algemeen aan bod:

- Toelichting op de ruimte en eventuele camera
- Introductie van de aanwezige personen met korte toelichting wat ze doen.
- Korte toelichting van het te evalueren product
- Uitleg op het gehele proces van evaluatie, inclusief scenario's en vragenlijsten en het niet geven van hulp (zie ook verderop)
Indien van toepassing vermelden hoe er moet worden gehandeld als de telefoon gaat o.i.d.
- Indien je dat wilt: Verzoek aan deelnemer om zoveel mogelijk hardop te denken. Hierdoor krijgt het testteam inzicht in wat er in iemands hoofd omgaat (zie ook verderop).
- De deelnemer eraan herinneren dat niet hij getest wordt, maar het product.

Hardop denken

Sommige mensen doen dit van nature, anderen zijn erg zwijgzaam, ook als het niet lukt. Vragen om hardop mee te denken, is voor hen dan onnatuurlijk. Een toelichting op het waarom en het nut kan dan helpen. Bijvoorbeeld:

“We willen graag weten wat u denkt dat er gaat gebeuren als u een keuze maakt en of wat er gebeurt, overeenkomt met wat u verwacht had. We willen graag weten wat u verrast, wat verhelderend is, wat u in de war maakt of frustreert e.d. en waarom dat zo is. We krijgen een beter begrip van hoe een en ander werkt voor u als u ons dit wilt vertellen terwijl u bezig bent.”

Er zijn mensen die vinden dat hardop denken ten koste van de tijd gaat die nodig is om de taak uit voeren. Onderzoeken hebben dit echter niet bevestigd.

Helpen tijdens de test?

Het kan gebeuren dat de deelnemer tijdens de test een vraag stelt of hij het goed deed bijvoorbeeld. Het is niet de bedoeling om daar antwoord op te geven, maar je kunt natuurlijk ook niet botweg zeggen dat je geen antwoord geeft. Een oplossing is om een wedervraag te stellen. Je krijgt zo meteen ook informatie over de gedachtengang van de deelnemer.

Hetzelfde geldt ook voor het geven van hulp. Je wilt weten hoe de gebruiker *zelfstandig* met het product kan omgaan. In een normale situatie zit er ook niet iemand bij die hem direct kan helpen. Een vraag om hulp is een duidelijk signaal dat er iets niet goed zit met de usability van het product.

Problemenlijst

Het is aan te bevelen om tijdens de afname van de testen een zogenaamde **problemenlijst** bij te houden. Op deze lijst verzamel je alle problemen die zich voordoen tijdens het uitvoeren van de testen (zoals vraag om hulp omdat men ergens niet uitkomt). In principe ook als een probleem zich slechts bij één deelnemer voordoet. Aangezien slechts een klein aantal gebruikers deelneemt aan de test, vertegenwoordigt één persoon toch een substantieel deel van de totale gebruikersgroep. Als er bijvoorbeeld vijf deelnemers aan de test zijn, dan vertegenwoordigt één deelnemer 20%.

§4.2 Ordenen en verwerken van de gegevens

Tijdens elke test-sessie worden er heel wat gegevens, zowel kwalitatieve als kwantitatieve, verzameld, namelijk: bij de pretest, posttask (eventueel bij alle scenario's!), posttest, de metingen tijdens de test, de waarnemingen tijdens de test en overige opmerkingen

Ordenen kwantitatieve gegevens en gegevens van ordinaal niveau

Tijdens de test worden er per taak een aantal zaken gemeten. In de verschillende vragenlijsten komen misschien ook een aantal vragen voor die je statistisch wilt verwerken. Deze vragen en metingen vormen elk een variabele (zie voor precieze definitie hoofdstuk 2). Voordat je deze gegevens kunt samenvatten (zie volgende paragrafen) en vervolgens analyseren, is het handig deze gegevens in bijvoorbeeld een Excel-werkblad in te voeren. De volgende punten zijn daarbij van belang:

1. Een korte, duidelijke afkorting voor de variabelen die je meet.
Als voorbeeld zou je kunnen denken aan de afkorting LFT voor de variabele leeftijd.
2. Het meetniveau van de variabele (zie hiervoor hoofdstuk 2).
3. Welke coderingen je gebruikt, bij ordinale variabelen, om de antwoorden te noteren. Het is namelijk sneller typen met codes zoals 1, 2, 3, 4, 5 in plaats van 'zeer makkelijk' etc.
4. Wat noteer je als een taak niet volbracht is in de beschikbare tijd? De maximale tijd is niet reëel. Dus niets invullen is de beste oplossing met later bij die taak een aantekening dat x personen die taak niet konden volbrengen.

Als dit gedaan is en de test is afgenomen, dan kunnen de resultaten van de test bij de verschillende deelnemers worden ingevoerd in Excel. Hieronder staat een mogelijke opbouw van een dergelijk werkblad.

Figuur 4.1 Mogelijke opbouw werkblad vóór verwerking

	A	B	C	D	E	F	G	H	I	J	K
1		var 1	var 2	var 3	var 4	var 5	var 6	var 7	var 8	var 9	var 10
2	deelnemer 1										
3	deelnemer 2										
4	deelnemer 3										
5	deelnemer 4										
6	deelnemer 5										
7	deelnemer 6										
8	NORM										

Het is aan te bevelen om de variabelen te groeperen per gemeten grootte (bijvoorbeeld tijd, aantal fouten) en dan per scenario. De reden hiervoor is dat het dan gemakkelijker is om de gegevens te kunnen samenvatten en te analyseren. Ook is het handig om ook direct de norm die je gesteld voor de betreffende grootte op te nemen. Verder is het aan te bevelen om de resultaten van de verschillende vragenlijsten en de testgegevens op verschillende tabbladen in hetzelfde werkblad te zetten.

Als je een indeling in subgroepen hebt gemaakt, kun je daar bij het invoeren van de gegevens rekening mee houden. Je bekijkt welke deelnemers bij elkaar horen en zet ze bij elkaar. Voor het berekenen van allerlei grootheden (zie hoofdstuk 5) kun je dan gemakkelijk onderscheid maken in de subgroepen. Dit kost even tijd: je moet zelf de objectieve criteria langs lopen. Als het goed is heb je die ook gemeten door middel van een of meerdere vragen. Het met de hand doen, kan fouten in de hand werken. Aangezien het toch gemeten is, komt de waarde terug in het werkblad. Wil je bij het berekenen van verschillende grootheden een uitsplitsing naar subgroep, kan dat gemakkelijk met behulp van formules gerealiseerd worden. Je moet dan een If-statement toevoegen voordat je iets gaat berekenen.

Verwerken kwalitatieve gegevens en gegevens van nominaal niveau

Dit is een moeilijker verhaal. Antwoorden verkregen op vragenlijsten en observaties per taak kunnen vaak ingedeeld worden in verschillende categorieën, zodat er een goed beeld ontstaat van deze zaken. Ze kunnen dan eventueel ook vastgelegd worden in het hiervoor genoemde werkblad. Zorg ervoor alle opmerkingen te ordenen en vast te leggen per scenario/taak.

Algemene opmerkingen kunnen ook gegroepeerd worden en voor de analyse later goede input leveren.

Soms kan het gebeuren dat je conflicterende gegevens hebt. Dan heb je een probleem. Je moet er dan achterzien te komen wat de oorzaak van dit verschil is en welke van de 2 waarschijnlijk de juiste is. Bekijk het volgende, veel voorkomende voorbeeld. Het testteam ziet dat een deelnemer veel moeite heeft met een taak of deze zelfs niet kan voltooien. De deelnemer geeft echter bij de posttask aan dat de taak 'makkelijk' was. Dit komt waarschijnlijk omdat de deelnemer graag positief wil zijn (het geven van gewenste antwoorden) en daarom een hogere waardering geeft dan feitelijk juist is. Bovendien neigen deelnemers ernaar om zichzelf de schuld te geven als iets niet goed ging. Kortom: geloof eerder wat je ziet dan wat deelnemers zeggen bij conflicterende gegevens.

Het noteren van positieve bevindingen

Er zijn 2 belangrijke redenen om positieve bevindingen te verzamelen en vast te leggen.

1. Iedereen hoort graag goed nieuws

Het doel van Usability Testing is natuurlijk om problemen boven tafel te krijgen. Dat neemt niet weg dat positieve opmerkingen genoemd mogen worden. Zeker bij het rapporteren later als blijkt dat een product grote veranderingen moet ondergaan. Dit is erg frustrerend voor de makers. In dat geval wil je ook positieve zaken kunnen vermelden. Het is een manier van feedback geven.

2. Als je positieve bevindingen niet vastlegt, dan kunnen ze worden veranderd.

Aangezien het product aangepast gaat worden op basis van de aanbevelingen van het testteam, bestaat het gevaar dat zaken die goed werkten voor de gebruikers veranderd worden tegelijk met de zaken die problemen veroorzaakten. Door vast te leggen wat goed is, is het duidelijk wat niet veranderd moet worden.

§4.3 Presenteren van de gegevens

Het is moeilijk om getalsmatige informatie op een effectieve manier duidelijk te maken in woorden. Naarmate de informatie complexer wordt en zaken meer met elkaar samenhangen wordt dit alleen maar lastiger. Een goed gebruik van tabellen en grafieken kan hierbij een enorme steun zijn. Niet voor niets wordt gezegd dat één plaatje meer kan vertellen dan duizend woorden. Maar dan moet het wel goed gebeuren.

Hoe functioneert een tabel of grafiek?

Het doel van het maken van tabellen is om de lezer een duidelijk overzicht te geven van het onderwerp van de tabel. Op een heldere en eerlijke manier moeten de gevonden uitkomsten worden gerapporteerd. Stijl en vormgeving richten de lezer op de kern van de zaak. Dingen die de aandacht kunnen afleiden worden vermeden.

Iets soortgelijks geldt voor grafieken. Wat moet een grafiek doen? Allereerst natuurlijk de gegevens laten zien, vaak veel getallen op een kleine ruimte. Vervolgens wil je in een grafiek samenhang creëren in de gegevens. En je wilt de lezer ertoe aanzetten om na te denken over de inhoud van de grafiek en niet over het ontwerp, de vorm, de franje en dergelijke. Ook wil je de lezer aansporen om vergelijkingen te maken, om overeenkomsten en verschillen te zoeken tussen diverse delen van gegevens.

Hoe kun je dit allemaal bereiken? Op deze vraag is geen pasklaar antwoord te geven. Wat we wel kunnen doen is bekijken welke kwaliteitscriteria je kunt gebruiken om een goede grafiek van een slechte grafiek te onderscheiden.

1. Nauwkeurigheid en aanzien

Een tabel of grafiek moet nauwkeurig worden gemaakt. Het mag niet zo zijn dat door slordigheid van de maker een misleidend of onvolledig beeld wordt geschapen.

Het aanzien van een tabel of grafiek bepaalt of zij belangstelling op zal wekken. Een professioneel uiterlijk en een harmonische, evenwichtige opbouw dragen daar toe bij.

2. Eenvoud en helderheid

Hoofddoel moet zijn het overbrengen van een stuk informatie, van een statistische boodschap. Niet terzake doende tekst en versiering moet je proberen te vermijden.

De gebruiker van de grafiek moet zonder al te veel moeite de juiste boodschap eruit kunnen halen. Een tabel of grafiek is bedoeld om tijd en inspanning te besparen en inzicht te geven. Het mag geen puzzel zijn en het is ook niet bedoeld ter decoratie.

3. Vormgeving

De grafiek of de tabel die je kiest moet in overeenstemming zijn met de boodschap die je brengt. Maak functioneel gebruik van grijsinten en kleuren. Verschillen in waargenomen contrast moeten een aanwijzing zijn voor verschillen in waarden. En zorg ervoor dat de visuele weergave van de data consistent is met de numerieke waarden. Optische illusies of vertekeningen moeten vermeden worden.

Valkuilen

Bij het maken van grafieken zijn er een aantal gevallen waarbij de plank volkomen wordt misgeslagen. Hieronder noemen we een aantal veelvoorkomende valkuilen

3D-grafieken

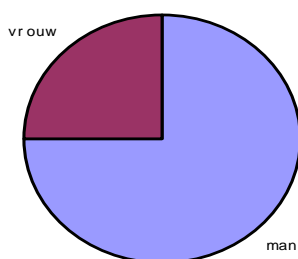
Dit is een van de gruwelijkste neveneffecten van de huidige computertechniek. Een 3D-grafiek ziet er vaak 'leuk' uit, maar dat is ook het enige. In 99% van de gevallen zijn de gegevens vertekend weergegeven en/of alleen met puzzelen af te lezen. Dit wordt veroorzaakt door het noodzakelijke perspectief in de grafiek voor het 3D-effect. Hiermee ondermijnt je het hierboven genoemde punt 2. Dus niet doen!

Een overdaad aan grafieken

Als je eenmaal ontdekt hebt hoe je makkelijk grafieken kunt maken, is vaak het gevolg dat van *alle* metingen een grafiek wordt gemaakt en vervolgens ook opgenomen in het rapport (al dan niet aan bijlage). Hiermee ondergraaf je het aantrekkelijke van een grafiek. Mensen worden overspoeld door de plaatjes en kijken er dan niet meer naar: het zal allemaal wel. Dus schiet je hiermee het doel van de getoonde grafiek voorbij. Je wilt met de grafiek een punt duidelijk maken, er de aandacht op vestigen. Je mag best overal grafieken van maken, maar zorg dat je alleen de belangrijkste opneemt in je rapport en zeker nooit grafieken in bijlagen stoppen. Als een grafiek niet interessant is voor het rapport zelf, dan is het opnemen in de bijlage onzin. Een tabel of de ruwe gegevens volstaan dan.

Cirkel- of staafdiagrammen bij 2 of 3 antwoordmogelijkheden

Bij nominale variabelen heb je niet veel keuze in grafiekmogelijkheden. Een optie is een cirkeldiagram. Het is echter onzin om een cirkeldiagram te maken als er maar 2 (of 3) antwoordmogelijkheden zijn. Kijk maar naar het volgende voorbeeld over de verdeling van mannen en vrouwen in een bepaalde gebruikersgroep. Links de uitbeelding daarvan in een grafiek en rechts in woorden:



In de gebruikersgroep is 75% man.

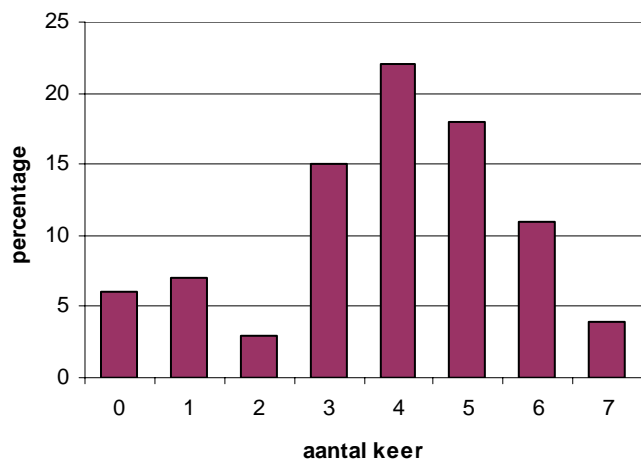
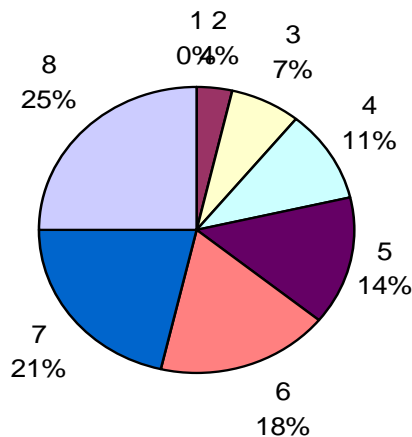
Het 2^e is net zo duidelijk en minder storend als het opnemen van de grafiek. Kies bij 2 (of 3) antwoordmogelijkheden dus voor het weergeven in tekst en niet in een grafiek.

Cirkeldiagrammen bij veel antwoordmogelijkheden

Een andere veel gemaakte fout bij cirkeldiagrammen is dat ze gebruikt worden als er veel antwoordmogelijkheden zijn. Het gevolg is dat je door de bomen het bos niet meer ziet en er percentages bij moet gaan vermelden om nog goed te zien wat vaker voorkomt e.d. Een alternatief (ook bij nominale variabelen) is een kolom- of staafdiagram.

Kijk naar onderstaand voorbeeld waar in beide grafieken hetzelfde wordt weergegeven:

De procentuele verdeling van studenten over het aantal keer vlees dat ze eten per week bij de avondmaaltijd.



Technische randvoorwaarden voor tabellen en grafieken

Tabellen en grafieken vertellen een verhaal. Maar dan moet dat verhaal wel compleet zijn. Daarom gelden de volgende randvoorwaarden waaraan elke goede tabel en grafiek moeten voldoen.

- Titel** boven de tabel of grafiek hoort een korte omschrijving te staan van het onderwerp dat wordt weergegeven. Wanneer in een wat groter verslag veel tabellen en grafieken gebruikt worden, is het zinvol om deze op een logische manier te nummeren. Dit maakt het verwijzen naar een grafiek een stuk makkelijker.
- Bron** als de gegevens afkomstig zijn van een instelling of uit een bepaald rapport, dan moet deze bron vermeld worden. Vaak gebeurt dit direct onder de tabel of grafiek. In een groter verslag kunnen alle bronvermeldingen ook aan het eind worden samengevoegd. Wanneer je gegevens van Internet afkomstig zijn, dan dien je niet alleen de URL te vermelden, maar ook het instituut of de persoon die daar verantwoordelijk voor is, en de datum waarop je de gegevens gekregen hebt (het Internet verandert namelijk voortdurend).
- Bijschrift** (dit geldt voor tabellen) boven in elke kolom en voorin elke regel van de tabel moet duidelijk vermeld worden om welk soort gegeven het gaat.
- Assen** (dit geldt voor grafieken) gebruik je in je grafiek assen, dan moet bij de assen vermeld worden welk soort verschijnsel langs die as is uitgezet. Op de assen hoort een duidelijke en regelmatige schaalverdeling te staan.
- Eenheden** er moet duidelijk vermeld zijn (in de bijschriften of langs de assen) in welke meet-eenheid de betreffende cijfers zijn gegeven. Als in een tabel voor alle kolommen of regels dezelfde eenheid geldt, dan kun je dat natuurlijk ook in de kop van de tabel vermelden.
- Legenda** als in een grafiek verschillende gegevensreeksen voorkomen, dan moeten die op verschillende wijzen worden aangegeven, bijvoorbeeld door wisselende lijntypen: ononderbroken, streepjes, stippels en dergelijke. De betekenis hiervan moet duidelijk in de grafiek of in een legenda worden aangegeven. Het gebruik van een legenda is aan te bevelen als het storend is om bij de lijnen zelf tekst te zetten.

§4.4 Enkele maatstaven

We bespreken in deze paragraaf een aantal locatiemaatstaven en een spreidingsmaatstaf. Bij een **locatiemaatstaf** gaat het om de plaats, de orde van grootte van de gegevens. Bij een **spreidingsmaatstaf** gaat het om de mate waarin de gegevens onderling verschillen. De locatiemaatstaven die besproken worden zijn: de mediaan en het rekenkundig gemiddelde. Van de spreidingsmaatstaven bespreken we alleen de variatiebreedte. Er zijn meer spreidingsmaatstaven. Deze bespreken we echter niet, omdat we bij Usability Testing te maken hebben met een beperkt aantal (hooguit 10, meestal ca. 5) gebruikers per taak. Het berekenen van de (vaak gebruikte) standaarddeviatie levert bij nadere beschouwing weinig meerwaarde op. Dezelfde conclusies t.a.v. verbeterpunten voor het te testen product kunnen gevonden worden bij het combineren en analyseren van enkele zinvolle grafieken en de bovengenoemde grootheden (mediaan, gemiddelde, variatiebreedte).

Voor het bepalen van de verschillende grootheden is het van belang om goed rekening te houden met het meetniveau van de gegevens. We hebben in hoofdstuk 2 gezien dat het meetniveau van de gegevens bepaalt welke berekeningen je mag uitvoeren met de gegevens. Zo is het rekenkundig gemiddelde berekenen van gegevens van een kwalitatief nominale variabele (bijvoorbeeld 'geslacht') klinkklare onzin.

Het is niet de bedoeling dat jullie de maatstaven met de hand berekenen. Daarvoor kun je gebruik maken van Excel.

Mediaan

Voor het berekenen van de *mediaan* is het nodig dat de variabele tenminste van ordinaal niveau is. De gegevens moeten op volgorde gezet kunnen worden.

Definitie

Als de gegevens op volgorde van grootte gerangschikt zijn, dan is de mediaan de middelste waarde. Anders geformuleerd: de mediaan is de waarde waarvoor geldt dat hoogstens 50% van de gegevens kleiner is dan deze waarde en hoogstens 50% van de gegevens groter is dan deze waarde.

Deze definitie leidt er toe dat bij een oneven aantal waarden de mediaan inderdaad de middelste waarde is wanneer de gegevens naar grootte gerangschikt zijn. Bij een even aantal waarden is er sprake van de "middelste twee". Beide waarden voldoen aan de definitie van de mediaan. Het is echter gebruikelijk om in een dergelijk geval het gemiddelde van deze twee middenwaarden te definiëren als de mediaan.

Voorbeeld 1

Gegeven is de volgende tabel met het aantal nieuwbouwwoningen in Nijmegen.

Tabel 4.1 Aantal nieuwbouwwoningen in Nijmegen per jaar

Jaar	Aantal woningen
1988	746
1989	943
1990	792
1991	609
1992	644

Bron: Statistisch Jaarbeeld gemeente Nijmegen 1993

Om de mediaan hiervan te kunnen bepalen, zetten we de gegevens van tabel 3 op volgorde van klein naar groot. We krijgen dan de reeks: 609 644 746 792 943.

Aangezien het aantal gegevens 5 is, is de mediaan de middelste waarde. Dit is de derde waarde. De mediaan is dus 746.

Voorbeeld 2

Neem aan dat een banenbureau van een Hbo-opleiding een vragenlijst stuurt aan een steekproef van afgestudeerden waarin het informatie vraagt over de aanvangssalarissen. In tabel 4.2 staan de verzamelde gegevens.

Tabel 4.2 Aanvangsmaandsalarissen van een steekproef afgestudeerden bedrijfskunde

Afgestudeerde	Maandsalaris (\$)	Afgestudeerde	Maandsalaris (\$)
1	2350	7	2390
2	2450	8	2630
3	2550	9	2440
4	2380	10	2825
5	2255	11	2420
6	2210	12	2380

We zetten de gegevens van tabel 4 op volgorde van klein naar groot. We krijgen dan de reeks:

2210 2255 2350 2380 2380 2390 2420 2440 2450 2550 2630 2835

Er zijn twaalf gegevens. De middelste twee zijn 2390 en 2420. We kunnen nu zowel 2390 als 2420 aanwijzen als de mediaan. Er zijn namelijk 5 gegevens kleiner dan 2390 ofwel 42%, daarnaast zijn er 6 gegevens groter dan 2390 ofwel 50%. Beide percentages zijn kleiner dan of gelijk aan 50, dus de waarde 2390 voldoet aan de definitie van de mediaan. Voor de waarde 2420 zijn de percentages respectievelijk 50 en 42, ook deze zijn kleiner dan of gelijk aan 50, dus ook de waarde 2420 voldoet aan de definitie.

Echter, zoals hierboven al vermeld staat, is het gebruikelijk om als de mediaan het gemiddelde van 2390 en 2420 te nemen, dus 2405.

De mediaan is in feite die waarneming die als het ware alle waarnemingen in twee gelijke stukken verdeelt. Men zegt ook wel eens dat de helft *links* van de mediaan ligt en de andere helft *rechts*.

Het nadeel van de mediaan is, dat die alleen rekening houdt met het *aantal* waarnemingen en niet zo zeer met de grootte van al die waarnemingen. Wel de kwantiteit, maar niet de kwaliteit.

Rekenkundig Gemiddelde

Het *rekenkundig gemiddelde*, of kortweg het gemiddelde, van een variabele is de meest bekende locatiemaatstaf. Deze maatstaf is alleen maar gedefinieerd voor kwantitatieve variabelen.

Bij gegevens uit een steekproef wordt het gemiddelde, het zgn. **steekproefgemiddelde**, aangegeven met \bar{x} .

Definitie

Het rekenkundig gemiddelde van een groep gegevens wordt berekend door alle waarde van de gegevens op te tellen en te delen door het aantal elementen.

Voorbeeld 3

We bekijken weer de gegevens van het aantal nieuwbouwwoningen in Nijmegen van voorbeeld 1. Het gemiddelde aantal nieuwbouwwoningen in Nijmegen over deze vijf jaren is dan gelijk aan:

$$\bar{x} = \frac{(746 + 943 + 792 + 609 + 644)}{5} = 746,8 \approx 747 \text{ woningen}$$

Voorbeeld 4

We bekijken nu weer de gegevens van de aanvangssalarissen van voorbeeld 2. Het gemiddelde aanvangsmaandsalaris voor deze steekproef van 12 afgestudeerden is dan:

$$\bar{x} = \frac{2350 + 2450 + \dots + 2380}{12} = \frac{29.280}{12} = 2440 \text{ dollar.}$$

Opmerking

Veel boeken over Usability Testing besteden summier of geen aandacht aan het samenvatten en analyseren van de gegevens. In een aantal boeken waarin dat wel gebeurt, wordt dan vaak de fout gemaakt om bij een ordinale variabele (dus bijvoorbeeld bij een vraag naar een mening over iets) het gemiddelde te bepalen. Er wordt geredeneerd (vanwege de codering) dat het om een kwantitatieve variabele gaat. Dit is echter fundamenteel fout. Bij een ordinale schaal is de afstand tussen bijvoorbeeld 'zeer mee eens' en 'mee eens' niet even groot als de afstand tussen 'mee eens' en 'niet mee eens/oneens'. Zeker niet als je het gaat bekijken voor verschillende personen. Bij een ordinale variabele kan dus alleen de mediaan bepaald worden!

Variatiebreedte

De *variatiebreedte* is de eenvoudigste spreidingsmaatstaf voor een gegevensverzameling.

Definitie

De variatiebreedte van een groep gegevens is het verschil tussen de hoogste en de laagste waarde.

Voorbeeld 5

We bekijken nogmaals de aanvangssalarissen uit voorbeeld 2. Het hoogste aanvangssalaris is 2825 en het laagste 2210. De variatiebreedte is dus $2825 - 2210 = 615$ dollar.

Hoewel de variatiebreedte eenvoudig te berekenen is, wordt deze binnen de statistiek (bij grote steekproeven) maar zelden als spreidingsmaatstaf gebruikt. Dit komt omdat hij maar op twee waarden is gebaseerd en daardoor erg vertekend kan worden door een extreme hoge en/of lage waarde. De variatiebreedte is dus niet erg informatief. Hij geeft alleen aan hoe breed/groot het gebied is waarbinnen alle waarden liggen. Daarbij is de variatiebreedte ook moeilijk te interpreteren, omdat hij afhankelijk is van het aantal waarnemingen: hoe groter het aantal waarnemingen, des te groter is in het algemeen de variatiebreedte. Aangezien we bij Usability Testing te maken hebben met slechts een beperkt aantal waarnemingen (per taak) speelt dit laatste voor ons niet.

§4.5 Gebruik van Excel

In deze paragraaf bespreken we een paar mogelijkheden van Excel. Er is veel meer mogelijk!

Het (laten) berekenen van percentages/gemiddelde/mediaan e.d.

Voor de analyse van de gegevens is het handig om bijvoorbeeld aan te kunnen geven hoeveel procent van de deelnemers de norm heeft overschreden. Je kunt dit voor alle taken en grootheden natuurlijk met de hand uitrekenen...., maar je kunt er ook Excel voor gebruiken.

Voorbeeld 1

Stel je hebt bij een bepaalde taak de volgende tijden (in seconden) genoteerd:

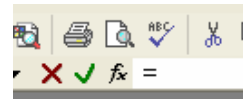
	A	B	C
1		var 1	var 2
2	deelnemer 1	100	
3	deelnemer 2	60	
4	deelnemer 3	75	
5	deelnemer 4	150	
6	deelnemer 5	140	
7	deelnemer 6	70	
8	NORM	120	
9			
10	perc > norm	0,33	

Je laat nu Excel tellen (de formule komt in cel B10) hoeveel cellen er boven de norm zitten. Hiervoor is er de functie: *Aantal.als*. Dit deel je door het totale aantal niet lege cellen. Dat doe je met de functie *Aantal*.

In cel B10 staat dan: =AANTAL.ALS(B2:B8;"> 120")/AANTAL(B2:B7)

Vervolgens kun je deze formule kopiëren (door het hoekje rechtsonder vast te houden en te slepen) naar alle andere cellen waar zoiets berekend moet worden.

Een functie maak je als volgt. Begin met het selecteren van de cel waarin je de functie wilt hebben staan. Klik op de knop (f_x) bovenin:

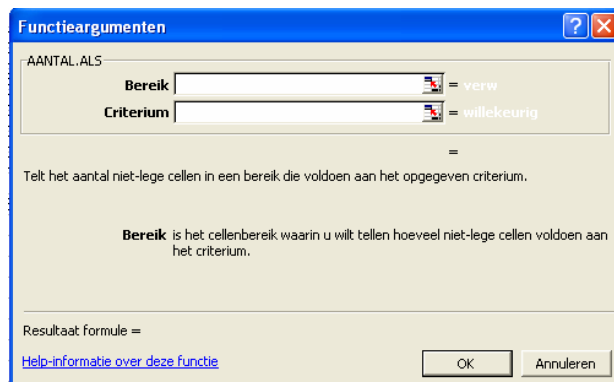


Je krijgt nu een overzicht van de categorieën functies waarover Excel beschikt.




Kies voor 'Alles' en daarbinnen voor 'Aantal.Als'. Druk dan op OK voor de volgende stap.

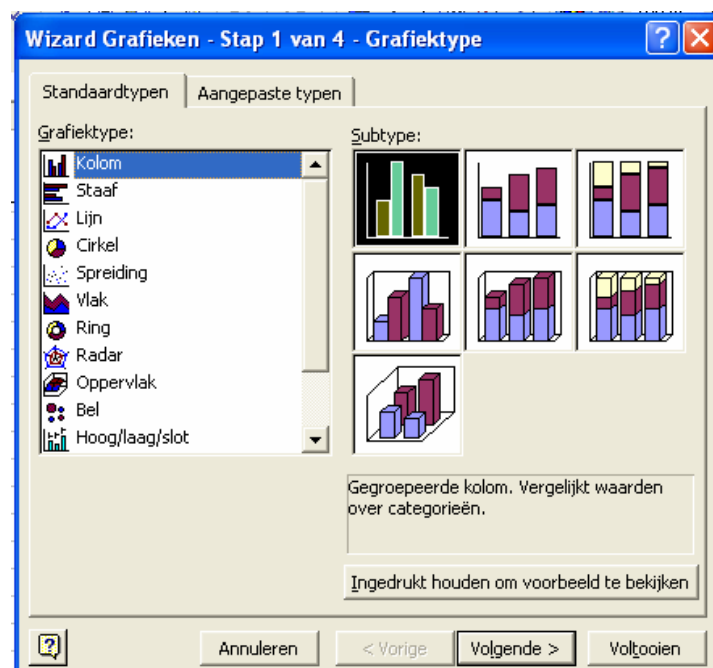
Nu moet je aangeven van welke getallen je het aantal wilt hebben. Je kunt het bereik bepalen via slepen met de muis of door het intypen (hier is het bereik B2:B7) in het aangegeven veld. Bij 'criterium' zet je de voorwaarde neer. Helaas werkt nu een celverwijzing niet. Door op OK te klikken, verschijnt het aantal in de cel die je in het begin hebt geselecteerd.



Op eenzelfde soort manier kun je gemiddelde en mediaan bepalen.

Het maken van grafieken

Voor het maken van grafieken maak je gebruik van de knop  Als je hierop klikt, verschijnt het volgende overzicht:



Niet alle grafieken zijn voor Usability Testing zinvol of bruikbaar. Zinnvolle grafieken (met aanpassingen) kunnen zijn:

- Kolomgrafieken
- Lijngrafieken (alleen de punten, de lijnen weghalen)
- Cirkeldiagrammen (bij percentages)

De bedoeling van deze reader is niet om een handleiding Excel te geven. Ga zelf wat aan het stoeien met de verschillende grafieksoorten. Ook in de les zal er wel eens wat voor gedaan worden, hoe je iets aanpakt en verder kun je hier altijd ook naar vragen.

Hoofdstuk 5 Analyseren van gegevens

Tijdens het uitvoeren van de Usability Testen heb je waarschijnlijk dingen gezien die je niet verwacht had (en dingen die je wel verwacht). Je handen jeuken om de problemen direct aan te gaan pakken. Het is echter zaak om niet te snel overhaaste conclusies te trekken over de aard van de problemen en hun oplossing. Na het uitvoeren van de Usability Test is het moment gekomen om wijs te worden uit wat je geobserveerd hebt, wat je hebt genoteerd en wat de gebruikers hebben geantwoord in de vragenlijsten. In dit hoofdstuk bespreken we hiervoor een manier.

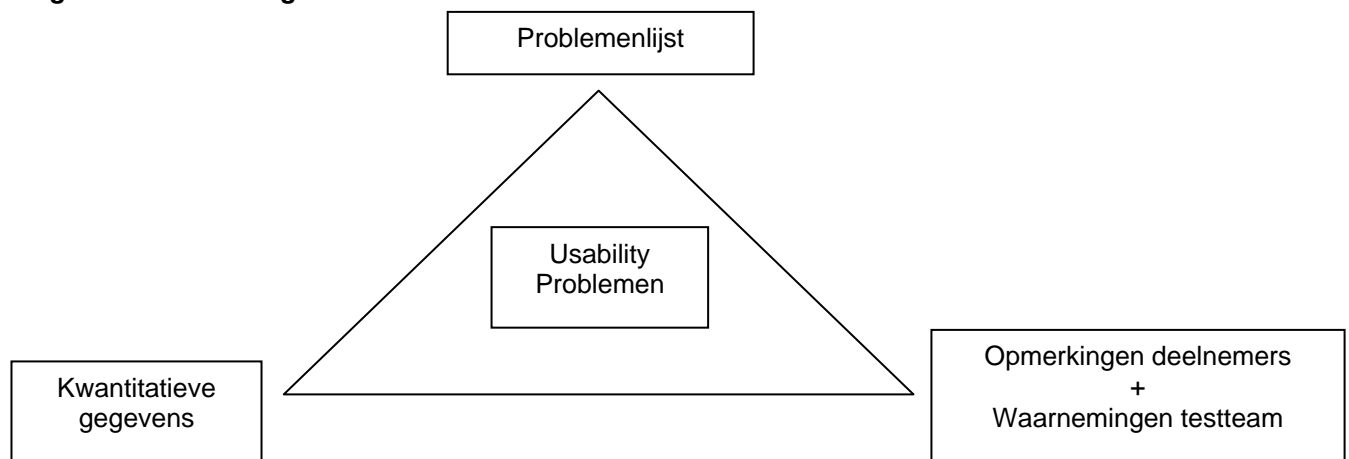
§5.1 Triangulatie

Na het uitvoeren van de Usability Test heb je een grote berg aan verschillende soorten gegevens van een klein aantal deelnemers gekregen. Nu is het moment om die te gaan analyseren. In de analyse bepaal je welke problemen er zijn en wat de oorzaak daarvan is en tot slot beveel je oplossingen aan voor de problemen.

Het signaleren van problemen is één ding. Moeilijker wordt het om de oorzaak van het probleem te achterhalen. Als bijvoorbeeld het probleem te maken heeft met de navigatie dank kan de oorzaak verschillende dingen zijn: slecht ontwerp van de pagina, inconsistentie, onbekende terminologie, een concept dat niet overeenkomt met het mentale model van de gebruiker etc.

Dit betekent dat je de verzamelde gegevens nader moet gaan bestuderen. Dit betekent dat je moet gaan kijken wat de gebruikers deden en wat ze zeiden, welke antwoorden ze gegeven hebben en wat je geobserveerd en genoteerd hebt. Deze manier van analyseren van zo'n grote hoeveelheid aan gegevens uit verschillende bronnen heet **triangulatie** (zie figuur 5.1): je kijkt naar alle gegevens samen om te zien hoe ze elkaar ondersteunen. Bijvoorbeeld: een lange benodigde tijd om taak te volbrengen, veel fouten, de opmerkingen van de deelnemers hierbij kunnen allemaal wijzen naar een en hetzelfde probleem.

Figuur 5.1 Triangulatie



Het komt er eigenlijk op neer dat je als volgt te werk gaat:

1. Analyseer/bekijk de kwantitatieve gegevens (zie §5.2). Waar zitten de problemen?
2. Analyseer/bekijk de kwalitatieve gegevens (zie §5.3). Waar zitten de problemen?
3. Bekijk de observaties en combineer dit met stap 1 en 2. Wat zijn de oorzaken van de problemen?
4. Wat is de omvang van elk probleem, hoe ernstig is het? Wat is de prioritering van de problemen en wat zijn de mogelijke oplossingen?

§5.2 Analyseren kwantitatieve gegevens

Als het goed is, heb je alle gegevens die je verkregen hebt uit de Usability Test in een Excel-werkblad gezet. Als alle gegevens ingevoerd zijn, kun je Excel er grafieken van laten maken en gemiddelde en zinvolle percentages laten berekenen. Kijk vervolgens kritisch naar de uitkomsten. Zie je trends of juist niet, veel overeenkomsten of juist grote verschillen tussen de gebruikers. Probeer te achterhalen waar die vandaan komen (triangulatie). Vermoedelijk zit daar een usability probleem achter, maar er kan ook een andere oorzaak zijn.

De analyse bestaat dus uit de volgende stappen.

1. Bepalen van minimum, maximum en gemiddelde van de verschillende grootheden. Bepaal percentages van de gebruikers die boven de norm zitten. Het gebruik van percentages is ook zinvol in indirecte metingen. Bijvoorbeeld: zoveel % van de deelnemers kon een taak niet foutloos uitvoeren of zoveel % van de deelnemers kon de taak niet binnen de maximum tijd uitvoeren. Dit kan soms ook belangrijke informatie geven.
2. Kijk kritisch naar de waarnemingen (per variabele/per taak). Hoe ver liggen die uit elkaar en hoe liggen de waarnemingen daar tussen in: vlak bij elkaar, aan de uiteinden, is er een uitschieter? Hierbij is een grafiek een handig hulpmiddel. Vergelijk de gegevens ook met de gestelde doelstellingen en criteria.

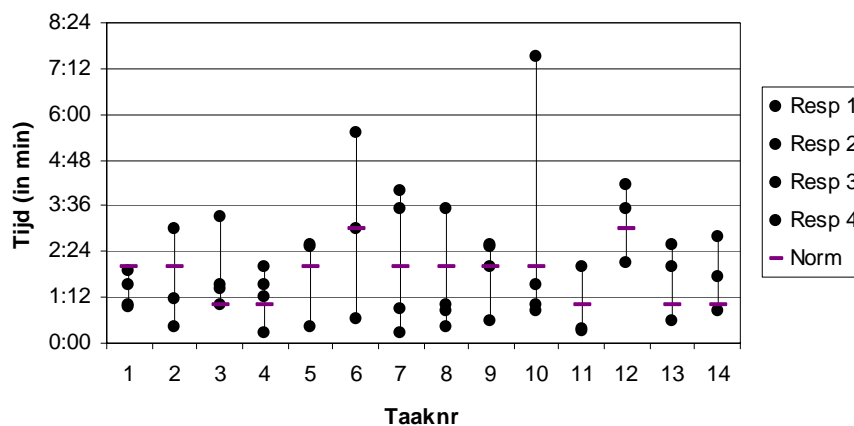
Stap 1

Spreekt voor zich, eventuele informatie is te vinden in het vorige hoofdstuk.

Stap 2

Voor het vervolg kan het handig zijn om overzichtsgrafieken te maken. Bijvoorbeeld een *puntgrafiek* met daarin de waargenomen tijden van alle taken met bijbehorende norm (NIET een staafdiagram). Het is extra duidelijk als ook de variatiebreedte wordt gevisualiseerd. Hieronder een voorbeeld

Figuur 5.2 Waargenomen tijden per taak met norm en variatiebreedte



Verder is bekend dat taak 2, 5, 6, 11 t/m 14 door 25% van de gebruikers niet is volbracht!

Als je deze grafiek bekijkt dan zie je bij bijvoorbeeld taak 10 dat 3 gebruikers onder de norm blijven en dat er één gebruiker extreem lang over doet. We hebben hier duidelijk te maken met een uitschieter. Wat hiermee te doen, bespreken we verderop. Iets dergelijks valt ook op bij taak 7. Andere opvallende zaken zijn, dat bij alle taken behalve taak 1 er één of meer gebruikers boven de norm zitten. Bij taak 3, 4 is dat zelfs 75%. Zo valt er nog veel meer te halen uit deze grafiek. Je signaleert hier dus duidelijk waar de problemen liggen.

Een andere soort grafiek is er een waar je de norm met het gemiddelde vergelijkt.

Wat te doen met uitschieters?

Het is zeer de moeite waard om (eventueel) gevonden uitschieters aan een nauwkeurig onderzoek te onderwerpen en proberen te achterhalen wat de oorzaak is. Soms is een uitschieter een gegeven dat niet juist is overgenomen. Als dat zo is, kan het gecorrigeerd worden voorafgaand aan de verdere analyse van de gegevens. Een andere mogelijkheid is dat er bijvoorbeeld een apparaat kapot ging (dat niets te maken had met de Usability Test zelf, maar wel nodig was) dat tijd kostte, die niet van de tijd nodig voor de taak is afgehaald. Ten slotte is het ook mogelijk dat een uitschieter niets anders is dan een ongebruikelijke waarde, die goed is weergegeven en dus thuishoort in de gegevensverzameling. In dat geval moet zij blijven staan. Er is dan een grote kans dat er een usability probleem is. Het is dan ook zaak deze uitschieters nader te bestuderen en terug te halen wat er precies tijdens de test gebeurde. Het verdient ook aanbeveling om niet alleen naar de desbetreffende taakuitvoering te kijken, maar ook of er andere gemeenschappelijke kenmerken zijn te vinden.

Hiervoor spreken we over het vinden van meerdere uitschieters bij een variabele. Wat nu als je maar één uitschieter vindt? Concluderen we dan: er was er maar één, dus die vergeten we maar? Bij grote steekproeven zoals bij een marktonderzoek, waar je honderden personen ondervraagt, kun je dat doen. Bij een Usability Test hebben we slechts per subgroep maar een beperkt aantal deelnemers. Één deelnemer vertegenwoordigt dan misschien wel 20% van de gebruikers. Het niet serieus nemen van zijn/haar probleem, kan dan grote gevolgen hebben. Kortom bij een Usability Test is zelfs één uitschieter genoeg om nader te bekijken en het achterliggende probleem serieus te nemen.

§5.3 Analyseren kwalitatieve (ordinale) gegevens

Behalve resultaten van allerlei kwantitatieve metingen zoals tijd en aantal fouten e.d., heb je ook informatie gekregen via de posttask-vragenlijsten en/of posttest-vragenlijst. Bij vragen naar een mening heb je een gegeven op ordinaal niveau. Die kun je op een rij zetten en hierbij eventueel een staafdiagram maken en/of de mediaan bepalen. Ook kun je in een aantal situaties weer percentages berekenen.

Het gebruik van percentages is ook zinvol in indirecte metingen. Bijvoorbeeld: zoveel % van de deelnemers % van de deelnemers vond de taak erg moeilijk. Dit kan soms ook belangrijke informatie geven.

Kijk ook hier vervolgens kritisch naar de uitkomsten. Zie je trends of juist niet, veel overeenkomsten of juist grote verschillen tussen de gebruikers. Probeer te achterhalen waar die vandaan komen (triangulatie). Vermoedelijk zit daar een usability probleem achter, maar er kan ook een andere oorzaak zijn.

De analyse bestaat hier dus ook weer uit de dezelfde drie stappen.

De analysestappen lijken erg veel op die van de vorige paragraaf. Het grootste verschil zit hem in de aard van de gegevens. Hieronder een paar voorbeelden. Aangezien bij de voorbeelden de context ontbreekt, kan de belangrijkste stap (stap 3) niet worden toegelicht (namelijk, wat is de *oorzaak* van de problemen).

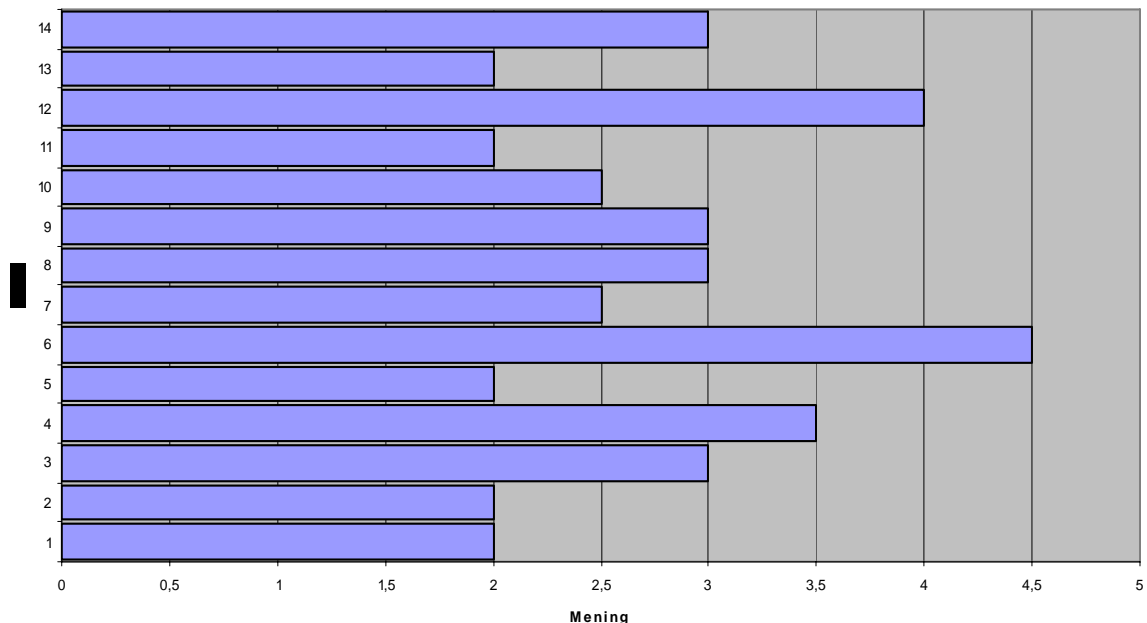
Voorbeeld 1

Bij een Usability Test is er na elke taak een posttask afgenomen. De eerste vraag was:

1. Hoe vond u de taak?
 - Zeer gemakkelijk
 - Gemakkelijk
 - Noch gemakkelijk / noch moeilijk
 - Moeilijk
 - Zeer moeilijk

Bij het invoeren heeft 'zeer gemakkelijk' de waarde 1 gekregen en zo oplopend tot 5 voor 'zeer moeilijk'. Van de mening van de 4 deelnemers is vervolgens per taak de mediaan bepaald en een grafiek gemaakt (zie figuur 5.2).

Figuur 5.2 Mediaan van de moeilijkheidsgraad van de taak

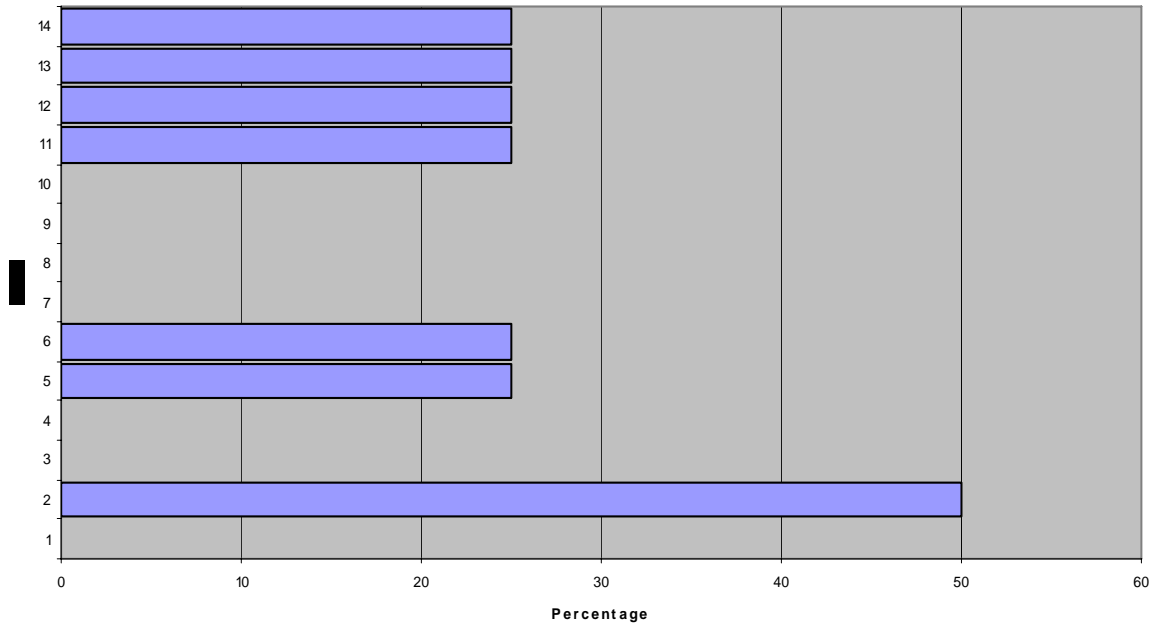


Direct valt taak 6 op die als moeilijk tot zeer moeilijk werd ervaren. Hier moet dus zeker wat aan gedaan worden. Daarna valt taak 12 op en taak 4.

Voorbeeld 2

Bij dezelfde Usability Test bleek niet elke taak door iedereen volbracht te zijn. Zie de grafiek in figuur 5.3 op de volgende bladzijde. Duidelijk is dat taak 2 erg moeilijk was. Hier is dus wat mee aan de hand. Het bleek dat de bijbehorende posttask-vraag door deze deelnemers niet was ingevuld, vandaar dat deze taak er hiervoor niet uitkwam. Belangrijk om dit mee te nemen in de evaluatie en het rapport! Verder waren taak 5, 6, 11 t/m 14 voor één persoon te moeilijk. De vraag is of dit steeds dezelfde persoon betrof? Nalopen van de gegevens blijkt dat het hier behalve bij taak 11 steeds om dezelfde persoon ging. Wat was er bijzonder aan? Geen tijd meer voor de laatste 3 taken? Het is duidelijk dat hier via de observaties en andere metingen dingen helder gekregen moeten worden.

Figuur 5.3 Percentage deelnemers dat taak niet kon volbrengen



§5.4 Bepalen van de omvang en de ernst van de problemen

Nu je op allerlei manieren de gegevens hebt bekeken en geanalyseerd, heb je een lijst gekregen van problemen met het product en met de verschillende (onderdelen) van de taken. Voor de goede orde: dit is niet de problemenlijst die in het begin van het hoofdstuk werd genoemd. Deze lijst zal er wel in verwerkt zijn. De lijst met problemen die hier bedoeld wordt, is uitgebreider naar aanleiding van de analyses. Je moet deze problemen nu ordenen naar belangrijkheid. Het ene probleem is namelijk veel 'erger' dan het andere. Hoe bepaal je nu welk probleem belangrijk is? Hier zitten 2 kanten aan:

1. De omvang van het probleem (*scope*)
2. De ernst van het probleem (*severity*)

De omvang van het probleem

Hoe wijd verspreid is het probleem? Is het slechts plaatselijk (*loca*): op 1 bladzijde, 1 scherm of menu, of op meerdere plekken (*global*).

Globale problemen zijn over het algemeen belangrijker dan plaatselijke problemen. De reikwijdte van een lokaal probleem is beperkt. Toch kan een lokaal probleem erg belangrijk zijn. Bijvoorbeeld als dit probleem zit bij een taak die vaak uitgevoerd moet worden. Of als het ernstige gevolgen heeft.

Bovendien kan een groep van een aantal lokale problemen duiden op een globaal probleem. Een van de voordelen van een Usability Test is, dat informatie verkregen met een test op een paar onderdelen van een product toegepast kan worden op andere delen van het product. Ook kunnen ontwerpers er weer van leren voor een volgend product.

De ernst van het probleem

Het indelen van de problemen in lokaal en globaal alleen is niet voldoende om te bepalen welke problemen als eerste opgelost moeten worden. Om een prioritering van de problemen te kunnen maken, moet je ook kijken naar de ernst van het probleem. Hoe kritiek is het probleem? Er worden hiervoor verschillende manieren gebruikt om de problemen te schalen. Alle schalen kennen 2 uitersten: aan de ene kant het probleem waardoor een taak niet uitgevoerd kan worden en aan de andere kant een aanpassing die aardig zou zijn, maar meestal cosmetisch van aard is.

We gebruiken hier de schaal van Dumas en Redish (A Practical Guide to Usability Testing).

- **Level 1 problems prevent completion of a task.** For example, users consistently select the wrong menu option and do not know where else to go, or participants give up after two tries to print an envelope because they can't figure out how to get the envelope into the printer correctly.
- **Level 2 problems create significant delay and frustration.** For example, a lack of feedback to users confirming what they have just done causes them to do the task over to make sure they have done it.
- **Level 3 problems have a minor effect on usability.** For example, using the same word to mean two different actions causes users to question, for a moment, whether they are making the correct choice.
- **Level 4 problems are more subtle and often point to an enhancement that can be added in the future.** For example, after working with a new software application for an hour, a participant suggests that initial ease of learning would be improved if there was a short online tutorial explaining three or four basic concepts. But be careful here. If participants say, "It would be nice if this product had a spell-checker," you might put that as a Level 4 recommendation for future enhancements. But if participants say, "I wouldn't use this product without a spell-checker," or "My e-mail has a spell-checker and I consider it essential," getting a spell-checker into the product may be a Level 1 problem.

Prioritering van de problemen

Als je al de problemen die je had gevonden, hebt ingedeeld naar hun omvang en de ernst van het probleem, dan kun je ze in volgorde van belangrijkheid gaan zetten; je kunt er een **prioriteit** aan toe gaan kennen. Meestal zijn de globale niveau 1 problemen het belangrijkste: die moeten verbeterd/gerepareerd worden. Indien mogelijk moet je ook een oplossing, of meerdere mogelijke oplossingen, aandragen voor de gevonden problemen. Dit wordt helder verteld in 'Usability Testing and Research', zie hieronder:

Making Recommendations

The last step in the analysis process is to recommend a solution for each problem or several possible solutions. Factors affecting the implementation of recommended solutions may be beyond the team's ability to control. One factor may involve the amount of time and resources that can be expended on implementing the recommendations. Another factor relates to the development stage of the product at the time of testing. If the testing is done early in product development, the recommendations can generally be made quickly and inexpensively. If, however, testing takes place late in the development cycle, then few changes may be permitted before the product releases. Still, the recommendations should be made and the arguments put forth for improving the usability of the product. One of the benefits of usability testing is to document improvements that should be made to increase the user's satisfaction with the product. If the changes cannot be made to the current product, seize the opportunity to make the case for earlier product testing next time to allow for faster, cheaper changes. Of course, the argument should also be made that any recommended changes be tested to confirm that the improvements do, in fact, increase the usability of the product. Typically, follow-up testing can be done at far less expense than initial testing because much, if not all, of the planning has already taken place: you know the user profile, the tasks you want to test, and the objectives of the test. You may even be able to bring in fewer participants to test, as the objective is to confirm improvements over previous data collected from testing.

Dan rest alleen nog het presenteren van je bevindingen in een rapport en/of presentatie aan de juiste mensen. Zie voor suggesties de volgende paragraaf.

§5.5 Rapporteren

Na het afnemen van de Usability Test en het verwerken en analyseren van de gegevens, zal er mondeling of schriftelijk verslag gedaan moeten worden van het hele gebeuren. Dit omvat ook een interpretatie van de resultaten van de analyse, het trekken van conclusies en het doen van aanbevelingen.

Bij de lessen vaardigheden wordt er aandacht besteed aan het schrijven van een rapport en aan presenteren

Voor de volledigheid wordt hier in het kort nog op het schrijven van een rapport ingegaan vanuit de invalshoek van Usability Testing.

In het rapport moet in ieder geval komen te staan wat de aandachtspunten en doelstellingen van de Usability test waren, de structuur/opzet van de test en uiteraard de uitkomsten. Over het algemeen is het rapport bedoeld voor het management en het ontwikkelteam. Dus voor personen die over het algemeen een leek zijn op het gebied van Usability Testing. Ze hebben meestal niet veel op met allerlei gedetailleerde informatie en te technische informatie over het testen en de resultaten. Aan de andere kant zijn er ook mensen die kennis van Usability Testing hebben, geïnteresseerd in het rapport. Zij willen juist wel graag weten wát je hebt gedaan en hoe en waarom en wat de resultaten zijn. Je hebt dus te maken met 2 verschillende groepen mensen voor wie je het rapport schrijft. Dit betekent dat je goed moet afwegen hoe je een en ander gaat aanpakken.

Uiteraard moet het rapport aan de technische randvoorwaarden van een goed rapport voldoen. Verder geldt dat je voor de hoofdstukindeling als leidraad de verschillende fases bij een Usability Test kunt nemen (zie hoofdstuk 1). In het rapport zal een verslag daarvan moeten komen waarbij ook aangegeven staat wat de motivatie is voor bepaalde keuze(s). Je zult hierbij moeten afwegen wat je in het verslag zelf zet en wat in de bijlage. Kijk hier kritisch naar! Een handvat is dat het rapport goed leesbaar moet zijn en dat de conclusies duidelijk moeten zijn zonder daar direct de bijlage(n) voor nodig te hebben.

Zo neem je bijvoorbeeld de interviews met iemand als bijlage op. De conclusies gebaseerd op die interviews komen echter in het verslag, waarbij je eventueel enkele zaken citeert uit de interviews om een en ander te verduidelijken. Verder verwijst je voor detailinformatie naar de bijlage.

Maak ook gebruik van tabellen, grafieken en screenshots om je verhaal te verduidelijken en te ondersteunen. Een plaatje zegt vaak meer dan duizend woorden.

Er wordt soms gedacht dat tabellen en grafieken niet in het rapport zelf zouden mogen, maar alleen in de bijlage. Dit is niet waar! Tabellen en grafieken dienen om verschillende gegevens overzichtelijk weer te geven en verduidelijken de bijbehorende tekst. Over het algemeen is de mens visueel ingesteld. Dit betekent echter niet dat het verslag overladen moet worden met tabellen en grafieken. Als het goed is, heb je veel gemeten tijdens de afname van de Usability Testen. De cijfertjes (én geen grafieken!) komen op een overzichtelijke manier in de bijlage. Bijzondere opvallende zaken illustreer je in het verslag zelf. Verder neem je in het verslag een analyse op van al die cijfertjes. Daarbij wil ik er, misschien ten overvloede, op wijzen dat het bekijken en verwoorden van wat te zien is niet voldoende is. Denk aan triangulatie: je bent op zoek naar de OORZAAK van het probleem.

Bijlage

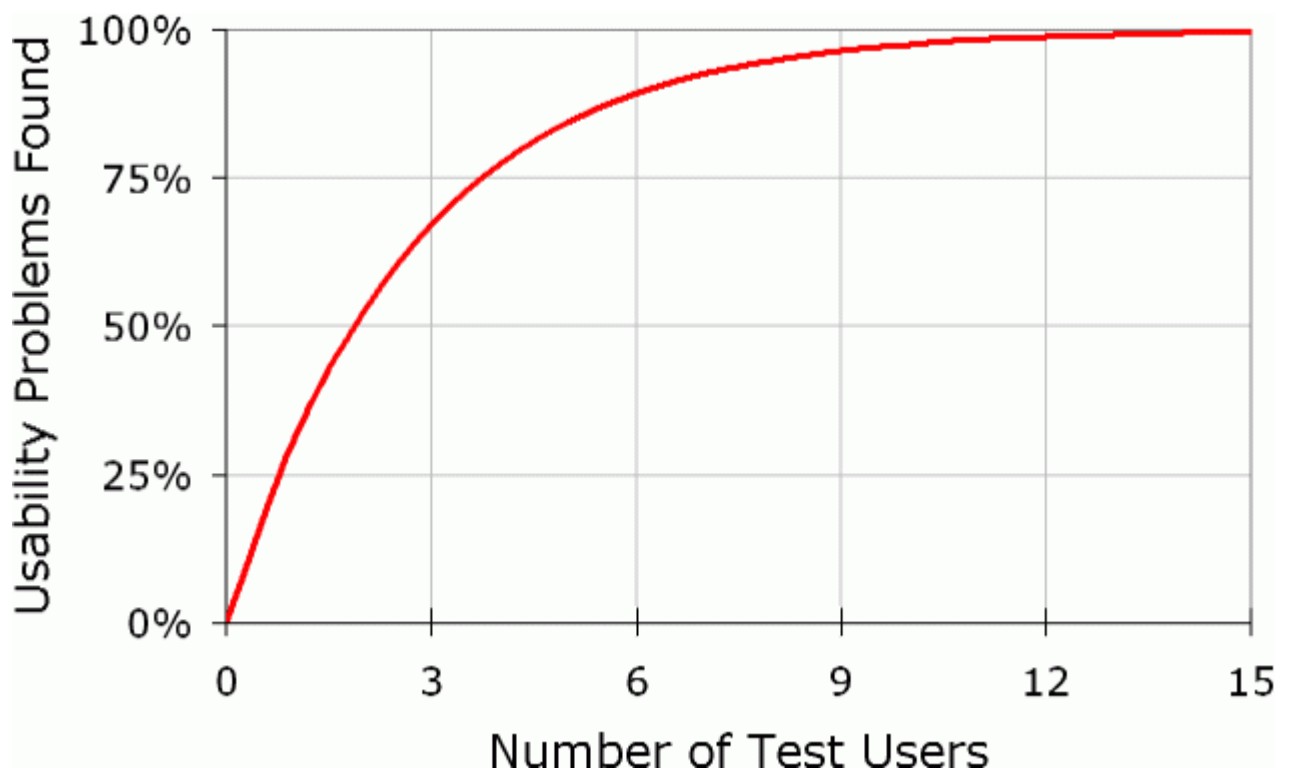
Why You Only Need to Test With 5 Users

Some people think that usability is very costly and complex and that user tests should be reserved for the rare web design project with a huge budget and a lavish time schedule. Not true. Elaborate usability tests are a waste of resources. The best results come from testing no more than 5 users and running as many small tests as you can afford.

In earlier research, Tom Landauer and I showed that the number of usability problems found in a usability test with n users is:

$$N(1-(1-L)^n)$$

where N is the total number of usability problems in the design and L is the proportion of usability problems discovered while testing a single user. The typical value of L is 31%, averaged across a large number of projects we studied. Plotting the curve for $L=31\%$ gives the following result:



The most striking truth of the curve is that **zero users give zero insights**.

As soon as you collect data from a **single test user**, your insights shoot up and you have already learned almost a third of all there is to know about the usability of the design. The difference between zero and even a little bit of data is astounding.

When you test the **second user**, you will discover that this person does some of the same things as the first user, so there is some overlap in what you learn. People are definitely different, so there will also be something new that the second user does that you did not observe with the first user. So the second user adds some amount of new insight, but not nearly as much as the first user did.

The **third user** will do many things that you already observed with the first user or with the second user and even some things that you have already seen twice. Plus, of course, the third user will generate a small amount of new data, even if not as much as the first and the second user did.

As you **add more and more users, you learn less and less** because you will keep seeing the same things again and again. There is no real need to keep observing the same thing multiple times, and you will be very motivated to go back to the drawing board and redesign the site to eliminate the usability problems.

After the fifth user, you are wasting your time by observing the same findings repeatedly but not learning much new.

Iterative Design

The curve clearly shows that you need to **test with at least 15 users to discover all the usability problems** in the design. So why do I recommend testing with a much smaller number of users?

The main reason is that it is better to distribute your budget for user testing across many small tests instead of blowing everything on a single, elaborate study. Let us say that you do have the funding to recruit 15 representative customers and have them test your design. Great. **Spend this budget on three tests with 5 users each!**

You want to run multiple tests because the real goal of usability engineering is to improve the design and not just to document its weaknesses. After the first study with 5 users has found 85% of the usability problems, you will want to fix these problems in a redesign. After creating the new design, you need to **test again**. Even though I said that the redesign should "fix" the problems found in the first study, the truth is that you *think* that the new design overcomes the problems. But since nobody can design the perfect user interface, there is no guarantee that the new design does in fact fix the problems. A second test will discover whether the fixes worked or whether they didn't. Also, in introducing a new design, there is always the risk of introducing a new usability problem, even if the old one did get fixed.

Also, the second test with 5 users will discover most of the remaining 15% of the original usability problems that were not found in the first test. (There will still be 2% of the original problems left - they will have to wait until the third test to be identified.)

Finally, the second test will be able to **probe deeper into the usability of the fundamental structure** of the site, assessing issues like information architecture, task flow, and match with user needs. These

important issues are often obscured in initial studies where the users are stumped by stupid surface-level usability problems that prevent them from really digging into the site.

So the second test will both serve as quality assurance of the outcome of the first study and help provide deeper insights as well. The second test will always lead to a new (but smaller) list of usability problems to fix in a redesign. And the same insight applies to this redesign: not all the fixes will work; some deeper issues will be uncovered after cleaning up the interface. Thus, a third test is needed as well.

The ultimate user experience is improved much more by three tests with 5 users than by a single test with 15 users.

Why Not Test With a Single User?

You might think that fifteen tests with a single user would be even better than three tests with 5 users. The curve does show that we learn much more from the first user than from any subsequent users, so why keep going? Two reasons:

- There is always a risk of being misled by the spurious behavior of a single person who may perform certain actions by accident or in an unrepresentative manner. Even three users are enough to get an idea of the diversity in user behavior and insight into what's unique and what can be generalized.
- The [cost-benefit analysis of user testing](#) provides the optimal ratio around three or five users, depending on the style of testing. There is always a fixed initial cost associated with planning and running a test: it is better to depreciate this start-up cost across the findings from multiple users.

When To Test More Users

You need to test additional users when a website has **several highly distinct groups of users**. The formula only holds for comparable users who will be using the site in fairly similar ways.

If, for example, you have a site that will be used by both children and parents, then the two groups of users will have sufficiently different behavior that it becomes necessary to test with people from both groups. The same would be true for a system aimed at connecting purchasing agents with sales staff.

Even when the groups of users are very different, there will still be great similarities between the observations from the two groups. All the users are human, after all. Also, many of the usability problems are related to the fundamental way people interact with the Web and the influence from other sites on user behavior.

In testing multiple groups of disparate users, you don't need to include as many members of each group as you would in a single test of a single group of users. The overlap between observations will ensure a better

outcome from testing a smaller number of people in each group. I recommend:

- 3-4 users from each category if testing two groups of users
- 3 users from each category if testing three or more groups of users (you always want at least 3 users to ensure that you have covered the diversity of behavior within the group)

Reference

Nielsen, Jakob, and Landauer, Thomas K.: "A mathematical model of the finding of usability problems," *Proceedings of ACM INTERCHI'93 Conference* (Amsterdam, The Netherlands, 24-29 April 1993), pp. 206-213.