Visualizing Financial Futures

Susanna Heyman
Akademisk avhandling som med tillstånd av KTH i Stockholm framlägges till offentlig granskning för avläggande av teknisk doktorsexamen fredagen den 8 september 2017 kl. 13:00 i sal F3, KTH, Lindstedtsvägen 26, Stockholm.
Abstract

Research on financial decision aids, systems designed to help people make financial decisions, is sparse. Previous research has often focused on the theoretical soundness of the advice that the systems provide.

The original contribution of this doctoral thesis is a set of empirical studies of how non-expert people understand the advice provided by financial decision aids. Since every piece of advice must be interpreted by a receiver, the accuracy of the advice can be corrupted along the way if the receiver does not understand complicated terminology, probabilistic reasoning, or abstract concepts.

The design concept resulting from the studies visualizes a unique combination of short-term and long-term variables that are usually treated as separate and not interacting with each other; loans and amortizations, insurance, retirement saving, and consumption. The aim is to visualize the consequences of different decisions and possible adverse events in terms of their effect on the user’s future consumption, rather than abstract numbers detached from the user’s lived experience.

The design concept was tested and evaluated by personal finance experts and professional financial advisors, as well as students and people without financial education, who represented the target users of the system. Results indicate that the system has a learning curve, but that once users understand how to read the graph, they find it more informative than conventional financial planning tools.

Keywords
Human-computer interaction; personal finance; fintech; behavioral economics; retirement saving; persuasive technology; financial literacy; data visualization; graph perception; external cognition; distributed cognition
Sammanfattning

Tidigare forskning om datorhjälpmedel för privatekonomiskt beslutsfattande har syftat till att bedöma hur korrekta råden är som genereras av systemen. Bedömningarna har gjorts av experter, som ibland har uttalat sig om systemens användarvänlighet men utan att testa det på vanliga användare utan utbildning i ekonomi.

Den här doktorsavhandlingen innehåller ett antal studier av hur vanliga människor förstår de råd de får av datorsystem för ekonomiskt beslutsfattande. Även om råden som genereras av systemet är korrekta, så är de inte användbara om användaren tenderar att misstolka eller bli förvirrad av exempelvis komplexa begrepp eller resonemang som bygger på sannolikheterna.

Designkonceptet som genererats av studierna utgör en modell som visualiserar en kombination av risker och parametrar som verkar på kort och lång sikt, och som mig veternligen inte har kombinerats tidigare i samma visualisering. Dessa parametrar, exempelvis bolån och amortering, försäkrings-, pensionssparande och konsumtion, brukar vanligtvis hållas åtskilda när man planerar sin ekonomi. Idén med designkonceptet är att få en större överblick över konsekvenserna av beslut och risker, och hur dessa påverkar användarens framtida konsumtionsutrymme.

Avhandlingen beskriver hur designkonceptet har testats och utvärderats av experter på privatekonomi, professionella ekonomirådgivare och studenter samt personer utan särskild ekonomisk utbildning, vilka representerar systemets målgrupp. Resultaten visar att det tar ett tag att lära sig hur man läser visualiseringsmodellen, men att när man väl har förstått det upplever man vanligen att det ger en tydligare bild av framtiden än konventionella planeringsverktyg.

Nyckelord
Människa-datorinteraktion; privatekonomi; fintech; behavioral economics; pensionssparande; persuasive technology; financial literacy; datavisualisering; graph perception; external cognition; distributed cognition
Preface

The idea to write a thesis like this came to me while I worked as an interaction designer in the finance sector. While we all agreed that our job was to help our customers understand their financial situation, there did not seem to exist any known best practices. What sort of information can be misleading? What type of design leads to what user behavior? What are some dos and don’ts for designing finance applications? I searched for handbooks on the design of personal financial management tools, but I did not find any. Not even an article or a blog post! Disappointed in the meager result from my search for guidelines and best practices, I decided to write the handbook that I wanted to read. This thesis is the end product.

In chronological order, I want to thank Kent Eriksson for believing in my idea for this thesis project and finding the funds. Thanks to VINNOVA for funding the early research and to KTH for funding the later. I also want to thank Johan Tjernell, my manager at my ordinary job, who supported me when I wanted to take time off for PhD studies.

My deepest thanks to Henrik Artman, my main supervisor who always had time, always was on my side, and showed commitment well beyond the expected. I could not have wished for a better supervisor.

Thanks to my co-supervisors Inga-Lill Söderberg and Mario Romero for many long and interesting discussions, as well as great teamwork with the practical aspects of studies. I have had hours and hours of fun discussing personal financial management tools and methodology in this interdisciplinary group.

Thank you Niklas Rönnblom, my husband and best friend who always encouraged me, believed in me, and pulled me through some of the darkest times of my life. I love you.

Stockholm, July 1st 2017
1 INTRODUCTION

In recent years, designers and Human-Computer Interaction (HCI) researchers have taken up an interest in using design to improve the world. There are computer systems and mobile apps designed to help us consume less energy1, quit smoking2, exercise more3, eat better4, and improve our mental health5. Within HCI as a research field, there are subfields dedicated to using computers and technology to save the climate6, end poverty7, and promote global peace8. Design for environmental conservation, often called sustainable HCI, and design for healthy living and sound medical practice, sometimes called e-health, are probably the two largest prosocial movements within HCI. Given this wave of idealism, it is remarkable how few HCI researchers have paid attention to the growing area called fintech – financial technology. Financial technology has given ordinary, non-expert people the power to trade stocks in competition with robots (Kirilenko & Lo, 2013), manage their own retirement savings without going through a professional9 (Cronqvist & Thaler, 2004), and choose between an increasingly complex flora of investment options (Brunnermeier & Oehmke, 2009) such as derivatives, crowd-funded start-ups10, and alternative currencies11. This empowerment of users can be both good and bad. It has made trading and investing cheaper and more accessible, but also more confusing and potentially dangerous.

Consider the case of an online stockbroker. Stockbrokers typically earn their profits by charging a fee on each trade their customers make, so naturally they want their customers to make as many trades as possible. Stockbrokers can encourage their customers to buy and sell on a regular

---

1 Examples: Kill-Ur-Watts, Energy Tracker, Green Outlet
2 Examples: Butt Out, Livestrong MyQuit Coach, Get Rich or Die Smoking
3 Examples: Strava, RunKeeper, C25k
4 Examples: MyFitnessPal, Lose It!, Noom Coach
5 Examples: Code Blue, Breathe2Relax, Operation Reach Out
6 ICT4S – Information and Communications Technology for Sustainability
7 ICT4D – Information and Communications Technology for Development
8 The Peace Innovation Lab at Stanford and the PeaceTech Lab at the government-funded United States Institute of Peace.
9 Although the idea that professionals achieve better results than laypeople is contested. See Hackethal, Haliasos and Jappelli (2012) and Chalmers and Reuter (2010).
10 Examples: Fundable, Seedrs, Tessin (specific to property development)
11 Examples: Bitcoin, Peercoin, Deutsche eMark
basis by actively notifying them whenever something interesting happens, such as a sudden stock market movement, a news story in the business world, or a press release from a public company. At first glance, this looks like a win-win situation; the customer keeps up with the market and the stockbroker earns trading fees. Intuition tells us that people who stay informed and act on new information in the financial markets must get better results. However, the opposite is in fact what happens. People who make many trades per month get worse results than people who buy stocks and hold them for longer times (Barber & Odean, 2000). The reason is that active traders are not as good as they think they are at beating the market, but they do incur trading fees and taxes, which lowers their net result. The effects of these fees and taxes add up. An annual fee of 1% lowers the portfolio value by 20% after 20 years (SEC, 2014).

Fintech and its role in society thus presents a complex entanglement of fast computers, vast amounts of data, human behavior and cognition, advanced and mobile interfaces, and ethical puzzles. In my opinion, it is unfortunate that critical analyses of fintech systems are so rare in HCI, a field that otherwise prides itself on being prosocial, user-centric, and reflective. People who make mistakes while using fintech systems may have to suffer the consequences for the rest of their lives. Making the wrong financial decision can lead to poverty after retirement, lost opportunities for leisure and enjoyment, and stress and anxiety over an uncertain future.

Can clever interaction design help people make better financial decisions? Others have argued (Norman, 2010; McCullagh, 2010), and I agree, that complex problems cannot be solved by designers who are only experts in design principles, user-centric methods, and “creativity.” Designers who are ignorant of the problem domain, in this case finance, risk coming up with “solutions” that are misguided, naïve, or insufficient. We should also know the limits of what design can do; better designed computer systems are not the answer to all of society’s problems (Baumer & Silberman, 2011; Morozov, 2013). Many people are poor due to factors that are beyond their own control and no interactive decision-support system in the world is going to help them. Thus, problems of poverty and inequality that are due to larger structures in society are outside the scope

---

12 Assuming an annual growth of 4%. Portfolio value is lost both through the subtraction of fees, and through the loss of the additional return that would have been earned, had the fees paid been invested instead.
of this thesis. Sometimes a better answer is new regulation or policy changes.

However, there are circumstances under which an individual does have some control, or at least influence, over their own financial future. Planning, saving, and not over-using credit are ways in which an individual can increase their chance of financial success. When a designer in the financial sector sets out to create a system that will help people make better decisions, they need to know the fundamentals of both normative and behavioral finance in order to understand what they are doing.

Normative finance describes how a rational actor would act in a given situation. This can help us set design goals; how we want our users to behave. Behavioral finance describes how people actually do act in the real world, which can further guide our design decisions. We know, for example, that people overreact or underreact to market information under special circumstances (Odean, 1998), that they overvalue the importance of negative information compared to positive information (Kahneman & Tversky, 1979), and that they tend to focus on relatively unimportant details (Lucey & Dowling, 2005). This helps us form hypotheses about what information should be salient and what should be hidden, or which actions should be quick and easy and which should require the user to go through more confirmation checks.

This can, of course, present an interesting dilemma, because the users themselves, unaware of the damaging effects of too much trading, may want access to live updates from the stock market and may want every action to be executed instantaneously with no interfering confirmation dialogs. The designer is now confronted by the question of how paternalistic they want to be.

Behavioral economics has a lot in common with HCI, in that they are both heavily empirical traditions in which researchers strive to observe people's real behaviors with all their quirks and mistakes, instead of constructing theoretical models of how idealized people should behave – and take care to design computer systems or economic policies for real people, not idealized people.

---

13 Though in practice, many behavioral economics studies are conducted in labs with artificial experiment situations. However, it is closer to reality than purely theoretical models of human behavior.
I am not the first to suggest that behavioral economics can be helpful for interaction designers working in the finance sector (Gunaratne & Nov, 2015) or in other domains (Lee, Kiesler & Forlizzi, 2011; Froehlich, 2009; Lockton, Harrison & Stanton, 2009). In the background chapter, I review some earlier studies where researchers set out to apply findings from behavioral economics and turn them into implications for design. A common view is that people have some innate systematic errors of judgment and that the right design can compensate for said errors and help people make more rational decisions.

When I first started my PhD project, my research question was how a graphical visualization (of something!) could help non-experts better understand the relationship between risk and return. It was silently agreed, or at least I perceived it so, that the risk we were talking about was market risk – the risk associated with ups and downs in the financial markets, which mainly affects the owners of investment portfolios. My first study focused mainly on market risk and I assumed that decisions around investing were central to people, or at least that they should be. Later, as I found more literature, I learned that the return on their stock portfolio is a relatively unimportant factor for most people except the wealthiest. Thus, if I wanted to tackle a problem of great significance, I was forced to re-evaluate what "risk" meant to me in the context of this doctoral work. Instead of risk as the standard deviation from the expected return on a stock portfolio, I decided to look at risk as potential adverse events that could threaten the financial well-being of an individual. It could be said that I moved away from the typical finance professional's technical definition of risk and toward a definition that is more in line with ordinary people's perception of the term; laypeople tend to think of risk in terms of potential consequences rather than probabilities (Slovic, 1987).

The result is a visual concept that strives to capture the main risks that may threaten an average Swedish citizen today: for example short-term catastrophic events, rising interest rates, a pension that is lower than expected, and declining house prices. The exact scenarios I chose to include are not what's important; the scenarios are meant to change dynamically in response to changes in the economic and political climate. This way, the concept can also be modified to suit the economic context of countries other than Sweden.
1.1 RESEARCH FIELDS

This thesis is positioned in Human-Computer Interaction (HCI). HCI traces its roots back to ergonomics. Traditionally, HCI researchers study the relationships between human users and computer systems, in order to learn how to make the computer system more efficient and user-friendly (Faulkner, 1998). I also draw from its subfield, Information Visualization (InfoVis). InfoVis focuses on how data structures such as graphs should be visually and interactively designed to suit a specific type of user who has concrete tasks relating to particular data. Both qualitative and quantitative methods are common in HCI and InfoVis, but quantitative studies often have fewer participants than is conventional in other fields. The nature of the problems studied in HCI and InfoVis makes it possible, even considered preferable, to use within-subject studies where the same participants see different variations of the same system, rather than between-subject studies with more participants (Munzner, 2016).

HCI and InfoVis studies are often embedded in a domain for which the systems in question are being developed. For InfoVis studies in particular, it is considered necessary to begin by a well-informed description of the problem which the visualization is meant to solve (Munzner, 2009).

For this thesis, HCI is the research field but personal finance is the domain. I therefore draw additional theories and concepts from financial planning and behavioral economics when describing the problem that my design is meant to address.

1.2 DELIMITATIONS

While personal finance includes many potential areas to explore, I have chosen to focus mainly on retirement planning. Figure 1 shows my conceptual model of this topic, which is also how I have structured the thesis; it begins by a general introduction to personal financial planning but delves deeper into retirement planning.

The term fintech describes areas that are undergoing fast innovation and transformation, including but not limited to the areas in the mid-level of Figure 1 (Investopedia, n.d.). Out of all the areas in fintech, I have focused on tools for advice and planning. Financial planning, in turn, is
1. INTRODUCTION

said to consist of the areas in the bottom level (Financial Planning Standards Board, 2015).

I have chosen to focus on users who lack financial expertise or training. The general level of financial literacy in Sweden is low – half of all Swedish adults lack what the Swedish Financial Supervisory Authority calls “basic financial knowledge” \(^{14}\). Specifically, they fail to correctly answer three questions about interest, inflation, and risk (Ekman, 2015).

While systems for experts can present a huge amount of information at the same time, and the focus is on providing shortcuts so that the expert can work efficiently, systems for novices need to break processes and information into piecemeal steps so as not to overwhelm the novice user (Faulkner, 1998).

Because of the differences between experts and novices, it is crucial to use people who are representative of the intended target group when evaluating systems.

\(^{14}\) My translation. Original: “grundläggande finanskunskap.”
1.3 RESEARCH QUESTIONS

My overarching goal was always to explore how data visualization might help people make long-term decisions that would make them more likely to be satisfied with their financial situation in retirement. However, my focus shifted somewhat along the way, which is reflected in the progression of the research questions.

Research Question 1. How comprehensible are today’s financial decision aids to non-expert users? Do they find them helpful? What are the systems’ strengths and weaknesses?

At this point, I still thought of a “financial decision aid” as something that helps a user choose an appropriate investment portfolio according to their risk preference. Hence, this study only tested so-called portfolio pickers. However, some of my findings in this study ought to generalize to all types of financial planning systems.

Research Question 2. What would a financial decision aid look like that addresses the weaknesses while preserving the strengths found in RQ1? Is a representational change necessary?

A representational change is a complete transformation of the model presented in a graph, as opposed to a minor tweak. For example, changing what data units are presented because the new data units have been found more relevant to decision making.

While working with this research question, which was largely a creative task, I came up with the idea to contextualize the expected return from an investment against the user’s needs. If we know how much the user consumes, we can translate a lump sum of, say, a million Swedish crowns, into a number of years of sustained consumption after retirement. If the prognosis is that the user’s retirement savings won’t last very long, then the user needs to do something in order to increase the expected return. At this point I thought that “something” was generally to invest more aggressively in riskier products.

Part of the inspiration for this idea came from watching participants in my first study appear to have no clear opinion of how satisfied they should be with the recommendations from the portfolio picker and part of it came from financial advisors in a workshop saying it was common for their clients not to have any defined goal for their investment. I later learned that I was not the first to invent this design concept, but that the concept as such had support from well-cited retirement specialists.
However, no one had tested how well non-expert users understood the concept.

**Research Question 3.** Do non-expert users understand how to use the representation constructed in answer to RQ2? Does it work better than systems that exist today?

I tested this empirically by having non-experts solve a task using the new representation and a control group using a representation typical of the systems I tested in RQ1. I also conducted a qualitative user study.

**Research Question 4.** Does the new representation “nudge”\(^\text{15}\) users to take more or less risk than they would with another representation?

This is a question of general interest to actors who might be interested in implementing my visual concept, because anyone who gives financial advice in Sweden must take the client’s risk preference into account (SFS 2003:862). However, I would never be able to make any normative claims as to whether the users of my system demonstrated more “appropriate” risk preferences than the control users, only whether or not they were different. It turned out that participants who used the new representation took somewhat less risk than did participants who used a control. The effect was only seen in a test scenario where participants were told to think of a young person with modest wealth. When the fictional test case was older and wealthier, the difference disappeared.

I returned to RQ2 and RQ3 in an iterative process as the visualization prototype evolved.

Table 1 shows a summary of all research questions, methods, and results. The wording of the research questions is somewhat shortened in order to fit in the table.

\(^{15}\) The concept of “nudging” is explained in section 2.2.6.
Table 1: Research process. The shading of table cells is for readability.

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Methods</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RQ1</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| How well do current financial decision aids work for non-expert users? | Reviews of existing systems | Heuristic evaluation | • Narrow focus on market risk  
|                   |                    | Usability test                               | • Assume knowledge of complex terminology  
|                   |                    | Workshop*                                    | • Fail to persuade users to save  
|                   | Requirements gathering |                             | • Too abstract  
|                   | Heuristic evaluation |                                             |
| **RQ2**           |                    |                                              |
| Can a new representation address the weaknesses while preserving the strengths found in RQ1? | Concept design | Workshop* | • Translate savings from lump sum to level of monthly consumption after retirement |
|                   | Expert evaluation | Sketching & prototyping                      | |
|                   | Financial advisor test | Interview                                     | |
|                   | Usability test     |                                             |
| **RQ3**           |                    |                                              |
| Do non-expert users understand how to use the new representation? | Comprehension test | 3 pilot experiments | • Significantly more success at solving a retirement planning task with the new representation than a control  
|                   |                     | Experiment                                    | • Perceived as more complex but a better overview of future consequences  
|                   | User test           | Usability test                               | |
| **RQ4**           |                    |                                              |
| How does the new representation influence users’ risk taking? | Risk test | Pilot experiment | • Take somewhat less risk than with a conventional tool, when imagining a scenario where the test case is a young person of modest wealth |
|                   |                     | Experiment                                    | |

* The workshops in RQ1 and RQ2 are the same; it served two purposes.
2 BACKGROUND

2.1 FINANCIAL PLANNING

The purpose of financial planning is to try to make sure that a person can afford to pay their expenses, before as well as after they are retired. This entails planning ahead for expensive life goals such as buying a home or travelling (Gutter, n.d.; UCI, n.d.), as well as possible negative events such as sickness or unemployment (Gutter, n.d.). The process of financial planning thus involves setting goals, protecting assets and earning potential through insurance and/or saved capital, and growing capital through long-term investment (Gutter, n.d.; UCI, n.d.).

It is typically recommended to save for an emergency fund for unforeseen expenses before starting any long-term investment. The emergency fund should cover at least a few months' worth of expenses (Gutter, n.d.). Very expensive catastrophes, for which it is difficult for most people to save up enough money, such as a house fire or a life-long disability, are handled through insurance (Gutter, n.d.; UCI, n.d.).

When the person has a sufficient emergency fund, they should start investing money into assets that grow over the long term, such as stocks and bonds (Gutter, n.d.). Because of the exponentially growing compound interest effect, there are enormous benefits to starting such investing early in life (Gutter, n.d.). Riskier investments will typically pay a higher return as compensation for the risk (Mehra & Prescott, 1985) and as stock market movements are said to smooth out over long time periods, long-term investments are considered suitable for taking some more risk.

There are two types of risk involved in investing on the stock market: Systemic risk and idiosyncratic risk. Systemic risk is difficult or impossible to guard against. It shocks the entire economic system. Examples are changing interest rates, stock market crashes, or instability in a country. Idiosyncratic risk only affects one company, for example the risk of bankruptcy. An investor can reduce their exposure to idiosyncratic risk through diversification. By diversifying across industries, regions, and company sizes, a stock portfolio can reduce risk without reducing return. In general, larger companies are more stable, while smaller

---

16 This conventional wisdom has been challenged by Bodie (2003), see chapter 2.3.2 on assumptions in retirement planning software.
companies grow more, and investors should have a mix of both (Gutter, n.d.).

Simply investing an equal amount of money in all available stocks or mutual funds is not considered the best way to diversify (Benartzi & Thaler, 2001). Modern portfolio theory states that when historic risk and historic return are considered together, some combinations of assets have proved to generate a better return for a given risk than others (Markowitz, 1952). These optimal combinations, or portfolios, are said to lie on the efficient frontier. A rational investor should always choose a portfolio on the efficient frontier that matches their risk preference.

Figure 2: The efficient frontier. Portfolios on the efficient frontier give the best expected return for a given risk.

Portfolios A and B in Figure 2 are both on the efficient frontier and thus are both good portfolios. The choice between A and B depends on how much risk the investor is prepared to take. Portfolio C is inferior because compared to portfolio B, it gives a lower return for the same risk, and compared to portfolio A, it takes higher risk for the same return.

The easiest way for a layperson with limited funds to compose a diversified portfolio is to buy shares in a mutual fund (Serlin, n.d.). However, mutual funds charge fees, and low fees are an important factor for the overall result over long time horizons. Index funds generally give the best combination of good performance and low fees for the stock market (Serlin, n.d.).
However, diversification is only one aspect of choosing a suitable investment portfolio. As illustrated by the efficient frontier, there are several portfolios that are all well-diversified, but with different risk. Among the portfolios on the efficient frontier, investors must choose one that matches their risk tolerance.

Risk tolerance has an objective aspect; it is generally considered to be dependent on the time horizon of the investment (if the time horizon is long, more risk is acceptable) and on whether or not the investor could afford to lose the investment (Gutter, n.d.; UCI, n.d.). Another factor is how much growth is needed to achieve the investor's goal (UCI, n.d.). Additionally, people can take more or less risk depending on their ability to adjust their income and their spending in response to financial loss (Bodie, Merton & Samuelson, 1992). For example, people with stable jobs who can choose to increase their working hours or retire later in order to compensate for a financial loss can take higher risks, as can people with low fixed costs and a large margin between income and spending.

However, risk tolerance also has a subjective aspect – what Gutter (n.d.) calls the "sleep at night factor." This depends on whether the investor is willing to risk some loss in order to get a chance at a higher return. An investor should not risk more than they feel comfortable with. For the remainder of the thesis, I will refer to the subjective component of risk tolerance as risk preference.

Assets, debt, and income are plain numbers that can simply be fetched from the user’s transaction records. However, the user’s goals, plans for the future, and risk preference constitute “soft” or “fuzzy” information that the system cannot deduce from records in a database. The user may have never thought of their goals or their risk preference before, but when they sit down in front of a financial decision aid it becomes necessary to formalize their risk preference into something concrete that can be fed into the system for processing. This is due to a Swedish legal requirement that all financial advice, including advice delivered through a computer algorithm, be suited to the individual advisee’s risk tolerance (SFS 2003:862).

---

17 This conventional wisdom has been challenged by Bodie (2003).
18 Risk preference is related to risk perception, but the two are not the same (Wang, Keller & Siegrist, 2011). For instance, a risk-averse person can unwittingly take high risks if they perceive risks to be lower than they are.
However, there is no generally scientifically accepted way to ask a person about their risk preference. In fact, there is no consensus that risk preferences even are stable personality traits (van Schaik, Kusev & Juliusson, 2011). Risk preference has been shown to depend upon the person’s current mood (Klinger & Levy, 2009; Loewenstein et al., 2001), how the question is asked (Kahneman & Tversky, 1979), and what the person sees other people doing (Trautmann & Vieider, 2012). People have also been shown to change their risk preference when trends on the stock market change (Kim & Nofsinger, 2007).

2.1.1 Retirement Planning

People who plan for their retirement are more satisfied with their life after retirement (Elder & Rudolph 1999). Yet many people avoid planning, especially people who need it the most, namely people with lower incomes, less qualified jobs, and lower levels of financial literacy. This has been confirmed in international studies (Taylor & Doverspike, 2003; Lusardi & Mitchell, 2005) as well as Swedish studies (Almenberg & Säve-Söderbergh, 2011).

People who employ financial advisory services save more and accumulate more wealth before retirement, but this effect is only seen if the advisor helps the clients calculate their retirement needs (Martin & Finke, 2012). Even the act of calculating retirement needs itself seems to increase retirement savings (Mayer, Zick & Marsden, 2011). People who succeed in their planning are more likely to use tools such as retirement calculators, while people who do not succeed are more likely to plan by discussing their strategies with family or friends (Lusardi & Mitchell 2005).

Several economists have argued that it is dangerous to focus on maximizing retirement funds without first calculating needs (Merton, 2003; Merton, 2014; Bodie et al., 2004; Siegel, 2015; Boshara & Emmons, 2015). Unsophisticated investors who are simply looking to grow their capital without any clear goal might feel satisfied so long as their net development has been positive (Merton, 2014), but many investors are in fact surprised to find that their savings are not enough to live on comfortably after retirement (Dimensional Fund Advisors, 2014). A needs calculation should take into account any government pension that will help pay for some of the expenses (UCI, n.d.). Current lifestyle is
considered a good-enough approximation for future lifestyle\textsuperscript{19} (UCI, n.d.; Merton, 2014).

The following risks have been named the most important for retirees and people planning their retirement (Turner & Witte, 2009). The relative importance of risks differs between socioeconomic groups; for people with more financial assets, financial market downturns pose a greater threat (Turner & Witte, 2009).

- **Longevity risk.** The risk of living longer than expected, thereby running out of money and spending the last years in poverty.
- **Retiring earlier than expected.** Early retirement, forced by job loss or health concerns, reduces retirement income.
- **The death of a spouse.** The death of a spouse will usually affect household income to an extent that is not fully compensated by the accompanying decrease in spending, since it is cheaper to live together than alone.
- **Financial market downturn.** Financial market downturns reduce the value of saved capital. This can be especially bad for a person who is in the process of drawing down from said capital, because it increases the rate of decline even further.
- **Rising interest rates.** For people with variable mortgage rates or consumer credit debt, this increases household expenses.
- **Declining interest rates.** For people who rely on bonds or other interest-bearing investment products, this means less income.
- **Increase in inflation.** For people with a fixed, non-inflation-protected income, this reduces their spending power.

\textsuperscript{19} The so-called life-cycle hypothesis is the dominant model of saving (Wärneryd, 1999). It posits that people want to smooth their consumption over their lifetime, rather than follow their income which is typically low in the beginning (as students) and end (as retirees). However, empirical evidence suggests that rather than a constant lifestyle, people may more typically want a continually improving lifestyle (Wärneryd, 1999). It has also been observed that people save more than predicted by the life-cycle hypothesis. A possible explanation is that people amass additional savings out of precaution (Wärneryd, 1999). Particularly, the elderly do not spend down their savings as quickly as predicted (Danziger et al., 1982; Bernheim, 1987). This could be due to precaution and uncertainty over longevity, or a desire to bequeath. However, though it may need additions and adjustments, the life-cycle hypothesis has not been abandoned (Wärneryd, 1999). For the purpose of retirement planning for an individual, I agree with Merton (2014) that current lifestyle is probably a pedagogical way to frame the discussion about future lifestyle.
- **High health care and long-term care costs.** Health care and drugs are one type of spending that tends to increase as a person ages.
- **Housing market decline.** For people who are planning to sell their house or borrow against it to finance their retirement, declining house prices restrict their options.

Financial risks can be transferred to another party through the purchase of insurance (Turner & Witte, 2009; Gutter, n.d.); mitigated through a voluntary reduction in spending which increases economic margins (Turner & Witte, 2009); accounted for by the use of conservative estimates when planning (Turner & Witte, 2009); or accepted, if the potential consequence of the risk is not catastrophic (Gutter, n.d.). One solution to the problem of a retirement plan that looks like it is not going to cover the person’s spending needs, is to take more investment risk in order to achieve a higher return. This solution is promoted by some retirement planning software (Turner & Witte, 2009), but it is a risky approach that can backfire. This is discussed further in section 2.3.2.

### 2.2 Psychology, Computers, and Personal Finance

Daniel Kahneman and Amos Tversky (1979) pioneered the field of behavioral economics when they showed that people did not really behave according to the rational principles of economics. However, the ways we deviate from rational behavior are often systematic and non-random; Dan Ariely (2010) called this being “predictably irrational.” Kahneman (2011) explains this by likening the human mind to two systems, System I and System II, which complement each other. System I is fast and resource-efficient. It makes quick, superficial judgments that usually provide “good enough” results. Herbert Simon called this concept “satisficing” in his theory of bounded rationality – sometimes the slightly better result that would be achieved by expending a lot more effort just isn’t worth it. System II, on the other hand, performs complex calculations and tries to take all available information into account. It is

---

20 While Kahneman popularized the term, he was did not coin it – he borrowed it from psychologists Keith Stanovich and Richard West.
considerably slower and requires more effort and therefore, claims Kahneman, we prefer to use System I whenever we can\textsuperscript{21}.

In cognitive psychology, the study of “higher mental processes such as attention, language use, memory, perception, problem solving, and thinking” (APA, 2002), the trend in recent decades has been to move away from viewing the human brain’s internal processes as the only interesting focal point and toward a view that includes how human mental processes are shaped by and interact with the human’s environment (Scaife & Rogers, 1996). Hutchins (1995) suggested that the behavior of complex, goal-oriented systems can be better understood when the unit of analysis is not the brains of individual people, but the entire system of collaborating people and artifacts. Hutchins illustrated his point through the example of an airplane cockpit, which requires a complex interaction between pilots, sensors, and gauges that measure and store values in order to maintain the proper velocity of the airplane. Artifacts\textsuperscript{22} that support human thinking, for example by storing values that would otherwise have to be stored in human memory, have been called cognitive artifacts (Norman, 1993). It is crucial that cognitive artifacts present information in a way that makes the task as easy as possible for the human (Norman, 1993). This entails presenting information in the right order, at the right level of detail, and in the right data unit.

Thus, when a person is using a financial planning system, the planning process can be thought of as a joint activity between the user, the computer, and the professional advisor, if one is present. The computer extends the user's cognition by performing calculations and predictions and by presenting information about investment options about which the user may have been unaware. At the same time, the system cannot make decisions on its own without the user, since the user has unique knowledge about their own goals and preferences. In order to make decisions easier for the user, the system must present options in a way that makes the consequences of the user’s choices visible in terms of the user's goals and preferences. In other words, the system must adapt to

\textsuperscript{21} This dichotomy has been criticized for not being fully empirically supported (Fiedler & Sydow, 2015). For example, System II can also produce biased judgments, especially in people of lower cognitive ability. However, just because System II is more meticulous does not mean that it’s infallible.

\textsuperscript{22} An artifact can be for example a post-it note, a calculator, a computer, or a thermometer.
the user's language and avoid jargon or concepts that are unfamiliar to the user.

![Figure 3](#)

A similar model has been used to describe the meeting between clients and human financial advisors (Jungermann, 1999)\(^\text{23}\), although it can be argued that the computer is even stronger in retrieving and processing vast amounts of information while human advisors have other strengths, such as an ability to adapt their communication style to the client's knowledge level and emotional state. Figure 3 may suggest that the client comes to the meeting with stable and fixed goals and preferences, but this isn't necessarily the case. However, the responsibility for deciding goals and preferences must necessarily fall on the client.

### 2.2.1 Graph Perception

Graphs are a clear example of cognitive artifacts and they are commonly seen in financial decision aids. The purpose of presenting financial data in graphs rather than tables or text is to take advantage of some strengths of the human visual apparatus: The visual channel is efficient at processing differences in length, position, and color, which allows us to

\(^{23}\) Jungermann proposes that, contrary to previous thinking, human advisors tend not to present their clients with a set of options but rather with only one option – the recommendation – and that the client’s role is not to choose between options but to accept or reject the recommendation from the advisor. For the purpose of this thesis, I find that distinction to be of secondary importance.
perceive patterns and find outliers quickly and without effort (Mazza, 2009). However, like any cognitive artifact, graphs must be designed in a way that suits the task.

Statisticians and cartographers have studied the suitability of different types of graphs since at least the early 20th century, but Cleveland and McGill (1984) were the first to systematically break down visual structures into what they called “elementary perceptual tasks,” and to rank them based on how accurately people can use them to read quantitative information.

Different types of graphs present the reader with different such elementary perceptual tasks – for example, in order to compare values in a bar chart, the reader must judge the position of the top of the bars relative to a common vertical scale. They can also judge the values based on the lengths of the bars, although it is more likely that the positions of the tops is what people look at (Cleveland & McGill, 1984). In contrast, when reading a pie chart, the reader must look at the angles of the different “pieces of pie” and perhaps also the areas of said pieces.

Consider Figure 4; most people find it easier to arrange all the values from smallest to largest in a bar chart than a pie chart. In a stacked bar chart, the bottom pieces are easy to compare since they share a common start position, but pieces that are higher up, for example the green pieces in the stacked bar chart in Figure 4, are harder to compare (Cleveland & McGill, 1984).

Cleveland and McGill found through experiments that people’s judgment of quantitative information is the most accurate when presented as a position on a scale. Accuracy decreases when people are asked to distinguish between lengths, angles, areas, or volumes. These results have later been supported by evidence from large-scale experiments (Heer & Bostock, 2010). The conclusion is that graphs based on position, length,
and sometimes angle are usually good representations, while graphs based on area or volume should be avoided.

Figure 5 exemplifies this. It is difficult to tell how much larger (in %) the green circle is than the blue circle. This is partly because the human eye is not good at comparing areas and partly because the task is ambiguous; did the designer intend us to compare the circles based on area or radius? The volume chart suffers from the same problem to an even greater degree (Cleveland & McGill, 1984).

![Figure 5. Charts that are usually unsuitable.](image)

This does not mean that one type of graph is the best choice for all tasks. Some graphs are better for reading exact numeric values; others are better for comparing the relative sizes of parts to the whole (Simkin & Hastie, 1987).

However, as Cleveland and McGill (1984) pointed out, the main reason for using graphs is not to achieve perfect accuracy; otherwise a table of exact numbers would always be the best representation. Graphs are used for their ability to reveal patterns and structures. However, a high degree of accuracy is important in order to make sure that a perceived pattern is real and not just an optical illusion.

Scaife and Rogers (1996) have criticized studies on graphical representations for being ignorant of the workings of mental processes, basing claims and methodology on the unfounded assumption that graphs work by depicting a model of reality and transferring that model into the viewer's head. Scaife and Rogers argue that an alternative mechanism might be that graphs are disambiguating – verbal descriptions are often vague and can be interpreted in different ways, while a picture “constrains” the number of possible interpretations to the
one intended by the designer\textsuperscript{24}. This might explain why some studies have shown success in improving participants' quantitative reasoning; the graph simply made the task less ambiguous. Scaife and Rogers note the difficulty in designing experimental tasks where the information contained in a graph and in a verbal description are equivalent.

Additionally, Scaife and Rogers speculate that the reason that older and well-known graphs often seem to outperform new, innovative graphs (flat vs. 3D, static vs. dynamic, black-and-white vs. color) may be not that the former are more naturally adapted to our visual apparatus, but rather that we have had years of experience in reading them.

While the quality criterion for a “good” graph according to visual perception theorists may be that it makes it easy for viewers to read quantitative numbers and find patterns, Scaife and Rogers, who looked at the problem from a learning perspective, rather asked the question: What helps a learner learn? From their perspective, a good graph helps the viewer learn about the subject domain. A certain amount of active cognitive effort may help the learner construct and remember a conceptual model of how the domain works. A deceptively simple graph can help people solve a task simply by making it graphically apparent “where [the] object wants to go” (p. 202), without the need to invoke deeper knowledge about what the graph represents in the subject domain. This will make the graph look successful in a test without necessarily enhancing the user’s understanding. Imagine an interface such as Figure 6; the example is exaggerated but it illustrates the principle. A user test might show that all users understood that the test conductor expected them to push the button, but it is possible to pass such a test without understanding anything about what happens “under the hood” when the button is pushed.

It could be debated whether the goal for a financial decision aid should be to educate the user about the intricacies of financial markets, or simply help them come to a decision that is

\textsuperscript{24} Ainsworth (2006) illustrates this point through the following example: The sentence “the cat is by the dog” can be left ambiguous as to whether the cat is on the dog’s left or right side. However, it is not possible to draw a picture of a cat and a dog without making a decision about which side of the dog the cat sits on.
likely to be good for them. While some have pointed out the folly of trying to teach every citizen how to manage their own finances and argued that it should be the responsibility of policymakers to ensure that everyone gets tolerable life outcomes (Willis, 2008; Merton, 2014), the Swedish law regulating financial advisory services is based on the view that what people need in order to make better decisions is more transparency and more information about different investment products (Andersson & Korling, 2012). This view clearly falls into the “educational” model of financial advice; that people can and should be taught how to make their own decisions rather than have someone make decisions for them. If this “educational” spirit of the law is to be followed, then it is better to strive for graphs that enhance deeper understanding rather than graphs that make it superficially easy to reach a quick decision.

Additionally, if we take the least-cognitive-effort approach to its logical conclusion, then visualization becomes unnecessary, since all decisions are made by the advisor and not the advisee25. I will return to this point in chapter 2.3.1, my review of academic studies on financial decision aids.

Scaife and Rogers (1996) suggest three analytical concepts for explaining external cognition. Computational offloading refers to the way that a cognitive artifact can reduce the mental effort required to solve a task. A chart can make a problem significantly easier by presenting it in such a way that viewers can “read” the answer from the chart rather than compute the answer themselves. Re-representation refers to how some representations are inherently easier to use than others. As proof, Scaife and Rogers cite Zhang and Norman (1994), who showed that it is much easier to multiply numbers represented through Arabic numerals than Roman numerals. Finally, graphical constraining refers to how a graphical representation can restrict the number of possible interpretations of a problem so that, ideally, only the intended interpretation is a logical choice. Scaife and Rogers admit that these concepts are closely related and sometimes hard to distinguish.

### 2.2.2 Mental Accounting

While knowledge about the internal mental structures that process graphical information may be immature26 (Scaife & Rogers, 1996), some

---

25 In this case, the advisor would be a computer that makes a number of “invisible” decisions on behalf of the user, in order to be able to present a sufficiently simple choice to the user.

26 At least at the time of Scaife and Rogers’ writing.
observations have been made of the mental structures that govern our use of money. *Mental accounting* (Thaler, 1985) is probably the most well-known (Belsky & Gilovich, 2010). Mental accounting describes how people mentally assign their money to implicit categories and only think about each category in isolation, sometimes in inconsistent ways; worrying about pennies in one category while spending frivolously in another. This violates the rational principle that money is fungible, or interchangeable; a unit of currency is worth the same whether it came from a lottery winning, a gift, a wage earned through labor, a tax return, or an unexpected saving on a planned purchase that turned out to be cheaper than expected. Yet, people regularly behave as if their categories are separate and money cannot be moved from one category to another. Heath and Soll (1996) illustrate mental accounting through the following colorful and relatable anecdotes:

Mr. P recently went shopping for a pair of slacks. When he could not find any slacks he liked, he spent a similar amount of money on a sweater that he normally would not have purchased. Ms. C reluctantly declined a Sunday invitation to dinner because she had "spent too much money" on tickets to the theater two days earlier. On questioning, she admitted that she would have enjoyed the dinner and was well able to afford it, but she felt compelled to decline because of the earlier theater expense. Neither person was entirely content. After her experience of "perilous" poverty, Ms. C found herself declining an enjoyable dinner with friends, and after his experience of "triumphant" wealth, Mr. P found himself the owner of an unnecessary sweater.

In the example, Mr P. had already “set aside” money for buying clothes and, instead of saving the money and mentally putting it back in his stash of general cash, he spent it on something unexpected from the same category. Ms. C felt guilty about spending too much on entertainment, but she might not have felt obligated to abstain from the dinner with friends had she spent money equivalent to the price of the theater tickets on a visit to the dentist instead.

The theory does not assume that mental accounts are fixed and stable; on the contrary, there is some evidence that we make up the categories as we go along – that mental accounts are constructed and deconstructed over time, in the moments when we make purchasing decisions (Cheema
2. BACKGROUND

This is especially true for items that are not easy to classify; for example, a restaurant dinner with friends could be classified as either “food” or “entertainment” (Cheema & Soman, 2006).

The “purpose,” or evolutionary advantage, of mental accounting is debated. Several researchers (Heath & Soll, 1996; Shefrin & Thaler, 2004; Prelec & Loewenstein, 1998) see mental accounting as a form of self-control; it limits a person’s spending on luxuries. At the same time, some experiments show that people construct their mental accounts in a way that “allows” more spending and makes them feel less bad about their purchases (Cheema & Soman, 2006; Shafir & Thaler, 2006). Perhaps mental accounting is simply a bounded rationality problem; the human brain has limited capacity for remembering and weighing all kinds of different purchases against each other. Separating our money into silos makes it easier.

Thaler, who coined the term, is reluctant to say whether mental accounting is good or bad (Thaler, 1999). It can be viewed as a cognitive deficiency that makes us behave irrationally, or as a useful tool of the mind that simplifies a complex task and helps us make sense of our finances. While the literature is full of examples of real or hypothetical problems stemming from mental accounting behavior, it has been suggested that designers of computer-based financial applications perhaps should not try to steer their users away from their natural tendency to do mental accounting (Ariely, 2013). Instead, mental accounting can be harvested as a positive trait that helps people stay focused and disciplined (Egan, n.d.; Ariely, 2013). For example, people are reluctant to spend money that has been set aside in a savings account (Thaler, 1999; Shefrin & Thaler, 2004). This can be exploited as a self-control device; for example by setting up an automated transfer to a savings account each month, only leaving enough cash in the checking account that can be safely spent.

That said, mental accounting does lead to irrational behavior. A classic example is the person who saves some money each month into an account with low interest and never touches their savings, but at the same time uses a credit card with high interest. That person would save money if

---

27 Bounded rationality: We act rationally with the information that we have, but our rationality is limited (bounded) by the amount of information our brains can process (Simon, 1972).

28 For the interested reader, there are several papers about self-control devices in behavioral economics literature. See e.g. Thaler & Shefrin (1981) and Ashraf, Karlan & Yin (2006).
they were to take money out of their savings when they are short on cash, instead of using credit. But the person probably thinks of consumption as separate from saving (Ariely, 2013). In all honesty, this behavior might not result in losses that are worth fretting over. But there are worse cases. Mental accounting has been observed among households in the Swedish owner-occupied apartment market (Almenberg & Karapetyan, 2009), where it skews apartment prices and prompts consumers to take on more debt in the tax-inefficient form of co-operative debt. Despite the fact that real estate constitutes over 70% of the typical Swedish households’ total assets (Campbell, 2006), people may not be thinking of their home as part of their investment portfolio. This can lead to people fretting over details in their stock portfolio while ignoring much bigger losses or opportunity costs related to their mortgage or home equity (Bolmeson, 2015).

2.2.3 ANCHORING
Another commonly used concept in behavioral economics is anchoring. Anchoring occurs when people are asked to give a numerical answer to a question to which they are unsure of the answer. People in that situation are easily influenced by random numbers to which they are exposed, even when they are aware that the numbers are random (Tversky & Kahneman, 1975; Ariely, Loewenstein & Prelec, 2003). In one experiment, participants were shown a bottle of wine along with a description of the wine’s country of origin, age, and other qualities, but no information about the wine’s retail price. They were then asked how much they were willing to pay for one bottle of that wine. At the start of the experiment, the participants were asked to write down the last two digits of their social security numbers. Participants with the highest social security numbers were willing to pay three times more for the same bottle of wine than participants with the lowest social security numbers (Ariely, Loewenstein & Prelec, 2003)²⁹.

The researchers’ explanation for this remarkable effect was that the mere exposure to numbers provided participants with a so-called anchor point, something against which to compare the price of a bottle of wine,

²⁹American social security numbers are not connected to birth date, so the effect was not due to age differences. The social security numbers simply served as “random” numbers. The experiment has been replicated with lottery wheels instead of social security numbers. The results were confirmed.
even when the participants must have known that their social security number had nothing to do with the price of wine.

It does not seem like a far leap to assume that abstract questions like “How much return do you want on your investment?” and “How much loss can you accept?” are similarly easy to influence through exposure to irrelevant numbers.

Like graphical perception theory, behavioral economics has come under criticism for not being an actual theory, but rather a collection of observations with no unifying explanation (Harford, 2014; Fiedler & von Sydow, 2015). As such, it does not offer much in ways of explaining underlying processes or predicting under what conditions its results will hold. Additionally, it has been argued that the typical experiments in behavioral economics are too artificial and devoid of context, such that the researchers who designed the experiments cannot really claim that their preferred outcomes are more rational than the “deviations” in judgment displayed by participants (Fielder & von Sydow, 2015).

Perhaps the harshest critic of behavioral economics is the German psychologist Gerd Gigerenzer (Fielder & von Sydow, 2015). Gigerenzer and Brighton (2009) showed empirically that some of the “suboptimal” strategies disdained by behavioral economists actually perform better than more sophisticated strategies under certain conditions. This breaks the assumption that if only we humans had more memory and more computational power in our brains, we could make better decisions by understanding statistics better. According to Gigerenzer and Brighton, this is because of a phenomenon called overfitting; the complex and statistically sophisticated strategies promoted by behavioral economists are good at describing data in hindsight, but are not good at predicting the future because they include too much noise in their models (Gigerenzer & Brighton, 2009). For a person making a decision, argue the authors, being able to make predictions about the future is much more important than being able to describe historical data in minute detail.

Gigerenzer has gone on to direct the Center for Adaptive Behavior and Cognition, where his research team studies what decision strategies perform best under what conditions, rather than assuming that certain

---

30 Different mechanisms have been suggested for how anchoring might come into play, but the existence of the effect itself has strong empirical support (Fielder & von Sydow, 2015).
strategies are always superior. He calls this concept ecological rationality and it has garnered some attention in HCI (Rogers, 2012).

However, behavioral economists do not claim that the heuristics employed by our brains are always suboptimal. It is precisely because they usually work well that they have become common heuristics.

2.2.4 INTERTEMPORAL CHOICE

Early economic models of intertemporal choice assumed that people are indifferent between a reward now and a similar reward in the future, discounted for the time until the reward. Empirical studies have shown that this is not how people actually behave; people have different intertemporal preferences depending on whether the temporal difference is near or far in the future (Laibson, 1998). The closer in time, the stronger is the discount effect (Thaler, 1981). For example, a person may prefer receiving $100 today over $110 tomorrow, but at the same time prefer $110 in 31 days over $100 in 30 days. The effect is stronger for small rewards than for big rewards, perhaps because people find it more worthwhile to be patient when more is at stake (Thaler, 1981). Additionally, the effect is stronger for gains than for losses; when asked if they would prefer to pay a smaller fine now, or a larger fine later, the preference for immediate gratification diminishes; most people would rather pay the smaller fine now (Thaler, 1981).

The relevance to retirement saving is obvious. Retirement saving is an example of behavior that entails abstaining from immediate gratification in the form of consumption, in order to have more disposable income later in life. Failure to save enough, even when one could afford it, is often described as a self-control problem (Thaler & Shefrin, 1981; Laibson et al., 1998). The self is modelled as consisting of multiple “selves” with different preferences, battling over what decision to make. Thaler and Shefrin (1981) likened the mind to a “planner” and a “doer”; the planner is farsighted and tries to impose different strategies of self-control on the doer, while the doer is impulsive and seeks immediate rewards without regard for the long-term consequences. The metaphor is analogous to Kahneman’s (2011) rational System I and impulsive System II.

---

31 Note that it is rational to prefer a reward now over an equal reward in the future; postponing the reward incurs some risk that the reward will not happen. In hypothetical experiments such as Thaler (1981), participants are asked to ignore that risk and pretend that they are certain to always get the reward. However, in real life, nothing is 100% certain. Additionally, a reward received now can be invested and start accumulating interest immediately.
Sophisticated people are aware that they have self-control problems and that their future preferences will also be present-biased; for example, that they are unlikely to start better habits (saving, dieting, exercising) tomorrow if they did not start today (O'Donoghue & Rabin, 1999). Naïve people believe that they will be able to stop procrastinating in the future, after just this one last time. Being aware of their own self-control problems allows sophisticated people to employ self-control strategies that shackle their future impulsive selves, which allows them to behave as if they did not have self-control problems. For example, they can join health clubs where they pledge their commitment to weight loss in front of others, join an addiction treatment program, or start an automated transfer to a savings account.

One such self-control strategy is to exploit the effects of mental accounting. People tend to think of their wealth as divided into current income, current assets, and future income, and are much more willing to spend current income than current assets or future income (Shefrin & Thaler, 1988). By having the “planner” set up an automated transfer from current income (this month’s paycheck) to a savings account from which money cannot be easily withdrawn, the “planner” makes less money available to the “doer” for spending. When experts in the USA suggested that the penalties for early withdrawal from retirement savings be lax ed so that people could have more freedom in how and when to spend their own savings, 60% of American consumers said they would rather keep the restrictions, out of fear that the money would be spent on the wrong things (Farkas & Johnson, 1997).

Programs designed to have people commit in advance to saving have shown success in increasing savings rates (Thaler & Benartzi, 2004; Ashraf, Karlan & Yin, 2006). The increase in savings seems to be long-lasting and drop-out from the programs is low.

### 2.2.5 Visualizing Risk and Uncertainty

The problem of planning for retirement involves making decisions that involve some risk which has to be weighed against a potential reward, as well as understanding how risk and return plays out over long periods of

---

32 O’Donoghue and Rabin (1999) argue from a theoretical model that this will only be true in certain situations; in other situations, sophisticated people can be even worse off than naive people.

33 It is not clear from the report whether consumers feared that they themselves would not be able to resist the temptation, or whether it was mostly a concern that other people would not.
time. These are complex relations which are not intuitive to the human brain and it has been suggested that visualization might be a good way to explain to non-expert people how risk works (Chandra, Krovi & Rajagopalan, 2008).

The problem of communicating the trade-off between risks and benefits in a choice situation is not unique to finance. Most studies on how people understand risk graphs are probably found in health informatics. Health informatics is concerned with explaining to patients of varying numerical and cognitive ability how different treatments will affect their chances of recovery and at the same time, their risk of side effects. This has obvious parallels to investment strategies which may have different chances of succeeding and also different consequences in terms of gains and losses.

Spiegelhalter, Pearson, and Short (2011) conducted a literature review on studies of risk communication to laypeople. They found that this field is still in its infancy and that there are few known rules and best practices. One general recommendation is to present both the upside and the downside – in other words, rather than “8 of 10 people recover,” say “8 of 10 people recover and 2 of 10 people do not recover.” Presenting only the up- or downside has been shown to bias people’s reaction to the information.

Parrot et al. (2005) found that visual graphics are not always the easiest format to grasp for people untrained in quantitative reasoning. The participants in their study understood verbal descriptions of statistical information better than visual graphics. Additionally, Gresh et al. (2011) found that some people have problems with the concept of a probability curve. The so-called fan chart (see Figure 7), a chart that shows a forecast that gets wider as the future gets more uncertain with time, has been shown to bias people
toward too much optimism; they expect the actual outcome to be in the upper part of the curve (Tak, Toet & van Erp, 2015). This is especially true for people of lower numerical ability (ibid).

It is important to remember that risk and uncertainty are two distinct concepts that are processed differently mentally (Ellsberg, 1961). Risk has known probabilities, such as a lottery with a known number of tickets, while uncertainty has unknown probabilities (Knight, 1921). Hence, though finance experts often use the word “risk,” it is actually uncertainty that is at play in financial markets. People are generally less willing to accept uncertainty than risk (Ellsberg, 1961).

To make matters more complicated, there are different ways to evaluate visual graphics and no one type of graph scores best on all possible tests. Ancker et al. (2006) reviewed graph studies in health care and found that the graphs that best support quantitative reasoning are not the best at inspiring behavior change. Presumably this is because the graphs that inspire behavior change do so by exaggerating risks. Additionally, the graphs that people prefer are not the ones that help them make the best decisions. People tend to like for graphs to be simple, but overly simplified graphs can lead people to draw the wrong conclusions.

Several researchers have had the idea that findings from behavioral economics ought to have implications for computer system design. The approach is usually to take a known bias and try to come up with a design solution that will counter the bias. For example, Gunaratne and Nov (2015) designed a portfolio picker that reduced the impact of the endowment effect and made people pick more efficient portfolios, while Inbar (2007) found that visual representation of risk reduced the impact of the certainty effect but not the Ellsberg paradox.

---

34 The experiment by Tak, Toet & van Erp (2015) was in not in finance but in meteorology. Their participants were asked to interpret a weather forecast shown in the form of a fan chart. I personally see no reason why the biasing effect of the fan chart would not generalize across domains, but the critical reader may remain unconvinced. Risk preference has shown to be heavily domain-dependent; for example, a person who enjoys parachuting does not necessarily enjoy taking financial risk (Weber, Blais & Betz, 2002). However, risk preference is a subjective inclination, like taste in food or music. It is thus a completely different concept from the ability to accurately read a chart without over- or underestimating the risk presented quantitatively in said chart.

35 Endowment effect: People tend to overvalue things that they own, compared to things that they do not own. (Marzilli Ericson & Fuster, 2014).

36 Certainty effect: People prefer a certain outcome over a gamble, even when the gamble has a higher expected outcome.

37 Ellsberg paradox: People prefer known odds over unknown odds, even when the known odds are poor and the unknown odds may well be better.
One question which has garnered some attention from researchers is whether investment account overviews should aggregate individual investment products (such as different mutual funds and stocks) into one portfolio and track its ups and downs, requiring the user to click down a level to see the individual products, or keep the products separated and show their individual performance upfront. This could be seen as a form of mental accounting; should users be encouraged to think of their investment portfolio as a whole, or as separate products whose success are to be evaluated independently of each other? Prospect theory\(^{38}\) predicts that users might react impulsively to negative change in any of the individual products and make rash decisions to sell. It also predicts that people who see aggregated returns will be willing to take more risk, since they risk seeing fewer potential “red numbers.” Shavit et al. (2010) seemed to confirm the theory when they found that people looked more at the performance of individual products than the performance of the overall portfolio when they were given an account overview and an eye-tracker recorded their gaze. On the other hand, Beshears et al. (2011) followed real mutual fund investors over a year and found that people who saw aggregated portfolio performance did not take discernibly more risk than people who saw individual product performance. The authors suggest that previous lab studies may not have taken in account all factors that go into real decisions.

A related question is whether the time period over which performance is measured will affect people’s decisions. Here the theory predicts that people who see their investment’s performance over a longer time will take more risk, since time tends to smooth out volatility. Beshears et al. (2011) disconfirmed this hypothesis, too. They note that earlier studies, which supported the hypothesis, were usually conducted in the lab with hypothetical money while Beshears and his co-authors tracked real investors managing their own money. They found no difference between investors who saw one-year charts versus investors who saw five-year charts. Beshears et al. suggest that people may be more cautious and less impulsive than the hypothesis assumes.

\(^{38}\) Prospect theory is one of the fundamental concepts in behavioral economics. It postulates that people evaluate their current state in terms of a loss or a gain compared to a historic reference point, and that losses have a greater emotional impact than gains (Kahneman & Tversky, 1979). Despite some later developments, notably the discovery that people make different choices when they get a description of the odds of winning or losing versus when they learn the odds from their own experience, the theory is still considered robust and empirically supported (Newell, 2015).
2.2.6 **Persuasive Design and Nudging**

*Persuasive design* is a school of thought pioneered by psychologist BJ Fogg (Fogg, 2002). Fogg had observed that people would often be “polite” to computers and that when they had perceived a computer as helpful, they were more willing to answer a survey prompted by the computer’s interface. Politeness and reciprocity are social functions, and this gave Fogg the idea that computers might sometimes function as social actors. If they are social actors, they might have the power to influence people in their “social network.” Thus, Fogg thought, computer systems might be able to change people’s opinions, attitudes, or behavior.

Prior to Fogg, the goal of user interface design was often thought to be to make it easier to perform a task, such as buying a product from a website. With persuasive design came the idea that user interface design could also increase users’ willingness to perform the task in the first place; it could make people more motivated to buy the product. *Gamification* is a special type of persuasive design, by which designers apply game mechanics to non-game contexts\(^{39}\).

While “persuasive” systems existed before Fogg, the term itself has surged in popularity and there is now an entire academic conference dedicated to persuasive technology.

A closely related idea in behavioral economics is called *nudging* (Thaler & Sunstein, 2008). The term denotes giving someone a “nudge” toward a particular choice without taking away their ability to make alternative choices. Nudging is thus used when the person who designs the choice situation, called a choice architect (Thaler, Sunstein & Balz, 2014), believes that they can help people make more rational choices than they otherwise would, but does not want to forbid them from making “less rational” choices (Carlsson, 2016)\(^{40}\).

The main difference from persuasive technology researchers is that behavioral economists do not necessarily focus on interactive technology; they usually study rules and policies such as national organ donor registries (Johnson & Goldstein, 2003) or occupational pension plans (Choi, Laibson & Madrian, 2004). One of the most powerful tools of nudging is the default option; unless people actively choose otherwise, the

---

\(^{39}\) The idea is that games are intrinsically motivating, and thus gamified applications and processes will increase user engagement if implemented correctly (Robson et al., 2015; Werbach, 2016).

\(^{40}\) The concept is called libertarian paternalism (Thaler & Sunstein, 2003).
“desired” choice is automatically selected for them (Thaler & Sunstein, 2008).

After a decade of persuasive design research, Fogg himself seems less convinced that computers can really change people’s opinions. Nowadays he has softened his claim to say that computers can encourage behavior that the person already agrees with (Fogg, 2012).

The thought of manipulating people’s choices through subtle design clues may raise suspicion. Among academics, it has been suggested that persuasive technology is only ethical if it is used to encourage people to do things that are in their own best interest (Berdichevsky & Neuenschwander, 1999). The thinking is the same in behavioral economics; nudging is only to be used for encouraging good choices for the chooser (Thaler & Sunstein, 2008). However, Thaler and Sunstein (2010) point out that no design is neutral and that it could be argued that the very knowledge that every design will influence people in some way creates a moral imperative to try to make that influence positive.

Berdichevsky and Neuenschwander (1999) argue that persuasive technology should follow the moral golden principle: Only persuade users to do things which the designers themselves would agree to be persuaded to do. In other words, they argue that persuasive systems should be strictly user-centric.

However, a strict user-centered perspective is not the only possible moral stance. At a seminar on nudging and the Swedish retirement system, Flodén (2016) raised the point that a choice architect must decide whether to focus on the interests of the individual, or the interests of society. The two can sometimes be nice enough to coincide, but not always. In the case of the Swedish retirement system, a person who has a short life expectancy, perhaps due to poor health, and high income and savings will probably be better off taking out most of their retirement funds as a lump sum at the beginning of their retirement, because that allows them to spend the money or bequeath it. On the other hand, a person with long life expectancy who will receive a low pension is probably better off annuitizing their funds, so that they are guaranteed an

---

41 Carlsson (2016) cautions that the “nudger” must be certain that they know better than the person they are trying to nudge. In the case of planning for retirement in Sweden, Carlsson points out that this isn’t always easy to know. For example, it isn’t rational for everyone to save more. Having too much capital at retirement can cause some people to lose their entitlement to government-provided housing allowance.
income even when they outlive their own savings. Should a retirement planning system advise people on how they can squeeze the most money out of the government? If everyone does what is the most rational for them personally, the retirement system will come under greater pressure and the risk increases that the system will not be able to fulfill its promises to everyone.

Thus, as Flodén (2016) points out, policy makers must decide whether it can be justified to “sacrifice” some individuals in order to protect stability for most other individuals. The same is true for an advisory system or any persuasive system; it is essentially a policy in algorithmic form.

A framework better equipped to handle goal conflicts than persuasive technology ethics is value-sensitive design (Friedman et al., 2013). Value-sensitive design methodology is based on an iterative process of empirically studying the goals and interests of all stakeholders to a system, not just its direct users, before making design decisions.

I have not made a systematic analysis of stakeholder values in retirement planning systems, but it could make an interesting future project. I will simply suggest that perhaps a parallel could be made to an innovative approach to the ethics of self-driving cars. Bonnefon, Shariff and Rahwan (2016) noted a current debate on whether an autonomous vehicle that finds itself in a situation where an accident is unavoidable should steer itself so as to always save its own passengers, or whether it should be programmed to save the most lives, even if the latter means driving off a bridge to save a group of pedestrians who are walking in the middle of the road. The authors decided to use a large survey to gauge the opinions of the general public; a sort of democratic approach to ethics.

Interestingly, most people favored the “utilitarian” vehicle when thinking about it in the abstract, but said they would themselves not buy a self-driving car unless it was programmed to always protect them as passengers. It is not unthinkable that this is how people would feel about advisory systems too; the user only trusts the system because they expect the system to be user-centric. It has been pointed out that “the most ethical decision may be the one that gets the most [self-driving cars] on

---

42 Everyone in the Swedish retirement system is guaranteed some income, but the minimum benefit is relatively low and some consider it to be below the poverty line.  
43 This is essentially a variant of the so-called trolley problem.
the road” (Lubin, 2016). If fewer people trust self-driving cars, fewer people will buy them; hence the roads will still be full of human drivers, who are much more accident-prone. Perhaps the same argument could be made for retirement planning systems: If fewer people trust them, fewer people will use them, leading perhaps to more people having too little savings and becoming more dependent on government assistance, thereby increasing the burden on the retirement system as a whole, compared to a situation where more people are helped to maximize their benefits and more people therefore trust the retirement planning system.

For the remainder of the thesis, I will assume that retirement planning software is usually user-centric, for three reasons: 1) The self-driving car analogy above 2) Users are likely to believe that a planning system is designed for their best interest and hidden motives would be deceptive 3) It follows conventional standards for what is otherwise considered good practice for retirement planning software, namely to help users make the most tax-efficient choices (Turner & Witte, 2009). That a utilitarian planning system would “fool” people into paying more taxes for the sake of the greater common good seems absurd.

2.3 PREVIOUS WORK ON FINANCIAL PLANNING SYSTEMS

2.3.1 FINANCIAL DECISION AIDS IN ACADEMIC LITERATURE

I found early in my PhD studies that academic studies on financial planning software were relatively rare and that the finance domain seemed underexplored. This chapter is an attempt at an exhaustive review of academic design papers about financial decision aids. By “design papers,” I mean papers where the authors have designed a prototype or a theoretical model of a financial decision aid, with a specific intended effect on its users. The papers were found by searching for relevant keywords in academic databases and by reading every abstract on financevis.net. Only decision support systems for novice users are

44 According to the stated goals of the cars’ designers.
45 Would this not result in the same situation – many retirees who depend on the government? I argue no, because while the system would show the user that they can be entitled to government assistance if their income and savings go below a certain point, there are two benefits to saving instead of relying on assistance: 1) The level of assistance in the future is uncertain and depends on the political climate and the economic development in the country. 2) Increased savings can be used to raise living standard, while government assistance typically only covers a rather minimal living standard.
included. Stock market visualizations, tools targeted at professional specialists such as fraud detectors, and literature reviews are excluded. I also excluded papers that measured the effects of different presentation formats, but did not have a normative design goal (e.g. Beshears et al., 2011).

None of the papers I found described systems that were intended to support the entire process of financial planning; most were portfolio-pickers. The designs were aimed at one or more of the following design goals (summarized in Table 2):

**Design Goal 1.** Help people maximize the return on their portfolio without particular regard for their personal risk preference. For example, in the system presented by Chandra, Krovi and Rajagopalan (2008), users were meant to readjust their risk preferences when they saw that they could have achieved better results had they chosen a different risk profile in the past. This evaluation method can be accused of retrofitting; the results depend on past developments on the (real or simulated) stock market but do not necessarily predict success in the future.

**Design Goal 2.** Help people choose a portfolio that matches their risk preference. The way to do this is usually by trying to find a better way to explain risk, so that participants can understand the consequences of their choices. The success of the system is measured through risk comprehension questions, satisfaction scores after the participants have seen the outcomes of their choices, or the internal consistency of the participant’s choices. People who act like they have wildly inconsistent preferences have probably not quite understood what they are doing.

**Design Goal 3.** Help people design portfolios on the efficient frontier. Recall that portfolios on the efficient frontier are considered to give the best return for a given risk. Thus, systems of this type are aimed at visualizing the efficiency of the user’s portfolio and the success of the systems is measured in terms of the efficiency of the portfolios created by study participants. One model (Csallner et al., 2003) was only designed to help people maximize their diversification without regard for efficiency and as such, it might prompt people to use the disputed 1/n heuristic46.

**Design Goal 4.** Help people diversify their portfolios, by visualizing the level of diversification in the user’s portfolio.

46 The 1/n heuristic is when an investor divides their funds evenly across all available options, without regard for the resulting risk.
Design Goal 5. Make people feel more confident, satisfied, or engaged while trading and designing their portfolios.

Table 2: The papers reviewed in this chapter.

<table>
<thead>
<tr>
<th>Design Goals</th>
<th>Maximize return</th>
<th>Consistency of risk preference</th>
<th>Risk comprehension</th>
<th>Diversification/Portfolio efficiency</th>
<th>Subjective feeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bateman et al. 2010</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chandra et al. 2008</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Csallner et al. 2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Goldstein et al. 2008</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Gunaratne &amp; Nov 2015*</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haisley et al. 2010</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Inbar 2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Rudolph et al. 2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Savikhin et al. 2011</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Sundali &amp; Guerrero 2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
</tr>
</tbody>
</table>

*Gunaratne and Nov (2015) set a moderate savings goal for their participants, so their study was not focused on return maximization per se. It measured whether participants reached the goal, but did not encourage them to take excessive risk.
None of the reviewed systems attempted to contextualize the risk and return of portfolios in terms of the user's needs and objective risk tolerance\(^{47}\). Participants were thus assumed to come to the experiment with an established view of their own risk preference and required return. However, the view that people have stable preferences that can be measured in experiments has been challenged by economists who instead argue that preferences are often constructed in the experiment situation (Ariely & Norton, 2008; Ariely, Loewenstein & Prelec, 2006). Goldstein, Johnson, and Sharpe (2008) were the only ones in this collection of articles to say that risk preferences may be constructed at the moment the user sees the visualization, rather than a data point that can be measured, and that the job of the visualization tool is to help the user construct preferences that are usable for the task at hand\(^{48}\). However, their prototype still asks users to construct preferences out of “thin air” – without grounds in income, assets, or debt. In all the reviewed papers, the participant’s stated risk preference is treated as the end-all be-all of what risk they ought to take, with no regard for other factors. The user’s gut feeling is thus what forms the base for financial advice in these systems.

Systems that measure their success in terms of the acquired return on participant’s portfolios can be problematic, especially if the return is not adjusted for risk. Pension economists have criticized contemporary visualization tools for being too narrowly focused on maximizing wealth while ignoring important risks that are relevant to the user’s decisions (Merton, 2003; Bodie et al., 2004). Retirement savings should not at all be about trying to maximize wealth (Merton, 2014; Merton, 2003; Bodie et al., 2004; Hatcher, 2003). Wealth-maximizing strategies tend to lead to ill-considered savings plans (Merton, 2014), where people choose a contribution rate based on heuristics such as saving whatever is the default rate in the system or saving a certain fraction of their income\(^{49}\) (Benartzi & Thaler, 2007). It is easy to monitor a person’s adherence to such a strategy and their progression toward their goals, which creates a feeling that the person is in control and on the right track; but because all

---

\(^{47}\) By “objective risk tolerance,” I mean whether a loss would have palpable consequences for the user. Losing 100,000 Swedish crowns can be catastrophic for some, and trivial for others, depending on their circumstances.

\(^{48}\) Goldstein, Johnson, and Sharpe also suggest that risk preference questions that ask people how they feel about one-time gambles may not be an accurate representation for how the same people want their retirement funds invested over long periods of time.

\(^{49}\) Typically a multiple of 5% (Benartzi & Thaler, 2007).
goals were chosen arbitrarily, it does not tell the person whether they are really saving enough. Instead, retirement saving should be about making sure that the person can still pay their expenses after they stop working (Merton, 2014; Merton, 2003; Bodie et al., 2004; Hatcher, 2003; Hamermesh, 1982).

Additionally, some problems are more suitable for visualization than others. Sedlmair et al. (2012) suggested a framework for analyzing whether visualization would be helpful in supporting a given task. Tasks that require some intuition to solve, and where the process for solving the task cannot be formalized into step-by-step procedures, are not good candidates for visualization. It is difficult to imagine a system that shows us visually how to paint an original masterpiece or write a best-selling novel. On the other hand, tasks that are so well-defined that a computer can be programmed to perform the task on its own are not good candidates for visualization either, because there is no reason to involve a human.

We see some design objectives in Table 2 where visualization was perhaps not the best solution. The efficiency of portfolios is one such objective. The efficiency is calculated by combining large amounts of data and comparing many different combinations, which is a task where computers clearly outperform humans. Since the computer has already determined which portfolios are the most efficient, in order to visualize said efficiency, what is the point of then having a human select among some efficient and several less-efficient portfolios? Compare this to Don Norman’s notion of design constraints; if there is something we wish to stop the user from doing, the design should make it impossible to do it (Norman, 2002). We have solved the problem of people ruining electronics by inserting cables the wrong way by designing the plugs and sockets such that only the right plug fits in the right socket. Why should then a financial visualization tool allow users to choose portfolios that are not efficient? Instead of having efficiency be a parameter to visualize, so that users can choose between degrees of efficiency when we know that all users want optimal efficiency, efficiency should be a hygiene factor. In other words: Only portfolios that are efficient should be presented to the user. The focus should instead lie on helping people understand the consequences of the different risk levels that they can choose between. Selecting a suitable risk level is, unlike efficiency, a task that can never be fully delegated to a computer. The human must always answer what their
goals are and what possible negative consequences they are willing to accept.

Another factor that would be a better candidate for automation is diversification. Diversification of assets, especially across sectors whose fluctuations correlate as little as possible with each other, is an important principle in reducing the overall risk of a portfolio. As previously mentioned, many people do not diversify their assets enough. However, visualizing diversification in order to make people diversify more is a suboptimal solution because 1) It is a simple operation that can easily be done by a computer and 2) Selecting the best diversification strategy is actually not something humans are good at. People tend to either invest too much in their own home market (Coval & Moskowitz, 1999), or go by the unsophisticated 1/n rule (Benartzi & Thaler, 2001). The 1/n rule, where n is the number of investment products such as stocks, bonds, and mutual funds that the user can choose from, means that the user simply takes their money and invests an equally large share of it in all of the available options. The resulting portfolio is simply a haphazard mix of whichever investment options were available at the time, which has nothing to do with the user’s risk preference or goals. An experiment where people were asked to choose between a set of mutual funds to invest their retirement savings in, showed that people who were shown five equity funds and one fixed-income fund ended up choosing much riskier portfolios than people who were shown one equity fund and four fixed-income funds (Benartzi & Thaler, 2001). This was simply because participants divided their money equally across all available funds without considering the differences in risk. It could be noted that this concept is closely related to efficiency; the efficient frontier shows the optimal ways of diversifying assets and it is usually not in 1/n proportions.

When portfolio efficiency is a hygiene factor, meaning that only efficient portfolios are possible choices, the problem of diversification is also automatically solved. Other hygiene factors are an aesthetically pleasing design, simple navigation, visual clarity, and constraining the selection of investment products to low fees\(^{50}\). However, I choose not to

---

\(^{50}\) While future return is uncertain, the fee is known and is therefore one of few factors that can be optimized with near certainty. Products with higher fees generally do not perform better than products with lower fees (Bazerman & Moore, 2012).
focus on visual design and navigation in this thesis because they are general HCI problems and not specific to financial planning systems.

To summarize, a financial decision aid should not primarily visualize parameters that can be better used as filters, such as portfolio efficiency\textsuperscript{51}. Instead, the decision aid should focus on helping the user choose an appropriate level of risk.

2.3.2 RETIREMENT PLANNING SOFTWARE IN THE INDUSTRY

Based on my experience as an interaction designer, I suspected that the industry might be ahead of academia in terms of innovation in financial services. I therefore decided to include a review of commercial systems as well. Here I had the opposite problem from when I conducted my academic review: There were so many systems to review that I had to limit my search. I decided to search for retirement planning software and exclude simpler systems such as portfolio pickers.

While searching for retirement planning software, I also came across some systematic reviews of such software which I had not previously discovered. These reviews were not made by HCI researchers, but by economists or actuaries. Their aim was typically to evaluate whether the software gave accurate advice. A common method was to find a sample of (sometimes free, sometimes a mix of free and premium) online retirement calculators and feed all calculators the same data in order to see how the different calculators responded. Judging whether retirement software gives accurate advice is no simple task. There is often no truly correct answer, since there are many factors involved where experts disagree on what the proper advice ought to be (Turner & Witte, 2009).

In one of the first studies, Ciccotello and Wood (2001) simply looked at whether the different calculators gave consistent advice. They found that the advice was reasonably consistent – or at least not less consistent than the advice from human advisors. However, the human advisors in their sample were graduate students in personal financial planning and might not have been representative for experienced financial advisors. The authors also looked at whether the calculators were better at handling any

\textsuperscript{51} Not everyone agrees that portfolio efficiency is a valid concept, because it is based on the historic performance of assets that may not perform similarly in the future. However, a system that visualizes portfolio efficiency must be designed from the assumption that portfolio efficiency is relevant, otherwise why visualize it? Hence, systems that visualize portfolio efficiency are pointless whether or not you believe in portfolio efficiency. The only valid purpose I can think of for such systems is as educational material about portfolio efficiency as a concept.
certain type of customer and found that they gave more consistent advice to people of low to middle net worth, who tend to have less complex finances and fewer unique issues than wealthier people.

Later, economists have provided more in-depth critiques of retirement planning software. Kotlikoff (2008) even challenged conventional financial planning theory while doing so (Turner & Witte, 2009). The main points of his and others’ critique are recounted in this chapter. In general, researchers find that retirement planning software does a poor job of including all the relevant risks faced by retirement planners and retirees (Turner & Witte 2009; Sondergeld et al., 2003; Bodie, 2003; Kotlikoff, 2008; Dowd & Town, 2008; Bi, 2015; Turner, 2010a; Turner, 2012).

2.3.2.1 DETERMINISTIC VS STOCHASTIC
There are two main types of retirement planning software: Deterministic and stochastic (Turner & Witte, 2009). Deterministic software makes assumptions, for example that the stock market is going to return 8% per year and that the yearly inflation rate is 3%, and then projects a future based on those assumptions. By assuming that everything will follow the historic average and presenting a single outcome as if it were certain, deterministic software does not demonstrate risk to the user; in fact, it hides risk (Sondergeld et al., 2003). In order to fulfill the basic requirement that the software explain to the user that the prognosis is uncertain, deterministic software often runs different scenarios under varying assumptions, for example one based on the historic average and one where the stock market does worse than usual (Turner & Witte, 2009). The software can then present a range of possible outcomes. Still, one weakness of deterministic software is that it assumes that every year will have the same growth as the average growth. In fact, actual annual growth varies widely around the average and the order of good and bad years matters. If an investor gets a streak of bad years in the beginning, it can have a devastating effect on their capital (Bodie, 2003). In other words, deterministic software might underestimate market risk. It is prudent for developers of deterministic software to make conservative assumptions in order to compensate for such systematic biases (Turner & Witte, 2009).

Stochastic software is based on Monte Carlo simulations or pseudo Monte Carlo. In a Monte Carlo algorithm, a variable such as stock market growth varies randomly within a range. The algorithm generates a
random result for the variable and uses that to make a prediction for the future. It does this hundreds or thousands of times. It then calculates the number of times that each future occurred and based on this, it calculates the probability of some scenario occurring, for example the probability that the user’s total capital reaches a target amount in 30 years. A quirk of Monte Carlo simulations is that the user can get different results when running the software twice with exactly the same input. Pseudo Monte Carlo methods handle this by generating the same value for the variable each time.

Stochastic software is considered better than deterministic software (Turner & Witte, 2009), but it too is susceptible to some biases. It is important to remember that Monte Carlo simulations are only as good as their assumptions and that the typical assumption is that the future will resemble the past. Additionally, Monte Carlo simulations are usually based on the assumption that variables such as stock market growth are normally distributed (Turner & Witte, 2009). However, it has been suggested that annual stock market development rather resembles a fat-tailed bell curve; or, in other words, that very high growth and very bad downturns are more common than implied by the classic normal distribution (Turner & Witte, 2009). Thus, even stochastic systems may underestimate the risk of bad stock market performance. Additionally, some have claimed that Monte Carlo systems that run too few simulations (for example, fewer than 1,000) underestimate risk due to there being too few bad cases in their sample (Turner & Witte, 2009).

A big advantage of stochastic systems over deterministic systems is that the former are better at simulating the actual effects of variability in outcomes. They can calculate the probability of worst-case scenarios where several adverse events occur simultaneously, such as changing interest rates and declining stock prices. In practice, though, most retirement planning programs treat only stock market growth as a

---

52 It could be noted that this trick doesn’t make the prognosis more true, only more convincing.

53 The probability function for the stochastic variable must come from somewhere. For example, it can be based on a table of historic data.

54 Certain events tend to be correlated. For example, there is a historical correlation between the price of small-cap stocks and long-term bonds. Several Monte Carlo-based retirement calculators base assumptions on these historical correlations. However, it has been observed that asset classes tend to correlate more during market crashes, presumably because they all go down. In other words, the normal assumptions seem to break down in times of crisis (Turner & Witte, 2009).

55 Rising interest rates are bad for people with debt, but falling interest rates are bad for people who are planning to live on interest-generating investments.
stochastic variable, while other important variables are still treated as deterministic (Turner & Witte, 2009). This suggests that stock market performance is the factor that gets treated with the most care, despite the fact that for most people who are retired, the importance of stock market performance pales in comparison to the importance of state pension (Turner & Witte, 2009).

It is not known how laypeople interpret Monte Carlo results or whether the presentation format influences their interpretation. Turner and Witte (2009) raise this as an important question for future research. I will not address it in this thesis since my design went in another direction.

There is no standard for how to visualize a Monte Carlo result. The presentation often resembles that of a deterministic result with a few different scenarios (Turner & Witte, 2009). A common graphic for retirement calculators is a bar chart of expected retirement wealth which is to be compared to a bar chart of calculated retirement need (Bi, 2015).

### 2.3.2.2 Retirement Need

Calculating how much money the user will have at retirement is only one piece of the puzzle. In order to tell the user whether they have saved enough, the software also needs to determine how much they will need.

One method is to tally up the user’s current expenses and then multiply them by the expected number of years that the user will spend in retirement (Turner & Witte, 2009). This assumes that the user’s lifestyle will be the same indefinitely. There are a few problems with this approach.

First, the cost of medical care increases at a rapid rate for most people as they age (Dowd & Town, 2008). Some calculators correct for this by assuming a higher rate of inflation for those categories of expenses that tend to increase with the user’s age (Turner & Witte, 2009). Additionally, some expenses such as drugs and medical care can be tax-deductible which is another good reason to keep categories of expenses separate (Turner & Witte, 2009).

Second, the user’s current level of consumption can be unsustainably high or low (Kotlikoff, 2008). If the user spends too much, they will need a lot of capital to sustain the same consumption after retirement. If they spend too little, it may not be realistic to expect them to keep to the same low consumption level in the future. Thus, making the wrong consumption estimate can lead to over- or undersaving. Kotlikoff (2008) argues that both are a problem because oversaving costs extra in forms of
fees and taxes, which by itself reduces the user's available resources, and undersaving puts the user at risk of not having properly funded their retirement.

An additional complication is that even a minor misestimation compounds into a large error, because the planning period is long (Kotlikoff, 2008). Kotlikoff’s solution is to calculate, based on the assumption that the user wants to smooth their living standard, how much the user can spend and what insurance they need in their last year of life, then calculate the corresponding plan for their penultimate year, and so on backwards until the current year is reached. Hence, he suggests prescribing to the user how much they can spend rather than taking their current level of spending as given. Kotlikoff admits that his approach is challenging in terms of the computational power required. His method is a breach from conventional financial planning theory (Kotlikoff, 2008; Turner & Witte, 2009).

Some retirement planning software allows the user to indicate which of their expenses they consider to be essential and which ones they could live without (Turner & Witte, 2009). Some also ask the user to what extent they are willing to reduce their non-essential expenses. This is a great feature that makes it possible to weigh a reduction in consumption against other possible solutions such as saving more or buying insurance (Turner & Witte, 2009).

A different approach, perhaps the most common, is to use a rule of thumb for what fraction of pre-retirement income the user should have after retirement. This is sometimes called the replacement rate (Turner & Witte, 2009). As with other “magic” numbers, experts disagree on what the replacement rate should be. It is usually somewhere around 70-80% of pre-retirement income. The number being lower than 100% is supposed to reflect the fact that certain expenses go away after retirement, such as retirement saving and work-related expenses (Turner & Witte, 2009). However, this ignores the fact that certain expenses, such as healthcare, tend to go up after retirement. In any case, it is clear that the ideal replacement rate cannot be the same for everybody. For example, people with high fixed costs, such as a large mortgage, might need a higher replacement rate, while people with children are considered

---

56 Smoothing living standard is not the same as smoothing spending. For example, Kotlikoff’s model would take into account that children grow up and leave home.
to need a lower replacement rate, since the children will have left home by the time the parents retire\textsuperscript{57} (Turner & Witte, 2009).

2.3.2.3 **Longevity**

Besides level of consumption, the other factor that determines how much capital the user is going to need is how long they are going to live after they retire. This is obviously an unknown factor and there are different ways of handling it.

Some calculators simply assume that the user will live to 95, or even 100 (Turner & Witte, 2009). The end age is set high in order to reduce the risk that the user outlives their means (Turner & Witte, 2009). However, not everyone can realistically expect to live to 100, and under that policy there is a risk that many people will live austerely and die with capital that they never had the chance to enjoy (Turner & Witte, 2009). People living poor and dying rich is not the goal of economic theory, since it does not constitute an optimal use of resources\textsuperscript{58}.

Other calculators solve the problem by having the user choose their own end date (Turner & Witte, 2009). However, most people underestimate their own life expectancy and a plan based on the user’s own guess is therefore likely to fall short (Turner & Witte, 2009).

Turner and Witte (2009) recommend that the end date be based on official statistics of the life expectancy of different demographic groups, considering that the statistical differences between some groups within the United States are more than 10 years in life expectancy. The difference is even greater when accounting for behaviors such as drinking and smoking.

It is important not to use life expectancy as the actual end date, since it is an average and many people will live longer, but to use life expectancy plus a margin of a few years (Turner & Witte, 2009). Ideally, longevity should be treated as a stochastic variable in the program’s simulation of different futures (Turner & Witte, 2009).

An alternative to smoothing retirement consumption over retirement years is to weight consumption for each year by the probability that the user will be alive that year (Turner & Witte, 2009). That way, the user will

\textsuperscript{57} This ought to apply to the current level of consumption, not the current income. I guess the reasoning is that the family is assumed to make ends meet on today’s income.

\textsuperscript{58} This assumes that there is no bequest motive.
have more capital available while they are younger and presumably more active.

2.3.2.4 **ANNUITIES**
The above reasoning assumes that the main way for the user to fund their retirement is by withdrawing from saved capital until it runs out. Indeed, this is the only method presented by most retirement calculators (Turner & Witte, 2009). But it can be difficult to accumulate all that capital, especially for women since they both live longer and have lower lifetime incomes (Cohen Birman et al., 2016), and even when the calculator says that there will be enough capital, there is still always the risk of living even longer than the calculator expected. An alternative solution, favored by economists (Benartzi, Previtero & Thaler, 2011) but vastly under-represented in retirement calculators (Turner, 2010; Turner & Witte, 2009) is to purchase an annuity. An annuity guarantees the retiree an income for the rest of their life, thus eliminating longevity risk altogether, especially if the annuity is protected against inflation (Bodie, 2003). There are even cheap longevity insurance products, which begin paying out if and when the user reaches an advanced age (Turner & Witte, 2009). Annuities are not very popular and economists are puzzled as to why, since they are often a rational choice (Benartzi, Previtero & Thaler, 2011). It has even been suggested that lack of information about annuities in retirement calculators may to some extent be to blame for the poor sales of annuities (Turner 2010). In Sweden, every retiree can be said to have an annuity since a large share of the government pension can only be withdrawn as lifelong payments. However, it has relatively recently become possible to withdraw occupational pension over a shorter period and this has led to some concern that people who should be choosing lifelong payments might instead choose to spend their capital quickly (Hagen, 2017).

2.3.2.5 **DICHOTOMOUS SUCCESS INDICATORS**
Many retirement calculators show a clear indication of projected success or failure, for example a smiling face or a frown (see Figure 8). The definition of success is usually that the user is not expected to run out of money before they die (Turner & Witte, 2009). Obviously, this could be calculated in different ways and there is no consensus on the “one, true formula.”
Empirical tests show that feeding the same data to different calculators will make some calculators predict success, others failure (Sondergeld et al., 2003; Turner & Witte, 2009; Bi, 2015). In one case, one calculator predicted a monthly retirement income of $0, while another calculator predicted $2,241 for the same person (Bi, 2015).

The reasons for the disparity in prognoses are, among other factors: Differences in underlying assumptions about variables such as stock market returns and inflation; differences in whether the calculators account for taxes, government-provided benefits, and home equity; differences in the software’s ability to handle special cases such as expected inheritance and big, one-time expenses; differences in target replacement rate; and for stochastic programs, differences in how high the probability of reaching the goal must be in order for the plan to be declared successful (Turner & Witte, 2009).

This practice of displaying a yes/no answer has been criticized for being too simplistic (Turner & Witte, 2009). As pointed out by Turner and Witte (2009), there is a big difference between running completely out of money 10 years into retirement, versus not quite being able to afford the same high standard after 35 years of retirement. Perhaps the latter outcome is acceptable to the person in question. Hence, the probability of failure is not the only interesting component of risk; the consequence of a failure is at least equally important (Turner & Witte, 2009; Bodie, 2003; Kotlikoff, 2008). For a person who relies more on state pension than on wealth drawdowns, such as is the case for the majority of Americans (and Swedes), a loss of capital can be less catastrophic.

Additionally, Kotlikoff (2008) points out that the prognosis of how long the money will last is based on the assumption that the family will keep up their normal level of consumption during a financial crisis, while empirical evidence shows that households reduce their consumption in uncertain times. He argues that this potential variability in consumption...
is in fact the relevant measure of risk for a layperson – not the probability of capital loss.

2.3.2.6 **Pro-risk Bias**

Bodie (2003) opened a conversation on whether online retirement calculators have a pro-risk bias, that is, whether they steer users toward investing too much in stocks. He tested four online calculators by entering data which should have resulted in recommendations for a conservative portfolio and found that the calculators still always recommended some investment in stocks, even for people who are already retired.

Bodie criticized not only the calculators as such, but some of the conventional wisdom behind their design. He pointed out that while most calculators assume that a young person can take more risk, this assertion has no basis in finance theory (Bodie, 2003). A risk averse person should always invest conservatively; a person who prefers high risk can take risk even if they are close to retirement. The reasoning behind the idea that stock market risk decreases with time, is that time smooths out the average. However, if a person happens to invest right before a big crash, they will have lost so much capital in the beginning that they will never get back to break-even, even if subsequent years are good. This is especially true for a person who is already retired, who is also drawing down from their capital at the same time it is depreciating due to stock market movements. A risk-averse person should invest in such a way as to minimize the volatility of their consumption – this means that a lifelong, risk-free annuity is a better choice for them than the stock market (Bodie, 2003).

However, Bodie admits that there is another, more valid reason for young people to take more risk, provided that they have a relatively secure job: Their ratio of future income to (potentially lost) capital is high. Older people risk more simply by not having as much income in their future, which would otherwise mitigate some of the damage done by capital loss.

Bodie also rejects the idea that investing on the stock market protects against inflation. The only time in the last century when inflation was high in the United States was in the 1970’s, at which point the stock market was also declining (Bodie, 2003).

The simplest and worst calculators are deterministic calculators that simply correlate the proportion of stocks in the portfolio to a higher
expected return (Turner, 2012). This makes it look like an underfunded retirement plan can be salvaged by taking more risk, when a more reasonable response might be to reduce spending, to move to a smaller house, or to buy insurance against worst-case scenarios (Kotlikoff, 2008; Turner, 2012). It also completely ignores the riskiness of risky investments and makes it look like stocks will always generate a positive return. Truly risk-free investments, such as inflation protected government bonds and annuities, are rarely recommended in retirement calculators (Bodie, 2003; Turner 2010a).

Simple calculators are also prone to overestimate the return that an average user can get from the stock market. This is because they do not account for taxes, fees, and the fact that average investors underperform the market by about 1% because they make mistakes and because of poor market timing (Turner & Witte, 2009).

Some retirement calculators recommend users to reallocate their portfolio without considering whether this action will trigger a tax effect (Turner & Witte, 2009).

### 2.3.2.7 STATE PENSION

Most retired Americans are dependent on Social Security for their monthly income (Turner & Witte, 2009). Therefore, their Social Security benefits are the most important factor in their planning, despite most retirement calculators devoting the most time and sophistication to their investment portfolio (Turner & Witte, 2009).

For Americans of low to middle income, the most important question is how early they can retire (Turner & Witte, 2009). Postponing retirement relative to the required minimum retirement age increases monthly benefits. Additionally, benefits are affected by whether the user has a working spouse, or a deceased spouse with a work history. Few retirement calculators do a proper job of calculating future Social Security benefits according to the actual rules (Kim, Hanna & Chen, 2014; Turner & Witte, 2009).

### 2.3.2.8 HOME EQUITY

For most Americans (Turner & Witte, 2009) and Swedes (Calvet, Campbell & Sodini, 2006), their home is by far their most valuable

---

59 Postponing retirement in Sweden one year increases state pension by about 7-8% (Pensionsmyndigheten, 2017).
financial asset. In spite of that, home equity was not widely considered part of the household’s assets and risk before the American house market crash in 2008 (Turner & Witte, 2009), in which the average American family lost two decades’ worth of financial gains (Baker & Rosnick, 2008).

Owning a mortgaged home brings mainly two risks to the household: A risk that the house loses value and a risk that a variable mortgage rate rises (Turner & Witte, 2009). When the home loses value, that affects people who were planning to sell it or borrow against it to finance their retirement; when mortgage rates rise, it affects the cash flow of the family. In the worst case, both can occur at the same time, at which point the family can find themselves unable to pay their bills and forced to sell their home at a loss. This happened in Sweden in the 1990’s and a recent estimate says that approximately 25,000 Swedes are still deep in debt due to mortgages they took in the 1990’s before the crash, despite having sold their homes (Aronsson & Palm, 2015).

The calculators that do account for home equity treat it in very different ways (Turner & Witte, 2009). Some assume that the home can be sold in order to finance retirement, hence count it as part of the user’s wealth. This is misleading if the user is in fact not willing to sell their home. Others assume that the home is completely illiquid and do not account for it at all, thereby restricting options for the user and rendering the home a practically useless asset. The most flexible calculators ask the user whether they are willing to sell the home if necessary (Turner & Witte, 2009).

2.3.2.9 Hidden Risk
Retirement calculators may give the impression that they visualize a user’s “risk,” but in fact there are many different types of risk of which stock market risk is only one. Even when a calculator accounts for more risks, for example longevity risk, there are usually other hidden risks of which the user may be unaware. Turner and Witte (2009) recommend that all retirement planning software warn their users that the software is only a tool and to use conservative estimates, cushion their plan with extra savings, and complement the software with self-study and

---

60 Most new mortgages in Sweden have a variable rate (Holmberg et al., 2015).
61 When selling a home at a loss in Sweden, the debtor is still required to pay back the remaining part of the debt. If they are unable to pay it back quickly, the debt will accrue interest and grow. Filing for debt restructuring is a long process and the appeal is sometimes declined (Kronofogden, 2016).
discussions with friends and financial advisors. Additionally, they recommend that the software state upfront what kinds of questions it answers and who it is suitable for, considering the different needs of different socioeconomic groups. Sondergeld et al. (2003) recommend people to run several different calculators instead of relying on just one, in order to get a view of the range of advice from different software.

2.3.2.10 **Evaluation Methods**

As shown by this meta-review, the majority of studies on retirement calculators are expert evaluations, with economists or actuaries evaluating the appropriateness of the advice and sometimes also the user-friendliness of the system. To my knowledge, there is only one attempt to study the actual effects of using retirement calculators for real people. Bi (2015) used data from the Survey of Consumer Finances in the United States to test whether people who used retirement calculators engaged in more positive economic behaviors and whether they had accumulated more retirement wealth. She found a positive correlation in both cases. Unfortunately, this study method does not prove that there is a causal link – it seems equally plausible that people who engage in positive economic behaviors can be more prone to use retirement calculators. To test cause and effect, an intervention study is needed.\(^{62}\)

Thus, there is a lack of research on how representative users, as opposed to experts, read and interpret retirement calculators and other financial decision aids. This thesis is aimed at addressing that lack.

---

\(^{62}\) Intervention studies on financial literacy education, not retirement calculators per se, have shown mixed results and newer studies appear less optimistic than older studies (Bernheim, Garrett & Maki, 2001; Bernheim & Garrett, 2003; Fox, Bartholomae & Lee, 2005; Mandell & Klein, 2009; Fernandes, Lynch & Netemeyer, 2014).
3 THE SWEDISH CONTEXT

All empirical studies in this thesis have been conducted in Sweden. Therefore, it is worth noting how Sweden is similar to and different from other countries.

Sweden has one of the highest internet penetrations in the world (Sanou, 2015). Out of the total population, 91% use the internet, 92% have a computer, 77% have smartphones, and 76% use the internet on their smartphones (Findahl & Davidsson, 2015). In the ages 26-45, more than 90% paid their bills online according to a survey from 2012 (Findahl, 2012) and there is no reason to believe that that trend has declined. Swedes can be expected to be relatively computer-savvy and familiar with the idea of online services.

Culturally as well as materially, Sweden is a highly modern society. In the World Values Survey (2014), the country scores at the extreme end on both secularity-rationality and self-expression values among countries surveyed. Swedes are generally trusting of authorities such as their government (Rothstein, 2004) and their banks (Eriksson, Persson & Söderberg, 2009). The Swedish mentality has been called “state individualism” – the welfare state provides support in times of need, so that the individual can be independent from family and church (Berggren & Trägårdh, 2015). Swedes generally do not live with their aging parents; the material welfare and caretaking for the elderly is considered the responsibility of the government.

Sweden has a large middle class and a relatively equal distribution of income. This is due to high taxes and extensive transfer programs (Calvet, Campbell & Sodini, 2006). However, the distribution of wealth is closer to other industrialized nations, with the lower 20% having almost no measurable wealth and the wealthiest households owning most of the financial assets, while middle class households have most of their wealth.

---

63 As opposed to traditionalism.
64 As opposed to survival values.
65 Bank accounts returning less than 100 Swedish crowns in annual interest rate are not reported to the Swedish Tax Authority, from which this data was pulled (Calvet, Campbell & Sodini, 2006).
invested in their homes. Real estate constitutes over 70% of the average household’s wealth\textsuperscript{66} (Calvet, Campbell & Sodini, 2006).

Sweden has four main types of households. The same pattern can be observed in other countries (Calvet, Campbell & Sodini, 2006):

- Households who only save in the form of cash.
- Households who are saving to buy a home and are investing partly in stocks and mutual funds.
- Households who own a home and have few financial assets.
- Households with large financial wealth.

In recent decades, Swedish housing prices and mortgage debt have increased at a rate that outpaces real wage increases and inflation (Finocchiaro et al., 2011). This has inflated Swedish households’ indebtedness to a level that might threaten the stability of the Swedish economy (Emanuelsson, Melander & Molin, 2015), despite the fact that a majority of Swedes say they are uncomfortable with debt (Almenberg et al., 2016). About half of the working age population has a mortgage, many of which are interest-only or are paid back slowly (Almenberg et al., 2016). The average repayment rate in 2012-2013 was 99 years (Winstrand & Ölcer, 2014).

Household risk due to inflated housing prices is “hidden,” i.e. not apparent to the average layperson from merely looking at their balance sheet. For example, a positive net worth might look positive because the household’s home is overvalued and a good margin between after-tax income and interest payments might look good because interest rates are currently extremely low (Emanuelsson, Melander & Molin, 2015). A combination of falling house prices and rising interest rates could put many Swedish households in a situation where they cannot afford to pay their mortgages and at the same time cannot sell their home at a price that would redeem the mortgage in full\textsuperscript{67}. Swedish mortgages are not tied to the property but to the borrower; a foreclosure does not relieve the

\textsuperscript{66} The numbers are from 1999-2002, but with recent increases in real estate prices, there is no reason to believe that it has gone down.

\textsuperscript{67} This happened in Sweden in the financial crisis of the 1990’s. An estimated 25,000 Swedes still live with large mortgages that they took out during the housing bubble before that crash, while at the same time having lost their homes (Aronsson & Palm, 2015). Currently, it looks like interest rates can only move up while housing prices might soon only be able to move down; the official bank rate is negative and the average debt ratio of Swedish households is 343% (van Santen & Ölcer, 2016).
borrower of debt if the value of the property has declined below the value of the mortgage. Thereby it is the borrower, not the bank, who carries the risk in an unstable housing market. This makes the mortgage a very important factor in the balance sheets of Swedish households.

The two most common investment mistakes made by Swedish households are under-diversification and non-participation in risky markets (Calvet, Campbell & Sodini, 2006). Stock market non-participation is more common among groups that score lower on numeracy and financial literacy tests: The elderly, women, people born outside Sweden, and people with lower education (Almenberg & Widmark, 2011). People who do not participate in the stock market at all while they are young and can bear the risk, can miss out on large gains in wealth (Cocco, Gomes & Maenhout, 2005). However, the cost of under-diversification is not very large for the typical Swedish household (Calvet, Campbell & Sodini, 2006). In other words: Even though a suboptimally composed stock portfolio is a common problem, it may not be a very important problem.

Sweden’s retirement system is currently rated as B by the Mercer Melbourne Global Pension Index (Mercer, 2016). This grade denotes a system which is basically sound, but has room for some improvement. In Sweden’s case, the researchers suggest that the state pension age ought to be raised to reflect increasing life expectancy, that some taxes and rules ought to be changed in order to encourage more people to engage in occupational pension plans and private savings, and that more measures should be taken to protect both parties in a divorce (Mercer, 2016). The Swedish retirement system is currently changing from a generous but underfunded system which is still operative for the oldest citizens today, to a more sustainable system based partly on income-based pension credits, partly on occupational pension and private savings. For the lowest income earners, there is a guaranteed minimum pension (Mercer, 2016).

Sweden shares the experience of shifting from a defined-benefit pension system to a defined-contribution system with many other countries, for example the U.K. and the U.S (Holzmann & Palmer, 2006). In a defined-contribution plan, more risk and responsibility is carried by the individual citizen, who is responsible for managing the investment of their own retirement funds. When the system was initially changed, the government encouraged citizens to be active and change their
investments from the government-provided default mutual funds. This led to such poor results that the government has now stopped recommending people to deviate from the default plan (Cronqvist & Thaler, 2004; Palme & Sundén, 2004). The difference in retirement wealth between the most and the least successful retirement investors is expected to grow in the future (Cohen Birman et al., 2016). Additionally, people who have been absent from the Swedish job market will be penalized more than the retirees of today, who still receive their pensions from the old system (Flood, 2014). This will be particularly noticeable for people who came to Sweden as adult refugees (Flood, 2014). The share of retirees in poverty is expected to grow in the future, especially if the default retirement age is not raised (Flood, 2014).

There are current discussions of removing the default standard of starting government pension payments at the age of 65, at which age many people are still capable of working (SNS, 2016). A relatively recent change to the occupational pension system has made it possible to withdraw occupational pension over as little as five years instead of lifelong payments. This has led to the concern that some people will withdraw and spend their pension too quickly, underestimating how much income they are going to need in later retirement (Hagen, 2017). People who spend their pension too quickly can later become dependent on extra government assistance, which would increase the burden on the system.

Swedish law requires financial advisors to adapt their recommendations to the client’s wishes, needs, and knowledge level (SFS 2003:862; MiFID II 2014/65/EU). They are, thus, legally bound to ask the client about their risk preference, even though there is no scientific consensus on how to measure risk preference (see section 2.1).

Andersson and Korling (2012) have reviewed contemporary information to consumers in the Swedish financial services sector. They found that while the earliest parts of the decision process – identifying needs, establishing expectations, setting goals – are by far the most important, since a mistake in that early stage will almost certainly lead to disappointment further down the line, most legal regulators have focused their efforts on later stages in the decision process, where consumers

---

68 The exact share of the pension that can be withdrawn at a custom rate varies by occupational contract. White-collar workers typically have more flexibility than blue-collar workers (Hagen, 2017).
already know their goals and are about to choose a specific investment product among other similar products. This is also the stage where portfolio pickers come in; the user is assumed to have a more or less clear idea of what return they expect and what risk they are willing to take, and now it is simply a matter of composing a portfolio that fulfills those requirements.
4 METHODS & APPROACH

While I may have delved deep into the domain of personal finance in the background chapter, it is worth reminding the reader that the purpose of this thesis is not to answer questions in economics or personal finance, but to design a useful system.

Behavioral economics, from which I have drawn much of my inspiration, is an empirical field. It studies actual decisions made by real people, as opposed to theoretical models of rational agents. However, behavioral economists often stop at describing, and sometimes explaining, curious “anomalies” in people’s behavior as it deviates from expectations of rationality. Research aimed at designing concrete interventions that help us overcome our irrational tendencies is rarer, although examples exist69.

Human-Computer Interaction (HCI), and its subfield Information Visualization (InfoVis), tend to focus heavily on testing and evaluating systems and their suitability for the target group. Although theoretical, technical, and aesthetic considerations are important in InfoVis, a solid understanding of the domain problem is considered equally important in order to design visualizations that are useful (Sedlmair, Meyer & Munzner, 2012; Munzner, 2009). The related field Interaction Design (IxD) tends to focus on the process of design, and it has been said that while other fields analyze things as they are, IxD imagines what could be (Cooper, Reimann & Cronin, 2007).

The controlled experiment is a common method in both behavioral economics and HCI. However, qualitative methods are more common in HCI than in behavioral economics. HCI studies typically have fewer participants and greater importance is placed on their verbal comments as they describe their experience and explain their thought process while using the system.

My empirical work follows a tradition that is more typical for HCI. However, I spent considerable time and effort trying to understand the domain problem before I started working on my own design idea. My early studies were aimed at understanding how current financial planning

---

69 See for example Save More Tomorrow (Thaler & Benartzi, 2004).
systems work and thereby find a problem description complementary to the one offered by behavioral economists. While behavioral economics literature describes many different financial mistakes made by typical people, it does not offer a map of how these mistakes interact in a usage situation such as when a person sits down in front of a computer to plan their retirement. It is not easy to find a systematic overview of which mistakes can have important consequences for people’s financial futures and which ones are trivial. Thus, some of my empirical work was aimed at forming a deeper understanding of the problem, which could be translated into design requirements. Some work was spent on generating a design idea. Later studies are dedicated to testing and developing the design idea.

This chapter is an overview and details about my methods are found under each respective study in the Results chapter.

4.1 Methodology

Figure 9 describes my process and the methods I used at the different stages. Sometimes results from an evaluation of the prototype in its then-current state forced me to revise my understanding of the problem, which prompted a new chain of exploring-creating-evaluating. Thus, the whole research project was iterative, as should normally be the case for user-centered design projects (Gulliksen et al., 2003).
4.2 EXPLORATORY METHODS

I use the term *exploratory methods* here to describe methods I used to gain deeper knowledge about the problem and to do more concrete requirements gathering. The term *exploratory* has been used elsewhere to denote early stage evaluations of systems (Rubin, 1994), but I prefer to reserve it for the process of exploring the domain and the problem, rather than exploring a new design.

I could have included the literature review here too, because it is a type of method and it constituted the bulk of my information-gathering at the beginning of the project, but I do not consider it an empirical method.

4.2.1 EXPERT INTERVIEWS

I interviewed two subject-matter experts in order to validate and nuance my understanding of personal finance in Sweden and in order to get early reality checks on my design idea. This is strongly recommended when the domain is complex, such as “medical, scientific, or financial services” (Cooper, Reimann & Cronin, 2007). Subject-matter experts provide knowledge about industry best practices and complex regulations, but they are not professional designers, nor are they representative of ordinary users (Cooper, Reimann & Cronin, 2007). They are therefore better suited to discuss the underlying problem than to create or evaluate concrete design details.

4.2.2 FOCUS GROUPS

Focus groups, or workshops, is a good alternative to one-on-one interviews when the researcher is interested in group dynamics and hearing the participants’ reactions to each other’s statements (Greenbaum, 2000). It is also a suitable format for collaborative and creative efforts such as generating requirements (Young, 2002) and design ideas (Muller, 2003). I used a focus group with financial advisors in order to hear their views on their clients’ typical problems and knowledge level. Hearing their discussions when I asked them to rank problems in order of importance was very informative.

4.3 CREATIVE METHODS

*Creative methods* are used to create something, in this case a design concept and its materialization into an interactive prototype that can be presented to others.
4.3.1 Sketching
Sketches are used very early in the design process, when radically different ideas are still being explored. Sketches are cheap and easy to make and can therefore be discarded without much concern (Buxton, 2010). They should not look too finished, but should have a rudimentary look and feel that encourages criticizing, re-thinking, and re-drawing (Buxton, 2010). However, they should be actual pictures of the interface and not just descriptions of functionality; otherwise it is difficult to know whether the design team is talking about the same thing (Gulliksen et al., 2003). Sketching could be called a pictorial form of brainstorming; its purpose is to produce a multitude of ideas without judging them, because the evaluation process comes later.

4.3.2 Prototyping
According to Buxton’s terminology, prototypes are more detailed than sketches. They can include working interactive elements and hence they take longer to build and are typically constructed at a stage when there are much fewer design alternatives left (Buxton, 2010). The main difference between a prototype and a real system is that the prototype does not do all the things that are required of a real system (Preece, Rogers & Sharp, 2002). The prototype might not have all the necessary functions, it can use mock data, and its underlying algorithms may be simplifications whose only purpose is to demonstrate a few key cases.

Buxton (2010) suggests that sketches are used to ask questions, while prototypes are used to resolve questions. I produced a number of sketches to discuss with my supervisors, but used exclusively interactive prototypes in the later evaluation studies.

4.4 Evaluative Methods
Evaluative methods are used to test a design, in order to verify that it meets requirements and to find potential for improvement.

4.4.1 Heuristic Evaluations
A heuristic evaluation is an expert review, where one or more experts look through the system in search of possible flaws. Many of the studies in the background chapter are heuristic evaluations, where economists have examined financial planning systems in order to evaluate the suitability of
the advice they provided (e.g. Bodie, 2003; Kotlikoff, 2008; Merton, 2003; Turner, 2010; Turner & Witte, 2009).

In HCI, a heuristic evaluation typically entails a usability expert inspecting the system in order to make sure that it complies with a list of known usability principles (Preece, Rogers & Sharp, 2002). The list of usability principles keeps evolving and changing as more empirical research is done, but it is often considered too general for specialized systems that may encounter complex, domain-specific problems (Preece, Rogers & Sharp, 2002).

I have made what I would call an informal heuristic evaluation of the different graphs found in retirement planning software; I did not use the usability principles from HCI, but instead used domain-specific knowledge about retirement planning derived from my literature studies to judge the suitability of different graphical representations.

4.4.2 Usability Testing
Usability testing is one of the core methods in HCI (Faulkner, 1998; Rubin, 1994; Preece, Rogers & Sharp, 2002). Variations of the technique go under names such as user observation and contextual inquiry. The basic idea is to evaluate systems and find their strengths and weaknesses by having representative users use the systems to solve relevant tasks. A researcher observes the user while they are solving the task, but does not tell the user how to do it. Sometimes the user is asked to verbalize their thought process as they proceed through the task (Nielsen, Clemmensen & Yssing, 2002; Ericsson & Simon, 1980). The user’s spontaneously uttered thoughts can be complemented by a post-test interview where the user reflects upon the experience they had while using the system and how they reasoned while solving the task (Fonteyn, Kuipers & Grobe, 1993). The number of users required to produce valid results is debated, but an empirically established minimum seems to be 10 (Faulkner, 2003).

4.4.3 Controlled Experiments
Controlled experiments are used to demonstrate cause and effect (Faulkner, 1998). In HCI, they can also be used to compare the performance of one system design to another (Faulkner, 1998; Rubin, 2002). In contrast to user observations, which can be quite exploratory and open-ended, experiments require one or several a priori hypotheses and a set of circumstances under which the hypotheses will be accepted
or rejected. When comparing two systems, for example, a threshold for when one system is considered to have outperformed the other must be determined before the experiment. An example of such a threshold can be that a statistically significantly (at the 5% level) higher number of users manage to complete a task in one of the systems. Provided that the user groups are equivalent, the difference in the systems is then assumed to have caused the difference in success rates among users.

However, a demonstration of causality should not be mistaken for an explanation. The demonstration of a causal link between system design and user success does not explain how the design caused the difference, only that the difference exists. Verbal comments from users can inspire the researcher to produce possible explanations, although the validity of study participants’ reports about their own thoughts as explanations for their behavior is debated (Nisbett & Wilson, 1977). Such speculative explanations can be discussed in relation to the study’s result, but should not be taken to be as strongly evidenced as the result itself. To explain the causal link in the most rigorous manner, subsequent experiments should be conducted to test the respective predictions of competing explanations.

I have focused mainly on determining whether there is an effect and collected only indicative data on possible explanations for the observed effect. Explaining the effect in terms of users’ mental processes would be an interesting future project.

To help the reader stay oriented through all the chapters and studies, the below image will recur throughout section 5 whenever a new project phase begins, highlighting the current phase. The image is not a figure that relates to my results or arguments; it is merely a visual aid. Hence, it has no figure number.
4.5 SUMMARY

Table 3 summarizes the methods and data used in this thesis.

Table 3: Methods and data in this thesis.

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Methods</th>
<th>Number of items or participants</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RQ1</strong> How well do current financial decision aids work for non-expert users?</td>
<td>Reviews of existing systems</td>
<td>26 example graphs</td>
</tr>
<tr>
<td></td>
<td>Heuristic evaluation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Usability test</td>
<td>17 participants</td>
</tr>
<tr>
<td></td>
<td>Requirements gathering</td>
<td>7 participants</td>
</tr>
<tr>
<td><strong>RQ2</strong> Can a new representation address the weaknesses while preserving the strengths found in RQ1?</td>
<td>Concept design</td>
<td>Same as above</td>
</tr>
<tr>
<td></td>
<td>Workshop*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sketching &amp; prototyping</td>
<td>~10 sketches, 1 prototype</td>
</tr>
<tr>
<td></td>
<td>Expert evaluation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interview</td>
<td>2 participants</td>
</tr>
<tr>
<td></td>
<td>Financial advisor evaluation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Usability test</td>
<td>6 participants</td>
</tr>
<tr>
<td><strong>RQ3</strong> Do non-expert users understand how to use the new representation?</td>
<td>Comprehension test</td>
<td>12 participant</td>
</tr>
<tr>
<td></td>
<td>3 pilot experiments</td>
<td>8 participants</td>
</tr>
<tr>
<td></td>
<td>Experiment</td>
<td>24 participants</td>
</tr>
<tr>
<td></td>
<td>User test</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Usability test</td>
<td>12 participants</td>
</tr>
<tr>
<td><strong>RQ4</strong> How does the new representation influence users’ risk taking?</td>
<td>Risk test</td>
<td>4 participants</td>
</tr>
<tr>
<td></td>
<td>Pilot experiment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Experiment</td>
<td>119 participants</td>
</tr>
</tbody>
</table>
5 RESULTS

This chapter contains my empirical results, which constitute the bulk of my contribution to knowledge about financial decision-support systems.

5.1 EXPLORING EXISTING FINANCIAL PLANNING SOFTWARE

5.1.1 GRAPHS IN RETIREMENT PLANNING SOFTWARE

While previous research has provided systematic overviews of the functionality of contemporary retirement planning software, no one has studied their graphical representations. Since an important task for such software is to properly communicate risk to users, I had my own look at the charts and graphics in retirement planning software.

5.1.1.1 METHOD

I excluded simple compound interest calculators and portfolio pickers; otherwise there would have been hundreds of cases to study. With the help of Wikipedia's lists of banks by country, I visited the websites of the largest banks in Sweden and the other Nordic countries, the U.K., the U.S., and Europe70. I explored the banks' websites in search of interactive retirement planning tools. I also visited the websites of well-known financial advisory firms and financial news publishers. Finally, I googled "retirement calculator," which turned up a few previously unseen results from (English-speaking) government authorities and private individuals. I was forced to exclude systems that were inaccessible to me due to restrictions on who could create a login account based on location or

---

https://en.wikipedia.org/wiki/Banking_in_the_United_States accessed 2016-10-25
citizenship\textsuperscript{71}. I also had to exclude systems that were inaccessible unless I became a paying customer\textsuperscript{72}.

5.1.1.2 Results

After this search, my sample consisted of 26 interactive or non-interactive\textsuperscript{73} retirement calculators; only one from Sweden, three from other Nordic countries, two from the U.K., one from the rest of Europe, 17 from the U.S. and one of undisclosed origin. Ten of the calculators came from banks, five from wealth management firms, one from an online broker, one from an investment company, one from a financial advisory firm, five from financial news publishers, one from an industry association of banks and insurance companies, and one from a group of anonymous “IT professionals” who create free online calculators\textsuperscript{74}. The latter was the calculator of undisclosed regional origin as previously mentioned. There are probably hundreds of calculators that I missed and the sample is obviously skewed toward the English-speaking world; this was a choice I made because of the language barrier. I did see some saturation in the sample; the calculators could be sorted into about five groups based on visual similarity.

The following examples are the graphs that were either the most common, or that best met the quality criteria established by Turner and Witte in section 2.3.2. The graphs are presented in order of frequency.

\textsuperscript{71} Betterment and BlackRock generously offered me access to their systems via demo accounts.

\textsuperscript{72} A couple of high-end Swedish financial advisory firms were excluded because of this.

\textsuperscript{73} Only one was non-interactive; it consisted of two example calculations with no way to change the input values.

\textsuperscript{74} These classifications of companies in the finance sector were not always easy to make, and I may have miscategorized a few. As far as possible, I tried to go by how the companies described themselves on their websites.
5.1.1.2.1 Stacked Bar Chart

Stacked Bar Chart was used by 12 (46%) of the retirement calculators in the sample.

![Stacked Bar Chart](image)

Stacked Bar Chart shows the user’s total future wealth at retirement, if they change nothing about their current savings strategy. Future wealth is compared to future need (exemplified here by a dotted green benchmark line), based on anticipated lifestyle and anticipated longevity. If future wealth is expected to fall short of future need, the deficit is shown as a stacked bar in another color.

Some calculators in the sample have only one bar, which makes them the type of deterministic calculator that Turner and Witte (2009) warned about. Some have two bars, representing two alternative future scenarios; one where the stock market performs as expected and one where it performs poorly. This gives some indication of investment risk, although only for the discrete scenarios selected by the designer. No indication of the uncertainty of the prognosis is given.

Stacked Bar Chart needs a fixed end date for the retirement period, in order to tell the user whether their savings will last long enough. Hence, it can be prone to the longevity risk problem mentioned in an earlier chapter; it does not show what happens if the user lives longer than expected.

In terms of known graph perception principles, Stacked Bar Chart is easy to read so long as its colors are easily distinguishable for people with color perception deficiencies. An alternative to the stacked bars would be to simply represent the expected wealth by a bar and the expected
deficiency by empty space – that would get rid of any problems for colorblind users.

Stacked Bar Chart shows whether there will be a shortfall and how big the shortfall will be, but it requires a lot of mental effort to translate this shortfall into lifestyle changes; how much spending would need to be cut on a monthly basis. I would therefore argue that Stacked Bar Chart does not do a sufficient job of computational offloading. Alternatively, it encourages an unreflected acceptance of the graph’s binary pass/fail verdict – an example of an oversimplified graph. Additionally, because the graph focuses on the size of wealth and collapses expenses into a thin benchmark line, I conjecture that it nudges the user to think about solutions that concern increasing their expected wealth at the time of retirement, while solutions based on a reduction in consumption do not come as readily to mind.

5.1.1.2.2 Wealth Burndown
Wealth Burndown was used by 6 (23%) of the retirement calculators in the sample.

![Wealth Burndown Graph](image)

Figure 11: Wealth Burndown.

Wealth Burndown shows how the user’s wealth grows over time, if they change nothing about their savings strategy. After retirement, they start consuming their wealth at a rate determined by their future anticipated need. Wealth Burndown can represent risk through a fuzziness of the edge of the blue area. This fuzziness can represent a combination of risk (the more pessimistic scenarios predicted by the model) and the uncertainty of the model itself (by adding a margin to the uncertain area).

Wealth Burndown solves the end date problem in a rather elegant way; it shows when the money will run out, without need for the user to specify
results

5. RESULTS

68

a target date. However, the user still needs to make the judgment whether the projected end date is at a late enough age.

A drawback, which Wealth Burndown shares with Stacked Bar Chart, is that neither is explicit about where income comes from and where the money goes. They both focus on wealth and ignore income, at least visually. If income from a state pension is accounted for, it slows the consumption of wealth, but that is not visible in the chart so it is hard for a user to know. Wealth Burndown makes it look like the user is broke when their wealth has been drained, but in fact, they may not be dependent on their wealth because they have a state pension or an annuity that covers their expenses. As such, Wealth Burndown makes it look like the only solution to longevity risk is to invest more or invest more aggressively. Like most retirement calculators (Turner & Witte, 2009), Wealth Burndown cannot suggest annuities or insurance, which is often a better solution to the longevity problem.

By not showing what expenses are consuming the wealth, Wealth Burndown and Stacked Bar Chart both make it impossible to see how spending might be reduced. Wealth Burndown does not say anything about what kind of lifestyle is being funded for all those years – affluent, satisfactory, or meager – only how long it can last. Thus, similar to Stacked Bar Chart, Wealth Burndown provides cognitive offloading to a point that gets deceptively simple – they both show whether the money will last to a given date but do not invoke deeper questions about lifestyle, priorities, and how severe the consequences would be of failing to meet the goal. However, I would argue that Wealth Burndown provides a more useful type of graphical constraining, by guiding the user toward the thought that the retirement planning puzzle is about making the money last through life, not about acquiring a big lump sum.

However, the unit of analysis in Wealth Burndown is still a lump of money that grows and shrinks. The focus on total wealth in Wealth Burndown produces big numbers that are far away from what normal users are used to thinking about or handling. There is a potential risk that users might be impressed by the big numbers and feel satisfied with their

---

75 Additionally, spending is either assumed to be static across all years, or it is difficult for a user to decide to e.g. spend less on transport after retirement and see the effect in an intuitive way. Such a change will be represented as a flatter curve (a slower rate of consumption) which is not the same data unit as the input (a lower monthly cost). In case the user makes several budget changes at different points in the chart, they will probably lose track of all those changes and their effects rather quickly.
forecast without translating the number into a monthly cash flow. Additionally, to my knowledge it has not been confirmed that users understand that the growing and shrinking curve represents a constant level of consumption.

5.1.1.2.3 Component Areas
Component Areas was used by 2 (8%) of the retirement calculators in the sample.

Component Areas shows the user's expenses as areas on a timeline and their income as a benchmark line. When the income line dips below the top of the expense areas, that means the user will no longer be able to afford their spending. In one version of Component Areas, it is the user's different income sources\footnote{Such as pension, post-retirement work, etc.} that are separated into areas and all expenses are aggregated into one line. Like Wealth Burndown, Component Areas solves the longevity problem by having the money last for as long as it lasts, without need for the user to specify an end date.

Component Areas makes it easy to see where the money goes, so that the user can decide whether they want to reduce consumption or try to increase income. It also makes it easy to see whether it is the income or the expenses that are affected in different alternative scenarios. Importantly, it shows when the user will manage on just state pension, without needing much saved capital\footnote{Except perhaps an emergency fund for unexpected expenses.}. As requested by Bodie (2003) and
Kotlikoff (2008), it shows risk and consequences in terms of reduced spending power rather than reduced stock portfolio value.

The perceptual task of reading the chart ought to be easy, or at least I thought so before my empirical studies. The line where savings end is to be compared to a point on the horizontal timeline, and the user must then determine whether this point is sufficiently late in life. If the money does not last long enough, the user can try strategies to increase their savings, or cut down on their spending.

Most graphs in the reviewed sample were the only graphics in the retirement planning system, but some were complemented by additional graphics. All focused on retirement savings and did not attempt to include other aspects of financial planning such as insurance and credit management. Although some of the charts were part of a suite of financial planning tools, the tools were always focused on one task at a time.
5.1.2 **ONLINE INVESTMENT ADVISORS AND NOVICE USERS**

An online investment advisor is not a person but a computer system. It asks users questions about their risk preference and their financial situation, runs their answers through an algorithm and generates a recommended investment portfolio which is supposed to provide the level of growth and risk that matches the user’s goals. It is similar to so-called robo-advisors, but while robo-advisors can act autonomously on the client’s behalf, for example by making trades without asking the client first, online investment advisors are typically just a form of recommender system.

Figure 13 shows a conceptual sketch of what an online investment advisor might look like. The sketch contains elements that were found in the systems tested in this study, although the elements looked different in the different systems.

This study was published in Heyman (2015), except for the results described in section 5.1.2.2.6, which I noted at the time but did not include in the conference paper, and some additional analysis.

![Figure 13: A conceptual sketch of the graphical interfaces of the online investment advisors tested in the study. The numbers in the figure are made up for illustrative purposes.](image)

### 5.1.2.1 AIM

This study served to explore research question 1: How comprehensible are today’s financial decision aids to non-expert users? Do they find them helpful? What are the systems’ strengths and weaknesses?

I would have liked to measure how well the systems’ recommendations matched the users’ financial goals, but there is no exact objective norm.
against which to make such a comparison. Instead, I looked for potential ways that the communication between user and system could go wrong; on the way into the system (when the user answers the system’s questions) and on the way out of the system (when the user reads and tries to make sense of the system’s recommendation).

5.1.2.2 **Method**

5.1.2.2.1 **Participants**
Test participants were recruited through a Facebook advertisement, which said that researchers at my university were looking for people willing to participate in a paid interview about their personal finances. Since Facebook was used by 66% of Swedish Internet users\(^{78}\) at the time (Mjömark, 2014), it was deemed mainstream enough that it would not severely skew the participant group. The advertisement was shown to people aged 18 or older who lived in Stockholm. It was only shown to desktop users and not to mobile users. Clicking on the advertisement led to a questionnaire that screened out people with a background in finance or user experience design. The questionnaire also asked for information about the participant’s occupation, education, household size, whether or not they owned stocks, and whether or not they had a mortgage. As young people were quicker to sign up for the interview, the target group for the advertisement was gradually narrowed to include only ages that were underrepresented in the pool of people who had signed up.

Twenty people were selected for the interviews. They were picked for maximum diversity with regards to age, gender, family situation, and financial experience (see Table 4). Out of the 20, 18 people came to the interviews. One did not respond after being invited to the interview and one did not show up after repeated tries to reschedule. One interview did not render usable results and was not transcribed. Hence, 17 interviews were collected in the end.

\(^{78}\) 91% of the Swedish population uses the internet (Findahl & Davidsson, 2015).
Table 4: Participants in the online investment advisors study.

<table>
<thead>
<tr>
<th>n</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Median 49, SD 13.54</td>
</tr>
</tbody>
</table>
| Gender | 58.8% M  
|       | 41.2% F |
| Level of education | 29.4% high school  
|                    | 17.6% vocational training  
|                    | 52.9% university |
| Occupation | 3 students  
|            | 3 elementary or preschool teachers  
|            | 2 artists  
|            | 2 care or nursing assistants  
|            | 1 journalist  
|            | 1 animal keeper  
|            | 1 small business owner  
|            | 1 coordinator  
|            | 1 business developer  
|            | 1 manager  
|            | 1 engineer (retired) |
| Family situation | Lives with another adult 64.7%  
|                  | Has children under 18 23.5% |
| Financial experience | Owns a stock portfolio 52.9%  
|                      | Owns a mortgage 47.1%  
|                      | Has ever met with a financial advisor 47.1%  
|                      | Has met with a financial advisor within the last 2 years 29.4% |

5.1.2.2.2 Stimuli
The study used fully functional online investment advisors as stimulus. In order to avoid focusing the analysis on idiosyncrasies from one single
online advisor, the aim was to include all online investment advisors that could be found in the Swedish language. The need to limit my search to Swedish systems was because I wanted Swedish study participants from a variety of educational backgrounds and I did not want the language barrier as a confounding factor.

The websites of all large Swedish banks were searched for online advisors that were open to the public without logging in. The websites of government authorities for consumer information were also browsed and an internet search was performed in order to find any interactive guides or simulators that might be published by some other third party. The search turned up only two Swedish online advisors, both published by medium-sized banks whose names are likely to be known by most Swedes. Both were essentially portfolio-pickers; they did not provide advice on insurance, credit, or other aspects of personal finance.

The two systems were quite different from each other (see Table 5). However, both were programmed in Flash and had a colorful look and feel. Both started with a needs elicitation questionnaire and then landed on a page that presented a recommended investment portfolio, where the user could manipulate the risk level to see how that changed the suggested portfolio.

The resulting recommendation was presented in a graphical format in both systems, although details in the presentations differed. Figure 13 shows a conceptual sketch of the recommendations page.

It is worth repeating that the purpose of the study was not to compare the two systems, but to get a diversity of opinions and reactions from the test participants.
Table 5: The two online investment advisors used as stimuli.

<table>
<thead>
<tr>
<th></th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question items</td>
<td>8</td>
<td>31</td>
</tr>
<tr>
<td>Questionnaire pages</td>
<td>7</td>
<td>22</td>
</tr>
<tr>
<td>Type of questions</td>
<td>Simple language; vague</td>
<td>Financial jargon; choices between probabilistic outcomes</td>
</tr>
</tbody>
</table>
| Graphics             | • User’s risk preference on a scale  
• Chart of proposed asset allocation incl. several asset classes  
• Expected return and worst case return | • Risk level on a scale*  
• Chart of proposed allocation between equity funds and fixed-income funds  
• Expected return, worst case return, and best case return  
• Probability of negative outcome |
| Variables manipulable by user in graphic simulation | Risk, time | Risk |

*Ambiguous whether it was the user’s preferred risk level or the recommended portfolio’s estimated risk level

5.1.2.2.3 Procedure

All participants consented to have audio recorded. They were semi-structurally interviewed about their views on their personal finances and financial risk for about 30 minutes, starting from a map that they drew of their personal finances; a method used by Kaye et al. (2014). The purpose of this initial session was to warm up the participants and get them to talk about money and financial risk, as well as getting a better picture of who they were in order to help me make better sense of their comments during the following usability evaluation.

After the interview, I asked the participant to help me evaluate one of the systems by trying it and saying what they thought of it. To avoid
fatiguing the participants, each participant evaluated only one system. Participants were told to answer the system’s questions the same way they would if they were alone in their homes, but they were allowed to turn the screen away when answering questions about their income and assets. Participants were told to think aloud throughout the test, a method commonly used in contextual interviews (Nielsen, Clemmensen & Yssing, 2002). To help participants feel more comfortable criticizing the online advisor, they were informed that I, the evaluator, had found the online advisor on the Internet and had not personally had any part in its development. The test of the online advisor, including filling out the needs elicitation questionnaire and talking about the participant’s impression of the results, took about 30 minutes, rendering the total interview about one hour long.

5.1.2.2.4 Analysis
I transcribed and read through all the audio tapes and identified recurring themes in the text according to conventional content analysis as described by Hsieh and Shannon (2005). According to Hsieh and Shannon, this is a suitable method when there is little existing theory on which to base interpretations of the data. Some patterns were common but were not deemed interesting for design decisions. For example, many respondents said they or someone in their family had had a bad experience with investments. This information was not found useful for design and hence it was excluded from further analysis.

After thematic coding, I selected 100 snippets from the text and had two independent researchers validate the thematic categories by sorting the snippets by category. The length of text snippets ranged from 19-1,113 words. Snippets that I found the most difficult to categorize were given precedence, in order to allow for a deeper discussion among coders about what had been observed in the study. The categories were explained prior to coding, in order to ensure that category labels had equivalent meanings to all coders. More than one category could be selected for each text snippet. The coders worked independently, but met for a discussion afterwards. Coders could modify their judgment after the discussion. An overall inter-coder reliability of 80% was reached after discussion and

---

Pilot testing had shown that evaluating just one system was quite time-consuming.
adjustments80. The average number of unassigned text snippets per coder was 3.33 (SD 3.51), which means that coders found the categories to fit most of the snippets.

5.1.2.2.5 Results
Initially there were 13 categories, but 9 of them contained fewer than 10 text snippets on average. Only the three largest categories (with more than 15 text snippets on average) are reported below. The size of the category is calculated as the number of text snippets averaged over coders, following the procedure described by Aryani, van Schaik & Barker (2012). Table 6 shows the sizes of the three largest categories.

Table 6: The three largest themes found in the interview transcripts.

<table>
<thead>
<tr>
<th>Category</th>
<th>Size (SD)</th>
<th>N of snippets agreed upon by coders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust</td>
<td>23.33 (8.08)</td>
<td>12</td>
</tr>
<tr>
<td>Financial literacy</td>
<td>20 (2.65)</td>
<td>11</td>
</tr>
<tr>
<td>Resignation</td>
<td>17.67 (3.79)</td>
<td>10</td>
</tr>
</tbody>
</table>

5.1.2.2.5.1 Trust
The trust category was defined as quotes where “the participant discusses trust or distrust toward banks, the computer, or the financial system.” One reason why this category was the largest was that quotes from the preceding interview were included in the analysis, where topics often concerned previous experiences with advisory meetings and general attitudes toward banks. Ten participants expressed distrust toward banks, for example suspicion that banks are more interested in selling profitable products than in giving customers unbiased advice. Two participants said they trusted the online advisor more, because a human advisor might get financial compensation for selling more profitable products while the computer could be more objective. However, most participants seemed to trust the online advisor no more than they trusted

80 A text snippet counted as 0% if no coders agreed on the appropriate category, 66% if two coders agreed and 100% if all coders agreed. If two coders agreed on more than one category for a text snippet, it still counted as only 66%. The total of 80% is an average across text snippets. This method was based on Ramberg, Artman and Karlsgren (2013).
the bank behind it. Six participants were not persuaded by the recommendations they got because they did not trust the stock market. They saw stocks as fleeting and abstract, whereas tangible assets such as gold, diamonds, real estate, or land were perceived as safer.

**Quote 1**
P9: But... some people say that... the best way to get rich is to buy gold. To own bullions. Because... it may go up and down, but at least it keeps its value. (Pause) Then the stock market can collapse completely... but the gold is like, still there. It could be diamonds, too. (Pause) But I hear gold is good. (Pause)

**Quote 2**
Researcher: Would you follow this recommendation if you had some money to invest?
P8: Er, I... Probably not. Because I'm not very... into this whole thing with saving in stocks or mutual funds. But... if I was to invest in something, I would probably invest in something more... solid. That is, gold, diamonds. That is... something that actually exists in the real world. (Laughs)

This lack of trust in the stock market might be partially due to poor financial literacy. These participants failed to recognize that the value of gold also fluctuates, even though the physical bullion is still there. Research by Almenberg and Widmark (2011) confirms that people with lower financial literacy are less prone to participate in the stock market. Perhaps the above quotes cast some light on why.

Six participants said they hesitated to save money because they perceived financial markets to be too chaotic and unpredictable. It is likely that they still had the financial crisis of 2009 in recent memory. Some said they saw no use in planning, since there was no guarantee that they would get the money back that they had invested.

Six participants said they saw the online advisor as one potential source of information, but that they wanted to collect information from more sources, such as other banks or family and friends before making a decision.

**5.1.2.2.5.2 Financial Literacy**
The financial literacy category described cases where the system presupposed knowledge that participants lacked. Many of the cases were
due to unexplained jargon in the user interface, such as “investment horizon,” “risk propensity,” or “required return.” There were also examples where participants understood the terminology as such but failed to make sense of the recommendation because of unspoken information that the system assumed to be common knowledge.

Researcher: Here it recommends for you to have twenty-two percent equity and seventy-eight percent fixed-income. What do you think of that recommendation?
P14: Hm, I’m not... Fixed-income funds are not riskier? No?

P14 does not know that a fixed-income fund is considered safer than an equity fund (perhaps they don’t even know what a fixed-income fund is) and hence they cannot make sense of the online advisor’s recommendation; they do not see the connection between their own preference for low risk and the recommendation that they invest in fixed-income funds.

5.1.2.2.5.3 Resignation
The resignation category was defined as quotes where “the participant feels dejected about their possibilities to influence their own [financial] situation, [and] perceives a lack of opportunities.” Eleven participants said there was no point for them to change their habits, to start saving or to change their asset allocation, because they had too low income or too little wealth for it to make much of a difference. Eight participants said they did not have the time required to be active investors or that they did not prioritize it. Five said they lacked the skill and expertise to trade stocks.

Researcher: Do you save for your retirement?
P7: No.
Researcher: Why not?
P7: Um, you mean to have my own retirement... investment? No... Ah... Why not... Well, I feel that I cannot afford it. (Smiles) And there are also so many stories that... it doesn’t work, that you don’t get the money.
Researcher: Mm. Where have you heard those stories?
P7: No but just that... Like... Er, the economy ... (Sighs) of the world collapses. And then those investments are worth nothing. But... ah, now I can’t find the words. You don’t get your money back.
RESULTS

Researcher: Mm. Do you think that’s what’s going to happen?
P7: You—- You don’t know. (Laughs)
Researcher: No.
P7: That’s the way it is, so... uncertain. So... I can... I don’t believe in it.
[inaudible] that it’s something that you can be sure about. So that’s why I don’t do it.

It could be noted that these interviews took place in 2014 and that the participants may have had the financial crisis of 2009 relatively fresh in their memories.

The three categories presented above seem interconnected. Due to low financial literacy and limited knowledge about financial markets and products, many participants viewed stock trading as a form of gambling. They reacted to recommendations of equity funds with suspicion, perhaps because they were ignorant of the principles of risk diversification within a mutual fund. They felt dejected about their own chance to accumulate some wealth, perhaps because they underestimated the power of compound interest.

5.1.2.2.6 Secondary Findings
I observed some behaviors that were less common, but in some ways more interesting than the somewhat expected findings reported under the proper Results section. The evidence in this section is weaker, because these were things I read into participants’ behavior, rather than them saying explicitly that this was what they were thinking. I still find it worthwhile to describe these secondary findings because of their important potential implications for financial decision aids.

5.1.2.2.6.1 Central Tendency Bias
When participants were unsure of how to answer a question, usually regarding their risk preference or their expectations on the return of their investment, they tended to choose the option in the middle (most questions were multiple-choice, like the example question in Figure 14)81.

81 It could be noted that the question in Figure 14 is probably also sensitive to anchoring; users may for example have heard sometime that 5% should be a minimal return requirement, and choose 5% return without considering the risk.
Figure 14. An example of a question that might invoke central tendency bias.

The following are quotes uttered when participants decided to choose the middle option because they did not know what to choose.

**Quote 1**

System A: [Your] experience with securities:
- No experience
- Some experience
- Average experience
- Extensive experience

P16: I have tremendous experience\(^{82}\), but that depends on how you look at it.

Researcher: Mhm, tell me more about that.

P16: Yeah I mean the terrible experience I had with those [securities] I inherited after my dad and stuff like that. So experience, I have.

Researcher: Mm. So how would you answer this question?

P16: “Average.” That’s always a great answer.

**Quote 2**

P16: (Reads) “It’s important to be able to predict an approximate amount at a certain point of time in the future.” Well, *important*, but it’s impossible. Just look at the housing market in Stockholm.

Researcher: Mm.

P16: Let’s put a 3.

---

\(^{82}\) Original: “en oerhörda erfarenhet”.

5. RESULTS

**Quote 3**
P3: (Reads) "I see risk as an opportunity to make money." (Long pause) Well...
Researcher: What do you think about that question?
P3: (Laughs) "I see risk as an opportunity to make money." I have to think about what they mean by that. But... since I don't like taking high risk, that doesn't feel like a question that I have a straight answer for. So I'll put a 3 on that one in that case.

**Quote 4**
P11: Um... (Reads) "Investment horizon"... [inaudible] but... yeah, no.
Researcher: Pardon?
P11: No, uh, investment horizon, I don't know what that is, but... "Limited."[^1]
[...]
P11: (Reads the next question) “How long is your investment horizon?” (Pause) Ah, I'll say... three to five years. Don’t know... it’s...
Researcher: What do you think investment horizon means?
P11: I don’t know, actually. I have no idea. Er... When the money is due to come back, maybe? When I... if I buy now, it could... bear fruit in these different years [intervals] maybe.
Researcher: Mm.
P11: And perhaps the shorter it is, the higher the risk and the higher is... what I can get, maybe? I don't know. So... it feels to me like somewhere here in the middle it might be a little... semi-safe. But I have no idea what it is.
(Laughs)

[^1]: "Limited" was one of the answer options.
Quote 5

P13: Wow, these things are so tricky\textsuperscript{84} to...
Researcher: Mm. What is it that you find tricky?
P13: My required return... and my risk propensity. Okay, uh... (Pause)
Researcher: What do you think they mean by required return?
P13: Well, what I think it should give [me]. [How much] it should increase.
Researcher: Mm. And risk propensity?
P13: Well, that it can... decrease. [inaudible] Right?
Researcher: I didn't write the questions. (Laughs) I'm just guessing at what they mean, too.
P13: Mm. But what \textit{my} requirements are, and... I guess that's a typical in-between. (Pause) We'll end up there.

It appears that some participants perceived the middle option to be an “average,” a “default,” or a “non-extreme” option. This phenomenon is a known problem for survey design (Tourangeau, Couper & Conrad, 2004) and the implication is that people who are unsure how to answer questions about their risk preference are at the hands of the designer who chooses the distribution of numbers on the risk scale. If the designer considers an average annual return of 5% at a risk of losing 15% of the capital to be “medium risk,” then that is the risk level that uncertain users will get. This cannot be said to have anything to do with user’s “true” risk preferences.

Paradoxically, a risk-averse user should use the option at the extreme end of the scale, when the scale itself represents risk. Yet, I observed participants who claimed to be risk-averse choose the middle option.

The fundamental problem is that people are unfamiliar with questions like “How much return do you require on your investment?” and “How much loss can you accept?” The answer options, framed as a percentage share of invested capital, are abstract and detached from any notion of what the real-world consequences would be of a loss or a failure to meet the expected return. In order to get more reliable answers about people’s risk preferences, the questions must be framed in terms of something they can relate to.

\textsuperscript{84} Original: “krångligt”
5. RESULTS

5.1.2.2.6.2 The Personality Test Effect
One of the tested systems displayed its result of the risk assessment in the form of a gauge that said “Your risk level” and ranked it from low to high\(^{85}\). See Figure 15 for a conceptual example of what that sort of illustration can look like.

![Gauge Illustration](image)

Our results indicate that you are a low risk taker.

Figure 15: An example of a gauge that seems to focus on the user’s risk preference rather than the risk of the portfolio.

The message could be read as a classification of users’ personalities\(^{86}\).

P2: At the moment I don’t do anything, I just study and work, but... If I were to decide to start working and not study, I would take risks. This is just boring.

Researcher: What type of risk?

P2: For example to invest, or do business... buy and sell goods, that [inaudible]

Not only might this confound the results by invoking impression-management issues\(^{87}\), depending on whether the participant wanted to seem like a high-risk taker or a low-risk taker, but can also steer the user’s attention away from the part that is actually relevant to their investment decisions – what they want to do with the money they are getting recommendations for. It should not matter whether the user is a “high-risk taker” in general, what matters is whether they can afford to lose this money, at this point in time.

---

\(^{85}\) The other system used the more neutral expression “Chosen risk level.”

\(^{86}\) In the beginning, I as test conductor went along with the systems’ language of personality testing, for example by asking participants whether they felt that the system had produced an accurate description of them. I did not see this as a problem until later.

\(^{87}\) This may have been exacerbated by my presence next to the participants when they were answering the system’s questions, but it is not unfeasible that people will still want to preserve their self-image when no one is looking. Who is a high-risk taker? Is that the sort of person I aspire to be? Are high-risk takers “cool” or are they stupid and reckless? Are low-risk takers boring or are they smart and responsible?
P11: (Reads) "Which of the following options is the closest to your own risk propensity?" (Pause) Well, in my life it has been one of these.
Researcher: You’re jumping between "Moderately high" and "Moderately low"?
P11: Yes. Now... nowadays I don’t dare to gamble that much, but back when I did that thing with the stocks I did dare to gamble, I bought some, so... here and there. So I guess I might be a bit risk taking, or at least I used to be. So I can say that. (Clicks)
Researcher: "Moderately high"?
P11: "Moderately high."

In the quote above, P11 searches their memory for historical indications of what sort of risk preference they seem to have displayed in the past. This is completely irrelevant to their investment decisions here and now. P11’s remark that they did not dare take many risks “nowadays” alluded to what P11 had told me in the preceding interview – they were currently in a situation with tight economic margins.

Additionally, some participants seemed to pay the most attention to the result of the “personality test,” perhaps because it was more relatable and naturally more interesting to a lot of people, while paying less attention to the actual investment recommendation.\footnote{An alternative explanation is that they were less interested in the recommendation than the “personality test” because they were not actually about to invest money.}

Risk preference in one domain tends not to correlate with risk preference in other domains (Weber, Blais & Betz, 2002). A person’s general behavior is therefore not a good enough indication of their risk preference specifically in the personal finance domain. Risk preference regarding investments should therefore be measured only through direct questions about the investment at hand; no proxy questions about personality, past behavior, or general attitudes.

5.1.2.3 Discussion
The study uncovered several communication problems between users and the systems. When the systems used jargon and complicated terminology, some participants did not understand what responses the system expected from them before it could compute its recommendation.
Additionally, when the system presented the recommendation, not all participants understood why they had received that particular recommendation.

5.1.2.3.1 Limitations
I do not know whether the tested systems were actually designed for novice users.

5.1.2.3.2 Reliability and Validity
The concepts of reliability and validity are generally understood as belonging in a quantitative research paradigm (Golafshani, 2003). However, qualitative research also needs to strive for rigor and trustworthiness and some would also include generalizability as a quality criterion for qualitative research (Golafshani, 2003; Leung, 2015). While some agree that the concepts of reliability and validity still have value for qualitative research, their definitions need to be adapted to the context of qualitative methods (Golafshani, 2003). Reliability in qualitative research is not understood as test-retest replicability, but rather as coherence in the interpretation of data (Leung, 2015). This can be pursued through the use of structured analysis methods and interrater comparisons such as was used in this study. The concept of validity is more accepted in qualitative research (Golafshani, 2003), although some researchers prefer terms such as credibility and confirmability (Shenton, 2004). Greater credibility is achieved by, among other things, using established research methods; using a random sample of informants; triangulation of methods and data sources; and by employing strategies to ensure honesty from informants, for example by assuring them that there are no right or wrong answers (Shenton, 2004), or by detaching the design from the evaluator, as I did in this study.

The section named Results follows an established standard for interrater reliability. The section named Secondary Findings does not, hence its evidential value is lower. While I find that section interesting, the observations reported there are my interpretations and could benefit from more rigorous testing. However, I do not see any harm in designing future systems as if central tendency bias and the personality test effect are real phenomena.

---

89 It is very difficult to fully replicate the conditions of a qualitative study.
5.1.2.3.3 Implications
This study was designed to address research question 1: How comprehensible are today's financial decision aids to non-expert users?

The tested online investment advisors appeared to assume too much of their users. While I do not know whether these systems were actually intended for a more financially sophisticated target group, they were available online for everyone to use. My study participants seemed like they needed to be convinced of the value of saving and investing at all, before they were ready to discuss optimal portfolio allocation. They also needed basic knowledge of concepts such as compound interest, which helps the small-scale investor grow a large fortune over time, the risk-minimizing effect of diversification, and the fact that tangible assets such as gold and real estate are by no means protected against value fluctuation. In other words, there are more fundamental pedagogical issues to address before going into the finer points of weighing the risk and return of different investment products.

Additionally, I considered the tone of the systems a poor match for these participants. Because several participants had stated that stock trading was not for them – it was for people who are more knowledgeable, interested, persistent, risk-loving, or greedy – I figured that a financial decision aid targeted at the average person should not be focused on "getting rich," but rather on financial security. While not everyone is interested in stockpiling money, most people ought to be interested in being able to cope financially with unexpected events such as sickness or unemployment. Perhaps if the system talked about security, it would feel less alien to non-experts and more people would be interested in its recommendations. This reflection, to an extent, is the central motivation and driving principle for the re-representation move of my body of design work in the following sections.

To summarize, the systems seemed in need of a re-representation if they were to appeal to this target group. If the system is to be used for retirement planning, then users need help judging whether 1 million in savings is a little or a lot – otherwise the user's target amount can be easily influenced by anchoring to irrelevant or outdated numbers. The system could help constrain the interpretation of the problem and provide computational offloading by showing users that their 1 million has to last them about 30 years after retirement, which comes down to less than 2,800 per month before taxes – suddenly the 1 million in savings sounds less impressive.
5.2 CONSTRUCTING A PROTOTYPE

The activities at this stage of the process were aimed at research question 2: What would a financial decision aid look like that addresses the weaknesses while preserving the strengths of contemporary financial decision aids?

5.2.1 FINANCIAL ADVISORS’ VIEWS ON THEIR CLIENT’S PROBLEMS

5.2.1.1 AIM

After my study on online investment advisors, I had learned that novice users did not seem ready to use a simple portfolio-picker without more educational information and tips and perhaps a re-representation. But was this a common problem or was it just my study participants? And if novice users needed help, what did they need help with?

I decided to run a workshop with professional financial advisors, in order to get a glimpse of their years of experience working with clients. My coworkers and I wanted to learn what areas of financial advice giving were perceived as the most problematic today. Our original aim was to also start creating a new design concept in collaboration with the financial advisors, but the creative part of the workshop proved less fruitful than the initial discussions.

5.2.1.2 METHOD

We designed the workshop as an adaptation of the Future Workshop (Apel, 2004), which deploys three stages where participants: 1) Describe what problems they currently experience; 2) Brainstorm a boundless scenario with perfect solutions fulfilling ideal requirements; and 3) Translate these requirements into workable solutions in practice. We ran a pilot workshop with a couple of our colleagues in order to ensure that our time plan seemed reasonable and that our workshop tasks rendered useful data.

We recruited seven financial advisors between 25 and 46 years old and with field experience ranging between 1.5 and 13 years. All of the advisors were co-workers amongst themselves. Their manager was not present for the workshop.

In the first stage of the workshop, we asked participants to pair up and list, on individual post-it notes, all the current problems they could devise focusing on: 1) Objectives paramount to clients; and 2) Concepts difficult to explain to clients. After 30 minutes, participants gathered in front of a
whiteboard and read their post-it notes aloud. With participant consent, we recorded audio for discourse analysis. The participants and a moderator then collaboratively ranked the post-it notes in order of importance to the client.

In the second stage, we attempted to collaboratively design a visual tool for financial advice. We instructed the participants to imagine any system that perfectly visualized every necessary concept and data item. My role was to constructively criticize the design and one of my supervisors moderated the discussion. In retrospect, this task was probably too foreign to our study participants, something we did not catch in our pilot study because we ran it with fellow designers and researchers who were more used to creative work. In the end, we did not use the result of the design session but focused our analysis on the list of problems discussed in the first stage of the workshop.

5.2.1.3 Results
Table 7 describes the post-it notes (translated from Swedish) and how they were collaboratively ranked in order of importance. The “Details” column is a synthesized transcription of the group discussion.

The literature in behavioral finance reports many of the problems described in Table 7, confirming our results and indicating their existence outside of the Swedish context. These problems include: people’s limited ability to focus on the bigger picture instead of being caught up in details (Kahneman & Riepe, 1998), impulsiveness and short-sightedness (Benartzi & Thaler, 1993), ignoring fees (Barber & Odean, 2000), herd behavior (Thaler & Sunstein, 2008), and unwarranted optimism (Bazerman & Moore, 2012).

The focus on details at the expense of the bigger picture (problem #2) seems like a mental accounting problem. Advisors believe that their clients should evaluate their situation at one particular level of granularity, while the clients see it differently. Since the advisors are trained experts in finance, I choose to trust the judgment of the advisors. Hence, clients have a problematic tendency to look at the performance and/or risk of individual investment products as opposed to seeing how the products combine to create a portfolio with a completely different risk profile. The overall risk of the portfolio can be low even though it contains a small portion of risky products.
Table 7: Problems financial advisors perceive to be common with their clients, ranked in order of importance.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Risk and return</td>
<td>Understanding risk and expected return in different scenarios, including correlations between different markets. Clients are often under-diversified.</td>
</tr>
<tr>
<td>2 Overview vs. details</td>
<td>Clients focus on details, such as the risk of an individual product, rather than seeing the bigger picture.</td>
</tr>
<tr>
<td>3 Incomplete Information</td>
<td>Clients make decisions based on too little or the wrong information, or fail to make any decision at all.</td>
</tr>
<tr>
<td>4 Information disclosure</td>
<td>Clients do not understand why the advisor needs to know personal details such as family situation, assets or debts in other banks, or plans for the future. Failure to disclose important information can potentially mislead the advisor to give unsuitable advice.</td>
</tr>
<tr>
<td>5 The retirement system</td>
<td>The Swedish retirement system is difficult to explain and failure to make the right decisions can have a large impact on the client's future.</td>
</tr>
<tr>
<td>6 Amortizing too little</td>
<td>It is common for clients to amortize too little on their mortgages.</td>
</tr>
<tr>
<td>7 Impact of fees</td>
<td>Clients don’t understand the magnitude of the impact of fees, such as mutual fund management fees, on the total return of the investment.</td>
</tr>
<tr>
<td>8 Poor foresight</td>
<td>Clients lack foresight and make impulsive decisions, such as buying a house without preparing financially.</td>
</tr>
<tr>
<td>9 Herd behavior</td>
<td>Clients follow others on the “market” rather than the advisors’ recommendations. They jump on the train too late and buy when prices are high and sell when prices are low.</td>
</tr>
<tr>
<td>10 No clear goal</td>
<td>Not many clients can answer questions about what they want their money for or when they will need it. Therefore, they find it difficult to discuss time horizons. However, this does not necessarily prevent clients from following some saving strategy.</td>
</tr>
<tr>
<td>11 Too high expectations</td>
<td>Clients expect unrealistically high returns on investments and low mortgage rates.</td>
</tr>
</tbody>
</table>

Problem #7, the lack of understanding of the impact of fees, is likely to be a computational problem, since fees can seem small when presented on an annual basis, but combine into an exponential overall effect.
The compound effect of fees is probably a good thing to offload through visualization, since its computation is hard and it is larger than people typically expect (Lusardi & Mitchell, 2007). However, I have not focused on visualizing the effect of fees in this project, since that entails the selection of specific investment products. Problems #3 (basing decisions on the wrong information), #4 (failure to disclose important information), and #11 (too high expectations) might be attributed to a poor understanding of financial planning. These clients might be helped by having the problem constrained, graphically or otherwise; by seeing what information is required to solve a planning problem, and typical values of parameters such as growth and interest, they might gain a more realistic understanding.

Our focus group added a dimension to this known set of behavioral problems by having professional advisors rank them in order of importance to the clients’ finances. Additionally, problem #4 was unknown to us prior to this study. Its high rank in the list indicates that the participating advisors perceive it as an important problem.

5.2.1.4 DISCUSSION
This focus group confirmed my idea that picking a stock portfolio was not the most important issue for people who are not experts in finance. Even though risk and return is at the top of the list, the advisors were talking about “the risk in different scenarios,” which I interpret to mean a broader view of risk than the standard deviation from average return on a financial product. Problem #3 (basing decisions on the wrong information) speaks to the difficulty clients have navigating the complex world of finance. Picking a stock portfolio is important, but it cannot be the sole focus of a comprehensive financial planning system.

5.2.1.4.1 Limitations
It is worth remembering that financial advisors may not discuss all sorts of financial problems that are important to their clients. The job of a bank-employed advisor is to discuss the bank’s products and services. Problems concerning social benefit or taxes may be less visible on this

---

90 I decided to focus on the step before the selection of products, namely determining what growth and risk to choose, because systems for selecting products are abundant while systems for determining needs are much rarer. At the same time, the selection of products depends entirely on needs.

91 Clients’ insurance, taxes, and advanced securities trading is sometimes handled by advisors who are specialists in that area.
list than they would have been, had we asked the clients directly instead of their advisors. The advantage of talking to advisors is that they are experts; they can make judgments about their clients’ knowledge level and potentially problematic behavior.

5.2.1.4.2 Reliability and Validity
The list of problems was created collaboratively by having participants write down their perceived problems and then discussing each problem in the larger group, sometimes discarding or downgrading a problem when it was not generally agreed to be important. Reliability could have been improved by running more focus groups with advisors working in other banks, since different banks might train their employees differently.

5.2.1.4.3 Implications
Reading Table 7 as a list of requirements rather than problems, it appears that financial advisors could make use of tools that:

- Visualize the risk in different scenarios.
- Focus on the bigger picture rather than the risk and return of individual products, because it is the risk and return of the overall portfolio that matters.
- Provide clients with all the relevant information needed to make decisions. This would require an analysis of what decisions the tool is intended to support.
5. RESULTS | 93

5.2.2 DESIGN REQUIREMENTS
After my literature studies, the online investment advisor test, and the advisor workshop, I had an idea of what characteristics I wanted my prototype to embody. On a high level, my vision was for the prototype to:

- Address as many of the problems identified in the financial advisor workshop as possible.
- Avoid the influence of anchoring and central tendency bias by modelling data in units that were as closely tied to users’ lived experiences as possible, thus avoiding abstract concepts whenever feasible\(^{92}\).
- Give the user freedom with regards to how to achieve their goals. For example, to let the user decide whether to cut spending or to try to increase their income. The system should not dictate the user’s lifestyle.
- Provide as much computational offloading as possible. In other words, be simple to use. Never require the user to do mental calculations in order to solve a task.
- Guide the user through the financial planning process through graphical constraining.
- Explore whether it is possible to counteract mental accounting by showing users that money is fungible; sometimes a more profitable alternative to investing is to use the money to pay off debt, for instance.
- Additionally, the design should follow known best practices for graphical design based on what is known about visual perception.

\(^{92}\) An example of what I consider an abstract concept is percentage annual growth of an investment portfolio. It is abstract because it is hard for an inexperienced person to judge whether a certain percentage is good or bad; it can only be judged in relation to other percentage growths and does not provide an obvious threshold at which the percentage is acceptable.
5.2.3 Creating and Evaluating a First Prototype
Since the workshop did not provide a design stub as planned, I had to come up with a design idea on my own.

I looked through the gallery of interactive chart examples at [https://github.com/d3/d3/wiki/Gallery](https://github.com/d3/d3/wiki/Gallery) and tried to imagine what financial data could be fed into each of them. This method resulted mostly in ideas that were technically plausible but did not appear to add much value to the user. I also tried approaching the creative task from the perspective of trying to counteract theoretical phenomena, of addressing known real-world problems, and of implementing renowned investment philosophies. I had about 10 different sketches that I discussed with my supervisors, but most were narrowly focused on a single problem and did not seem like they would provide the comprehensive, informative overview that the participants in my early studies seemed to need.

The actual idea for a re-representation came to me in a sort of epiphany during a long car ride.

What if financial risk, typically conceptualized as a probabilistic variation in wealth, could instead be thought of as a probabilistic variation in time before the money runs out?

It would remove a cognitive step for the user, namely to calculate (or guess) how much wealth they should set as their goal. I also had a feeling, still unproven by this stage, that time would be easier for people to reason about than money. My idea was very similar to Component Areas (see chapter 5.1.1.2.3), but I did not know that at the time.

I set to work drawing a sketch that visualized a person’s current spending, assumed it would be the same after retirement, and showed

---

93 Mental accounting, loss aversion, hyperbolic discounting.
94 Too little retirement savings, too little emergency savings, too much mortgage debt.
95 The 7tweleve portfolio, [http://www.7tweleveportfolio.com](http://www.7tweleveportfolio.com).
how long the person’s savings would last given this spending behavior and the person’s retirement income. I assumed that data about the retirement income could be pulled from the national pension agency’s website, where I knew they had a personalized retirement simulator. I made an interactive chart using the javascript library d3.js, where I made it possible to change the asset allocation of the person’s savings, to show what effect increased risk would have on the expected number of years of sustained consumption. I also exposed parameters such as mortgage interest rate, income tax, and working time (full time vs. part time) for the user to manipulate in order to simulate different scenarios. The user could change the amortization rate, which changed short-term costs and the time before the mortgage was repaid. Thus, there were three potential ways to make retirement savings last longer: To save more, to take more risk\textsuperscript{96}, and to amortize more before retirement in order to reduce living costs after retirement.

Figure 16 shows a screenshot of the prototype and Figure 17 is a simplified redrawing made to be readable in print.

The prototype had a timeline from today throughout the rest of the user’s life. The colored areas represented a user’s average monthly expenses, pulled from their bank statement and assumed to be constant into the future with one exception: Loans that get paid down over time. The example data in Figure 16 and Figure 17 has a mortgage, which explains the decrease in spending with advancing age.

The green line represented the user’s after-tax income. After retirement, income goes down to an amount predicted by the government-provided service minpension.se, from which I assumed that a real system could fetch data. In the prototype, retirement income was simply predicted to be a percentage of before-retirement income according to a recommended rule-of-thumb (Pensionsmyndigheten, 2012). To the right were parameters that the user could manipulate in order to simulate different scenarios such as increased tax rate or mortgage interest rate.

\textsuperscript{96} Albeit with the caveat that this only increased the expected number of years that savings would last; it actually decreased the number of years that savings were relatively certain to last. This was my conception of (market) risk.
Figure 16: A screenshot of the first prototype as it looked on a computer screen.

![Cash per month diagram](image)

- **Savings**
- **Mortgage**
- **Groceries**
- **Hobbies & sports**
- **Student loan**
- **Miscellaneous**

**Economic scenarios**
- Income tax +10%
- Unemployment
- Sickness
- Separation from partner
- Mortgage interest +5 ppt

**Life goals**

**Work and retirement**

**Financial strategies**

Figure 17: A simplified drawing of the prototype, made to be legible in the printed thesis.
Figure 18: The prototype with insufficient retirement savings and a simulated future period of unemployment. Recreated for print.

Figure 18 shows a case where the person’s income after retirement would not be enough to cover their expenses. Additionally, the user has clicked on *Unemployment* and selected to see what happens if they were to be unemployed for several years (the gray area, which is movable and resizable). The difference between income and expenses was covered by withdrawing from savings, which was represented by the orange line at the top of the colorful monthly costs. Savings thus acted as a type of “income” which was exactly as high as need be in order to cover spending. Each year where the orange line was present was covered by savings, but the savings would eventually run out. The number of years that the
savings would last depended on how much money was saved and how much would be consumed each year.

The amount of money saved was predicted by the same algorithm that would be used for a fan chart; uncertainty increases with time and with the share of risky products in the portfolio. In this version of the prototype, the solid part of the orange line represented years where there is almost certain to be enough money left. The line turned dashed when the algorithm predicted that there might be savings left, but it was no longer certain.

Increasing the share of risky products in the investment portfolio would widen the fan chart and give it a steeper slope to represent a larger expected return but with greater uncertainty. Here, the dashed line (expected number of years the money will last) grew while the solid line (number of years the money will almost certainly last) shrunk. The decision task for the user was thus transformed from one of judging the suitability of a range of outcomes in terms of money, to judging the suitability of a range of outcomes in terms of the age at which savings would run out. My hypothesis was that the latter would be easier for most people.

In Figure 18, saved capital gets consumed both while the person is unemployed and after retirement. It is not certain to last even throughout the period of unemployment.

I later learned that my concept was similar to a theoretical model for deciding when to retire, suggested by Hatcher (2003). Hatcher was also looking to match saved capital to capital need, but he had assumed that the return on savings was constant and that the two parameters to be manipulated were consumption level and time of retirement, while my design concept was based on the idea that the consumption is constant and the return on savings is a variable whose value depends on the asset allocation.
5.2.3.1 **Feedback from Personal Finance Experts**

In order to ensure that my design concept made sense from a personal finance perspective, I asked two personal finance experts independently to take a look at the prototype and give me their feedback. Both experts are prominent specialists on financial issues for the general public in Sweden and have often been featured in the media.

My purpose with showing the prototype to the experts was not to validate whether this would be a good design for laypeople, for which these experts were probably the wrong informants, but to confirm my assumptions about personal finance. I wanted to make sure that the predictions made by the prototype could theoretically be accurate enough to be useful for planning, that the scenarios I had picked were the most relevant for an average Swedish person, and that I had not left out any crucial factors.

I met the personal finance experts at two separate occasions and when the prototype was in different stages of development. This happened naturally as I incorporated the feedback from the meeting with the first expert and changed the prototype before meeting with the second expert. In the beginning of the meeting, I showed the prototype to the expert, explained how it worked and asked for their comments. To confirm that I had included the most relevant factors in my prototype, I asked the second expert to prepare an answer to the question “What are the most important things for an adult Swede to manage today, regarding their personal finance?” in advance of our meeting. I did not think to do this with the first expert.

After getting the experts’ first impressions of the prototype, I asked them specific questions about my assumptions that I was uncertain about. Specifically, I asked whether current spending could be used as a proxy for future lifestyle and whether it was okay to ignore inflation and future pay raises. I was concerned that accounting for inflation in the
visualization would invoke money illusion\textsuperscript{97} (Shafir, Diamond & Tversky, 1997), especially considering that many Swedes have trouble understanding how inflation works (Almenberg & Säve-Söderbergh, 2011). As for future pay raises, I preferred to be conservative in my prognosis rather than optimistic.

Both experts agreed that it was okay to ignore inflation and possible salary increases and also that current spending could be used as a proxy for future lifestyle.

The first expert thought the design looked messy with too many colors. They also thought the tool should have more action affordance; that things that need urgent action should “jump out” at the user. While the prototype could simulate different scenarios, it gave no indication of how the user could solve the problem in cases where savings would not last.

After receiving their comments, I simplified the color scheme so that the colored areas only shifted between one light and one dark shade of the same color. I chose different colors for fixed and variable expenses, though; since there are recommendations to keep fixed expenses low (Serlin, n.d.), I figured that the share of fixed expenses could be useful information to the user. To address their comment that the tool didn’t help the user create an “action plan,” I changed the resolution of the timeline from years to quarters and set more realistic time limits on scenarios. It was no longer possible to simulate being unemployed for several years, but only for a number of months. With the help of these realistic time periods for scenarios, I could now calculate how much money in total would be needed in an emergency buffer in order to manage such short-term predicaments. I also implemented a very rudimentary “alternative futures” function, where a person with no unemployment insurance could see how much smaller their monthly shortage would be if they had unemployment insurance when they became unemployed.

The second expert advised me to keep emergency savings separate from retirement savings. Up until now, I had considered all savings as one lump of money. Perhaps I had taken the thought of abstaining from mental accounting one step too far; there are clear advantages to keeping emergency savings and long-term savings as separate silos. Specifically,

\textsuperscript{97} Money illusion is when people mistake nominal values for real values. In other words, they would ignore inflation and interpret upward sloping curves in the prototype as real increases in their salary, pension, and costs of living.
the expert said, emergency savings are not part of a person’s wealth; it is meant to be drained whenever need arises, and so one has to count on that to happen every once in a while. Emergency savings cannot be relied upon to last all the way until retirement.

I took the expert’s advice and separated emergency savings from retirement savings. They were kept as different values in the underlying algorithm and they were represented by lines of different colors in the chart. Retirement savings could not be used to cover up for emergency savings and vice versa. I think this made the prototype conceptually easier to understand and it was also more in line with mainstream financial advice: To save up for an emergency buffer first, and only then get into long-term savings. Additionally, long-term and emergency savings are usually saved in different formats, with long-term savings sometimes having penalties for withdrawing early.

Figure 19 and Figure 20 show the prototype after I addressed the comments from the experts. The light blue line at the beginning of the timeline represents a period where the emergency buffer is being consumed. In the years after retirement with a purple line, retirement savings are being consumed.

The blue bar chart to the left shows how much money the user has in their emergency buffer (dark blue bar), versus how much would be needed in the selected scenario (total height of the bar; the light blue part represents a shortage). Scenarios are not mutually exclusive; the user can see what happens if they lose their job and there is a rise in interest rates at the same time.

I added a few more functions as inspiration came to me; goals could be inserted at points in time, such as having a child or buying a new home, to see how that affected costs and income. A very simple function for suggesting spending cuts compared the user’s data to recommended spending for the user’s type of household according to the Swedish Consumer Agency (Konsumentverket, 2016). This function was only visible if actively chosen, as I figured some users might otherwise find it annoying.

Since many people have self-control problems, simply seeing their spending habits might not have any effect beyond creating an “aha” moment. However, in combination with a commitment-saving product, the “planner” in the user’s mind can look at the visualization, decide to cut spending, setup automated saving transfers, and then later check how
well they have adhered to the plan. Thus, the prototype reveals areas in the user’s budget where spending could be cut and money could be saved.

Figure 19: The prototype after the experts’ suggested changes.

Figure 20: Recreated for print.
5.3 Testing the Prototype

5.3.1 Usability Testing with Financial Advisors

5.3.1.1 Aim
After the personal finance experts, I wanted to know whether the prototype had succeeded in addressing most of the client problems described by financial advisors in the focus group. I decided to test the prototype in a mock advisory meeting, where I would pretend to be the client.

The aim of this study was to verify that the prototype provided a way of working with financial planning that made sense to professional financial advisors, that no important aspect of the planning was left out, and that the representational model in the prototype added value.

5.3.1.2 Method

5.3.1.2.1 Participants
Six financial advisors were recruited for the study. Half of them had attended the workshop where the design requirements were constructed. Table 8 provides an overview of the participants.

Table 8: Participants in the usability study with financial advisors

<table>
<thead>
<tr>
<th>n</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Average 37.7, SD 13.3</td>
</tr>
<tr>
<td>Gender</td>
<td>1 M 5 F</td>
</tr>
<tr>
<td>Years of field experience</td>
<td>Average 10.3, SD 7.5</td>
</tr>
</tbody>
</table>

5.3.1.2.2 Stimuli
The data in the prototype at the test was chosen such that the example user would not have enough emergency savings to manage unexpected situations; not enough retirement savings to last throughout expected life as retiree; had some room between income and expenses that could be used to increase savings, but not so much that that would be enough to meet the goal without also increasing investment risk; had enough time left until retirement that investing in stocks would still not be too risky,
but not so much that increasing risk a little would easily solve the problem; and could choose to amortize more on the mortgage as an alternative strategy to saving more, but with today’s low mortgage rates, that would give less expected return on the money\textsuperscript{98}.

5.3.1.2.3 Procedure
The test took place at the study participants’ place of work. Participants were told before the study that they were allowed to work in pairs in order to promote more discussion. One session involved one advisor, one involved two and one involved three (the three who had attended the original design workshop). The shortest session, the one with just one participant, lasted around 40 minutes.

The longest session, which involved three participants, lasted approximately 2 hours.

In the beginning of each session, I demonstrated the prototype to the participants and explained all its features. The focus was on the usefulness of the concept rather than learnability and intuitiveness. All participants consented to having a camera pointed at the computer screen where the prototype was being demonstrated and to have their voices and hand gestures recorded. I then let participants explore the prototype freely for as long as they liked and give spontaneous comments.

When participants indicated that they had explored enough, we started the mock meeting. I played a client coming into the office for the first time and told participants to pretend that the example data shown in the prototype was my personal data. They were then to act like this was an ordinary client meeting and the prototype was their new tool provided by their employer. I had prepared three questions for the advisors:

- How is my financial situation looking?
- Is there anything I should change?
- What investment risk should I take with my retirement savings?

These questions were designed to indicate whether participants thought the prototype gave them a good overview over the client’s financial situation. I told the participants to let me know if they felt like they could

\textsuperscript{98} As is apparent, I still thought of increasing risk as the main method of salvaging an underfunded retirement plan, next after increasing monthly savings.
not give financial advice based on the information they saw in the prototype.

After the roleplay was finished, I asked the participants which functions they found most and least useful for their work, if and when they would like to use a system like this, and whether there was anything about the suggested design that concerned them.

5.3.1.2.4 Analysis
The recording from the longest session was transcribed word for word, including screenshots at points when participants were talking about something onscreen that cannot be understood from the transcription alone. The recording equipment failed in the two shorter sessions and I was left with only the written notes that were taken during the session as backup, plus details written down from memory immediately after the sessions.

I analyzed the data using a focused coding scheme where I looked for answers to the following questions:

- How well does this tool fit the advisors' current practices?
- What features of the tool would be the most and least helpful to the advisors?
- How well does this tool support the advisors in meeting their goals for the client meeting?

5.3.1.3 RESULTS
All participants were enthusiastic about the design concept and said they would like to use it in meetings with most types of clients. One participant said it was good for getting a first overview of a new client. Two participants said in discussion that the one group that this system would be less helpful for were retirees, since they are no longer in a position to change their pensions. However, one participant suggested that the system could show how much money is expected to be left over after the client’s expected lifetime, so that older clients could think about how they want to distribute their estate.

The participants at one session said they wished they could start using this system immediately. The participants at another session said they would like to try the system for their own personal use.

None of the participants seemed to have trouble answering the question about how the fictitious client was doing. Some of them did,
however, ask clarifying questions about the client’s family situation before answering and one participant asked whether the client owned a home before noticing that there was a mortgage visible in the system. All of them asked about the client’s plans for the future, specifically whether the client was planning to do anything that would require significant funds. All participants asked at some point about the market value of the client’s home and whether its value might have increased after the client bought it. Some participants said they would have liked to see an indicator of the loan-to-value ratio of the client’s home, based on a standardized value for similar homes in the client’s area. Knowing the loan-to-value ratio is useful for the advisor when deciding what options might be open to the client.

The question about whether the client should make any changes rendered rather different responses from advisors. The participants in one session focused immediately on the fact that retirement savings were not going to last throughout the client’s old age. They suggested both saving more and increasing risk. They also looked over the situation simulators and discussed the client’s lack of protection against the different scenarios. The participants in a different session looked at the client’s suggested spending cuts and said that this client should be able to relatively easily cut down on some of her restaurant visits and spend a little less on groceries, in order to put those savings toward the emergency buffer which was insufficient for most simulated catastrophic scenarios. Thus, they moved money between categories (from consumption to saving) with an ease that suggests that they did not see consumption and saving as separate mental accounts. Perhaps this is easier to do from an outsider’s view than for the clients themselves. However, none of the participants seemed to think to experiment with amortization as an alternative to investing. One participant said the client should make a long term plan for her savings, but generally preferred to book a separate meeting for discussing long-term saving in detail.

When discussing investment risk, the participants did not appear to look to the prototype for support. They did not use the visualization to illustrate how more stocks means higher expected return, but lower guaranteed return. Instead, they asked me in my role as client whether I prioritized a higher return or certainty that the money would not lose value. They also asked how I feel when my invested money jumps up and
down in value. This was probably how they were used to talking about risk.

Most participants said they wanted a function for setting savings goals for the medium turn, not just emergency buffer and retirement savings. They wanted to set a goal that lay a variable number of years into the future where they could visualize the effect of risk and the size of monthly savings, like in the retirement savings simulator. Participants in one session said they wanted to show the client how much easier it gets to gather enough capital to make a purchase if they can wait a few years. Participants in two sessions suggested a function that illustrates what would happen if the client cuts spending according to the prototype's suggestions and puts that money toward saving instead. These are both forms of computational offloading; they show users the effects of decisions quickly without manual calculation.

When asked which feature they found most useful, all participants mentioned the situation simulators. They believed the situation simulators would help them sell insurance products, especially life and mortgage insurance. In other words, they believed the system to be effective at persuading users that they were insufficiently protected against short-term catastrophes. The participants also liked having everything in the same view as opposed to separate silos; this is a testament to my efforts to combine investing, insurance, and amortization into a coherent whole.

The participants in one session, the ones who had attended the original design workshop where the issue of information disclosure was raised, said they hoped this design concept would persuade clients to share more information about assets and debts in other banks, since the clients would be able to see that the “diagnosis” would otherwise be incomplete. This may or may not be what happens with real clients. One person who attended a demonstration of the prototype but who was not one of the study participants was concerned about clients' privacy and said their expenses should not be visible in too much detail, since clients might find some purchases embarrassing. This person suggested expenses remain on aggregate levels, such as food, transport, clothes, etc. Only aggregate spending categories were present in the prototype, but we had discussed the possibility of clicking on a spending category to see all specific transactions.
As for potential improvements, participants said they would have liked to see concrete values at points on the timeline, rather than just the cash flow. For example, they wanted to see the value of the client’s home, how much savings was expected to remain after 90 (which was the end of the timeline), and how much was left on the mortgage at retirement. Thus, they were not fully onboard with a complete re-representation from lump sum to cash flow; they still wanted some remnants of the old representation. They also wanted a way to set savings goals other than retirement and to set up periodic savings for the client directly in the interface. Additionally, they wanted the system to visualize the effect of partial goal fulfillment, so that the clients could see that they are better off if they do their best to save and pay down debt, even if they do not succeed in fulfilling their goals completely. Participants also wanted a clearer comparison between alternative futures in the situation simulators, for example by seeing how big of an emergency buffer the client would need if they don’t have unemployment insurance, compared to when they do have it. At the time, the prototype only showed the difference in income but not the difference in required emergency buffer.

5.3.1.4 Discussion
The enthusiasm displayed by the participants was certainly encouraging, although they wanted to preserve certain elements of the lump sum paradigm, such as the value of the home, retirement savings, and mortgage at certain points in time.

I realized at this point that the design concept had a bias toward conservative investments. This was because the timeline had to be clipped somewhere and I had set the maximum age to be 90. Thus, if savings lasted longer than 90, the system did not show how much longer. In other words, there was no visible difference between a sufficient amount of capital and a surplus. However, I was more comfortable with the idea of my design having a conservative bias, so long as the users met their goals, than a bias toward risk like many other retirement planning systems (Bodie, 2003). The issue could perhaps be mitigated by showing total accumulated wealth somewhere in the system, like the study participants had wanted. That is, if it is to be considered an issue. Merton (2014) has argued that a person saving for retirement should not take more risk than is required to meet their target and that, as soon as the target is met, they should reallocate to safe investments in order to avoid losing capital and
ending up below the target again. Thus, the prototype is in line with Merton's investment philosophy.

5.3.1.4.1 Limitations
It could be noted that all participants were employed by the same bank and were thus accustomed to the same company culture, goals, policies, and rhetoric toward clients.

5.3.1.4.2 Reliability and Validity
I have strived to report only results where participants across sessions agreed, or otherwise state the instances where this is not the case. Validity was improved by the inclusion of participants who had not attended the original design workshop, as they provided some corroboration of the usefulness of the functionality. I did not have access to other data coders this time, but one of my supervisors was present at one of the test sessions and could confirm that the results described were consistent with her impression.

5.3.1.4.3 Implications
The results of my evaluations were promising enough that I did not feel the need to pursue other, radically different design ideas. Instead I would continue to explore the effects of this design.
5.3.2 LUMP SUM VS CASH FLOW: A RETIREMENT PLANNING EXPERIMENT

5.3.2.1 AIM
This study was addressed at research question 3: Do non-expert users understand how to use the new representation? Does it really work better than systems that exist today?

I had had my design concept evaluated by personal finance experts and financial advisors and it was now time to try it on the actual target group: Laypeople. I wanted to test whether the design concept would perform better than the online investment advisors I had previously tested. I expected my graph to improve users' understanding by means of the graphical constraining which shows users that the problem is about making the money last through retirement and to help them solve the task by providing computational offloading when calculating how to reach their goal.

However, my design concept was so different and had so many new functions that a straight up comparison to existing systems seemed meaningless. For example, since my prototype had functions for testing the need for emergency savings, and existing online investment advisors didn’t, comparing them with regards to how well they supported users with their questions about emergency savings seemed unproductive.

Instead, I decided to start at the beginning and test the core concept: The translation from money to time. Even though others before me had described the idea, no one had tested it empirically on non-experts. Would they understand how to read the graph?

5.3.2.2 METHOD
In order to measure whether people understood my model better than the wealth-maximizing model used in the online investment advisors, I had to invent an experiment task that could be solved in both systems. The task also needed to have an objectively right or wrong answer, as opposed to depending upon participants’ individual risk preferences; otherwise I would introduce a validity problem when trying to measure the participants’ risk preferences in order to establish a baseline. I formulated the following two hypotheses:

H1 More participants will solve the task in my model than in the control model.
H2 Participants will report that it required less mental effort to solve the task in my model than in the control model.

5.3.2.2.1 Participants
I recruited 24 university employees who were not students or faculty. Most were administrative staff. Participants were compensated for their time. Table 9 provides an overview of the participants.

Table 9: Participants in the retirement planning experiment

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n</strong></td>
<td>24</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>Median 48, SD 12.9</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td>50% M 50% F</td>
</tr>
<tr>
<td><strong>Level of education</strong></td>
<td>0% elementary school 8.3% high school 4.2% vocational training 83.3% university 4.2% postgraduate</td>
</tr>
<tr>
<td><strong>How much experience do you have with investments and securities?</strong></td>
<td>37.5% Almost nothing 25% I have traded mutual funds or other securities with help from an advisor 33.3% I have traded mutual funds or other securities on my own 4.2% I have extensive experience and great knowledge of different financial theories and terminology</td>
</tr>
<tr>
<td><strong>How experienced are you at reading graphs and charts?</strong></td>
<td>29.2% Not very experienced 41.7% Fairly experienced 29.2% Very experienced (e.g. have technical or mathematical training)</td>
</tr>
</tbody>
</table>

5.3.2.2.2 Study Design
To keep the required sample size down, I designed the test as a within-subjects experiment. However, this created the need to have two different datasets; otherwise the participants’ task would have the same answer in both conditions and solving the task a second time would be trivial, if the results agreed, or confusing, if the results disagreed. Thus, the experiment was a 2×2 factorial design – graph×dataset.

The task was to choose an allocation between stocks and bonds such that the resulting predicted wealth would cover a given level of spending
during a given number of years. In other words, participants had to take enough risk to meet the target but no more. The task was designed to resemble the situation that a layperson will typically encounter in real life; planning for retirement with little guidance in regard to how to think and what factors are important to consider (Hershey & Walsh, 2000).

5.3.2.2.3 Stimuli

I built a prototype from the JavaScript libraries D3.js and jQueryUI. It was optimized for the screen size used in the experiment (1680×1050). The prototype fetched synthetic data from an XML file and rendered a graph representing either my model or the control model, depending on a choice made in a pop-up box when the prototype page was started or refreshed. In the same popup box, I could also choose between datasets 1 and 2.

The baseline is pictured in Figure 21, a screenshot of the actual prototype, and Figure 22, which is created to be readable in the printed thesis. The baseline was designed to resemble a simplified version of FinVis, a well-cited financial analytics tool for laypeople (Rudolph, Savikhin & Ebert, 2009). It showed the growth of the simulated portfolio in a traditional fan chart with the worst case outcome as a red line (bottom), the expected outcome as a blue line (middle), and the best outcome as a green line (top). Increasing the proportion of stocks in the portfolio increased the predicted value of the best case and the expected case outcome, but decreased the value of the worst case outcome. In other words, the fan got wider as risk increased.

The treatment chart is pictured in Figure 23 and Figure 24. To make the chart more self-explanatory, I added flags that indicated the different points on the line representing savings; how long it would last in the worst case, the expected case, and the best case. I kept colors and typefaces as similar as possible across conditions.
5. RESULTS

Figure 21: The baseline condition as it looked in the experiment.

50% Stocks

Wealth

50% Bonds

1,952,925 kr

1,364,553 kr

622,372 kr

Age

Figure 22: Recreated for print.
Figure 23: The treatment condition as it looked in the experiment. The dramatic reduction in costs happens because the fictitious person has a mortgage which is repaid at age 72.

Figure 24: Recreated for print.

The prototype predicted growth in the following manner: Bonds were set to grow by 1% per year, based on a historical average (Mehra & Prescott, 1985) and stocks had three cases: The worst case, based on the worst 10-year run in stock market history and corresponding to an annual loss of 1.33% (Al, 2009); the expected case, based on the historical average of 7%
annual growth (Mehra & Prescott, 1985); and the best case, based on the best 10-year run in stock market history and corresponding to an annual growth of 19% (Al, 2009). Test participants could change the allocation between stocks and bonds and see how that affected predicted wealth at retirement. The allocation was the only variable that participants could manipulate. They could change it from 0% to 100% stocks in 5% increments by pulling a slider. Data and scales were selected such that each 5% increment made a visible difference in both graph versions.

The two datasets represented two fictitious people with different ages, assets, debts, income, etc. Both of them also had a mortgage which would be paid off in a few decades, after which their monthly costs would decrease. For the purpose of this experiment, there were only two types of assets: Stocks and bonds. The prototype showed participants an introductory text explaining the difference between stocks and bonds, along with a warning that the growth prediction algorithm in the prototype was simulated and not a reliable basis for financial decisions in real life.

5.3.2.2.4 Procedure
First, participants were assigned by lottery to one of the four test groups until each group had the same number of participants.

Because this was partly a test of the learnability and intuitiveness of the two graphical representations, participants were not given any explanation of how to read the charts. They were informed that the prediction model was a mock-up and that they should not base any real decisions on it, but for the purpose of the experiment they were told to pretend that it was reliable enough.

Participants were given a description of a fictitious person with a given income, wealth, debt, and spending habits. Some of the information was irrelevant, such as loan-to-value ratio and income tax rates. Participants were informed that not all information would be relevant and that they were only to use the information they thought necessary to solve the task.

The fictitious person had some money saved for retirement in a portfolio consisting of 50% stocks and 50% bonds. Participants were instructed to change the allocation between stocks and bonds such that the money would last exactly until the fictitious person was 80 years old;
no more and no less. The asset allocation was the only parameter the participants could control – they could not make the fictitious person save more or consume less. Participants were not explicitly told whether they should go by the expected, the best or the worst stock market scenario, but I had chosen the data such that the task was not possible to solve in the worst case scenario. This was my way of constraining and disambiguating the test task, since basing a calculation on the worst case outcome is arguably an equally valid strategy as is basing it on the expected outcome. Solving the task by relying on the historically best stock market performance was technically possible but counted as a wrong answer.

To make the task more realistic, participants were allowed to use a calculator and pen and paper in both conditions, since they would have access to such cognitive artifacts if they were sitting at home planning their own retirement. However, the calculator should not be necessary in the treatment condition. I decided to take note of whether they chose to use the calculator and/or pen and paper, as a secondary indication of the mental load imposed on the participants when solving the task in the two conditions.

To solve the task in the control condition, participants had to add up all monthly costs, multiply them by 12 to get annual costs, then multiply them by the number of years between retirement and the target age in order to get the total needed amount, and then drag the slider until the expected value of the portfolio hit the closest value above the needed amount. The task was complicated by the fact that the fictitious person had a mortgage; participants had to calculate the point in time at which the mortgage would be repaid and reduce costs for the remaining time period. This is the expected situation for people with mortgages. However, I simplified the task by allowing participants to use a fixed monthly cost for the mortgage, as opposed to mortgage interest being a function of remaining debt.

In the treatment condition, all participants had to do was drag the slider until the edge of the “expected value” line hit the target age. The system automatically compensated for changing expenses such as the disappearing mortgage.

---

99 80 in one dataset and 85 in the other. This was to make the tasks feel different so that participants would not perfunctorily re-use the answer from the first condition in the second condition.
Participants were told that they could think aloud while solving the task and that if they preferred not to, I was going to ask them afterward how they reasoned while solving the task. I wrote down their calculation process. I also wrote down their other verbal comments. This way, I could distinguish between cases where they had genuinely solved the task and cases where they had just guessed at the right answer by coincidence – since the prototype could only move between 0% stocks and 100% stocks in 5% increments, there were only 20 possible answers, meaning that a few participants might provide the right answer by pure chance.

After solving the task in the first condition, participants were asked to answer a shortened version of the NASA Task Load Index (TLX) (Hart, 2006). Question items about physical strain and temporal stress were deemed irrelevant for this experiment and were thus taken out of the questionnaire. Participants were asked to rate their experience on an 18 degree scale regarding the following:

- **Mental Demand.** How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?

- **Performance.** How successful do you think you were in accomplishing the goals of the task set by the experimenter? How satisfied were you with your performance in accomplishing these goals?

- **Frustration Level.** How insecure, discouraged, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Participants then solved the task in the second condition with the other dataset and subsequently answered the TLX questions again. Finally, we briefly discussed the participant’s experience and impressions of the two versions and I asked them which version they would prefer if they were planning their own retirement.

After each test session where the test participant had gotten the wrong answer, I compared the participant’s solutions to the task to their verbal explanations of their reasoning, their scribbles on paper, and the history.

---

100 The questionnaire was translated from English to Swedish.
in the calculator they had used, in order to pinpoint where participants' reasoning or calculations had gone wrong.

5.3.2.2.5 Pilot Studies
I ran three pilot studies in order to work out any problems with the experiment design and to get an indication of the effect size. The first pilot was run on colleagues, who tended to have a technical background, and the second and third on people from the target group, namely people of mixed backgrounds. People in the second and third pilot were employed by the university but were not faculty or students.

The first pilot study showed no significant difference between the two test conditions. The second pilot study, on the other hand, indicated a large effect size for which only 12 participants would be needed to achieve a statistical power of 80%, if the same effect size was observed in the real test.\footnote{Calculated with the StatsToDo sample size calculator, http://www.statstodo.com/SSizMcNemar_Pgm.php.} I ran a third pilot study with 24 participants, but it again failed to show significant results (McNemar's \( p = 0.0736 \)). I noticed that about half the time participants failed the task, it was because they had ignored the test task and done something else, for example picked an asset allocation they thought would "suit" the fictitious person instead of trying to hit a target age. This was true in both test conditions. One participant admitted that they had not really "looked at the numbers." As a result, I clarified the instructions and changed the experiment procedure such that participants had to read the instructions out loud before beginning the task.

In the early pilot studies, I noticed that the treatment condition had a quirk: The "best case" point would often disappear beyond the maximum age, which was set to 120. I therefore changed the maximum age to 260. While 260 is an unrealistic human age, I prioritized visibility of the full line at all times. This resulted mostly in confusion, as participants became uncertain whether they had misunderstood what the timeline represented when it went all the way up to 260. For the full experiment, I therefore changed the maximum age back to 120.\footnote{As noted in the background chapter, there are also other problems with having users choose their own maximum age, as people tend to underestimate their own longevity. At this point, I did not know that.}
5.3.2.3 Results

5.3.2.3.1 Success Rate

The result in Table 10 shows a clear advantage for the treatment graph over the control graph in helping participants solve the task. None of the participants who succeeded with the task in the control condition failed in the treatment condition, whereas 12 participants who failed at the control succeeded with the treatment.

Table 10: Success rate. The cells show the number of participants who succeeded or failed at each condition.

<table>
<thead>
<tr>
<th></th>
<th>Succeeded Control</th>
<th>Failed Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Succeeded Treatment</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Failed Treatment</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Significance (McNemar's)</td>
<td>&lt; 0.01</td>
<td></td>
</tr>
</tbody>
</table>

Therefore, H₁ was not rejected. The records of whether participants used calculators or pen and paper were disregarded in the analysis because many participants would read the task and start calculating before they looked at the graph and realized that manual calculations were unnecessary.

![Figure 25: Reasons participants failed the task.](image-url)
Figure 25 shows the reasons participants failed the task. Participants can make more than one mistake, which is why the total number of errors is greater than the number of participants who failed the task.

The most common reasons for failing the task in the control condition were mistakes in the calculations. For example, participants would forget to account for the expected pension income after retirement, forget to lower expected cost after the age at which mortgage was repaid, or reduce expected cost by mortgage amortization payments only and forget to also subtract mortgage interest. Typically, participants would make a combination of several of these errors. This illustrates the extra cognitive burden and potential for errors when users of wealth-maximizing graphs are left to calculate their own future capital needs. The treatment graph clearly provided impactful cognitive offloading in the context of this test task.

Misconceptions of the control graph itself were seen in three participants who optimized the best case outcome instead of the expected case and in one participant who chose 100% stocks because they did not notice that the worst case outcome grew worse with increasing risk. One participant did not understand the control graph at all and chose 100% stocks in order to make the best case “as big as possible” and the worst case “as small as possible.”

Three participants did not understand the treatment graph and failed the task as a result. A few more had misconceptions about the treatment graph but managed to solve the task anyway. It was clear that participants were not used to reading change on the horizontal axis – several participants said they expected spending level to go up as they increased risk, rather than the number of years. At least three participants read the best, expected, and worst case points as if their value on the vertical axis was the relevant variable; one participant expressed surprise that the worst case was that the fictitious person would have 25,000 a month while the best case was that they would only have 15,500 (see Figure 26). This confusion happened because there were two levels of spending; before and after the mortgage was repaid. When participants pulled the slider to change the allocation between stocks and bonds, the best, expected, and worst case flags would move in the horizontal direction and when the flags passed the point on the timeline where the mortgage would be repaid, they would jump up and down between the higher and lower level of spending. Some participants seemed to focus only on the
resulting movement in the vertical direction, probably because this is the most common way of reading graphs.

![Diagram](image.png)

Figure 26: The worst case and expected case points are located above the best case on the vertical axis, because the red line follows spending and spending goes down after 75. Some participants found this counter-intuitive, because they took the points’ positions on the vertical axis to be what was important, when in fact it was their position on the horizontal axis. Figure recreated for print.

One participant thought the treatment graph showed that the worst case would be that the person would have 25,000 for three years and that the best case was that they would have 15,500 for 27 years. While closer to the intended use, this is still a misconception. The graph shows how long the money will last given expected future needs. In this case, that is 25,000 until the mortgage is repaid and 15,500 thereafter.

At least one more participant thought the red line showed that withdrawal from savings would be 25,000 a month, when it is actually 25,000 minus pension income – 12,000 a month.

One participant thought the best, expected, and worst case were points to be placed at important points on the person's timeline:

No, the best case must happen when he retires, not like this when he is 80, because [when he retires] is when he needs [the money]. Now I’m just reading this here, “best case,” and now we have to decide when the best case should happen.

Thus, making sense of the treatment graph was not as easy for participants as I had expected, because it required them to perform an unusual elementary perceptual task, namely to compare points (the end...
of a line versus a point on a timeline) on a common scale. As mentioned in the background chapter, this perceptual task is normally considered easy for the human eye (Cleveland & McGill, 1984), but it was complicated by its breaking of a rather ubiquitous convention; such comparisons are normally made on the vertical axis, not on the horizontal.

Additionally, I found it hard to choose good labels for the worst, expected, and best outcomes as well as the red line in its entirety. I tried several variants between the different pilot studies and experiment runs, but I never found labels that completely satisfied me. Every label I tried resulted in questions from participants. This could be a sign that the underlying concepts themselves are not easy to understand, if they cannot be easily described.

A few participants erroneously based their answers on the best case outcome instead of the expected outcome. This could be due to a lack of understanding of what the term “best case outcome” entails – that it signifies something that is statistically unlikely. Perhaps “best case” sounds positive to a mathematically untrained user. Several participants also wanted to base their answers on the worst case outcome, which makes more sense from a financial perspective, but I had designed the synthetic data such that it was not possible to meet the target with the worst case. I do not consider it a problem if people want to base their calculation on a worst case outcome – but a best case calculation should be discouraged, so I drew the conclusion that a real system should not show the “best case” outcome.

One participant was confused by the dashed line that represented uncertainty. The participant read the dashed line as though it literally represented savings going off and on. Keeping transparency as the sole representation of uncertainty may therefore be a better choice than using dashed lines.

An interesting phenomenon, observed in both conditions, was that some participants added a “safety margin” by overshooting their target for the expected value by about one step (a 5% increment) on the graphic slider that allocated savings between stocks and bonds. In the control condition, this caused the expected amount to go up by about 30,000 SEK and in the treatment condition, it caused the expected burnout age to go up by about a year. Obviously, the participants perceived this as caution. However, what they were in fact doing was allocating more of the
savings portfolio to stocks – a riskier choice. Thus, this is a conceptual trap that both conditions seemed to lead participants into. This happened twice in each condition, but was much more common in earlier pilot tests when the instructions did not specifically tell participants to make the money last exactly until the target age. It appears that participants who made this error focused exclusively on the expected scenario and did not notice that the “safety margin” actually worsened the worst case outcome. Part of the reason why participants did this was probably the inertia of the worst case line in the case of the treatment condition, as mentioned above. Another reason, for both conditions, might be the lack of transparency regarding the certainty levels of the best case, expected case, and worst case scenarios. Several participants remarked that they were not sure how to use the system when they did not understand the basis for the “expected outcome.”

The problem is clearly illustrated by one participant who had understood the treatment model correctly and also performed all the calculations in the control condition correctly, chosen the correct allocation, but then drastically overshot the correct result by adding a personal assumption about the certainty of the model:

So if there is a 50/50 chance [to reach the expected result], then I would place myself so that there is at least a 70/30 chance that I will meet my needs.

The participant then pulled the slider to increase the proportion of stocks by 25-35 percentage points from the correct allocation. The result was that this participant took much more risk than required\textsuperscript{103}.

Thus, this participant must have understood the task within the narrow context of the experiment, but failed to understand the wider implications of risk and uncertainty; the kind of superficial understanding detached from actual domain knowledge that Scaife and Rogers (1996) warned about. The participant, and others like them, probably read the graph too literally; they took a longer “expected value” line to mean an overall better result, ignoring the worse “worst case” outcome. In this sense, both graphs hid risk in the way criticized by

\textsuperscript{103} This would probably not happen in a Monte Carlo system, which would have showed the participant if their probability of meeting the target went down as a result of the change.
5. RESULTS

Sondergeld et al. (2003), by highlighting the potential advantage of taking more risk while downplaying the potential downside.

5.3.2.3.2 Subjective Experience

Table 11 shows the median scores on the TLX items, although the statistical testing is performed on a per participant basis. A Wilcoxon signed-rank test on the different items in the TLX scale showed no difference in any of the items mental load (p=0.085), task difficulty (p=0.303), satisfaction (p=0.493), or frustration (p=0.461). H2 was therefore rejected.

Table 11: The median scores (standard deviation in parentheses) on the TLX test.

<table>
<thead>
<tr>
<th>How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving? 1 Low effort – 18 High effort</th>
<th>Treatment 11.5 (4.8)</th>
<th>Control 13.5 (4.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>How easy or difficult was it to find the right answer to the task? 1 Easy – 18 Difficult</td>
<td>Treatment 12.5 (5.0)</td>
<td>Control 13 (4.9)</td>
</tr>
<tr>
<td>How successful do you think you were in accomplishing the goals of the task set by the experimenter? How satisfied were you with your performance in accomplishing these goals? 1 Good – 18 Poor</td>
<td>Treatment 8.5 (4.5)</td>
<td>Control 8.5 (4.7)</td>
</tr>
<tr>
<td>How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task? 1 Low frustration – 18 High frustration</td>
<td>Treatment 12 (4.9)</td>
<td>Control 12.5 (5.3)</td>
</tr>
</tbody>
</table>

Most participants would read the task and start calculating retirement needs before even looking at the graph. It is likely that this affected the TLX scores for the treatment graph negatively, because participants would not realize that they did not have to perform the calculations until they had already done it. Since the TLX asks about participants’ experience while solving the task rather than their impression of the system, it is likely that the treatment graph received somewhat unfairly negative scores. Additionally, several participants realized when they saw the treatment graph that they had ignored factors while solving the task.
in the control condition, thus having perceived the control task to be easier than it was.

5.3.2.3.3 Subjective Preference
Table 12 shows which graph was the most preferred by participants. About twice as many preferred the treatment as preferred the control. While a substantial difference, this was less than I had expected given the experiment task.

Table 12: Subjective preference.

<table>
<thead>
<tr>
<th>Preferred the treatment</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preferred the control</td>
<td>7</td>
</tr>
<tr>
<td>Did not like or understand either</td>
<td>1</td>
</tr>
<tr>
<td>Liked both as complementaries</td>
<td>1</td>
</tr>
</tbody>
</table>

One of the most commonly mentioned reasons for preferring the treatment was that it put the numbers into a context which made the decision easier; it became apparent where the money would go and what would be the consequences of different decisions. Of course, another common reason for preferring treatment was that it spared the participants the tedium of calculating retirement needs by hand.

Participants who preferred the control fell into two categories: On the one hand participants who did not quite understand the graphs and failed the task, who preferred the control because it was cleaner, simpler, and contained less information, and on the other hand mathematically sophisticated participants who felt more in control when they could perform their own calculations. The latter group appreciated the way the control showed exact figures.

Additionally, the control was more familiar to some participants. In the words of one participant:

[The control] is based on the way you would normally think. You are accustomed to thinking that you have a certain [amount of] capital. That is also the way the banks have it. But [the treatment ] is actually better. Because here you see all the variables and how long [the money] will last.
5. RESULTS

[The control] is the way you have been taught to think, but [the treatment] is actually better, it's crystal clear once you take the time to look.

One participant who had failed the task in both conditions preferred the control but said the control graph should contain information about the person’s income and spending included in the graph, which indicates that this participant may have wanted something like the treatment but had failed to grasp that this was actually what the treatment represented. Five participants remarked that the treatment graph required more initial cognitive effort to process because it was unfamiliar or because it contained more information. Four of them said they still preferred the treatment once they had figured it out.

Some participants spontaneously talked about how they thought the different graphs influenced their willingness to take risk, without being specifically asked. Interestingly, they seemed to feel “nudged” by the graph, but not all in the same direction. Two said the treatment prompted them to take more risk because the best case and expected case lines were clearly seen to increase while the worst case line looked like it hardly moved. This is a characteristic of the treatment graph that occurs because each step on the timeline is divided by monthly spending; it can require quite a large amount of money to move one step, while on the other hand the lines in the control graph move smoothly with each change. Since the spending in this case had two levels, before and after the mortgage was repaid, lines on the timeline also moved faster after the mortgage, which made it look like the best case line increased much faster than the other ones. One participant asked why one should not simply choose 100% stocks since it looked like the worst case line hardly moved. However, that participant made exactly the same mistake in the control graph too. On the other hand, one participant remarked that the control seemed more focused on growth, while the treatment seemed more focused on security:

[The treatment graph] doesn't look like I will be able to afford a Rolex, it looks like I will be able to afford my rent.

The same participant expressed a wish for the graph to show what could be possible to afford in the future beyond the bare necessities. This indicated to me that there might be a demand for a more sophisticated
version of the treatment graph where spending levels can also be manipulated.

Four participants remarked that it was strictly not necessary to even know the figures for income, spending, or accumulated wealth to solve the task in the treatment condition. They reacted differently to this fact. One participant said, in regards to the control:

This feels like an unnecessary intermediate step, I'm not interested in knowing the total amount, I'm interested in knowing that I can make it.

However, some participants implied that they did not fully understand what they were doing when using the treatment graph. As the treatment graph probably disambiguated the test task and made it easier for the participants to understand what was expected of them in the experiment, there is a risk that the treatment graph encourages the sort of surface-level understanding based on graphical clues that Scaife and Rogers (1996) warned about.

Three participants expressed that the control graph had failed to support their thinking and decision-making, for example:

And here I see that there is no underlying algorithm that helps me, I only see the [predicted] growth. I see a total amount which is difficult to relate to over time. It's easier to see it broken down. Here you must think for yourself and make decisions regarding each of the three outcomes one by one.

On the other hand, four participants said the treatment graph helped them understand how to solve the task and five participants expressed feeling a deeper understanding of the problem, for example by seeing what expenses the money will cover or what will be the boundaries on their time of sustained spending. One participant said it was difficult to know what information was important while solving the task in the control condition, which was the intention behind the experiment design since it resembles the situation in real life.

The problem of choosing where to clip the timeline was still apparent. This is a weakness of the treatment model. I had now tried both alternatives:
• Scale the timeline to the maximum outcome of the best case at 100% stocks, so that the best case is always visible. This caused confusion when test participants questioned whether they had really correctly interpreted the timeline as representing age, when 260 (in this case) isn't a plausible human age.

• Clip the timeline at a more natural human life expectancy. This can make the difference between medium growth and large growth invisible because best outcome and perhaps also expected outcome disappear beyond the maximum age. This is likely to affect the user's willingness to take risk.

However, graphs of the control type show total predicted amounts that can look impressive and may prompt fantasies about what the user could do with all that money. Thus, since graphs with growth curves tend to invoke unwarranted optimism (Tak, Toet & van Erp, 2015), the choice seems to be between a graph that is biased toward more risk and a graph that is biased against risk.

5.3.2.4 DISCUSSION
At the time of this experiment, I had not yet found the literature on retirement planning software reviewed in chapter 2.3.2 or the commercial software described in chapter 5.1.1. I was under the impression that the standard for real retirement planning systems was like the control graph I had used in the experiment, because that seemed to be the state of the art for financial advisory systems in academia.

However, I later learned that contextualizing a lump sum of money in terms of retirement spending needs is actually quite mainstream among pension economists and retirement software developers.

I still learned some important lessons regarding how people read graphs like my design concept when they are unfamiliar with them. Particularly interesting were participants' comments that the graph seemed to make them want to take more or less risk compared to the control. Their statements were contradictory; some participants could present reasons for why the treatment graph made them want to take more risk, while others argued the opposite. I decided to test the influence of the graph on people's risk taking in a new experiment.
5.3.2.4.1 Limitations
An experiment design where participants have some time to familiarize themselves with the different graphs before being given the task would probably show more favorable TLX scores for treatment, but I intentionally avoided teaching the participants how to read the graphs, since I suspected that lack of intuitiveness might be my prototype’s weakest point. I was hoping to see a strong result despite this, thereby making an overall more convincing case. However, I do not want to make excuses; I accept that there was no detectable difference in subjective experience when using the two graphs. Financial planning systems are not used very frequently in a person’s life, so users cannot be expected to remember how they work from time to time. Therefore, learnability is an important factor in the overall usability of a financial planning system.

5.3.2.4.2 Reliability and validity
While not all pilot studies had statistically significant results, the fact that the effect size always pointed in the same direction provides some indication that the result is reliable and replicable. It is also consistent with intuition; a tool that provides substantial computational offloading for the task should get better results.

The question of whether test participants really understood what they were doing is trickier. Can the result be taken to mean that the treatment graph really was better at explaining retirement planning to participants, or were they simply taking clues from the interface that they were expected to make a line stretch to a certain point (see section 2.2.1)? In other words, was it possible to solve the task without fully understanding what the task represented? The statistical result in isolation cannot answer this question, which is why I also collected think-aloud data from participants while they were solving the task. Participants who admitted that they were merely guessing at the answer but happened to choose the right answer by coincidence counted as fails, which was an a priori criterion\(^{104}\). However, participants who expressed a general feeling of uncertainty as to whether they had understood what they were doing but who managed to get the right answer anyway counted as hits. There was not a clear-cut line between participants who seemed confused and participants who merely expressed some uncertainty or self-doubt, which

\(^{104}\) In the final study, this was only one participant and it happened in the control condition.
is why it is difficult to quantify exactly how many participants this concerned. Additionally, I did not record these test sessions, so there are no exact transcripts, only notes taken by hand.

However, the most interesting result is not the statistical hypothesis testing in itself, but the deeper explorations of how participants read and interpreted the graph; that readability was hampered by the way that the treatment graph broke certain conventions. This informed later iterations of the design concept.

5.3.2.4.3 Implications
This study was aimed at answering research question 3: Do non-expert users understand how to use the representation in the design prototype?

Measuring genuine comprehension is difficult. Still, I consider this test a validation that the graph was not completely incomprehensible to most test participants, even though they had not seen it before and received no explanation of how to read it – a purposely harsh test. In subsequent studies, participants received instructions for the graph. Thus, there is some triangulation between studies where participants had to make sense of the graph on their own and studies where they were informed by instructions.
5.3.3 NUDGING TOWARD MORE OR LESS RISK?

5.3.3.1 AIM
This study was targeted at research question 4: Does the new representation “nudge” user to take more or less risk than they would with another representation? The question was inspired by previous experiments, where participants had anecdotally talked about how they felt that the two graphs put them in different frames of mind when choosing an asset allocation. Was any such nudging actually taking place?

For the purpose of this study, I prefer to talk about risk taking as a behavior displayed in the particular context of this experiment, rather than risk preference which connotes a personality trait.\textsuperscript{105}

5.3.3.2 METHOD
Based on the comments from participants in previous studies, I posed the following hypotheses:

- $H_0$ People who use the new graph as decision support for an investment task will not take more or less risk than people who use a representative control graph.
- $H_1$ People who use the new graph will reason more in terms of security and protection against loss than people who use a wealth-maximizing graph.
- $H_2$ People who use a wealth-maximizing graph will reason more in terms of potential profit than people who use the new graph.

5.3.3.2.1 Participants
The participants were students at a technical university. They were invited to participate via email. Additionally, participants were invited to forward the link to the study to their friends. Students at the university were very likely to be more experienced graph readers than the general population, but since this study was going to measure an effect across groups, I was more concerned with getting a homogenous sample and reducing the number of variables than with representing the entire Swedish population. Thus, the results of this study are only indicative of

---

\textsuperscript{105} It could be noted that there was no actual risk taking per se, as the participants did not invest real money. “Risk taking” can therefore be read as “hypothetical risk taking.”
this particular target group. For greater generalizability, similar studies need to be done on other target groups.

Table 13: Participants in the risk taking study.

<table>
<thead>
<tr>
<th>n</th>
<th>119</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Median 24, SD 3.6&lt;sup&gt;106&lt;/sup&gt;</td>
</tr>
<tr>
<td>Gender</td>
<td>45.4% M  54.6% F</td>
</tr>
<tr>
<td>Level of education</td>
<td>5.9% elementary school  51.3% high school  2.5% vocational training  31.9% bachelor’s degree  8.4% master’s degree or higher</td>
</tr>
<tr>
<td>Pre-test risk preference&lt;sup&gt;107&lt;/sup&gt;</td>
<td>14.3% would bet 0 SEK  30.3% would bet 250 SEK  31.9% would bet 500 SEK  10.9% would bet 750 SEK  12.6% would bet 1000 SEK</td>
</tr>
</tbody>
</table>

**Numeracy and financial literacy**

Suppose you have 200 SEK in a savings account. The interest is 10 per cent per year and is paid into the same account. How much will you have in the account after two years?<sup>108</sup>

| Distribution of answers for participants: | Correct 84.5%  Incorrect 14.3%  Don’t know 0.8% |
| Distribution of answers for Swedes:<sup>109</sup> | Correct 35.2%  Incorrect 49.2%  Don’t know 15.6% |

Suppose the interest on your bank account is 1 per cent and inflation is 2 per cent. If you keep your money in the account for a year, will you be able to buy more, as much, or less at the end of the year?

| Distribution of answers for participants: | Correct 71.4%  Incorrect 16.0%  Don’t know 12.6% |
| Distribution of answers for Swedes: | |

<sup>106</sup> Plus five participants older than 47.

<sup>107</sup> See the methods section for the definition of this measurement.

<sup>108</sup> The question in the table is the original English version quoted by Almenberg & Widmark (2011), used here for clarity for non-Swedish readers. The participants in my study answered a Swedish version translated and used by Almenberg & Widmark (2011).

<sup>109</sup> Almenberg & Widmark (2011)
5. RESULTS

| Which has had the higher historical long-run return, stocks or bonds? | Correct 59.4%  
Incorrect 24.0%  
Don't know 16.5%  

Distribution of answers for participants:  
Correct 47.1%  
Incorrect 31.9%  
Don't know 21.0 %  

Distribution of answers for Swedes:  
Correct 66.4%  
Incorrect 22.6 %  
Don't know 11.1% |

As seen in Table 13, all possible risk preferences were represented among participants. There was no statistically significant difference in level (Kruskal-Wallis p=0.982) or variance (Levene's p=0.871) of pre-existing risk preference between experiment groups.

As might be expected from these participants, they scored higher than the general population on the literacy questions that required mathematical ability but worse on the question that required financial knowledge or experience.

5.3.3.2.2 Stimuli

Due to limitations on the number of study participants that I could accommodate and pay, I decided to compare only two graph types. Recall that in section 4.1.1, I found three main types of graphs on the Internet, which I called Stacked Bar Chart, Wealth Burndown, and Component Areas (see Figure 27). Based on my reading of the literature on retirement planning systems, I argued that Wealth Burndown and Component Areas were superior to Stacked Bar Chart. My own prototype was similar to Component Areas, with a few differences. I therefore selected Wealth Burndown as the comparison chart for this test.

Figure 27: Left to right: Stacked Bar Chart, Wealth Burndown, Component Areas.
I realized that any decisions participants made would be dependent not only on the visual design of the chart, but also on the data presented in the chart. I therefore designed two task scenarios; one where conventional wisdom prescribes that a rational person would take high risk and one with less risk. Thus, the experiment was a $2 \times 2$ design – graph $\times$ dataset. The test cases are seen in Table 14.

Table 14: The two test cases.

| Test scenario 1: High risk | Age 25  
Retirement age 65  
Retirement savings 70,000 SEK  
(+ income, retirement income, spending)$^{110}$ |
| Test scenario 2: Low risk | Age 50  
Retirement age 65  
Retirement savings 1,300,000 SEK  
(+ income, retirement income, spending) |

The high risk case was designed such that retirement savings would not last more than a couple of years unless the person took some risk, due to the lower expected growth of bonds. Even though the outcome in case of bad luck on the stock market looked bad, taking some risk was that person’s only chance to acquire enough capital.

The low risk case was designed such that the person could live comfortably on their capital to a high age without taking much risk at all. Taking more risk would only lower the age to which they could live on their capital in the case of a bad stock market run, while pushing the “runout” age in case of normal stock market growth further into the future – toward unrealistic human ages.

I built an online test platform where participants could perform the experiment, at a time and place of their own choosing. The prototype did not accommodate mobile users since I needed people to look at a large enough screen to see the graphs.

Figure 28 and Figure 29 show the baseline graph, Wealth Burndown.

---

$^{110}$ In order to keep the task simple, both fictitious persons were said to spend all of their income and not accumulate more savings beyond what they already had. Their spending habits were equal to their incomes distributed by the percentage of income that the average Swede spends on food, clothing, and other categories according to SCB.
5. RESULTS

Figure 28: The control graph (Wealth Burndown) as it looked in the experiment.

Figure 29: Recreated for print.

Figure 30 and Figure 31 show the treatment graph, Component Areas.
5. RESULTS

5.3.3.2.3 Procedure
Since the dependent variable in the experiment was going to be participants’ risk taking, their pre-existing risk preference was an important factor to control for. I searched the literature on risk preference and found that it cannot be easily measured and reduced to a number. However, for the purpose of this test, I strived to keep test groups as homogeneous as possible. If this pre-existing variable corresponds to

---

111 The alert reader may notice that I have previously expressed skepticism against the idea that risk preference can be easily measured and reduced to a number. However, for the purpose of this test, I strived to keep test groups as homogeneous as possible. If this pre-existing variable corresponds to
preference tests and found in a literature review by Charness, Gneezy and Imas (2013) that the Gneezy and Potters method is a simple and satisfactory means of measuring risk preferences. As such, it is suitable for a pre-test questionnaire. The Gneezy and Potters method gives a simple, numeric measurement of risk preference, namely how much of a gifted amount of money that a person would bet in a lottery\textsuperscript{112}.

The test began with a questionnaire about the participant’s age, gender, education, and risk preference. An algorithm assigned test participants to groups. The algorithm simply checked the database for existing data records of people who had responded with the same risk preference and assigned the new participant to the group with the smallest number of identical risk preference responses.

Next, the system asked the participants three questions measuring numeracy and financial literacy. The questions were taken from Almenberg and Widmark (2011), who have in turn translated them to Swedish from the well-known international studies of financial literacy by Lusardi and Mitchell (2011). I hesitated to use the full numeracy and financial literacy indices; partly because there are 13 questions and I did not want to deter people from participating in the study because of a measurement that was not a key part of the research question\textsuperscript{113}, and partly because I had no way to stop participants from using a calculator when they were doing the experiment online. The original Almenberg and Widmark study was done through telephone interviews and I suspected that the respondents in that study were caught more “off guard” when given math problems. Asking people who are sitting at a computer to answer simple arithmetic problems would almost certainly render mostly correct responses, which makes those problems useless for distinguishing people by math proficiency. I decided to keep one numeracy question about compound interest, which requires such a level of understanding anything real, then at least I have distributed risk-lovers and risk-haters as evenly as possible though the measurement may not be perfect. If it doesn’t, then no harm is done.

\textsuperscript{112} The amount and the odds vary, but the important thing is that the expected value of playing the lottery is higher than the expected value of keeping the money without playing the lottery. I chose 1,000 SEK as the initial amount and odds of 50\% chance of winning 2.5 times the money and 50\% chance of winning nothing. I restricted the number of possible answers to the risk preference question to 0, 250, 500, 750, or 1,000 SEK. This lowers the preciseness of the measurement, but I figured that it was already a merely ordinal scale and that the initial choice of every number between 0 and 1,000 was itself arbitrary. In the original Gneezy and Potters method, participants receive actual money that they bet. Since this was not possible in this case, I asked participants how they would place a hypothetical bet.

\textsuperscript{113} The longer the questionnaire, the bigger the risk that people will lose patience and quit the experiment.
that an uninformed person with a calculator will still get it wrong, and two financial literacy questions. The financial literacy questions were multiple-choice, so I needed more than one literacy question in order to distinguish knowledgeable people from those who just made a lucky guess\textsuperscript{114}. I picked questions of medium difficulty according to Almenberg’s and Widmark’s data about the percentage of Swedes who answer these questions correctly. I later learned that Lusardi and Mitchell (2011) had already decided to condense their financial literacy instrument down to only three questions. Luckily, the questions they had kept were equivalent to those I had chosen\textsuperscript{115}.

After agreeing to the terms of the study and answering the background questions, participants were given a short introduction to stocks and bonds, in order to ensure that all participants were on at least a minimal level of knowledge. This minimal level of knowledge was that bonds are considered safer than stocks, but that stocks tend to grow more over long periods. After this introduction, participants were hopefully more equipped to understand the choices that were about to be asked of them. After that introduction, participants were introduced to a wizard that showed them step by step how to read the chart in the subsequent experiment task. Since this was not a test of the charts’ intuitiveness and readability, but rather their influence on risk taking, I wanted to make sure that participants understood what they saw. After participants closed the wizard, they were presented with the experimental task. In order to move on to the next stage, they first had to choose an allocation and explain their choice in a text field. They were then shown a wizard explaining the next graph and the next experiment task was conducted. Finally, test participants were asked to choose which chart they liked best. They were reminded of what both charts looked like by screenshots of the graph, which were presented next to each other in random order. Participants could choose to explain their preference in a text field.

\textsuperscript{114} Almenberg and Widmark recorded correct answers, incorrect answers, and missing answers. I take the latter to mean that the respondent in the telephone interview didn’t know and declined to answer. To mimic this, I added the option “Don’t know” to the multiple-choice questions, though I expected that most people who were unsure of the answer would make a guess rather than saying “Don’t know.”

\textsuperscript{115} The main difference was that Lusardi and Mitchell kept a question about risk diversification (whether an equity fund is safer than individual stocks), while I kept a question about asset classes (which is safer, stocks or bonds). Additionally, Lusardi and Mitchell made all questions multiple-choice (the compound interest question had the answer options whether the resulting balance on the account would be more than, equal to, or less than $102), while I kept the compound interest question open-ended in order to maintain comparability to Almenberg’s and Widmark’s (2011) data.
On the final page of the experiment platform, participants were given a code that they could send to me to redeem a movie ticket as compensation for their time in the experiment. This code was a protection against people falsely claiming that they had participated in the study. The code was generated according to a pattern known only to
me, in order to protect against fake codes. For extra security, I could have recorded all generated codes in the database to confirm that they had indeed been generated in the study. I decided against that, because I had promised my participants that I would not connect their identity to their exact responses.

Figure 32 summarizes the test procedure.

5.3.3.2.4 Pilot Study
I ran a pilot study where I observed people doing the experiment and asked them to think-aloud about the experiment task and the platform. These participants are not included among the final results. They were a convenience sample of four people of different ages and backgrounds.

The pilot testers’ reasoning was clearly influenced by their previous knowledge and attitudes to stocks and bonds. One participant remarked that the growth of stocks depends on which stocks you choose – presumably, the participant alluded to the fact that bad stocks will perform worse than average. Another participant asked whether “bonds” referred to corporate or government bonds\(^ {116}\) and added that government bonds might be untrustworthy in times of political turbulence. I briefly entertained the idea of anonymizing the assets into “Asset type A” with a known growth rate which is stable but low and “Asset type B” with a more volatile but higher growth. This would steer participants’ attention away from the asset types themselves and away from rules of thumb they may have heard of\(^ {117}\) and perhaps allow them to focus on the graphs; thus a more “pure” comparison of graphical representations of risk, less confounded by pre-existing attitudes. In the end, I thought that experiment design would be too artificial and decided to keep using the names of recognizable asset types, for the sake of ecological validity. I clarified the task so that instead of allocating between “stocks” and “bonds,” participants were asked to divide their retirement investment between an “equity fund” and a “fixed-income fund.” I also made their (made-up) historic performance explicit: The instruction now told participants that the equity fund had an annual average growth of 8% and the fixed-income fund 2%.

\(^ {116}\) Corporate bonds can be quite risky.\(^ {117}\) For example, the principle that you should keep 100 - (your age)% of your portfolio in stocks and the rest in bonds (CNN Money, n.d.)
Like participants in earlier studies, some pilot testers wanted to manipulate the monthly consumption level instead of the length of the retirement period. One pilot tester explicitly said they wanted the system to suggest their expected longevity based on a statistical table, just as Turner and Witte (2009) recommended. I made a note of remembering this as a future development of the prototype.

Figure 33: The pilot testers reacted to the fact that “Alcohol and tobacco” was one of the categories that visually exceeded the income. (Detail, recreated for print)

Interestingly, several pilot testers noticed that in the prototype of Component Areas, the category “Alcohol and tobacco” was one of the areas that visually exceeded the post-retirement income line (see Figure 33). This was a coincidence and not something I had purposefully designed. The category Alcohol and tobacco was, along with all the other categories, fetched from Statistics Sweden’s (SCB) data about average
Swedes’ spending habits. As alcohol and tobacco might be deemed an “unnecessary” expense, several pilot testers implied that an alternative solution to adjusting their investment plan might just be to drink and smoke less after retirement.

This was interesting because I had expected participants to view the fictional person’s spending as a whole, which can be or not be covered by their income. I did not expect them to look at which categories happened to be above the income line. For comparison, no participant mentioned the category Hotels, cafés, and restaurants which was of similar size to Alcohol and tobacco and arguably as non-essential. But Hotels, cafés, and restaurants was near the bottom of the chart and not visually “above” the fictional person’s income.

An interesting future study would be to explore how the order of spending categories in the chart influences people’s attitudes and decisions. I conjecture that a chart with spending categories that are considered essential or non-negotiable at the top, such as mortgage, transport, or healthcare, will have fewer users think about how they can cut their spending and more users who focus strictly on how they can increase their post-retirement income.

Alternative explanations for the pilot tester’s commenting on alcohol and tobacco, but not on restaurants and cafés, might be that they simply read the first couple of categories from top to bottom, regardless of the income line. Or they might have viewed restaurants and cafés as a more justifiable expense than alcohol and tobacco. In either case, they would not necessarily read categories above the income line as “causing” the shortfall. Further experiments could determine whether the effect I observed really exists, but that is outside the scope of this thesis.

### 5.3.3.3 Analysis
Recall that each participant was required to explain their choice of asset allocation, once for the low risk scenario and once for the high-risk scenario. The explanation was in the form of a free-text comment. Participants could also choose to leave a comment detailing why they preferred one graph over the other. All but six participants provided a rationale for their preference.

---

118 I do not have exact transcripts of what the pilot testers said. Some may have just laughed at the idea that alcohol and tobacco was what they were supposed to strive to afford after retirement.
After looking through all comments, I created a coding scheme based on themes I saw in the data. I coded a subset of the comments and had another researcher code the same subset, as described in Hruschka et al. (2004). We checked the Cohen’s $\kappa$ for this subset and discussed the places where we had disagreed. This led to a gradual refinement of the codes with increasing clarity and decreasing ambiguity. We repeated this process until Cohen’s $\kappa$ exceeded 0.6 for all codes.

5.3.3.4 RESULTS

5.3.3.4.1 Influence on Risk Taking

5.3.3.4.1.1 Low Risk Scenario

For the low risk scenario, there was no statistically significant difference in risk taking among participants who looked at a Wealth Burndown chart, compared to participants who looked at a Component Areas chart. Details are found in Table 15. Figure 34 shows that the variances were similar, too – there were no visible differences in how the data clustered.

Table 15: Results for the low risk scenario.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Median risk Wealth Burndown (n=55)*</td>
<td>35.00</td>
</tr>
<tr>
<td>Median risk Component Areas (n=64)</td>
<td>40.00</td>
</tr>
<tr>
<td>Mann-Whitney U</td>
<td>1,745.500</td>
</tr>
<tr>
<td>Wilcoxon W</td>
<td>3,825.500</td>
</tr>
<tr>
<td>Z</td>
<td>-0.078</td>
</tr>
<tr>
<td>Significance (two-tailed)</td>
<td>0.938</td>
</tr>
</tbody>
</table>

*Risk here means share of portfolio allocated to stocks.
5. RESULTS

A power analysis\textsuperscript{119} revealed that an effect of this size would require about 2,750 participants to detect. Thus, it is possible that there is a detectable difference on the scale of thousands of users, but this difference is likely to be so small as to be of no practical significance.

5.3.3.4.1.2 High Risk Scenario

In the high risk scenario, there was a statistically significant difference whereby users of the treatment graph Component Areas took less risk. Details are in Table 16. Figure 35 shows the difference in variance between groups.

Table 16: Results for the high risk scenario.

<table>
<thead>
<tr>
<th></th>
<th>Median risk Wealth Burndown (n=64)</th>
<th>Median risk Component Areas (n=55)</th>
<th>Mann-Whitney U</th>
<th>Wilcoxon W</th>
<th>Z</th>
<th>Significance (two-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>92.50</td>
<td>80.00</td>
<td>1,322.500</td>
<td>2,862.500</td>
<td>-2.402</td>
<td>0.016</td>
</tr>
</tbody>
</table>

\textsuperscript{119} The power analysis was done by calculating the required sample size for a t-test through the calculator at https://www.stat.ubc.ca/~rollin/stats/ssize/n2.html, and then increasing it by 15% (see IBM Support, n.d.). The calculation was based on a significance level of 5% and a statistical power of 80%.
5. RESULTS

Figure 35: The variance of responses in the high risk scenario. Note that there are fewer responses in the treatment group, so the heights of the bars are not directly comparable.

The effect size, calculated as $r = Z/\sqrt{n}$ (Fritz, Morris & Richler, 2012), was 0.22. According to guidelines, this is considered between a small effect ($r=0.1$) and a medium effect ($r=0.3$) (Fritz, Morris & Richler, 2012).

Thus, $H_0$ was rejected. The median portfolio in the control group had 12.5 percentage points more funds allocated to stocks than the median portfolio in the treatment group. The treatment graph appears to encourage less risk taking than the control graph in the high risk scenario.

5.3.3.4.2 Influence on Reasoning

The thematic codes that emerged from the coding process are found in Table 17. Although some codes look like pairs of opposites, they were not all mutually exclusive.

Table 17: Coding scheme for participants’ rationales for their asset allocation decision.

<table>
<thead>
<tr>
<th>Code</th>
<th>Definition</th>
<th>Example Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic</td>
<td>Does not appear to use the information in the graph</td>
<td>I choose to use 100 - age i.e. 100 - 50 = 50. 50% equity funds and 50% fixed-income funds.</td>
</tr>
<tr>
<td>Specific</td>
<td>Could not have made this choice without looking at the graph</td>
<td>I just figured that with poor stock market development the money will last until age 90 which is a fairly advanced age, probably above life expectancy and additionally the money will last a long time if stock market development is average.</td>
</tr>
<tr>
<td>Security</td>
<td>Says explicitly that they want security or that they want to avoid risk</td>
<td>To minimize the risks and secure a certain income for a longer time.</td>
</tr>
</tbody>
</table>
Gamble

Says explicitly that they are willing to take risk. Participants who say they feel they have no choice but to take risk are not included. Since savings will last less than a year regardless of asset allocation, I figure that a higher risk is worth taking.

Sustainability

Reasons in terms of years of sustained consumption, or talks about maintaining the same lifestyle. I figured that my life expectancy ought to be about 94 so I adjusted the allocation by worst case scenario. I.e. so that savings would last only until 94 with poor stock market development.

Profit

Reasons in terms of monetary losses or gains and appears to focus on profits, or talks about a more luxurious lifestyle. Going by the graph, I choose equity funds because the return is higher. Since I'm only 25 [in the example] I will take more risk since the chance of winning back lost money is greater.

5.3.3.4.2.1 Low Risk Reasoning

Table 18 shows the relative frequencies of utterances between the two graph conditions in the low risk scenario.

Table 18: Frequency of codes per free-text comment in the low risk scenario.

<table>
<thead>
<tr>
<th></th>
<th>Wealth Burndown (n=55)</th>
<th>Component Areas (n=64)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic</td>
<td>18.2%</td>
<td>23.4%</td>
</tr>
<tr>
<td>Specific</td>
<td>65.5%</td>
<td>67.2%</td>
</tr>
<tr>
<td>Security</td>
<td>45.5%</td>
<td>37.5%</td>
</tr>
<tr>
<td>Gamble</td>
<td>10.9%</td>
<td>9.4%</td>
</tr>
<tr>
<td>Sustainability</td>
<td>54.5%</td>
<td>65.6%</td>
</tr>
<tr>
<td>Profit</td>
<td>21.8%</td>
<td>25.0%</td>
</tr>
</tbody>
</table>

An independent samples t-test per code showed that there were no statistically significant differences in participants' reasoning in the low risk scenario.
5.3.3.4.2.2  **High Risk Reasoning**
Table 19 shows the relative frequencies of utterances between the two graph conditions in the high risk scenario.

<table>
<thead>
<tr>
<th></th>
<th>Wealth Burndown (n=64)</th>
<th>Component Areas (n=55)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic</td>
<td>32.8%</td>
<td>38.2%</td>
</tr>
<tr>
<td>Specific</td>
<td>43.8%</td>
<td>43.6%</td>
</tr>
<tr>
<td>Security</td>
<td>25%</td>
<td>38.2%</td>
</tr>
<tr>
<td>Gamble</td>
<td>35.9%</td>
<td>27.3%</td>
</tr>
<tr>
<td>Sustainability</td>
<td>31.3%</td>
<td>36.4%</td>
</tr>
<tr>
<td>Profit</td>
<td>46.9%*</td>
<td>27.3%*</td>
</tr>
</tbody>
</table>

An independent samples t-test showed that the only statistically significant difference was in the Profit category, p=0.027 (equal variance not assumed). If participants who looked at the Wealth Burndown chart were more likely to reason in terms of profit in the high risk scenario, then that is consistent with the fact that they also chose higher risk in that scenario. It is also in line with expectations; it makes sense that people who see wealth grow would be more inclined to think about money, profits, and a wealthier lifestyle. However, the difference in reasoning is not significant after a Bonferroni correction\textsuperscript{120}.

Thus, both H\textsubscript{1} (security) and H\textsubscript{2} (profit) were rejected. No difference in how test participants reasoned could be demonstrated.

5.3.3.4.3  **Subjective Preference**
Participants were asked which graph they personally preferred. Table 20 shows that preferences were so evenly distributed that statistical testing is unnecessary. There is clearly no graph that is commonly preferred over the other.

\textsuperscript{120} Significance level 0.05, 12 comparisons (6 codes per scenario). The corrected significance level is 0.0042.
5. RESULTS

Table 20: Proportion of participants who preferred one graph over the other

<table>
<thead>
<tr>
<th>Preferred Component Areas</th>
<th>61</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preferred Wealth Burndown</td>
<td>58</td>
</tr>
</tbody>
</table>

We coded participants’ explanations to why they preferred one graph over the other and discussed the codes until we reached a Cohen’s \( \kappa > 0.6 \). Codes were not mutually exclusive; participants could say that they found a graph both more readable and more informative, for example. Table 21 explains the definitions of the codes we used and Table 22 shows the frequencies of the codes for the two graphs.

Table 21: Coding scheme for participants’ rationales for their graph preference.

<table>
<thead>
<tr>
<th>Code</th>
<th>Definition</th>
<th>Example Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readable</td>
<td>Prefers this graph for aesthetic reasons, or because the graph itself is easier to understand.</td>
<td>It’s simpler, with fewer elements and colors.</td>
</tr>
<tr>
<td>Informative</td>
<td>Considers this graph better at explaining the real-world consequences of decisions.</td>
<td>It shows clearly what type of standard of living to expect.</td>
</tr>
<tr>
<td>Distrusts spending assumptions</td>
<td>Does not want to mix capital growth with spending predictions in the same graph, or points out that they are unlikely to retain the same spending habits in old age. Therefore prefers the graph that they perceive not to depend on this assumption.</td>
<td>Your spending can vary so much, so you can’t base too much on that. The upper graph, on the other hand, shows much more clearly the money relative to age.</td>
</tr>
</tbody>
</table>

Table 22: Reasons why participants preferred one graph over the other.

<table>
<thead>
<tr>
<th></th>
<th>Wealth Burndown (n=58)</th>
<th>Component Areas (n=61)</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readable</td>
<td>50%</td>
<td>26.2%</td>
<td>0.007</td>
</tr>
<tr>
<td>Informative</td>
<td>37.9%</td>
<td>73.8 %</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Distrusts spending</td>
<td>29.3%</td>
<td>0%</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

* Equal variances not assumed
All of the coded dimensions were significantly different in a Bonferroni-corrected independent samples t-test. Thus, participants who preferred the Wealth Burndown chart tended to do so because it was simpler and easier to understand at first glance, while participants who preferred Component Areas did so because they perceived it as more informative and better at explaining the actual consequences of decisions.

Interestingly, only Component Areas provoked comments questioning the assumption that spending could be held constant. Wealth Burndown was based on the exact same assumption in order to calculate the future date when wealth would have “burned down,” or run out. However, Wealth Burndown does allow users to see how capital grows independently of spending assumptions up until the point when the user retires and capital starts decumulating.

5.3.3.5 DISCUSSION
Previous literature has raised the concern that retirement planning systems have a pro-risk bias (Bodie, 2003). This bias arises when increasing investment risk (and thereby expected growth) looks like it will save an underfunded retirement plan.

Due to the way that Component Areas cuts off “excessive” wealth and makes it invisible after the maximum age in the chart, I speculated that Component Areas would instead have a conservative bias compared to conventional graphs focused on building wealth. That idea found some support in this study; in a fictitious test case designed to encourage risk-taking, where the person was young and had modest wealth, participants took less risk when they looked at Component Areas than participants who looked at a wealth-maximizing control graph. The effect was not seen when the fictitious test case was older and wealthier.

It is difficult to make normative claims about which level of risk-taking was more “appropriate.” While the high risk scenario was designed to encourage more risk-taking compared to the low risk scenario, both participant groups behaved as expected with regards to choosing more risk in the high risk scenario than in the low risk scenario.

It can be noted that a clear majority of participants in the previous study preferred a graph of Component Areas type, while no favorite was discernible in this study. The reason is most likely that the control task in the previous experiment required manual calculations, while the mental effort required to solve the task in this experiment was much more similar between conditions.
Participants who preferred Wealth Burndown tended to do so because it was simpler, while participants who preferred Component Areas found it more informative. This confirmed results from the previous experiment, where the same pattern was observed. It is possible that Component Areas needs some getting used to, but that it is more helpful once the viewer understands it. A parallel could also be drawn to Ancker et al. (2006), who showed that the graphs that people say that they prefer are not necessarily the ones that help them draw the best conclusions, since people tend to like oversimplified graphs (see section 2.2.5).

5.3.3.5.1 Limitations
This study suffers from the same problem as does any study that involves hypothetical money: It cannot be known how well participants’ behavior at the experiment reflected their risk preferences with their real retirement savings. The absolute risk levels that participants chose in this experiment should not be taken at face value. However, my intention was not to measure absolute risk preference but the difference in risk taking between conditions. Since both conditions involved hypothetical money, any observed difference ought to be attributable to the graph visualization.

5.3.3.5.2 Reliability and Validity
The replicability of this study ought to be high, since it was distributed online and all code, including the wording of instructions, can be distributed to other researchers. I also eliminated my own presence as a possible biasing factor.

The downside is that I could not observe the participants and do not know how much time they spent on the task or how carefully they read the instructions. Again, the free-text comments were intended to compensate for this, in the same way that I had used think-aloud data in previous studies. The free-text data showed that most participants seem to have taken the task seriously. Since the sample size was larger this time, I chose not to eliminate the very few participants who admitted in the comments that they had simply clicked without thinking.

In this study, I had access to a co-coder and we validated the coding of the free-text comments by means of the Cohen’s $\kappa$, which increases the reliability of the interpretation of the free-text comments.

The validity of the study – that is, did it really measure risk preference? – has already been discussed under the Limitations section.
5.3.3.5.3 Implications
This study was designed to answer research question 4: Does the new representation “nudge” people to take more or less risk than they would with a different representation? In a test scenario where the participants imagine themselves to be younger and poorer, participants who used the new representation took less risk than did participants who used a control. The effect was small to medium in size – both groups took high risk. When participants imagined themselves as older and wealthier, there was no observable difference between groups and both groups took less risk.

Both groups behaved as expected according to conventional financial planning principles, which suggest that risk taking should decrease with age. If the designer of a financial decision aid thinks that a young person of modest wealth should allocate 100% of their investment funds on the stock market, then they might want to consider using a wealth-maximizing graph or complementing Component Areas with information that “nudges” young users toward more risk. This decision is up to the judgment of the designer.
5.4 IMPROVING THE Prototype

In my empirical studies, I had seen people struggle to understand a graph that moved in the horizontal direction, at least without explanation. Additionally, participants in several studies had remarked that it was difficult to choose an end date and requested a function that would allow them to change their monthly spending instead of their maximum age.

However, what made me definitively decide to change my design concept was Turner and Witte's (2009) thorough analysis of the problems with asking users to choose their own target maximum age. I could not ignore studies that showed that people systematically underestimate their own longevity. Additionally, removing the possibility to choose the target date would give users one less variable where they could be influenced by anchoring to irrelevant numbers.

It was relatively simple to see what this meant for the prototype: Where previously the monthly spending had been held constant and the target age variable, target age should now be constant and spending variable. This meant that there would be two lines, one for the expected case and one for the worst case, at different points on the vertical axis. Manipulating savings, growth, and risk would cause the lines to move on the vertical axis, as is more in line with convention than the old design, which had the two lines at the same height but growing or shrinking to different lengths in the horizontal dimension. The new concept is pictured in Figure 36 and Figure 37. The purple lines are now a fixed length and increasing or decreasing savings will cause them to move up or down.
Figure 36: The prototype after a conceptual change.

Figure 37: Recreated for print.

While a simple graphical change, this constitutes a fundamental change to the underlying model; instead of risk in terms of years of sustained consumption, the new prototype presents risk in terms of monthly allowed spending. However, this new model of risk fits much better with what Kotlikoff (2008) argued: That risk should be thought of as potential impoverishment of lifestyle.

The new model allows more flexibility in exploring different solutions to the retirement puzzle; it can cut spending as well as increase savings.
To this end, I added a function where users can make custom adjustments to their budget (upwards as well as downwards) between two arbitrarily chosen years. For example, users can see the effect of setting their spending on transport by car to zero. For essential spending categories, such as groceries, I implemented a function that warns users if they try to enter a spending level that goes below recommendations from the Swedish Consumer Agency, so that they do not unwittingly make a budget that is unrealistic. All the risk scenarios from the previous version are preserved and it is still possible to experiment with amortization as an alternative to investing, which to my knowledge is a unique feature found in no other financial planning system.

While making these changes, I also strived to incorporate all the intricacies of the Swedish retirement system in the model. This necessarily made the system more complex, but I had realized that the previous version was oversimplified. The new version separated state pension from occupational pension and private savings, which can all be invested in different asset allocations and which can all be withdrawn at different rates. Private savings can be withdrawn completely at the consumer’s discretion, but for the sake of simplicity I implemented only two withdrawal functions in the prototype: One that withdrew the same amount each year and one that based its withdrawal amount on the expected spending for each year.

The state pension depends on the person’s salary and years of participation in the workforce, and only a small portion is invested according to the user’s preferences; the rest is considered in the model to be an extension of the user’s income that continues after retirement. The model does not currently account for “pensionsbromsen”121, the risk that state pensions are stunted in such a way that they do not keep up with inflation in times of economic turbulence.

In order to account for the risk of a house market crash, which has previously been ignored in financial planning systems even though it is an important risk (Turner & Witte, 2009) I added a function for selling the user’s home in order to fund retirement and exposed the price of the home as a parameter that could decline, which showed the overall effect on the user’s retirement plan if the user could not get their expected selling price.

---

121 Literal translation: “The pension brake.”
5.4.1 **EXPERT VALIDATION**

I asked a personal finance expert to review the prototype in order to confirm that the new model made sense. The expert had seen the prototype in an earlier stage and was the one who had recommended me to separate retirement savings from buffer savings.

The expert confirmed that the new model was sound, with the addition that occupational pensions can come in many forms and not all of them allow lifelong withdrawals. However, an additional comment was that it was good to encourage as long withdrawal times as possible. Moreover, the prototype should inform the user that their real costs can change if the relation between an increasing salary and inflation changes.

The expert called the prototype “tasteful, well-thought-out, and suitable.”

5.4.2 **NON-EXPERT USERS USING A FULL VERSION OF THE PROTOTYPE**

5.4.2.1 **AIM**

This study served to further explore research question 3: Do users understand how to use the representation in the prototype and does it serve their needs better than other tools on the market?

For this study, I decided to focus on the latter part of research question 3; not so much whether the prototype was intuitive and easy to use at-a-glance, but rather how well it supported users once they had understood how to use it. In other words, these study participants would be instructed on how to read the graph before they took the test. This would allow me to observe the behavior of users who were fairly familiar with the prototype and get their relatively more informed opinions on how well the prototype supported their needs.

I focused this aim to answer the research question: Can study participants identify problems in a person’s financial situation by looking at the prototype?
A control representation of the same information was needed in order to ensure that exemplary financial problems were not so trivial as to be obvious from merely looking at figures that can be found in contemporary financial management tools: Income, spending, debt, and projected retirement income, among others.

5.4.2.2 Method

5.4.2.2.1 Participants
I recruited 12 people aged 45-62 to participate in the study. They were compensated for their time. The sample was a combination of volunteers who had signed up to previous studies but not yet participated, people from the social networks of people around me, and people recommended by other participants (snowball sampling). I did not know any of the participants personally. Participants were told that I was primarily looking for people who were not finance professionals. Details about the participants are found in Table 23.

I was interested in this target group because due to their age, they were likely to have thought about retirement before coming to the test and they might offer interesting perspectives on what support they were missing in their planning efforts. People who are already retired were not the target group of the system and might be less motivated to explore and discuss it.

All of the participants were women. Women are an important target group because they typically get lower pensions, while at the same time needing savings to last a longer time after retirement because of their longer life expectancies.

Table 23: Participants in the final usability study.

<table>
<thead>
<tr>
<th>n</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Median 56, SD 3.95 (1 did not answer)</td>
</tr>
<tr>
<td>Gender</td>
<td>100% F</td>
</tr>
<tr>
<td>Level of education</td>
<td>0 elementary school 3 high school 1 vocational training 5 bachelors 3 masters or higher</td>
</tr>
</tbody>
</table>
## Occupation

<table>
<thead>
<tr>
<th>Occupation</th>
<th>2 teachers/professors</th>
<th>2 managerial positions</th>
<th>2 administrators</th>
<th>1 IT</th>
<th>1 cultural worker</th>
<th>1 accountant</th>
<th>1 scientific assistant</th>
<th>1 public relations officer</th>
<th>(1 left blank)</th>
</tr>
</thead>
</table>

## Financial experience

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Have you ever traded stocks or mutual funds on your own, without help from an advisor?</td>
<td>42% Yes</td>
</tr>
<tr>
<td>Have you thought about planning your retirement?</td>
<td>75% Yes</td>
</tr>
<tr>
<td>Have you done anything concrete to plan your retirement (Calculations, gathered information, asked for advice, etc.)?</td>
<td>58% Yes</td>
</tr>
</tbody>
</table>

## Numeracy and financial literacy

<table>
<thead>
<tr>
<th>Question</th>
<th>Participants</th>
<th>Swedes</th>
<th></th>
</tr>
</thead>
</table>
| Suppose you have 200 SEK in a savings account. The interest is 10 per cent per year and is paid into the same account. How much will you have in the account after two years? | Distribution of answers for participants:  
Correct 58.3%  
Incorrect 41.7%  
Don’t know 0.0%  
Distribution of answers for Swedes:  
Correct 35.2%  
Incorrect 49.2%  
Don’t know 15.6% | |
| Suppose the interest on your bank account is 1% and inflation is 2%. If you keep your money in the account for a year, will you be able to buy more, as much, or less at the end of the year? | Distribution of answers for participants:  
Correct 91.7%  
Incorrect 0.0%  
Don’t know 8.3%  
Distribution of answers for Swedes:  
Correct 59.4%  
Incorrect 24.0%  
Don’t know 16.5% | |
| Which has had the higher historical long-run return, stocks or bonds?     | Distribution of answers for participants:  
Correct 83.3% | |

---

122 Almenberg & Widmark (2011)
It is worth noting that the cohort that is currently middle-aged performs better at numeracy and financial literacy tests than both the younger and the older (Almenberg & Säve-Söderbergh, 2011). These participants thus belong to the most financially competent age group. Table 23 reveals that the participants had higher levels of numeracy and financial literacy than the average Swedish adult. The participants were highly educated and worked white-collar jobs.

Testing on younger users and on users with less education might reveal greater difficulties reading graphs and/or understanding financial concepts.

5.4.2.2.2 Stimuli
Based on the literature on risks in personal finance and in particular the situation in Sweden, I designed three test cases with numbers representing three fictional people. All fictional people were middle-aged women, because I did not want differences in test participants’ judgments to be attributable to differences in the test cases’ age or gender. I did not give the fictional people names, in order to avoid any associations to socioeconomic class. The test cases were intended to represent people with the financial situations in Table 24.

---

123 I chose to use middle-aged test cases because I was interested in using test participants in that age range, since they were likely to have started thinking about retirement. Additionally, I wanted to represent problems that could not always be easily solved simply by saving a little more (young people have the advantage of a long time before retirement). The test cases were women rather than men because they are a more important target group: low pensions are more common for women and at the same time, their retirement savings need to last them through a longer life as retirees.
Table 24: Fictitious test cases.

<table>
<thead>
<tr>
<th>Person A</th>
<th>The control case. Average income and average spending according to data from SCB\textsuperscript{124}. No particular problems.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person B</td>
<td>Mortgage too large. Currently making ends meet only because of historically low interest rate. Very sensitive to rising interest rate.</td>
</tr>
<tr>
<td>Person C</td>
<td>Dependent on her partner’s income, because the household’s fixed costs are higher than her net income. Could not afford a divorce.</td>
</tr>
</tbody>
</table>

I had a personal finance consultant look over the numbers representing the fictional people’s income, spending, etc., and confirm that they were relevant to typical Swedish household problems. The consultant suggested minor changes which I implemented.

I also asked the consultant whether the control material I had designed had enough information about the fictional test cases for an observer to be able to base a judgment of the test cases’ financial situation. The consultant confirmed that the information in the control material included all the relevant factors.

Additionally, I checked the literature. Hershey et al. (1990) identified three sub-problems that are important for retirement planning:

- Whether the person needs additional savings during retirement.
- Whether budget changes can be made so that more funds become available for retirement investing.
- What sort of investment account is suitable. This problem comes from an American context\textsuperscript{125}. It could be adapted to Swedish conditions, but it is out of scope for this study.

\textsuperscript{124} SCB Hushållens utgifter 2012.

\textsuperscript{125} Part of the retirement planning task in Hershey et al. (1990) was to determine whether the person should open an Individual Retirement Account (IRA), a type of account with tax advantages.
5. RESULTS

Person C
Female, 49, income after tax: 18,670 kr

C’s individual and shared expenses

C’s projected retirement income

Mortgage: None, lives in a rented apartment
Retirement savings: 500 kr/month; retirement savings capital: 20,000 kr, of which 25% is invested in stocks
Emergency buffer: 5,000 kr

Figure 38: The control material as it looked for the first half of participants, translated and recreated for print.
Figure 38 shows what the control material looked like. Some changes were made after pilot tests and after half of the study; Figure 38 shows the version used in the first half.

The blue bar chart shows household spending as it is often presented in personal financial management tools such as Mint, Tink, Utgiftskollen, etc.126

The yellow and blue bar chart shows projected future retirement income in a format that resembles the chart seen at Minpension.se, a retirement forecast website provided by the Swedish government in cooperation with private financial services companies. I simplified the chart somewhat by assuming that occupational pension would be lifelong127 and by adding a red line that showed after-tax retirement income128, since the charts at Minpension.se are based on pre-tax income. This made the control task easier than it would have been in real life, but since my prototype was designed to show after-tax retirement income, I needed both test materials to contain the same information. The first six participants saw only the fictional woman’s retirement income. For the last six, I included her partner’s as well.

Information for which I could not find graphical representations in existing tools or apps, was simply put in the test material as text. This included pre-retirement after-tax income, mortgage debt, mortgage interest rate, amortization rate, retirement savings, and emergency buffer savings.

The control material was printed on paper and handed to the participants during the study. For the last half of participants, the control material was instead presented on a computer screen in order to reduce variability between the test conditions. However, the control material was still not interactive.

The interactive prototype was the recently redesigned concept described in section 5.4 Improving the Prototype, see Figure 39.

---

126 However, those tools are personal and thus show spending on an individual level instead of a household level. Co-owned accounts are presented as if the individual pays their charges in full. Hence, the bar chart used here gives a fuller picture of household spending than one would get from a personal app. The first six test participants saw only the fictional person’s own spending plus half of the charges on co-owned accounts. This made it look like all fictional women paid more than their partners. For the last six participants, I added the partner’s full spending to the picture because the first round of feedback changed my design concept from more individual-oriented toward more household-oriented

127 It is more typically assumed to be paid out over five years.

128 The tax rate was based on today’s income tax for retirees.
5. RESULTS

162

Figure 39: The prototype as it looked for the first half of participants.

5.4.2.2.3 Procedure
Each test took place with one single participant at a time. The test took roughly one hour. Participants began by reading and signing a consent form, which took extra care to explain that the prototype was based on simplified rules and calculations; that it had not been quality controlled; that it might contain errors; and that it should not be used as a base for decisions in real life. Participants were also asked whether they consented to have audio recorded. All participants but one agreed. For that participant, I took notes instead.

After they had signed the consent form, participants filled out a form containing some background variables and took a short financial literacy test, the same one used in the risk taking study in chapter 5.3.3.

After they had filled out the form, I asked them to write down all the risks they could think of that might threaten a person’s financial future. The purpose of this exercise was to detect whether they had learned of any new risks after the study which they had not previously thought of. However, it proved difficult to determine the extent to which any new
insights had come from the prototype as opposed to just the general discussion and I dropped this part of the test after the first six participants.

Participants were assigned to their test condition by lottery (see section 5.4.2.2.4 on study design). In order to avoid learning effects, half of participants were first handed the control material on paper and asked for their impression of the fictional person's situation. I then explained to them how to read the chart in the prototype (using demo data which had nothing to do with any of the three test cases) and then presented them with the prototype showing the same fictional person's data. I repeated the same questions I had asked when they looked at the data on paper. I then repeated the same procedure, minus the demonstration, for the remaining two test cases. For the other half of the participants, I showed them the prototype first, then the paper. For the last six participants, as I had tweaked the study design, I also explained how to read the control material using demo data.

The questions I asked the participants for each fictional test case were the following:

- Will this person [household] be able to pay their expenses after retirement?
- Could this person [household] manage unforeseen events?
- What, if anything, should the person [household] change about their financial situation?

I purposely left the second question open with regards to what unforeseen events to consider\(^{129}\).

After going through all three test cases, I asked the participants whether they would like to use a system similar to the prototype for planning their own retirement. I also asked them how they would improve the prototype. For the one participant that did not want to be recorded, we did not have time to go through all test cases. I prioritized asking the final questions over going through the last test case.

It was apparent to the participants in this study that I was the designer of the prototype and that the control material was there for comparison.

---

\(^{129}\) For participants who had written down risks at the start of the test, I used their own list of risks as examples of unforeseen events.
Therefore, their praise must be interpreted with caution since there is a risk that they were being cordial. I excluded from analysis vague compliments such as “this is really cool,” or “I bet this would be great for [other target group]”. I only considered comments about concrete advantages or disadvantages.

Since the test focused on the usefulness of the prototype rather than usability, I sometimes helped participants when they asked how to do something or whether something could be done. However, I did not give them clues as to what information to look for in order to solve the task. I did however quite often have to encourage participants to not just look at the chart and make judgments based on the static picture, but to actively experiment with the different controls in order to actually see the effects of the improvements they suggested and be able to produce more informed suggestions.

5.4.2.2.4 Study Design
Since there were three fictional test cases (A, B, and C), there were six possible permutations. To ensure an even distribution of test sequences, the number of participants needed to be a multiple of six. Half of participants saw the control graph first, the other half saw the prototype first. Thus, the study had 12 participants. They were assigned through lottery to the test sequences in Table 25.

Table 25: The design of the final usability study.

<table>
<thead>
<tr>
<th>Test sequence</th>
<th>Control first</th>
<th>Prototype first</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>Participant 5</td>
<td>Participant 10</td>
</tr>
<tr>
<td>ACB</td>
<td>Participant 3</td>
<td>Participant 11</td>
</tr>
<tr>
<td>BAC</td>
<td>Participant 9</td>
<td>Participant 4</td>
</tr>
<tr>
<td>BCA</td>
<td>Participant 12</td>
<td>Participant 6</td>
</tr>
<tr>
<td>CAB</td>
<td>Participant 1</td>
<td>Participant 8</td>
</tr>
<tr>
<td>CBA</td>
<td>Participant 7</td>
<td>Participant 2</td>
</tr>
</tbody>
</table>
5.4.2.2.5  Pilot Study
I ran a pilot study on a convenience sample of five people, four of whom were in the right age range. The pilot study revealed that it was easy for participants to spot the problems in the fictional people’s situations even with just the control material. One reason was that all variables had been kept constant across conditions, except the “odd one out” which represented the problem; for example, the mortgage debt for B was considerably higher than for A and C, while they all had the same income.

In order to hide which variable was the odd factor, I had to change the numbers so that everything was slightly different across conditions; ages, incomes, spending in different categories, etc. The test cases were no longer strictly “comparable,” but this was a necessary change.

5.4.2.3  ANALYSIS
I transcribed all audio tapes from the tests. For the one participant who did not want to be recorded, I transcribed the hand-written notes into more detailed text immediately after the test. The result of the transcriptions was a document containing 73,041 words.

I sorted the participants’ comments into tables where they were grouped by question (such as, “Will this person be able to pay their expenses after retirement?”) and by test case (A, B, or C). I compared the answers prompted by the control material against the answers prompted by the prototype. The purpose of this was to see whether the different graphical representations seemed to inspire different answers from participants to the same question about the same data.

I also had comments and reflections that I had written down during the tests or while transcribing. These were usually about things that struck me as surprising during the test, things that ought to be improved about the prototype or the test method, or things that were recurring across participants. When I had finished all transcriptions I checked these reflections against the data to confirm that what I had perceived as recurring really was as common as I thought.

I made an intermediate analysis after six participants. It became apparent that the prototype had some problems that needed fixing. I decided to fix them before the next six participants, in order to get some validation of the fixed version. It is normal in information visualization projects to modify prototypes as soon as lessons are learned (Shneiderman & Plaisant, 2006). This was also when I took the opportunity to make some improvements to the test method. The new
version of the prototype is pictured in the next section (Figure 40 and Figure 41).

When all transcriptions were finished, I read through the document and coded meaningful chunks in a grounded theory-inspired style. I used the codes to find common or interesting patterns in the data.

### 5.4.2.4 Results

#### 5.4.2.4.1 The First Prototype

Below are tables of responses from the six participants who saw the first version of the prototype. Participants did not always answer by a simple “yes” or “no,” so in some cases I had to determine from a lengthier comment whether they leaned more toward yes or no, in order to fit the response in a table.

Some responses are missing, because I did not ask the question when the participant had already talked about the subject in response to a previous question. For example, if a participant answered the question “Will this person be able to pay their expenses after retirement?” by naming things they could do in order to better their situation, I did not ask them again what the fictitious person should do.

Table 26: Responses to person A from the first six participants.

<table>
<thead>
<tr>
<th>Person A (typical Swedish adult, no particular financial troubles)</th>
<th>Will this person be able to pay her expenses after retirement?</th>
<th>Could this person manage unforeseen events?</th>
<th>Should this person change anything about her situation?**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control Material</td>
<td>Prototype</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 Yes, 2 Qualified yes*, 1 Must cut costs</td>
<td>5 Yes, 1 Qualified yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

130 The unit of analysis is utterances, not individuals (participants).
It appears that participants were somewhat more confident to say that A was doing okay when they saw her data in the prototype than when they saw the control graphs. The control material seemed to prompt more just-in-case cost cutting, perhaps because the control material did not make it as clear that A had a relatively good margin between income and spending. A common strategy for participants using the control material was to look at the spending charts in search of expense categories that stood out as abnormally high and then suggest cutting down on those categories. However, spending cuts are not necessary for a person who can afford their lifestyle.

Table 27: Responses to person B from the first six participants.
For person B, it was fairly obvious to most participants that her situation was not ideal, regardless of visualization technique. However, when the question was what B could do about her situation, the responses prompted by the control material seem slightly more defeatist and less constructive. Three participants suggested that B should have made better choices earlier in life so that she would not have ended up in this situation. Two participants remarked that it was difficult to tell from the control material whether it was possible for B to increase her savings rate today.

In contrast, all but one participant experimented with different solutions in the prototype. Some solutions were tested and disregarded when participants saw that they did not have the intended effect. For example, one participant tried increasing risk in B’s retirement savings but changed her mind when she saw that it did not make much of a difference. One participant started by postponing retirement, but eventually found a way to increase savings so much that she could move the retirement date back to its original position (age 67). Two participants recommended increasing amortization rate (for one participant the motivation was to lessen B’s exposure to the risk of raised interest rates, for the other participant it was to get rid of the mortgage faster and lower costs after retirement). Another participant recommended decreasing the amortization rate in order to lower fixed costs and have a greater margin between income and spending at the current time, even though it came at the price of having higher fixed costs for the rest of B’s life.

Table 28: Responses to person C from the first six participants.

<table>
<thead>
<tr>
<th>Person C (low income and high spending, sensitive to divorce)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Will this person be able to pay her expenses after retirement?</strong></td>
</tr>
<tr>
<td>Control Material</td>
</tr>
<tr>
<td>4 No, 1 Hesitates, decides no, 1 Qualified</td>
</tr>
</tbody>
</table>
Could this person manage unforeseen events?

<table>
<thead>
<tr>
<th>Control</th>
<th>Prototype</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 No (2 concern divorce)</td>
<td>2 No (1 concerns divorce), 1 Qualified yes</td>
</tr>
</tbody>
</table>

What can this person do in order to improve her situation?

<table>
<thead>
<tr>
<th>Control Material</th>
<th>Prototype</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Cut costs, 3 Save more, 1 Get insurance, 1 Amortize more, 1 Transfer pension from partner, 1 Move to a cheaper place</td>
<td>5 Cut costs, 2 Save more, 1 Increase investment risk, 1 Postpone retirement, 1 Change occupational pension withdrawal rate, 1 Amortize more, 1 Move to a cheaper place, 1 Get a second job</td>
</tr>
</tbody>
</table>

Participants considered person C to be in a precarious position financially. Most deemed her to be living beyond her means. However, fewer than I expected spotted her specific vulnerability to divorce. In fact, the control material seems to have been better at showing that a partner was carrying a large part of household expenses.

P1: This [the prototype] looks more positive than when you look at that one.

(Points at the control material)

Researcher: What makes it look more positive?

P1: Because here [in the prototype] you only see hers. You don’t see what someone else… Here is her own income so to speak. Here [the control material] you involve her husband, who you can see takes a large share of the costs. So… you don’t see that here [in the prototype].

Essentially, the prototype hid a large share of the household’s actual costs by only presenting the share that the individual pays. Total costs only became visible when the user tested the divorce scenario, but few participants did.

On the plus side, participants using the prototype came up with a wider range of options for what C could do, though they sometimes expressed frustration that C could not be completely “saved,” only improved. One participant expressed surprise that C was harder to “save” than B, even though B had a lower income. The explanation is that C had higher fixed costs and paid rent rather than a mortgage which would eventually
shrink. One participant suggested amortizing more even though C lived in a rented apartment. This is a mistake that users in real life would not make about themselves.

When participants had clearly stated that C would not be able to pay her expenses after retirement even in a “sunny day” scenario, I did not ask them whether C could manage unforeseen events. This explains the low number of responses to the second question in Table 28.

In general, participants could suggest solutions for the test cases, but the control material did not allow them to test whether their solutions were sufficient, due to its lack of interactivity. They were then forced into vague hypothetical reasoning about e.g. “cutting costs” without specifying which costs and how much. However, even when they had the prototype, it was common for participants to base their judgments only on the static graph in front of them. They did not experiment with different parameters and test different scenarios as much as I had expected, and when they did, it was often after I had specifically asked them to.

Table 29: Degree of interaction with the prototype. The table counts cases (fictitious person×test participant), not participants.

<table>
<thead>
<tr>
<th>Question</th>
<th>Experimented with parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will this person be able to pay her expenses after retirement?</td>
<td>17 No, 1 When prompted</td>
</tr>
<tr>
<td>Could this person manage unforeseen events?</td>
<td>5 Yes, 6 No, 6 When prompted</td>
</tr>
<tr>
<td>What can/should this person do to improve her situation?</td>
<td>5 Yes, 2 No, 5 When prompted</td>
</tr>
</tbody>
</table>

There could be at least two explanations for the low level of interaction with the prototype in response to the first question:

- The participants expected the initial view to reveal any financial vulnerabilities without them having to actively click anywhere to dig deeper into detailed scenarios.
- Once I had asked them to test scenarios, they knew for the remainder of the test that I expected them to interact with the prototype, but they did not do so spontaneously.
To summarize, the prototype suffered from the design problem that risks were not sufficiently salient. Users were required to click through different scenarios and actively experiment with different parameters to discover vulnerabilities. This seemed particularly true for the divorce scenario, which was actually an umbrella term for anything that would cause the loss of a partner’s income. The invisibility of the partner also prevented the testing of different solutions on the household level, such as transfer of pension credits, bequest, and life insurance.

5.4.2.4.2 The Second Prototype
To address the problems with the first prototype, I created a second prototype for the remaining six test participants. The new prototype is pictured in Figure 40 and Figure 41.

5.4.2.4.2.1 Major Design Changes
The biggest conceptual change was that I incorporated the entire household into the model, with one common view for both spouses, complemented by separate individual views. I did not prioritize perfecting the design or implementing the full functionality of the previous prototype, so the new prototype was visibly less finished.

Additionally, I added functions that checked for certain vulnerabilities and flagged them to the user in case of projected deficiencies, without the user having to click anywhere. These functions were presented as green checkmarks (for success) or red crosses (for warnings) below the chart.\(^{131}\) These automatic checks were meant to be different in the common household view than in the individual views. The individual views would focus on individual-specific risks, such how the spouse in question would

\(^{131}\) The alert reader might ask: How is this different from the binary “Thumbs up or down” criticized by Turner and Witte (2009)? My answer would be that the thumbs up or down attempts to summarize an entire financial life into a simple binary success/failure, while my auto-checks are limited to much more well-defined events. Additionally, the thumbs up or down is based on an arbitrarily drawn line of probability, while my auto-checks do not involve the probability of events occurring; they simply check whether income will be enough if the event occurs.
fare in a divorce. However, the individual views were not implemented at the time of the test.

Figure 40: The second prototype in the household view.

Figure 41: Translated and recreated for print.
The prototype auto-checked for:

- Whether either spouse could afford to be unemployed for six months.
- Whether the household could afford an 8% mortgage rate.
- Whether the household could afford their expenses after retirement.

The auto-checks were responsive; they updated when other parameters, such as household spending, changed.

Additionally, I hid parameters and controls in pop-ups that were opened by buttons named “Test more scenarios.” This was to provide a cleaner look and prevent information overload.

My idea was that any risks deemed vital by the designer would be auto-checked and that less important risks could be optional for the user to experiment with. This would make more information available at-a-glance and reduce the need for active interaction. Thus, the new design was intended to provide more computational offloading and graphical constraining than the previous version; the user would not have to know which risks are the most important to look out for.

5.4.2.4.2.2  Minor Design Changes
I also made a few minor design changes that were less fundamental to the concept, but made the prototype easier to read.

- I gave the area between income and spending its own color and marked it “Surplus”, as opposed to leaving it empty. This was in response to feedback from participants. See Figure 42.
- In the previous prototype, the line representing income was green, while the area representing monthly retirement savings (always at the top of monthly spending) was a striped green. One participant read the striped green as an alternative level of income. To avoid this confusion, I changed the color of the retirement savings to something completely unrelated to the income line. See Figure 42.
Since participants had sometimes failed to notice points where income fell below spending, I colored areas of deficit red and added a warning text that said “Deficit!” I also moved labels to the end of the line so they did not occlude the graph at the time of retirement. This made small deficits more visible. See Figure 43.

I simplified the color scheme by giving all expense categories the same hue, regardless of whether they were fixed or variable costs. This seemed like a necessary change after the auto-check functions and the surplus areas had introduced more colors to the chart. See Figure 44.

I changed the sort order so that deprecating costs (loans; in this case the mortgage) were always at the top of the chart, which allowed all constant costs to remain as straight lines. The fact that the entire chart sloped had confused some participants. See Figure 44.
5.4.2.4.2.3  Method Changes
For the remaining six participants, I changed the interview question from focusing on persons A, B, and C, to focusing on their households. In other words, I asked whether both spouses together could afford to pay their expenses after retirement.

Because the new design made it possible to switch between a household view and a view for each individual, I gave the fictitious people in the test cases names so that I could write their names on buttons and labels in the interface. In order to keep avoiding socio-economic connotations as far as possible, I gave them names from the top 10 list of names for Swedes born in the 1970’s. They did not have last names.

5.4.2.4.2.4  Results
To reiterate, spouses in household A had average salaries for Swedish adults. Their spending was similar to the average spending of Swedish two-income households without children. Their auto-checks indicated that they could not afford to become unemployed, but that they could afford 8% mortgage interest and that they would be able to pay their expenses after retirement.
Table 30: Responses to household A from the last six participants.

| Household A (typical Swedish adults, no particular financial troubles) |  
| --- | --- |
| **Will these people be able to pay their expenses after retirement?** |  
| Control Material | Prototype |
| 1 Yes, 1 Qualified yes, 1 Barely, 1 No, 1 Cannot tell | 5 Yes, 1 No |
| **Could these people manage unforeseen events?** |  
| Control | Prototype |
| 4 Yes, 1 For a while, depends on what it is | 3 Yes, 1 Not if it happens before retirement, 1 No (unemployment), 1 Cannot say |
| **Should these people change anything about their situation?** |  
| Control Material | Prototype |
| 3 Cut costs, 3 Review invested assets, 2 Save more, 1 Buy real estate, 1 Buy life insurance, 1 Reconsider retirement age, 1 Amortize less | 3 Save more, 2 Review invested assets, 2 Cut costs, 1 Get unemployment insurance, 1 Amortize less |

The one participant who said household A could not afford to pay their expenses after retirement was looking at the green line representing retirement pension and ignored occupational and private pension. This is obviously a risk when the design separates income from different sources into different lines. As a future development, the two lines might be combined into one.

The reason why so many participants who looked at the control material did not think that A would be able to pay their expenses after retirement is because the static spending charts lack a time dimension; they show A’s spending as of today, but as both spouses in A have some time left until retirement, they will have paid down their mortgage by the time they retire and fixed costs will have gone down.
Two participants using the prototype ignored the auto-checks and only looked at surplus income when answering whether household A could manage unforeseen events. One participant noticed the auto-checks but still answered yes, because “nobody manages that” in reference to the scenario of becoming unemployed without having unemployment insurance. Only one participant mentioned that household A should get unemployment insurance, despite the auto-check clearly indicating that they could not afford to lose their jobs. Perhaps the system needs to be even more explicit in its recommendations, by providing users with a to-do list. There might currently be too many potential risks and problems for users to keep track of in their heads.

It is worth noting that almost all participants looking at the control material said that household A could manage unexpected events. They usually based this judgment on the perceived high margin between A’s income and spending. The one participant (P9 below) who hesitated alluded to risks she had seen in the prototype with previous test cases.

Researcher: Is it possible to say how long they would manage?

P9: Yes, but that depends on what unexpected event it is.

Researcher: Yes.

P9: Yes. But let’s say it’s the example that was, six months of unemployment… But then I need to look at the other thing [the prototype] so I can start experimenting. (Laughter)

The participant who said she could not tell from the prototype whether household A could manage unforeseen events wanted to know how much they had in emergency savings, which was not apparent from the interface. The first version of the prototype did show the emergency buffer, but not in the main chart. It made use of the emergency buffer by nibbling from it in times of deficit and warning the user when it would run out. That function was not implemented in the second prototype. Still, I had noticed that the presentation of the emergency buffer in the first prototype was too subtle; hardly any participants noticed it.
Table 31: Responses to household B from the last six participants.

<table>
<thead>
<tr>
<th>Household B (too high mortgage, sensitive to raised interest rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will these people be able to pay their expenses after retirement?</td>
</tr>
<tr>
<td>Control Material</td>
</tr>
<tr>
<td>4 No, 1 Yes</td>
</tr>
<tr>
<td>Could these people manage unforeseen events?</td>
</tr>
<tr>
<td>Control</td>
</tr>
<tr>
<td>2 No, 2 A small event, 1 Yes</td>
</tr>
<tr>
<td>Should these people change anything about their situation?</td>
</tr>
<tr>
<td>Control Material</td>
</tr>
<tr>
<td>5 Cut costs, 2 Save more, 1 Move somewhere cheaper, 1 Raise income</td>
</tr>
</tbody>
</table>

At one test, we ran out of time before we could go through the control case with household B, which explains why there are only five participants in that column.

Almost all participants agreed that household B would be unable to pay their expenses after retirement. The one participant who said they could, made the mistake of looking at pre-tax retirement income in the control material.

Three participants using the prototype asked if household B had an emergency buffer, which further confirmed the finding that the emergency buffer needs to be visible in the chart.

The participant who said that household B could manage unexpected events only looked for a margin between income and expenses, ignoring the auto-checks which otherwise revealed that household B could cope with neither unemployment nor raised mortgage interest.
Perhaps unsurprisingly, all participants recommended that household B cut down on their spending. Participants using the prototype came up with a slightly wider range of additional solutions.

Table 32: Responses to household C from the last six participants.

<table>
<thead>
<tr>
<th>Household C (low income for the wife, high spending, sensitive to divorce)</th>
<th>Will these people be able to pay their expenses after retirement?</th>
<th>Could these people manage unforeseen events?</th>
<th>Should these people change anything about their situation?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Material</td>
<td>Prototype</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 No, 1 Cannot say</td>
<td>6 No</td>
<td>2 A small event, 1 Yes, 1 As long as both work, 1 No, 1 Cannot say</td>
<td>4 Cut costs, 3 Save more, 2 Move somewhere cheaper, 1 Increase investment risk, 1 Transfer pension credits, 1 Not get divorced</td>
</tr>
<tr>
<td>Prototype</td>
<td></td>
<td>3 No, 1 Yes, 1 A small event, 1 Cannot say</td>
<td>6 Cut costs, 2 Increase investment risk, 1 Get unemployment insurance, 1 Get health insurance, 1 Postpone retirement, 1 Husband saves more, 1 Wife moves somewhere cheaper after husband’s death, 1 Save more</td>
</tr>
</tbody>
</table>

The participant in the prototype condition who said that household C could manage unforeseen expenses noticed that they would not be able to cope with unemployment for the husband, but answered “yes” nevertheless. She seems to have taken the question to mean “Can they handle unexpected events reasonably well?” rather than something that required a binary response. She had also started experimenting with parameters and lowered costs before I asked the question, so she saw a more positive prognosis. The participant who said she could not say whether they could manage unforeseen events wanted to know whether they had an emergency buffer.
Two participants tried increasing monthly savings for household C but changed their minds when they saw that it had a surprisingly small effect, due to the small size of the surplus that was available for extra saving. One participant expressed concern that this might discourage real users from saving. Part of the problem is that the available surplus depends on the order in which the user makes changes; if they start by cutting costs, a much larger surplus becomes available. A future improvement would therefore be to automatically assume that any money saved by cost cutting goes toward savings, rather than requiring users to actively put the newly freed up money toward savings.

5.4.2.4.3 General Impressions
All test participants were positive toward the prototype and said they would use a similar tool if it were available. Six (50%) expressed great enthusiasm such as “I would stay up all night pulling those levers” (P6), “Yes, I would be incredibly interested” (P10), and “Yes, I think this would be super great. Super great.” (P2). However, two participants (both of whom saw the older prototype) admitted that they probably would not look at it that often in practice and two participants (also with the older prototype) mostly emphasized how good it would be for a younger target group.

Eight participants talked about the prototype’s ability to visualize the long-term effects of the user’s decisions and lifestyle. Two said the prototype would help them understand what pensions they could get better than the information that is available today and two said it could help them understand whether they could make it through retirement.

Several participants said that the prototype was fun, but P9 was particularly relevant because she said she was not usually interested in finance.

P9: Yes actually, since I never plan financially at all, I have to say that this was... fairly accessible. And I think that if you sit with it for a while and see that it makes a difference... then it might even be kind of fun, I think. Actually.

---

133 P6’s statement is obviously tongue-in-cheek, but still.
134 P10 used an untranslatable Swedish colloquialism: “Så det bara dånar om det skulle jag vara intresserad.”
135 We cannot know whether it actually suits young users until that has been tested.
P10 provided an account of how she would use the tool, which
effectively fit very well with my intention.

P10: Then I would feed it my fixed costs and my income. And… yes and my
savings of course… (pause) on all levels with occupational pensions or what
I my numbers from the orange envelope and from [pensions company] and
all of it… er… so that I get a picture of how it will look. And when… could I
manage to retire before 67, even? Or do I have to work until I’m 70 or so?
What can I do if the interest rate, what happens if the interest rate goes up?
If it just doesn't work with my mortgages anymore? That would be very
interesting to see. If the interest rate goes up by 1%, what happens to my
costs then? What can I no longer do then? Like in a very simple way, how
will it affect my life if the interest rate goes up… yes. Yes.

Like several other participants, P10 was under the impression that she
would have to enter the data into the tool herself. I did not bother
correcting the participants on this minor detail unless they perceived data
entry as an important hurdle for using the tool.

5.4.2.4.4 Advantages

5.4.2.4.4.1 Graphical Presentation Format
Eight participants said, without being specifically asked, that they found
the prototype easier to read or understand than the control material.
Several sighed or frowned when they had to go back to the control
material and some made comments indicating that they viewed the
prototype as the “answer” to the puzzle posed to them in the control
material.

Researcher (hands P3 the control material): This is person B. So will person
B be able to pay their expenses after retirement?
(Pause)
P3: Yeah here he is right at eleven and a half, and they are at twelve… the
costs. No, just barely\(^{136}\). Yeah no so like this…. This (points at the computer)
will show that. (Laughter)

\(^{136}\) “Nå det är på håret.”
One participant, P12, initially said she preferred the graphical presentation format of the control material, if only the control material had been interactive too, but she took some of that back toward the end of the test.

P12: Here you see the surplus so that’s actually better… In a way this one is a bit more advanced than the bars [chart]. So right there… you kind of missed this space that goes up to… [the income]. Then you have to calculate for yourself, yes they get this much and they spend that, then you have to calculate that yourself. So of course now that I’ve seen this a little, this is much more advanced. Since you can experiment with all this.

In principle, the bar charts in the control material could also show the surplus if a line representing income was added and they could also be made interactive. However, the bar charts lack a crucial dimension: Time. Seeing future income and spending on a timeline made it easy for participants to reason around different points in the future. For example, they talked about the different conditions for couples when one spouse was retired, versus a few years later when both were retired. They were also made aware of the extra funds that were continuously being freed up as the couple’s mortgage was repaid. This allowed participants to design plans where spending would only have to be cut for a few years when margins were extra tight and then go back to a more normal budget.

Perhaps unsurprisingly, participants also appreciated seeing everything in one view rather than having expenses in one graph and retirement income in another.

5.4.2.4.4.2 Interactivity
The interactivity of the prototype was an obvious advantage over the control material; if the intention was to compare only the graphical presentation formats as such, the control material should also have been interactive. However, the purpose was rather to compare the prototype to the information that a Swedish user would have fairly easy access to today.

Participants appreciated seeing the effects of different parameters and how they affected the whole. Some were surprised at the effects they saw and were forced to question their assumptions.
Below are two different times when P1 assumes that a variable will have a certain effect, but is forced to change her mind when she sees the actual effect.

**Quote 1**
P1: But it's if she gets sick here (Points at timeline before retirement). Then she might have to stop saving, that's how I think about risks. That when you become poor, then people... often perhaps what you might cut down on, is savings among other things.
Researcher: Mm, could you try and see what happens if she gets sick?
P1: Let's see, where is sickness... There is sickness. (Reads) "How long could you manage sickness"... six months... That didn't make much of a difference.

**Quote 2**
P1: But of course as soon as you start, that is if you start to see that if interest on the loans rises, then other things might happen. I think so... (Increases interest rate parameter) That wasn't too bad.

Surprise such as the one seen above is probably due to computational offloading; accurately calculating the effects of different variables on-the-fly is overwhelming for the human brain.

P12: OK, then I would... since this has leisure activities, sports, car and restaurants and all that. Then I would try to limit myself to this. And then I would do the same here. (Types in new, lower costs)
Researcher: Mm.
P12: And then I [press] OK, right? (Pause) This is like... you wouldn't think so. That it would have this effect. (Smiles) But there is an incredible effect. Almost so it seems weird. Because the only thing that has been cut, this is still quite a lot if you, like, don’t... You don't have to buy furniture *that* often.
Researcher: No, I don’t know, maybe they buy something super expensive once a year and then it averages.
P12. No, exactly. Yes, exactly. But still, if you have that then you can manage if you're an adult. Because here you get much, much more security then. And you can feel like, not have to worry. So that's actually... that’s...
because I don’t think people imagine that, because now we cut down one and a half… yes, four and a half. Yes.\textsuperscript{137}

Participants also appreciated that the prototype helped them understand the relative importance of different factors, although they did not behave as I expected they would if they had grasped their relative importance to the full extent. For example, many participants spent a lot of time cutting the test cases’ budgets when postponing retirement for a few years would have dramatically improved their financial outlooks even with the same lifestyle\textsuperscript{138}. I did see some encouraging examples of participants who were able to weigh different options and find different paths to the same goal, for example:

P3: Yes, he will have to work for two more years, that will make it better. There, solved that one.
Researcher: Mm. You made it so the person retires at 69 instead of 67.
P3: Yes. He has been saving so poorly. Alternatively, he will have to look over his costs and try to free up some more.
[...]
P3: Let’s pull this one up to 200 and see.
[...]
P3: But is this enough that it could… But hey, maybe this isn’t so bad. That got you another year [of retirement], my friend. (Laughs) Mm. Then you can go down a little [in working years], so save that money.
Researcher: You increased savings so they could retire at 68?
P3: Mm.
Researcher: Instead of 69.
P3: Mm. And if they work a little more at saving because they reduce their costs… Let’s save 1,000 here… […] Maybe I should become a pensions… sell pension insurance. […]
Researcher: Now you went back to 67?
P3: Now I went back to 67 and I raised his monthly savings and still have some [surplus] left over there.

\textsuperscript{137} The reason why a rather modest (in absolute value) spending cut had such a dramatic effect on the test case’s situation was that their economy was small to begin with (low income, low spending). The cut was actually substantial in relative terms.

\textsuperscript{138} I saw indications that some, perhaps many, test participants valued early retirement over high spending, so this may have been a conscious choice. A more systematic study is needed.
P3: Yes, if they make some a few minor adjustments, it will look much better for the first few years [after retirement] than it did before.

Perhaps most importantly, two participants said the prototype made them feel like they had some control over their future as retirees, like the choices they made mattered.

**Quote 1**

P7: Yes, this [tool] could be something. Absolutely. Er... Like a positive balance to... the orange envelope\(^{139}\). Which you receive as a punishment each year. If you are a [occupation] like me and have been doing too many other things [than working].

Researcher: How do you think this is different from the orange envelope?

P7: But that one is like, it says... if you earn as much as you did two years ago, this is the pension you will get. [...] And that feels like I am just receiving information that feels very inert\(^{140}\) and that feels very destructive. While with this one I could see that okay... now there is too much IKEA [shopping]\(^{141}\) [...]

Researcher: What did you mean by it being destructive? You said the information is destructive.

P7: Yes, it feels like that to me, because I am going to get such an unbelievably small pension. And then I find out about that at the age of 52. And then it feels like... since I feel like I cannot change that number, I might as well not know it.

**Quote 2**

P11: The idea is to raise awareness, no? With a model like this you can see that there is actually something you can do, instead of thinking that you won’t get any pension.

---

\(^{139}\) “The orange envelope” is the pension prognosis that Swedish citizens receive from the government each year.

\(^{140}\) P7 used the word “stum” which could be translated to “inert” or “mute.”

\(^{141}\) P7 probably referred back to the previous discussion, where all of the test cases spent a lot of money in the category “Furniture and decorating.” I think IKEA was a metonym for overspending.
5.4.2.4.5 Disadvantages
Since the model shows only cash flow and not capital, capital that is too small to have a substantial effect on monthly cash flow becomes barely noticeable. Capital that exists before retirement is invisible.

Researcher: What were you looking for before, when you were looking at him whether he saved more or less?
P12: Well, the difference... I mean because if you... It don't think it becomes... Let's say that... now he saves 600. Then they end up... somewhere here. (points at level of retirement income) You should have like more, smaller intervals, like. And if you then save... like... here, 3,000 more. Did that make that much of a difference? No, let's see.

Researcher: No, it didn't make much of a difference. That's probably because there are relatively few years left that he saves 3,000 a month, but there are many years after retirement. Now it [the prototype] thinks he should divide it equally across all years. So it's being divided over many years.
P12: Yes, exactly, mm. But actually, it's like this, right, that you should think that you will... yes... Because if you are a bit of a spendthrift and not very frugal, you might think that whatever, it doesn't make any difference.

Researcher: Yes. You're thinking there's a risk that people think there's no point in saving?
P12: Yes, that you, like, think that it... If I save more, it makes such a small difference that it's neither... so.142

This problem can probably be addressed with a few improvements. Note that these improvements are untested and need a new round of user tests, but that is out of scope for this thesis.

- The emergency buffer should be visible in the interface. It was in fact visible in the first version of the prototype (see Figure 45)143, but only one participant noticed it, and only toward the end of the test. The design of the emergency buffer needs to be re-

---

142 I normally ignore test participants' speculations about how other people might react, but P12 does have a point. The statement can be read as an indication that the graph gave her the impression that saving is pointless, but that that goes against her better judgment.

143 A more readable version of the emergency buffer is seen in Figure 20.
conceptualized in such a way that it can be incorporated in the big chart, where users are looking.

Figure 45: The emergency buffer in the first version of the prototype. The system warns the user if a scenario occurs where the emergency buffer is not enough to save the situation, but few participants noticed that the emergency buffer existed.

- Funds that are freed through cost cutting should be automatically transferred to the emergency buffer, with the choice to put them toward long-term saving instead. In the current version, freed funds disappear unless they are actively put into long-term saving.
- Changes in income or spending that occur as a result of users experimenting with parameters need to be highlighted so that the differences between different scenarios are more visible. This was a common problem; eight participants had trouble at some point remembering what the old scenario looked like before they changed something. A list of changes, perhaps with "undo" functionality, would take care of that problem.
The problem of invisible capital also exists on the other side of the fence: Invisible debt.

[Looking at the control material for B, the over-mortgaged household]
Researcher: What could they do to improve their situation?
P10: Well… […] Perhaps they live in an expensive home, but it doesn’t seem like that, because they have such little… The debt is 1.3 [million]… but the interest rate is low, 48,000 [amortization] per year…. […] Yeah I don’t know… Because when you looked at the previous picture [the prototype], it looked like the total mortgage was pretty cheap. And then I feel like… Can they really move… Are there any cheaper homes to move to?

To reiterate, household B had low incomes and a budget that depended on interest rates remaining very low. Thus, household B’s mortgage looks cheap in the monthly cash flow, but there is an inherent risk to being over-mortgaged, which I was hoping more participants would discover. One of the auto-checks did warn that household B could not manage raised interest rates (see Figure 46), but this rarely prompted test participants to recommend they move. Some of that may be due to the current state of the housing market in the city where this test took place; as P10 alludes to, finding a cheaper place might not be that easy, so participants might not have thought that was a realistic solution.

Figure 46: The auto-check warning users that they will not be able to pay their expenses at 8% mortgage interest.
Another disadvantage is that the current model is based on average monthly spending. If margins are tight, that leaves little room for months that are more expensive than average. This could be solved by incorporating the emergency buffer in the model, as previously mentioned, and by adding a safety margin to the level of spending that needs to be covered.

The model does not account for the need to divest from risky financial markets as retirement draws nearer. It is possible that with an underlying Monte Carlo-based algorithm, the model would simply return poor results for risky short-term investments and better for risky long-term investments. This problem needs further investigation.

5.4.2.4.6 Potential for Improvement
As previously mentioned, participants found it hard to see and remember the differences when they made changes, for example the difference between amortizing 1,000 per month versus amortizing 3,000 per month. This could be helped by keeping a visible history of changes. The history would add additional value if it could turn into a printable to-do list which would include actionable items such as signing up for unemployment insurance.

One participant seemed to want extra encouragement and confirmation that she had made the right choices.

P10: Yes, at least now I’m on the right side of the line\textsuperscript{144}. But I mean, if I were to do something smarter here that I can’t do on my own. That then you would go and talk to someone who says, yeah but if you do this, if Elisabet\textsuperscript{145}, if you think like this that you reallocate, then that will mean blabla, and you withdraw your occupational pension like this, blabla, and that will mean that… I can’t do that, I can’t manage it on my own. So like I can’t contribute anything to that particular field because I can’t do it. […]

P10: I want to feel safe and know that if I do this… If I cut down, if I sell the car and start riding public transport for whatever it was, 1,400 per month, then we’ll make it. Then I have it in black and white, that’s what I want, very perspicuous in black and white. But that you have to see that… Or you say “I

\begin{footnotes}
\item[144] Spending was below the income line after P10 had made budget cuts.
\item[145] Elisabet was the name of one of the fictitious people.
\end{footnotes}
can’t sell the car. We need the car for going out of town every other weekend to see my cousin who needs”... Yeah fine, but then you can never buy another carver chair again, then that’s the deal. So it’s like, you can’t have it all. That’s what I want.

Researcher: And you don’t feel like you get that from this picture?
P10: Er... yes. It’s just that I don’t quite understand how I myself should behave, like, in order to...

Researcher: Make it do what you want?
P10: Yes, exactly, or like in order to... I can’t, I can’t tell... I can’t change this apart from saying like this or that, because the money is the money. (Points at investment asset allocation)

One interpretation of P10’s statement is that she suspected there might be other, better solutions lurking that she had not discovered. Indeed, that was sometimes true.

P6: But hey, they have to work longer of course. Or, well... right, that’s what you have to do of course. You’ll have to work until 69. (Moves retirement to 69). That’s still not... No, this is completely unreasonable, I give up. This is a disaster. Public assistance. (Laughs) Can’t work till 70...¹⁴⁶

In fact, there was a threshold effect at 70 which would have sharply increased the fictitious person’s public pension, had P6 postponed retirement for just one more year. A solution was thus right around the corner. Thus, while the second version of the prototype guided users more firmly by indicating some risks that were important, perhaps the system could go even further, by detecting and suggesting possible solutions. Alternatively, the system could simply indicate to users where there are threshold effects.

The double lines after retirement, one for public pension and one for occupational and private pension (see Figure 47), caused some confusion. My intention behind separating the two was to distinguish defined-benefit income from the less secure income that depends on the person’s success in investing in financial markets, but perhaps that distinction is

¹⁴⁶ Employees in some sectors are not allowed to work till 70, but I don’t think that’s what P6 alluded to. From her tone of voice, I think she meant that it would be inhumane to force the fictitious person to work till 70.
unnecessary. However, I still believe there needs to be a range of outcomes represented in order to account for risk.

![Figure 47: The green line at the bottom (public pension) and the purple lines at the top (defined-contribution pensions) could be combined into just one income, which would have two levels (the worst case and the expected case).](image)

5.4.2.4.7 Possible Future Functions
There were many good suggestions for additional functionality. Those are found in Appendix 1.

5.4.2.4.8 Design Dilemmas
The data revealed a number of what I like to call design dilemmas – choices that have to be made, where it is not obvious which option is the most appropriate, objective, or ethical, but where the choice unavoidably must depend on the personal values of the designer.

*Design Dilemma 1.* The fact that capital was invisible, as previously mentioned, means that users who decide to retire early when they can, might not realize the magnitude of cash that they would be missing out on. Some might argue that this could lead to suboptimal decisions.

P4: Er, and I have learned this much [through my own research], that it’s unwise to retire at 65. It’s better to wait until about 66 and a half – then I would get the maximum return. Waiting an additional six months won’t generate that much more. According to minpension.se, which is where I checked.
The model in the prototype does not easily lend itself to the sort of optimizing P4 describes. On the other hand, this is the sort of capital-centric reasoning that Merton (2014) and others have criticized. Should users be made aware of opportunity costs, or is it better to let them solve their retirement puzzle in a way that makes them satisfied?

**Design Dilemma 2.** I had arbitrarily made the choice to emphasize the statistically expected outcome of financial market investing, indicating the worst-case scenario only as a thin shadow (see Figure 48), but I could have done it the other way around. Only one participant, P12, used the worst-case scenario as her benchmark. It is likely that more participants would have done so too, had the design looked different.

Researcher: So would these people manage to pay their expenses after retirement?
P12: As it looks now?
Researcher: Mm.
P12: Er… (Pause) No. And they have absolutely no surplus. Of course, they won’t have to save for retirement [after retirement]. But they're right… on the edge here.
Researcher: You’re looking at the dashed line?
P12: Yes, let’s say it goes [better on the stock market], but even here, it’s not… (Points)
Researcher: Yes, the solid [line]?
P12: Yes. That’s like assuming that it goes as it does now.
Researcher: Yes. Do you prefer to focus on the dashed line? That is, the worst, more worst case?
P12: Yes, that’s what I would do, who am… but that’s personality.
Researcher: Yes.
P12: (Laughs) So that you always assume like… the worst. Because then you have margins.
Researcher: Yes.
P12: Because you can’t be convinced that it goes as it does now. You have to, like, guard against the worst outcome.

The fact that I had visually differentiated the worst-case line and the expected-case line in such a way that one looked stronger than the other, rather than just different, was in fact a form of graphical constraining; it signaled to users that the strong line was the important one to look at. A
more neutral design would for example use different colors but not different thickness, transparency, or dashes for the two lines.

![Diagram of Maria retires]

Figure 48: The line representing the expected outcome was strong and the line representing the worst case outcome was weak, but it could have been the other way around.

However, focusing exclusively on the worst case means lowering risk, perhaps to the point that a large and relatively safe gain is missed. Most conventional financial advice says that young people should take risk. This problem might be alleviated in a future Monte Carlo-based system, where risk becomes more harmful the closer the person is to retirement.

5.4.2.4.9 Unexpected Behavior
If the panacea for underfunded retirement plans in other systems is to take more investment risk (see chapter 2.3.2), in this system it was to cut spending. Reducing costs was suggested three times more often than increasing savings. This was true for both the control material and the prototype. This did not appear to be driven by laziness on the part of participants – they could have just suggested minimizing costs as a way to solve the task and be done with it, but they often expressed empathy toward the fictitious people and seemed careful about designing budgets and retirement plans that a real person could live with.

[From notes, as P11 did not want to be recorded]
Checks when they [the fictional household] were planning to retire by opening the scenario simulator. Thinks that 67, which is the default, is a reasonable retirement age. Does not want to increase it, does not think it feels right to plan on working until 70. Therefore thinks that there isn’t
much that can be changed in their case. It depends on one’s working life and how worn-out one’s body is.

However, it is quite possible that it is easier for an outsider to make the decision to cut “unnecessary” spending than for the person living it. I had hoped to see more suggestions geared towards increasing income in catastrophic scenarios, such as getting sickness and unemployment insurance. I had also expected participants to value a continued lifestyle over early retirement, but I realize now that that may have been a personal presumption on my part. On the other hand, it cannot be ruled out that participants suggested spending cuts when they saw no other way. Is budget restrictions in the face of insufficient income a solution, or is it a capitulation?

Recall that a previous study showed that a graph similar to this had a slight conservative bias; people using this graph appeared more risk-averse than people using a control graph. In addition to that, maybe this graph has a conservative bias that works on another level: By nudging people to plan for cost-cutting rather than trying to raise their post-retirement income.

It was common for participants to look at the surplus between income and spending when answering questions about whether the fictitious people could handle unexpected events. Sometimes they even ignored the auto-checks that were in front of them. To my surprise, some participants seemed to judge risk in terms of social norms rather than objective danger. When they perceived a risk to be common in the community, the risk was brushed off.

**Quote 1**
P1: (Looking at household B, the low-income earners) Yeah… over a million in debt when you buy a house, that’s nothing strange today.

**Quote 2**
Researcher: Would these people manage unexpected events today?
P7: Absolutely. It says “You cannot manage six months of unemployment.” But then that’s that they don’t have unemployment insurance or something.
Researcher: Yes.
P7: And nobody manages that.
Thus, it appears that the prototype was not completely successful in providing an objective framework against which to judge whether an amount of money is small or large and whether a risk is acceptable or unacceptable, without resorting to peer comparison or generic rules-of-thumb. But I believe that it took a small step on the way there.

5.4.2.5 DISCUSSION
The research question of this study was: Can study participants identify problems in a person's financial situation by looking at the prototype? The question is phrased as yes or no, but it requires a more nuanced answer.

People could generally tell a person or household who was doing well (A) from persons or households who were doing poorly (B and C). This was not unique to the prototype but was almost equally true for the control material, although participants looking at the control material seemed slightly more hesitant to affirm that A was doing well and did not need to make changes\footnote{This conclusion has not been statistically tested and should be taken with caution.}. However, participants did not often point to the intended problems – that B was over-mortgaged and that C was sensitive to divorce – in either case. They tended to reason in more general terms, around the fact that incomes were too low and spending too high.

Recall Hershey et al’s (1990) retirement planning sub-problems:

- Does the person need additional savings during retirement?
- Can budget changes be made so that more funds become available for retirement investing?

The prototype did a good job at both. Participants easily identified cases where the projected pension would not be enough and they were proficient at making budget cuts that put total spending below projected income.

5.4.2.5.1 Limitations
This target group has some characteristics that prohibit generalization to other target groups. As previously mentioned, they belong to the most financially literate cohort in Sweden (Almenberg & Säve-Söderbergh, 2011). Additionally, they have experienced interest rates at 15%, which
several of them mentioned during the test. Young users may never have seen interest rates above 5% and might thus be less wary of interest risk.

It is unknown whether participants would have behaved differently, had they seen their own data in the prototype. Perhaps they would not have been quite as quick to turn to budget cuts as the primary solution. Unfortunately, I did not have the resources to produce a prototype quality-controlled enough that I dared put it in the hands of people who might take its projections personally.

5.4.2.5.2 Reliability and Validity
While this study had relatively few participants, it is well within what is generally accepted for this type of study (Faulkner, 2003). When the goal of the study is to inform design, a single instance of a participant misinterpreting the design, along with a description of how and why they misinterpreted it, can be enough for the designer to learn how to improve the design. However, the “improvement” needs to be validated through a new test. In this study, I went through two iterations of the design-evaluation cycle. Since the new design did not generate the same confusion regarding for example what was a surplus and what was a deficit, I consider the “improvements” to be real improvements.

5.4.2.5.3 Implications
The general aim of this study was to further explore research question 3: Do users understand how to use the representation in the prototype and does it serve their needs better than other tools on the market?

My impression is that they do and that it does. Granted I gave them a thorough introduction to the tool, and that further studies can be done to fine-tune the intuitiveness and user-friendliness of the system, but since I have conducted previous experiments where I did put users in front of the graph without explanation, I feel relatively confident that it is not incomprehensible, though it might look a little overwhelming at first glance. Once users get the hang of it, it allows them to make informed and well-weighed choices in a way that no other tool that I am aware of does.
P12: This is actually better than those [graphs/systems] that those advisors use.

Researcher: Hold on, let me turn the tape recorder back on.

P12: These different types, no matter what type of advisory meeting, you always get... partly... but of course, I am... I am pretty knowledgeable myself. But they always talk about... they never talk about trying to look at your expenses. And think... let's see, how would it be if this happened? No, it's always like this: How much do you want to save? And how do you want to save? How much risk do you want to take? Kind of. So this advice, it is in my opinion, not advice. It feels like it's usually, and I have met a number of them, it's like the idea is to... get people to invest in these different mutual funds and like, these fund managers here and there. Always. And it doesn't matter if you're at a bank or if you're like through work or so. They always, always focus solely on... your savings, almost. No matter how little or how much you save. Instead of seeing... can you do something about the foundation, so that you achieve security instead of like... it's about saving as much as possible. So I have stopped going to those. (Laughs)
6  DISCUSSION & CONTRIBUTION

I have designed a prototype intended to help people who are not experts in finance understand their financial situation and make informed choices about their financial future, specifically their retirement planning. My research is only about the presentation format and not about the forecasting algorithm, which is simplistic in my prototype. Thus, my main result is the design itself. However, I discuss some additional points and suggest directions for future research in this chapter.

6.1  THE RESEARCH QUESTIONS

The design process was guided by the following research questions:

Research Question 1. How comprehensible are today’s financial decision aids to non-expert users? Do they find them helpful? What are the systems’ strengths and weaknesses?

The participants in my study tended not to trust the advice from the tested financial decision aids, nor did everyone find them relevant. Several participants said they could not afford to save, some said they were not the type to invest on the stock market, and several said they would rather invest in tangible assets such as gold or real estate. These were all rather vague and general statements, as opposed to being grounded in facts and figures about their financial situation.

The decision aids failed to persuade the participants of the utility of saving. Additionally, they assumed knowledge of complex terminology and financial concepts, such as the fact that fixed-income funds are considered safer than equity funds. The systems required users to choose between different outcomes that were expressed as lump sums of money that could be smaller or bigger. This was not very helpful to users who did not know how much money they would need.

In addition to the empirical test of portfolio pickers, I also did a review of existing retirement planning systems. Most of these systems did take into account not only retirement wealth but also retirement need. However, most systems did not show explicitly what the needs were, or in other words, how much money the system assumed that the user would be spending on different categories in the future. This makes it hard for a user to criticize the system’s assumptions.
Research Question 2. What would a financial decision aid look like that addresses the weaknesses while preserving the strengths found in RQ1? Is a representational change necessary?

I designed a prototype that translates retirement savings from a lump sum to level of affordable consumption after retirement. This does constitute a representational change.

The model also incorporates different risks that can threaten the user’s retirement savings, besides the normally accounted-for market risk.

Research Question 3. Do non-expert users understand how to use the representation constructed in answer to RQ2? Does it really work better than systems that exist today?

The new graph is slightly more complicated than traditional graphs at first glance. Learning how to read it may require instruction. However, once viewers have understood how to read it, they typically feel it gives them a more informed understanding of what different financial decisions will actually mean for their lives.

Research Question 4. Does the new representation nudge users toward taking more or less risk than they would with another representation?

The study in this thesis showed that participants who imagined a young person with modest wealth took less risk with the new representation than with a control representation. The effect was not seen when participants imagined an older and wealthier person. However, the graph that was tested was an older version of the design concept, in which the time that money needs to last after retirement was variable and consumption was assumed constant. In the newer version, time is fixed to life expectancy while consumption is variable. Strictly speaking, the possible influence of the newest version on risk taking has not been tested.

6.2 THE AIM OF THE THESIS

My general aim has been to explore how data visualization might help people make long-term decisions that would make them more likely to be satisfied with their financial situation in retirement. This can be broken down into two parts:

1. What should people do in order to have a financially secure retirement? For this part, I have relied entirely on literature studies as it is not my area of expertise.
2. How can people be encouraged to do what they should? This is where my contribution lies, as I have explored the special kind of educational intervention that is data visualization. Research on design aspects of finance visualization for non-experts is scarce and tends to be focused on narrowly defined problems rather than broad.

I have been forced to reconsider several assumptions that I had at the beginning of the project. The first was that market risk was the most important type of risk to explain to users. This is not an objective fact but a value judgment, but it is one that I changed after learning that visualizing stock portfolio performance to non-experts was not only difficult, but also not a strong motivator for behavior change. I learned that visualizations and explanations of financial forecasts for non-experts should be grounded in concepts that non-experts recognize; what does this future scenario mean for me, for my future opportunities, freedom, and standard of living? For example, a lump sum of money can be translated to a monthly cash flow which is a unit that most adults already use in their daily lives.

My design philosophy was based on my readings about mental accounting; money should be treated as completely fungible\textsuperscript{148}. Additionally, I wanted the system to be value neutral; as long as the retirement plan works out mathematically, it should be up to the user how to solve the equation. For example, the system should not tell the user to save more if the user is happy to cut spending instead. In light of the user studies, I have been forced to soften this design philosophy a bit. The backside of a completely “neutral” system\textsuperscript{149} is that it by definition provides less guidance and recommendations to the user; it just presents an assortment of parameters and leaves it up to the user to decide which ones to try manipulating. In my qualitative user tests I realized that after all, there were steps that I wanted all users to go through. For example, everyone should make sure that sudden illness or temporary unemployment does not lead to financial catastrophe. Additionally, the test participants heavily favored spending cuts as a solution to underfunded retirement plans, which is essentially the same as accepting

\textsuperscript{148} Although not everyone agrees with this; some see mental accounting as a helpful self-control mechanism (Ariely, 2013; Egan, n.d.). See chapter 2.2.2 for more details.
\textsuperscript{149} Obviously, no visualization is neutral in the strictest sense. The designer has made choices, for example which parameters to include and which ones to ignore. Here I mean neutral as in striving to avoid giving preferential treatment to one type of retirement plan over another.
defeat; adapting to a lower income, when working a few years longer could have solved the deficit while maintaining the same future standard of living. After these observations, I modified the design concept so that the prototype highlighted prominent risk factors and broke the planning process down into clearer steps.

Based on my literature readings, my review of academic and existing systems, and my own empirical studies, I propose that a good retirement planning system targeted at novice users should possess the qualities pictured in Figure 49.

**Degrees of freedom**
- Do let users choose whether to raise income or cut spending
- Do not let users make choices that could be filtered out by the computer (inefficient portfolios, high fees, etc.)

**Risk preference elicitation**
- Do show users the real consequences of risk in terms of potential reduction in consumption, and let them decide whether those consequences are acceptable
- Do not ask users to choose between percentage changes on their portfolios
- Do not ask users whether they are "risk takers"

Figure 49: Characteristics of a good financial decision aid for novice users.

### 6.3 CONTRIBUTIONS TO RESEARCH

#### 6.3.1 HCI AND INTERDISCIPLINARY WORK

The common procedure for many research projects is to start from a field, choose a theory, test some deeper aspect of that theory, and then report findings. The findings in such projects will naturally constitute contributions to the theory whence they came.

Interdisciplinary projects such as this one, however, require breadth that must necessarily come at the expense of depth. A design project draws from several fields and its product is a synthesis of those fields rather than deeper knowledge within any one of the fields. Therefore, it is difficult for a researcher in an interdisciplinary project to make a
contribution to any of the base fields that would be regarded as original and meaningful within that field. Attempts to claim such contributions in retrospect risk coming across as superficial and naïve to people who are well-versed in one of the base fields. Thus, I consider my main contribution with this thesis to be the design itself, which is a synthesis of the knowledge I gathered and produced through the design process. Additionally, I suggest a few humble contributions to the fields that inspired me.

6.3.2 INFORMATION VISUALIZATION
Case studies are a common type of contribution to InfoVis; examples of how a particular representation worked for a particular user group, a particular type of data, and a particular task that the users are trying to do. As such case studies, the studies in this thesis constitute a straightforward contribution to InfoVis.

6.3.3 EXTERNAL COGNITION
Previous models for external cognition (Scaife & Rogers, 1996; Hutchins 1995) are based on an information-processing perspective. Scaife and Rogers (1996) suggested mechanisms through which graphs can help make tasks easier through computational offloading, which is a way of reducing mental effort for the viewer. Graphs can provide computational offloading through for example re-representation or graphical constraining.

Scaife and Rogers make no mention of whether they consider their model to be universal for all humans or whether it is limited to certain target groups and Hutchins (1995) studied professional pilots.

Based on my empirical work, I would like to suggest some additional aspects of cognitive artifacts that pertain to users who are not confident in the subject domain. Given the changes in the Swedish retirement system, people who are not confident in their own financial capability and who did not choose a career in finance, will nevertheless have to make complicated financial decisions. Through the following mechanisms, interactive graphs can reduce the cognitive and emotional burden on insecure users:

- Assurance that the user has reached the best solution according to their own priorities and that all relevant candidate solutions have been explored.
• “Hard” and “soft” constraints – for example, warning the user if they try to cut spending down to an unrealistic level while not stopping them from doing so, but never allowing them to base their plan on unrealistically high expectations on the stock market. The soft constraints serve as a guide for how an experienced person would think.

• Previews of threshold effects, which are discoverable without the user actively searching for them.

6.3.4 PERSUASIVE DESIGN

Persuasive design and its cousins, nudging and choice architecture, usually proceed from the assumption that the designer wants to influence people to make one choice over another. However, the findings from choice architecture studies make it painfully obvious to designers that we cannot abdicate from this power, even when we do not want it. How should we handle design dilemmas such as the ones I discovered in the last study, where my design influenced participants’ propensity to base their plans on the assumption that the stock market will perform according to its historical average, rather than a scenario with poor growth? To reiterate, the dilemma is that reasonable experts disagree on what the optimal choice is. People who take less risk may be exposed to less volatility, but on the other hand, they can miss out on relatively safe growth and actually reduce their likelihood of reaching their goals.

I would like to see a discussion within persuasive technology on how we can minimize our own influence in situations where we do not want it. When there are alternative options and we as designers do not favor one over the other, can we make them all equally visible without causing information overload? How should we inform users that there are alternative paths and that the path they choose should depend on their own values?

6.3.5 FINANCIAL PLANNING

People have different tolerance for risk and financial advice must be suited to the individual customer’s risk tolerance according to law. In order to produce valid responses, questions eliciting a person’s risk preference must have reliability and be psychometrically validated (Callan & Johnson, 2002), ask only about financial risk and not risk in other domains (Weber, Blais & Betz, 2002), and be simple enough that
the person being asked understands the question (Callan & Johnson, 2002) – the latter requirement may not always be easy to meet.

Risk preference questions suggested in the literature are typically validated through either their correlations to other risk scales (Weber, Blais & Betz, 2002; Hanna & Lindamood, 2004; Grable & Lytton, 1999; Grable & Lytton, 2001) or through the correlation between a customer’s score on the preference test and their actual portfolio allocation (Guillemette, Finke & Gilliam, 2012). The former approach depends on the validity of the instrument used as baseline and the latter assumes that the customer holds a portfolio that is consistent with their actual preferences. If the customer’s portfolio was chosen as a result of previous risk preference elicitations, then the logic is circular.

In my studies that used think-aloud methods, I discovered a few misconceptions in how people interpret questions about their risk preference.

- Even risk-averse people seemed to have a central tendency bias when ranking their return requirements and their loss aversion. This could be due to a failure to realize that the safest option is at the extreme end of the scale and not in the middle.
- Questions about people’s general risk preference prompted them to think about their personalities and their past behavior, rather than their goals with the investment under discussion.

I suspect that more such biases would be found if risk preference questionnaires were systematically evaluated with qualitative methods. When new risk preference instruments are developed, they should routinely be tested in a think-aloud style with people representative of the target group before quantitative tests and factor analysis happens.

With tools such as my prototype, which can compute personalized risk questions based on the customer’s own data, customers can instead be asked what possible consequences in terms of reduced consumption they are prepared to accept. This can be a fruitful approach for risk preference elicitation in the future.

---

150 Further study is required with participants that are ascertained to be risk-averse. My assessment is based on how they described their situations and their general attitudes in an unstructured manner.
6.3.6 Behavioral Economics
My design is inspired by the concepts of mental accounting and anchoring. The one big timeline of the user's monthly cash flow, which can be manipulated by changing savings rate as well as amortization rate, is intended to highlight the fungibility of money; a Swedish crown can be used to pay down debt just as easily as it can be put toward savings. The user should choose to use it in the way that shapes the future timeline the way the user likes best. By training the user to think of money as fungible, the design is intended to lessen the biasing effect of mental accounting. Furthermore, the design focuses on the visual timeline and not on percentage growth on investments, which is meant to steer the user away from fixating on numbers which can be influenced by anchoring.

However, I have not measured the effect of the design on neither mental accounting nor anchoring per se. This thesis draws inspiration from behavioral economics but does not exactly contribute to it.

6.4 Contributions to Society

6.4.1 Individuals
The increasing need to save for retirement creates a need for support for making financial decisions. The average Swedish adult is poorly equipped to make such decisions, and while the complexity of the financial world might not be the only reason why people fail to plan and save for retirement, it is a barrier that needs to be overcome.

The individual user is the primary person for whom this design was intended and the primary goal was to show users how much they need to save for retirement and what risk they can and should take. Secondary goals were to show users whether they are underinsured against dangerous risks such as sickness and unemployment and whether they have enough savings in their emergency buffer. By showing what a difference saving makes, and what expenses the person can cut down on in order to save for a more secure retirement, the intended effect of the design is to provide a more positive and constructive outlook on the future, which makes the person feel like the decisions they make really matter. In a sense, the system is intended as support for the user's inner "planner" to set guidelines for their inner "doer" and decide what self-control devices might be needed to reach the user's goals.

Housing prices in Sweden have been climbing faster than real wages and inflation for years and there is concern that the market might be in a
6. DISCUSSION

bubble (Dermani, Lindé & Walentin, 2016). By incorporating current and future mortgage debt and showing its effect on cash flow, I hope to make some people think twice before taking on too much mortgage debt.

I also had in mind the gender inequality after retirement. It is more common for women to work part-time, which lowers their pensions considerably\(^{151}\) (SCB, 2016). Since men tend to have higher incomes after retirement, women can be financially dependent on their spouses and ill prepared for a divorce. This was my reason for including a parameter for working time and a scenario for divorce. A future version could also include a function to transfer pension rights to the spouse with the lower income, since that is possible today but very few people do it (Eklund, 2013).

6.4.2 FINANCIAL ADVISORY SERVICE PROVIDERS

A secondary group of users for this system would be financial advisors, who can use it in their discussions with clients in order to establish the client’s risk preference, manage their expectations, and get a view of their overall financial health. The tool is best suited at the beginning of the advisory meeting, when determining what growth and risk the client needs, whether they need insurance, and perhaps whether they are creditworthy. The tool does not help select individual investment products, but there are other tools that do that.

With a visualization that speaks the client’s language, rather than the language of percentages and probabilities, the client is more likely to leave the meeting feeling informed and confident rather than confused. This reflects well on the competence of the advisor. At the same time, clients who understand everything that is being discussed at the advisory meeting are more empowered to take initiatives and make decisions, thus the information asymmetry between advisor and client is lessened. Hopefully, this reduces the risk that clients unwittingly agree to more risk than they truly are prepared to take and seek litigation later.

6.4.3 GOVERNMENT

There is current concern that Swedes are surprised to find that their pensions are lower than they expected (Hellekant, 2014) and that some retirees choose to spend their occupational pension in a short time

---

\(^{151}\) There is debate over the extent to which this is voluntary.
If these people were to be made aware of their prospects as retirees while they still had time to do something about it, they could probably avoid some future problems. The Swedish Pension Agency provides an electronic service that shows every citizen their expected income as retirees and they also send them their prognosis in paper mail once a year. This does not seem to be enough. It is possible that more people would be exposed to their pension prognosis if it was provided through their electronic bank, where most Swedes log in to do banking errands several times a month (Findahl, 2012), than through the government hosted single-purpose website.

If people had more private savings and took out most of their pensions in life-long form rather than in short periods, that would decrease the burden on the Swedish social safety net. It is therefore in the interest of the government that people refrain from spending their pensions too quickly only to later become dependent on government support. It ought to also be in the interest of environmental sustainability; people who spend their pensions quickly and then receive assistance have consumed more over a lifetime than people who live within their means. However, I stated earlier in the thesis that I have assumed that a financial planning system should always put its users’ best interest first. It could be argued that this entails showing users that they are eligible for government assistance if they consume more in the present and have less savings by the time they retire. Speaking of environmental sustainability, it has been said that society needs a paradigm shift toward less consumption by means of shorter working hours (Schor, 2005). Unfortunately, working shorter hours is currently heavily penalized by the Swedish retirement system, and as long as that is true, my system will make it abundantly clear what a terrible financial deal it is for an individual to work part-time for extended periods of time.

In general, whenever rules and their effects are made more apparent through visualization, the visualization itself should not be faulted for any resulting unwanted behavioral changes. If the visualization exposes tax loopholes, for example, then the onus is on policy makers to fix the loophole, not on the designer to manipulate the visualization.

---

152 Save for the risk that the rules will have changed when the person retires.
6.5 LIMITATIONS

6.5.1 LIMITATIONS OF THE METHODS
Although common, it is generally not considered ideal for a designer to be in charge of the evaluation and testing of their own design. It can be difficult to remain objective regarding the merits of one’s own ideas and the designer-evaluator can subconsciously engage in motivated reasoning when interpreting data and drawing conclusions.153 I have strived to minimize my own bias by setting up a priori success criteria for quantitative studies and by designing qualitative studies to be goal-oriented so that it would be apparent whether or not test participants understood how to use the system. I have tried to avoid indulging in the over-reporting of flattering but vague comments about the system being “nice,” “cool,” or “good.” If a participant said they liked the system, I asked follow-up questions and asked them to explain what it was they liked about it. I only valued comments about concrete advantages, for example that the system explained how much future income was worth in terms of consumption. I approached the testing procedure with the mindset that the goal of the testing was to find flaws, not to validate the existing design. Finding flaws is the only way to improve. Still, there is no arguing that an external evaluator in a double-blinded testing procedure could have been even less biased.

None of the studies, except the one that tested real existing online investment advisors, used participants’ real data.154 Hence, their reactions might have been somewhat artificial and their engagement lower. Additionally, it has previously been established that people who invest hypothetical money in experiments do not behave the same as they would with real money (Vlaev, Stewart & Chater, 2008). People of high intelligence tend to see the experiment as a game that they can win by taking higher risks in the experiment than they would in real life (Taylor, 2013). If I were measuring their risk taking in absolute terms, this would have been a serious methodological flaw. However, my studies only compared relative risk taking in different controlled conditions – I was interested in the differences, not in the absolute levels.

153 Motivated reasoning is a cognitive style whereby a person wants to reach a certain conclusion, and tends to evaluate evidence in the way that best supports their desired conclusion (Kunda, 1990).
154 And while the participants in the online investment advisor study were asked to answer the systems’ questions truthfully, they did not invest real money as part of the study.
Creating logically equivalent experiment tasks was difficult. The tasks were possible to solve in both conditions with the information given, as evidenced by the fact that some participants succeeded. However, the experiment tasks sometimes seemed deliberately designed to show off the advantages of the treatment system and expose the weaknesses of the control system. That was not the case; the experiment tasks were designed to resemble a real-life retirement planning process and the treatment graph was designed to support the ideal planning process, as I viewed it at the time. On the contrary, the experiment task in the retirement planning study was generous toward the control system because the task was more clearly defined than it would have been in real life; participants were told that they were to make the money last to a given age given certain assumptions about the fictitious person’s spending. Had I simply shown them the control graph and told them to choose a suitable investment strategy, it is likely that they would have arbitrarily picked a round number for the allocation between stocks and bonds, without considering retirement need.

Still, Scaife and Rogers (1996) were right to point out that it is almost impossible to create two alternative graphs for comparison, where both truly do contain the same information. However, for the purpose of improving the design of my prototype, many valuable design lessons were learnt through comparing it to an alternative, even though the comparison could not be perfect. Alternatively, I could have evaluated the prototype without a baseline. But I prefer to do it this way, because letting study participants see different versions of a system can elicit rich responses as they reason about the advantages and disadvantages of both versions.

6.5.2 LIMITATIONS OF THE DESIGN
My prototype is obviously based on a deterministic algorithm, but it is meant to be more sophisticated if it were ever to be implemented. The prediction algorithm itself was not the focus of my work. The prototype does not present a definite solution to the question of how to display Monte Carlo results, but it approaches the problem in a different way. Recall that Monte Carlo results can account for many different variables, for example the probability that the stock market goes down, the probability that the user loses their job, and the probability that interest rates go up, all taken together into one large probability that everything goes well. It would be quite a challenge to come up with a visual display
of such a probability that would at the same time be simple to understand at a glance, as well as transparent regarding what the different variables and their probabilities are. My model does not account for the probabilities of different events; it simply shows the consequences if one or several of them were to happen. It has been argued before that consequences are as important, if not more so, than probabilities (Turner & Witte, 2009; Bodie, 2003; Kotlikoff, 2008). In a system that focuses on probabilities, events with very low probabilities can be ignored — that is, left without any further action such as buying insurance. However, it might not be safe to ignore events with very low probabilities if they have catastrophic consequences. In my system, that relation is turned on its head: Events with only minor consequences can be ignored, even if they are fairly likely to occur.

While the first version of the system, which displayed number of years of sustained constant consumption, might have had a bias toward lower risk since the reward of higher risk-taking could become invisible, the newest version of the system is back in the same class as traditional growth curves: More risk means more readily visible expected wealth, regardless of whether it is expressed as a lump sum or as disposable monthly income. Therefore, it is possible that my system also has a pro-risk bias. I have tried to mitigate this in a couple of different ways. Risk-biased systems give users the impression that an underfunded retirement plan can be salvaged simply by taking more risk (Kotlikoff, 2008; Turner, 2012). This is especially true for systems that show a linear correlation between proportion of stocks in the portfolio and wealth at retirement. My prototype displays not only the expected wealth, but also a worst case wealth that shrinks with the proportion of stocks. The idea is that the user should not take so much risk that the worst case outcome falls below an acceptable level. While this relation is still very simplistic, a real implementation of the system should use conservative growth estimates, as recommended by Turner and Witte (2009), and be based on Monte Carlo simulations that do not assume a uniform growth across all years.

There are three ways that a user of my system can improve their retirement plan, which do not involve the uncertainty inherent in taking more investment risk: They can increase their monthly savings; they can reduce their spending; and they can postpone retirement. The option to increase investment risk could be hidden further down, if the designer wants to discourage users from increasing risk as other than a last resort.
However, not everyone would agree that people should take as little risk as possible. Almenberg and Widmark (2011), for example, have problematized certain demographic groups’ non-participation in the stock market and the wealth that they miss out on due to exaggerated risk aversion. In the end, it is unavoidable that the goals of the system come down to the designer’s own risk philosophy.

One big design question remains unsolved after all studies and meetings with finance experts and advisors. How should the prototype treat people who do not appear to have any need to save additional money? Is it okay for the system to tell users not to save? Kotlikoff (2008) says that saving too much is not optimal, even though saving too little is possibly worse. One of the personal finance experts I met with remarked that elderly people with a low income may lose some of their benefits if they have saved capital, which renders the act of saving meaningless for that group. But still, telling people to save less or to not start saving at all seems like it would open up precarious questions of advisory responsibility.

I have not spent a lot of time on the visual design in order to make it aesthetically pleasing. That work would be done at a later development stage, were this system actually to get built. I have made brief efforts to make the latest version conform to WCAG\textsuperscript{155} standards for color contrast, which improves readability for people with color perception deficiencies. However, the color scheme is still very complex and could be improved considerably by an artistically talented designer.

### 6.6 Future Work

A financial planning system might be expected to recommend specific products. I have not studied or developed that part of the system, but it would be a logical next step for the system to recommend an investment portfolio with the agreed-upon risk and expected return.

\textsuperscript{155} Web Content Accessibility Guidelines, a global standard for web accessibility for people with disabilities. https://www.w3.org/WAI/intro/wcag
6.7 Future Research

The studies in this thesis have only begun to scratch the surface of the many complex interactions between financial planning systems and users who are not finance experts.

Because the prototype was not quality controlled by an external inspector, neither with regards to the intricate effects of taxes and social benefit rules, nor pure software bugs, I did not find it ethically defensible to let people try the prototype using their own financial data, even though I always warned my study participants that the system was a prototype and as such not reliable for real prognoses. It is likely that participants would have treated the system differently had it contained their own data. Such a study could reveal cases where the participant has gone through life not realizing their own risk exposure until they see it visualized, providing the opportunity for the researcher to capture real “aha moments.” As attractive as that kind of study is, the prototype needs to go through thorough review and testing before it can be put in the hands of people who might take actual advice from it.

Another interesting aspect of letting participants use their own data would be to see how they value their own spending habits. The participants in my studies were quick to suggest spending cuts to the fictitious test cases, oftentimes as the primary or only solution to an underfunded retirement plan. Would they have been as quick to prescribe spending cuts to themselves, or would they have spent more effort trying to find other solutions? Is it perhaps easy to view the “luxury” habits of others as frivolous and unnecessary, while viewing one’s own as an essential part of what gives life meaning? Additionally, without conducting longitudinal studies, we cannot know how effective people are at keeping their self-imposed budget restrictions. It is possible to imagine an “all-knowing” system, presumably provided by a bank which also controls the user’s credit and debit cards, which could warn or stop the user from making purchases when they are about exceed their own self-imposed spending limits for different categories. While I would be very interested in reading about the outcomes of testing such a system, I am skeptical that people would really like to be halted in the moment of making a purchase. This is why I am weary of solutions that focus heavily on spending cuts.

I have only conducted qualitative user testing on one target group, namely people who are nearing retirement. The next step would be
younger people. Young people are an interesting target group because they have more time to plan and make changes. A small change in savings rate that compounds over many years can result in a large effect. Thus, young people might be the best target group for educational interventions in the finance area. Additionally, young people have intermediary goals between today and retirement, such as buying their first home. As such, they have a complex and perhaps conflicting set of goals for the future, which a future version of the system might show them how to prioritize and accomplish. Additionally, since the last version of the prototype included the entire household as opposed to one individual, it would be interesting to test the prototype on couples and see how it influences their discussions and negotiations of fairness, goals for the future, and acceptable risk. Another relevant target group is people with lower education than the participants in my study.

Likewise, I have only tested the prototype in a Swedish context. People in other countries with other levels of income and welfare might have completely different needs as well as attitudes and trust toward computer systems and information or forecasts from the government, banks, or other authorities. The basic concept of translating a lump sum to a level of consumption for a period of time ought to generalize to other cultures, but we cannot know for sure without testing it in other cultures. Obviously, the specific scenarios that can be simulated in the prototype are tailor-made for Sweden and would be different for other countries.

Although the design of the prototype was informed and inspired by the literature on mental accounting, I did not actually test whether the design counteracted mental accounting. My goal was first and foremost to create a useful and usable system which included all of the most relevant risk factors; therefore, I focused on testing the overall relevance of the prototype to financial decision-making. A study that operationalized mental accounting and its effects on decision-making when people use different kinds of visualization would be of academic interest, but I deemed the more practical issue of creating a useful system more important.

As was seen in the background chapter, previous researchers on retirement planning systems have requested studies on how Monte Carlo simulations can be visualized and explained to non-expert users. I did not ultimately use a Monte Carlo-based model; while Monte Carlo simulations focus on the probability of different scenarios, my design
focuses on the effects of different scenarios regardless of their probability. My model is more in line with the reasoning of economists who have argued that the consequences of different decisions should be explained in terms of potentially reduced consumption, not in expected wealth by the time of retirement (Turner & Witte, 2009; Bodie, 2003; Kotlikoff, 2008; Merton, 2014). Thus, the question of how Monte Carlo simulations can and should be visualized remains for another day.

6.8 FINTECH CRITIQUE: A RESEARCH AGENDA

As stated in the beginning, fintech is currently undergoing explosive growth and academic research has not kept up. We should develop analytical tools for critiquing fintech systems and apps and their effects on individual users and society at large.

One phenomenon that I would like to see critiqued is the practice of publishing aggregated data to uninformed viewers, in a way that makes it resemble advice but does not fall under the legal definition of advice. I call this concept “non-advisory advice.” I came across several examples during my search for financial decision aids: Top-ten lists of the most sold stocks or funds; average spending or other behavior of peers; stock tips from analysts; and blogs or discussion forums hosted or endorsed but not curated by banks. What these examples have in common is that they appear informative to the reader, but do not take the reader’s personal risk tolerance into account. Additionally, there is no guarantee that the behavior suggested by the information (buy this stock, adjust your spending to be similar to that of your peers, etc.) is prudent. At the same time, non-advisory advice is not classified as advice because advice by its legal definition only occurs between an advisor and an individual who knows that the advice is targeted at them personally. Mass communication does not count as advice. Hence, since giving actual advice carries legal obligations of documentation and following protocol, there is an incentive for banks and insurance companies to provide non-advisory advice instead; it can drive sales without the legal risks and responsibilities inherent to advisory services. This topic did not fit into my thesis, but I mention it here as a possible area for future critical research.

I have met HCI researchers who seem to abhor fintech systems because they do not like capitalism, neo-liberalism, or banks – but studying something is not the same as endorsing it. In fact, the critical voices of
those researchers might be exactly what is needed in a world where algorithms continue to take more power over our financial futures.
7 REFERENCES


7. REFERENCES


APPENDIX 1: ADDITIONAL FUNCTIONALITY

Suggested by financial advisors in the first evaluation study:

- Be able to set medium-term goals for future expensive purchases, in a way that visualizes the uncertainty in the likelihood of meeting the goal the same way as for retirement.
- Show how much less monthly savings is required if the purchase can be postponed for a few years.
- Show the effect of following the system’s suggested spending cuts and putting the money toward savings.
- See some concrete values, such as value of home, mortgage remaining after retirement, savings remaining after 90 (the cut-off of the timeline).
- See the effect of partial goal fulfillment.
- See alternative scenarios simultaneously for better comparison.
- Setup automated savings and transfers directly in the interface.

Suggested by participants or discovered in the last study:

- See additional loans besides mortgage and student loan.
- Pay off loans other than the mortgage quicker, e.g. the student loan.
- See if one can afford a new loan.
- See several homes, e.g. holiday cottages.
- See how moving to a cheaper place affects the budget.
- See what interest compensation has to be paid if one moves while having a fixed mortgage interest.
- Account for the need to fund regular repairs if one lives in a house.
- Account for the fact that groceries and other expenses will cost less if partner disappears.
- Account for the fact that the cost for children changes somewhat predictably as they grow older.
- Account for salary reduction plans.
- See subsistence level for comparison.
- See liquidity of funds.