MASTER THESIS RESEARCH PROJECT

EXAMINING FINANCIAL LITERACY OVERCONFIDENCE DISCREPANCIES AMONG DUTCH RESIDENTS

STUDENT NAME

T.G. (THIJS) DAGGENVOORDE

STUDY- AND INSTITUTIONAL INFORMATION

MSC BUSINESS ADMINISTRATION, FINANCIAL MANAGEMENT TRACK

FACULTY OF BEHAVIOURAL, MANAGEMENT AND SOCIAL SCIENCES

UNIVERSITY OF TWENTE

GUIDANCE FROM THE UNIVERSITY

FIRST SUPERVISOR: POLINA KHRENNIKOVA

SECOND SUPERVISOR: BEREND ROORDA

DATE

06 FEBRUARY 2025

UNIVERSITY OF TWENTE.

Acknowledgements

The writing of this thesis “Examining Financial Literacy Overconfidence Discrepancies among Dutch Residents” has been the final assignment for obtaining my master’s degree in Business Administration, with a specialization in Financial Management, at the University of Twente. There are several people, departments, and institutions that I would like to thank.

First of all, I would like to express my gratitude to the supervisors for their guidance and feedback throughout the writing process. With their expertise on/experience with the domains of behavioural finance and aspects of decision making, they have helped me gain valuable insights into the subject and enhance my understanding of research processes. As someone who had never written a thesis before, I feel like I have been able to learn a lot.

Regarding the partial replication of Moore and Healy’s (2008) illustrative study in a different context, as part of this thesis’ analysis, I would like to thank the CCC RightsLink Copyright Department and the APA Copyright Department for their formal permission to do so, as well as for addressing related queries regarding this replication and the online publication hereof (permission obtained). I also want to thank Mr. Moore for granting this permission. Besides, I would like to thank the UT’s Information Specialists and the Copyright Specialists of Elsevier for their help with obtaining the rights to translate and reuse the yet existing financial literacy knowledge question sets by Van Rooij et al. (2011), and analyse the results. Especially a big thanks to those specialists who helped me with queries on associated terms and conditions.

Moreover, I would like to convey my appreciation to the UT Traffic Department for the selfless provision of templates and house-style designs, which can be found back in this thesis, and thank the UT Ethics Committee and the BMS PCP for the ethical review of my survey and the provided information/answered queries regarding GDPR-regulations.

Additionally, I would like to thank all survey respondents for their time and effort to fill in the survey, and thank everyone who has helped me distribute the survey to a relatively diverse sample. Writing this thesis would not have been possible without your help. Irrespective of the attainment of high or low financial literacy scores, exhibition of stronger or lesser judgment abilities, and differences in required time to finish the questions and make the judgments: the final dataset reflects the effortful, collective, experiences of many individuals in the same setting, which is what the data collection has been all about. These contributions have led to valuable insights, in which each respondent has played a precious part.

Lastly, I would like to extend a special thanks to my parents and brother for their continuous support while writing this thesis, and thank all other family members, friends, classmates, acquaintances, or whosoever, who have shown interest in the project and/or provided personal support along the process. Your encouragements have been really important to me, motivating and enabling me to achieve this professional milestone.

Thank you, and enjoy the read!

Thijs Daggenvoorde
Enschede, February 2025

Abstract

This thesis examines the impact of a question difficulty categorization on 3 forms of overconfidence, as identified in past research, in the context of financial literacy. It also explores various behavioural aspects related to answering the knowledge questions and making the estimates. The first part of the analysis looks into the emerging levels of financial literacy, (mis)perceptions and (mis)calibrations, expected overconfidence patterns regarding task difficulty effects, and socio-demographic dissimilarities. This is done by partially replicating past research on financial literacy levels by Van Rooij et al. (2011) and judgmental overconfidence in question sets of distinct difficulty by Moore and Healy (2008). Regarding the latter, the applied financial literacy context is distinct. Previous research indicates the desire of an overconfidence analysis in this specific context from multiple angles, in which the examination of the therefore necessary financial literacy levels is a resulting intermediate step. The second part of the analysis revolves around respondents' behaviour in making these estimates and answering the knowledge questions, taking into account distinctions in click and time data, breaks, perceived confidence in the assembled estimates, and their demonstrated variances. Consequently, it extends past research by offering further related behavioural-specific insights. By conducting an online survey with a 2x3 within-subjects experimental design, relying on fair measurement of the knowledge questions, it finds patterns of estimation and placement misperceptions that are mostly in line with the expected task difficulty effects. Overprecision appears on a more varying basis, even within the more difficult question sets. The findings are somewhat in line, but also in part deviating from past research, considering that the difficulty level is found not to be dispersed across the entire possible scale; questions were generally perceived easier than expected. Following descriptions and explanations from the literature, there are nevertheless some statistical and methodological considerations that should be taken with caution when interpreting these results. The second part of the analysis finds that both click and time data show quite comparable mean values across the distinct question difficulty levels and the SPIES estimates, and are distributed fairly evenly across the distinct SPIES estimates per set as well. The latter is especially interesting, taking into account the higher uncertainty in the estimates for others, and the therefrom potentially resulting respondent fatigue or (increased) misunderstanding. Furthermore, the data indicates a pattern of converging estimation and placement misperceptions when respondents take a break between the sets. Regarding reported confidence in both SPIES estimates, estimation misperceptions differed significantly between respondents who indicated having more confidence in their SPIES1 (for oneself) than in their SPIES2 (for others) estimates, and those who showed either an opposing confidence relationship or equal levels. After totalling, respondents from the primer group expressed significantly less underestimation, while being closer to the ceiling scale-end in all aspects. No significant differences in overplacement have therein been observed. This is also considered intriguing, recalling the distinct financial literacy levels and SPIES perceptions across the groups. Lastly, no significant differences in full-estimates were found in this comparison, neither were differences observed when comparing SPIES1 and SPIES2.

Keywords: Financial Literacy, Overestimation, Overplacement, Overprecision, Hard-easy Effect, Dunning-Kruger Effect, Question Difficulty, Click Data, Time Influence, SPIES Certainty, Full-estimates, Behavioural Characteristics

Table of contents

List of tables and figures

Tables

Figures

List of abbreviations

ABBREVIATION	DEFINITION
AFL	Advanced Financial Literacy
BFL	Basic Financial Literacy
BTA (WTA)	Better-Than-Average (Worse-Than-Average)
CONFSPIES	Confidence in estimate
CRT	Cognitive Reflection Test
D-K effect	Dunning-Kruger effect
DK response	Don’t Know response
FL	Financial Literacy
OC (UC)	Overconfidence (Underconfidence)
OE (UE)	Overestimation (Underestimation)
OFL	Objective Financial Literacy
OoT response	Out of Time response
OP (UP)	Overplacement (Underplacement)
OPR (UPR)	Overprecision (Underprecision)
SCT	Social Comparison Theory
SFL	Subjective Financial Literacy
SPD	Subjective Probability Distribution
SPIES	Subjective Probability Interval EStimate

1 Introduction

Regarding personal finances, we all seem to know people with an apparent unfaltering confidence in their financial decision making. Maybe you have that one friend, who has just started investing in the stock market with seemingly full confidence, or that aunt, who has confidently put her money in a foreign savings account. The confidence in the taken decision may be justified with knowledge or experience, although likewise you might get the impression that the person does not even have a basic understanding of half of the subject. In the latter scenario, in case this person is not specifically risk-seeking but genuinely convinced of own related capacities, a situation of overconfidence might have arisen.

Overconfidence (OC) can manifest itself in multiple contexts of decision-making and judgment (Hadar et al., 2013; Kahneman, 2011). This thesis does not look into people's decisional overconfidence (e.g. trading stocks profitably, or choosing a financially attractive savings account), but it examines the overarching (over)confidence considering one's knowledge, abilities, and the judgment thereof in the context of financial literacy. The concept of financial literacy (FL) is considered to consist mainly of one's knowledge about financial concepts, which' sense is approximated and worded in line with the approach of Hilgert et al. (2003). For the examination of OC, this thesis builds upon three forms: overestimation (OE), overplacement (OP), and overprecision (OPR), as distinguished and applied in more recent psychology research on the overconfidence phenomenon (Moore & Healy, 2008).

Past research has indicated relatively limited FL among households (Hilgert et al., 2003; Lusardi & Mitchell, 2007), leading to more space for OE to arise when personal perceptions are inaccurate. As both objective (e.g., Van Rooij et al., 2011) and subjective FL (e.g., Lind et al., 2020) are often found to affect one's financial decision-making (Balasubramnian & Sargent, 2020), calibration capacities seem to have a considerable impact on the ability to make well-informed monetary choices. It is intriguing to look more into the 3 forms of OC appearing in FL-judgment to address knowledge gaps about the existence of discrepancies in several contextual variations. A more elaborate understanding might ultimately even be useful to get insights into possibilities to lower the occurrence and impact of decisional OC. This aims to avoid suboptimal and irrational decisions that may arise of it in terms of returns and risk-taking (Barber & Odean, 2001; Glaser & Weber, 2007; Russo & Schoemaker, 1992).

As OC forms have often been considered separately from each other, an examination of OE, OP, and OPR simultaneously has, to the best of knowledge, only been performed previously in a FL context by Hamurcu and Hamurcu (2020) and Vörös et al. (2021). The effect of the question difficulty level on misperceptions and miscalibrations has not extensively been taken into account in these studies, although literature normally considers task difficulty to be an influential factor. Both studies consequently referred to this examination in their future research suggestions. By differentiating between basic and advanced FL knowledge questions, following difficulty assignment of queries by Van Rooij et al. (2011) (permission obtained, see paragraph 3.3.1), this thesis tries to obtain more insights in when certain forms of OC seem to occur in this context. This is performed in line with the methodological design (estimates design and OC calculations in question sets of distinct difficulty) of the quiz stage and the interim stage in the study by Moore and Healy (2008), for which permission[1] was

[1] Copyright © 2008 by APA. Adapted with permission. Moore, D. A. & Healy, P.J. (2008). The Trouble With Overconfidence. Psychological Review, 115, 502-517. https://doi.org/10.1037/0033-295X.115.2.502. No further reproduction or distribution is permitted without written permission from the American Psychological Association.

obtained. This is also substantiated by their indication to be reserving some judgment about the broader appliance of their theory and ideas. Following them, the impact of the difficulty level on the occurrence of OC emergences will specifically be addressed by analysing the contextual existence of the hard-easy effect, further difficulty effect expectations, and their associated hypotheses on overestimation and overplacement across knowledge question sets in a distinct FL context following a within-subjects design. For the survey and analysis, the estimates design and OC calculations have, by the way, initially (before having a general overview) been deduced from the first study in Prims and Moore (2017). These are the same, except for the main OPR calculation and the name of the estimates (SPIES, see below). This has hence been applied. Further elaboration can be found in paragraph 3.3.3.

Although the observed overconfidence levels and the hard-easy effect appear at least partly due to statistical artifacts (within a question set) and methodological decisions (e.g., Erev et al., 1994; Juslin et al., 1999; Juslin et al., 2000; Klayman et al., 1999; Soll, 1996), see Olsson (2014) for a summary, it is intriguing to check whether the phenomena of the alleged effects occur and differ in the selected context to be able to compare with earlier findings. Additionally, a sub-analysis on respondents' characteristics can be performed to indicate socio-demographic differences. Secondly, respondents' click and time data, confidence in the made estimates, and styles of answering the Subjective Probability Interval Estimate (SPIES) questions (which is a term from Haran et al. (2010), adopted by Prims and Moore (2017)) will be looked into, aiming to get further insights into one's behaviour while answering the knowledge questions and making performance estimates using SPIES. Consequently, this second part of the analysis is mostly new.

Facilitating a comprehensive exploration of the dynamics of OC in FL, the overarching research question of this thesis is: "To what extent do contextual factors, such as the difficulty level of the questions, respondents' socio-demographic variables, click and time data, and styles of answering the SPIES questions, contribute to understanding the 3 forms of OC in FL and one's behaviour in answering the knowledge questions and making the estimates?". This query can be separated into the following sub-questions: (1) To what extent do the expected difficulty effects manifest in the context of OE and OP within the FL domain? (2) To what extent do the socio-demographic differences relate to the occurrence of OC variances in FL? (3) To what extent do differences in respondents' click and time data add to the understanding of answering and estimation behaviour? (4) To what extent do differences in one's confidence in -and extreme variabilities of- the estimated probabilities relate to OC variances in FL and say something about estimation behaviour?

This thesis contributes to financial literacy and calibration studies, enriching literature on self-assessment of financial knowledge. By making the distinction between OE, OP, and OPR, it tends to acknowledge the existence of multiple forms of OC in FL, which until now limitedly has been done. As indicated, the first goal of the thesis is to further examine the thin line between the multiple forms of OC and one's actual competence in FL by including an analytical categorization of the difficulty level of knowledge questions, suggested to be examined in this context by both Hamurcu and Hamurcu (2020) and Vörös et al. (2021). Although in both studies proposed to be related to other individual variables as well (narcissism and financial well-being), this thesis will only look into the task difficulty categorization due to the inclusion of the second part of the analysis. The examination of click and time data, confidence differences in estimates for oneself and others, and styles of answering the SPIES questions has, to the best of knowledge, not been assessed thoroughly

before either in this context. This part of the analysis has been inspired by the methodological limitation of respondent unfamiliarity with filling in SPIES questions, as formulated and explained by Prims and Moore (2017), together with the ideas of unequal difficulty between estimates for oneself and others (Moore & Healy, 2008) and unequal consideration of the competence of self and others when comparing (Kruger, 1999). It is hypothesized that this uncertainty, particularly regarding estimates for others, will be of behavioural impact that might potentially lead to respondent misunderstanding, fatigue or reconsiderations. Hereby, design choices might also be of influence. Lastly, this thesis will try to validate earlier findings with a reappraisal of FL levels among the sample, and by performing socio-demographic analyses on these FL levels and the overconfidence forms.

As the first part of the main analysis (parts of paragraph 4.2 and paragraph 4.3) is mostly a contextually distinct replication of part of Moore and Healy's (2008) illustrative study, multiple theoretical concepts in especially paragraphs 2.3 and 2.4 overlap with their theoretical elaboration. Some of the by them (either directly or indirectly) shortly mentioned or additional concepts, like measurement variations applied to the 3 OC forms, potentially influential statistical artefacts and their operations within the question sets, and the distinction into the objective and subjective components of task difficulty have been tried to be described either more extensively or from a distinct point of view. This has been done to elaborate further on methodological and analytical considerations and give a broad related context. In turn, multiple of these additions align with works on error models and methodological dependency, which were summarized by Olsson (2014). As this work indicated to be mainly looking into the OPR form of OC, this was anew tried to be extended by looking at the broader applicability of these concepts. As the second part of the analysis is mostly new, the hereto connected theoretical paragraphs (2.6 and 2.7) focus on the related behavioural expectations and their associated theoretical foundations.

The subsequent sections of the thesis consist of the theoretical background (II), methodology (III), results (IV), and discussion and conclusion (V). The theoretical background explores the relevant concepts as mentioned. Primarily, the distinction between objective, subjective, basic, and advanced FL will be made. Additionally, an elaboration of cognitive processes and biases, the lengthy history of research on OC, task difficulty and the hard-easy effect, the Dunning-Kruger (D-K) effect, expected behavioural conditions, and the considered socio-demographic variables takes place. The methodological section explains the research design, sampling procedure, survey design, and chosen data analysis structure. Afterwards, the findings are presented, setting the stage for interpretation and discussion. In this section, expectations are compared with actual outcomes, trying to provide insights into the formulated questions. The discussion and conclusion section includes the main findings, the practical and academic contribution of the results, an elaboration of this thesis' limitations, and some possible directions for future research.

2 Theoretical background

2.1 Objective, subjective, basic, and advanced financial literacy

Financial literacy examinations are a relatively young, but rapidly emerging field of research (Kaiser & Lusardi, 2024; Lusardi & Mitchell, 2014). Several definitions of FL exist, which are all varying in their reach and scope. Due to the large variety in conceptualizations, there have been numerous debates among scholars about what applies to its definition (Remund, 2010). Besides of this ambiguity and the therefrom emerging pluri-interpretability, concepts like financial knowledge, financial literacy, and financial education are often confounded, leading to even more confusion in this discussion (Kimiyaghalam & Safari, 2015). Speaking of FL along with perceptions and calibrations, a crucial distinction is the one between objective financial literacy (OFL) and subjective financial literacy (SFL) (Van Rooij et al., 2011).

According to Hung et al. (2009), FL has in past research been defined highly diverse, as: "(a) a specific form of knowledge, (b) the ability or skills to apply that knowledge, (c) perceived knowledge, (d) good financial behavior, and even (e) financial experiences." (chapter 2). This thesis approaches OFL mainly from a perspective of financial knowledge, as considered to be the most common view (Remund, 2010). This is also substantiated by the homonymous approach of Hilgert et al. (2003). Besides of pure financial knowledge, multiple abilities have additionally been taken into account, as examined related to the derived knowledge queries in the question sets. Although limited in scope, the conceptual specification of "knowledge and abilities" entails the essence of the concept in the chosen data-collection method, and facilitates best the goal and method of the research to apply an examination of differences regarding the difficulty level of questions in the situation of a judgmental FL OC examination.

Speaking of the degree of difficulty, financial literacy includes an extensive spectrum of knowledge and abilities, as it relates to terms and situations from all over the finance and economics field (Karaa & Kuğu, 2016). FL levels have in previous research broadly been classified into two groups of (1) basic- and (2) advanced financial literacy (BFL and AFL), in which BFL is more often related to day-to-day activities, while AFL related terms often appear in specific financial contexts and conditionally need more elaborate knowledge (Lusardi, 2008; Van Rooij et al., 2011). Support for a significant positive relationship between basic and advanced FL has previously been observed by Karaa and Kuğu (2016).

With the inclusion of OC as an important benchmark, the additionally mentioned aspect of perceived knowledge also comes into examination in the sense of SFL. Where OFL was designed by an attribution of knowledge and abilities, SFL can be defined as someone's self-perception about these attributes (Hadar et al., 2013). Generally speaking, objective knowledge and subjective knowledge are often weakly correlated (Alba & Hutchinson, 2000). Subjective financial literacy has previously been found to be the most influencing factor of the two in several contexts of financial decision making (e.g., Anderson et al., 2017; Lind et al., 2020). Findings about the role of OFL here are often considered to be somewhat dependent on the topic and/or one's personal traits. Some studies state that OFL clearly affects certain decision making, while others report no (or barely any) effect in (other) behaviours. For instance, Guiso and Jappelli (2008) found a positive effect of OFL on portfolio spread (risk-assessment), and Van Rooij et al. (2011) and Bucher-Koenen et al. (2021) found a positive relationship between OFL and stock market participation (which could be bi-directional), while a meta-analysis by Fernandes et al. (2014) expressed a limited effect of OFL improvement on decision making regarding (among other) debt- and saving-behaviour,

controlling for personal factors. Due to the apparent more important role of SFL in decision making, often found to be distinct from OFL effects, an examination of the mutual relationship between OFL and SFL is interesting. This relationship is expressed in calibration studies between the two. Recent research has further demonstrated the conceptual difference between both forms of FL, the resulting misjudgements in several contexts of financial households and consumers, and the correlations thereof with personal actions (like financial advisory and advice-seeking) and conditions (like financial well-being) (e.g., Balasubramnian & Sargent, 2020; Lind et al., 2020; Nejad & Javid, 2018; Vörös et al., 2021).

2.2 Cognitive processes, -biases, and some alternative explanations for OC

Prior to the introduction of the overconfidence concept (underconfidence as counterpart), the conceptual illustration of cognitive processes and -biases will be discussed. This is of importance in the explanation and understanding on the occurrence of OC, as cognitive biases and heuristics form part on often used explanations on this phenomenon (Skała, 2008). Additionally, it provides some theoretical depth on the possible appearance of the hard-easy effect and the D-K effect (which contain additional insights), as all terms are often linked to overconfidence and the explanatory term of cognitive biases. Although theoretical motivations and reasons behind the occurrence of overconfidence are way more extensive and can impossibly be fully expressed here, it seems appropriate to shortly discuss some background on this topic. Hereby, several works by Kahneman and Tversky have been taken as a starting point, while some additional explanations are briefly discussed as well.

The concept of cognitive biases was primarily introduced by Kahneman and Tversky in the 1970s. Through multiple fundamental studies, they demonstrated that one's perception can be inaccurate and misleading in certain situations involving judgment or decision making under uncertainty, among other due to the occurrence of heuristics. Several heuristics, which can be seen as intuitive mental shortcuts, were found to play an important role in the development of overconfidence bias; although often very useful to simplify complex judgmental and decision making processes, the use of these simplifications may lead to judgmental errors as well (Kahneman & Tversky, 1972; Tversky & Kahneman, 1974, 1983). This possible influence of heuristics was however not entirely new, as it was foundationally introduced by Simon (1947, 1955) a few decades earlier. Simon reasoned that although people try to make entirely rational choices, the possibility thereto is limited by the capacities of one's cognition, and that therefore simplifications are developed and used.

After their initial exploration, research on cognition gained more popularity, resulting in further elaboration of the courses of cognitive processes as well. With the increasing interest in this subject, cognitive processes were more often differentiated into two ways, often named as dual-process theories (e.g., Evans & Over, 1996; Smith & DeCoster, 2000; Stanovich & West, 2000). Stanovich and West (2000) differentiated the two processes by naming them System 1 and System 2. Kahneman (2003), who referred to these systems as intuition and reasoning, mapped the characteristics of the two systems following the general view. This mapping shows that System 1 (intuition) can, among other things, be seen as a "fast", "effortless", and "automatic" process, while System 2 (reasoning) can be attributed the characteristics that it is "slow", "effortful", and "controlled" (p. 698). Therefrom concluding, people's limited cognitive load is charged less by intuitive judgment and decisions. In the view of the dual-process theory, to indicate people's ability to reason analytically rather than solely using intuition, the cognitive reflection test (CRT) by Frederick (2005) is used often.

This test differentiates between the two systems, as it consists of questions that initially motivate to give an intuitive and impulsive answer, but lead to a distinct outcome after applying reasoning to their solving. One of the main findings of the initial appliance of this test is “that men are more likely to reflect on their answers and [are] less inclined to go with their intuitive responses.” (Frederick, 2005, p. 37).

In the popular book *Thinking, Fast and Slow*, Kahneman (2011) explored the distinction between the two systems further. In this publication, he described that System 1 is more susceptible to cognitive biases due to its reliance on heuristics, while System 2 can, on the other hand, help to reduce cognitive biases because of the critical thinking and logical reasoning that is applied. Additionally, it was described that System 1 is used far more often than System 2, but that as both systems can work both separately and together, it is important to note that the division should be seen as a metaphoric wording, as they cannot be strictly distinguished. Applying the distinction between the systems to a context of judgment without any form of feedback or objective evidence, it can thus be expected that quite intuitive actions (containing more uncertainty) lead to occurrences of misjudgment.

Although explained quite extensively in this paragraph, it should be noted that the influence of heuristics and cognitive limitations are often-used modelled explanations for the systematic occurrence of OC, but that past research has provided further framework developments and additional explanations for findings of OC and the considered side-effects. While the works of Kahneman and Tversky (1972, 1974, 1983) described heuristics as frequent sources of error, Gigerenzer (1991) argued that heuristics can also be useful and rational, depending on the context/environment, aligning with the Brunswikian theory of ecological rationality. Accordingly, this takes a distinct view on the concept of rationality, and the appearances and influences of heuristics compared to the described principles by Kahneman and Tversky. A distinct explanation for OC, which does not directly rely on heuristics, can be found in error models. These rely on the idea that OC findings may arise from regressive errors in the judgment process, and are often supplemented by a methodological dependency perspective (which can also independently lead to distinct OC findings) (e.g., Erev et al., 1994; Juslin et al., 1999; Juslin et al., 2000; Klayman et al., 1999). These alternatives have more recently been discussed/summarized by Moore and Healy (2008) and Olsson (2014). While writing, the concepts and most original sources of the error models were derived from Olsson. The theoretical assignment has been described by both.

The differential or distinct explanations do not necessarily neglect the systematic occurrence of cognitive overconfidence bias, but also rely (partly or entirely) on a substantiation of imperfect environmentally dependent judgments instead of cognitional limitations, and the often-found methodological distinctions and weaknesses/critiques in measuring OC. The methodological differences will be discussed further in the paragraph on overconfidence forms (2.3), while the weaknesses/critiques will mainly be described along the theoretical sections of the hard-easy (2.4) and the Dunning-Kruger effect (2.5). Besides, there are several other factors that might play a role in the emergence of overconfidence, like one’s personal characteristics, one’s experience and expertise with the subject (nevertheless closely related to the environmentally dependent judgments) (Griffin & Tversky, 1992), and the cultural environment. As these variables can also be considered influential on the amount of confidence one has in the selected context, and can be taken into account at a more observable level, these will be considered in the socio-demographic section (2.8).

2.3 The concept of overconfidence

As Benjamin Franklin wrote yet in 1750: “There are three things extreamly hard: steel, a diamond, and to know one’s self.” (Benjamin Franklin, 1750, in Poor Richard’s Almanack, as cited in Sitzmann et al., 2010, p. 169). One of the most prominent findings in earlier confidence research is that most people contain more confidence in the actions they undertake than their actual performance justifies (Griffin & Tversky, 1992; Russo & Schoemaker, 1992). Overconfidence can be seen as an overly optimistic view of own competency and abilities, which is therewith not accurately reflecting reality (Fischhoff et al., 1977; Russo & Schoemaker, 1992). However, no all-embracing definition of the term exists due to the large amount of examinations that are literarily attributed to this subject (Glaser & Weber, 2007). Overconfidence should nevertheless be distinguished from the concept of overoptimism, as these are often confounded, but in reality reflect distinct concepts (Alicke & Govorun, 2005; Herz et al., 2014; Moore & Dev, 2018). Occurrences of overconfidence have in previous research often been assigned as either miscalibrations or additional effects, like the better-than-average effect (Skała, 2008). Compared to research on FL, the branch of research concerned with OC is way older and more elaborate, as can be seen in the long-term research developments on cognitive biases and additional explanations. Effects of OC have previously been tested among applications in several more settings than the one of FL, concerning both economic and non-economic subjects.

2.3.1 Three forms of overconfidence

OC has in earlier research often been assessed as a single construct, while measurement and meaning appeared to be distinct compared to each other, making the representation in the form of a single construct seemingly too abstract (Moore & Healy, 2008). To add more nuance in this construct, Moore and Healy identified 3 separate applications in literature, which are: overestimation, overplacement, and overprecision. As they provided support for the existence of a conceptual difference between these forms, this division is now more often applied in calibration research. As described earlier, Hamurcu and Hamurcu (2020) and Vörös et al. (2021) found support for/substantiated a separation of these forms specifically applied to a FL context, considering objective and subjective dimensions. Consequently, the distinguishment between the 3 forms also sets this thesis’ examinational context.

Overestimation is the easiest form to understand, as it examines the alignment between one’s perceptual and actual performance (in case of a question set). If perceptions overrule the actual performance, a situation of overestimation takes place, and in case the actual performance is higher, the phenomenon is labelled as underestimation (Moore & Healy, 2008). When appearing to be at the same level, the personal performance assessment is assumed to be aligned correctly. Contextually seen, this implies comparing OFL and SFL.

According to past research, overplacement is the kind of overconfidence that is often perceived to be most intuition-based. This is motivated by the perception that one can reason better about own experiences and beliefs, rather than making rational assessments about others’ internal thoughts and performances (Moore & Healy, 2008). This comparison with others is the focal point of overplacement. As described in Larrick et al. (2007), the origin of comparison with others can be theorized related to social comparison theories (e.g., the SCT by Festinger (1954)), according to which multiple motivations can be suggested that declare the connection with, and basis of, external comparison, from which placement misperceptions can arise. Accordingly, this form of overconfidence is a potential corollary of

social comparison. An in literature often assessed effect in this category is the better-than-average (BTA) effect, with the worse-than average (WTA) concept as counterpart (Alicke & Govorun, 2005; Brown, 2012; Larrick et al., 2007; Svenson, 1981; Taylor & Brown, 1988). As the name says, this effect means that one is convinced to perform better than others (the average). Emergences of OP and the better-than-average phenomenon may have both positive (e.g. happiness) and negative (e.g. suboptimal behaviour) effects, depending on the related setting and the characteristic or behaviour that is assessed (Xu et al., 2024). It seems like in behavioural assessment the negative effects generally weigh more heavily in the end. A distinction between OP in performance and ability can be made (Kruger & Dunning, 1999).

“It ain’t what you don’t know that gets you into trouble. It’s what you know for sure that just ain’t so”, a statement that is often credited to Mark Twain (n.d.), describes the essence of the recognition of overprecision (OPR) as a third form of overconfidence. Especially since this credit to Twain can, ironically enough, not be confirmed, and is possibly incorrect (Seybold, 2016). Despite having been attributed (in part) to various other people, full assignation of this statement remains unknown up until now (Quote Investigator, 2018), but that is a distinct topic. OPR is the miscalibrated conviction of “the accuracy of one’s beliefs” (Moore & Healy, 2008, p. 502). In relation to estimations, this can imply that one is overly convinced of the given estimates for oneself and/or others compared to the actual situation. Specifically applied to question set examinations that intent to distinguish the 3 forms, an intriguing difference is the conviction of knowing the truth on one’s own score versus the conviction of the accuracy of estimate judgment for others (Moore & Healy, 2008). Consequently, OPR has multiple application possibilities. An applied example of OPR in this thesis’ context: when one is totally convinced of the estimates for oneself and/or others for the FL question sets, regardless of whether this appears to be OC, UC, or good judgment, an additional form of miscalibrations can arise in the form of OPR; that personal conviction of being right.

2.3.2 Measurement varieties in the forms of overconfidence

As overconfidence has frequently been assessed as a single construct in past research, measurements and definitions have been thrown on one pile in these studies as well, named: ‘overconfidence’ (Moore & Healy, 2008). OC has frequently been measured by comparing self-assessment of knowledge with actual knowledge (e.g., Fischhoff et al., 1977; Glaser et al., 2005; Kruger & Dunning, 1999). First instance, OC might therewith seem to be a relatively unambiguous construct, with a lot of theoretical substantiation. As can be observed in previous research, several forms of measurement have, however, been applied to assess different dimensions of the subjective perception of knowledge questions and sets. Although sometimes appearing to be subtilities, the measurement therewith often yet distinguishes between the 3 forms of OC, not specifically recognizing, differentiating, and naming them. Following reasoning by Glaser and Weber (2007), the findings on a certain form of overconfidence should not be used as a substantiation for other forms of overconfidence.

As it can be inferred that the different forms may occur simultaneously, non-distinction leads to a potential pitfall. Hence, the above-mentioned differentiation is important. While distinguishing, it should be noted that observed levels of OC can differ based on the form of measurement that is used (Juslin et al., 1999; Klayman et al., 1999), with even distinct measurement forms per type of overconfidence that can be applied. Although these studies were mainly targeted at the OPR form of OC, this measurement dependence is considered to be of importance more generally in estimation and placement judgments as well.

Conditionally, it is important to emphasize the existence of differences and methodological dependence, which will be done by discussing some elaborated options and choices for all forms (hence, extending their views). It is important to note that this has been done to give insight into some often-applied varieties, but that possibilities are not limited to the examples given in this overview, as literature is more elaborate in its variations.

An often-applied form of measuring overconfidence in knowledge is a subjective assessment after multiple knowledge questions. For instance, with raw estimates per set of questions, in which one is asked about the expected number of correctly answered objective questions. This has for instance been done in the second and third study of the popular publication of Kruger and Dunning (1999), yet combined with a comparison with others (related to OP, see paragraph 2.5). Specifically applied to a FL-context, this has for example been done by Hamurcu and Hamurcu (2020). Besides of these often-applied raw estimates, estimates have in several studies also been made through Subjective Probability Interval EStimates (SPIES), in which a respondent expresses the self-estimated probabilities of quantitative correctness regarding a number of items in a group of questions. Introduced by Moore and Healy (2008) as Subjective Probability Distributions (SPD), and named in Haran et al. (2010), this has been applied by e.g. Prims and Moore (2017) and Vörös et al. (2021), where in the latter in the context of FL as well. SPIES estimates provide respondents with the opportunity to be more detailed in their expressions and are hence expected to give a more elaborate insight into personal convictions. On the other hand, they are inferred to be quite difficult for some respondents to understand (Prims & Moore, 2017). As these approaches initially focus on the assessment of solely own performance, they can be seen as forms of measuring OE.

Specifically applied to a comparison with others' performance after multiple knowledge questions, both raw estimates and SPIES estimates have also been used in measuring OP. Considering SPIES judgments, own SPIES(1) has been measured and compared with the expected performance of a random other respondent, after which adjusted for one's own overperformance (Prims & Moore, 2017; Vörös et al., 2021). This adjustment is, however, not always made. The BTA effect can as an absolute number also be assessed by simply comparing the estimates for self and others (Brown, 2012). For raw estimates, the starting point has sometimes been changed slightly, as a percentile compared to others can be estimated (for instance: "I believe I am in the top 40% of people answering these questions") instead of an absolute number. For a single person, the effect can in this case be examined by comparing the expected with the actual percentile, and consequently derive a conclusion on whether one was too confident in the comparative statement (Larrick et al., 2007).

Besides of an estimation of performance after multiple questions, on certain occasions one is asked about certainty estimates per question. In that case, a person is asked about the self-estimated certainty that a specific question has been answered correctly. For instance, in the research by Fischhoff et al. (1977), who asked knowledge questions with 2 answer options (that one had to either choose, or compare in the form of a statement), it was looked at both limited (probabilities) and unlimited (likeliness odds) estimates of confidence after individual queries in two of their experiments. In the experiment with probabilities, one was asked about the certainty of correctness in 4 different formats, leading to distinct levels of miscalibration. In the experiment with odd-estimates, one was asked about the likeliness to be correct in comparison with the likeliness to be incorrect, expressed in odds and estimated with unrestricted values. The limited and unlimited estimates were found to differ in several ways, especially in the expression of extremely high OC, yet indicating the importance of the

measurement method. Another application of question-specific measurement of OC can be seen in e.g. the studies of Russo and Schoemaker (1992) and Glaser et al. (2005), who let people estimate lower and higher bound confidence intervals on the question-answers. Regarding knowledge questions with numeric answers, people had to give a lower bound and a higher bound answer to a question with which they were to a certain extent (e.g. 90%) confident that the by them indicated range included the correct answer. Juslin et al. (2007) described that this format generally finds high levels of OC. Being related to the degree of conviction in correctness of the applied estimated range, they can be seen as a ways of measuring OPR. A more elaborate exploration of some of these earlier OPR measures can be found in Olsson (2014). OPR has more recently been measured using the difference between the variance of others' actual scores and the variance of the perceived scores of others when using SPIES (Prims & Moore, 2017; Vörös et al., 2021), or by comparing average confidence indications with the percentual objective score (Hamurcu & Hamurcu, 2020). Hence, although to the best of knowledge not described explicitly in the literature, it seems like at least two distinct applications of OPR can be distinguished: (1) being right on a knowledge question, and (2) being right on an estimate. In the latter, the distinction between the conviction of being right on one's own score, and being right on estimate judgments for others can be made (Moore & Healy, 2008). This broader applicability of OPR is important.

Besides of subjective questions that are referring to the objective questions, regardless of whether these are asked per question or per 'set', subjective knowledge has in the past also been assessed based on queries that were asked on unrelated basis to the objective questions. For instance, Nejad and Javid (2018) talked about misperceptions between objective and subjective knowledge in FL as well, and measured SFL by 3 items on subjective knowledge beforehand (derived from Flynn and Goldsmith (1999)), which were non-referring to the objective questions. Van Rooij et al. (2011) measured it using just 1 item on the expected knowledge before-, and Balasubramnian and Sargent (2020) measured it using just 1 item on subjective performance after the objective questions, asking in general what is perceived to be one's level of FL. In the latter, non-relatedness of the objective and subjective question(s) can nevertheless be countered to a certain extent. The distinction between measurement beforehand and afterwards is in this case important, as people might adjust their perception based on obtained information while answering the questions (Nejad & Javid, 2018). The level of confidence might therewith differ from their initial level due to the gained perception of the questions, indicating the influence of survey order on the observed confidence judgements and the therefrom resulting (mis)judgments.

2.4 Task difficulty and the hard-easy effect

This thesis will take into account the level of task difficulty when examining OC in FL to find out whether differences in overconfidence appearances can be found in this contextual branch. Understanding when a task is deemed to be (more) difficult is therein a crucial aspect. According to a recent concluding view by Krawczyk and Wilamowski (2019), who examined the role of task difficulty on overconfidence in running activities: "task difficulty reflects an interaction between the nature of the task itself and the competence of the person undertaking it" (p. 1). This implies that the determination of the difficulty level contains both an objective aspect and a subjective aspect, where the latter is defined by individual differences. As described, the difficulty level will first instance be taken into account with the distinction between basic and more advanced financial literacy queries. The AFL questions

are expected to be more difficult to answer correctly, as these have been described in Van Rooij et al. (2011) to be “clearly much more complex” (p. 453). Therewith they are affecting task nature, as increased complexity will lead to more uncertainty and doubt. The subjective aspect will be taken into account by means of the measurement of the considered socio-demographic variables, like daily and educational affinity with finance (see paragraph 2.8). Perceived difficulty is measured by taking one’s estimates for others (SPIES2) as a proxy.

Considering differences by task difficulty, the hard-easy effect is a bias closely related to the overconfidence bias, seen its focus on the calibration of self-assessment. Originally assessed with overconfidence as a single construct, the effect of question difficulty (although not immediately assessed under this name) implies that overconfidence is likely to diminish and eventually turn into underconfidence in easy tasks, while in the more difficult tasks the amount of OC is expected to increase (Lichtenstein & Fischhoff, 1977). When distinguishing between the 3 forms, Lichtenstein and Fischhoff looked at OC in a form that in the applied distinction is categorized as OPR. Following a more recent view on this by Moore and Healy (2008), repeated in Moore and Dev (2018), the format of many of these questions (confidence assessed at an individual item level) does not only measure OPR, but OE as well, as these are identical things in these type of questions that are asked.

The study by Lichtenstein and Fischhoff therewith created a starting point for the hypotheses of OE depending on the difficulty level, but they did not say much about OP, neither did they separate the 3 forms from each other (unidentified at that point). Taking a look at the OP form, Kruger (1999) theorized that people generally tend to consider their own abilities, but do not accurately consider the abilities of others when comparing with them. He reasoned that among other due to this phenomenon, the better-than-average effect (closely related to OP, as described earlier) varies with task difficulty: when people are proficient in a task (in absolute terms; an easy task), they generally tend to have a higher expectation of performing better than average, and reversed. This was examined across distinct abilities/domains. This inspection, but categorized on relative terms within a domain and within a set, is also part of the D-K effect examinations, as discussed in the next paragraph. More recently, the effect of task difficulty has also been tested applied to the multiple types of OC, simultaneous and apart from each other. Studies by e.g. Larrick et al. (2007) and Moore and Healy (2008) found support for the hypotheses that at easy tasks people tend to underestimate themselves, but believe they are better than average (BTA) and/or express overplacement, while at difficult tasks people tend to overestimate themselves, but believe their performance is worse than average (WTA) and/or express underplacement. Moore and Healy looked in their illustration at this across question sets of distinct difficulty within a domain, examining groups in general. Despite not always necessarily mentioning the hard-easy effect, the above-mentioned studies provide hypotheses on these forms of OC regarding task difficulty, sometimes substantiated by statistical influences as described below. These difficulty expectations are also expected to be found in the first part of this thesis’ analysis.

As the hard-easy effect is an often-considered effect in calibration research, there are also literary critiques on this effect regarding its added value and the often-applied categorization to cognitive biases (the error models), especially applicable to such set calibrations with their restricted scales. The main critique against the hard-easy effect, considered in contexts of OC examinations that in the current distinction would be put under the OPR form in individual queries, is that the observed levels and patterns are likely to appear at least partly due to statistical grounds instead of cognitional miscalibrations, in which regression towards the

mean, linear dependency, and scale-end effects are the most prevalent ones (e.g., Erev et al., 1994; Juslin et al., 2000; Klayman et al., 1999; Larrick et al., 2007, who discussed one or multiple of these artefacts). These artefacts have been summarized in, and were also partially derived from, Olsson (2014). Larrick et al. (2007) described that the statistical effects of the hard-easy effect mainly occur when applying a difficulty categorization based on objective performance. In that case, so-called 'double dipping' is applied: the objective performance is used to determine both the difficulty level and the overconfidence level. Although distinct in their appearance due to methodological changes in the examination of the 3 OC forms, multiple of these artefacts are also deemed relevant for this thesis' context. Especially when reconsidering the approximate equality between OE and OPR in those earlier studies, as referred to above, and the fact that a set bundling also consists of minimum and maximum levels (scale-ends in the individual sets, see also the comparable D-K effect elaboration on this in the next paragraph). The expected impact of the statistical effects can be found in paragraph 3.3.5.

Some researchers consider findings of the discussed OC patterns mostly unimportant in case of finding statistical substantiation as it would not have a psychological or cognitive background, or leave this out of discussion. Moore and Healy (2008), however, theorize the difficulty patterns to occur as respondents are adjusting their estimate to the expectation of an error as a result of task perception (easy or difficult) according to one's personal "best unbiased estimate" (p. 505). They theorize that this adjustment would be even stronger when estimating for others, as one has even less insights about them. Taking the study of Erev et al. (1994) as main point for their thinking of these regressive effects, their theory, hypothesizing the effects to occur due to informational limitations, seemingly considers the patterns to be more valuable.

2.5 The Dunning-Kruger effect

The Dunning-Kruger effect is often described as a cognitive bias as well, and can be seen as a partial manifestation of illusory superiority (Muller et al., 2021). The basic elaboration of the effect can be explained as thinking that you are better than you actually are, when having a relatively low competence level regarding the subject (Kruger & Dunning, 1999). The appearance of the effect might depend on the (scope of the) assessed subject (Dunning, 2011; Dunning et al., 1989; Kruger & Dunning, 1999). While carrying out examinations of absolute performance estimations in a question set -also done in the hard-easy effect-, this is not the effects' main consideration. The effect itself mostly concerns the large differences in one's perceived performance and ability level on the assessed subject, displayed as a percentile compared to others, related to one's actual competence (Kruger & Dunning, 1999). Dunning and Kruger presumed that people tend to overrate themselves when they know relatively little about the subject, and that their misalignment decreases once they know a bit more, which is due to the at that point obtained insight that they do actually not know that much, or because they can get better insights due to their knowledge. Often added to this theory, but not as specifically mentioned in the original publication, is that people who know a lot about the subject will be undervaluing their competence (Magnus & Peresetsky, 2022). This effect is, nevertheless, often less explicit, leading to a relatively better alignment.

To the extent their examinations look into OE in the realm of the 3 discussed forms of OC (related to set-performance; raw estimates as mentioned in 2.3.2), study 2 and 3 from Kruger and Dunning (1999) find OE at low performance, (moving towards) UE at high performance.

This is reasonably in line with the hard-easy pattern. The quartile ranking on performance concerns both the objective and subjective aspect of task difficulty. Quiz-specific placement has been assessed in these studies as well, being a percentile rating. BTA-expressions emerged in all quartiles without clearly changing patterns across the performance categories. Nevertheless, as categorizing the quartiles on the actual performance (which increases), the percentiles seemingly show OP at low performance and UP at high performance. The difference of the set-specific assessment on a more individual level, categorized on relative performance, when examining the D-K effect should nevertheless be noted compared to the described examination of patterns across question sets to examine the task difficulty patterns in Moore and Healy (2008), in which the sets are mutually compared to get to the estimation and placement expectations about a group in general. Furthermore, the calculational and assessment components (of OP measurement) differ between these studies.

The D-K effect has previously been assessed in the realm of a FL-context by Gignac (2022). Although found when applying the 'original design', no support was found for its existence in this specific context when applying analyses accounting for limitations of the D-K effects' measurement. Just like the measurement of OC patterns in the form of the hard-easy effect, the D-K effect is not uncontroversially described as a (meta)cognitive phenomenon. It is likely to appear at least partly due to regression towards the mean, scale-end effects, and the BTA-effect (e.g., Ackerman et al., 2002; Gignac, 2022; Krajc & Ortmann, 2008; Krueger & Mueller, 2002; Magnus & Peresetsky, 2022, who discussed one or multiple of these artefacts/related concepts). According to these studies and theories, regression towards the mean combined with the BTA-effect can lead to the pattern as described, and occur due to the imperfect relationship between objective and subjective abilities, while performance is always bound to a certain level. The bounds of scale-ends can impact the possibility of the appearance of certain misperceptions during personal assessment, particularly in relation to the actual performance. To clarify: when somebody answers (almost) all questions correctly, the subjective assessment can under no circumstance lead to high overestimation, as this is calculated as the difference between objective and subjective performance, while contrary an objectively weak performance can never lead to a lot of underestimation (Krajc & Ortmann, 2008; Magnus & Peresetsky, 2022). Although some limitations have been recognized by Dunning and Kruger in their initial publication and in later discussions, Dunning also told about the effect in 2022: "They [critics] fail to notice that the pattern of self-misjudgements remains regardless of what may be producing it."

2.6 Click and time data in FL-queries, SPIES, and breaks (hypotheses)

Answering the knowledge questions and making the estimates also comes with some actional characteristics, following variables of click and time data and their relation to the expression of OC levels. Focusing on these behavioural aspects, it can be hypothesized that click and time data in the FL questions might differ depending on the tasks' difficulty level. It is contemplated that respondents will be more insecure about their answers on the more difficult questions. Consequently, they are expected to more often change their answers on these queries. Click data can thus be taken into account in this realm, hypothesizing to see a higher number of changes/clicks in the questions of the more advanced sets. Considering the fact that question length will impact time results, a set or question comparison on time is not appraised to be a valid approach to demonstrate one's uncertainty or reconsideration. Time data of the knowledge questions can, however, be looked into meaningfully applied to a

socio-demographic group division within a set, or when examined at a correlational level. For the knowledge questions, this consequently only leads to the concrete expectation that the number of clicks will be higher in the more difficult sets. If remained unobserved, this might be due to a consideration before selecting the answer(s). This does, nevertheless, still give some indication about one's behaviour when answering questions in such a setting.

Time data differences can be assessed more extensively in SPIES estimates, as question length approximately equals in this appraisal. Thereby, it should be taken into account that more elaborate SPIES estimates, with a higher variance, logically take more time. These estimates could be sorted based on the number of selected components and be compared between the basic and advanced sets based on this sorting to compare fairly (a time/click ratio). Sole click data in SPIES estimates is considered to be less meaningful, concerning increasements in clicks can occur rapidly when re-adjusting performance (e.g. one might adjust a .7-.3 estimate to a .6-.4 estimate (2 additional clicks), or adjust this to a .6-.3-.1 estimate (3 additional clicks)). This is highly susceptible. People with a high OPR are hypothesized to have less clicks and need less time on both individual questions and SPIES estimates. Herein, especially the individual questions are interesting, as for the SPIES questions this expectation is related to the OPR calculation (variance of estimates, a logical consequence). Furthermore, for the SPIES estimates, it is hypothesized that both click and time data will be lower for SPIES2 than SPIES1 concerning possible respondent fatigue as a result of unreasonable uncertainty (see paragraph 2.7) for a certain group, while another group might actually express higher levels due to increased consideration. To summarize, distributional differences are expected to be found here.

Regarding breaks, time data is intriguing as it informs about the scope of the break. Time data, combined with the personal indication of taking a break, shows the duration of the break and one's personal perception hereof. It is hypothesized that people who take a break make more accurate judgments due to the clear distinction between the sets, for instance substantiated by ideas of bounded rationality (Simon, 1947, 1955), limited cognitive load (e.g., Sweller, 1988), and dual-process theories (e.g., Kahneman, 2011), as described earlier. Especially considering survey design, with the limited number of queries per set (see methodology/Appendix 1-3); also proposed influential for FL levels (Anderson et al., 2017).

2.7 CONFSPIES and full-estimates (hypotheses)

As will be clarified later, respondents are asked about the confidence they have in their SPIES estimates after taking the 3 sets (CONFSPIES). Considering the hypothesis, but the simultaneously strong statement, from Moore and Healy (2008) that people know more about their own performance than the performance of others, and the idea of Kruger (1999) of mostly focusing on oneself when comparing with others (relevant in the realm of indirect comparison in case of the SPIES2 estimates), it is reasoned that the CONFSPIES indication should normally (base-case scenario) be higher for SPIES1. While many respondents are expected to meet this criterion, there might also be a group of respondents that does not. It is hypothesized that this group might express some misunderstanding on either SPIES or CONFSPIES questions. Particularly when considering the unfamiliarity with making the SPIES estimates (Prims & Moore, 2017). It might also occur due to fatigue. Consequently, both groups are split and compared, leading up to the cautious expectation that misperceptions might be higher for the group that deviates from the expected relationship sign. Additionally, it is intended to look at the number of full-estimates that are made. These

might indeed occur due to full confidence, but can also be a sign of fatigue, e.g. from a point of misunderstanding or indifference. Especially when considering that no incentives have been provided for participation. Consequently, these full-estimates are related to the CONFSPIES distinction as described above. No specific expectations have been formed for this data collection beforehand, although it was reasoned that this data could be of influence in the analysis as full-estimates are particularly doubtful when made for SPIES2 estimates.

2.8 Socio-demographic variables

When looking at OC in FL, there are several socio-demographic variables related to personal characteristics, one's experience and expertise, and the cultural environment that can be considered to have an impact on levels of financial literacy and the occurrence of OC in this context. Looking for substantiation of variables in these categories, it is again important to differentiate between studies taking into account overconfidence as a single construct, and studies taking into account OC as an umbrella term for the sub-forms OE, OP, and OPR. The separate forms might demonstrate distinct effects per socio-demographic characteristic.

First of all, the personal characteristics gender, age, and income are taken into account as presumably introducing differences in the level of financial literacy (Lusardi & Mitchell, 2014) and the appearance of the forms of overconfidence. Bucher-Koenen et al. (2017) quite recently described the existence of a large FL gender-gap, with gender differences in the amount of confidence as well: men were found to have generally higher levels of both financial literacy and confidence. The confidence level was therein assessed based on "Don't know"-responses of the questions, and therewith not related to a specific type of OC. This makes it somewhat unclear what sign to expect applied to the 3 forms of overconfidence related to gender (overconfidence is distinct from confidence (Moore & Dev, 2018)). The existence of this gender difference in OC has on a more general level been observed earlier in Lundeberg et al. (1994), in a context of filling in test questions, in which they also indicated the existence of domain dependent differences. In a finance-unrelated context, Prims and Moore (2017) found some effect of gender on differences in all three overconfidence types, although appearing on a varying basis. Furthermore, they did not find support for moderation by age for OE and OP, but they found a positive correlation with OPR. In this thesis, this will be tested in the applied FL context. The expected relationship of age with levels of financial literacy is bell-shaped (Lusardi & Mitchell, 2014; Van Rooij et al., 2011). Regarding income, a positive correlation with FL is expected (Lusardi & Mitchell, 2011b). The expected relation of income with overconfidence is not entirely clear. It has been assessed earlier by Isidore and Christie (2019) regarding OC in decision making, who found a positive relationship.

Furthermore, one's highest level of education, the relation of education with financial subjects, and one's daily affinity with financial activities are assessed, as earlier found to be (and consequently expected to be) positively related with OFL (Lusardi & Mitchell, 2014; Van Rooij et al., 2011). Contrasting the latter two variables, which were derived from Van Rooij et al. (2011, pp. 469-470), wording and scale items (and thereby the variables themselves) have been changed slightly. These variables are partly relating to personal characteristics, but mainly intriguing due to their role in financial experience and expertise, which might affect levels of OC as well, being part of the subjective part of task difficulty. People who are well-informed about a subject, often containing higher levels of experience and/or expertise, have previously found to be better at predicting what they know than people who were less-

informed (Kruger & Dunning, 1999). For the umpteenth time, it is therein important how these variables are conceptualized and measured (Sanchez & Dunning, 2023); just like FL, difficulty levels, and OC forms. Following suggestions and findings by Lusardi and Mitchell (2011b), Lusardi and Mitchell (2014), and Balasubramnian and Sargent (2020), self-employment and marital status have also been considered, as these might affect FL levels as well. The impact of these variables on the overconfidence forms is unknown.

Furthermore, a factor of risk-assessment is considered. The attitude towards risk-taking might influence the way the subjective questions are answered, and therewith influence OC levels. This is taken into account by asking respondents directly about their perceived risk aversion and by focusing on perceived uncertainty afterwards. Herewith, it is distinct from Van Rooij et al. (2011), who looked more specifically at FL risk taking using an entirely distinct measurement approach (3 FL-related questions). Additionally, competitiveness might affect the mutual relationship between one's estimates of own results, and others' performance. These are only viewed shortly (out of interest); individual characteristics are more extensive. Self-enhancement, as potentially influential variable following its focus on presenting oneself better than one would actually predict (Krueger, 1998; Moore et al., 2018), was not taken into account due to the individual and anonymous character of the survey.

Regarding the cultural environment, multiple scholars found an influence of cultural aspects on the amounts of overconfidence. For instance, Yates et al. (1998) found cross-cultural variations in OC that were likely to be more generalizable. Furthermore, in more recent studies, Heine et al. (1999) and Chui et al. (2010) argued that the prevalence of individualism -which is more common in certain cultures (Hofstede, 1988, as cited in Hofstede, 2011)- is of importance in the amount of (decisional) OC in comparison with others (OP). Individualism may also affect OPR (Moore et al., 2018). Lechuga and Wiebe (2011) found support for an impact of cultural differences on the OPR(/OE) form(s) of OC. There are also studies that dispute the often-generalized robustness of findings on cultural distinctions. This has e.g. been done by Moore et al. (2018) with regards to the 3 OC types in relation to the difficulty level. To be on the safe side, trying to avoid unconsidered side-effects, this thesis' survey has solely been directed to Dutch residents to limit the impact of large cultural differences.

3 Methodology

3.1 Research design

This study uses a quantitative data-collection approach to examine judgment levels among respondents and look into the behavioural aspects of performance estimations. The data is collected using a closed-structured online survey. It was chosen to set up a survey with a single-group comparative 2x3 within-subjects design, which implies that all respondents received the same knowledge questions, in which they were subjected to multiple conditions. Hence, respondents were asked to answer both the BFL ánd AFL questions (2 groups), and judge their performances in all sets, from which the OC forms could be calculated (3 groups).

By choosing this design, socio-demographic characteristics between the sets are equal, facilitating a reliable environment for fair comparison. This decision was made beforehand anticipating the expected limited access to respondents, considered to be one of the main difficulties of the data collection procedure. In case of a comparative cross-sectional design with separate groups (between-subjects design) for the BFL and AFL questions, more respondents would be needed to achieve the same statistical power (Charness et al., 2012).

The survey has been designed and executed in Qualtrics' survey software, as the online survey tool of the University of Twente. After implementing feedback and recommendations by the first supervisor and obtaining ethical approval from the UT Ethics Commission (nr. 240993), the survey has been distributed. No incentives were offered for participation. The survey has been included in Appendix 1 (in Dutch).

3.2 Sampling procedure

Considering the selected within-subjects design, which affects survey length, attaining a sufficient number of respondents remained one of the main concerns. Due to constraints in available time and resources, it was chosen not to limit the sample to certain characteristics that might limit the total number of respondents extensively. The non-specification of the sample is justified by the nature of the subject, as everyone contains FL to a greater or lesser extent and therefore might be prone to judgmental OC bias in this context. Further specification is obviously possible, but would in many cases appear to be unsubstantiated. A convenience sample was used to further address this challenge of respondent recruitment.

The main drawback of this non-probability form of sampling is that it leads to reduced generalizability compared to probability forms, due to the increased possibility that the addressed group of people is an erroneous representation of the population (Andrade, 2021; Jager et al., 2017). According to Jager et al. (2017), this disadvantage can be limited by conducting a homogeneous instead of a heterogeneous sample, although in homogeneous samples it is more difficult (or, in case of tight narrowing: almost impossible) to compare socio-demographic groups. As this comparison forms part of the desired analysis, a middle path has been chosen in which two sampling-criteria were put on the data collection. This contains (1) that the respondent should live in the Netherlands, and (2) that the respondent should have reached at least the age of 18 when filling in the survey. The first constraint was set to limit the impact of potential cultural differences, as described in paragraph 2.8. Regarding the second constraint: reaching this age is the moment that a person is defined to be financially responsible in the Netherlands, as a consequence of the legal capacity that is in principle obtained (the exception for below the age of 18 is no longer applicable) at the age

of majority (18+) according to Articles 1:233 and 1:234 (*Book 1, Title 13 – Minderjarigheid*) of the Dutch Civil Code, and Article 3:32 (*Book 3, Title 2 – Rechtshandelingen)* of the Dutch Civil Code. Derived from these articles, this means that financial decisions taken by an adult person can in principle be taken without the explicit permission of (a) parent(s) or caregiver(s) and will legally fall under one's own responsibility. As it can thus be expected that one's FL has been somewhat developed at this age, and the questions of the survey can be reasonably understood, this legal boundary was set as the limit for survey participation. If one has no knowledge about FL at a basic level of understanding, one might also not differentiate between easier and more difficult questions, as every question appears to be difficult. Furthermore, this was adopted due to ethical considerations. To ensure compliance with the chosen constrains, participation criteria were explicitly stated in the consent form.

3.3 Survey design

3.3.1 Survey components

The survey begins with a discussion of data collection, privacy, and ethical considerations. These are outlined towards respondents in the representation of a consent form. In this form, respondents are provided information about the anonymity and confidentiality by which their data will be treated. Additionally, it is explained that the collected data will be used for the purposes of this thesis, and that the thesis will be published in the university's online repository. Furthermore, the consent form informs about the voluntary form of participation and the option to withdraw at any time. Respondents had to specifically agree with the described considerations, declare they met the established participation criteria (18+, and living in Netherlands), and give permission for the use of their data in this research, to start with the survey. Hereby, the collected data has been summed. After starting, respondents were informed on how to make the estimates, and were provided with an example to increase the probability the assignment was understood before starting the first question set.

The data collection for the first part consists of 2 main types of questions: (1) Financial literacy knowledge questions to measure OFL in the two difficulty categories of BFL and AFL, and (2) perception-related questions to measure SFL and the three forms of overconfidence. Key papers that have been important in setting up this main structure can be seen in Figure 1. For the finance-based knowledge questions, the question sets have been adopted from, and are translated from, 2 developed modules of the De Nederlandsche Bank Household Survey (DHS) by Van Rooij et al. (2011, pp. 452, 454). Explicit permission from Elsevier, as copyright holder of their paper, has been obtained for reuse, serving purpose adjustments, and (data) reporting of both question sets in both English and Dutch.[2] Permission has additionally been checked at CentERdata, as executor of the DHS. The sets originally

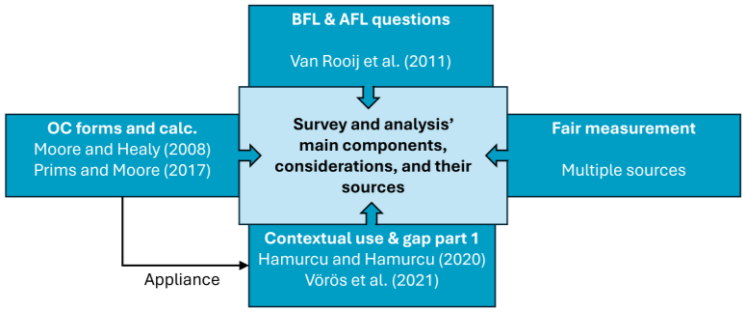


Figure 1: Survey and analysis' main components, considerations, and their sources

[2] From Journal of Financial Economics, 101(2), van Rooij, M., Lusardi, A. & Alessie, R., Financial literacy and stock market participation, pp. 452, 454, Copyright © Elsevier (2011). Explicit written permission from Elsevier has been obtained for reuse, small adjustments, and (result) reporting of the FL questions from Box 1 and Box 2.

exist of 5 basic FL questions and 11 advanced FL questions. As described in Van Rooij et al., some of these questions are derived from earlier surveys and studies. An oversight of question origin and inspirations (before their use in the DHS) can be found in Appendix 2. This has been done to give credit to these sources as well. Content wise, no problems are expected in terms of regulations or comprehensibility, as the questions have been asked before in the Netherlands.

Multiple of the questions follow the design building blocks of: simplicity, relevance, brevity, and differentiation, as described in Lusardi and Mitchell (2014). After reconsidering these building blocks for all questions, it was chosen to remove one of the AFL questions. This specific question, about characteristics of bonds, is perceived to intentionally measure the same perception as one of the other questions, which asks about the characteristics of stocks. Although both questions asked about distinct financial concepts, they shared identical answer options, leading up to a situation that responding to the one question in a certain way likely influenced the given answer to the other question. To clarify: one answer option becomes basically unapplicable as it has been chosen as the answer for the other question. This leads to a situation where the confidence in a certain answer might influence the confidence in the other interrelatedly. Besides of the removal of this question, the “Refusal” answer-option as included in the original setting has been deleted for all questions, as this option would not provide any useful information for the data analysis in this context. This option was exchanged for “Insufficient time”, and combined with the “Don’t know”-option, as the questions are assessed time-bound (see paragraph 3.3.2).

Furthermore, the question about the time value of money (question 4) has been adjusted to own terms, following reasoning and discussion by Zytek (2018). The adjusted question by Zytek (2018) has been optimized further by keeping the question simple (not specifying the investment asset or giving a specific return, solely mentioning the investment and the sign of this return), as the question is intended to be part of the BFL set. Furthermore, it has been optimized by specifying the statement of: “richer today”, as this was unspecified to the context of inheritance as mentioned in the original question. This cause-and-effect relationship is considered important here. In Q14, the term “company fund” was specified, as this can be interpreted broadly. It was expected that only mentioning this term might confuse respondents. Lastly, the answer option: “None of the above” in the question about the relationship between interest rates and bond prices (Q15) has been removed, as the other answer options (rise, fall, and stay the same) yet present the entire possible range.

After the implementation of these rearrangements, the question total consists of 5 basic questions (1 adjusted) and 10 advanced questions (2 adjusted). It was chosen to split the AFL questions into two blocks of questions, ending up with three sets of five questions in total. This has been done as the number of questions in a set might influence the extent to which performance can be judged well. It was expected that the performance on a set of ten questions would be more difficult to judge than a set of five questions, as one might yet have forgotten the first question at the point one arrives at the last one, e.g. due to the limited cognitive load as discussed in paragraph 2.2. As we want to assess solely the influence of the difficulty of the questions on the accuracy of the judgment, unrelated to the influence of difficulty due to the number of questions in a set, it was chosen for this design. Where other examinations often looked at ten- or even twenty-item sets, the sets of this thesis provide the large advantage that respondents are less likely to have forgotten about questions, which is expected to limit measurement bias. Furthermore, the set distinction provides the option to

give respondents structure and pauses while answering the questions, likely improving clarity, consciousness, and respondent satisfaction as well. The 3 sets are presented to the respondent in random order to avoid respondent certainty about which are the BFL and AFL sets, as knowledge thereof might impact the results. Moreover, respondents are beforehand not familiar with this strict distinction between the sets, as they are only provided with the information that some questions might be experienced as more difficult (considered important to motivate); this might also occur in mixed order. All questions consist of 3, 4 or 5 answer options (2, 3 or 4 real options, and 1 option: "I don't know / Insufficient time"). On each occasion, only one of the real options is correct. Although merged in the same answer option, responses are seen as "Don't know" if (1) selected and (2) clicked through within the time-limit of 50 seconds, and categorized as "Out of Time" when (1) selected and (2) automatic advancement after 50 seconds is applied. Following Xia et al. (2014) and Balasubramnian and Sargent (2020), "Don't know"-responses (DK) are considered to be incorrect. The same reasoning has been applied to the "Out of Time"-responses (OoT).

According to Perreault (1975), the survey questions should be asked in random order to prevent from Question Order-Effect bias. Furthermore, randomization of questions helps to mitigate effects of respondent fatigue, in which answers to questions are influenced by motivational factors of respondents due to the position of the questions in the survey (Hochheimer et al., 2016). Without randomization, the knowledge questions that are instantly placed at the end of the survey might show worse performance in both question and judgment performance due to a reduced focus and/or motivation of the respondent. The desired randomization has been adopted within the realm of the 3 question sets, randomizing the order of the sets, the order of the questions within these sets, and the order of the answer options within these MC questions. Complete randomization is not possible as a result of the desired perception-related questions after each set (difficulty categorizations should be preserved). After each of the sets, respondents are asked about their perception of their own performance and their perception of the performance of others to measure OE, OP, and OPR in the way as described in paragraph 3.3.3. The question-order considerations were later found to be mostly in line with Moore and Healy (2008). To further prevent from Question Order-Effect bias, socio-demographic questions should ideally be placed at the end of the survey (Perreault, 1975); this was implemented. Appendix 3 shows the survey flow.

3.3.2 Measurement considerations for fair measurement of OFL questions

As every survey setting has its pros and cons, this thesis' survey is no different. Setting up an online survey introduces some uncertainty. In the specific setting of this thesis, this uncertainty mainly relates to the nature of the questions. As the data collection is partly focused on knowledge questions, people might try to look up the answers to the questions if they don't know them or want to verify their initial thoughts, which would disrupt the estimation processes afterwards. This especially creates some risk among the group of younger respondents, who generally have a higher attitude towards and are generally more skilled at using digital sources (Czaja & Sharit, 1998). Beforehand, one could expect answer sets of the questions to be available online, as some of the questions are popular and therewith have been used often in the past. After some online search, no answer set for the totality of the translated questions could be found. This basically excluded the possibility of easy cheating. The answers to several questions could, however, be found by including (parts of) the English version of the questions into search engines. Hence, a huge advantage

of this thesis' survey is that it is conducted in Dutch. As the questions are put under the cover of this language barrier, questions and related answers did barely come up in the search engines, making cheating in this seemingly effortless way quite difficult.

While the possibility of cheating using search engines has mainly been ruled out, there is another possibility to look up answers, which is with the use of Artificial Intelligence (AI). In times with the rise of personal access to AI-tools, obtaining the answers online in other ways gets easier rapidly. Applying the above described findings of Czaja and Sharit (1998) to more recent developments, younger respondents may hence more likely possess the abilities to use AI for this purpose of cheating. Usage of additional online tools can, nevertheless, never be ruled out entirely without the use of e.g. personal tracking tools or a restricted online environment. As conducted by a student, it lacks resources to mitigate the risk by taking one of these measures. The possibility and motivation to look things up can nevertheless be reduced largely. This was intended by applying the following 4 measures:

1. Commitment request – At the start of the survey, respondents were asked about their commitment not to cheat, as commitment language turns out to be effective in such requests (Mazar et al., 2008). The way of stating this query might affect the given answer, and one's behaviour (Krosnick, 1991). The respondent had to respond to this request in a positive way to continue with the survey. Past research on (political) knowledge questions has found an effect of reducing yet more than half of the reported cheating by applying this measure: from 14% to 6% (Clifford & Jerit, 2016). This request is in line with Van Rooij et al. (2011).

2. Time limitations – A time limit of 50 seconds per question was put on the knowledge questions to encourage a quick reply, let people focus on the survey, and limit the time to look up answers. This time limit should, considering expected reading and consideration time, for most people be sufficient to give an informed answer. Average silent reading times for English text are considered to be between 175 and 300 words per minute (Brysbaert, 2019), with generally quite similar results for Dutch texts (Brysbaert et al., 2021). Considering that each question, including answer options, contains between 20 and 60 relatively complex words, the reading time per question will likely take between 6 and 20 seconds. This gives respondents an additional 30+ seconds to consider the right answer. Clifford and Jerit (2016) found that time limits are also an effective approach to avoid reported cheating, with a reduction of almost half as well: from 14% to 8%. Although a combination of approaches was not addressed in their study, it is sequacious to expect that a combined-measures approach with both commitment requests and time limitations is even more effective.

3. Disabling return option – The option for respondents to come back to the survey at a later time after filling in a few questions has been turned off, as this discourages switching from the survey to look things up. No substantial problems in terms of data collection are expected in applying this measure, as filling in the survey takes respondents about 12 minutes. Although relatively long for a student survey, this should for most people be a reasonable time to finish it in a sole attempt, in which respondents can (and are explicitly given the opportunity to) take a break between the sets. Besides, this measure has been necessary due to limitations of the Qualtrics platform, restarting question time when re-entering the survey. Consequently, this return option would basically provide respondents with unlimited time to answer the knowledge questions. Furthermore, the option to return to previous questions has also been disabled, as this resulted in a similar situation.

4. Sequential question display – Lastly, it was chosen to display the 15 knowledge questions sequentially to make the use of additional tools for obtaining answers inconvenient for respondents. Simultaneously, this measure should make respondents who cheat behaviourally conscious, as the undesired behaviour 'has to be' repeated over and over again. This approach may prompt respondents to consider the ethical implications of cheating more in depth, conceivably making them change their behaviour (if necessary).

Besides of preventive measures, cheating has also been checked during data analysis, in which dubious cases could be identified based on combined extreme outlier responses of time per set and levels of OFL. With these additional measures, cheating can only occur if a respondent (1) ignores the instruction, (2) is untruthful about the usage of additional tools, and (3) deliberately tries to manipulate the results while reconsidering the wrong actions at every question. This combination of bad intentions is simply unlikely to apply to the vast majority of respondents. Together with the informal setting in which the survey is likely to be answered, and the fact that no incentives are offered, it makes looking up the answers of the same calibre as deliberately filling in survey questions incorrectly while having a malicious intent to manipulate the outcomes. The latter is possible in basically every survey setting, which makes the remaining risk acceptable. Considering the possibility of looking up answers is important, but assuming instant occurrence is more like defeatism.

3.3.3 Measurement considerations for OC forms

Considering the opportunities and pitfalls of the several measurement varieties as described in paragraph 2.3.3, it quickly becomes clear that the methodological design is more complex and comprehensive than one might initially think. Combined with the OC forms in paragraph 2.3.2, there are several options to consider in choosing measurement. Luckily, past research has already provided fairly clear suggestions in other contexts, although individual considerations and preferences regarding the specific situation, design, and personal convictions (developed using the theoretical background) must be taken into account. For instance, it had primarily been decided that the expression of confidence levels should be limited to a certain extent (no unlimited odd estimates). This has been chosen due to the conviction of arbitrariness in situations of high confidence. For instance, how can one objectively state to be 1000 times more certain than uncertain, instead of 500 or 2000 times? This becomes even more affective on overconfidence levels when applying this to even larger numbers, e.g. millions. Furthermore, it was intended to relate the SFL questions to the OFL questions. If placed independently, respondents might get confused about the actual scope of the concept while answering the SFL question(s), especially beforehand. Besides, this has been necessary to measure the 3 forms of overconfidence simultaneously.

As these considerations and preferences suit with the measurement descriptions of OE, OP, and OPR as mostly described by Moore and Healy (2008, pp. 508-509), further developed in Prims and Moore (2017, pp. 31-32), and likewise applied to a FL context in Vörös et al. (2021, pp. 1295-1296), there is no substantiated reason to deviate from this measurement approach. Their measures and recommendations have been followed for all 3 forms. Overestimation is calculated as the difference between SFL and OFL, in which SFL is constructed and calculated from SPIES data (SPIES1). Although indirectly theorized that it is more difficult to make rational judgments about others than about oneself (Moore & Healy, 2008), and multiple studies have shown that "when people compare themselves with their peers, they focus egocentrically on their own skills" (Kruger, 1999, p. 221), SPIES are also

applied to the measurement of the expected performance of others (SPIES2), as following the above-mentioned references. This uncertainty sets the main stage for the analysis' second part. To improve the comprehensibility of respondents on their task of estimating the SPIES question probabilities, it was chosen to let the answer options on questions add up to a maximum (and minimum) sum of 100%. The amount of OP has been measured in line using both one's own actual and perceived score, the perceived score of a random other, and the average score. The calculation consists of comparing one's own expected performance with the expected performance of a random respondent (SPIES1 and SPIES2), while taking into account and adjusting it for (subtracting) actual overperformance (OFL minus average FL). OPR of an individual was calculated in accordance with aforementioned studies by examining the difference between the individual variance of the SPIES2 estimate and the general variance in the distribution of the actual scores. Considering smaller set size than the earlier studies, both components are likely to decrease. Note: very slight difference in OP and OPR calculations in line with Vörös et al. (2021), using total average and variance.

Considering the corresponding within-subjects design, the beforementioned idea of splitting the basic and advanced FL questions into several sets, and the corresponding measurement of OE and OP, it should be noted that the first part of the analysis (task difficulty effects on OC in a FL quiz) closely aligns with Moore and Healy's (2008) design of the quiz stage and interim stage in their illustrative study. It is, however, distinct in the way it calculates OPR as described above (used in their study solely as a robustness check; here the changes from Prims and Moore (2017) come into play), is conducted in the Netherlands, uses only 5 questions per set (limiting memory constraints, considered in line with the bounded rationality ideas of Simon (1947, 1955) and the Cognitive Load Theory by Sweller (1988)), contains 2 'more difficult' sets of approximate equal difficulty, gives no rewards, and specifically applies the examination to a FL context. Moore and Healy write: "We cannot claim that the pattern we observe would hold regardless of context, task domain, or subject population, although our theory does not suggest that these factors should matter" (p. 514). In other words: they appeared to be somewhat uncertain, or at least reserving some judgement, about this more general applicability of their findings. Conditionally, putting it in this FL context seems to be substantiated and useful, especially combined with the past-years calls to distinguish and compare easy and difficult categories in FL misjudgments, the applied contextual distinctions as described above, and the extent to which segments of the second part of the analysis build on this. It also differs in result reporting, looking more into calculational components.

As described, explicit permission from the American Psychological Association, as copyright holder of their paper, has been obtained for reuse of their instruments (SPDs in multiple sets) and OC calculations/measures, together with the earlier discussed methodological factors leading to a partial contextually distinct replication.[3] Following terms, permission for the reuse hereof -and the publication in the university's online repository- has also been obtained from the authors. "Wholehearted permission" was granted (D.A. Moore, personal communication). To ensure compliance with the conditions, some unspecified terms have additionally been checked with the APA Copyright Department. The article by Prims and Moore (2017), with SPD named as SPIES, which is using a distinct main OPR calculation, is licensed under the Creative Commons Attribution 3.0 License (https://creativecommons.org/licenses/by/3.0/).

[3] Copyright © 2008 by APA. Adapted with permission. Moore, D. A. & Healy, P.J. (2008). The Trouble With Overconfidence. Psychological Review, 115, 502-517. https://doi.org/10.1037/0033-295X.115.2.502. No further reproduction or distribution is permitted without written permission from the American Psychological Association.

3.3.4 Measurement considerations for other behavioural analyses

During each of the FL-knowledge questions and perception-related questions, click and time data are monitored to be able to perform this second part of the analysis on additional respondent behaviour. As the consent form yet clearly informs respondents about this data collection, this part is performed in the back of the survey to provide a neutral environment to the respondent. While answering the questions, respondents should not instantly be distracted by the monitoring of this data, as shifting focus. Hence, respondents are by no means influenced in the time they take and the clicks they make, except for the predefined time limit of 50 seconds which might give some pressure. Seen the elaboration of this limit, unexpected to push respondents (see paragraph 3.3.2), this effect should be limited as well.

The click and time data consist of the collection of 4 variables for each of the knowledge questions and SPIES estimates: (1) first click, (2) last click, (3) total clicks, (4) total time, being the standard data collection approach in Qualtrics. In the realm of the behavioural analyses, total clicks and total time are hereof the most intriguing variables to look into. Click and time data have also been collected during the breaks to be able to objectively check whether respondents really took their eventually indicated breaks, monitor the duration, and consequently derive whether this contradicts with later statements to clarify the threshold.

For the insight in one's confidence in SPIES1 and SPIES2 estimates (CONFSPIES1 and CONSPIES2), question order (first confidence in own estimates, then confidence in estimates for others) and comparison by respondent (displayed on the same page, directly after each other) are considered to be important. The analysis on full-estimates does not require any additional data collection. Altogether, the examination of these behavioural aspects in making the estimates is mostly initial and explorative, focusing on the nature of the SPIES estimates and the applied set-divisions in combination with design choices.

3.3.5 Additional deliberation to limit and convey methodological deficiencies

As described in paragraphs 2.4 and 2.5, the examination of OC, the hard-easy effect, and the D-K effect contain statistical influences that are likely to impact the findings, depending on the method. This thesis focuses on a categorization between easy and difficult questions based on a predefined categorization. Manipulation of the difficulty level beforehand was theorized by Larrick et al. (2007) to limit the impact of statistical artifacts. Considering the performance of the FL questions in Van Rooij et al. (2011), the categorization between BFL and AFL is mostly in line with objective performance. Although task difficulty differences are made on theoretical grounds and delimitations following an existing question set, it does not entirely get rid of a ranking by performance; knowledge questions are never equal (Table 3, paragraph 4.2). It does, however, make a difference that the 3 sets follow this categorization.

The scale-end effects are expected to be more likely to occur in these sets of 5 questions than in sets with 10 or 20 questions, as the 'in-between space' between the extreme values is larger in these bigger sets. To limit the possible influence of scale-end effects to a certain extent, it was anticipated to use the basic and advanced questions of a yet existing question set, of which results could be reasonably estimated and considered beforehand. As both the basic and advanced questions seemed to be answered correctly at a relatively acceptable distance from the scale-ends (78.7% and 53.8% average correctness in Van Rooij et al. (2011)), it was expected that these would not be affected significantly by these limits. Furthermore, to act transparently, the results sections of estimation and placement

judgments contain descriptions of the likelihood of their occurrences within the sets. There seems to be a difference between scale-end effects in actual scores and estimated scores.

3.4 Data analysis procedure

The data analysis is performed in IBM SPSS version 29.0.2.0. Extreme outlier responses, identified by applying combinations of extreme time (short) and OFL scores (high ánd low), are critically evaluated to prevent from cheating and respondent fatigue. Based on the therefrom arising valid responses, the results section consists of multiple segments of descriptive statistics, visual analyses, and differential significance calculations.

The analysis starts with an examination of the sample characteristics and actual levels of OFL. Although not considered to be the main focus of the results section, it is inevitable to contemplate and analyse the values of these upbuilding and informational sections. Sample characteristics and OFL levels are intriguing in comparison with population characteristics, as representability and generalizability can be deduced from this to a certain extent, while the objective performance also acts as a cornerstone for the calculation of some OC forms. Introducing OC results, the correlation between OC forms has first been looked at to assess the conceptual difference applied to the FL context and compare it with earlier findings. Regarding the difficulty analysis, the initial comparison takes place between the two levels of difficulty (1 basic set, 2 more advanced sets) and the 3 forms of OC (OE, OP, and OPR), as drawn in Figure 2, forming the 2x3 within-subjects design. The difficulty level forms the independent variable, while OC forms are considered the dependent ones.

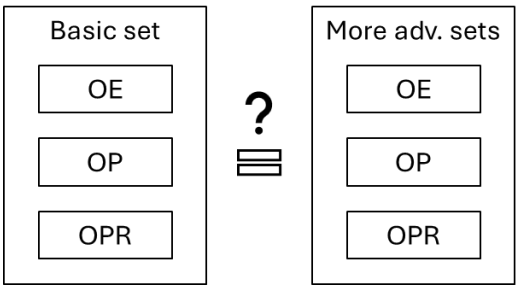


Figure 2: Starting point of the difficulty analysis

This is extended as schematized in Figure 3. Within the course of the OE and OP-curve, patterns in line with the task difficulty effects (across sets) are tried to be examined in this specific context (question 1). Thus far, the OC analysis is basically in line with Moore and Healy (2008). Furthermore, socio-demographic variables that potentially influence OC differences are taken into account, using t-tests and non-parametric alternatives (considering the within-subjects design, and the therefrom emerging dependence of observations), applied to the difficulty categorization as well (question 2). The variables gender (GEN), age (AGE), income (INC), education level (EDU), highest education finance affinity (Likert item) (FINEDU), and daily affinity with economics (Likert item) (DAYECO), see paragraph 2.8 for origin and expectations, are included.

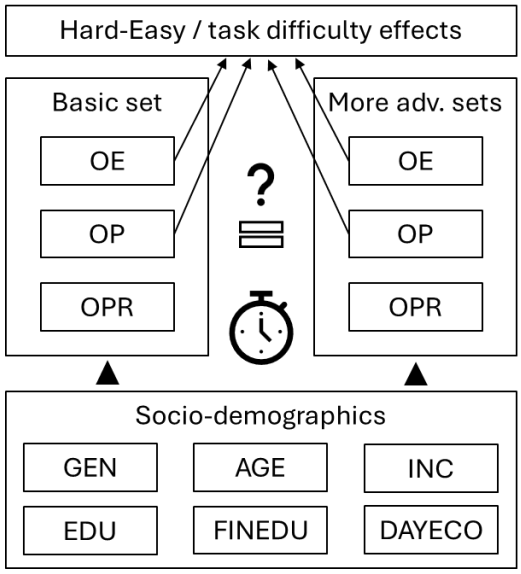


Figure 3: Extension of the difficulty analysis

As described, the second part of the analysis revolves around one's behaviour in answering the knowledge questions and making these estimates. This starts with the inclusion of click and time data (indicated by the clock-pictogram in Figure 3), as potentially providing

information about respondents' certainty and effort in the given answers and estimates (question 3). This data has, as far as known, not been collected before in this context and consequently provides new ideas and insights. For instance, this data makes it possible to examine differences in the appearance of forms of OC between in-time and automatically selected (out-of-time) answers, and between seemingly certain answers (1 click) and less-certain answers (>1 click). With this data collection, it could also be looked at OC differences related to categorizations in taken time to make the judgments of the SFL questions and the needed time to answer the question sets and make the SPIES estimates. This has been done to provide further information on whether, and in which circumstances, forms of confidence misperceptions and miscalibrations seem to occur. Click and time data are also prevalent in the analysis on the relationship between the OC forms and the voluntary breaks.

Lastly, additional analysis is performed on respondents' behaviour in making the estimates (question 4), especially regarding SPIES2 uncertainty. Respondents' confidence in their SPIES estimates is examined, in which Likert items on one's confidence in SPIES1 and SPIES2 estimates are asked for directly after each other to facilitate an initial comparison. The analysis is also partly based on the variance of the estimates, differentiating between full-estimates (variance = 0), and split-estimates (variance > 0). To facilitate this, dummies have been created. These variables (CONFSPIES and full-estimates) have also been examined mutually to e.g. check whether people with equal or less confidence in their own estimates than in their estimates for others (illogical indication) showed significantly different behaviour in the variance of the provided estimates. This has been done to take a look at potential misunderstanding or survey fatigue when making the estimates.

4 Results

4.1 Descriptive statistics

This thesis' sample comprises a total of 107 respondents, who spent on average about 12.7 minutes to complete the survey, which is relatively long for a student survey. Respondents' characteristics are shown in Table 1. The socio-demographic variables gender, age, and highest education are compared with estimated population characteristics (calculated using data from Centraal Bureau voor de Statistiek (2024a, 2024b), see footnote 4 for further specification) to roughly demonstrate commonalities and dissimilarities between the compositions of the reached sample and the addressed population on a principal level.

		N	%	EPop%[4]
Gender	Male	52	48.6%	49.4%
	Female	53	49.5%	50.6%
	Prefer not to say	2	1.9%	-
Age	18-24 years	34	31.8%	11.0%
	25-34 years	25	23.3%	16.2%
	35-44 years	6	5.6%	15.1%
	45-54 years	20	18.7%	15.7%
	55-64 years	19	17.8%	16.9%
	65+ years	3	2.8%	25.1%
Highest educ. level	Primary education (primary school)	0	0.0%	9.3%
	Secondary education (high school)	8	7.5%	55.1%
	Secondary vocational education (MBO)	38	35.5%	
	Higher professional education (HBO) associate degree	6	5.6%	21.2%
	Higher professional education (HBO) bachelor's degree	25	23.4%	
	Scientific education (WO) bachelor's degree	11	10.3%	
	Higher professional education (HBO) master's degree	4	3.7%	13.8%
	Scientific education (WO) master's degree or 'higher'	14	13.1%	
	Prefer not to say / Unknown	1	0.9%	0.6%

		N	%
Income	€0-€19.999	25	23.4%
	€20.000-€39.999	29	27.1%
	€40.000-€59.999	24	22.4%
	€60.000-€79.999	8	7.5%
	€80.000-€99.999	0	0.0%
	€100.000+	4	3.7%
	Prefer not to say	17	15.9%
Main employ.	Student	21	19.6%
	Employed	71	66.4%
	Self-employed	7	6.5%
	Unemployed	3	2.8%
	Retired	2	1.9%
	Other	3	2.8%
Highest educ. finance affinity	Very little	34	31.8%
	Little	18	16.8%
	Some	24	22.4%
	Much	14	13.1%
	Very much	17	15.9%
Daily finance affinity	Very little	17	15.9%
	Little	29	27.1%
	Some	29	27.1%
	Much	20	18.7%
	Very much	12	11.2%
Com-petitive	Strongly disagree	4	3.8%
	Disagree	15	14.0%
	Neutral	34	31.8%
	Agree	41	38.3%
	Strongly agree	13	12.1%

Table 1: Descriptive statistics

[4] Estimated population percentages (EPop%) are calculated based on 2024 CBS Statline data. Gender and age data were retrieved from Centraal Bureau voor de Statistiek (2024a), and highest education data was retrieved from Centraal Bureau voor de Statistiek (2024b) (Q1 2024 data). For the latter, a population between 15 and 90 years old is assumed due to filter limitations in that dataset. Although not entirely accurate, this provides estimates that should be fairly close to the proportions in the actual relevant population.

Table 1 shows that the ratio between men and women rather well satisfies the estimated population composition. Furthermore, respondents' age groups are reasonably spread for a relatively small convenience sample, although two comments must be made to the demonstrated dispersion. Primarily, it should be noted that the age groups of 35-44 and 65+ are represented inadequately. This can be declared by means of the sampling strategy, and/or as an outcome of the digital form in which the survey was distributed. People aged 65 and older are generally expected to be less digitally literate, and to use online platforms to a lesser extent. Therefore, this group was presumably less likely to come across the online survey. Secondly, it should be noted that the youngest age groups (18-24 and 25-35) are relatively overrepresented, which is likely a result of the convenience sampling as well.

Highest education levels are limitedly in line with the estimated population characteristics, despite the fact that the sample contains a reasonable spread for such size. The sampling strategy has led to a situation of overrepresentation of the more theoretically based education levels: 57.0% of the sample has obtained a professional degree or 'higher'. Income appears to be considerably well-spread in the lower groups. Taking into account that the sample is relatively young, this is in line with the expectations. Nevertheless, higher income groups are therewith consequently represented substandard. The large number of "Prefer not to say" answers (15.9%) for this question is deemed interesting. Looking at main employment, most of the sample consists of employed people (66.4%) and students (19.6%).

4.2 Objective financial literacy levels

4.2.1 Check on the appearance of self-selection bias

Regarding the subject of interest, financial literacy, one might expect self-selection bias to play a role among respondents in the decision to participate. Preliminary, it can be reasoned that people with higher financial education and/or a higher financial knowledge in general may be more likely to participate in this study than people who perceive themselves as less financially literate. A large occurrence of this bias could be detrimental to the findings in this sample. In this realm, the data collection purpose might actually give somewhat of an advantage to limit the occurrence of this bias. Since potential participants are approached with a clear request for assistance in completing this thesis, sometimes paired with an emphasis on the challenges of finding respondents as a student or an effort-based motivational statement on participation, respondents are likely motivated mostly by this intention to help somebody. As a result, they might be less inclined to actually drop out or not participate at all if they possess lower financial knowledge.

To check the existence of a reasonable spread regarding finance enthusiasts and people that are less concerned with this topic, trying to get insights into potential overdominance of ultimately financially literate people, the earlier indicated variables about the affinity of one's highest education level and the affinity of the daily activities with financial subjects are useful. These were purposely asked at the end of the survey to increase the probability that respondents were familiar with what this matter of "financial subjects" entails. Furthermore, as no data could be found on the populations' degree of financial interest, and therefore the findings on educational and daily financial affinity are somewhat limited in expressiveness, the FL outcomes of Van Rooij et al. (2011) have been used as a benchmark (considering their large sample). The findings on these two checks are described in the sections below. Please note that these are only explorative checks, not a final conclusion.

Highest educational- and daily affinity with financial subjects

As could yet be seen in Table 1, levels of highest education and daily affinity with financial subjects are broadly dispersed across respondents. With respectively 29.0% of respondents showing “much” or “very much” highest education finance affinity, and 29.9% showing “much” or “very much” daily activity finance affinity, it can fortunately be deduced the sample does not solely comprise people who possess high affinity with financial subjects. These mutual relationships are further elaborated on in Table 2. Categories merged, the data shows that 23.4% of the respondents show high (much & very much) affinity levels in both groups, and that only 10.3% of the total shows really high affinity twofold (very much). Opposing to this, 30.8% shows low (little & very little) affinity in both groups, including 12.1% of the total really low (very little). Both affinity variables show significant correlations with OFL at all sets, in which correlations and significance increase in line with the difficulty level according to a .2-.3 range at the 5% level for BFL and a .35-.50 range at the .1% level for the AFL sets (also combined). Correlations with the SPIES1 questions (.3-.5 range at the .1% level) remain relatively equal for daily affinity, while enlarging with the difficulty level for H Edu affinity.

H Edu affinity \ Daily affinity	(Very) little	Some	(Very) Much
(Very) little	33 (30.8%)	12 (11.2%)	7 (6.5%)
Some	11 (10.3%)	13 (12.2%)	0 (0.0%)
(Very) Much	2 (1.9%)	4 (3.7%)	25 (23.4%)

Table 2: Mutual affinity dispersion

Benchmark comparison of financial literacy levels

As the FL questions have been taken from Van Rooij et al. (2011), results can be compared. This has been done in Table 3. Together with the comparison with estimated population characteristics, and the dispersion of affinity levels as described above, the outcomes inform about differences, similarities, and the results’ possible generalizability to a certain extent.

There are several differences in both the results and the implementation of the questions that might have led to the percentual distinctions as described in the table: (1) Considering the samples’ demographic composition as observed in Table 1, it could yet be expected that the FL levels of this sample would turn out to be relatively high. Especially when considering the relatively high levels of education, combined with the moderately high levels of educational and daily affinity with financial subjects. Although the entire spectrum demonstrates higher levels of FL, performance differences mainly emerged in the third set (+17.8%). (2) Dissimilarities in the adjusted questions (*) might be caused by formulation changes. All adjustments

QUESTION	RESULTS OWN SAMPLE	VAN ROOIJ ET AL. (2011)	ABS DIFF.
Q1	94.4%	90.8%	+3.6%
Q2	77.6%	76.2%	+1.4%
Q3	87.8%	82.6%	+5.2%
Q4*	87.8%	72.3%	+15.5%
Q5	87.8%	71.8%	+16.0%
SET 1	**87.1%**	**78.7%**	**+8.4%**
Q6	78.5%	67.0%	+11.5%
Q7	59.8%	62.2%	-2.4%
Q8	72.9%	66.7%	+6.2%
Q9	60.7%	47.2%	+13.5%
Q10	73.8%	68.5%	+5.3%
SET 2	**69.1%**	**62.3%**	**+6.8%**
Q11	86.9%	63.3%	+23.6%
Q12	48.6%	30.0%	+18.6%
Q13	74.7%	60.2%	+14.5%
Q14*	71.9%	48.2%	+23.7%
Q15*	33.6%	24.6%	+9.0%
SET 3	**63.1%**	**45.3%**	**+17.8%**
TOTAL	**73.1%**	**62.1%**	**+11.0%**

Table 3: Correctly answered FL questions, compared with van Rooij et al. (2011)

were made to clarify terms or correct small mistakes. Increased correctness levels are in line with expectations, as respondents are provided with more clarity or less MC-options. (3) This study shows dissimilar design choices in terms of the 50 second time-limit per question, the slightly different population (18+, instead of 22+), the deviation of the questions into 3 sets (introducing set specific confidence and increasing structure), and the further adoption of cheat prevention measures.

Despite these differences, there are multiple similarities and benefits as well: (1) Besides the fact that the questions are mainly the same, it seems like the difficulty distribution among the sets is quite accurate. The first set appears to be by far the easiest with its average question' correctness percentage of 87.1%, while the other two sets follow at a considerable distance (69.1% and 63.1%; 66.1% on average). The convergence of the performance on the two more advanced sets compared to its expectation (difference of 6.0%, instead of 17.0%) might actually be beneficial for the sake of this thesis, as a closer related performance would expectedly lead to closer related judgments and an easier comparison with the basic set. On the other hand, this will lead to more of a 'moderate-easy effect' consideration, instead of the hard-easy effect. (2) 14 out of 15 questions (93.3%) show an increase in performance, indicating the same reaction to the higher FL levels in this sample. Considering the relatively small sample size, these increases seem to be distributed quite evenly. Only 1 of the 12 non-adjusted questions shows an absolute difference of >+20%.

4.2.2 Distribution of OFL levels, and measuring Cronbach's Alpha

Zooming in on respondents' set dispersion, correctness of answers is distributed as shown in Table 4. The basic set demonstrates a left skewed distribution (-2.010) (capped at the max) with a high kurtosis (4.109), while the more advanced sets are less skewed (-.653 and -.168) and show a significantly lower kurtosis (-.538 and -.712). The distributions hence shift from clearly skewed to close to normal. Especially set 3 shows this close to normal distribution.

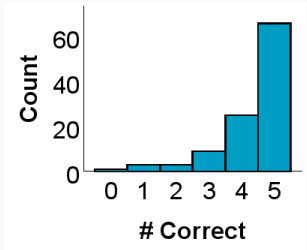
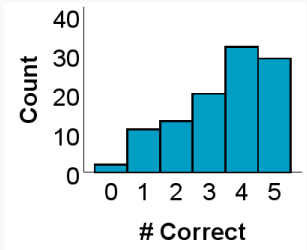
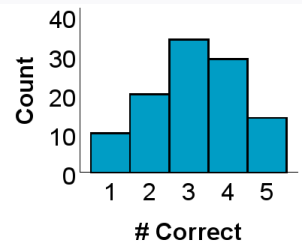
		OFL SET 1 (BASIC)			OFL SET 2 (MORE ADVANCED)			OFL SET 3 (MORE ADVANCED)		
		N	%	Cum. %	N	%	Cum. %	N	%	Cum. %
Correct	**0**	1	.9	.9	2	1.9	1.9	0	.0	.0
	1	3	2.8	3.7	11	10.2	12.1	10	9.3	9.3
	2	3	2.8	6.5	13	12.1	24.3	20	18.7	28.0
	3	9	8.4	14.9	20	18.7	43.0	34	31.8	59.8
	4	25	23.4	38.3	32	29.9	72.9	29	27.1	86.9
	5	66	61.7	100.0	29	27.1	100.0	14	13.1	100.0
	Total	107	100.0		107	100.0		107	100.0	
Avg. correct ~ incorrect		4.355 ~ 0.645			3.458 ~ 1.542			3.159 ~ 1.841		
Visualization										

Table 4: Set-specific composition of OFL levels

As described in Van Rooij et al. (2011), BFL and AFL questions are “noisy proxies” (p. 454) to measure the actual constructs. The Cronbach’s Alpha’s of both show that these question sets are indeed suboptimal in measuring the two constructs, as found to be <.7. Differences with this threshold are nevertheless relatively small for the totals (.644 for the 5 basic questions, and .647 for the 10 advanced questions). With these values, they both belong to an in research vaguely described range of moderately acceptable levels (Taber, 2018). No value is placed on the lower Cronbach’s Alpha of the individual sets within the “advanced group” (.558 for set 2, and .332 for set 3), as both sets consist of some relatively difficult (e.g. 75% correct), and some very difficult (e.g. 50% correct) questions. Lower levels of Cronbach’s Alpha are inevitable and obvious results of this yet expected variability within these smaller sets, especially considering their range in subject diversity. The only implication is that the two sets from now on will be placed under the header of “more advanced sets” while making the set comparison, instead of using the concept of “AFL”. This is because the internal reliability does not provide sufficient support for subcategorization or reuse of the conceptual term for the separate sets in this way. The two more advanced sets are thus not supposed to perfectly represent the AFL construct or measure its dimensions in any way; they should appear to be significantly more difficult than the basic set in order to perform the analysis on differences in OC forms when filtering on difficulty level. Applying visual analysis on Table 3 and 4, it seems like this purpose has been achieved. The actual occurrence of this emergence will be tested to add statistical support for this premise. Also regarding one’s perception hereof. The origin and approach of these comparisons are in line with Moore and Healy (2008), but the applied tests and the use of SPIES2 for the latter differ.

Performing a rank comparison between the sets, the OFL level of set 1 shows to be significantly different from those of sets 2 and 3, while the latter two do not significantly differ from each other (see Table 5). This was tested using a Related-Samples Friedman’s Two-way Analysis of Variance with Bonferroni correction, as both the differences between the groups ánd the OFL levels per group were not distributed normally (SW < .001, QQ-plots unaligned). The Bonferroni correction was applied to this test, as otherwise probabilities on a type I error would increase while making multiple comparisons (Armstrong, 2014). The mean rank of set 1 was 2.50, while those of set 2 and 3 were respectively 1.86 and 1.64. Friedman’s test finds plausible support for the idea that set 1 differs from set 2 and 3 (p_{adj} < .001) in difficulty, while set 2 and 3 seem like they have produced relatively similar outcomes to each other (p_{adj} = .282). This finding is in line with the a priori categorization and the emerged expectations based on mean differences and graphs as displayed in Table 4.

Pairwise Comparisons

Comparison	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
OFL set 3-OFL set 2	.229	.137	1.675	.094	.282
OFL set 3-OFL set 1	.864	.137	6.323	<.001	.000
OFL set 2-OFL set 1	.636	.137	4.648	<.001	.000

Each row tests the null hypothesis that the distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .050.
a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

Table 5: Related-Samples Friedman’s Two-way ANOVA of OFL

Getting acquainted that the 2 more advanced sets can objectively be considered more difficult than the basic set, and are distributed approximately equal concerning objective performance, it is important to check whether this has subjectively also been the case. People might perform objectively speaking significantly different on a set, while not

recognizing or perceiving this difference in difficulty level. We also want to look into whether this perceptual difference exists. SPIES2 estimates were taken as a proxy for this perceived difficulty, as for these estimates one tends to consider the question sets in their totalities, disregarding own perceptions and assumptions that are included in SPIES1 estimates (e.g. in case of guesses). The SPIES2 ranks of set 1 (with $\bar{x}$ = 3.200, s = .103) were found to differ significantly from those of set 2 (with $\bar{x}$ = 2.616, s = .096) and 3 (with $\bar{x}$ = 2.661, s = .085), while the latter sets do differ very little (see Table 6). In this Friedman's ANOVA, the mean rank of set 1 was 2.46, while those of set 2 and 3 were both 1.77. The null hypothesis of equal (rank)distributions between SPIES2-values of the sets has hence been rejected for combinations set 1-2 and 1-3. Again combined with a consideration of mean differences, this suggests that the difference in objective difficulty has also been experienced as such.

Pairwise Comparisons

Comparison	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
SPIES2 set 3-SPIES2 set 2	.005	.137	.034	.973	1.000
SPIES2 set 3-SPIES2 set 1	.696	.137	5.093	<.001	.000
SPIES2 set 2-SPIES2 set 1	.692	.137	5.059	<.001	.000

Each row tests the null hypothesis that the distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .050.
a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

Table 6: Related-Samples Friedman's Two-way ANOVA of SPIES2, as proxy of difficulty

4.2.3 Correlations between OFL and the socio-demographic variables

This sample finds a negative correlation (r) between age and OFL in all separate sets, more advanced sets combined (AFL), and the total. On the one hand, this is quite an unexpected finding, since past research often talks about a bell-shaped distribution when relating the two (Lusardi & Mitchell, 2014; Van Rooij et al., 2011). On the other hand, this can actually be explained reasonably well, considering the distribution of the sample. The largely represented younger groups (18-24 and 25-34) showed significantly higher levels of education, highest educational affinity with FL and daily finance affinity. These variables were found to correlate positively with OFL levels in all sets, with as a matter of fact an even more prevalent relationship in the more advanced sets, as mostly described earlier. This distinct finding is therewith likely unassignable to this specific context, but more of a consequence of the imperfect sampling. Looking at income, a positive significant correlation with FL was observed, as found earlier by Lusardi and Mitchell (2011b) as well. This was only found for the more advanced sets, both separately and together, and the total. Not for the easier set, which is an interesting finding in the realm of OFL set dispersion. In line with Karaa and Kuğu (2016), a small positive correlation between BFL and AFL appeared (r = .238*, p = .014).

4.3 Overconfidence expressions, correlations, and patterns

Correlation coefficients

Introductory to the results of the 3 overconfidence forms between the sets, it is intriguing to inspect absolute OC levels and look into their mutual correlations. The

	OVERESTIMATION	OVERPLACEMENT	OVERPRECISION
Set 1	-.764	.391	.610
Set 2	-.681	.161	1.349
Set 3	-.397	.101	.733
Sum	**-1.842**	**.653**	**2.692**
Avg.	**-.614**	**.217**	**.897**

Table 7: Overconfidence expressions per set

absolute values are summarized in Table 7. The table shows that estimation misperceptions were actually found in the form of UE (underestimation) in all 3 sets. OP and OPR have both been observed to varying 'positive degrees'. A positive and highly significant correlation between OE and OP (r = .534***) in their totalities was found. Splitting into the separate sets, it was observed that this correlation increased in line with increasement of the difficulty level (set 1: .639***; set 2: .693***; set 3: .708***). No significant correlations of OE and OP with OPR are found in any of the sets. Nevertheless, the mutual correlations of OPR between the sets (all between .5 and .75) are striking, especially considering the weaker (often in the range .2-.3) or even non-significant mutual correlations of OE and OP across the sets. Applying visual inspection using histograms and QQ-plots, OE and OP seem to be approximately normally distributed in all sets and in the total amounts, while OPR follows a uniform left-skewed distribution in the sets. Below, the main findings on the emerged misperceptions are summarized per form, considering the calculational aspects, possible appearance of scale-end effects, relevant socio-demographic variables, and the presence of patterns aligning the difficulty effects. The implementation of RM ANOVA's for the estimation and placement analyses was derived from Moore and Healy's (2008) analysis-design.

Estimation judgments

Examining levels of OE, Table 7 reveals that respondents (on average) actually expressed amounts of underestimation in this sample for all sets. The correlations between OFL and SPIES1 per set are respectively: .548***, .665***, .572***, and their total correlation even measures .725***. Due to the high absolute levels (larger than anticipated) of OFL and SPIES1, scale-end effects will likely have influenced estimation levels in especially the first set. Sixty-six respondents showed maximum BFL levels, creating a situation that these could only underestimate or be calibrated correctly. In set 2 and 3 this was way less (see Table 4).

Please note that this is a different perspective than scale-end effects in one's estimates. The possibility hereto should be considered reasonable as well, seen the substantial number of 23 respondents who estimated to have answered all questions correct, and thus could only be calibrated perfectly or express OE. These respondents *might* have overestimated their performance (more) if this were possible (in a situation in which a scale limit does not restrict the misperception). A big sidenote must however be placed to this division, as people with a perfect estimate and score are assigned to this group as well, while these respondents theoretically show among the ones with the best judgments (15 respondents). Belonging to this perfect calibration group does obviously not imply one cannot show OE, but there is, more importantly, no reason to assume one systematically would. For the more advanced sets, this probable additional OE was the case for solely 9 (7 perfectly calibrated) and 7 (4 perfectly calibrated) respondents, making the impact on these estimates quite negligible. The values show a proportional distance between the sets of 3.3:1.3:1.0 (3.7:1.7:1.0).

Considering the large number of respondents in each of the sets that showed a perfectly aligning perception of estimates (two-thirds of respondents with 5-out-of-5 estimates), it is expected that the impact of estimate scale-end effects has not been substantial. Even when considering a highly unlikely hypothetical scenario in which all 39 cases show an additional 3 points of OE (severe), the main finding would remain to be UE, even in all sets separately. Hardly any scale-end effects are anticipated in the lower bound, considering the low number of respondents with 0-out-of-5 estimates (set 1: 0 (0); set 2: 3 (1); set 3: 0 (0)). The components of estimation assessment calculations are displayed in Table 8 and Figure 4.

	Estimates for self (SPIES1)	Objective FL (from Table 4)	OE level (from Table 7)
Set 1	3.591	4.355	-.764
Set 2	2.777	3.458	-.681
Set 3	2.762	3.159	-.397
Total	**9.130**	**10.972**	**-1.842**
Avg.	**3.043**	**3.657**	**-.614**

Table 8: Components of estimation assessment

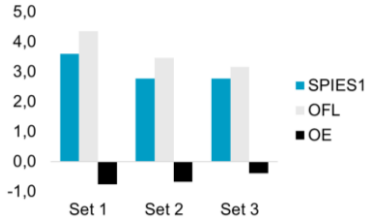


Figure 4: OE comp. visualization

As the differences between the sets are approximately normally distributed (SW = .052 (set 1~2), .175 (set 2~3), .501 (set 1~3)), and Mauchly's Test of Sphericity is found to be non-significant (p = .080; sphericity assumption is not violated), significant differences between the sets can be tested using a Repeated Measures ANOVA. The slight non-normality of one of the dependent variables (OE set 3, according to the SW-tests) is handled with more flexibility, as this test is generally found to be quite robust against non-normality in case of the finding of sphericity (Blanca et al., 2023). Besides, the QQ-plot seems to be reasonably aligned. The Within-Subjects effects table of the RM ANOVA shows a significant difference in UE between the groups (F = 3.824, p = .023). Performing post-hoc tests with Bonferroni adjustment (see Table 9), the mean value of UE in set 1 is found to be significantly different from set 3 (p = .015). The difference is found to be non-significant when comparing set 1 and 2 (p = 1.000), and set 2 and 3 (p = .114).

Pairwise Comparisons (Post-hoc OE)

(I) Set	(J) Set	Mean Diff. (I-J)	Std. Error	Sig. [b]	95% Confidence Interval for Difference[b]	
					Lower bound	Upper bound
1	2	-.082	.153	1.000	-.454	.289
	3	-.366*	.128	.015	-.677	-.056
2	1	.082	.153	1.000	-.289	.454
	3	-.284	.135	.114	-.613	.045
3	1	.366*	.128	.015	.056	.677
	2	.284	.135	.114	-.045	.613

Based on estimated marginal means
*. The mean difference is significant at the .050 level.
b. Adjustment for multiple comparisons: Bonferroni.

Table 9: Post-hoc analysis of mean difference in OE

The expressed pattern in UE across the sets is reasonably in line with the pattern of the hard-easy effect. The pattern is matching the expected effect sign, as UE levels change significantly from .764 to .397 comparing set 1 and 3, but not necessarily its predicted effect size, as UE remains prevalent for all sets. Respondents kept estimating their results were lower than they actually were, which likely occurred due to the higher than anticipated OFL levels in set 2 and 3 (due to reasonably knowledgeable participants). Although described as advanced financial literacy, about two-thirds of these questions were answered correctly as well. Therewith, this finding of remaining UE across the sets actually becomes a quite logical one. Furthermore, it might stand out that the difference between set 1 and 2 is clearly non-significant, especially when considering the significant difference in both objective and subjective difficulty between these sets as found earlier (Table 5 and 6). Concerning difficulty level, one would expect this second set to behave in line with set 3. Taking into account that the more advanced sets were perceived very similar according to Table 6, this finding is therewith not of such deviant kind, as differences are small. SPIES1 estimates between the more advanced sets are basically equal as well, while OFL levels differ only slightly between

these sets. This .3 difference is nevertheless yet enough to make the underestimation in the second set differ non-significantly with the first set.

Looking at socio-demographic differences, a significant difference in UE was shown for gender at the .05 level in the basic set (BFL), in which women expressed clearly more UE (difference of .44 point) than men (p = .044), while their actual BFL levels (x̄ men = 4.37, x̄ women = 4.41) were approximately the same. The approximate equality of OFL levels indicates that this difference does likely not occur due to the possible influence of statistical influences, like scale-end effects or regression towards the mean, as discussed earlier. Using Cohens' d for interpretation of the effect size (d = .398), the effect is considered to be small to moderate (Cohen, 1988; Sullivan & Feinn, 2012). The finding does in part align with Bucher-Koenen et al. (2017), since they found men to express more confidence (measured in DK-responses) than women when answering FL knowledge questions. On the other hand, Bucher-Koenen et al. (2017) found higher FL levels for men as well, and one could reason DK-responses to lead to more certainty, hence better judgments. Significantly higher FL for men has only been observed in set 2. The gender difference in estimation perceptions has not been observed in the more advanced sets (p = .621 and p = .223). For the basic set, a small positive correlation with both affinity variables was found. The other socio-demographic factors did not show significant correlations with estimations within the sets or total.

Placement judgments

Respondents show levels of OP in all sets. It should, however, be noted that this OC form is more prevalent in the BFL set (p = .003) than in the more advanced sets, as in the latter the appearance can be considered non-significant (p = .225; p = .365). Considering the number of 0-out-of-5 and 5-out-of-5 estimates for others, it is expected that within SPIES2 estimates barely any scale-end effects will have influenced the results. For set 1, such extraordinary estimations were made by only 7 respondents (upper and lower bound combined), for the second set only 3 respondents did this, and for set 3 there were none. The components of calculation are displayed in Table 10. Calculated for sets instead of individuals, OFL levels here directly reflect set averages, making the OP calculation consist of solely the difference between SPIES1 and SPIES2. Hence, average OFL levels are displayed within parentheses.

	Estimates for self (SPIES1)	Estimates for others (SPIES2)	(Avg.) OFL (from Table 4)	OP level (from Table 7)
Set 1	3.591	3.200	4.355	.391
Set 2	2.777	2.616	3.458	.161
Set 3	2.762	2.661	3.159	.101
Total	**9.130**	**8.477**	**10.972**	**.653**
Avg.	**3.043**	**2.826**	**3.657**	**.217**

Table 10: Components of placement assessment

As all differences between the set-specific OP levels are approximately normally distributed (SW= .227 (set 1~2), .747 (set 2~3), .262 (set 1~3)), sphericity was not violated (p = .058) after one extreme outlier was located and removed, the dependent variables are close to normally distributed as well (anew, following the QQ-plot), and respondent dependency and a continuous dependent variable are present, the assumptions for a Repeated Measures ANOVA are met. The Repeated Measures ANOVA gives a p-value of .247, designating that the difference in mean values of OP are considered to be non-significant between the sets. As it is uncertain whether the extreme outlier is an actual representation of cases to be

occurring in the population (the data does not give a clear conviction of fatigue), it is intriguing to check what would happen if the outlier remains included. In this event, the sphericity assumption would be violated. As the epsilon is above .75 (ε = .929), the Huynh-Feldt correction should therefore be used to adjust the degrees of freedom (Huynh & Feldt, 1976). This would lead to a non-significant p-value as well, although clearly closer (p = .154).

Considering the expectation on task difficulty effects, the expressed pattern is again approximately in line with the expectations. The demonstrated pattern matches the expected effect sign, as levels drop from .391 to .161 and .101. Nevertheless, this time there are two striking features with the prospects of the effects. The first is that the OP differences are found to be non-significant, while the sets were found to differ significantly in their difficulty. The fact that the effect is not expected to be exactly linear (Moore & Healy, 2008) can account for the non-significant OP differences. The other lies again in the predicted effect size, as OP remains prevalent for all sets and does not turn into underplacement at any point. This, again, likely occurs as task difficulty is not dispersed across the entire possible range (moderate-easy), making the findings align expectations. No significant correlations of socio-demographic variables with OP were found in any of the sets, nor in the total amount.

Although OP calculations are extensive in indicating the exact differences, a large disadvantage is that individual information gets lost quickly as the belief of being better (worse) than others, which is the first part of its calculation (difference SPIES1 and SPIES2), is directly combined with whether this expression is justified, being the second part (adjustment for own overperformance compared to average). The data of these separate components is, especially in the realm of task difficulty, quite intriguing to look into as it might indicate changes in the patterns along with distinct difficulties. This delves deeper into where differences come from, and where they eventually balance each other out. A schematic description of perceived and actual performances can be found in Appendix 4. Please note that BTA here is based on a mean comparison, instead of respondent percentiles. BTA percentages can become deviant from the standard 50% expectation. This might occur as the set-distributions are not (perfectly) normally distributed (Harris & Hahn, 2011). Despite providing additional insights, absolute values/differences here are lost due to categorization.

It turns out that for the basic set 63 respondents (58.9%) believed they performed better than a random other, 21 respondents (19.6%) believed they performed equal, and 23 (21.5%) thought they did worse. In total, 66 people (61.7%) overplaced themselves in this set (as OP on an individual level can emerge due to underperformance as well), which is equal to the number of people that had a performance that was actually better than average. It shows that people were about 70% accurate for both the better- and worse-than-others estimates. For the more advanced sets, the idea of being better than others was expressed by 56 (52.3%) and 52 (48.6%) respondents. OP was expressed by 51 (47.7%) and 64 (59.8%) respondents in these sets. Being BTA was the case for 61 (57.1%) and 43 (40.2%) respondents. More systematically, these data on a categorized level are displayed in Table 11.

Measure	Set 1	Set 2	Set 3
% of respondents with SPIES1 > SPIES2 (part 1 OP calc.)	58.9%	52.3%	48.6%
% of respondents with OFL > average FL (part 2 OP calc.)	61.7%	57.1%	40.2%
% correctness BTA (compared with BTO estimate)	71.4%	75.0%	59.6%
% correctness WTA (compared with WTO estimate)	69.6%	68.6%	79.5%
% of respondents with overplacement	61.7%	47.7%	59.8%

Table 11: OP components on respondent level

Considering percentages, no ultimately clear patterns emerge when putting set 1 on the one side, and set 2 and 3 on the other side. Regarding correctness of BTA and WTA estimations, one might reason that the relatively lower BTA and higher WTA correctness in set 3 might occur due to the higher difficulty. In that case, this would nevertheless also be expected in the second set. Set 2 is actually quite comparable with the basic set on these percentages. Concludingly, one's mutual relationship between own estimates and estimates of others moved not exactly, but reasonably, in line with the actual situation in quantitative terms. This can be seen in the mutual dispersion on BTO and WTO estimates: increasing WTO in the more difficult sets, and the amount to which BTA/WTA were correctly indicated quite equally.

Precision miscalibrations

Examining both components of OPR calculation, with on the one hand the variance (σ^2) in actual OFL scores, and on the other hand the subtraction of the variance in the estimate for others (SPIES2) per respondent, Table 12 and Figure 5 are formed. In this table, the sum and average values are derived from the variances of the individual sets (!). As variances are calculated using a non-linear function, they do therewith not equal variances as can be found when looking at the total of the scores for all sets together (as for OE and OP). These are provided for consistency purposes, but are informationally seen less meaningful.

	σ^2 actual scores	σ^2 estimate for others (avg.)	OPR level (from Table 7)
Set 1	1.099	.489	.610
Set 2	1.892	.543	1.349
Set 3	1.342	.609	.733
Sum (!)	**4.333**	**1.641**	**2.692**
Avg. (!)	**1.444**	**.547**	**.897**

Table 12: Components of precision calibrations

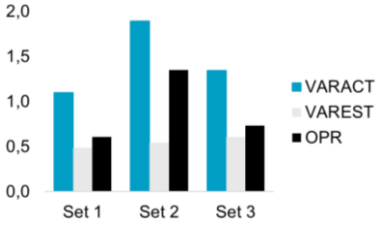


Figure 5: OPR comp. visualization

To perform a test on the statistical difference of overprecision between the sets, Repeated Measures ANOVA was again intended to be used. Nevertheless, this time, the assumptions were not met as the differences between the groups were not distributed approximately normally, and the sphericity assumption was violated ($p = .002$). Consequently, the Related-Samples Friedman's Two-way Analysis of Variance with Bonferroni correction has been used to assess rank differences instead of mean differences (Table 13). The mean rank of set 1 is 1.35, the mean rank of set 2 is 2.80, and the mean rank of set 3 is 1.85. The difference was obviously found to be highly significant for set 2 with the other sets ($p < .001$). The test, nevertheless, shows a highly significant p-value ($p = .001$) for a rank difference between set 1 and 3 as well. Hence, the sets show the erratic course of OPR, regardless of the task difficulty level.

Pairwise Comparisons

Comparison	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
OPR set 1-OPR set 3	-.505	.137	-3.691	<.001	.001
OPR set 1-OPR set 2	-1.458	.137	-10.664	<.001	.000
OPR set 3-OPR set 2	.953	.137	6.973	<.001	.000

Each row tests the null hypothesis that the distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .050.
a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

Table 13: Related-Samples Friedman's Two-way ANOVA of OPR

As the observed differences in OPR levels are relatively large, and no continuous patterns in set difficulty can be seen, it is intriguing to see where these differences come from. Table 12 shows that OPR differences between the sets mainly arise from a point of variance in the actual scores, which are found to be significantly different between the sets. The variance of the estimates of others increases steadily, but slowly. This presents some remarkable findings, since one would initially expect: (1) the variance in the actual scores of the more difficult sets close to similar, and (2) the variance in the estimates for others to move in line with the variance of the actual scores. Both expectations were thus not found in this realm. Regarding socio-demographic variables, the on forehand expected positive correlation between age and OPR (Prims & Moore, 2017) has not been found in any of the sets. However, it was found that OPR in set 3 differed significantly at a 10% level for gender ($p = .098$). Women (.858) expressed for this set significantly more OPR than men (.599). With a Cohens' d of .326, the effect size can be considered small to moderate. Considering the clearly non-significant difference in set 2, this does not seem to be linked to the difficulty level. Taking into account the significance level, this was more likely found by chance.

4.4 Click and time data related set differences

The number of clicks on the questions can be interpreted in two distinct ways. Primarily as a form of doubt, and secondly as a form of reconsideration and carefulness. The same goes for the time it takes to answer the knowledge questions, and especially in making the SPIES estimates. It is therefore intriguing to see whether differences arise between the sets, categorized on these variables. This paragraph will first look into the click and time data on knowledge questions (set level), after which extended with appliance of these variables to the SPIES questions. Lastly, the influence of breaks will be discussed. Hereby it follows the relevance considerations as discussed in paragraph 2.6, while assessing the methodological components of paragraph 3.3.4.

Click data in knowledge question sets

Examining along the sets, it is found that the click data is distributed right skewed for all sets after deleting 9 extreme outliers. The same goes for the total distribution. This is a logical distribution to find, as most respondents will change none or few answers in a set (e.g. 5-8 clicks), while there are also people who have more doubts during their consideration time, even when yet having selected an answer, and therewith place in the longer tale of the distribution (e.g. 9-12 clicks). Totalling below 5 clicks per set is only possible if one does not know the answer and/or runs out of time. Continuing with the initially selected "Don't know / Insufficient time" option, without selecting any other answer option meanwhile, leads to 0 clicks on that question. Clicks increased with age and were significantly higher for women.

The cut-off point for excluding outliers has been put at a maximum of 30 clicks in total (on average 2 clicks per question), which was met by the remaining 98 respondents. The extreme outliers are likely respondents that just clicked randomly through the answers while considering the right option. This is made very plausible considering their repeated outliers over the sets. As these cases were of such distinct nature (multiple respondents showed values far above this threshold, e.g. one person even 88 clicks), these could be identified and removed easily. This does, however, not mean one cannot express this behaviour incidentally, which is a weakness of this form of data. There is, nevertheless, no indication this has happened frequently, checked in combination with time data: there are few cases

with a high number of clicks in a short time span (an excessive click/time ratio). Additionally, in case of occurrence, this is expected to appear randomized due to question randomization.

Mutual correlation coefficients of clicks appear in a moderate range of .35 and .55. These correlations are the strongest between the more advanced sets. Although clicks were hypothesized to increase while sets get more difficult due to increased doubt and more reconsideration, an apparent finding is that this is not the case. Click data actually remains fairly constant across the sets ($\bar{x}$ set 1: 6.07, $\bar{x}$ set 2: 6.36, $\bar{x}$ set 3: 5.62).

Second thought, one might reason this to occur due to a larger amount of non-selection in the more advanced sets as a result of respondents running out of time or simply indicating that they did not know the answer. This could hypothetically balance a higher number of clicks in the other questions out. Where in the basic set respondents answered with the "Don't know / Insufficient time" option only 16 times (6 DK, 10 OoT) out of 535 answers (3.0%), this was found to be significantly higher in the advanced sets: 45 (8.4%) (43 DK, 2 OoT) and 51 (9.5%) (50 DK, 1 OoT) times. The relative sum of "Don't Know Responses" (DK) and "Out of Time Responses" (OoT) compared to the number of incorrect questions per set did not change regarding the questions' difficulty level (remained about 25%). Excluding cases with OoT or DK responses, but maintaining the within-subjects design, only 65 respondents could be retained for this sub-analysis. Comparing means again, it (obviously) shows higher levels for each of the sets, but the mutual relationship, with no clear distinction between the basic and more advanced sets, remains (set 1: 6.83, set 2: 7.44, set 3: 6.25). Therewith, it seems like, although respondents recognized the difficulty difference between the sets, they did not express distinct behaviour in the selection of their answers once they had selected an answer (Figure 6). The number of clicks did not correlate significantly with any of the OC forms, neither when looking at the absolute difference from perfect alignment.

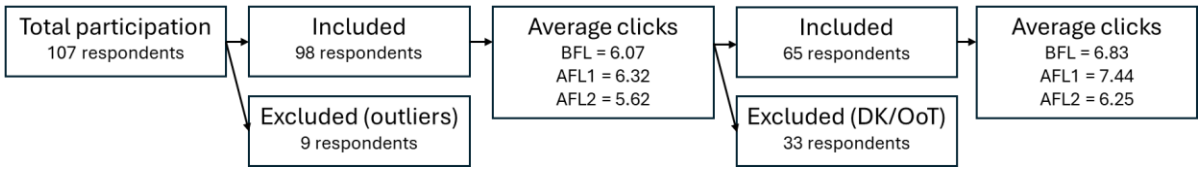


Figure 6: Analysis process on click data

Time data in knowledge question sets

Regarding taken time on answering the question sets, all distributions are right skewed as well. Few responses were found to express clear fatigue (really short time) or to give extreme outliers on the upside (really long time). What stands out is that the taken time to answer the knowledge questions is clearly higher for the basic set ($\bar{x}$ = 114 seconds, s = 42.58) than for the more advanced sets on average ($\bar{x}$ = 84 seconds, s = 32.93). The data shows that especially the calculation questions took more time for some respondents, while the more advanced questions were often more of a "I know this (/I think so), or not". This is in line with the higher number of OoT responses for the basic set, and the higher number of DK answers in the more advanced sets. The correlation between taken time and objective scores was found to be significant for the basic set (r = -.250**, p = .010), but non-significant for the more advanced sets, including for a consolidated comparison of set 2 and 3 to examine the AFL concept. Examining totals, there again turned out to be a weak negative correlation between the two (r = -.196*, p = .020). This anew emphasizes the importance of a difficulty-specific consideration, as it can lead to distinct results.

Time data of all sets correlate strongly with each other (correlation range .70 - .90), indicating that respondents were relatively seen consistent in their times of answering the knowledge questions. Being relatively fast in the one set often meant one was relatively fast in the other sets as well. The correlation between clicks and taken time is weakly positive in set 1 and 2 (r = .236*, p = .015; r = .248**, p = .010), but surprisingly non-significant for the last set (r = .128, p = .190). Concerning socio-demographic variables, significant negative correlations of taken time are found with education level, highest education finance affinity and daily finance affinity, while a positive correlation was found with age. These correlations remain for both BFL and AFL (set 2 + 3) and are also prevalent in the totals. No correlations were found between taken time and the OC forms (neither in the difference from 0, but OE set 2) in any of the sets. Having given some correlational results in general, further set comparison of time data has little added value as this depends on question length as well. Consequently, the next part will focus on time in SPIES, as these are questions of approximately equal length.

Time data (and click data) in SPIES estimates

The distribution of taken time in SPIES estimates does obviously not stand on its own either. For this distribution, it is highly relevant to what extent respondents variance their estimates and make changes, again captured in the number of clicks on this question. Therefore, time data is coupled with clicks for a proper comparison, expressed in the Time/Click (T/C) ratio. Categorizing time data on clicks, one would expect (logically, and as mostly observed in the FL questions) a significant positive correlation between the two variables. This correlation should expectedly be way stronger than for the knowledge questions, considering the direct independence of question difficulty and length in making these estimates. The data in Table 14 (time in seconds) shows that this is indeed the case, accounting for approximately 33% to 55% of the variance (correlation range .55 - .75). As it is considered important to maintain the within-subjects design, but meanwhile some significant outliers emerge in clicks and time, it was contemplated most intriguing to test the averages and patterns without these outliers. As this sub-analysis consists of 2 (SPIES) x 2 (T&C) x 3 (Sets) conditions, all containing a few extreme upside outliers, the number of cases for this sub-analysis had to be lowered to 89.

N = 89	Basic set				More advanced set 1				More advanced set 2			
	Avg. time	Avg. clicks	T/C ratio	r	Avg. time	Avg. clicks	T/C ratio	r	Avg. time	Avg. clicks	T/C ratio	r
SPIES1	14.77	5.74	2.57	.754***	16.31	7.00	2.33	.702***	16.00	6.53	2.45	.560***
SPIES2	13.39	6.37	2.10	.676***	15.13	6.56	2.31	.743***	15.56	7.22	2.16	.700***

Table 14: Average time and click data per SPIES estimate (excl. outliers), sorted on sets

The data shows that, although small differences can be observed, respondents generally seemed to take approximately equal time and put in an equal amount of effort to answer the SPIES1 and SPIES2 questions across all sets. The comparable behaviour is expressed in the T/C ratios as well. Some small, non-significant, differences can be observed when categorizing on the difficulty level, with the slightly lower values for the basic set. Differences in taken time and clicks between SPIES1 and SPIES2 in each of the sets are pretty much negligible. Furthermore, no significant differences in distributions (histograms) were detected for click and time data (and T/C) between SPIES1 and SPIES2. The latter is mainly important in the realm of the uncertainty of using SPIES2 as a measurement method, providing few to no indication of the expected fatigue or reconsideration for these more difficult estimates (see paragraph 4.5 for further elaboration). It seems like respondents tried their best on the SPIES2 estimates as well, despite the fact that they had less information about this.

Influence of breaks

Lastly, quite intriguing to check as well, it can be examined whether people who took (a) break(s) between the sets expressed distinct levels of OC. As explained, respondents were upfront informed and motivated (to a low extent) they could take a voluntary break before each of the sets. They were reminded of this during the survey. It is hypothesized that people who took this break were more likely to differentiate between the sets and start the sets more consciously, leading to better estimation and placement judgments as not confusing the sets with each other and remembering the questions. Although not many respondents indicated to have taken this break each time (10 respondents) or sometimes (1/2 times) (23 respondents) in the latter part of the survey (directly asked), time data has also been monitored. Breaks can thus objectively be checked, raising discussion on where to put the threshold. Herein, it was hypothesized to see somewhat of a dual-sided logarithmic or linear calibration curve towards good judgment (0) between break time and estimation/placement judgments. Roughly speaking, this would hypothesize converging misperception boundaries when taking a break. Considering the graphs of the 3 sets, this idea seems not to be too far off (see Appendix 5). It is nevertheless difficult to really draw conclusions on this phenomenon, as groups of people who took a break were relatively small: on average only about a quarter of the respondents when putting the threshold at 10 seconds. Therewith, findings might be a coincidence, instead of an actual emerging pattern that can be assigned to breaks (e.g. due to unequal FL distributions across the groups, of which converging patterns can be a side-effect in a certain composition). Besides, it might be that other characteristics or motivations play a role, underlying the motive to take breaks. It thus solely provides some initial support for the idea that people who take a break, certainly differentiating the sets, express better judgments. Future research, with more sampling options, might want to dive deeper into this.

4.5 Confidence in SPIES estimates, and the appearance of full-estimates

SPIES estimates are used in the measurement of both ones' own estimates and the estimates of others. As yet mentioned, past research has (indirectly) theorized that estimates for others are more difficult to make than estimates for oneself (Moore & Healy, 2008) and that one mainly centres the own performance while comparing with others (Kruger, 1999). The usage of this form of subjective assessment in both cases therewith comes with some uncertainty, especially since it is applied to a format people are generally unfamiliar with (Prims & Moore, 2017). Using the same measurement method at a distinct difficulty level (note: talking about the difficulty level of making the estimates, which is distinct from the difficulty level of the knowledge questions) might impact the based-on assumptions, one's behaviour, and consequently the judgmental outcomes. Considering that this uncertainty is methodologically and outcome-wise often not taken into account, this thesis tends to put some analysis on the examination of this inequality, especially regarding respondents' perception. Relevant questions here are: "Is this inequality being perceived as such?" and "How do respondents behave in response to this uncertainty?".

The appliance of the expected unequal perception has yet initially been looked into regarding click and time data as described in the previous paragraph, which on its own actually found quite comparable behaviour, without traces of fatigue. This will thus be extended here. Considering the higher difficulty for SPIES2, one might nevertheless reason to expect to see some misunderstanding or fatigue. For the analysis in this paragraph, respondents have been asked directly about the perception of their made estimates. Confidence has in both cases, own estimates (CONFSPIES1) and estimates for others (CONFSPIES2), been

measured using a 5-point Likert Item. These items were asked directly after each other in order to obtain a valuable comparison by respondents. Confidence levels in SPIES can be considered a measurement method for individual precision perception, containing the overarching uncertainty of multiple question sets in both SPIES1 and SPIES2. The independent measures, the split into two groups as described below, and the fact that misunderstanding regarding the measurement format can be indicated, give additional insights compared to earlier studies by introducing a way of difficulty-related comparison. CONFSPIES1 and CONFSPIES2 were found to correlate significantly with each other.

Differences in confidence afterwards (dummy composition)

In total, 93.5% of respondents took a neutral or better position in their CONFSPIES1 estimates, compared to 85.0% who took this position for the estimates they made for others (CONFSPIES2). This shows that respondents expressed little direct aversion to the way the estimates could be made. Based on these adjacent percentages, one might initially derive that differences in confidence for the SPIES estimates are not as substantial and unilateral as one would expect beforehand. Nonetheless, taking a look at the distributions and mutual relationships of the 3 higher categories (neutral+), it is apparent that the dispersion among the groups has been relatively diverse, opposing each other in multiple ways. Initiating an examination among the gathering in two highest groups ("agree" and "totally agree"), percentages decline to respectively 65.4% (CONFSPIES1) and 32.7% (CONFSPIES2). Spearman's correlation between both variables is .338***. The distribution of absolute differences between the items can be seen in Figure 7.

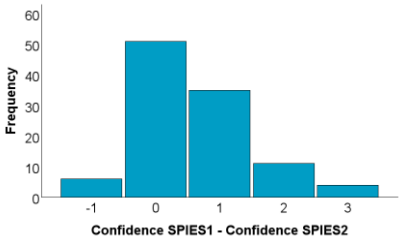


Figure 7: Confidence difference between SPIES1 and SPIES2 estimates

As the differences between both variables, and the actual variables, are presumed not to be distributed normally by the Shapiro-Wilk test and the QQ-plots, a within-subject design with dependent samples is applied, and the confidence is measured using a single Likert Item (considered to be ordinal data), the Wilcoxon Signed Rank test has been used to test for differential significance of the median from zero (McDonald, 2014). The median of the differences in CONFSPIES1 and CONFSPIES2 was found to differ highly significantly from 0 ($p < .001$). This result is nevertheless considered to be somewhat controversial, given the asymmetry of the histogram (skewness = .75). This asymmetric distribution appears due to the large number of ties that have arisen, combined with the small ordinal range (5-point Likert items). Looking at the histogram, differences mainly appear to be positive. Although this histogram is quite self-explanatory about with which sign differences mainly occur, a Sign test has been performed as additional substantiation to get certainty on upside significance. It shows a highly significant result as well ($z = 5.746$, $p < .001$). Considering mean values, CONFSPIES1 shows a mean value of 3.77, while CONFSPIES2 has a mean value of 3.18. Support was thus found for the expectation that respondents generally had more confidence in the estimates they made for themselves, than in the estimates they made for others. The large number of ties nevertheless remains an important sidenote in the realm of one's answering of (CONF)SPIES questions and the total perception after multiple sets.

As it is thus considered the 'normal' situation to have more confidence in the estimates for oneself than in the estimates for others, these groups can logically be distinguished and tested to differ. The dispersion into a dummy variable gives 2 independent groups for these

tests: (1) Confidence SPIES1 > Confidence SPIES2 (n = 50; expectation / control group), and (2) Confidence SPIES1 <= Confidence SPIES2 (n = 57). It should be kept in mind that the statistical power decreases due to this split. Primarily, it is intriguing to examine whether people who told to express higher or equal confidence in their estimates for others show significantly higher levels of estimation and placement misjudgements than people who stated to have higher confidence in their own estimates (expected). Solely total OE and OP levels, and not those of separate sets, have been considered in these tests, as CONFSPIES levels were asked for in their totalities as well. OPR has not been taken into account, considering that the sum of the individual variances does not equal the variance of the total.

SPIES confidence dummies and OE

Total OE is normally distributed in both categories according to both the Shapiro-Wilk test and the QQ-plots. Looking at Pearsons' correlation coefficient, it appears that the difference in SPIES confidence (either as a dummy as in absolute levels) significantly and positively correlates with the total level of OE. Homogeneity of variances does not appear between the groups, and no significant outliers were found. Therefore, the Welch t-test has been used to determine whether there are significant differences in this sample. This test (Table 15) finds support for a significant difference of estimation expression between the 2 groups (t = -2.359, p = .020). Glass's delta, selected as effect size measure due to the relatively large difference in standard deviation (2.64 v. 1.61), gives a moderate point estimate of .610.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					
						Significance		Mean	Std. Error
		F	Sig.	t	df	One-Sided p	Two-Sided p	Difference	Difference
OE	Equal variances assumed	7.373	.008	-2.289	105	.012	.024	-.983754	.429837
	Equal variances not assumed			-2.359	94.341	.010	.020	-.983754	.417099

Table 15: Welch t-test of OE among dummy categorized SPIES-confidence dispersions

This sample thus finds a significantly higher estimation misperception for respondents who declared to express equal or higher confidence in the estimates for others than in the estimates for themselves. As underestimation of performance was found in both groups, it manifests itself in the form of almost a full point of UE (mean values of -1.32 v. -2.30). This is especially interesting considering the fact that FL levels of these groups differ significantly as well, with approximately 12 FL points for the control/expectation group and 10 for the other. This difference continues in the SPIES1 (10.73 v. 7.72) and SPIES2 (9.21 v. 7.83) estimates. From that point of view, one would expect to see more UE for the control group as they perceived the sets as easier; the opposite is the case. The distribution of the individual sets might herein be of importance, which is for future research to explore.

SPIES confidence dummies and OP

Total OP is approximately normally distributed in both categories as well. Primarily looking at Pearsons' correlation coefficient, it appears that differences in confidence in one's estimates do not correlate significantly with total OP. Homogeneity of variance is anew not present between the two groups, while simultaneously no significant outliers were observed.

Consequently, the Welch t-test has again been used to determine whether there are significant differences between the selected groups according to this sample. This test (Table 16) finds no support for a significant difference of OP means (t = .787, p = .433).

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					
						Significance		Mean	Std. Error
		F	Sig.	t	df	One-Sided p	Two-Sided p	Difference	Difference
OP	Equal variances assumed	15.618	<.001	.761	105	.224	.448	.389404	.511745
	Equal variances not assumed			.787	90.791	.217	.433	.389404	.494749

Table 16: Welch t-test of OP among dummy categorized SPIES-confidence dispersions

Hence, the data of this sample shows no support for the idea that people with a higher or equal confidence in the estimates for others than the estimates for themselves exhibit significantly higher levels of placement misperceptions in any form (OP or UP). Considering the higher FL levels, SPIES estimates, and deviating OE levels for the control group as described above, this non-significance can however still be considered remarkable (value of .44 for control group, and .83 for others). Just like in OE, the role of the distribution of the individual sets should be pointed out as potentially influential.

SPIES confidence dummies and dispersions in the way of answering

As estimation misjudgements differed significantly between the CONFSPIES groups with an unexpected sign, it is interesting to test whether the groups show distinct patterns in how the SPIES estimates were filled in, especially focussing on SPIES1 as estimation measure, but also looking at SPIES2 in the realm of its expected higher uncertainty. For instance, respondents might have filled in whole numbers (full-estimates) more often, instead of percentages. On the one hand this might simply indicate full-confidence (most likely the basic set), on the other hand this might indicate indifference or misunderstanding regarding SPIES measurement. Although large differences are unexpected to be found, following the equal and relatively high amounts of click and time data together with the equal distributions as described in paragraph 4.4, it is good to take a look at this from another point of view. Also in combination with the CONFSPIES measures.

Group Statistics

	Dummy SPIES-confidence groups	N	Mean	Std. Deviation	Std. Error Mean
FullSPIES1BFL	0	57	.491	.504	.067
	1	50	.440	.501	.071
FullSPIES2BFL	0	57	.421	.498	.066
	1	50	.300	.463	.065
FullSPIES1AFL1	0	57	.456	.503	.067
	1	50	.320	.471	.067
FullSPIES2AFL1	0	57	.439	.501	.066
	1	50	.300	.463	.065
FullSPIES1AFL2	0	57	.439	.501	.066
	1	50	.340	.479	.068
FullSPIES2AFL2	0	57	.421	.498	.066
	1	50	.260	.443	.063

Table 17: Mean comparison on respondents' variance dispersion (dummy-based)

This has been done by exploring 6 created dummy variables. These dummies are based on whether respondents spread their chances using decimal estimates (0) or filled in full-estimates in their SPIES (1). In Table 17, these are expressed as a mean value and compared to the confidence difference as indicated, using the created groups dividing CONFSPIES1 > CONFSPIES2 (1) and rest (0). It turns out that the latter group answered

more often with full-estimates in all sets. Nevertheless, this difference has in none of the cases found to be significant (Table 18). Especially since Bonferroni adjustment should be applied, leading to an even stricter threshold. Differences are yet found to be non-significant.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					
						Significance			
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference
FullSPIES1BFL	Equal var. assumed	.764	.384	.526	105	.300	.600	.051	.097
	Equal var. not assumed			.526	103.345	.300	.600	.051	.097
FullSPIES2BFL	Equal var. assumed	6.295	.014	1.296	105	.099	.198	.121	.093
	Equal var. not assumed			1.302	104.636	.098	.196	.121	.093
FullSPIES1AFL1	Equal var. assumed	6.841	.010	1.439	105	.077	.153	.136	.095
	Equal var. not assumed			1.445	104.516	.076	.151	.136	.094
FullSPIES2AFL1	Equal var. assumed	7.762	.006	1.480	105	.071	.142	.139	.094
	Equal var. not assumed			1.487	104.695	.070	.140	.139	.093
FullSPIES1AFL2	Equal var. assumed	3.918	.050	1.038	105	.151	.302	.099	.095
	Equal var. not assumed			1.041	104.208	.150	.300	.099	.095
FullSPIES2AFL2	Equal var. assumed	11.518	<.001	1.756	105	.041	.082	.161	.092
	Equal var. not assumed			1.770	104.976	.040	.080	.161	.091

Table 18: t-tests on significant differences in full-estimates per SPIES per set

Following Table 17, it should be kept in mind full-estimates were a common phenomenon, with between 34% and 47% of the estimates per SPIES. They correlate moderately to strongly with each other (range: .52 - .82), where the correlations between the SPIES2 estimates are especially striking (all >.7). Full-estimates of the BFL set correlate significantly with the indicated SPIES-confidence for both SPIES1 ($r = .214^*$, $p = .027$) and SPIES2 ($r = .275^{**}$, $p = .004$). This was, however, not found for the other sets. Checking the dispersions in the way of answering related to actual OC levels, it gives some additional intriguing results. Estimation misjudgements are significantly less for the group that estimated using a full-estimate in both the first ($t = -3.924$, $p < .001$) and third set ($t = -2.401$, $p = .018$). The second set shows a non-significant difference ($t = -.631$, $p = .529$) (nevertheless still in favour of the full-estimates). The effect is most prevalent in the basic FL set, as it gives a Cohen's d effect size of .760, indicating a moderate to large effect. For set 3, it gives a moderate effect size of .416. While it should be interpreted as a mainly initial comparison, as the results can also occur due to chance or distinct average FL levels, it seems like people knew what they were doing when making full-estimates for themselves. For SPIES2 estimates, the accuracy between these groups is checked by comparing the (mean) SPIES2 values of the groups with the actual FL average. In the first and third set, the full-estimates are again significantly more accurate ($t = -2.350$, $p = .021$ and $t = -3.054$, $p = .003$). Anew, the second set appears to be non-significant ($t = -.466$, $p = .665$), but in favour of full-estimates. This again provides an indication one fills in full-estimates from a view of actual conviction; one might doubt how. Being more cautious: no support for the excessive use of full-estimates for fatigue was found.

5 Discussion and conclusion

5.1 Elaboration of main results

This thesis has looked into aspects of the overarching research question: “To what extent do contextual factors, such as the difficulty level of the questions, respondents’ socio-demographic variables, click and time data, and styles of answering the SPIES questions, contribute to understanding the 3 forms of OC in FL and one’s behaviour in answering the knowledge questions and making the estimates?” As described in the introduction, four sub-questions have been formulated to cover parts of this extensive question.

Considering the observed misperceptions and miscalibrations in the sets, some conclusions about the appearance of task difficulty patterns in the overconfidence forms can be drawn in this FL context, applying it to the realm of this exact methodology. This methodological aspect is highly important in answering sub-question 1, as mostly following (the quiz stage and interim stage, with related OC calculations, of) Moore and Healy (2008) for this analysis. It has been observed that the patterns are in line with their expectations and findings in earlier contexts; UE and OP for the easier set, moving towards OE and UP for the more difficult sets. Full elaboration could nevertheless not be performed, considering the limitation of question difficulty perceptions by respondents, leading to more of a moderate-easy examination. Although patterns mainly align the expected patterns as described in paragraph 2.4, no statements can be made on whether the observed confidence misperceptions are actual misperceptions of one’s financial literacy (in general), or whether these occur due to/are influenced by the set divisions and statistical artifacts. It might even be the case that all aspects play a role. This is the tension that has been discussed multiple times, but which cannot simply be avoided when applying this set-focused methodology. It is for a reason that Moore and Healy (2008) have yet theorized the effects to occur regressive in general due to informational limitations. Getting aware of this, the implications of the first part of the analysis should reserve judgement on statements regarding OC in FL, apart from the comparable patterns in this methodology (for which hence no support was found that they belong to, or deviate from the general expectations in, this specific context). Regarding sub-question 2, only few socio-demographic correlations with the overconfidence forms have been observed. The intended examination of differences was difficult due to the sample composition, which led to unequal and small groups. A fair comparison could only be made for gender.

The second part of the analysis has been more innovative and can be more expressive in its findings. Sample size, composition, and measurement limitations should nevertheless be noted and considered in the meantime. Analyzing click and time data on all individual questions (summed: question sets) and SPIES estimates, various insights into respondents’ behaviour were gathered. Expected higher levels of click data, due to increased doubt and uncertainty, in the more difficult sets have not been observed. Neither when excluding OoT and DK responses. Regarding time data on the knowledge questions, an important finding is that mutual set time is highly correlated in a .7 to .9 range, and that the correlation with click data is apparent, but not as strong for the SPIES estimates. Click and time data in the knowledge questions were both uncorrelated to any OC form. Quite even levels of click and time data were found for the SPIES across the sets. No significant differences in the distributions of these, and the hereto related (T/C ratio), variables was found. This gives the indication that the SPIES2 estimates were generally not considered much different than SPIES1 estimates (but answered seriously), although often indeed indicated by respondents

to be more difficult to make (with the lower CONFSPIES2). Due to the higher estimation difficulty in an unfamiliar context, one might expect more uncertainty and distinct behaviour, e.g. in forms of reconsideration or fatigue/rush. The latter has additionally been checked in the number of full-estimates that were made, being the fastest estimation option. Full-estimates remained relatively constant across the sets. It was found that the appearance of full-estimates was not significantly distinct between SPIES1 and SPIES2, which can be called remarkable. A significant difference was only the case for the basic set, in which selected more often for SPIES1. This is most likely because of actual certainty. Full-estimates in SPIES1 and SPIES2 were, however, found to be better in their capacity of reality alignment than the varianced ones, which is especially thought-provoking for SPIES2.

Furthermore, in the CONFSPIES comparison, in which a higher CONFSPIES1 is considered to be the baseline, it was found that full-estimates repeatedly differed in the same direction (less often made by the control group). Zooming in on these one-sided differences, they nevertheless appeared to be non-significant. Estimation perception alignment was found to be significantly better for the baseline-group, which is remarkable considering the related FL levels and perceptions, combined with the expected task difficulty and ability patterns. For OP, a non-significant difference appeared. The composition of the individual sets could however not be considered in these analyses. Lastly, the data expresses a pattern of converging estimation and placement misperceptions towards accurate judgment when respondents take a break between the sets. The power of this latter part is nevertheless diminished by the unequal distribution of cases (75%-25%), making the findings more sensitive to unequal distributions of FL levels.

As the analyses in the second part of the analysis have mainly been focused on an initial exploration of several behavioural variables and measurements in this context, the findings present some statements on parts of the overarching question and the therefrom derived sub-questions (3 and 4). These are interesting, but mostly insufficient to formulate a real answer to the questions. Optimization and further elaboration in the future is recommended.

5.2 Theoretical and practical implications

Primarily, although not considered to be the main focus of the project, this thesis provides some support and contradictories on OFL levels and their correlations with the selected socio-demographic variables. As this data collection was necessary to be able to add to calibration and behavioural research as described below, this has led to a for this thesis mostly unintentional, but useful check of these variables/levels, in which the results could be compared with the findings of past research (see paragraph 4.2). Therewith, it adds to findings of e.g. Karaa and Kuğu (2016), Lusardi (2008), Lusardi and Mitchell (2011a, 2011b), and Van Rooij et al. (2011) on financial literacy and socio-demographic dispersions, by reperforming analysis on the BFL and AFL question sets as retrieved from van Rooij et al. (2011). This chiefly provided further support for the findings of the earlier studies on this theme, but also showed some small contradictories. Furthermore, it adds by extending research on the financial literacy topic with the consideration of click and time data on these questions, which to the best of knowledge had not been performed earlier. These aspects provide information on how respondents answered the questions across the sets and are therewith of behavioural knowledge importance.

By examining the 3 forms of overconfidence in multiple categories of FL difficulty levels, this thesis has taken action to extend past studies that combine the OC forms with FL themes, as had been suggested by Vörös et al. (2021). It has slightly added to them by directly exploring patterns of the task difficulty level in this specific context (as this effect had been indicated interesting by Hamurcu and Hamurcu (2020) as well, but remained unexplored by them either). This part of the analysis has not only been relevant for OC insights, but also for the awareness of the often-used FL measurement method (limitations), together with these OC forms. One should carefully consider the difficulty level of the FL queries, especially when the OC level is determined on set dispersions and related to other variables (e.g. financial decision-making). It might be that the observed OC is not specifically FL-related.

This thesis differentiates (like Moore and Healy (2008) and Vörös et al. (2021)) from study 1 of Prims and Moore (2017) by applying the methodological approach to a context of answering (financial literacy) question sets with knowledgably correct answers one is reasonably likely to know. Applying it to a MC test-setting, the measurement of SPIES is put to a more recognizable context for most respondents, as basically everyone has ever experienced the doubt of correctness of MC question(s) (sets) while answering a test. This recognizable context is of utter importance, as respondents are consequently familiar with the fact that one has a reasonable possibility to (have the feeling to) be entirely correct on an answer option. This is distinct from Prims and Moore (2017), where this possibility, in a context of weight guessing, is considered way smaller since answers have to be guessed from a picture. In their study, points of correctness were gained when estimating within a range of 40 (easy) or 3 (hard) pounds from the actual weight. Consequently, the chance of being entirely correct was merely theoretical, and, more importantly, correct answers contained way higher uncertainty as the result could simply not be known. The possibility of being correct ánd knowing for sure was therewith way smaller. The within-subjects design minimizes variability between respondents (0) and consequently contains less measurement noise due to respondent' characteristics (Charness et al., 2012). This facilitates a broader view on the subject. Besides, it should again be noted that the methodology for this part of the analysis is mostly a contextually distinct replication of the quiz stage and the interim stage from Moore and Healy's 2008-article, of which the design has been adopted, but that some distinctions emerged due to this specific contextual difference. This has been used to check whether the expectations hold under the conditions as described in paragraph 3.3.3, on 3 of the by them indicated considerations: domain, population, and context (Netherlands, 18+, FL, 5-question sets, 2 equally perceived more difficult sets, no rewards).

This thesis found support that this is indeed the case. It finds the expressed patterns across sets to be mainly in line with the expected task difficulty patterns. Appliance to, and testing it in this specific setting with BFL and AFL (or, more advanced sets) has provided more insights on how and whether this effect manifests in this context. It does not show any unexpected deviations. As in many situations questions are not extremely easy or extremely difficult, an examination of the difficulty effects in these question sets was not supposed to find the exact patterns of the expectation; there is also something in-between. The results show this trade-off to be the case for the more advanced FL questions, as both correctness volume and OC patterns express as such. The set dispersion, with the two more difficult sets, has also facilitated to look into specific OPR differences at approximately the same difficulty level. This is also new. The calculational components showed that the outcomes of this sample mainly depended on the variance of the actual scores, and that these (and thus the OPR levels) can be highly distinct; even between 2 sets of almost equal difficulty.

By analysing the calculational components of OC in all forms, it has also looked more into relative patterns and proportions. Furthermore, this thesis adds to existing research with an initial exploration of other research design and measurement limitations, like the uncertainty of SPIES estimates (especially for others) and one's related behaviour. Besides, it considers earlier described effects of formational decisions like question number dependency, possible cheating in knowledge questions, and the likeliness of the occurrence of scale-end effects.

The behavioural analyses have extended data on the SPIES estimates with an examination of click and time data differences, compared perceived confidence in the SPIES estimates, and the appearances of full-estimates. Especially here, the difference with Moore and Healy (2008) regarding incentives provision is considered to be potentially influential. The fact that few behavioural differences were found between SPIES1 and SPIES2 in click data, time data, and the distributions of these variables indicates that respondents took SPIES2 estimations generally quite seriously, but that one did not express very distinct behaviour due to the increased SPIES2 uncertainty. Besides, it is worth noting that full-estimates occurred relatively equal between SPIES1 and SPIES2. They generally showed better alignment than split-estimates, also for SPIES2. The high number of full-estimates for SPIES2 (despite showing good results) and their strong mutual correlations across the sets, the opposed OE and OP levels compared to general task difficulty expectations in the CONFSPIES dummy exploration, and the existence of the CONFSPIES dummy in general nevertheless remain somewhat remarkable and seem to reveal some kind of misunderstanding or fatigue.

Regarding the behavioural part on SPIES confidence (CONFSPIES): this measure looks at precision perception, instead of actual precision. Therewith, it adds to the OPR measurement of many of the aforementioned calibration studies in this paragraph, as these mainly looked at the consequences of the by them assumed actual precision calibration, simultaneously assuming understanding of measurement by respondents. CONFSPIES can basically be seen as applying the knowledge question certainty estimates (see paragraph 2.3.2 for a short overview), but translated into a Likert item, to the made estimate. Hence, the confidence in the estimates for oneself and the confidence in the estimates for others are both assessed, while indirectly being compared with each other. CONFSPIES measurement therewith also creates a space for respondents to actually indicate uncertainty or misunderstanding regarding these questions, clearly extending insights. By providing the same transparent situation to all sets, a fair comparison could be made. This does, however, not mean that all limitations on the topic have been remedied and that there is no room for improvement in its measurement. This will be discussed in the next paragraph.

5.3 Limitations and suggestions for future research

This thesis has one main limitation on its contribution regarding OC levels across FL sets of distinct difficulty. It remains unclear whether levels of OC appear due to cognitive bias, psychological aspects, or due to statistical effects and ways of stating the questions (set dispersion), and whether these are (somewhat) topic-related. For the time being, it seems like such set dispersion, regardless of topic, produces the general levels/effects on the 2 OC forms (OE and OP), as yet mostly theorized in Moore and Healy (2008). Getting to know this, the observed patterns across the sets are not of a very surprising character. Regarding topic dependency, the difference between judgmental and decisional confidence miscalibrations is herein highly important, alike the methodology used. By relating SFL to OFL in a question set, and not to general ability perceptions, one's derived perception does not necessarily

have to correspond with the general perception of financial knowledge and abilities; especially when considering that OFL is measured using “noisy proxies” (Van Rooij et al., 2011, p. 454). This makes the translation of the calibration to the impact on decision-making difficult. Considering this dependency of the OC forms on the (perceived) task difficulty level (especially OE and OP), future research might want to look further into people that are at the same ability level, and whether/why they show distinct OC levels. This is distinct from the D-K effect as it can look into individual components and estimates, instead of making a comparison with the entire ability-range. Besides, it might want to elaborate further on the calculational components of the overconfidence forms.

Another limitation of this thesis, related to the applied research design, is the adoption of data from a relatively small convenience sample. Due to this non-probability form of sampling, combined with the moderate sample size, generalizability of the results is limited as individual differences might play more of a role (Andrade, 2021; Jager et al., 2017). Just like the self-selection bias, that might have had its impact on the results’ generalizability as well. Nevertheless, considering the for a student available resources, no better alternatives were identified to perform the project in an effective and timely manner. As described, the sample limitations have put a limit on the extent socio-demographic differences could be examined. Future research might want to take these sampling limitations into account by preferably changing the approach to a form of random sampling, and by getting the insights of a larger group of respondents.

A third limitation relates to the nature of the FL questions, in combination with the survey design. As knowledge questions form part of the data collection, the online survey setting makes it easier for respondents to look up answers to these questions, which has been considered more concerning following the rise of especially AI-tools in the last few years. Multiple measures have been taken to prevent from this behavior, mainly based on research concerned reducing reported cheating rates. Actual cheating rates might, however, respond differently and exhibit less effective outcomes at worst. The results would give an even more trustworthy reflection of reality in a more transparent setting, where the environment can be controlled. This should ideally be measured on the same device type as well (nevertheless, 90% of the respondents indicated to have filled in the survey on their smartphone; large differences are not expected here). A more transparent setting would additionally lead to a situation in which the question answer option of OoT/DK could be removed, as this option was included as one of the anti-cheating measures. Despite offering some interesting insights, removal of this option would lead to less disturbance of the task difficulty perception considering the fact that respondents could choose for this option in case of really high uncertainty, bowing it into full certainty (that one is wrong on the specific question). Handling both limitations simultaneously was not possible considering the available resources.

Additionally, the questions of the survey that are designated as financial literacy are mainly limited to investment, risk, savings and calculation questions. There are, however, more subjects that can be assigned to the header of FL. It was chosen to follow the specification approach as applied in previous research, since a more extensive assessment would lead to an even longer survey (12+ minutes is already quite long for a student survey). It is, however, important to acknowledge that an assessment of additional financial themes might lead to distinct results. Topics and question settings might be of influence in calibration capacities. OC levels might also differ based on the number of questions in a set. Consequently, AFL questions were split into two separate sets, leading to a situation in which these sets could

not be referred to as AFL due to the limited internal reliability of the sets apart from each other. Future research might want to focus on developing specific criteria and metrics to have both equal set size and actual constructs on both sides of the comparison.

Regarding measurement, the usage of SPIES in both estimates for self and estimates for others is definitely considered a limitation, but one that has been adopted consciously. The initial exploration of SPIES-confidence and behaviour is a first step in indicating difficulty and behavioural differences, although future research might want to focus on the broader applicability of findings. For instance, future research might want to examine whether findings hold when including a more elaborate measurement of this confidence, since this is currently measured using a single Likert item. In the ideal situation, this measurement would be extended to increase validity, and the questions would be applied at the timepoint of the SPIES questions after each set, instead of asking only once at the end. In that case, it might also be interesting to calculate an actual OPR value out of this, instead of only looking at the perception, and one might want to check its correlation with the currently adapted OPR measure. This set-specific appliance would also provide a solution to the potentially influential distribution of the individual sets in the OE and OP analyses, as CONFSPIES and estimate accuracy (for both SPIES1 and SPIES2) can be bundled for each set.

Regarding click and time data in the knowledge questions, question length plays a role as well, as this obviously differs across the sets. The difficulty of handling this data can also be seen in the number of outliers for each of the cases. Future research might want to make a clearer distinction between doubt and reconsideration to be able to say more about these variables. Besides, it might want to look at click dynamics at an individual level, following sequential behaviour in case of apparent uncertainty. Furthermore, it might want to take more measures to keep respondents surely focussed on the questions and estimates, and provide the equal circumstances as described. Although respondents' focus was attempted to be stimulated by including breaks and time-limits, respondents might still have gotten distracted during the sets, especially since data collection occurred online (Clifford & Jerit, 2016). Additionally, it might want to look further into these breaks, as the number of people that took these voluntary breaks was limited. This could for example be done by forming an experimental and control group in a larger and random sample, in which one of the groups has to take a mandatory break. Hereby, performance levels are considered to be important.

References

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences*, *33*(4), 587-605. https://doi.org/10.1016/S0191-8869(01)00174-X

Alba, J. W., & Hutchinson, J. W. (2000). Knowledge calibration: What consumers know and what they think they know. *Journal of Consumer Research*, *27*(2), 123-156. https://doi.org/10.1086/314317

Alicke, M. D., & Govorun, O. (2005). The Better-Than-Average Effect. In *The Self in Social Judgment.* (pp. 85-106). Psychology Press.

Anderson, A., Baker, F., & Robinson, D. T. (2017). Precautionary savings, retirement planning and misperceptions of financial literacy. *Journal of Financial Economics*, *126*(2), 383-398. https://doi.org/10.1016/j.jfineco.2017.07.008

Andrade, C. (2021). The Inconvenient Truth About Convenience and Purposive Samples. *Indian J Psychol Med*, *43*(1), 86-88. https://doi.org/10.1177/0253717620977000

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic Physiol Opt*, *34*(5), 502-508. https://doi.org/10.1111/opo.12131

Balasubramnian, B., & Sargent, C. (2020). Impact of Inflated Perceptions of Financial Literacy on Financial Decision Making. *Journal of Economic Psychology*, *80*, 102306. https://doi.org/10.1016/j.joep.2020.102306

Barber, B. M., & Odean, T. (2001). Boys will be Boys: Gender, Overconfidence, and Common Stock Investment. *The Quarterly Journal of Economics*, *116*(1), 261-292. https://doi.org/10.1162/003355301556400

Blanca, M. J., Arnau, J., García-Castro, F. J., Alarcón, R., & Bono, R. (2023). Non-normal Data in Repeated Measures ANOVA: Impact on Type I Error and Power. *Psicothema*, *35*(1), 21-29. https://doi.org/10.7334/psicothema2022.292

Brown, J. D. (2012). Understanding the Better Than Average Effect:Motives (Still) Matter. *Personality and Social Psychology Bulletin*, *38*(2), 209-219. https://doi.org/10.1177/0146167211432763

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, *109*. https://doi.org/10.1016/j.jml.2019.104047

Brysbaert, M., Sui, L., Duyck, W., & Dirix, N. (2021). Improving reading rate prediction with word length information: Evidence from Dutch. *Q J Exp Psychol (Hove)*, *74*(11), 2013-2018. https://doi.org/10.1177/17470218211017100

Bucher-Koenen, T., Alessie, R., Lusardi, A., & Rooij, M. (2021). Fearless Woman: Financial Literacy and Stock Market Participation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3805715

Bucher-Koenen, T., Lusardi, A., Alessie, R. O. B., & Van Rooij, M. (2017). How Financially Literate Are Women? An Overview and New Insights. *The Journal of Consumer Affairs*, *51*(2), 255-283. https://doi.org/10.1111/joca.12121

Centraal Bureau voor de Statistiek. (2024a). *Bevolking op 1 januari en gemiddeld; geslacht, leeftijd en regio* [Data set]. CBS. Retrieved on September 5, 2024, from https://opendata.cbs.nl/statline/CBS/nl/dataset/03759ned/table?fromstatweb

Centraal Bureau voor de Statistiek. (2024b). *Bevolking; hoogstbehaald onderwijsniveau en onderwijsrichting* [Data set]. CBS. Retrieved on September 5, 2024, from https://opendata.cbs.nl/statline/CBS/nl/dataset/85313NED/table?ts=1720175968004

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1-8. https://doi.org/10.1016/j.jebo.2011.08.009

Chui, A. C. W., Titman, S., & Wei, K. C. J. (2010). Individualism and momentum around the world [Article]. *Journal of Finance*, *65*(1), 361-392. https://doi.org/10.1111/j.1540-6261.2009.01532.x
Clifford, S., & Jerit, J. (2016). Cheating on Political Knowledge Questions in Online Surveys: An Assessment of the Problem and Solutions. *Public Opinion Quarterly*, *80*(4), 858-887. https://doi.org/10.1093/poq/nfw030
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. http://site.ebrary.com/id/10713862
Czaja, S., & Sharit, J. (1998). Age Differences in Attitudes Toward Computers. *The journals of gerontology. Series B, Psychological sciences and social sciences*, *53*, P329-340. https://doi.org/10.1093/geronb/53B.5.P329
Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology, Vol 44.* (pp. 247-296). Academic Press. https://doi.org/10.1016/B978-0-12-385522-0.00005-6
Dunning, D. (2022, March 7th). *The Dunning-Kruger effect and its discontents*. The British Psychological Society. https://www.bps.org.uk/psychologist/dunning-kruger-effect-and-its-discontents
Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology*, *57*(6), 1082-1090. https://doi.org/10.1037/0022-3514.57.6.1082
Dutch Civil Code, Book 1. (1958, last amended 2024). Title 13 - Minderjarigheid, Arts. 1:233 & 1:234. Overheid.nl. https://wetten.overheid.nl/BWBR0002656/
Dutch Civil Code, Book 3. (1980, last amended 2024). Title 2 - Rechtshandelingen, Art. 3:32. Overheid.nl. https://wetten.overheid.nl/BWBR0005291/
Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519-527. https://doi.org/10.1037/0033-295X.101.3.519
Evans, J. S. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Psychology/Erlbaum (Uk) Taylor & Fr.
Fernandes, D., Lynch, J., & Netemeyer, R. (2014). Financial Literacy, Financial Education, and Downstream Financial Behaviors. *Management Science*. https://doi.org/10.1287/mnsc.2013.1849
Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117-140. https://doi.org/10.1177/001872675400700202
Financial Industry Regulatory Authority. (2007). *NASD and NYSE Member Regulation Combine to Form the Financial Industry Regulatory Authority*. FINRA. https://www.finra.org/media-center/news-releases/2007/nasd-and-nyse-member-regulation-combine-form-financial-industry
Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), 552-564. https://doi.org/10.1037/0096-1523.3.4.552
Flynn, L. R., & Goldsmith, R. E. (1999). A Short, Reliable Measure of Subjective Knowledge. *Journal of Business Research*, *46*(1), 57-66. https://doi.org/10.1016/S0148-2963(98)00057-5
Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25-42. https://doi.org/10.1257/089533005775196732
Gigerenzer, G. (1991). How to Make Cognitive Illusions Disappear: Beyond “Heuristics and Biases”. *European Review of Social Psychology*, *2*(1), 83-115. https://doi.org/10.1080/14792779143000033

Gignac, G. E. (2022). The association between objective and subjective financial literacy: Failure to observe the Dunning-Kruger effect. *Personality and Individual Differences*, *184*, 111224. https://doi.org/10.1016/j.paid.2021.111224
Glaser, M., Langer, T., & Weber, M. (2005). Overconfidence of Professionals and Lay Men: Individual Differences Within and Between Tasks?
Glaser, M., & Weber, M. (2007). Overconfidence and trading volume. *The Geneva Risk and Insurance Review*, *32*(1), 1-36. https://doi.org/10.1007/s10713-007-0003-3
Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411-435. https://doi.org/10.1016/0010-0285(92)90013-R
Guiso, L., & Jappelli, T. (2008). Financial literacy and portfolio diversification. *European University Institute (Working Papers No. 31)*.
Hadar, L., Sood, S., & Fox, C. (2013). Subjective Knowledge in Consumer Financial Decisions. *Journal of Marketing Research*, *50*, 303-316. https://doi.org/10.1509/jmr.10.0518
Hamurcu, C., & Hamurcu, H. (2020). Can financial literacy overconfidence be predicted by narcissistic tendencies? *Review of Behavioral Finance*, *ahead-of-print*. https://doi.org/10.1108/RBF-05-2020-0113
Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making*, *5*(7), 467-476. https://doi.org/10.1017/S1930297500001637
Harris, A. J. L., & Hahn, U. (2011). Unrealistic optimism about future life events: A cautionary note. *Psychological Review*, *118*(1), 135-154. https://doi.org/10.1037/a0020997
Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychol Rev*, *106*(4), 766-794. https://doi.org/10.1037/0033-295x.106.4.766
Herz, H., Schunk, D., & Zehnder, C. (2014). How do judgmental overconfidence and overoptimism shape innovative activity? *Games and Economic Behavior*, *83*, 1-23. https://doi.org/10.1016/j.geb.2013.11.001
Hilgert, M., Hogarth, J., & Beverly, S. (2003). Household Financial Management: The Connection Between Knowledge and Behavior. *Federal Reserve Bulletin*, *89*, 309-322.
Hochheimer, C. J., Sabo, R. T., Krist, A. H., Day, T., Cyrus, J., & Woolf, S. H. (2016). Methods for Evaluating Respondent Attrition in Web-Based Surveys. *J Med Internet Res*, *18*(11), e301. https://doi.org/10.2196/jmir.6342
Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture*, *2*(1). https://doi.org/10.9707/2307-0919.1014
Hung, A., Parker, A., & Yoong, J. (2009). Defining and Measuring Financial Literacy. *RAND Corporation Publications Department, Working Papers*, *708*. https://doi.org/10.2139/ssrn.1498674
Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, *1*(1), 69-82. https://doi.org/10.2307/1164736
Isidore, R., & Christie, P. (2019). The relationship between the income and behavioural biases. *Journal of Economics, Finance and Administrative Science*, *24*(47), 127-144. https://doi.org/10.1108/JEFAS-10-2018-0111
Jager, J., Putnick, D. L., & Bornstein, M. H. (2017). II. More than just convenient: the scientific merits of homogeneous convenience samples. *Monogr Soc Res Child Dev*, *82*(2), 13-30. https://doi.org/10.1111/mono.12296
Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 1038-1052. https://doi.org/10.1037/0278-7393.25.4.1038

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: a naïve sampling model of intuitive confidence intervals. *Psychol Rev*, *114*(3), 678-703. https://doi.org/10.1037/0033-295x.114.3.678
Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: a critical examination of the hard-easy effect. *Psychol Rev*, *107*(2), 384-396. https://doi.org/10.1037/0033-295x.107.2.384
Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697-720. https://doi.org/10.1037/0003-066X.58.9.697
Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430-454. https://doi.org/10.1016/0010-0285(72)90016-3
Kaiser, T., & Lusardi, A. (2024). Financial Literacy and Financial Education: An Overview. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4800263
Karaa, I., & Kuğu, T. (2016). Determining Advanced and Basic Financial Literacy Relations and Overconfidence, and Informative Social Media Association of University Students in Turkey. *Educational Sciences: Theory & Practice*, *16*. https://doi.org/10.12738/estp.2016.6.0415
Kimiyaghalam, F., & Safari, M. (2015). Review papers on definition of financial literacy and its measurement. *SEGi Review*, *8*, 81-94.
Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It Depends on How, What, and Whom You Ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216-247. https://doi.org/10.1006/obhd.1999.2847
Krajc, M., & Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. *Journal of Economic Psychology*, *29*(5), 724-738. https://doi.org/10.1016/j.joep.2007.12.006
Krawczyk, M., & Wilamowski, M. (2019). Task difficulty and overconfidence. Evidence from distance running. *Journal of Economic Psychology*, *75*, 102128. https://doi.org/10.1016/j.joep.2018.12.002
Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213-236. https://doi.org/10.1002/acp.2350050305
Krueger, J. (1998). Enhancement Bias in Descriptions of Self and Others. *Personality and Social Psychology Bulletin*, *24*(5), 505-516. https://doi.org/10.1177/0146167298245006
Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *J Pers Soc Psychol*, *82*(2), 180-188.
Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *J Pers Soc Psychol*, *77*(2), 221-232. https://doi.org/10.1037//0022-3514.77.2.221
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol*, *77*(6), 1121-1134. https://doi.org/10.1037//0022-3514.77.6.1121
Larrick, R. P., Burson, K. A., & Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes*, *102*(1), 76-94. https://doi.org/10.1016/j.obhdp.2006.10.002
Lechuga, J., & Wiebe, J. S. (2011). Culture and Probability Judgment Accuracy: The Influence of Holistic Reasoning. *J Cross Cult Psychol*, *42*(6), 1054-1065. https://doi.org/10.1177/0022022111407914

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159-183. https://doi.org/10.1016/0030-5073(77)90001-0
Lind, T., Ahmed, A., Skagerlund, K., Strömbäck, C., Västfjäll, D., & Tinghög, G. (2020). Competence, Confidence, and Gender: The Role of Objective and Subjective Financial Knowledge in Household Finance. *Journal of Family and Economic Issues*, *41*(4), 626-638. https://doi.org/10.1007/s10834-020-09678-9
Lundeberg, M. A., Fox, P. W., & Punćcohaŕ, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, *86*(1), 114-121. https://doi.org/10.1037/0022-0663.86.1.114
Lusardi, A. (2008). Financial Literacy: An Essential Tool for Informed Consumer Choice? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1336389
Lusardi, A., & Mitchell, O. S. (2007). Financial Literacy and Retirement Preparedness: Evidence and Implications for Financial Education. *Business Economics*, *42*(1), 35-44. https://doi.org/10.2145/20070104
Lusardi, A., & Mitchell, O. S. (2011a). Financial Literacy and Planning: Implications for Retirement Wellbeing. *National Bureau of Economic Research Working Paper Series*, *Working Paper n. 17078*. https://doi.org/10.3386/w17078
Lusardi, A., & Mitchell, O. S. (2011b). Financial literacy around the world: an overview. *Journal of Pension Economics and Finance*, *10*(4), 497-508. https://doi.org/10.1017/S1474747211000448
Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence [Article]. *Journal of Economic Literature*, *52*(1), 5-44. https://doi.org/10.1257/jel.52.1.5
Magnus, J. R., & Peresetsky, A. A. (2022). A Statistical Explanation of the Dunning-Kruger Effect. *Front Psychol*, *13*, 840180. https://doi.org/10.3389/fpsyg.2022.840180
Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, *45*(6), 633-644. https://doi.org/10.1509/jmkr.45.6.633
McDonald, J. H. (2014). *Handbook of Biological Statistics* (3rd ed.). Sparky House Publishing, Baltimore, Maryland.
Moore, D. A., & Dev, A. S. (2018). Individual differences in overconfidence. In V. Zeigler-Hill and T. K. Shackelford (Eds.), Encyclopedia of Personality and Individual Differences. New York: Springer.
Moore, D. A., Dev, A. S., & Goncharova, E. (2018). Overconfidence Across Cultures. *Collabra: Psychology*, *4*, 36. https://doi.org/10.1525/collabra.153
Moore, D. A., & Healy, P. J. (2008). The Trouble With Overconfidence. *Psychological Review*, *115*, 502-517. https://doi.org/10.1037/0033-295X.115.2.502
Muller, A., Sirianni, L. A., & Addante, R. J. (2021). Neural correlates of the Dunning-Kruger effect. *Eur J Neurosci*, *53*(2), 460-484. https://doi.org/10.1111/ejn.14935
Nejad, M. G., & Javid, K. (2018). Subjective and objective financial literacy, opinion leadership, and the use of retail banking services. *International Journal of Bank Marketing*, *36*(4), 784-804. https://doi.org/10.1108/IJBM-07-2017-0153
Olsson, H. (2014). Measuring overconfidence: Methodological problems and statistical artifacts. *Journal of Business Research*, *67*. https://doi.org/10.1016/j.jbusres.2014.03.002
Perreault, W. D. (1975). Controlling order-effect bias. *Public Opinion Quarterly*, *39*(4), 544-551. https://doi.org/10.1086/268251
Prims, J. P., & Moore, D. A. (2017). Overconfidence over the lifespan. *Judgment and Decision Making*, *12*(1), 29-41. https://doi.org/10.1017/S1930297500005222

Quote Investigator. (2018, November 18). *It Ain't What You Don't Know That Gets You Into Trouble. It's What You Know for Sure That Just Ain't So*. https://quoteinvestigator.com/2018/11/18/know-trouble
Remund, D. (2010). Financial Literacy Explicated: The Case for a Clearer Definition in an Increasingly Complex Economy. *Journal of Consumer Affairs*, *44*, 276-295. https://doi.org/10.1111/j.1745-6606.2010.01169.x
Russo, J., & Schoemaker, P. (1992). Managing Overconfidence. *Sloan Management Review*, *33*, 7-17.
Sanchez, C., & Dunning, D. (2023). Are experts overconfident?: An interdisciplinary review. *Research in Organizational Behavior*, *43*, 100195. https://doi.org/10.1016/j.riob.2023.100195
Seybold, M. (2016, October 6th). *The Apocryphal Twain: "Things we know that just ain't so."*. Center for Mark Twain Studies. https://marktwainstudies.com/the-apocryphal-twain-things-we-know-that-just-aint-so/
Simon, H. A. (1947). *Administrative behavior; a study of decision-making processes in administrative organization*. Macmillan.
Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99-118. https://doi.org/10.2307/1884852
Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education*, *9*(2), 169-191. https://doi.org/10.5465/AMLE.2010.51428542
Skała, D. (2008). Overconfidence in psychology and finance - an interdisciplinary literature review. *Bank i Kredyt, 4*, 33-50.
Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. 108-131. https://doi.org/10.1207/S15327957PSPR0402_01
Soll, J. B. (1996). Determinants of Overconfidence and Miscalibration: The Roles of Random Error and Ecological Structure. *Organizational Behavior and Human Decision Processes*, *65*(2), 117-137. https://doi.org/10.1006/obhd.1996.0011
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645-665. https://doi.org/10.1017/S0140525X00003435
Sullivan, G. M., & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. *J Grad Med Educ*, *4*(3), 279-282. https://doi.org/10.4300/jgme-d-12-00156.1
Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, *47*(2), 143-148. https://doi.org/10.1016/0001-6918(81)90005-6
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257-285. https://doi.org/10.1016/0364-0213(88)90023-7
Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273-1296. https://doi.org/10.1007/s11165-016-9602-2
Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193-210. https://doi.org/10.1037/0033-2909.103.2.193
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases [Article]. *Science*, *185*(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293-315. https://doi.org/10.1037/0033-295X.90.4.293

Van Rooij, M., Lusardi, A., & Alessie, R. (2011). Financial literacy and stock market participation. *Journal of Financial Economics*, *101*(2), 449-472. https://doi.org/10.1016/j.jfineco.2011.03.006

Vörös, Z., Szabó, Z., Kehl, D., Kovacs, O., Papp, T., & Schepp, Z. (2021). The forms of financial literacy overconfidence and their role in financial well-being. *International Journal of Consumer Studies*, *45*, 1292-1308. https://doi.org/10.1111/ijcs.12734

Xia, T., Wang, Z., & Li, K. (2014). Financial Literacy Overconfidence and Stock Market Participation. *Social Indicators Research*, *119*(3), 1233-1245. https://doi.org/10.1007/s11205-013-0555-9

Xu, Y., Huang, G.-H., Xiao, Y., Li, S., Wang, W., & Liang, Z.-Y. (2024). A Double-Edged-Sword Effect of Overplacement: Social Comparison Bias Predicts Gambling Motivations and Behaviors in Chinese Casino Gamblers. *Journal of Gambling Studies*, *40*(3), 1-20. https://doi.org/10.1007/s10899-024-10293-8

Yates, J. F., Lee, J. W., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-Cultural Variations in Probability Judgment Accuracy: Beyond General Knowledge Overconfidence? *Organ Behav Hum Decis Process*, *74*(2), 89-117. https://doi.org/10.1006/obhd.1998.2771

Zytek, R. (2018). Time Value of Money in a FinLit Survey. https://www.researchgate.net/publication/340507481_Time_Value_of_Money_in_a_FinLit_Survey

Appendices

Appendix 1 – Survey (in Dutch)

Zelfinschatting in financiële geletterdheid & gedragsaspecten

Start of Block: Toestemmingsformulier

Toestemmingsformulier ethische procedure. **Welkom!**
U bent uitgenodigd om deel te nemen aan een masterscriptie onderzoek uitgevoerd door Thijs Daggenvoorde, student Bedrijfskunde aan de Universiteit Twente. Voordat u deelneemt is het belangrijk om het hoofddoel van het onderzoek en de procedures voor deelname te begrijpen.

Dit onderzoek is allereerst gericht op het onderzoeken van verschillende vormen van vertrouwensmiskalibraties in de context van financiële geletterdheid. Dit klinkt in eerste instantie misschien ingewikkeld, maar houdt simpelweg in dat er zal worden gekeken naar de nauwkeurigheid van prestatie inschattingen bij het beantwoorden van vragen over financiële thema's. Binnen deze selectie van vragen zal onderscheid worden gemaakt tussen verschillende moeilijkheidsniveaus. Er zullen dus ook vragen tussen zitten die als lastig kunnen worden ervaren. Ook zal worden gekeken naar de invloed van gedragsmatige aspecten (bijv. tijd per vraag of per set) op de gemaakte inschattingen en de hieruit voortvloeiende verschillen. De dataverzameling bestaat enkel uit deze enquête, waarbij de gegeven antwoorden, klikgegevens en tijdsdata van u als deelnemer worden verzameld, opgeslagen en geanalyseerd.

De enquête is gericht aan inwoners van Nederland, 18 jaar of ouder. Vult u hem alstublieft alleen in wanneer u aan deze 2 criteria voldoet. Alle gegevens worden vertrouwelijk behandeld. Deelname aan het onderzoek is vrijwillig en anoniem. Het invullen van de enquête zal naar verwachting ongeveer 10-12 minuten duren. U kunt zich op elk moment, zonder opgaaf van reden, terugtrekken van deelname. Hiervoor, en voor andere vragen, kunt u contact opnemen per mail: [e-mailadres]. De scriptie zal na afronding worden gepubliceerd op de hiervoor bestemde website van de universiteit.

Door “Ik ga akkoord” te selecteren gaat u akkoord met de hierboven beschreven voorwaarden, erkent u dat u voldoet aan de gestelde deelnamecriteria, en verleent u toestemming aan de onderzoeker om uw data te gebruiken in het kader van dit onderzoek.

- ○ Ik ga akkoord (1)
- ○ Ik ga niet akkoord (2)

Skip To: End of Survey If Welkom! U bent uitgenodigd om deel te nemen aan een masterscriptie onderzoek uitgevoerd door Thij... = Ik ga niet akkoord

End of Block: Toestemmingsformulier

Start of Block: Hulpmiddelen procedure

Hulpmiddelen. **Ter info**
Het is voor mij belangrijk dat u geen externe bronnen (zoals het internet of een rekenmachine) gebruikt bij het beantwoorden van de vragen in deze enquête. Belooft u de vragen in deze enquête te beantwoorden zonder de hulp van externe bronnen?

- ○ Ja (1)
- ○ Nee (2)

Skip To: End of Survey If Ter info Het is voor mij belangrijk dat u geen externe bronnen (zoals het internet of een rekenma... = Nee

End of Block: Hulpmiddelen procedure

Start of Block: Instructie

Instructie. **Instructie (belangrijk!)**
U krijgt zo 15 multiple choice vragen* te zien over financiële onderwerpen, verdeeld in 3 sets van 5 vragen. Probeert u deze alstublieft zo goed mogelijk te beantwoorden. Voor iedere vraag krijgt u 50 seconden om het antwoord te selecteren dat volgens u juist is. Wanneer de tijd om is zal automatisch door worden gegaan naar de volgende vraag. Indien u eerder klaar bent kunt u doorgaan via het pijltje rechtsonder in beeld.

Naast de antwoordopties kunt u ook aangeven wanneer u het antwoord niet weet of wanneer u niet genoeg tijd had om de vraag te beantwoorden. Deze optie staat in eerste instantie altijd standaard geselecteerd, zodat altijd iets wordt ingevuld (i.v.m. de tijdslimiet). Het is dus de bedoeling dat u deze verandert naar het antwoord dat u denkt dat goed is, wanneer u het antwoord weet (of denkt te weten). We beginnen met een voorbeeld.

*Vragen zijn afkomstig en vertaald uit de publicatie in Journal of Financial Economics, 101(2), van Rooij, M., Lusardi, A. & Alessie, R., Financial literacy and stock market participation, pp. 449-472, Copyright Elsevier (2011). In sommige gevallen zijn de vragen aangepast. Toestemming is verkregen voor hergebruik en aanpassingen.

End of Block: Instructie

Start of Block: Voorbeeldvraag

TimerVB. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Voorbeeldvraag. **Dit is een voorbeeldvraag.** Stel dat u aandelen koopt voor €100 en er 7% rendement over behaalt, wat is dan uw rendement in euro's?

- ○ €7 (1)
- ○ €107 (2)
- ○ €700 (3)
- ○ Weet ik niet / Niet genoeg tijd (4)

End of Block: Voorbeeldvraag

Start of Block: Verdere instructie

Verdere instructie. **Verdere instructie (belangrijk!)**
Elke set bestaat uit 5 van dit soort vragen. Hierna wordt u (per set) gevraagd een inschatting van uw eigen prestaties en de prestaties van een willekeurige andere deelnemer te maken. Dit kunt u doen door kansen in te schatten.

Stel bijvoorbeeld dat u na een set denkt dat u zeker 2 vragen goed heeft, 2 vragen fout heeft, maar erg twijfelt over de vijfde vraag. Dan zou u kunnen kiezen voor 50% kans op 'precies 2 goed' en 50% kans op 'precies 3 goed'. U denkt immers dat u er zeker 2 goed heeft, met een 50/50 kans op de 3e goed. Twijfelt u, maar neigt u toch meer naar 2 goed, dan kunt u bijvoorbeeld ook kiezen voor 70% kans op 'precies 2 goed', en 30% kans op 'precies 3 goed'. Op deze manier zijn veel scenario's mogelijk, afhankelijk van uw eigen inschatting. Het totaal van de percentages dient uiteraard altijd 100% te zijn.

Let op! Decimale getallen in deze vragen duiden op een percentage (bijvoorbeeld: 0,2 betekent 20%).

Voorbeeld SPIES. Hoe groot schat u de kans dat u in dit voorbeeld...

_____ precies 0 vragen goed heeft beantwoord (0)
_____ precies 1 vraag goed heeft beantwoord (1)

End of Block: Verdere instructie

Start of Block: Basisvragen

Pauze BFL. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Aanduiding begin. Neem nu eventueel een korte pauze om de volgende set vragen fris en geconcentreerd in te gaan (maar klik de enquête niet weg). Bij het klikken op "Start volgende set" zal 1 van de 3 sets worden gestart. Iedere set bevat weer andere vragen.

End of Block: Basisvragen

Start of Block: BFL 1

TimerBFL1. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 1. Stel, u heeft € 100,- op een spaarrekening staan en de rente bedraagt 2% per jaar. Hoeveel denkt u dat u na 5 jaar op deze rekening zou hebben als u het geld zou laten groeien?

- ○ Meer dan €102 (1)
- ○ Precies €102 (2)
- ○ Minder dan €102 (3)
- ○ Weet ik niet / Niet genoeg tijd (4)

End of Block: BFL 1

Start of Block: BFL 2

TimerBFL2. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 2. Stel, u heeft € 100,- op een spaarrekening staan, de rente bedraagt 20% per jaar en u haalt het spaargeld en de rentebetalingen er niet af. Hoeveel denkt u dat u in totaal na 5 jaar op deze rekening heeft staan?

- Meer dan €200 (1)
- Precies €200 (2)
- Minder dan €200 (3)
- Weet ik niet / Niet genoeg tijd (4)

End of Block: BFL 2

Start of Block: BFL 3

TimerBFL3. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 3. Stel dat de rente op uw spaarrekening 1% per jaar bedraagt en de inflatie 2% per jaar is. Na 1 jaar, hoeveel kunt u kopen met het geld dat op deze spaarrekening staat?

- Meer dan vandaag (1)
- Precies hetzelfde (2)
- Minder dan vandaag (3)
- Weet ik niet / Niet genoeg tijd (4)

End of Block: BFL 3

Start of Block: BFL 4

TimerBFL4. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 4. Stel dat een vriend 3 jaar geleden €10.000 heeft geërfd en dat zijn broer vandaag €10.000 erft. De vriend heeft de erfenis in de tussentijd geïnvesteerd en hierover een positief rendement behaald. Wie is nu rijker door de erfenis, gezien de tijdswaarde van geld?

- ○ Mijn vriend (1)
- ○ Zijn broer (2)
- ○ Ze zijn even rijk (3)
- ○ Weet ik niet / Niet genoeg tijd (4)

End of Block: BFL 4

Start of Block: BFL 5

TimerBFL5. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 5. Stel dat in het jaar 2040 uw inkomen is verdubbeld en de prijzen van alles wat u koopt ook zijn verdubbeld. Hoeveel kunt u dan kopen met uw inkomen?

- ○ Meer dan vandaag (1)
- ○ Hetzelfde (2)
- ○ Minder dan vandaag (3)
- ○ Weet ik niet / Niet genoeg tijd (4)

End of Block: BFL 5

Start of Block: Perceptie basisvragen

Timer SPIES1. BFL Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

SPIES1 BFL. Hoe groot schat u de kans dat **u** van de **afgelopen 5 vragen**...

_____ precies 0 vragen goed heeft beantwoord (0)
_____ precies 1 vraag goed heeft beantwoord (1)
_____ precies 2 vragen goed heeft beantwoord (2)
_____ precies 3 vragen goed heeft beantwoord (3)
_____ precies 4 vragen goed heeft beantwoord (4)
_____ precies 5 vragen goed heeft beantwoord (5)

Page Break

Timer SPIES2. BFL Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

SPIES2 BFL. Hoe groot schat u de kans dat **een willekeurig gekozen andere deelnemer** van de **afgelopen 5 vragen**...

_____ precies 0 vragen goed heeft beantwoord (0)
_____ precies 1 vraag goed heeft beantwoord (1)
_____ precies 2 vragen goed heeft beantwoord (2)
_____ precies 3 vragen goed heeft beantwoord (3)
_____ precies 4 vragen goed heeft beantwoord (4)
_____ precies 5 vragen goed heeft beantwoord (5)

End of Block: Perceptie basisvragen

Start of Block: Geavanceerde vragen 1

Pauze AFL1. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Aanduiding vervolg. Neem nu eventueel een korte pauze om de volgende set vragen fris en geconcentreerd in te gaan (maar klik de enquête niet weg). Bij het klikken op "Start volgende set" zal 1 van de 3 sets worden gestart. Iedere set bevat weer andere vragen.

End of Block: Geavanceerde vragen 1

Start of Block: AFL 1

TimerAFL1. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 6. Welk van deze uitspraken beschrijft de belangrijkste functie van de aandelenmarkt?

- ○ De aandelenmarkt helpt de aandelenwinsten te voorspellen (1)
- ○ De aandelenmarkt resulteert in een stijging van de prijs van aandelen (2)
- ○ De aandelenmarkt brengt mensen die aandelen willen kopen samen met mensen die aandelen willen verkopen (3)
- ○ Geen van bovenstaande opties (4)
- ○ Weet ik niet / Niet genoeg tijd (5)

End of Block: AFL 1

Start of Block: AFL 2

TimerAFL2. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 7. Welk van de volgende uitspraken is correct? Als iemand aandelen koopt van bedrijf B op de aandelenmarkt, dan...

- ○ is deze persoon mede-eigenaar van dit bedrijf (1)
- ○ heeft deze persoon geld geleend aan dit bedrijf (2)
- ○ is deze persoon aansprakelijk voor de schulden van dit bedrijf (3)
- ○ Geen van bovenstaande opties (4)
- ○ Weet ik niet / Niet genoeg tijd (5)

End of Block: AFL 2

Start of Block: AFL 3

TimerAFL3. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 8. Welk van de volgende uitspraken is correct?

- ○ Wanneer je in een beleggingsfonds investeert kan je het geld er in het eerste jaar over het algemeen niet uit opnemen (1)
- ○ Beleggingsfondsen kunnen in meerdere activa investeren, bijv. in zowel aandelen als obligaties (2)
- ○ Beleggingsfondsen betalen een gegarandeerd rendement, gebaseerd op prestaties in het verleden (3)
- ○ Geen van bovenstaande opties (4)
- ○ Weet ik niet / Niet genoeg tijd (5)

End of Block: AFL 3

Start of Block: AFL 4

TimerAFL4. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 9. Welk activum geeft, rekening houdend met een lange tijdsperiode (bijvoorbeeld 10 of 20 jaar), doorgaans het hoogste rendement?

- ○ Spaarrekeningen (1)
- ○ Obligaties (2)
- ○ Aandelen (3)
- ○ Weet ik niet / Niet genoeg tijd (4)

End of Block: AFL 4

Start of Block: AFL 5

TimerAFL5. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 10. Welk activum vertoont doorgaans de meeste fluctuatie over tijd?

- ○ Spaarrekeningen (1)
- ○ Obligaties (2)
- ○ Aandelen (3)
- ○ Weet ik niet / Niet genoeg tijd (4)

End of Block: AFL 5

Start of Block: Perceptie geavanceerde vragen 1

Timer SPIES1 AFL1. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

SPIES1 AFL1. Hoe groot schat u de kans dat u van de **afgelopen 5 vragen...**

_____ precies 0 vragen goed heeft beantwoord (0)
_____ precies 1 vraag goed heeft beantwoord (1)
_____ precies 2 vragen goed heeft beantwoord (2)
_____ precies 3 vragen goed heeft beantwoord (3)
_____ precies 4 vragen goed heeft beantwoord (4)
_____ precies 5 vragen goed heeft beantwoord (5)

Page Break

Timer SPIES2 AFL1. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

SPIES2 AFL1. Hoe groot schat u de kans dat een **willekeurig gekozen andere deelnemer** van de **afgelopen 5 vragen...**

_____ precies 0 vragen goed heeft beantwoord (0)
_____ precies 1 vraag goed heeft beantwoord (1)
_____ precies 2 vragen goed heeft beantwoord (2)
_____ precies 3 vragen goed heeft beantwoord (3)
_____ precies 4 vragen goed heeft beantwoord (4)
_____ precies 5 vragen goed heeft beantwoord (5)

End of Block: Perceptie geavanceerde vragen 1

Start of Block: Geavanceerde vragen 2

Pauze AFL2. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Aanduiding vervolg. Neem nu eventueel een korte pauze om de volgende set vragen fris en geconcentreerd in te gaan (maar klik de enquête niet weg). Bij het klikken op "Start volgende set" zal 1 van de 3 sets worden gestart. Iedere set bevat weer andere vragen.

End of Block: Geavanceerde vragen 2

Start of Block: AFL 6

TimerAFL6. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 11. Wanneer een investeerder zijn geld spreidt over meerdere activa, dan...

- ○ neemt het risico op het verliezen van geld doorgaans toe (1)
- ○ neemt het risico op het verliezen van geld doorgaans af (2)
- ○ blijft het risico op het verliezen van geld doorgaans hetzelfde (3)
- ○ Weet ik niet / Niet genoeg tijd (4)

End of Block: AFL 6

Start of Block: AFL 7

TimerAFL7. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 12. Waar of niet waar? Als je een 10-jaar lopende obligatie koopt, betekent dat dat je hem niet na 5 jaar kunt verkopen zonder een hoge boete.

- ○ Waar (1)
- ○ Niet waar (2)
- ○ Weet ik niet / Niet genoeg tijd (3)

End of Block: AFL 7

Start of Block: AFL 8

TimerAFL8. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 13. Waar of niet waar? Aandelen zijn normaal gesproken risicovoller dan obligaties.

- ○ Waar (1)
- ○ Niet waar (2)
- ○ Weet ik niet / Niet genoeg tijd (3)

End of Block: AFL 8

Start of Block: AFL 9

TimerAFL9. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 14. Waar of niet waar? Het kopen van bedrijfsaandelen biedt doorgaans een veiliger rendement dan een aandelenbeleggingsfonds.

- ○ Waar (1)
- ○ Niet waar (2)
- ○ Weet ik niet / Niet genoeg tijd (3)

End of Block: AFL 9

Start of Block: AFL 10

TimerAFL10. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Kennisvraag 15. Als de marktrente daalt, wat zal er dan waarschijnlijk gebeuren met de prijzen van obligaties?

- ○ Die gaan omhoog (1)
- ○ Die gaan omlaag (2)
- ○ Die blijven gelijk (3)
- ○ Weet ik niet / Niet genoeg tijd (4)

End of Block: AFL 10

Start of Block: Perceptie geavanceerde vragen 2

Timer SPIES1 AFL2. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

SPIES1 AFL2. Hoe groot schat u de kans dat **u** van de **afgelopen 5 vragen...**

_____ precies 0 vragen goed heeft beantwoord (0)
_____ precies 1 vraag goed heeft beantwoord (1)
_____ precies 2 vragen goed heeft beantwoord (2)
_____ precies 3 vragen goed heeft beantwoord (3)
_____ precies 4 vragen goed heeft beantwoord (4)
_____ precies 5 vragen goed heeft beantwoord (5)

Page Break

Timer SPIES2 AFL2. Timing

First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

SPIES2 AFL2. Hoe groot schat u de kans dat **een willekeurig gekozen andere deelnemer** van de **afgelopen 5 vragen...**

_____ precies 0 vragen goed heeft beantwoord (0)
_____ precies 1 vraag goed heeft beantwoord (1)
_____ precies 2 vragen goed heeft beantwoord (2)
_____ precies 3 vragen goed heeft beantwoord (3)
_____ precies 4 vragen goed heeft beantwoord (4)
_____ precies 5 vragen goed heeft beantwoord (5)

End of Block: Perceptie geavanceerde vragen 2

Start of Block: Socio-demografische en overige vragen

Afsluitende vragen
Zodoende de 3 sets met vragen. Ter afronding enkele vragen over uzelf en over uw overwegingen gedurende deze enquête.

Geslacht. Wat is uw gender?

- ○ Man (1)
- ○ Vrouw (2)
- ○ Anders, ik identificeer mij als [specificeren] (3)
- ○ Zeg ik liever niet (4)

Leeftijd. In welke leeftijdscategorie valt u?

- ○ Jonger dan 18 jaar (1)
- ○ 18 tot 24 jaar (2)
- ○ 25 tot 34 jaar (3)
- ○ 35 tot 44 jaar (4)
- ○ 45 tot 54 jaar (5)
- ○ 55 tot 64 jaar (6)
- ○ 65 jaar of ouder (7)

Educatie. Wat is het hoogste opleidingsniveau dat u heeft afgerond, of is hiermee vergelijkbaar?

- ○ Primair onderwijs (basisschool) (1)
- ○ Voortgezet onderwijs (middelbare school) (2)
- ○ Middelbaar beroepsonderwijs (MBO) (3)
- ○ Hoger beroepsonderwijs (HBO) associate degree (4)
- ○ Hoger beroepsonderwijs (HBO) bachelor's degree (5)
- ○ Wetenschappelijk onderwijs (WO) bachelor's degree (6)
- ○ Hoger beroepsonderwijs (HBO) master's degree (7)
- ○ Wetenschappelijk onderwijs (WO) master's degree (8)
- ○ Kandidaat / PHD (9)
- ○ Zeg ik liever niet (10)

Affiniteit educatie. In hoeverre was de hierboven geselecteerde opleiding gerelateerd aan financiële onderwerpen?

- ○ Heel weinig (1)
- ○ Weinig (2)
- ○ Een beetje (3)
- ○ Veel (4)
- ○ Heel veel (5)
- ○ Zeg ik liever niet (6)

Inkomen. Wat is uw persoonlijke inkomenscategorie per jaar (bruto)?

- ○ €0 tot €19.999 (1)
- ○ €20.000 tot €39.999 (2)
- ○ €40.000 tot €59.999 (3)
- ○ €60.000 tot €79.999 (4)
- ○ €80.000 tot €99.999 (5)
- ○ €100.000+ (6)
- ○ Zeg ik liever niet (7)

Werk. Hoe ziet uw huidige werksituatie er uit?

- ○ In loondienst (1)
- ○ Zelfstandige (2)
- ○ Student (3)
- ○ Werkloos/-zoekend (4)
- ○ Gepensioneerd (5)
- ○ Anders [specificeer] (6)
- ○ Zeg ik liever niet (7)

Affiniteit dagelijks. In hoeverre gebruikt u financiële onderwerpen in uw dagelijkse activiteiten? (Studie, werk & privé)

- ○ Heel weinig (1)
- ○ Weinig (2)
- ○ Een beetje (3)
- ○ Veel (4)
- ○ Heel veel (5)

Gezinssituatie. Welke gezinssituatie is het best op u van toepassing?

- ○ Woonachtig bij ouder(s)/verzorger(s) (1)
- ○ Eenpersoonshuishouden (2)
- ○ Samenwonend, met partner (ongehuwd) (3)
- ○ Samenwonend, met partner (gehuwd) (4)
- ○ Samenwonend, met partner (ongehuwd) en kind(eren) (5)
- ○ Samenwonend, met partner (gehuwd) en kind(eren) (6)
- ○ Samenwonend, met kind(eren) (7)
- ○ Anders [specificeer] (8)
- ○ Zeg ik liever niet (9)

Persoonlijkheid. In welke mate zijn de volgende stellingen op u van toepassing?

	Zeer oneens (1)	Oneens (2)	Neutraal (3)	Eens (4)	Zeer eens (5)
Ik heb vertrouwen in de inschattingen die ik **voor mijzelf** heb gemaakt na de 3 sets (1)	○	○	○	○	○
Ik heb vertrouwen in de inschattingen die ik **voor anderen** heb gemaakt na de 3 sets (2)	○	○	○	○	○
Ik zie mijzelf als een risicomijdend persoon (3)	○	○	○	○	○
Ik zie mijzelf als een competitief persoon (4)	○	○	○	○	○

Pauzes. Heeft u (korte) pauzes genomen vóór het starten van de verschillende sets?

- ○ Ja, alle keren (1)
- ○ Soms (één of twee keer) (2)
- ○ Nee, geen enkele keer (3)

Medium. Middels welk apparaat heeft u deze enquête ingevuld?

- ○ Smartphone (1)
- ○ Computer/laptop (2)
- ○ Anders [specificeer] (3)

End of Block: Socio-demografische en overige vragen

Start of Block: Einde van de enquête

Bedankt voor het invullen van de enquête. Uw antwoorden zijn opgeslagen. U kunt het tabblad sluiten.

End of Block: Einde van de enquête

Appendix 2 – Origins and inspirations of financial literacy questions

Before their use in the FL modules of the DHS, the questions were taken from several sources as described below. Allocation of these sources is logically derived from van Rooij et al. (2011) as source article: in-text (p. 452) and footnotes 4 (p. 452) and 6 (p. 453). Named in footnote 4, Lusardi and Mitchell (2011a) has also been consulted for some sources. This table was created to give credit to the primary sources as well, following a copyright specialists' instructions. These sources are provided for attribution; I did not directly read them myself. For clarification, please note that Moore (2003) is a different author than Moore (as in Moore and Healy (2008), and Prims and Moore (2017)) in overconfidence studies.

Query	Set**	Category	Original sources and/or -inspirations of questions
Q1	1	BFL	Health and Retirement Study (2004), as cited in Lusardi and Mitchell (2011a)***
Q2	1	BFL	New in the FL modules of the DHS
Q3	1	BFL	Health and Retirement Study (2004), as cited in Lusardi and Mitchell (2011a)
Q4	1	BFL	New in the FL modules of the DHS
Q5	1	BFL	New in the FL modules of the DHS
Q6	2	AFL	National Council on Economic Education (2005), as cited in Van Rooij et al. (2011)
Q7	2	AFL	National Association of Securities Dealers (2003), as cited in Van Rooij et al. (2011)****
Q8	2	AFL	Hogarth and Hilgert (2002), John Hancock Financial Services (2002), and Moore (2003), as cited in Van Rooij et al. (2011)
Q9*	2	AFL	Hogarth and Hilgert (2002), John Hancock Financial Services (2002), and Moore (2003), as cited in Van Rooij et al. (2011)
Q10	2	AFL	Hogarth and Hilgert (2002), John Hancock Financial Services (2002), and Moore (2003), as cited in Van Rooij et al. (2011)
Q11	3	AFL	Hogarth and Hilgert (2002), John Hancock Financial Services (2002), and Moore (2003), as cited in Van Rooij et al. (2011)
Q12	3	AFL	Hogarth and Hilgert (2002), John Hancock Financial Services (2002), and Moore (2003), as cited in Van Rooij et al. (2011)
Q13	3	AFL	Hogarth and Hilgert (2002), John Hancock Financial Services (2002), and Moore (2003), as cited in Van Rooij et al. (2011)
Q14	3	AFL	Health and Retirement Study (2004), as cited in Lusardi and Mitchell (2011a)
Q15	3	AFL	Hogarth and Hilgert (2002), John Hancock Financial Services (2002), and Moore (2003), as cited in Van Rooij et al. (2011)

Table 19: Origins and inspirations of FL questions, as derived from van Rooij et al. (2011)

*Note 1: Question 9 of the original question set as composed by the source article has been removed from this survey due to concerns of answer dependency with question 7 (see explanation in paragraph 3.3.1). Consequently, the subsequent numbering of the questions in this thesis, starting from question 9, has been adjusted by 1 compared to the numbering in the source article. For instance, what was initially labelled as question 10 became question 9, question 11 became question 10, et cetera.

**Note 2: Sets have been assigned for the purpose of this thesis. These have not necessarily been adopted from the source article (BFL was adopted, AFL was split into 2 sets).

***Note 3: The questions attributed to the referred source of Lusardi and Mitchell (1, 3, and 14) have obviously been developed earlier than the article publication in 2011. Therefore, they can also be seen in several earlier papers/studies by Lusardi and Mitchell. The existence of these queries in the HRS has been demonstrated by this citation, as following assignment by the source article.

****Note 4: The National Association of Securities Dealers (NASD) has consolidated into the Financial Industry Regulatory Authority (FINRA) (Financial Industry Regulatory Authority, 2007).

Appendix 3 – Schematic survey flow

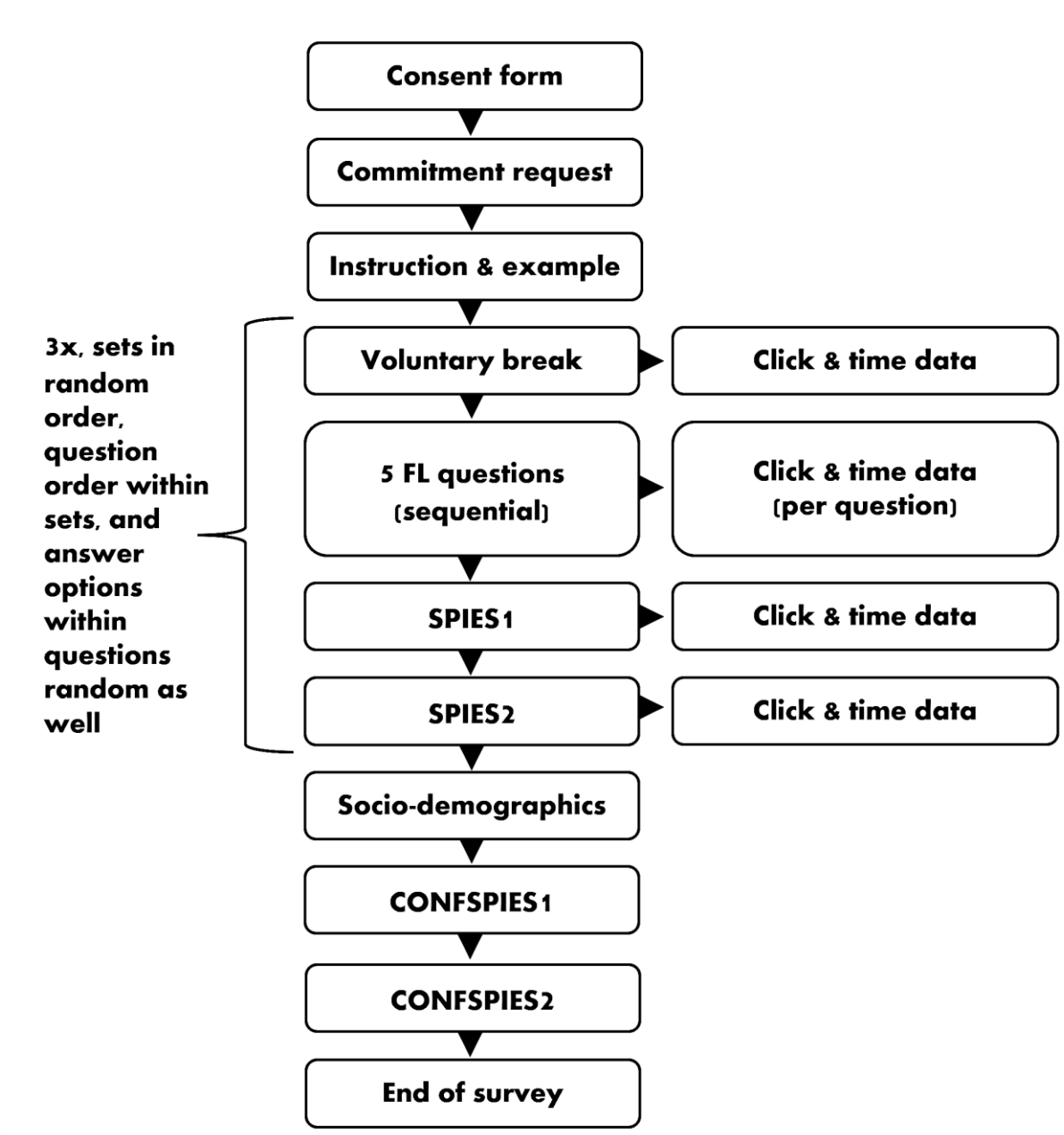


Figure 8: Schematic survey flow

Appendix 4 – Division into individual components of overplacement

As the calculation of OP solely looks at absolute levels, it does not take into account its components. As these might be of importance if differing between set difficulty, these are schematically drawn in this Appendix. The bottom part of the scheme (beliefs) instantly describes the mutual relationship between SPIES1 and SPIES2, which is the first part of the OP calculation, while the upper part (actual) tells whether the respondents' performance was better or worse than average (based on mean comparison, therefore possible to deviate from 50%), which is the second part of the OP calculation. Please note that the percentages at the references (arrows) relate to the relative proportion of the beliefs.

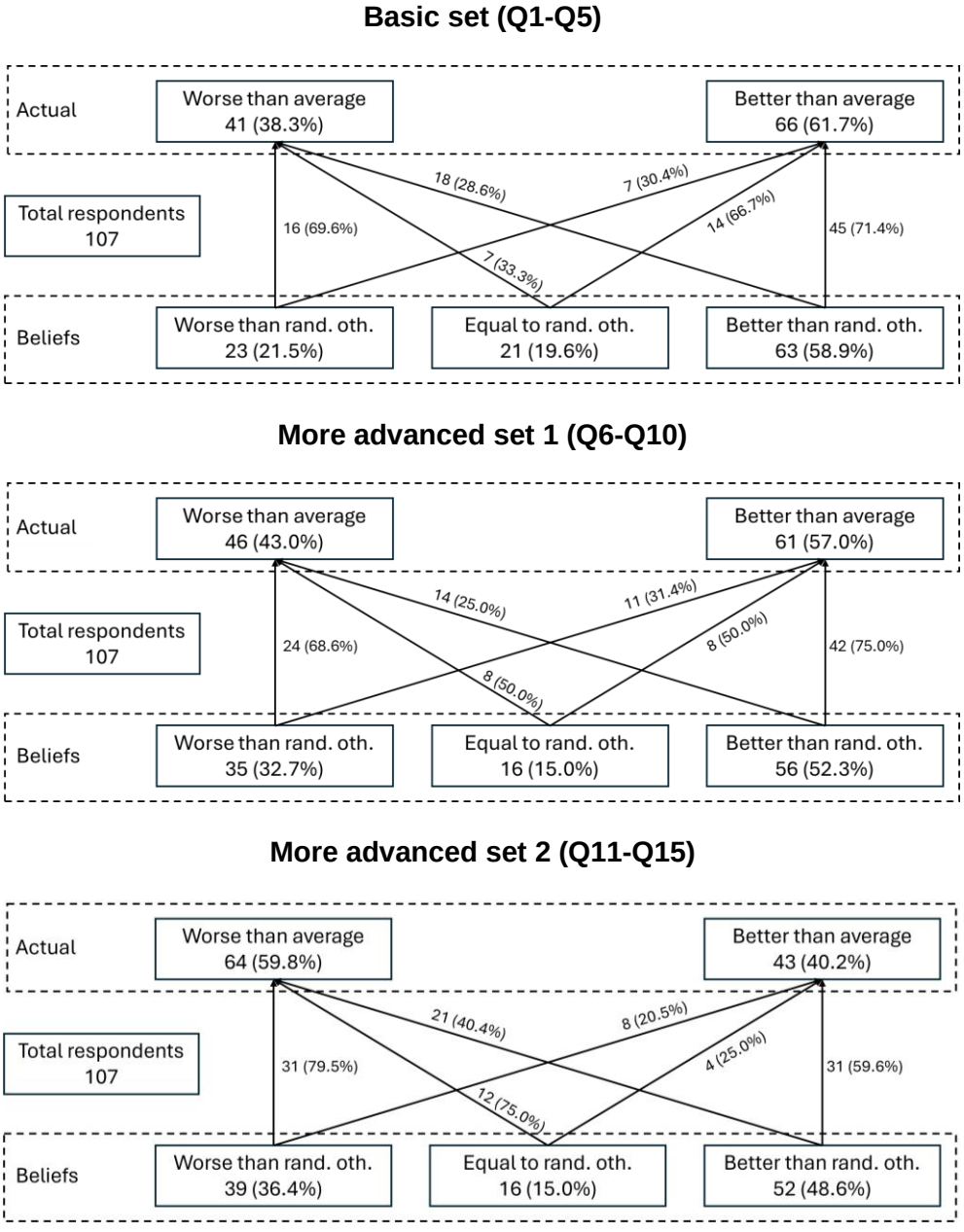
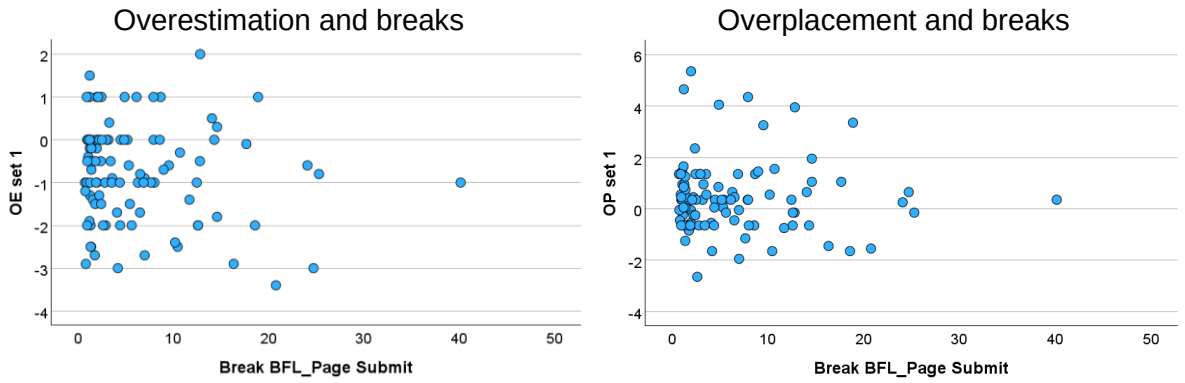


Figure 9: Set specific OP components at an individual level

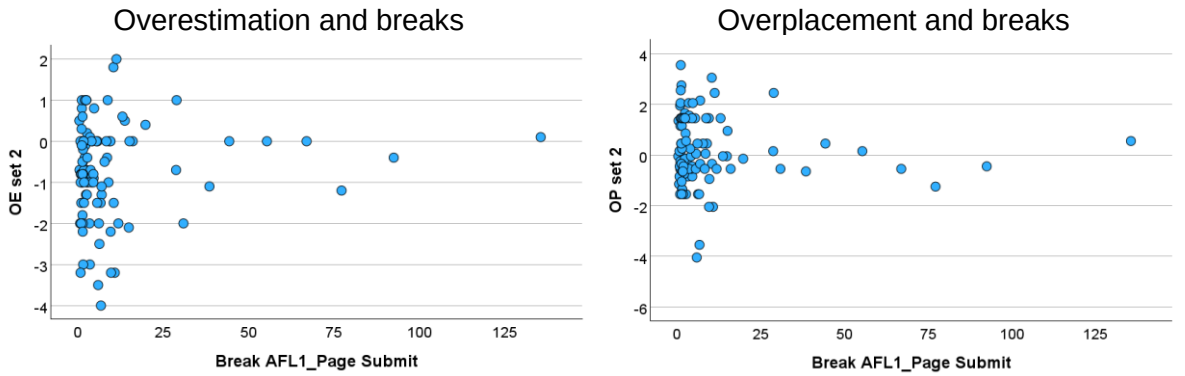
Appendix 5 – Correlations between breaks and OC forms (OE and OP)

The correlations between breaks and OC forms are drawn in this appendix. Extreme outliers in break length have been removed for visualization purposes. The x-axis shows the time-interval of the break (seconds), while the y-axis demonstrates respondents' set-specific level of OE or OP. As misperceptions mostly converge, patterns align expectations. However, the number of people that took breaks (put at 10 sec. threshold) is of such small size that it is difficult to really draw conclusions on it. This should only be considered an initial exploration.

Basic set (Q1-Q5)



More advanced set 1 (Q6-Q10)



More advanced set 2 (Q11-Q15)

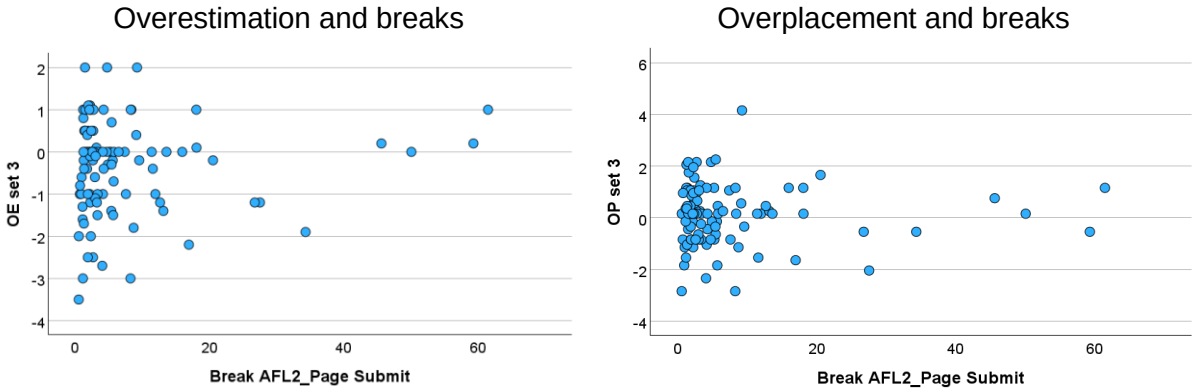


Figure 10: Correlations between breaks and OC forms (OE and OP)