Noname manuscript No.
(will be inserted by the editor)

Much more than a prediction: Expert-based software effort estimation as a behavioral act

Patrícia G. F. Matsubara · Igor Steinmacher · Bruno Gadelha · Tayana Conte

Received: date / Accepted: date

Abstract Traditionally, Software Effort Estimation (SEE) has been portrayed as a technical prediction task, for which we seek accuracy through improved estimation methods and a thorough consideration of effort predictors. In this article, our objective to make explicit the perspective of SEE as a behavioral act, bringing attention to the fact that human biases and noise are relevant components in estimation errors, acknowledging that SEE is more than a prediction task. We employed a thematic analysis of factors affecting expert judgment software estimates to satisfy this objective. We show that estimators do not necessarily behave entirely rationally given the information they have as input for estimation. The reception of estimation requests, the communication of software estimates, and their use also impact the estimation values — something unexpected if estimators were solely focused on SEE as a prediction task. Based on this, we also matched SEE interventions to behavioral ones from Behavioral Economics showing that, although we are already adopting behavioral insights to improve our estimation practices, there are still gaps to build upon. Furthermore, we assessed the strength of evidence for each of our review findings to derive recommendations for practitioners on the SEE interventions they can confidently adopt to improve their estimation processes.

Corresponding author: Patrícia G. F. Matsubara
Federal University of Mato Grosso do Sul (UFMS)
Campo Grande-Brazil
E-mail: patricia.gomes@ufms.br

Igor Steinmacher
Northern Arizona University (NAU)
Flagstaff-USA
E-mail: igor.steinmacher@nau.edu

Bruno Gadelha and Tayana Conte
Federal University of Amazonas (UFAM)
Manaus-Brazil
E-mail: bruno@icomp.ufam.edu.br
E-mail: tayana@icomp.ufam.edu.br

Moreover, in assessing the strength of evidence, we adopted the GRADE-CERQual (Confidence in the Evidence from Reviews of Qualitative research) approach. It enabled us to point concrete research paths to strengthen the existing evidence about SEE interventions based on the dimensions of the GRADE-CERQual evaluation scheme.

Keywords Software Effort Estimation · Behavioral Software Engineering · Bias · Noise

1 Introduction

To forecast is "to calculate or predict (some future event or condition) usually due to study and analysis of available pertinent data" (Merriam-Webster 2021). Forecasts support decision-makers when there is uncertainty about the future, like when estimating software projects or tasks (Halkjelsvik and Jørgensen 2018c).

Forecasting in social sciences differs from forecasting in the physical sciences, and SEE resembles more the first context than the second one. For instance, in the social sciences, the forecasts made affect what one is trying to forecast (Makridakis et al. 2020). Likewise, empirical results suggest that estimates affect the work executed in software projects (Grimstad et al. 2005). Moreover, there is little theoretical or quantitative basis for representing a causal or underlying relationship in the social sciences (Makridakis et al. 2020). In SEE, the relation between software size and effort is unstable (Jørgensen et al. 2009) and varies with the context (Halkjelsvik and Jørgensen 2018b). Brooks (1995) also criticized the unit of effort we commonly use, calling it a myth, as it is based on the false assumption that men and months are interchangeable. Similarities like these remind us of the social side of Software Engineering, and how many of the significant problems of failed software projects are sociological in nature (DeMarco et al. 2013).

To address such problems, SE researchers should explore more cognitive, behavioral, and social sciences concepts, using their knowledge as leverage to more realistic notions of software practitioners' behavior (Lenberg et al. 2014). Lenberg et al. (2015) proposed a definition and a model of such area: Behavioral Software Engineering (BSE), rooting it in the areas of Work and Organizational Psychology, Psychology in Programming, and Behavioral Economics. From these areas, Behavioral Economics is a sub-field of economics that emerged based on models of decision-making under uncertainty (Lenberg et al. 2014), which can also be deemed relevant in the context of SEE.

Behavioral Economics focuses on how real people act and decide, as opposed to the view of the idealized rational and self-interested agent from neoclassical economics (Yamagishi et al. 2014). Tversky and Kahneman (1974) demonstrated that people use heuristics as part of their judgment, either to help them in assessing the probability of uncertain events or predicting the values of uncertain quantities. While such heuristics save time and reduce complexity, they also lead people to systematic biases in their decision-making

(Frid-Nielsen and Jensen 2021). SE researchers recognized many such biases in our field, although overlooked categories of biases remain (Mohanani et al. 2020). For instance, regarding the social biases category in SE literature, Mohanani et al. (2020) identified studies concerning only the bandwagon effect: people's tendency to align with the majority opinion and to do or believe things because the majority of other people are doing or believing it (Vanden-Bos 2015). However, this category also includes other biases, such as cultural bias, stereotypical bias, and attribution error (Fleischmann et al. 2014). Given the social-technical nature of our field, all of these might be worthy of further investigation. Moreover, Mohanani et al. (2020) reports that few debiasing techniques are explored in SE literature, which ultimately means we need to improve on investigations of interventions to deal with the biases that negatively impact SE practice.

In this work, we address the behavioral side of SEE, looking more specifically at judgment-based estimates, which are the most used type of estimates in the software industry (Trendowicz et al. 2011; Usman et al. 2015). We focused on two research questions:

- **RQ1:** What are the latent themes in SEE interventions?
- **RQ2:** Which interventions from behavioral science are explored in the software estimation context?

To answer them, we analyzed the many factors that affect expert judgment software estimates (Matsubara et al. 2022) through reflexive thematic analysis (Braun and Clarke 2021). We found two overarching themes regarding perspectives about software effort estimation: SEE as a **technical prediction task** and as a **behavioral act**. In this work, we focus on the last one: SEE as a behavioral act. Our results show that such a perspective exists both in the academic and industrial context. For instance, on one side, researchers are investigating the impact of psychological biases such as anchoring (Løhre and Jørgensen 2016). On the other side, practitioners are reporting that goals and targets influence estimates (Magazinius et al. 2012) — when they should not if people are supposed to act rationally to get the best estimation result for a given task. Therefore, practitioners need to shape their estimation processes to account for the human behavior that can affect software estimates negatively. For example, one can reframe the first estimate from an estimator as an ideal value, immediately asking for a second estimate and reframing it as a most likely value. Such a small change in the estimation process improves results because it leads to systematically higher values (Jørgensen 2011), assuming that underestimating is a bad outcome, although frequent.

Therefore, looking for effective interventions to deal with human behavior, such as debiasing techniques, to incorporate into our SEE contexts is a must. To further investigate this matter, we focused on interventions to deal with the challenges and issues we face in SEE, making our results of more practical value and immediate application to the software industry. We refer to this subset of the factors affecting expert judgment software estimates as SEE interventions.

Moreover, we investigated interventions from behavioral economics that support decision-making, which we refer to from now on as behavioral interventions. We matched the SEE interventions to the behavioral ones by using codebook thematic analysis (Braun and Clarke 2021), making the link between them explicit. We envision that such effort makes the following contributions to SEE research:

- Strengthen the theoretical ground on which SEE interventions stand, by associating them with empirical results and theories from other disciplines. This can give more confidence to practitioners in choosing such interventions for software process improvement initiatives and to researchers in deriving enhanced guidelines for practitioners;
- Reveals SEE interventions that are unaligned with interventions from other domains and may require further investigation;
- Clarifies potential research gaps, highlighting interventions yet to be tested in the SEE domain. This can guide software engineering researchers in future research efforts.

The remaining of this work is organized as follows. Section 2 presents relevant definitions of Behavioral Economics and behavioral interventions. Section 3 describes our research methodology. Section 4 provides the results answering the two research questions we proposed: the latent themes of SEE interventions and their mapping to behavioral ones. Next, Section 5 presents our discussion of the results. Section 6 discusses the limitations of our work. Section 7 provides the final considerations and future work.

2 Background

Roughly, Economics is a social science focused on optimizing the allocation of scarce resources, with theories rooted in preferences, beliefs, and constraints (Buyalskaya et al. 2021). For a long time, the field was dominated by the view of people as rational decision-makers — a perspective now challenged in a sub-field termed Behavioral Economics, as Section 2.1 shows. Furthermore, Section 2.3 discusses which interventions researchers in Behavioral Economics are investigating to deal with the problems of noise and biases in judgment.

2.1 The Rise and Impact of Behavioral Economics

Classical economics assumes that humans are entirely rational: making decisions and acting to maximize subjective values by integrating relevant information and accounting for risks correctly (Buyalskaya et al. 2021). A famous representative of this branch of thought is the expected utility theory (Thaler 2018), which recommends the alternative with the highest expected utility, or value, considering the preferences of the decision-maker (Briggs 2019).

Simon (2000) is one of the earliest critics of models of rational choice, such as the expected utility theory, questioning their assumptions and capacity

to predict and explain human decision-making. But the limitations of these models gained relevance among scholars in the 70's (Brzezicka and Wisniewski 2014), with the works of psychologists Daniel Kahneman and Amos Tverstky (Frid-Nielsen and Jensen 2021). They showed that people use heuristics (or rules of thumbs) when facing complex assessment and prediction tasks, such as the **availability heuristics**, in which people assess the frequency of a class or the probability of an event by *how easy* it is to bring to mind their instances or occurrences (Tversky and Kahneman 1974). Availability is an ecologically valid clue, in the sense that instances of larger classes and frequent events are easier to recall or to imagine than smaller or infrequent ones — but the problem is that availability is also affected by other factors unrelated to real size or frequencies (Tversky and Kahneman 1973). One such factor can be the salience of an event: if it occurred recently and is fresh in one's mind, one can mistake it for a frequent event, even when it is not. For instance, subjective probabilities of car accidents rise when people see an overturned car on the road (Tversky and Kahneman 1974).

Therefore, Kahneman and Tverstky demonstrated that the use of heuristics led people to make systematic errors — or to *biases* — in judgments (Frid-Nielsen and Jensen 2021), showing that humans adopt behaviors that consistently deviate from the accepted rational theory (Brzezicka and Wisniewski 2014). The discovery of these biases revealed an opportunity to use psychological realism to improve the explanatory power of economic theory (Thaler 2018). The field of Behavioral Economics developed by providing more realistic explanations of how real people act and decide, relegating theories of an entirely rational human as normative — describing how people *should make decisions* (Briggs 2019).

2.2 Error, Bias, and Noise

Like in economic decisions and predictions, expert-judgment SEE is also subject to errors. Anytime one provides an estimated effort value for a software task or project, which equals the actual value from the task' or project' execution, we observe **accuracy** in estimation. When there is a difference between estimated and actual values, we observe an **error** or an inaccuracy. Such errors have two **components** which get in the way of good judgment when estimating: **bias** and **noise** (Kahneman et al. 2016).

Bias is a *systematical deviation* between estimated and actual values (Satopää et al. 2021); a tendency to predict too low or too high values (Halkjelsvik and Jørgensen 2018b). In principle, it is predictable: we can identify its direction and magnitude (Satopää et al. 2021). This suggests that bias has a causal explanation (Kahneman et al. 2021) that we can address. Moreover, we can assess single judgments (like estimates from one person) to conclude how much (un)biased they are, making it easy to notice when bias is causing problems. For instance, an optimistic estimator will likely give consistently too low esti-

mates. Consequently, they can cause damage when estimating large software tasks and projects, which typically require higher estimates due to their size.

Noise is *undesired variability* in judgments that should otherwise be identical Kahneman et al. (2021). It involves unpredictable deviations uncorrelated with the outcome; therefore, it is impossible to anticipate either its direction or magnitude (Satopää et al. 2021). Noise is a problem already detected in the SEE domain, where estimators provided inconsistent estimates for the same tasks on different occasions, based on the same information, and made under similar conditions (Grimstad and Jørgensen 2007).

We cannot directly observe noise in a single judgment, so it might be harder to grasp its presence when only one person estimates one specific task — as is often the case in SEE. However, this does not mean noise or its damages are absent Kahneman et al. (2021). Halkjelsvik and Jørgensen (2018b, p.8) gives us a notion of how noise can be detrimental, noting that an *unbiased* set of estimates is not necessarily an *accurate* set of estimates[1]: a half of the predictions can be far above the actual values while the other half is far bellow. Together they cancel each other's effect. We can only imagine the dissatisfaction of the recipients of such estimates, even though estimators exhibit the excellent quality of being unbiased on average. A much better situation would be one with unbiased and noise-free estimators.

Therefore, in the presence of errors in human judgment and decision-making, a matter of interest is what one can do to overcome them. Therefore, the following section introduces the idea of behavioral interventions to deal with bias and noise, as they are components of error.

2.3 Behavioral Interventions

One alternative is "choice architecture", which studies how to structure decision-making situations to influence choices over alternatives (Münscher et al. 2016). It brings the figure of the choice architect: the person responsible for organizing the context in which decisions occur (Thaler and Sunstein 2021), resorting to various interventions to guide people to make better decisions. Such interventions were popularized under the term nudge: any aspect of the choice architecture that predictably changes people's behavior without resorting to the prohibition of alternatives or changes in incentives. We refer here to large changes in financial incentives. Small economic incentives are acceptable to create a nudge, as well as social incentives (Münscher et al. 2016). (Thaler and Sunstein 2021). Nudges, or choice architecture techniques, as we call them in this work, either help people overcome their biases or take advantage of them. Münscher et al. (2016) organized the choice architecture techniques in a framework useful for their successful development and transfer. Their taxonomy involves three overarching categories: (i) decision information, regarding the presentation of information; (ii) decision structure, regarding the arrangement

[1] However, (Halkjelsvik and Jørgensen 2018b) do not mention noise explicitly in their discussion.

of alternatives and the decision-making format; and (iii) decision assistance, regarding aids to help decision-makers to follow through with their intentions and with beneficial behaviors.

However, biases are not the only problem in human judgment. As Section 2.2 discusses, another relevant error component is noise Kahneman et al. (2021). Such variability happens when professionals contradict their previous judgments when given the same data on different occasions or when different people provide diverging judgments when evaluating or estimating the same cases (Kahneman et al. 2016). Indeed, a little variation is acceptable. The problem lies in significant and intolerable levels of variation (Kahneman et al. 2016). In a recent study in the judgmental forecasting domain, researchers found that noise reduction played a dominant role in the effectiveness of the studied interventions to improve forecasters' performance (Satopää et al. 2021). Therefore, noise reduction strategies can also be considered to reduce estimation errors in the SEE domain.

3 Research Methodology

In this section, we present the research methodology that we adopted. Figure 1 shows that we started with the results from a prior Systematic Literature Mapping (SLM), reported in Matsubara et al. (2022), which resulted in a map of factors affecting software estimates using expert judgment. We used this map as input for the two major phases of our study.

Next, we detail the activities we executed in each phase in Sections 3.1 and 3.2. We also describe how we assessed the strength of evidence of our review findings in Section 3.3.

3.1 Phase 1: Reflexive Thematic Analysis

In the prior SLM, analyzing the factors from the raw data of the primary studies involved an inductive approach (Braun and Clarke 2006). We derived codes from such raw data. Those codes represented candidate factors, which later derived the final factors (Matsubara et al. 2022). During this process, we had prolonged engagement with the data for over one year and a half. We filled data extraction forms provided as supplementary material (Matsubara et al. 2021a).

However, we focused on data from primary studies mostly on a semantic level: we remained at the surface of their meanings without moving beyond what was written (Braun and Clarke 2006). We left out the latent themes — where the researcher identifies and examines underlying ideas, assumptions, and conceptualizations that are theorized as shaping or informing the semantic content of data (Braun and Clarke 2006). Therefore, searching for the latent themes was a research opportunity to comprehend better the phenomenon of the perspectives about SEE both from the research and practice points of view,

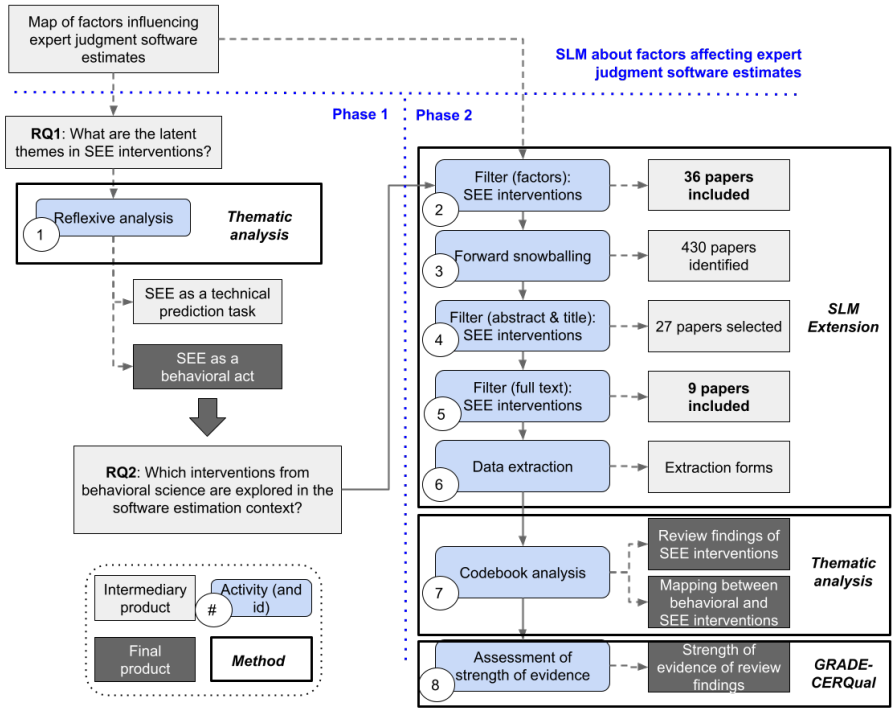

Fig. 1 Research methodology overview presenting the two phases of the study: a first phase focused on the reflexive thematic analysis of SEE interventions and a second phase focused on a codebook thematic analysis matching behavioral and SEE interventions.

so far unexplored in the previous SLM. We also aimed to move from identifying factors affecting the estimates to practical and actionable interventions that practitioners can use to improve estimation processes in the software industry. We proposed our first research question: **RQ1:** What are the latent themes in SEE interventions? To answer it, we revisited the chunks and codes once more.

As Figure 1 (Activity 1) shows, the first phase resulted in two major latent themes: SEE as a **technical prediction task** and as a **behavioral act**. The identification of the two above-mentioned latent themes was a result of revisiting the codes and categories of the previous SLM as an exercise of reflexive thematic analysis (Braun and Clarke 2021). Thematic analysis is a search across a dataset (including a dataset of texts) to find repeated patterns of meaning (Braun and Clarke 2006), united by a central concept (Braun and Clarke 2021). Reflexive thematic analysis is an interpretative reflexive process with no need for a coding framework, that fully embraces qualitative values and the researchers' subjective skills (Braun and Clarke 2021).

Next, we revisited the codes and extracts once more, searching for other additional themes. We did not find any other themes, possibly because the ones we had were already overarching. We were also able to refine and enrich the

description of our current themes. In this phase, as our objective was to show the existence of the themes, we focused only on the factors reported in more than one paper, adopting the same strategy of the original SLM (Matsubara et al. 2022). We realized had we stopped the analysis at that point, it would lead to a premature closure: when the researcher stops to analyze data at a superficial level (Connelly and Peltzer 2016). Therefore, we moved to Phase 2, considering the research question that we derived from the SEE as **behavioral act** theme. We no longer used the approach of reflexive thematic analysis at this phase. Instead, we adopted codebook thematic analysis, as we explain in the next section.

3.2 Phase 2: Codebook Thematic Analysis

To make explicit the perception of SEE as a behavioral act means to bring attention to the fact that human biases and noise are relevant components in prediction error (Satopää et al. 2021). Acknowledging their existence and impact on human judgment, we need interventions to address both to reduce estimating errors. As the field of behavioral economics proposes this kind of intervention, the first phase of our study led to one additional research question: **RQ 2:** Which interventions from behavioral science are explored in the software estimation context? We use the term "behavioral science" to mean the study of what people usually do and how they make decisions in varied and common situations, ultimately to change human behavior — and with Behavioral Economics at its core (Timon 2020).

To answer this question, we need to identify behavioral interventions from a reliable source in behavioral science literature. We explain our steps to do so in Section 3.2.1. We also need to identify the SEE interventions explored in the SE literature — a step we explain in Section 3.2.2. Moreover, we need to detail the analysis procedures we adopted to match these two kinds of interventions, as we do in Section 3.2.3.

3.2.1 The Analytical Framework of Behavioral Interventions

To answer RQ2, we need a theoretical framework of interventions for the reduction of bias and noise. We chose the book by Kahneman et al. (2021) as a reference for such a framework because:

- It provides behavioral interventions for both noise and bias.
- While bias has been addressed in several scientific papers and books, noise has not (Kahneman et al. 2021). The book is a recent attempt to redress the balance and call attention to noise.
- The book gathers interventions scattered in the literature of many different judgment domains where noise can emerge, such as forensic sciences, medicine, hiring, and others.

– The authors are highly knowledgeable and respected in the behavioral and judgment domain. Daniel Kahneman is a Nobel Prize winner[2]. Olivier Sibony is a professor specializing in strategic thinking and decision-making, who worked for McKinsey & Company[3]. Cass Sunstein is a Holberg Prize winner, who served the United States government and the World Health Organization[4].

The book by Kahneman et al. (2021) proposes three categories for bias and noise reduction: **better judges** (for both bias and noise reduction), **debiasing** (for bias reduction), and **decision hygiene** (for noise reduction). However, the authors focus more on noise. To balance the attention between bias and noise, we complemented the theoretical framework with choice architecture techniques presented by Münscher et al. (2016). Such techniques make use of our knowledge of the biases and their impact in creating deviations from rational behavior (Münscher et al. 2016), to nudge people into making the best decisions for themselves (Thaler and Sunstein 2021).

This step culminated with a codebook that we provide as part of our supplementary material (Matsubara et al. 2023, Online Resource 2). It contains categories and strategies representing all the interventions that we cataloged from Kahneman et al. (2021) and from Münscher et al. (2016) (from now on referred to as behavioral interventions), along with their descriptions and some examples. This forms our final analytical framework (Gale et al. 2013).

3.2.2 The Identification of SEE Interventions

Answering RQ2 also requires the identification of SEE interventions to match the behavioral ones. In Phase 1, described in Section 3.1, we focused on the factors affecting estimates reported in more than one paper. In Phase 2, we focused on any factor, including the ones reported in one paper only. However, we decided to focus on interventions (such as the use of checklists) instead of challenges (such as pressure). While a challenge is a statement of a problem, an intervention describes one or more actions to deal with a problem or improve a situation. We justify this focus because it leads to more practical findings that practitioners can apply in their daily practice. Also, we focus on interventions about SEE more specifically, discarding those from other processes in Software Engineering, such as Software Project Management (like risk assessment), as they are out of scope considering RQ2. Thus, we reduced the number of factors to consider. There were 265 factors in the original SLM (69 factors in the paper and 166 factors in the supplementary material) (Matsubara et al. 2022).

We filtered the factors that represent SEE interventions (Figure 1, Activity 2). This step resulted in 36 papers included in the current study (coming from the first SLM (Matsubara et al. 2022)[5]). Then, we updated the list of

[2] `https://scholar.princeton.edu/kahneman/home`

[3] `https://oliviersibony.com/about/`

[4] `https://hls.harvard.edu/faculty/directory/10871/Sunstein`

[5] The SLM included 131 papers in total.

included papers through forward snowballing. We identified 430 papers (Figure 1, Activity 3), selecting 27 as relevant during the first filter (totaling 63 papers at this step), when we read the titles and the abstracts of papers (Figure 1, Activity 4). In the second filter, we decided to include nine papers after reading their full-texts (Figure 1, Activity 5). We considered the snowballing step relevant because we executed the search step of the previous SLM in 2020 and were aware of new relevant papers (we found all of them in this snowballing step). We ended up with 45 papers to be analyzed in total. The list of selected papers is in Appendix A.

3.2.3 The Analysis Procedures

Next, we moved from reflexive thematic analysis to a codebook approach by using a structured coding framework (Braun and Clarke 2021) (Figure 1, Activity 7), which we described and justified in Section 3.2.1. This means that we changed from an inductive to a deductive (or theoretical) approach to thematic analysis (Braun and Clarke 2006), reflecting our interest in specific aspects of our data (Fereday and Muir-Cochrane 2006). More specifically, we chose the approach of Framework Analysis, with the central idea of using an analytical framework to reduce and summarize the data to support answering the research question (Gale et al. 2013). We adapted the method by using Evidence Profile (EP) and Summary of Qualitative Finding (SoQF) Tables, which are the typical results from the assessment of the strength of evidence using GRADE-CERQual (as we discuss in Section 3.3), instead of their recommended matrix, to avoid redundant data during analysis. We provide the SoQF tables in Section 4 and the EP tables as part of our supplementary material (Matsubara et al. 2023, Online Resource 3).

Next, we wrote down the review findings concerned with the SEE interventions and associated them with each category and strategy from our analytical framework. We interpreted the data, analyzing SEE and behavioral interventions, to identify the extent of their matching and potential gaps. The authors held regular meetings to discuss data charting to categories and strategies, and analyze the codes, descriptions of the review findings, and the chunks from the papers describing the interventions. We constantly checked our codebooks, ensuring the matching of the categories and review findings.

Although our description here is sequential, our approach was iterative. Moreover, we report our findings in Section 4 and provide extensive supporting documentation in the supplementary materials (mapping of latent themes and factors, codebook with theoretical framework, Evidence Profile tables, and quality assessment forms) (Matsubara et al. 2023).

3.3 Strength of evidence

A central issue for systematic literature reviews readers is how much confidence to place in their conclusions and recommendations (Dybå and Dingsøyr

2008). This means that the knowledge of which interventions exist and the existing evidence about them are not enough: we need to understand how confident we can be in this evidence, considering the importance of credible information for an evidence-based discipline Wohlin and Rainer (2021). The GRADE (Grades of Recommendation, Assessment, Development, and Evaluation) system (Guyatt et al. 2011) was proposed as the reference tool for assessing the strength of evidence in SE (Dybå and Dingsøyr 2008). However, for qualitative reviews (as ours), GRADE-CERQual is better-suited (Lewin et al. 2018a). Therefore, we used it to assess the strength of evidence of our review findings (Figure 1, Activity 8).

The evaluation using GRADE-CERQual involved the following steps:

1. Writing the review finding keeping it grounded on the data from the primary studies and focusing on the phenomenon of interest;
2. For each review finding, summarizing relevant data for each dimension;
3. For each review finding, assessing data for each of CERQual's dimensions, looking for concerns, and making a judgment on confidence.

GRADE-CERQual evaluations focus on four dimensions that contribute to an overall assessment of how much confidence the reader of a review can have that the findings represent a phenomenon or one of its aspects (Lewin et al. 2018a). The dimensions are:

- **Methodological limitations**: regards the design and execution of the primary studies that contributed to the review finding (Munthe-Kaas et al. 2018). It requires the choice of a quality assessment tool, and we chose an adaptation of the criteria proposed by Dybå and Dingsøyr (2008), which is inspired by CASP (Critical Appraisal Skills Programme). The adaptation was proposed and used in a previous review (Mendes et al. 2019). We provide quality assessments of the papers as part of our supplementary materials (Matsubara et al. 2023, Online Resource 4).
- **Coherence**: regards how clear and cogent is the fit between the data from the primary studies and the review finding (Colvin et al. 2018). For this dimension, we assessed whether there was (i) any primary study with data contradicting the review finding statement, (ii) any primary study for which it is unclear whether data directly supports the review finding, or (iii) alternative interpretations or explanations for the data (apart from what the review finding states).
- **Adequacy of data**: regards how rich and how much data support the review finding (Glenton et al. 2018). We collected data on the number of studies from which this data comes from, and the number of participants or observations associated with it. The number of participants was especially relevant in the case of controlled experiments. The number of projects/tasks was relevant in the case of data studies. We also assessed the richness of details in the case of qualitative studies.
- **Relevance**: regards the extent that the body of evidence for the review finding applies to the research context, as specified in the research question (Noyes et al. 2018). We collected data on the type of participants, the

publication year of the study, whether the estimation task was realistic, and the range of software engineering contexts included (like business area and company size).

At the beginning of the assessment of one specific review finding, we first assign it high confidence, downgrading it as our assessments raise concerns for the CERQual dimensions we discussed previously. We described our levels of concerns using the categories that Lewin et al. (2018b) proposed:

- No or very minor concerns: unlikely to reduce confidence.
- Minor concerns: may reduce confidence.
- Moderate concerns: will probably reduce confidence.
- Serious concerns: very likely to reduce confidence.

For instance, for a review finding supported by only one paper and very few participants, we had serious concerns about adequacy. Supposing we have no concerns about the other dimensions, we would downgrade the confidence in this finding from high to low — representing two levels of downgrading.

Moreover, evaluators need to describe all their concerns as part of their assessments, typically in an EP table. Evaluators should look for relevant concerns that may reduce confidence and do not need to list very minor ones (Lewin et al. 2018b). We provided such detailed descriptions of our assessments of the review findings in EP and SoQF tables (Lewin et al. 2018b), as part of our supplementary material (Matsubara et al. 2023, Online Resource 3) and in Section 4.2, respectively. The authors held regular meetings to assess the dimensions of each review finding and ensure consistency.

4 Results

In this section, we present the results of our study. Section 4.1 presents the latent themes regarding the **perspective about software effort estimation, thus answering RQ1**. Next, Section 4.2 presents the mapping of SEE interventions to the behavioral ones, thus answering RQ2.

4.1 SEE: Technical Prediction Task and Behavioral Act?

Figure 2 presents the two latent themes we found applying reflexive thematic analysis (Activity 1 in Figure 1) as gray rectangles: SEE as a **technical prediction task** and/or as a **behavioral act**. The gray notes present a brief description of each of them. The rounded blue rectangles represent factors associated with the SEE as a **technical prediction task** theme, and the pink ones represent factors associated with the SEE as a **behavioral act** theme. The purple rounded rectangle is a factor shared by both themes. Figure 2 is not exhaustive: we included some illustrative examples from the map of factors (see Figure 1) — the complete mapping of factors and themes is part of our supplementary material (Matsubara et al. 2023, Online Resource 1).

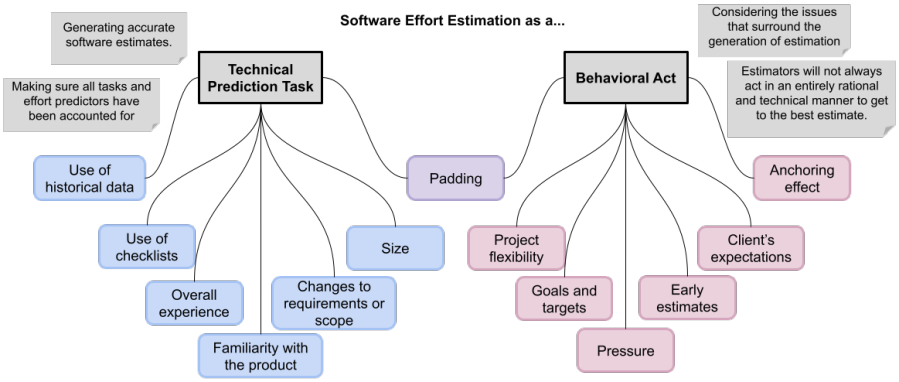

Fig. 2 Latent themes: SEE as a technical prediction task and SEE as a behavioral act.

The SEE as a **technical prediction task** theme aggregates the factors that relate to the issue of generating accurate software estimates from a purely technical point of view. It includes concerns about ensuring all tasks and effort predictors have been accounted for. For instance, size is an effort predictor that estimators must consider (He et al. 2010; Usman et al. 2017; Lagerström et al. 2012). The overall experience of the estimators also matters (Karna and Gotovac 2014), as well as the familiarity with the product (Lee et al. 2011). In addition, situational characteristics also impact estimation accuracy. For instance, changes to requirements and scope can increase estimation error (Layman et al. 2008; Usman et al. 2015; Zapata and Chaudron 2013). On the other hand, one needs to take concrete technical steps to improve our estimation processes, such as using historical data (Furulund and Molkken-stvold 2007; Conoscenti et al. 2019; Rahikkala et al. 2018) and checklists (Furulund and Molkken-stvold 2007; Usman et al. 2018b; Jorgensen and Molokken-Ostvold 2004), to get more accurate estimates. Therefore, the central idea of the theme is that if one can consider all relevant predictors and use the appropriate techniques and methods, one can more accurately predict software tasks and projects.

SEE as a **behavioral act** aggregates the factors that relate to a broader perspective, considering the issues surrounding the estimation generation. It includes factors that bring unexpected influences over the estimates, revealing that there is more to it than the technical side. It also includes factors that reveal that estimators will not always act in an entirely rational and technical manner to get to the best estimate given the information at hand at the estimation moment: they are affected by cognitive and social biases. For instance, a well-documented cognitive bias that can affect software estimates is the anchoring effect (Aranda and Easterbrook 2005; Løhre and Jørgensen 2016; Shepperd et al. 2018). Customer expectations (Jørgensen and Sjøberg 2004), early estimates (Jorgensen and Carelius 2004; Jørgensen and Sjøberg 2001), and goals and targets (Magazinovic and Pernstål 2008; Magazinius et al. 2012) can also affect software estimates, even when unrealistic. Furthermore, studies report pressure for reductions in software estimates (Yang et al. 2008;

Magazinius et al. 2012; Zarour and Zein 2019) and that flexibility in the implementation of requirements allows the creation of a perception of accuracy (Jorgensen and Molokken-Ostvold 2004; Grimstad et al. 2005), enabling the pursuit of the estimate as a target to be hit. Thus, the central idea of the theme is that non-technical issues, but behavioral ones that go beyond the objectives of the estimation process, are also highly influential.

The map of factors had a total of 69 factors, including only those reported in more than one paper (Matsubara et al. 2022). The perspective of SEE as a **behavioral act** aggregates only fourteen of these — a rough indication of how underexplored this perspective is. Thus, in the next section, we focus on the perspective of SEE as a **behavioral act**, investigating the SEE interventions from a behavioral point of view.

4.2 The Matching of Interventions

In **RQ 2**, we asked: "Which interventions from behavioral science are explored in the software estimation context?" To answer it, this section presents the results of the codebook analysis and assessment of the strength of evidence (Activities 7 and 8 in Figure 1, respectively). Figure 3 presents an overview of all categories of behavioral interventions for bias and noise reduction that compose our analytical framework.

Figure 3 shows an upper dark gray rectangle aggregating all categories and strategies under the umbrella super-category of **bias and noise reduction strategies**, derived from Kahneman et al. (2021). The medium gray rectangles in Part a of Figure 3 represent the first set of categories in our framework: **better judges** (for both bias and noise reduction), **debiasing** (for bias reduction), and **decision hygiene** (for noise reduction), as proposed also by Kahneman et al. (2021). We identified each of these categories with a number of the form X. The light gray rectangles represent their associated general strategies, with an identifier in the form of X.Y, where X specifies its parent category and Y represents the general strategy. We omitted specific strategies, presenting them in the coming subsections. They have no identifiers also.

Moreover, the debiasing ex-ante general strategy is not largely explored in (Kahneman et al. 2021). Therefore, we complemented it with the framework of (Münscher et al. 2016), which focus on **choice architecture**. We represented **choice architecture** in Part b of Figure 3 as a dark gray rectangle as it also works as an umbrella super-category aggregating a wide set of general and specific strategies. Therefore, Figure 3 connects ex-ante debiasing with the choice architecture set of strategies with a dotted line. This super-category is composed of three categories: **decision information**, **decision structure**, and **decision assistance**. They are represented and identified as the elements in part "a" of Figure 3, together with their general strategies.

In the following sections, we examine our review findings, as detailed in Section 3.3. They are presented inside Summary of Qualitative Finding (SoQF) tables, identified with an ID, its summary, the set of supporting papers, their

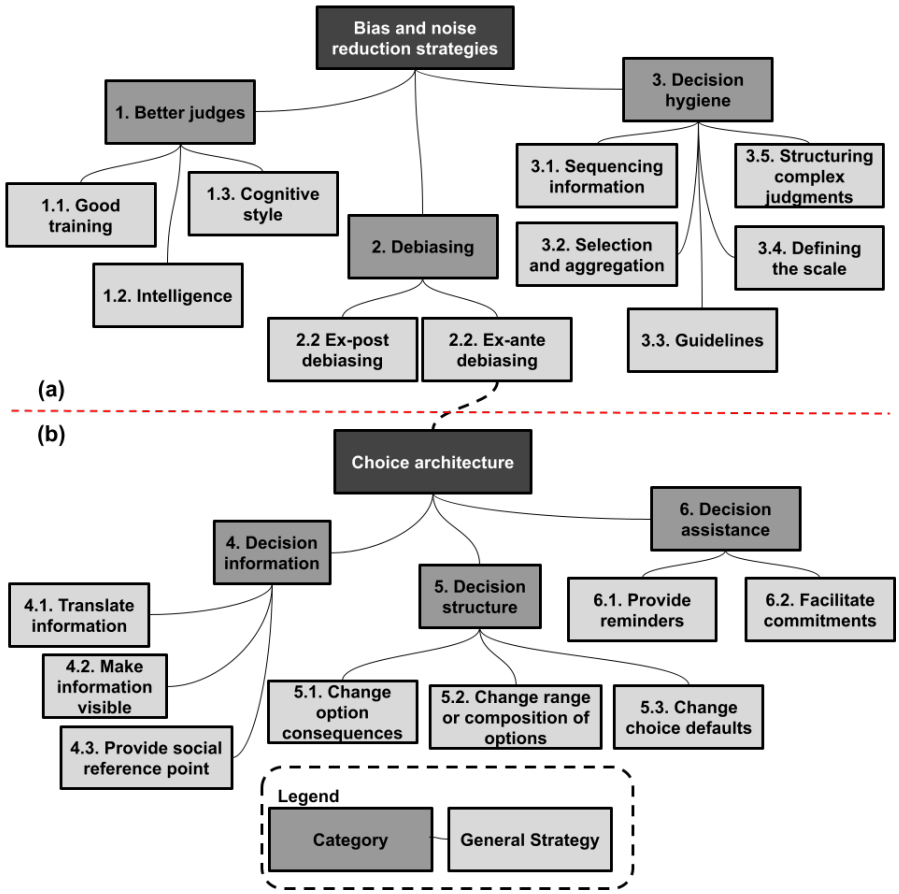

Fig. 3 Overview of categories of behavioral interventions. Each super-category (e.g. bias and noise reduction strategies) aggregates categories (e.g. better judges). Each category aggregates general strategies (e.g. good training). Some general strategies (such as defining the scale) aggregate specific strategies (such as anchored scales, not shown in the figure). The ex-ante debiasing general strategy has a special class of specific strategies so large that we put it as a separate super-category: choice architecture.

CERQual assessment of confidence, and a brief explanation of the CERQual assessment. Moreover, the SoQF tables are structured in accordance with the categories, general and specific strategies that Figure 3 shows. The review findings' IDs have a structure of the type X.Y.Z, where X specifies its parent category, Y represents its general strategy, and Z represents the review finding. We start with strategies for bias and noise reduction, focusing on the **better judges** category in Section 4.2.1. Next, we present the findings regarding the **debiasing** category in Section 4.2.2 and the findings for the **decision hygiene** category in Section 4.2.3. Then, we shift to the findings for the choice architecture interventions in Section 4.2.4.

We also present the assessment of the strength of evidence of the findings using GRADE-CERQual, as stated in Section 3.3. The assessment method results in an overall evaluation of how confident one can be that the review findings depict the investigated phenomenon or some of its aspects (Lewin et al. 2018a). It differs from the quality assessment in SE Systematic Literature Reviews (SLRs) because it does not focus on each primary study but on the review finding resulting from the SLR (Lewin et al. 2018a). The assessment includes the quality of primary studies through the dimension of methodological limitations but is much broader, including how much the results from multiple primary studies agree (coherence dimension), how rich the data supporting the review finding is (adequacy dimension), and how much the context of the primary studies corresponds to our context of interest (relevance dimension). CERQual recommends that all review findings start with high confidence, and as the evaluators identify concerns in each dimension, they can downgrade the confidence level accordingly. Evaluators can indicate no, minor, moderate, or serious concerns in each dimension. Minor or moderate concerns in one dimension do not necessarily lead to downgrading the confidence (Lewin et al. 2018b).

4.2.1 Better Judges

There are three categories for reduction of bias and noise (middle gray rectangles in Figure 3): **better judges**, **debiasing**, and **decision hygiene**. Table 1 shows the strategies that concern the better judges category.

ID	Review finding	Papers	Confid.
1.1 Good training			
1.1.1	Estimation competence improves estimates	14, 29, 38	Moderate
1.1.2	Estimation training improves estimates	4, 29	High
1.2 Intelligence: no findings			
1.3 Cognitive style			
1.3.1	A higher level of interdependence is connected with a higher level of estimation bias and with lower estimates	27	Moderate

Table 1: Summary of Qualitative Findings (SoQF) for the better judges category. It includes a summary of the review finding, together with the contributing papers, the confidence (Confid.), and their GRADE-CERQual assessments. The review findings are also organized by behavioral intervention—notably by the general/specific strategies. General/specific strategies with no matching review finding are marked with **"no findings"**.

According to Kahneman et al. (2021), judgments can be less noisy and less biased when people are well-trained, are more intelligent, and have the

appropriate cognitive style for the judgment task. In other words, judgments improve with **better judges**. One way to get them is through **good training**. This strategy includes knowledge of shared norms and experience and the ability to make and explain judgments confidently.

Two review findings concern good training. First, estimation competence improves estimates (Magazinovic and Pernstål 2008; Rahikkala et al. 2018; Keaveney and Conboy 2006). We have moderate confidence in this finding. No, or only minor concerns for coherence, adequacy, and relevance led us to strong confidence. However, moderate or serious methodological limitations led us to downgrade confidence at one level. Second, estimation training improves estimates too (Yang et al. 2008; Rahikkala et al. 2018). We have high confidence in this finding, as we found no or only minor concerns on coherence, adequacy, and relevance. One study with minor methodological limitations and another with moderate limitations led to no downgrading. Training in estimation focused on developing estimation skills and raising competence are essential to implementing the strategy of getting **better judges** in the SEE domain.

We found no studies investigating issues regarding intelligence in the SEE domain. By intelligence, we mean a multidimensional construct composed of a set of human cognitive abilities (Schneider and McGrew 2018). It can include abilities regarding verbal comprehension/knowledge, breadth, and depth of acquired cultural knowledge, fluid reasoning, visual-spacing processing, and short-term memory, among others (Schneider and McGrew 2018) — some of which might be relevant to a greater extent to SEE than others.

Another relevant issue in the **better judges** category is the cognitive style: the approach to the judgment task (Kahneman et al. 2021). There are many different measures of cognitive styles, and their relevance varies with the judgment task. In the case of SEE, one study investigated the matter, focusing on three variables: self-construal perspectives (interdependent vs. independent), levels of holism (holistic vs. analytical thinking), and degrees of need for cognition (Jorgensen and Grimstad 2012). According to Jorgensen and Grimstad (2012), those who score high in interdependence see themselves on the context of others; those with a high degree of holistic thinking include more of the context in their judgments; those with a higher need for cognition are more fond of engaging with effortful information processing. The review finding states that a higher level of interdependence connects with a higher estimation bias and lower estimates. Interdependence dominated the other variables — holism, and need for cognition (Jorgensen and Grimstad 2012). We have moderate confidence in this finding, as we have no or only minor concerns for all dimensions except adequacy. Moderate concerns for adequacy (only one paper supporting the finding) lowered the confidence at one level.

4.2.2 Debiasing

The second category is **debiasing**, which has two main approaches as Table 2 shows: ex-post or ex-ante. The first works by correcting judgments after they

have been made, while the latter works by intervening before the judgment occurs (Kahneman et al. 2021).

ID	Review finding	Papers	Confid.
2.1 Ex-post debiasing			
2.1.1	Use of padding prevents estimation problems	8, 24, 42, 44	High
2.1.2	The removal of padding is a cause of problems in estimating	14, 15, 41	High
2.2 Ex-ante debiasing: Training decision-makers to overcome biases			
2.2.1	Debiasing workshops reduce the impact of high anchors in productivity estimates	11	Moderate

Table 2: SoQF for the debiasing category.

Ex-post debiasing matches one SEE intervention: *padding*, which involves adding a buffer to the software effort estimate, to deal with unexpected events or to hold back reserves. One review finding states that the use of padding prevents estimation problems (Jorgensen and Molokken-Ostvold 2004; Lederer and Mirani 1990; Glass et al. 2008; Lederer and Prasad 1991; Matsubara et al. 2021b), and we have high confidence in it. No concerns for methodological limitations, adequacy of data, and relevance kept confidence high. Minor concerns about coherence were enough to lower the confidence. Another finding states that its removal is a cause of problems in estimating (Magazinovic and Pernstål 2008; Lederer and Prasad 1995; Lederer and Mirani 1990). We also have high confidence in this, as we have no or minor concerns for all dimensions. Moderate methodological limitations for one study were not enough to downgrade the confidence.

Ex-ante debiasing has two main possibilities: through **choice architecture techniques** (Section 4.2.4) or **training decision-makers to overcome biases**. We matched the latter to one intervention in the SEE domain: *debiasing workshops*, in which estimators are exposed to knowledge about cognitive biases in decision-making and their effects over estimates, raising their awareness (Shepperd et al. 2018). The review finding states that debiasing workshops reduce the impact of high anchors in productivity estimates, and we have moderate confidence in it. We have mostly no or minor concerns for CERQual dimensions. Moderate concerns about adequacy (only one supporting paper supporting the finding) led to downgrading the confidence.

4.2.3 Decision Hygiene

The third category is **decision hygiene**. Table 3 presents the general and specific strategies regarding **decision hygiene**, together with our review findings.

ID	Review finding	Papers	Confid.
3.1 Sequencing information			
3.1.1	The sequence of tasks to estimate impact the results of estimates	2, 30, 34	High
3.2 Selection and aggregation (Averaging x discussion-based)			
3.2.1	Estimates made by only one person hinder estimation results compared with group estimation	1, 3, 5, 29, 45	Moderate
3.2.2	The statistical combination of individual estimates is more optimistic and less accurate than Planning Poker estimates	22, 33, 45	Moderate
3.2.3	The statistical combination of individual estimates is more optimistic than unstructured group-based estimates	40	Low
3.2.4	Planning Poker is more accurate compared with unstructured discussion of estimates	37	Very low
3.2.5	Group consensus estimates using the Delphi method are not better than the statistical combination of individual expert estimates	9	Very low
3.2 Selection and aggregation (Select-crowd)			
3.2.6	The involvement of technical staff improves estimation results	4, 6, 15	High
3.2.7	The lack of customer involvement leads to estimation problems	19	Low
3.2.8	The involvement of mature teams improves accuracy	20	Very low
3.2.9	The involvement of estimators participating in the project improves accuracy	24	Low
3.2.10	The lack of involvement of the person responsible for the task is a reason for error	31	Low
3.2.11	Crowd workers using Crowd Planning Poker perform as well as specialists in estimations	12	Low
3.3 Guidelines: no findings			
3.4 Defining the scale - Anchored rating scales			
3.4.1	Absolute estimation leads to improved estimates compared with relative estimation	13, 32	High
3.4.2	Relative estimation is affected by sequence effects (i.e., the choice of the reference task impacts results)	13, 43	High
3.4.3	The choice of estimate for the reference task in relative estimation is not straightforward	7	Very low
3.4 Defining the scale - Frame of reference training: no findings			

Table 3 continued from previous page			
ID	**Review finding**	**Papers**	**Confid.**
3.5 Structuring complex judgments: no findings			

Table 3: SoQF for the decision hygiene category.

One general strategy is **sequencing information**, by gradually revealing the ones relevant for a given judgment, avoiding early exposures to biasing information. It works by limiting the formation of premature intuitions and protecting the independence of the judgment (Kahneman et al. 2021). The related review finding states that the sequence of tasks to estimate impacts the results of estimates (Grimstad and Jørgensen 2008; Jørgensen 2016; Jørgensen and Halkjelsvik 2020). In other words, there is a *sequence effect*. In the estimation of tasks of similar sizes in a sequence, estimators tend to estimate the target task as larger compared to the reference (first) task (due to a contrast effect). Additionally, in estimating tasks of different sizes in a sequence, a small reference task will lead to a lower estimate of the target task (due to an assimilation effect) — and the opposite happens when the first task is large. No or minor concerns for the dimensions led to a high confidence level in this review finding.

The next general strategy is **selection and aggregation**. Table 3 presents first the findings related to the aggregation of estimates, which according to Kahneman et al. (2021) is composed of five specific strategies:

- **Straight averaging**: the easiest way to aggregate forecasts: averaging them. We also refer to it as the statistical combination of estimates.
- **Select-crowd strategy**: selecting a small set of the best judges, based on the accuracy of their recent judgments, and averaging their forecasts.
- **Prediction markets**: individuals bet on likely outcomes, collectively producing estimates of probabilities of events.
- **Delphi method**: participants submit votes to a moderator and provide reasons for their estimates at each round or respond to the reasons of others. Multiple rounds are allowed.
- **Estimate-talk-estimate**: an adaptation of Delphi, where participants first generate independent (separate and silent) estimates and then justify them. After the explanations, participants make a new estimate, and the result is the average of the individual estimates in the second round.

The primary studies in our review rarely focused on one of these strategies alone. The first review finding states that estimates made by only one person hinder estimation results compared with group estimation (Rahikkala et al. 2018; Moløkken-Østvold et al. 2008; Gandomani et al. 2014; Altaleb et al. 2020; Fægri 2010). Therefore, the evidence in the SEE domain also favors the aggregation of estimates of multiple people over having one single person provide estimates. Moreover, we have high confidence in this finding. We found papers with varying levels of methodological limitations and no concerns for adequacy and relevance. We had moderate concerns about coherence (results

in the opposite direction in one paper and heterogeneity of group-based estimation methods). Concerns about coherence and methodological limitations led to the downgrading of confidence.

We also found a series of review findings comparing the statistical combination of estimates (i.e., straight averaging) with discussion-based methods, such as Planning Poker, Delphi, or unstructured group discussions. In summary, such findings focus on aggregation issues and state that:

- The statistical combination of individual estimates is more optimistic and less accurate than Planning Poker estimates (Mahnič and Hovelja 2012; Gandomani et al. 2019; Moløkken-Østvold et al. 2008). We have moderate confidence in this finding. Only one paper has moderate methodological limitations. We also have no concerns about the adequacy and minor concerns about relevance. Moderate concerns about coherence (results in the opposite direction in one paper) led us to downgrade confidence.
- The statistical combination of individual estimates is more optimistic than unstructured group-based estimates (Moløkken-Østvold and Jørgensen 2004). We have low confidence in this review finding. Although we had no concerns about methodological limitations and coherence, we have serious concerns about adequacy (one paper supporting the finding and few participants) and moderate concerns for relevance (narrow range of contexts and because not so recent paper), leading us to downgrade the confidence in two levels.
- Planning Poker is more accurate compared with an unstructured discussion of estimates (Haugen 2006). We have very low confidence in this review finding. We have no concerns about coherence. However, we have serious concerns about methodological limitations and adequacy (only one paper supporting the finding and an unclear number of participants) and moderate concerns about relevance (narrow range of contexts and not-so-recent paper), leading us to downgrade the confidence in three levels.
- Group consensus estimates using the Delphi method are not better than the statistical combination of individual expert estimates (Arnuphaptrairong 2021). We have very low confidence in this review finding. Although we have no concerns about coherence and relevance, the serious concerns about methodological limitations and adequacy (only one supporting paper and few participants) led down the confidence in three levels.

We also had findings about the selection part of the **selection and aggregation** general strategy. These findings point to the importance of selecting people for the estimation tasks in the SEE domain, stating that:

- The involvement of technical staff improves estimation results (Yang et al. 2008; Rahikkala et al. 2015; Lederer and Prasad 1995). We have high confidence in this finding, as we have no or minor concerns on all dimensions. The serious methodological limitations of only one paper were insufficient to downgrade the confidence.
- The lack of customer involvement leads to estimation problems (Usman et al. 2015). More specifically, the study suggests customer participation

during story sizing sessions to clarify when details are missing. We have low confidence in this finding. We have no concerns about methodological limitations and coherence. However, the serious concerns about adequacy (only one paper supporting the finding and few participants) and the moderate concerns for relevance (unclear context) led to downgrading the confidence.
- The involvement of mature teams improves accuracy (Usman et al. 2018a). We have very low confidence in this finding. We have no concerns about coherence and minor concerns about relevance. The serious concerns about methodological limitations and adequacy (only one supporting paper and few participants) led us to downgrade our confidence.
- The involvement of estimators participating in the project improves accuracy (Jorgensen and Molokken-Ostvold 2004). In other words, it is beneficial to have estimators estimate their own work instead of others'. We have low confidence in this finding. We have no concerns about methodological limitations and coherence. The serious concerns about adequacy (only one supporting paper and few participants) and the moderate concerns about relevance (narrow range of contexts and not so recent paper) led us to downgrade the confidence
- The lack of involvement of the person responsible for the task is a reason for error (Altaleb and Gravell 2019). We have low confidence in this review finding. We have no concerns about coherence and relevance. However, the moderate concerns about methodological limitations and the serious concerns about adequacy (only one supporting paper and few participants) led us to downgrade our confidence.
- Crowd workers using Crowd Planning Poker perform as well as specialists in estimations (Alhamed and Storer 2021). Crowd Planning Poker is an adaptation of Planning Poker where participants are crowd workers with software engineering experience from platforms like Mechanical Turk (Alhamed and Storer 2021). We have low confidence in this review finding. We have no concerns about methodological limitations and coherence and minor concerns about relevance. The serious concerns about adequacy (only one paper supporting the review finding and a few projects) led to a downgrade in the confidence level.

Although all these studies try to address whom to select, they do not mention basing such selection on previous accuracy results, as recommended in the **select-crowd** strategy (Kahneman et al. 2021). We discuss this issue in Section 5.2.3.

Other two specific strategies regarding selection and aggregation are **prediction markets** or the **estimate-talk-estimate** strategies. We found no findings to match them. However, Planning Poker resembles the last strategy, except it is not restricted to one round of justification and uses no averaging of estimates.

The general strategy of **guidelines** aims to decompose a complex judgment into more straightforward assessments on predefined dimensions (Kahneman et al. 2021). The technique guides the estimators by helping them to focus

on the relevant dimensions to consider; and defining how to score, weigh, and add the judgments for each dimension, leading to the overall value (Kahneman et al. 2021). The final decision is a judgment based on evaluating each guideline element, not a straightforward computation. This means that disagreements among raters can occur. In the expert-judgment part of the SEE domain, we matched no findings to guidelines.

Defining the scale is a general strategy with two specific strategies associated: **anchored rating scales** and **frame of reference training**. The first was originally called behaviorally anchored rating scales because they were created for performance ratings (Kahneman et al. 2021). We dropped the "behaviorally" on intention: we do not mean scales to rate behaviors. It concerns improving the rating format with a scale that establishes a common frame of reference. Each degree corresponds to a description of specific situations or behaviors. The second strategy regards training raters to recognize different dimensions. For instance, in the case of performance evaluations, raters are exposed to videotaped vignettes with reference cases, where they can compare their ratings of the cases with "true" ratings given by experts. The scale has anchor points (case scale), so each new rating compares with the anchor cases. Therefore, **frame of reference training** complements **anchored rating scales**. Together, these strategies promote relative judgments. In SEE, relative estimation can be implemented when using story points. For example, estimators choose a story as the frame of reference, and all others are estimated using its value as a reference. It contrasts with absolute estimation (Halkjelsvik and Jørgensen 2018b).

However, in SEE, absolute estimation leads to better estimates than relative estimation (Arifin et al. 2017; Jørgensen and Escott 2022). We have high confidence in this review finding. It is supported by one paper with no methodological limitations and one with moderate. We have no concerns about coherence and adequacy and minor concerns for relevance, which were not enough to rate down the confidence.

Another relevant finding is that relative estimation is affected by sequence effects (i.e., the choice of the reference task impacts results) (Jørgensen 2013; Jørgensen and Escott 2022). For this finding, we have high confidence due to no concerns for methodological limitations and only minor concerns for the other dimensions. Łabędzki et al. (2017) also found that the choice of estimate for the reference task in relative estimation is not straightforward. We have very low confidence in this finding. Although we have no concerns about coherence, serious concerns about methodological limitations and adequacy (only one supporting paper and an unknown number of participants) and moderate concerns about relevance (lack of information about participants) hinder confidence in the finding.

The last general strategy is **structuring complex judgments**. According to Kahneman et al. (2021), it has three guiding principles (that are not perfect for all cases and must be adapted to the judgment situation):

- **Decomposition:** Breaks down the decision in its components. It is like the dimensions in guidelines, aiding the judge to focus on what is important;
- **Independence:** Information on each component is collected independently to avoid that judgment on one dimension influencing the judgments of others (halo effect);
- **Delayed holistic judgment:** Intuition is not excluded. It is delayed until judges have all the information they need. Also, the judgment is not a computation of a final score but rather a decision holistically weighted based on the information collected independently on each relevant component.

We found no studies to categorize as proposing something similar to **structuring complex judgments**.

4.2.4 Choice architecture

As we mentioned earlier, a class of ex-ante debiasing is through choice architecture techniques aimed at reducing the effect of biases or even enlisting biases in the direction of better decision-making (Kahneman et al. 2021). Therefore, we relied on a framework that breaks up and organizes these techniques, as Figure 3 (Part b) shows. The framework of Münscher et al. (2016) divides the techniques into three categories, represented by dark gray rectangles in Figure 3 (Part b): **decision information**, **decision structure**, and **decision assistance**. We start with decision information, as Table 4 shows.

ID	Review finding	Papers	Confid.
4.1 Translate Information - Reframe			
4.1.1	The use of the alternative format leads to more optimistic estimates compared with the use of the traditional one	27	Moderate
4.1.2	Estimating ideal effort followed by the most likely effort produces more realistic estimates compared to estimating in the opposite order	17	Moderate
4.1 Translate Information - Simplify: no findings			
4.2 Make information visible - Make own behavior visible (feedback)			
4.2.1	Bringing attention to previous estimation performance prevents estimation problems	10, 15, 16	High
4.2.2	Structured lessons learned do not improve estimation accuracy or overconfidence when estimating own tasks	28	Moderate
4.2 Make information visible - Make external information visible			
4.2.3	The annotation of user stories leads to more accurate and less biased estimates compared with the use of Planning Poker alone	26	Low

Table 4 continued from previous page			
ID	**Review finding**	**Papers**	**Confid.**
4.2.4	The anticipation of project' participants' skills improves estimation	4, 6, 20, 23, 41	High
4.3 Provide a social reference point - Refer to descriptive norm: no findings			
4.3 Provide a social reference point - Refer to opinion leader			
4.3.1	The lack of careful examination of estimates by management is a reason for inaccuracy	15	Very low

Table 4: SoQF for the decision information category.

A general strategy in the **decision information** category is to **translate information**: changing the format of the presentation, without changing the content. Two of its specific strategies are to **reframe** and to **simplify**. The first concerns translating the existing information by presenting it in different ways (formally, logically, or mathematically). It requires no strict equivalence: linguistic re-definitions of the same decision problem also count (Münscher et al. 2016). There is one SEE intervention that regards that: the request format. Researchers investigated the format of how estimates are asked for, comparing a traditional format request ("How much effort is required to complete X?") with an alternative one ("How much can be completed in Y work-hours?"). The corresponding review finding states that using the alternative format leads to more optimistic estimates than using the traditional format (Jørgensen and Halkjelsvik 2010). We have moderate confidence in this review finding. We have no or minor concerns for most dimensions. Moderate concerns about adequacy (only one paper supporting the finding) led to downgrading one level.

Another finding concerning **reframe** involves redefining the type of estimate given by estimators. This involves two steps. First, the receiver of an estimate needs to frame the first estimate that an estimator provides as an ideal effort value. Next, the receiver has to ask the estimator for a second estimate, framing it as a most likely effort value. Researchers found that this redefinition produces more realistic estimates than the opposite order (most likely first, ideal second) (Jørgensen 2011). We have moderate confidence in this review finding. We have no or minor concerns for most dimensions. However, moderate concerns about adequacy (only one paper supporting the finding) led us to downgrade one level.

We found no interventions regarding the specific technique called **simplify**, which concerns reducing the burden of cognitive effort needed to process the information available and increase its usefulness (Münscher et al. 2016).

Another general strategy is to **make information visible** by providing easier access to information that is usually invisible. The first of its specific strategies is to **make own behavior visible (feedback)**. When feedback is infrequent or disconnected from decision-making, behavioral consequences

remain invisible (Münscher et al. 2016). In the SEE domain, bringing attention to previous estimation performance (through recalling past effort usage or providing feedback) prevents estimation problems (Lederer and Prasad 1995; Jørgensen et al. 2007; Hughes 1996). We have high confidence in this finding. We found variations among studies, from no methodological limitations to moderate ones. We have no or minor concerns about all other dimensions. Only one paper with moderate methodological limitations is not enough to downgrading.

Another review finding involves a more complex type of feedback: structured lessons learned, with feedback on estimation error, the realism of uncertainty assessments, and mean estimation error from all previous tasks, among others. Such structured lessons learned do not improve estimation accuracy nor reduce overconfidence when estimating own tasks (Jørgensen and Gruschke 2009). We have moderate confidence in this finding, as we have no concerns about most dimensions. However, moderate concerns about adequacy (only one supporting paper) lower the confidence in the evidence.

The second specific strategy of **make external information visible** focuses on existing relevant external information for decision-making. One related SEE intervention is the annotation of stories, which involves using Interaction Room annotation of user stories before estimating with Planning Poker. Participants' annotations are affixed on story cards and can mark: high load, flexibility, reliability, time limit, usability, complexity, and uncertainty (Grapenthin et al. 2016). For instance, a story marked with the time limit annotation must observe a fixed deadline or be executed within specified time constraints. Therefore, it is a technique that can aid in making information visible, supporting SEE. One review finding concerns it: the annotation of user stories leads to more accurate and less biased estimates than Planning Poker alone (Grapenthin et al. 2016). We have low confidence in it. The only supporting study has moderate methodological limitations. There are no concerns about coherence. Moderate concerns about adequacy (only one supporting paper) and relevance (participants were students) lower the confidence in the finding.

The anticipation of project' participants' skills involves knowing who will execute the estimated tasks/projects and their characteristics (skills and abilities). The corresponding review finding states that the anticipation of project' participants' skills improves estimation (Yang et al. 2008; Rahikkala et al. 2015; Usman et al. 2018a; Matos et al. 2013; Lederer and Mirani 1990). Although most papers report that the lack of such anticipation is a reason for errors, one reports that its presence increases optimism (Usman et al. 2018a), an outcome typically considered to lead to estimation problems. However, the estimates where such anticipation happened were also more accurate (Usman et al. 2018a), suggesting that such optimism improved results in this specific case. The participating company in the study produced estimates at two stages: quotation and analysis estimates. The increase in optimism was noticed in analysis estimates when estimators had more detailed information — explaining why they were more accurate despite being also more optimistic.

Therefore, in some contexts, postponing the information of who will be responsible for executing a task is beneficial for reducing the optimism bias — which is relevant when such bias leads to inaccuracies. No or only minor concerns in all dimensions leads to high confidence in the finding.

The third general strategy of decision information is to **provide a social reference point**, which concerns giving access to other people's behavior as a social reference (Münscher et al. 2016). The first of its specific strategies is to **refer to descriptive norm**: the observable behavior of other people (what they do), contrasting with injunctive norm (what they should do). We found no interventions in the SEE domain that we could match this.

The second specific strategy is to **refer to opinion leader** because highly respected people can influence the behavior of others. Therefore, changing their behavior is powerful (Münscher et al. 2016). The finding from the SEE domain matched to this behavioral intervention states that the lack of careful examination of estimates by management is a reason for inaccuracy. Lederer and Prasad (1995) provide data for this finding, showing that accuracy decreases when management fails to review estimates. Therefore, it is not about how the managers act when estimating but refers to their examination of estimation results. We have very low confidence in this review finding. The supporting study has minor methodological limitations. We have no concerns about coherence. However, moderate concerns about the adequacy of data (only one supporting paper) and serious concerns about relevance (very old paper) led to lower confidence.

The next category of strategies that we examine is the **decision structure**. Table 5 presents the related review findings.

ID	Review finding	Papers	Confid.
5.1 Change choice defaults - Set no-action default			
5.1.1	Shorter time frames lead to more optimistic estimates compared to larger time frames	36	Moderate
5.1.2	Asking for estimates using a lower granularity unit leads to lower estimates compared with higher granularity units	30, 35	High
5.1 Change choice defaults - Use prompted choice: no findings			
5.2 Change range or composition of options - Change categories/grouping of options			
5.2.1	The use of Fibonacci scale leads to lower estimates compared to linear scales	39	Moderate
5.3 Change option consequences - Connect decision to benefit/cost: no findings			
5.3 Change option consequences - Change social consequences of the decision: no findings			

Table 5: SoQF for the decision structure category.

The first general strategy of decision structure is to **change choice defaults**. Defaults are pre-selected options that give people the freedom to choose another alternative if they wish to. Their choice is impactful because many people accept the default without giving much thought to other options (Münscher et al. 2016). One of the specific strategies is to **set no-action defaults**. Considering that even the size of a unit can serve as a default (Münscher et al. 2016), we associated two factors with this strategy: time frame size and the unit effect. The first involves situations where people use the alternative format ("How much can be completed in Y work-hours?"). In such situations, shorter time frames lead to more optimistic estimates compared to larger time frames (Jørgensen and Halkjelsvik 2010; Halkjelsvik and Jorgensen 2011). Moreover, the time frame size will act as a default on how much time is expected for the team to work to achieve a deliverable result. We have moderate confidence in this review finding. We have no or minor concerns for most dimensions. Moderate concerns about adequacy (only one paper supporting the finding) led to lower confidence in the finding

The unit effect regards the unit used when asking estimators for an estimate. For instance, work-hours is a lower granularity time unit than work-days. Using a lower granularity unit leads to lower estimates compared with higher ones (Jørgensen 2016, 2015). Again, the unit acts as a default on how much work a task is expected to take: hours, days, months, or years. No or only minor concerns in all dimensions leads to high confidence in this finding.

The second specific strategy when working to **change choice defaults** is to **use prompted choice**: forcing people to decide (no defaults). It can be helpful when the target group is heterogeneous (Münscher et al. 2016). We found no interventions about it in the SEE domain.

The first general strategy in the decision structure category is to **change the range or composition of options** by changing what is presented (the alternatives) to decision-makers. The only specific strategy it has is **change categories/grouping of options**. The SEE intervention that we associated with this is the *use of Fibonacci scale*, which regards substituting linear scales for Fibonacci ones when estimating. Therefore, it involves rearranging the options of values that estimators can choose. In SEE, the use of the Fibonacci scale leads to lower estimates compared to linear scales (Tamrakar and Jørgensen 2012). We have moderate confidence in this finding. We have no concerns about most dimensions. However, moderate concerns about adequacy (only one paper supporting the finding) and relevance (most participants are students) lower the confidence in the evidence.

The last strategy in the decision structure category is to **change option consequences** while ensuring that such changes are insignificant from a rational perspective. We found no SEE interventions for any of its specific techniques. The first strategy, **connect decision to benefit/cost**, involves connecting the desired behavior to a small benefit or an undesired behavior to a small cost. The second, **change social consequences of the decision**, regards connecting a choice with the consequence to be regarded more positively or negatively by others (Münscher et al. 2016).

We removed one general strategy from this category (according to the original framework) from our analysis (**change option-related effort**) because it regards physical and financial effort. The techniques regarding cognitive effort, which is the one relevant for SEE, are covered by the decision information category (Münscher et al. 2016).

Next, we examine the last category in the choice architecture framework: **decision assistance**. Table 6 presents its findings.

ID	Review finding	Papers	Confid.
6.1 Provide reminders			
6.1.1	The use of checklists improves estimation results	18, 20, 24	High
6.2 Facilitate commitments - Support self-commitment: no findings			
6.2 Facilitate commitments - Support public commitment			
6.2.1	The lack of justifications of estimates is a reason for estimation error	24	Very low

Table 6: SoQF for the decision assistance category.

Its first general strategy is to **provide reminders** that make desired options more salient or suppress cues that remind people of less desired ones (Münscher et al. 2016). The closest SEE intervention to match it is the *use of checklists*: using a list of items that document the factors and activities that estimators should consider during estimation. The review finding indicates that using checklists improves estimation results (Usman et al. 2018b; Furulund and Molkken-stvold 2007; Jorgensen and Molokken-Ostvold 2004). No or minor for all dimensions leads to high confidence in this finding.

Another general strategy identified is to **facilitate commitments**, which helps to deal with self-control problems. Deviations lead to cognitive dissonance or a need to justify the behavior to others (Münscher et al. 2016). The first specific strategy associated with it is to **support self-commitment** through arrangements to help decision-makers to fulfill a plan. We found no SEE intervention associated with it. The second specific technique is to **support public commitment**, creating possibilities for and supporting the commitments in front of others; also creating external pressure and, possibly, negative consequences in case they are broken. The commitment must be voluntary, preserving freedom of choice (Münscher et al. 2016). The intervention we found in the SEE domain is the justification of estimates. Such justifications can act as public commitments to fit the estimate, especially when the estimator is also responsible for executing the estimated task or delivering the associated product. Our review finding states that the lack of justifications for estimates is a reason for estimation error (Jorgensen and Molokken-Ostvold 2004). We have very low confidence in this review finding. Although we have no concerns about methodological limitations and coherence, serious concerns about adequacy (only one paper supporting the finding and few participants)

and relevance (narrow range of contexts and not so recent paper) lower the confidence.

5 Discussion

Section 5.1 discusses more on the impact of the perspectives of SEE as a **technical prediction task** and as a **behavioral act** to SEE practice and research, in an answer to RQ1: What are the latent themes in SEE interventions?. We also discuss our review findings presented in Section 4.2 answering RQ2: Which interventions from behavioral economics are explored in the software estimation context? This makes explicit the link between such interventions. Finally, we discuss the confidence levels of review findings to stress how to improve the current research about SEE.

5.1 Acknowledging the Behavioral Act

From a rigorous perspective, the estimation process includes only the generation and aggregation of software estimates. However, focusing research efforts entirely on it adopts a myopic view. It assumes that software estimation is a rational prediction task only — a view we refer to as **software estimation as a prediction task**, inspired by previous definitions of software estimation (Kitchenham and Linkman 1997). The focus is entirely on the technicalities of estimating as the main tool to reach accuracy: the effort predictors to consider (as size), adding technical steps to the estimation process (as the use of historical data), or even accounting for expected events that can lead to a higher need of effort (as changes to requirements or scope). According to this perspective, estimating has a well-defined target: an actual value that one wants to predict — and all one has to do is act objectively to find it. Like in any other forecasting domain, this perspective assumes estimating should be an objective activity that builds upon facts, sound reasoning, and methods (Petropoulos et al. 2022).

We argue that one should go beyond the technicalities of generating an estimate and consider the (i) human and (ii) social aspects that can impact the results, either negatively or positively. Regarding (i), software estimates are created by humans and therefore are affected by humans' "*bounded rationality*", which is the idea that human judgment and decision are determined not only by people's goals in a task and the characteristics of the external world, but also by people's severely limited knowledge of the world and limited abilities to "*evoke that knowledge when it is relevant, to work out the consequences of their actions, to conjure up possible courses of action, to cope with uncertainty (including uncertainty deriving from the possible responses of other actors), and to adjudicate among their many competing wants*" (Simon 2000).

Regarding (ii), any forecast is created in a social setting and is subject to politics and personal agenda (Petropoulos et al. 2022). The results Section 4.1 presents show that SEE is no different. We term this view **software**

estimation as a behavioral act, following the idea of Behavioral Software Engineering (Lenberg et al. 2015). One can and should use other tools to reach accuracy, considering this activity's human and social sides. This perspective embraces the fact that the generation of estimates is affected by cognitive biases. It also embraces the surrounding activities of the generation of estimates, acknowledging that the reception of an estimation request (such as its format), the communication of the estimates, and their uses can also lead to biases and other social issues (such as pressure to change estimates). Therefore, our work intends to enrich the SEE research and practice landscape, calling attention to the fact that SEE is more than a purely technical prediction task: it is also a behavioral act. Ultimately, that means that improving estimation methods is not enough: practitioners need behavioral interventions to address the human and social sides too.

Implications for practice and research: SEE is more than a purely technical prediction task: it is also a behavioral act, meaning its results are affected by cognitive and social biases. Behavioral interventions are tools one can use to address the SEE human and social sides in the path to improve estimation results.

5.2 Embracing Behavioral Interventions

During the second phase of our research, we aimed to answer RQ2: Which interventions from behavioral science are explored in the software estimation context? Our results show that the SEE interventions cover several behavioral ones, although some remain to be investigated. One can use such interventions to design or improve SEE practices, also considering that some of them might not be feasible or useful in the SEE domain.

One important observation from the strength assessment of the evidence (as in Section 4.2) is that we had high or moderate confidence in many of our review findings. This means we are confident that they represent the phenomenon we are interested in: the effect of SEE interventions (as matched to behavioral interventions) in estimation results — either improving or harming them. Ultimately, researchers investigate interventions not only to generate evidence or identify gaps but also to draw recommendations to practitioners. GRADE supports both the grading of evidence and also of recommendations, considering how confident we are that the benefits of one intervention outweigh its undesired effects (Andrews et al. 2013a). We should base such confidence assessment on the confidence of estimates of effects (both desirable and undesirable ones), among other things (Andrews et al. 2013b). However, we do not have such estimates considering the qualitative nature of our review and many of the primary studies that support our findings. That is one of the reasons for using GRADE-CERQual instead.

Moreover, GRADE-CERQual lacks instructions for rating recommendations for practice. Therefore, the recommendations we made for practitioners in this section are based solely on the review findings we rated with high and

moderate confidence — and Section 5.3 discusses more on the findings rated as low or very low confidence. Finally, our recommendations for practice can be considered at the same level as the discretionary ones described in the GRADE guidelines (Andrews et al. 2013a): to be used at the discretion of the software organizations and professionals using this review results to inform their improvement decisions about their estimation processes. This recognizes that their adoption can vary among organizations and individuals. We also refrain from making any recommendations from findings where our results collide with our theoretical framework, assuming SEE can benefit from more research in such matters.

5.2.1 Better Judges

As Section 4.2.1 shows, the **better judges** category was explored mostly regarding good training and cognitive style. We have moderate and high confidence that estimation competence and training positively impact estimation results, respectively — as in Review Findings 1.1.1 and 1.1.2 in Table 1. Such results show the importance of **good training** to get superior estimation results in the software industry practice. In addition, the specific results of the primary studies mention training about estimation methods (Yang et al. 2008) and practices (Rahikkala et al. 2018), suggesting these are relevant as content for estimation training. However, the studies did not report the specific practices to explore. These practices probably vary in importance from one organization to another.

We could not find any study focused on the impact of **intelligence** on estimation results, although it is part of the **better judges** category. This can represent a gap because intelligence is a multidimensional construct (Schneider and McGrew 2018), and some of its dimensions can be especially relevant for software estimation. For instance, fluid reasoning is a broad ability defined as using deliberate and controlled procedures for solving new problems that one cannot solve with previous known schemes or habits (Schneider and McGrew 2018). Given that many software projects are new to people building them, it seems credible that narrow and specific abilities concerning fluid reasoning can play a role in SEE and would be worth investigating.

Regarding **cognitive style**, interdependence is a characteristic that impairs estimation results. It was connected with a higher effect of the anchoring bias and with lower estimates, with moderate confidence — see Review Finding 1.1.3 in Table 1. That happened possibly because individuals scoring high on interdependence may look to have more socially desirable behavior (Jorgensen and Grimstad 2012). In this context, influential stakeholders, like managers and customers, may desire lower estimates.

However, more research on cognitive styles can also benefit software estimation results. The only study about it explored only three variables: interdependence, holism, and need for cognition (Jorgensen and Grimstad 2012). Other variables representing different cognitive styles can be relevant, but we remain unaware of them due to a shortage of studies. For instance, actively

open-minded thinking can lead to better results in the political forecasting domain (Mellers et al. 2015a) and in estimating uncertain quantities (Haran et al. 2013). It involves the disposition to examine issues from a multitude of perspectives instead of only generating arguments to support a favored belief (Svedholm-Häkkinen and Lindeman 2018). We can assume this kind of thinking style can impact software estimation, but we found no studies in the SEE context.

Therefore, more research on what makes people **better judges** — especially regarding intelligence and relevant cognitive styles in the SEE domain — can help in the identification of what to train people in and what to look for when selecting estimators.

Recommendations for practice:

- Train estimators in software estimation good practice to increase their competence. Obs.: No undesirable outcomes. Potentially high resource usage.
- Select people with lower interdependence characteristics. Obs.: Unclear undesirable outcomes. Requires knowledge and resources (like access to scales) to assess interdependence.

Recommendations for research:

- Investigate the impact of intelligence on estimation results.
- Investigate other cognitive style variables beyond interdependence that can be relevant to the SEE context. Assess their impact on estimation results.

5.2.2 Debiasing Interventions

Section 4.2.2 shows that SEE interventions cover all types of debiasing interventions from our theoretical framework. This seems to answer the call of SE researchers for more studies proposing and investigating debiasing techniques, instead of studies that only demonstrate the existence of biases (Mohanani et al. 2020). Our findings of debiasing interventions show that ex-post debiasing, in the form of padding, can be a very effective intervention in the SEE domain, with high confidence — see Review Findings 2.1.1 and 2.1.2 in Table 2. This can be especially interesting for larger software projects, where it seems safe to say people are prone to underestimate: there is an overoptimism with a median time overrun of 20% (Halkjelsvik and Jørgensen 2018a). A recommendation for padding large software projects seems to be appropriate.

In addition, an **ex-ante debiasing** intervention is **training decision-makers to overcome their biases**. It is helpful in the SEE context with moderate confidence, as in Review Finding 2.2.1 in Table 2 — enough to make a recommendation of this kind of training. Still, it can be more challenging than it seems (Kahneman et al. 2021). For instance, the debiasing workshop for reducing biases in the SEE domain successfully reduced the anchoring effect (Shepperd et al. 2018).

One problem is that debiasing interventions focus on specific biases. However, in a given situation, multiple biases may be at play. In addition, it might

be hard to know precisely which biases are these and in which directions people are biased to (Kahneman et al. 2021). So, although we know that we tend to underestimate larger projects (Halkjelsvik and Jørgensen 2018a), the issue of what is the exact set of psychological biases that produce this effect remains to be empirically investigated. Our current answer for SEE is currently partial — for instance, we know software estimates are affected by biases such as the anchoring bias and overconfidence, as well as antecedents of psychological biases like optimism (Mohanani et al. 2020). Moreover, this answer probably varies from one company to another in practice.

Recommendations for practice:

- Use padding for debiasing in large projects. Obs.: This can lead to overestimation, which is an undesirable outcome that decreases accuracy. Low resource usage and easy to implement. However, it can be hard to decide how much to pad to avoid overestimation.
- Train estimators on biases that can affect estimates to avoid anchoring on high productivity values. Obs.: Unclear undesirable outcomes. Potentially high resource usage. It requires specialized knowledge and resources for training.

Recommendations for research:

- Other than anchoring bias and overconfidence, investigate additional psychological biases affecting software estimates.

5.2.3 Decision Hygiene Interventions

Decision hygiene interventions aim in reducing noise and can be effective tools when there is a knowledge shortage about the specific psychological biases affecting a task (Kahneman et al. 2021), which is the case of SEE. They include interventions like teaming, training and selecting the best forecasters, which reduce forecasting error by reducing noise instead of bias (Satopää et al. 2021). Moreover, as Section 4.2.3 shows that SE researchers are already investigating many interventions addressing noise in SEE.

To start with, we have results regarding **sequencing information**. More specifically, we have high confidence in one review finding showing that the sequencing of tasks to estimate impact estimation results, considering their size — see Review Finding 3.1.1 in Table 3. Results suggest that relative estimation can benefit from using medium-tasks first, as they would lead to smaller overall estimation bias (Jørgensen and Halkjelsvik 2020). However, although it concerns sequencing, this finding is not truly about sequencing *information*. We found no research regarding sequencing information presented to estimators, even though previous research results from the SEE context suggest that keeping some information from estimators can be beneficial, such as any information regarding expectations from customers (Jørgensen and Sjøberg 2004), and other kinds of irrelevant and misleading information (Jørgensen and Grimstad 2011).

Regarding **selection and aggregation**, aggregating estimates from multiple individuals is better than relying on individual estimates. These findings holds in the SEE domain with high confidence — see Review Finding 3.2.1 in Table 3. We also found studies investigating varied aggregation strategies: statistical combination, unstructured group discussions, Planning Poker, and using the Delphi Method. However, looking closer at the review findings, some issues about the best way to aggregate software estimates arise. For instance, when compared to the statistical combination of individual estimates, one review finding (3.2.2) favors Planning Poker, while another (Review Finding 3.2.5) states there is no evidence that the Delphi method is superior. That might come as a surprise and a contradiction, given the similarities between Planning Poker and the Delphi method. However, the similarity is not sameness and implies differences to some extent, such as the total avoidance of face-to-face interaction through several iterations, keeping estimators anonymous trough the whole process, which is part of Delphi method Moløkken-Østvold and Jørgensen (2004) but not required in Planning Poker. These differences might be sufficient to secure the diverging review findings regarding these two group discussion techniques when compared to the averaging of individual estimates. We discuss more in Section 5.3, focusing on which aspects of the current research about the different aggregation strategies need to improve to make us confident enough to recommend one of them to practitioners over the others.

Regarding the unexplored aggregation strategies from behavioral economics, the **estimate-talk-estimate** might be a lighter alternative (or variation) for Planning Poker. It can provide a solution for inexperienced teams, where Planning Poker can lead to slow justifications rounds, and long estimation sessions (Matsubara et al. 2021b). In addition, **prediction markets** can provide an alternative method to be empirically evaluated in the SEE domain, especially when estimating confidence levels associated with the estimation values. They are successfully employed in various domains, including presidential elections (Mann 2016). How to structure a prediction market for the SEE domain would be an interesting question, especially considering that people can alter the execution of a project to fit an estimate (Grimstad et al. 2005; Lederer and Prasad 1995).

As for the selection part of **selection and aggregation**, studies in the SEE domain indicate the importance of selecting appropriate people to participate in the SEE activities. For instance, we have high confidence that involving technical staff in estimation improves results — see Review Finding 3.2.6 in Table 3 in Appendix. We also have review findings about the involvement of customers, mature teams, people participating in the project, people responsible for the task, or even crowd workers for Planning Poker. Therefore, researchers investigated whom to select to participate in estimation based on their role and involvement with the task. Selection based on previous accuracy results, as recommended in the **select-crowd** strategy (Kahneman et al. 2021), remains unexplored. In the geopolitical forecasting domain, the selection of superforecasters based on their accuracy led to the best results compared

to training people (in cognitive debiasing) and teaming (grouping) forecasters (Mellers et al. 2015b; Satopää et al. 2021). Therefore, SEE researchers should consider more thoughtfully the **select-crowd strategy** even though, in practice, often the people making the estimates are the ones responsible for the task, and there might be little room for selecting other people. Moreover, it might not be easy because measuring accuracy in the SEE domain is not a straightforward task, especially due to the "moving target" problem (Matsubara et al. 2022).

We had no findings regarding the strategy of **guidelines**. That possibly happened because of our focus on expert judgment estimates. When looking for the broader literature on SEE, it has guidelines-based methods. For instance, we can consider that COSMIC Function Points propose dimensions: the data movements. Their measurement manual establishes what a data movement is and the procedures to get from them to the functional size of functional processes (Organization 2021).

Another interesting strategy is **defining the scale**, with the specific strategies of **anchored rating scales** and **frame of reference training**. Relative estimation employs the overall idea of anchored rating scales, through story points, for instance, (Halkjelsvik and Jørgensen 2018b). After all, it involves choosing a reference case (a story or a task), estimating it by giving a certain amount of story points and then estimating the remaining cases by comparing with it (Cohn 2005; Halkjelsvik and Jørgensen 2018b). Contrary to the results in behavioral economics, in the SEE domain, the use of relative estimation did not lead to improvements compared with absolute estimation — and we have high confidence in this finding — see Review Finding 3.4.1 in Table 3. We argue that a possible explanation for this result is that the SEE community has not completely explored the anchored rating scales strategy. Using only one story as a reference case may not be enough. Researchers can investigate the impact of using multiple stories as references, representing multiple values on a more sophisticated scale. In addition, estimation results could benefit from integrating frame of reference training to such a scale. Perhaps, companies and teams can use a few anchor stories, tasks, or requirements to support the training of their estimators. It can be complex and time-consuming, requiring customization to the organization/unit and constant updates of cases (Kahneman et al. 2021). However, noisy and biased estimation is also costly.

We also have no findings regarding structuring complex judgments. Kahneman et al. (2021) originally discussed this strategy for hiring decisions: a type of evaluative judgment — contrasting with predictive judgments, the type in which we classify software estimates. The structuring they propose is based on the decomposition on relevant dimensions, independence of judgment of each dimension (to avoid judgments in one component to affect the judgment of others), and delayed holistic judgment (not a computation, but a decision after gathering all information). SEE researchers can use these guiding principles to find ways to improve expert judgment estimation.

Recommendations for practice:

- Sequence tasks to estimate, putting medium ones first in the estimation sequence. Obs.: Potentially high resource usage and hard to implement. It requires pre-processing of tasks to estimate, which can be time-consuming.
- Keep irrelevant and misleading information, such as clients' expectations and future opportunities, from the estimators. Obs.: Potentially high resource usage and hard to implement. It requires pre-processing of tasks to estimate, which can be time-consuming. It also requires some protection of the environment to avoid exposition to such information, such as avoiding contact with customers.
- Aggregate estimates from multiple individuals instead of relying on only one person. Obs.: Potentially high resource usage.
- Involve technical staff in estimating. Obs.: Medium to high usage of valuable resources.

Recommendations for research:

- Investigate how to structure prediction markets for SEE and how worthy they are in improving estimation results.
- Propose and investigate the impact of more sophisticated anchored rating scales associated with a frame of reference training on estimation results when using relative estimation.

5.2.4 Choice Architecture Interventions

Regarding choice architecture techniques, most of the research in the SEE domain concentrates on the decision information category, with SEE factors covering almost all its general strategies — as Section 4.2.4 shows. Regarding the **reframe** strategy, the findings indicate that reframing how one asks for estimates and the type of estimate one is asking for can bring benefits for the estimation process — and we have moderate confidence in these findings (see 4.1.1 and 4.1.2 in Table 4). More specifically, the current results support a recommendation of using the traditional format ("How much effort is required to complete X?") and asking for a second estimate, portraying it as the most likely estimate of effort and the first one that the estimator gave as an ideal estimate.

In addition, we found no results for the specific strategy of **simplify**: reducing the cognitive burden to process existing information. In the case of SEE, the information to process can be the descriptions of tasks to estimate, the current systems' state (in the case of maintenance tasks), and the available resources. SEE researchers can investigate how to convey this information better to make it simpler to understand for estimation purposes.

As for **making own behavior visible**, the findings indicate that bringing attention to previous estimation performance prevents estimation problems, (through recalling past effort usage or providing feedback). We had high confidence in this finding — as in Review Finding 4.2.1 in Table 4. However, another review finding shows that complex feedback (through structured lessons learned, for instance) does not yield better results.

Regarding **making external information visible**, one review finding involves the annotation of user stories during Planning Poker: an intervention that improves estimation results. Another related finding for which we have high confidence is the anticipation of project' participants' skills — see Review Finding 4.2.4 in Table 4. These findings reveal the kind of information that estimators can benefit from: about the task and about people performing it. With information about the task, estimators can better grasp the dimension of the work. With information about people performing the task, they can have better productivity expectations, increasing accuracy for time estimates.

We also had a finding about to **refer to opinion leader**, stating that the lack of careful examination of estimates by management is a reason for inaccuracy. However, this technique can backfire in the estimation context: it can lead to pressure to reduce the estimates. For instance, management pressure is one of the top reasons for intentional increases in estimates (Magazinius et al. 2012). Moreover, agile methods that are now widely spread recommend self-organizing teams, which are supposed to have external autonomy, i.e., protection against the influence from external parties, including management (Hoda and Murugesan 2016). This sounds incompatible with managers examining estimates. The study supporting this finding was published in 1995, years before the rise of agile methods. Thus, it represents part of the management mentality of a much different time than the one we now live.

We found no results for the specific strategy of to refer to descriptive norm, which regards depicting the observable behavior of other people: what they do. This strategy can be helpful in spreading a target behavior, such as politely resisting pressure over estimates.

The category of decision structure is less explored in the SEE domain. Our findings cover only two of its specific strategies. The first is the **set no-action default**: when practitioners choose their default time frame size and their estimation unit, they are choosing defaults. Shorter time frames and lower granularity units can hamper results by increasing underestimation — and we have moderate and high confidence in these review findings, respectively (as in Review Findings 5.1.1 and 5.1.2 in Table 5). The findings suggest that using the same time frame sizes and units for all types of projects, large and small, can jeopardize accuracy. We recommend using larger granularity time units and larger time frames in larger projects, which typically suffer from an underestimation bias. Thus, one avoids underestimation that can happen by anchoring on low values. An alternative is to adopt **prompted choice** of time frame sizes and estimation units as part of project planning instead of using a one-size-fits-all default. Prompted choice is a specific technique currently unexplored in the SEE context.

The second explored strategy in the SEE context is to **change range or composition of options**. Our review finding compares the rearranging of estimation values' options with Fibonacci scales instead of the arrangement with linear ones, suggesting that the first are worse than the latter for estimation software tasks and projects because it leads to lower values systematically. We have moderate confidence in it — see Review Finding 5.2.1 in Table 5. Al-

though this finding might be surprising for some because Fibonacci scales are recommended as good estimation practice (Cohn 2005), it can be explained by the diversification bias: the tendency for even allocation of resources over the available options—a type of variety seeking behavior (Fox et al. 2005). Consider that each value of an estimation scale represents an option that an estimator can pick. In Fibonacci-based estimation scales, there is a higher number of values below than above the medium value of the scale. For instance, Tamrakar and Jørgensen (2012) used a Fibonacci scale with six options below the medium value of 20 (1, 2, 3, 5, 8, 13) and three options higher (30, 40, over 40 work-hours). For the linear scale, researchers used values going from 1 to 40, plus an option representing over 40 work-hours — thus providing estimators an almost equal number of options below and above the medium value of 20. In this case, the diversification bias explains the underestimation bias when using the Fibonacci scale because there are twice more small value options as large ones. If people choose values for their user stories evenly from such a scale, they will consistently choose a higher number of smaller estimation values.

Moreover, one might think that estimators are assuming that tasks (or stories) to estimate are small because large ones were broken down into smaller units. This can apparently explain estimators choosing smaller values. However, this would lead estimators to do so both for Fibonacci and linear scales, providing no explanations of their differences.

In addition, it seems that SEE researchers and practitioners are not taking real advantage of what the technique of **change range or composition of options** has to offer. Choice architects can use it by partitioning desirable options into diverse categories (leading people to allocate more resources to them). They take into account the many diverse biases that come with decisions with multiple options (allocation biases, variety seeking, mental accounting, the denomination effect, and others) (Münscher et al. 2016). For instance, SEE researchers and practitioners can work on taking advantage of such biases in breaking tasks to estimate and ensuring that software engineering activities that traditionally receive less effort than needed are corrected for this. If people severely underestimate testing activities in one company, breaking them into a larger set of sub-tasks creates more options. This can lead estimators to allocate a more significant total time to the set of such activities than they would otherwise do.

We had no findings regarding the overall technique of to **change option consequences**. The idea is to provide "micro-incentives": changes of consequences that are insignificant from a rational perspective (Münscher et al. 2016). It has two specific techniques: **connect decision to benefit/cost**, which requires connecting the desired behavior to a small benefit or an undesired behavior to a small cost; and **change social consequences of the decision**, which regards connecting a choice with the consequence to be regarded more positively or negatively by others. We assess that such techniques are not easily implemented in the SEE context. For instance, higher management and project planners tend to prefer lower estimates (Magazinius et al.

2012). In other words, such estimates are more socially desirable. Is there any way we can make accurate estimates more socially desirable than lower ones? A challenge comes with this idea: people can expand their work to fit a large estimate, reducing productivity (Jorgensen 2014). How to deal with this?

Finally, our findings cover almost all strategies in the **decision assistance** category. The first is to **provide reminders**, which we do in the SEE context with the use of checklists. It improves estimations results with high confidence — see Review Finding 6.1.1 in Table 6. Therefore, it yields a good recommendation for enhancing estimation practice.

The second general strategy is to **facilitate commitments**, of which we have some issues to resolve as researchers. It requires making private or public commitments to deal with self-control problems. Deviations lead to cognitive dissonance or a need to justify to others (Kahneman et al. 2021). The first of its strategies is to **support self-commitment**, which involves commitment devices: arrangements to help decision-makers to fulfill a plan. We found no results to associate with this technique. The matter here is: is that adaptable to the software estimation context?

The second specific strategy is to **support public commitment**, which requires creating possibilities for and supporting the commitments in front of others, creating external pressure and, possibly, negative consequences in case it is broken. Our review finding states that the lack of justifications of estimates is a reason for estimation error, as shown in one study (Jorgensen and Molokken-Ostvold 2004). In this case, the justifications can explain how one can attain the estimate — artificially creating the expectation for a commitment to hit it.

Recommendations for practice:

- Prefer the use of the traditional format when asking for estimates instead of the alternative format to avoid the underestimation bias. Obs.: Potentially medium resource usage in contexts where people are used to working in fixed-time iterations, as it will require someone to fit the tasks into the iteration.
- Frame first estimates as ideal ones, and ask for a second estimate framing it as a most likely estimate — Obs.: Low resource usage and easy to implement.
- Bring attention to previous estimation performance, either by asking people to remember it or by providing simple feedback on it. Obs.: Potentially high resource usage (if providing feedback based on data).
- Anticipate participants' skills to estimators. Obs.: Low resource usage and easy to implement. It may be unfeasible in some contexts.
- Decide on the time frame size (when using the alternative format for asking for estimates) and estimation unit on a project basis instead of defining defaults. Obs.: Low resource usage.
- Prefer linear scales instead of Fibonacci scales to reduce underestimation bias. Obs.: Low resource usage.

– Use checklists to remind estimators of often overlooked tasks or predictors. Obs.: Potentially high resource usage if the checklist is long.

Recommendations for research:

– Investigate how to simplify information for estimators and how much it can affect estimation results.
– Assess specific techniques that build on descriptive norms of behavior to deal with specific biases and how they impact estimation results.
– Study alternative breaking of options for estimators (like the breaking of values of scales or the breaking down of tasks they estimate) and how it impacts estimation results.
– Provide more studies to address the adequacy and/or relevance concerns for promising review findings, such as the ones regarding request formats, sequences of types of estimates (ideal x most likely), time frame sizes, and Fibonacci versus linear scales (and possibly other types of scales used in the software industry).

5.3 Raising the Confidence in the Evidence

In Section 5.2 we presented a discussion of our review findings from the perspective of the matching between SEE and behavioral interventions, providing recommendations for practice based on review findings with high and moderate confidence. This section discusses how research on SEE can improve, considering how confident we are in each of our review findings. We provide specific guidance for research whenever we envision improvement opportunities regarding the GRADE-CERQual dimensions: methodological limitations, coherence, adequacy, and relevance.

One issue that Section 5.2 discusses is that we are not confident in making recommendations for practice about the best aggregation strategy for software estimates due to a variety of concerns. First, when comparing the statistical combination of individual estimates and Planning Poker, our review finding states that the first lead to more optimistic and less accurate estimates than the latter — indicating that Planning Poker is a better aggregation strategy. However, at least one study revealed that when participants are students averaging fares better (Mahnič and Hovelja 2012). This raised moderate concerns about the coherence of the review finding because of one contradictory result (as in Review Finding 3.2.2 in Table 3). Second, when comparing the statistical combination of individual estimates with unstructured group-based estimates, the latter also leads to less optimistic estimates, and again a discussion-based strategy seems better. The issue is that we have low confidence in this finding due to serious concerns about adequacy (because of only one supporting paper with few participants) and relevance (because of only one participating organization) — as we show in Review Finding 3.2.3 in Table 3. Third, in another review finding, Planning Poker was more accurate than unstructured discussions of estimates. However, we have very low confidence in this finding

due to serious concerns about methodological limitations regarding the only study about it, serious adequacy concerns, and moderate relevance concerns because of the narrow range of contexts and not-so recent paper — see Review Finding 3.2.4 in Table 3. Fourth, the statistical combination of estimates is compared to — and considered no worse than — the Delphi's highly structured method (and another discussion-based method). Once more, we have very low confidence in the finding due to serious concerns with methodological limitations (due to the tiny sample in an experimental design and the introduction of confounding factors) and adequacy (one paper with very few participants) — as in Review Finding 3.2.5 in Table 3. These findings and the assessment of their strength of evidence do not make us confident that group discussion methods are better to recommend than averaging individual estimates. We would rather recommend more research to clarify whether averaging makes better (or is no worse) in some specific contexts.

As for the selection part of **selection and aggregation**, we have very low to low confidence for most of the review findings (about the involvement of customers, mature teams, people not participating in the project, people responsible for the task, or even crowd workers for Planning Poker). The reasons range from concerns with methodological limitations to concerns with the adequacy of data (all of them were supported by one paper only) or concerns with relevance (due to narrow SE contexts) — as in Review Findings 3.2.7, 3.2.8, 3.2.9, 3.2.10, and 3.2.11 in Table 3.

Regarding the **reframe** technique, we recommended the use of the traditional format ("How much effort is required to complete X?") and asking for a second estimate, portraying it as a most likely estimate of effort and the first one the estimator gave us as an ideal estimate. However, these review findings were reported in one paper only, leading to moderate concerns about data adequacy. Although the number of participants is high enough to avoid more adequacy concerns (as in Review Findings 4.1.1 and 4.1.2 in Table 4, we can strengthen the confidence in these findings by having more researchers investigate these issues.

As for **making own behavior visible**, the review finding of complex feedback (through structured lessons learned, for instance) does not yield better results. Again, only one paper supports the latter finding, decreasing the confidence from high to moderate (as in Review Finding 4.2.2 in Table 4) and suggesting that more research can increase the adequacy of data.

Regarding **making external information visible**, one review finding regards the annotation of user stories during Planning Poker. We had low confidence in this finding because of moderate concerns in three dimensions — as in Review Finding 4.2.3 in Table 4). First, we had concerns about methodological limitations because the estimation tasks were not standardized; second, we had concerns about adequacy because only one paper supported the results; and third, we had concerns about relevance because participants were students. All these concerns point to improvements in research to strengthen the evidence. Another finding is the anticipation of project' participants' skills. One of the supporting papers found that it can lead to increased optimism in some

situations: the opposite results compared with the other papers, leading to minor concerns about the coherence of the review finding. Still, such concerns were minor, and we kept high confidence that it is adequate to recommend anticipating participants' skills to improve estimation results.

We also had a finding regarding **refer to opinion leader**. The corresponding review finding in the SEE domains (The lack of careful examination of estimates by management is a reason for inaccuracy — as in Review Finding 4.3.1 in Table 4) is supported by only one paper, leading to moderate concerns about data adequacy. We also had serious concerns about relevance because the paper is too old (dating from 1995), predating the agile movement and its recommendations of software teams' autonomy, which are now accepted as good management practice. These assessments led us to rate the review finding with very low confidence, reflecting that it does not hold in many contexts nowadays, although it might have been valid in the past. We would not recommend it for the estimation process of software organizations.

Many review findings led us to draw recommendations for practice in Section 5.2 even though we did not have high confidence in them. Our confidence was moderate, suggesting there is still room for improvement in research investigating them and acknowledging these review findings were promising. Many of these review findings had their confidence levels downgraded due to adequacy concerns because only one paper supported them. That was the case of the recommendation for traditional request formats, sequences of types of estimates (ideal x most likely), larger time frames, and the use of linear scales (instead of Fibonacci ones). We need more researchers investigating the matter to raise our confidence in these findings from moderate to high.

Another review finding was about the lack of justifications, in the context of the **support public commitment** category. We have very low confidence that we can use this finding to draw a recommendation (as in Review Finding 6.2.1 in Table 6). The results come from one study collecting data in one company only, raising serious concerns about adequacy. It also leads to serious concerns about relevance due to the narrow range of SE contexts.

Recommendations for research:

- Execute more comparative studies between diverse aggregation strategies in different SE contexts. We need more comparisons between straight averaging (statistical combination) and discussion-based methods (like Planning Poker, unstructured group discussion methods, Delphi, and possibly the estimate-talk-estimate procedure).
- Provide more studies to address the adequacy and/or relevance concerns for promising review findings, such as the ones regarding request formats, sequences of types of estimates (ideal x most likely), time frame sizes, and Fibonacci versus linear scales (and possibly other types of scales used in the software industry).

6 Limitations

For thematic analysis, we can use the credibility, transferability, dependability, and confirmability criteria as pragmatic choices to assess research results and establish trustworthiness (Nowell et al. 2017). Credibility refers to the fit between the data and process of analysis with the intended focus (Cruzes and Dyba 2011). It can be addressed by varied means, such as prolonged engagement with data, triangulation of data collection modes, documenting thoughts about potential codes/themes and theoretical and reflexive thoughts, storing raw data in well-organized archives, and keeping records of field notes and reflexive journals (Nowell et al. 2017). We engaged with data for over two years and organized data into extraction forms and codebooks. Also, we chose the data from a previous SLM based on our intended focus to understand the perceptions about SEE research and practice from a broader perspective — something that a single primary study, such as a case study or a survey, was unlikely to provide us. In addition, we purposefully chose the two types of thematic analysis based on the reflexive nature of the first phase of our study. We then moved on with a codebook approach to address the research question we generated for the second phase, which required us to adopt a more well-defined theoretical lens. Although we chose a strong theoretical framework based on literature from consolidated researchers in the behavioral science arena, there is a risk our matching of SEE and behavioral interventions would be different if we had a behavioral science researcher in our team.

The transferability criteria refer to the generalizability of the inquiry. The researcher needs to offer thick descriptions to allow the readers to decide about transferability to their context (Nowell et al. 2017). In the case of research synthesis, it involves describing the selection and characteristics of the primary studies (Cruzes and Dyba 2011). Regarding the selection of primary studies, we selected papers from a previous SLM, which raises two issues. First, the previous SLM used a search string to answer different research questions. However, we did not consider this to be a major issue, as such questions focused on factors affecting expert-judgment software estimates, of which SEE interventions are a subset. Second, the search from the previous SLM was carried out in 2020. To mitigate this, we carried out one round of snowballing using the set of papers from the previous SLM as a seed set. Still, there is a risk that we might have missed relevant primary studies in our review, representing a threat to study selection validity (Ampatzoglou et al. 2019). Regarding the characteristics of our primary studies, we provided information in our extraction forms and supplementary material regarding their research strategy, context, and results.

Dependability concerns whether the research process is logical, traceable, and documented — demonstrated by making the research process auditable (Nowell et al. 2017). We documented all data extracted and created a detailed codebook to explain all categories and techniques we considered to address this. We also documented the details of our assessment of the strength of

evidence in EP tables provided as supplementary material (Matsubara et al. 2023, Online Resource 3).

Confirmability regards how the extracted data is coded and sorted (Cruzes and Dyba 2011). It is also about whether the researchers' interpretations and findings are derived from the data (Nowell et al. 2017). We addressed this criterion by using GRADE-CERQual, which brings to qualitative and mixed-methods systematic reviews the discussion on data coherence and adequacy — concepts that regard how clear and cogent is the fit between data and the review findings and how rich and how much data support them, respectively. The explicit evaluation of such dimensions might differ between researchers, but the idea of GRADE-CERQual is rather transparency of judgments instead of agreement, allowing people to disagree on objective grounds. To enable this, we preserved the links between the primary studies, the review findings (in the SoQF and EP tables), the categories and techniques described in our codebooks, and the themes. We published them as supplementary material (Matsubara et al. 2023, Online Resource 2).

7 Conclusions

In this study, we analyzed SEE factors to emphasize that SEE is more than a **prediction task**: it is also a **behavioral act**. Through reflexive thematic analysis, we show how this perspective on SEE requires that researchers and practitioners consider more than the technicalities of estimating software tasks and projects to improve accuracy. They need to account for human behavior too. This requires us to consider the surroundings of the estimation process: how one prepare, ask for, communicate, and use the software estimates also matters. Ultimately, that means that improving estimation methods is not enough: practitioners also need behavioral interventions to address the human and social sides.

Therefore, we investigated behavioral interventions from other domains to compare them with the SEE interventions in the context of judgment-based estimation. We carried out a codebook thematic analysis using a framework built from recent works from behavioral sciences, to which we matched the SEE interventions. This allowed the identification of the SEE interventions rooted in the knowledge of other behavioral interventions, even if only partially. With such knowledge in their hands, SE practitioners can assess which SEE interventions can be helpful in their contexts, supporting process improvement initiatives focused on their estimation processes. Some recommendations for practice we can draw from our review relate to preparing for estimation, such as training estimators to make them knowledgeable on estimation methods, practices, and biases affecting their estimates. We can also make a handful of recommendations concerning how to ask for estimates, such as choosing the appropriate request format (the traditional format reduces underestimation). Moreover, some recommendations concern the generation or the communica-

tion of software estimates, like the one of providing checklists to remember estimators of tasks and effort predictors.

All such recommendations are rooted in review findings that we assessed with high or moderate confidence considering GRADE-CERQual for the strength of evidence evaluations. However, our results also include review findings with low and very low ratings. SE researchers can use such findings to guide future research efforts. Future work includes more studies to address the adequacy and/or relevance concerns for promising review findings, such as the ones regarding request formats. Moreover, we now see the gaps represented by behavioral interventions unexplored or partially explored in the SEE domain, such as the impact of intelligence and other cognitive styles beyond interdependence, more sophisticated anchored scales, alternative breaking of tasks to estimate or options given to estimators to select as estimating values, among others. The investigation of such interventions is also part of future work. Finally, future research efforts can investigate further developments and nuances coming from other purposes that SEE is used for rather than just providing a prediction value (as in the "SEE as a technical prediction task"). For instance, SEE can also increase communication among team members, enabling a better team understanding of tasks to complete or the project roadmap. It can also be regarded as a social act, which can be seen as a development of SEE as a behavioral act.

Funding

We thank the reviewers for all their suggestions, many of which we incorporated into the paper and significantly improved it. The present work is the result of the Research and Development (R&D) project 001/2020, signed with the Federal University of Amazonas and FAEPI, Brazil, which has funding from Samsung, using resources from the Informatics Law for the Western Amazon (Federal Law nº 8.387/1991), and its disclosure is in accordance with article 39 of Decree No. 10.521/2020. Also supported by the Federal University of Mato Grosso do Sul (UFMS), the Federal University of Amazonas (UFAM), CAPES - Financing Code 001, CNPq processes 314174/2020-6 and 313067/2020-1, and FAPEAM process 062.00150/2020, and grant #2020/05191-2 São Paulo Research Foundation (FAPESP).

Conflict of interest

The authors declare that they have no conflict of interest.

Data availability

All material generated during the current study is available at Figshare (`https://doi.org/10.6084/m9.figshare.19406945.v1`), as we describe here:

– Online Resource 1 presents the relationships between factors and latent themes.
– Online Resource 2 presents the codebook with the categories, general and specific strategies, and their descriptions, composing the analytical framework.
– Online Resource 3 presents the list of papers included in the current study, along with the Evidence Profile and Summary of Qualitative Findings Tables.
– Online Resource 4 presents the quality assessment for each paper we included in the current study.

References

Alhamed M, Storer T (2021) Playing Planning Poker in Crowds: Human Computation of Software Effort Estimates. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), IEEE, Madrid, ES, pp 1–12, DOI 10.1109/ICSE43902.2021.00014, iSSN: 1558-1225

Altaleb A, Gravell A (2019) An Empirical Investigation of Effort Estimation in Mobile Apps Using Agile Development Process. Journal of Software 14(8):356–369, URL http://www.jsoftware.us/index.php?m=content&c=index&a=show&catid=211&id=2959

Altaleb A, Alhashimi H, Gravell A (2020) A Case Study Validation of the Pair-estimation Technique in Effort Estimation of Mobile App Development Using Agile Processes. In: 2020 10th International Conference on Advanced Computer Information Technologies (ACIT), IEEE, Deggendorf, Germany, pp 469–473, DOI 10.1109/ACIT49673.2020.9208985

Ampatzoglou A, Bibi S, Avgeriou P, Verbeek M, Chatzigeorgiou A (2019) Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. Information and Software Technology 106:201–230, DOI 10.1016/j.infsof.2018.10.006, URL http://www.sciencedirect.com/science/article/pii/S0950584918302106

Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, Nasser M, Meerpohl J, Post PN, Kunz R, Brozek J, Vist G, Rind D, Akl EA, Schünemann HJ (2013a) GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. Journal of Clinical Epidemiology 66(7):719–725, DOI 10.1016/j.jclinepi.2012.03.013

Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, Rind D, Montori VM, Brito JP, Norris S, Elbarbary M, Post P, Nasser M, Shukla V, Jaeschke R, Brozek J, Djulbegovic B, Guyatt G (2013b) GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. Journal of Clinical Epidemiology 66(7):726–735, DOI 10.1016/j.jclinepi.2013.02.003

Aranda J, Easterbrook S (2005) Anchoring and Adjustment in Software Estimation. In: Proceedings of the 10th European Software Engineering Conference Held Jointly with 13th ACM SIGSOFT International Symposium on Foundations of Software Engineering, ACM, New York, NY, USA, ESEC/FSE-13, pp 346–355, DOI 10.1145/1081706.1081761, URL http://doi.acm.org/10.1145/1081706.1081761, event-place: Lisbon, Portugal

Arifin HH, Daengdej J, Khanh NT (2017) An Empirical Study of Effort-Size and Effort-Time in Expert-Based Estimations. In: 2017 8th International Workshop on Empirical Software Engineering in Practice (IWESEP), IEEE, Tokyo, Japan, pp 35–40, DOI 10.1109/IWESEP.2017.21

Arnuphaptrairong T (2021) Enhancing Delphi Method with Algorithmic Estimates for Software Effort Estimation: An Experimental Study. SSRN Scholarly Paper ID 3898965, Social Science Research Network, Rochester, NY, URL https://papers.ssrn.com/abstract=3898965

Braun V, Clarke V (2006) Using thematic analysis in psychology. Qualitative Research in Psychology 3(2):77–101, DOI 10.1191/1478088706qp063oa, URL https://www.

tandfonline.com/doi/abs/10.1191/1478088706qp063oa, publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

Braun V, Clarke V (2021) One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qualitative Research in Psychology 18(3):328–352, DOI 10.1080/14780887.2020.1769238, URL https://doi.org/10.1080/14780887.2020.1769238, publisher: Routledge _eprint: https://doi.org/10.1080/14780887.2020.1769238

Briggs RA (2019) Normative Theories of Rational Choice: Expected Utility. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, fall 2019 edn, Metaphysics Research Lab, Stanford University, URL https://plato.stanford.edu/archives/fall2019/entries/rationality-normative-utility/

Brooks F (1995) Mythical Man-Month, The: Essays on Software Engineering, Anniversary Edition, anniversary edition edn. Addison-Wesley Professional, Reading, Mass

Brzezicka J, Wisniewski R (2014) Homo Oeconomicus and Behavioral Economics. Contemporary Economics 8(4):353–364, DOI 10.5709/ce.1897-9254.150

Buyalskaya A, Gallo M, Camerer CF (2021) The golden age of social science. Proceedings of the National Academy of Sciences 118(5):e2002923118, DOI 10.1073/pnas.2002923118, URL https://www.pnas.org/doi/10.1073/pnas.2002923118, publisher: Proceedings of the National Academy of Sciences

Cohn M (2005) Agile Estimating and Planning, 1st edn. Robert C. Martin Series, Pearson

Colvin CJ, Garside R, Wainwright M, Munthe-Kaas H, Glenton C, Bohren MA, Carlsen B, Tunçalp O, Noyes J, Booth A, Rashidian A, Flottorp S, Lewin S (2018) Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 4: how to assess coherence. Implementation Science 13(1):13, DOI 10.1186/s13012-017-0691-8, URL https://doi.org/10.1186/s13012-017-0691-8

Connelly LM, Peltzer JN (2016) Underdeveloped Themes in Qualitative Research: Relationship With Interviews and Analysis. Clinical nurse specialist CNS 30(1):52–57, DOI 10.1097/NUR.0000000000000173

Conoscenti M, Besner V, Vetrò A, Fernández DM (2019) Combining data analytics and developers feedback for identifying reasons of inaccurate estimations in agile software development. Journal of Systems and Software 156:126–135, DOI 10.1016/j.jss.2019.06.075, URL http://www.sciencedirect.com/science/article/pii/S0164121219301372

Cruzes DS, Dyba T (2011) Recommended Steps for Thematic Synthesis in Software Engineering. In: Proceedings of the 2011 International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society, Washington, DC, USA, ESEM '11, pp 275–284, DOI 10.1109/ESEM.2011.36, URL https://doi.org/10.1109/ESEM.2011.36

DeMarco T, Lister T, House D (2013) Peopleware: Productive Projects and Teams, 3rd edn. Addison-Wesley Professional, Upper Saddle River, NJ

Dybå T, Dingsøyr T (2008) Strength of evidence in systematic reviews in software engineering. In: Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement, Association for Computing Machinery, New York, NY, USA, ESEM '08, pp 178–187, DOI 10.1145/1414004.1414034, URL https://doi.org/10.1145/1414004.1414034

Fereday J, Muir-Cochrane E (2006) Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. International Journal of Qualitative Methods 5(1):80–92, DOI https://doi.org/10.1177/160940690600500107

Fleischmann M, Amirpur M, Benlian A, Hess T (2014) Cognitive Biaes in Information Systems Research: A Scientometric Analysis. ECIS 2014 Proceedings URL https://aisel.aisnet.org/ecis2014/proceedings/track02/5

Fox CR, Ratner RK, Lieb DS (2005) How subjective grouping of options influences choice and allocation: diversification bias and the phenomenon of partition dependence. Journal of Experimental Psychology General 134(4):538–551, DOI 10.1037/0096-3445.134.4.538

Frid-Nielsen SS, Jensen MD (2021) Maps of Behavioural Economics: Evidence from the Field. Journal of Interdisciplinary Economics 33(2):226–250, DOI 10.1177/0260107920925675, URL https://doi.org/10.1177/0260107920925675, publisher: SAGE Publications India

Furulund KM, Molkken-stvold K (2007) Increasing Software Effort Estimation Accuracy Using Experience Data, Estimation Models and Checklists. In: Seventh International Conference on Quality Software (QSIC 2007), IEEE, Portland, OR, USA, pp 342–347, DOI 10.1109/QSIC.2007.4385518, iSSN: 2332-662X

Fægri TE (2010) Adoption of Team Estimation in a Specialist Organizational Environment. In: Sillitti A, Martin A, Wang X, Whitworth E (eds) Agile Processes in Software Engineering and Extreme Programming, Springer, Berlin, Heidelberg, Lecture Notes in Business Information Processing, pp 28–42, DOI 10.1007/978-3-642-13054-0_3

Gale NK, Heath G, Cameron E, Rashid S, Redwood S (2013) Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Medical Research Methodology 13(1):117, DOI 10.1186/1471-2288-13-117, URL https://doi.org/10.1186/1471-2288-13-117

Gandomani TJ, Koh TW, Binhamid AK (2014) A case study research on software cost estimation using experts' estimates, Wideband Delphi, and Planning Poker technique. International Journal of Software Engineering and Its Applications 8(11):173–182, DOI http://dx.doi.org/10.14257/ijseia.2014.8.11.16

Gandomani TJ, Faraji H, Radnejad M (2019) Planning Poker in cost estimation in Agile methods: Averaging Vs. Consensus. In: 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI), IEEE, Tehran, Iran, pp 066–071, DOI 10.1109/KBEI.2019.8734960

Glass RL, Rost J, Matook MS (2008) Lying on Software Projects. IEEE Software 25(6):90–95, DOI 10.1109/MS.2008.150

Glenton C, Carlsen B, Lewin S, Munthe-Kaas H, Colvin CJ, Tunçalp O, Bohren MA, Noyes J, Booth A, Garside R, Rashidian A, Flottorp S, Wainwright M (2018) Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 5: how to assess adequacy of data. Implementation Science 13(1):14, DOI 10.1186/s13012-017-0692-7, URL https://doi.org/10.1186/s13012-017-0692-7

Grapenthin S, Book M, Richter T, Gruhn V (2016) Supporting Feature Estimation with Risk and Effort Annotations. In: 2016 42th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), IEEE, Limassol, Cyprus, pp 17–24, DOI 10.1109/SEAA.2016.24, iSSN: 2376-9505

Grimstad S, Jørgensen M (2007) Inconsistency of expert judgment-based estimates of software development effort. Journal of Systems and Software 80(11):1770–1777, DOI 10.1016/j.jss.2007.03.001, URL https://www.sciencedirect.com/science/article/pii/S0164121207000714

Grimstad S, Jørgensen M (2008) A Preliminary Study of Sequence Effects in Judgment-based Software Development Work-effort Estimation. In: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, BCS Learning & Development Ltd., Swindon, UK, EASE'08, pp 129–135, URL http://dl.acm.org/citation.cfm?id=2227115.2227129, event-place: Italy

Grimstad S, Jørgensen M (2009) Preliminary study of sequence effects in judgment-based software development work-effort estimation. IET Software 3(5):435–441, DOI 10.1049/iet-sen.2008.0110, conference Name: IET Software

Grimstad S, Jorgensen M, Molokken-Ostvold K (2005) The clients' impact on effort estimation accuracy in software development projects. In: 11th IEEE International Software Metrics Symposium (METRICS'05), IEEE, Como, Italy, pp 10 pp.–10, DOI 10.1109/METRICS.2005.30, iSSN: 1530-1435

Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A (2011) GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. Journal of Clinical Epidemiology 64(4):380–382, DOI 10.1016/j.jclinepi.2010.09.011, URL https://www.sciencedirect.com/science/article/pii/S089543561000329X

Halkjelsvik T, Jorgensen M (2011) To read two pages, I need 5 minutes, but give me 5 minutes and I will read four: how to change productivity estimates by inverting the question. Applied Cognitive Psychology 25(2):314–323, DOI 10.1002/acp.1693

Halkjelsvik T, Jørgensen M (2018a) How We Predict Time Usage. In: Halkjelsvik T, Jørgensen M (eds) Time Predictions: Understanding and Avoiding Unrealism in Project Planning and Everyday Life, Simula SpringerBriefs on Computing, Springer International Publishing, Cham, pp 5–11, DOI 10.1007/978-3-319-74953-2_2, URL https:

//doi.org/10.1007/978-3-319-74953-2_2

Halkjelsvik T, Jørgensen M (2018b) Time Predictions: Understanding and Avoiding Unrealism in Project Planning and Everyday Life. Simula SpringerBriefs on Computing, Springer International Publishing, Cham, Switzerland, DOI 10.1007/978-3-319-74953-2, URL https://www.springer.com/gp/book/9783319749525

Halkjelsvik T, Jørgensen M (2018c) Uncertainty of Time Predictions. In: Halkjelsvik T, Jørgensen M (eds) Time Predictions: Understanding and Avoiding Unrealism in Project Planning and Everyday Life, Simula SpringerBriefs on Computing, Springer International Publishing, Cham, pp 71–79, DOI 10.1007/978-3-319-74953-2_5, URL https://doi.org/10.1007/978-3-319-74953-2_5

Haran U, Ritov I, Mellers BA (2013) The role of actively open-minded thinking in information acquisition, accuracy, and calibration. Judgment and Decision Making 8(3):188–201, URL https://ideas.repec.org/a/jdm/journl/v8y2013i3p188-201.html, publisher: Society for Judgment and Decision Making

Haugen NC (2006) An empirical study of using planning poker for user story estimation. In: AGILE 2006 (AGILE'06), IEEE, Minneapolis, MN, USA, pp 9 pp.–34, DOI 10.1109/AGILE.2006.16

He M, Zhang H, Yang Y, Wang Q, Li M (2010) Understanding the Influential Factors to Development Effort in Chinese Software Industry. In: Ali Babar M, Vierimaa M, Oivo M (eds) Product-Focused Software Process Improvement, Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, pp 306–320, DOI 10.1007/978-3-642-13792-1_24

Hoda R, Murugesan LK (2016) Multi-level agile project management challenges: A self-organizing team perspective. Journal of Systems and Software 117:245–257, DOI 10.1016/j.jss.2016.02.049, URL https://www.sciencedirect.com/science/article/pii/S0164121216000807

Hughes RT (1996) Expert judgement as an estimating method. Information and Software Technology 38(2):67–75, DOI 10.1016/0950-5849(95)01045-9, URL http://www.sciencedirect.com/science/article/pii/0950584995010459

Jorgensen M (2014) What We Do and Don't Know about Software Development Effort Estimation. IEEE Software 31(2):37–40, DOI 10.1109/MS.2014.49, URL http://ieeexplore.ieee.org/document/6774376/

Jorgensen M, Carelius GJ (2004) An Empirical Study of Software Project Bidding. IEEE Trans Softw Eng 30(12):953–969, DOI 10.1109/TSE.2004.92, URL https://doi.org/10.1109/TSE.2004.92

Jorgensen M, Grimstad S (2012) Software Development Estimation Biases: The Role of Interdependence. IEEE Transactions on Software Engineering 38(3):677–693, DOI 10.1109/TSE.2011.40, conference Name: IEEE Transactions on Software Engineering

Jorgensen M, Molokken-Ostvold K (2004) Reasons for software effort estimation error: impact of respondent role, information collection approach, and data analysis method. IEEE Transactions on Software Engineering 30(12):993–1007, DOI 10.1109/TSE.2004.103, URL http://ieeexplore.ieee.org/document/1377193/

Jørgensen M (2011) Contrasting ideal and realistic conditions as a means to improve judgment-based software development effort estimation. Information and Software Technology 53(12):1382–1390, DOI 10.1016/j.infsof.2011.07.001, URL https://doi.org/10.1016/j.infsof.2011.07.001

Jørgensen M (2013) Relative Estimation of Software Development Effort: It Matters with What and How You Compare. IEEE Software 30(2):74–79, DOI 10.1109/MS.2012.70, conference Name: IEEE Software

Jørgensen M (2015) The effect of the time unit on software development effort estimates. In: 2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), IEEE, Kathmandu, Nepal, pp 1–5, DOI 10.1109/SKIMA.2015.7399992

Jørgensen M (2016) Unit effects in software project effort estimation: Work-hours gives lower effort estimates than workdays. Journal of Systems and Software 117:274–281, DOI 10.1016/j.jss.2016.03.048, URL http://www.sciencedirect.com/science/article/pii/S0164121216300085

Jørgensen M, Escott E (2022) Relative estimates of software development effort: Are they more accurate or less time-consuming to produce than absolute estimates, and

to what extent are they person-independent? Information and Software Technology 143:106782, DOI 10.1016/j.infsof.2021.106782, URL https://www.sciencedirect.com/science/article/pii/S0950584921002251

Jørgensen M, Grimstad S (2011) The Impact of Irrelevant and Misleading Information on Software Development Effort Estimates: A Randomized Controlled Field Experiment. IEEE Transactions on Software Engineering 37(5):695–707, DOI 10.1109/TSE.2010.78

Jørgensen M, Gruschke TM (2009) The Impact of Lessons-Learned Sessions on Effort Estimation and Uncertainty Assessments. IEEE Transactions on Software Engineering 35(3):368–383, DOI 10.1109/TSE.2009.2

Jørgensen M, Halkjelsvik T (2010) The effects of request formats on judgment-based effort estimation. Journal of Systems and Software 83(1):29–36, DOI 10.1016/j.jss.2009.03.076, URL http://www.sciencedirect.com/science/article/pii/S0164121209000879

Jørgensen M, Halkjelsvik T (2020) Sequence effects in the estimation of software development effort. Journal of Systems and Software 159:110448, DOI 10.1016/j.jss.2019.110448, URL http://www.sciencedirect.com/science/article/pii/S0164121219302225

Jørgensen M, Sjøberg DIK (2001) Impact of effort estimates on software project work. Information and Software Technology 43:10

Jørgensen M, Sjøberg DIK (2004) The impact of customer expectation on software development effort estimates. International Journal of Project Management 22(4):317–325, DOI 10.1016/S0263-7863(03)00085-1, URL http://www.sciencedirect.com/science/article/pii/S0263786303000851

Jørgensen M, Faugli B, Gruschke T (2007) Characteristics of software engineers with optimistic predictions. Journal of Systems and Software 80(9):1472–1482, DOI 10.1016/j.jss.2006.09.047, URL http://www.sciencedirect.com/science/article/pii/S0164121206002986

Jørgensen M, Boehm B, Rifkin S (2009) Software Development Effort Estimation: Formal Models or Expert Judgment? IEEE Software 26(2):14–19, DOI 10.1109/MS.2009.47, conference Name: IEEE Software

Kahneman D, Rosenfield AM, Gandhi L, Blaser T (2016) Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making. Harvard Business Review October:36–43, URL https://hbr.org/2016/10/noise, section: Decision making and problem solving

Kahneman D, Sibony O, Sunstein CR (2021) Noise: A Flaw in Human Judgment, vol 1, 1st edn. Little, Brown Spark, New York

Karna H, Gotovac S (2014) Estimators characteristics and effort estimation of software projects. In: 2014 9th International Conference on Software Engineering and Applications (ICSOFT-EA), IEEE, Vienna, Austria, pp 26–35

Keaveney S, Conboy K (2006) Cost estimation in agile development projects. In: ECIS 2006 Proceedings, AIS Library, Göteborg, Sweden, p 16

Kitchenham B, Linkman S (1997) Estimates, Uncertainty, and Risk. IEEE Software 14(3):69–74, DOI 10.1109/52.589239, URL https://doi.org/10.1109/52.589239

Lagerström R, von Würtemberg LM, Holm H, Luczak O (2012) Identifying factors affecting software development cost and productivity. Software Quality Journal 20(2):395–417, DOI 10.1007/s11219-011-9137-8, URL https://doi.org/10.1007/s11219-011-9137-8

Layman L, Nagappan N, Guckenheimer S, Beehler J, Begel A (2008) Mining software effort data: preliminary analysis of visual studio team system data. In: Proceedings of the 2008 international working conference on Mining software repositories, Association for Computing Machinery, New York, NY, USA, MSR '08, pp 43–46, DOI 10.1145/1370750.1370762, URL https://doi.org/10.1145/1370750.1370762

Lederer A, Mirani R (1990) Information System Cost Estimating: A Management Perspective. Management Information Systems Quarterly 14(2):159–176, URL https://aisel.aisnet.org/misq/vol14/iss2/3

Lederer AL, Prasad J (1991) The validation of a political model of information systems development cost estimating. ACM SIGCPR Computer Personnel 13(2):47–57, DOI 10.1145/122393.122398, URL https://doi.org/10.1145/122393.122398

Lederer AL, Prasad J (1995) Causes of inaccurate software development cost estimates. Journal of Systems and Software 31(2):125–134, DOI 10.1016/0164-1212(94)00092-2, URL https://linkinghub.elsevier.com/retrieve/pii/0164121294000922

Lee M, Rothenberger M, Peffers K (2011) Identifying Effort Estimation Factors for Corrective Maintenance in Object-Oriented Systems. In: AMCIS 2011 Proceedings, p 186, URL https://aisel.aisnet.org/amcis2011_submissions/186

Lenberg P, Feldt R, Wallgren LG (2014) Towards a behavioral software engineering. In: Proceedings of the 7th International Workshop on Cooperative and Human Aspects of Software Engineering, Association for Computing Machinery, New York, NY, USA, CHASE 2014, pp 48–55, DOI 10.1145/2593702.2593711, URL https://doi.org/10.1145/2593702.2593711

Lenberg P, Feldt R, Wallgren LG (2015) Behavioral software engineering: A definition and systematic literature review. Journal of Systems and Software 107:15–37, DOI 10.1016/j.jss.2015.04.084

Lewin S, Bohren M, Rashidian A, Munthe-Kaas H, Glenton C, Colvin CJ, Garside R, Noyes J, Booth A, Tunçalp O, Wainwright M, Flottorp S, Tucker JD, Carlsen B (2018a) Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 2: how to make an overall CERQual assessment of confidence and create a Summary of Qualitative Findings table. Implementation Science 13(1):10, DOI 10.1186/s13012-017-0689-2, URL https://doi.org/10.1186/s13012-017-0689-2

Lewin S, Booth A, Glenton C, Munthe-Kaas H, Rashidian A, Wainwright M, Bohren MA, Tunçalp O, Colvin CJ, Garside R, Carlsen B, Langlois EV, Noyes J (2018b) Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. Implementation Science 13(1):2, DOI 10.1186/s13012-017-0688-3, URL https://doi.org/10.1186/s13012-017-0688-3

Løhre E, Jørgensen M (2016) Numerical anchors and their strong effects on software development effort estimates. Journal of Systems and Software 116:49–56, DOI 10.1016/j.jss.2015.03.015, URL http://www.sciencedirect.com/science/article/pii/S0164121215000618

Magazinius A, Börjesson S, Feldt R (2012) Investigating intentional distortions in software cost estimation – An exploratory study. Journal of Systems and Software 85(8):1770–1781, DOI 10.1016/j.jss.2012.03.026

Magazinovic A, Pernstål J (2008) Any other cost estimation inhibitors? In: Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement - ESEM '08, ACM Press, Kaiserslautern, Germany, p 233, DOI 10.1145/1414004.1414042, URL http://portal.acm.org/citation.cfm?doid=1414004.1414042

Mahnič V, Hovelja T (2012) On using planning poker for estimating user stories. Journal of Systems and Software 85(9):2086–2095, DOI 10.1016/j.jss.2012.04.005, URL http://www.sciencedirect.com/science/article/pii/S0164121212001021

Makridakis S, Hyndman RJ, Petropoulos F (2020) Forecasting in social settings: The state of the art. International Journal of Forecasting 36(1):15–28, DOI 10.1016/j.ijforecast.2019.05.011, URL https://www.sciencedirect.com/science/article/pii/S0169207019301876

Mann A (2016) The power of prediction markets. Nature 538(7625):308–310, DOI 10.1038/538308a, URL https://www.nature.com/articles/538308a, bandiera_abtest: a Cg_type: Nature Research Journals Number: 7625 Primary_atype: News Publisher: Nature Publishing Group Subject_term: Economics;Politics;Research management;Society Subject_term_id: economics;politics;research-management;society

Matos O, Fortaleza L, Conte T, Mendes E (2013) Realising web effort estimation. In: Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering, Association for Computing Machinery, Porto de Galinhas, pp 12–23, DOI 10.1145/2460999.2461002

Matsubara P, Gadelha B, Steinmacher I, Conte T (2021a) Supplementary material for the SEXTAMT. URL https://doi.org/10.6084/m9.figshare.14502405.v2

Matsubara P, Steinmacher I, Gadelha B, Conte T (2021b) Buying time in software development: how estimates become commitments? In: Proceedings of the 14th International Conference on Cooperative and Human Aspects of Software Engineering, IEEE, Madrid, Spain, pp 61–70

Matsubara P, Gadelha B, Steinmacher I, Conte T (2022) SEXTAMT: A systematic map to navigate the wide seas of factors affecting expert judgment software estimates. The Journal of Systems and Software 185:111148, DOI 10.1016/j.jss.2021.111148

Matsubara P, Gadelha B, Steinmacher I, Conte T (2023) Material for Much More Than a Prediction. URL https://doi.org/10.6084/m9.figshare.19406945.v1

Mellers B, Stone E, Atanasov P, Rohrbaugh N, Metz SE, Ungar L, Bishop MM, Horowitz M, Merkle E, Tetlock P (2015a) The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. Journal of Experimental Psychology: Applied 21(1):1–14, DOI 10.1037/xap0000040, place: US Publisher: American Psychological Association

Mellers B, Stone E, Murray T, Minster A, Rohrbaugh N, Bishop M, Chen E, Baker J, Hou Y, Horowitz M, Ungar L, Tetlock P (2015b) Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. Perspectives on Psychological Science 10(3):267–281, DOI 10.1177/1745691615577794, URL https://doi.org/10.1177/1745691615577794, publisher: SAGE Publications Inc

Mendes FF, Mendes E, Salleh N (2019) The relationship between personality and decision-making: A Systematic literature review. Information and Software Technology 111:50–71, DOI 10.1016/j.infsof.2019.03.010, URL https://www.sciencedirect.com/science/article/pii/S0950584919300576

Merriam-Webster (2021) Forecast. In Merriam-Webster.com dictionary. URL https://www.merriam-webster.com/dictionary/forecast

Mohanani R, Salman I, Turhan B, Rodríguez P, Ralph P (2020) Cognitive Biases in Software Engineering: A Systematic Mapping Study. IEEE Transactions on Software Engineering 46(12):1318–1339, DOI 10.1109/TSE.2018.2877759, conference Name: IEEE Transactions on Software Engineering

Moløkken-Østvold K, Jørgensen M (2004) Group Processes in Software Effort Estimation. Empirical Software Engineering 9(4):315–334, DOI 10.1023/B:EMSE.0000039882.39206.5a, URL https://doi.org/10.1023/B:EMSE.0000039882.39206.5a

Moløkken-Østvold K, Haugen NC, Benestad HC (2008) Using planning poker for combining expert estimates in software projects. Journal of Systems and Software 81(12):2106–2117, DOI 10.1016/j.jss.2008.03.058, URL http://www.sciencedirect.com/science/article/pii/S0164121208000885

Munthe-Kaas H, Bohren MA, Glenton C, Lewin S, Noyes J, Tunçalp O, Booth A, Garside R, Colvin CJ, Wainwright M, Rashidian A, Flottorp S, Carlsen B (2018) Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 3: how to assess methodological limitations. Implementation Science 13(1):9, DOI 10.1186/s13012-017-0690-9, URL https://doi.org/10.1186/s13012-017-0690-9

Münscher R, Vetter M, Scheuerle T (2016) A Review and Taxonomy of Choice Architecture Techniques. Journal of Behavioral Decision Making 29(5):511–524, DOI 10.1002/bdm.1897, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.1897, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.1897

Nowell LS, Norris JM, White DE, Moules NJ (2017) Thematic Analysis: Striving to Meet the Trustworthiness Criteria. International Journal of Qualitative Methods 16(1):1609406917733847, DOI 10.1177/1609406917733847, URL https://doi.org/10.1177/1609406917733847, publisher: SAGE Publications Inc

Noyes J, Booth A, Lewin S, Carlsen B, Glenton C, Colvin CJ, Garside R, Bohren MA, Rashidian A, Wainwright M, Tunçalp O, Chandler J, Flottorp S, Pantoja T, Tucker JD, Munthe-Kaas H (2018) Applying GRADE-CERQual to qualitative evidence synthesis findings–paper 6: how to assess relevance of the data. Implementation Science 13(1):4, DOI 10.1186/s13012-017-0693-6, URL https://doi.org/10.1186/s13012-017-0693-6

Organization C (2021) Measurement Manual v5.0 Part 1 Principles, Defs. & Rules. Tech. rep., COSMIC Organization, URL https://cosmic-sizing.org/publications/measurement-manual-v5-0-may-2020-part-1-principles-definitions-rules/

Petropoulos F, Apiletti D, Assimakopoulos V, Babai MZ, Barrow DK, Ben Taieb S, Bergmeir C, Bessa RJ, Bijak J, Boylan JE, Browell J, Carnevale C, Castle JL, Cirillo P, Clements MP, Cordeiro C, Cyrino Oliveira FL, De Baets S, Dokumentov A, Ellison J, Fiszeder P, Franses PH, Frazier DT, Gilliland M, Gönül MS, Goodwin P, Grossi L, Grushka-Cockayne Y, Guidolin M, Guidolin M, Gunter U, Guo X, Guseo R, Harvey N, Hendry DF, Hollyman R, Januschowski T, Jeon J, Jose VRR, Kang Y, Koehler AB, Kolassa S, Kourentzes N, Leva S, Li F, Litsiou K, Makridakis S, Martin GM, Martinez AB, Meeran S, Modis T, Nikolopoulos K, Önkal D, Paccagnini A, Panagiotelis A, Panapakidis I, Pavía JM, Pedio M, Pedregal DJ, Pinson P, Ramos P, Rapach DE, Reade JJ, Rostami-

Tabar B, Rubaszek M, Sermpinis G, Shang HL, Spiliotis E, Syntetos AA, Talagala PD, Talagala TS, Tashman L, Thomakos D, Thorarinsdottir T, Todini E, Trapero Arenas JR, Wang X, Winkler RL, Yusupova A, Ziel F (2022) Forecasting: theory and practice. International Journal of Forecasting In press, DOI 10.1016/j.ijforecast.2021.11.001, URL https://www.sciencedirect.com/science/article/pii/S0169207021001758

Rahikkala J, Hyrynsalmi S, Leppänen V (2015) Accounting Testing in Software Cost Estimation: A Case Study of the Current Practice and Impacts. In: 14th Symposium on Programming Languages and Software Tools, Tampere, Finland, p 15

Rahikkala J, Hyrynsalmi S, Leppänen V, Porres I (2018) The Role of Organisational Phenomena in Software Cost Estimation: A Case Study of Supporting and Hindering Factors. e-Informatica Software Engineering Journal 12(1):167–198, DOI 10.5277/e-Inf180107, URL http://www.e-informatyka.pl/attach/e-Informatica_-_Volume_12/eInformatica2018Art7.pdf

Satopää VA, Salikhov M, Tetlock PE, Mellers B (2021) Bias, Information, Noise: The BIN Model of Forecasting. Management Science 0(0):1–20, DOI 10.1287/mnsc.2020.3882, URL https://pubsonline.informs.org/doi/10.1287/mnsc.2020.3882, publisher: INFORMS

Schneider WJ, McGrew KS (2018) The Cattell–Horn–Carroll theory of cognitive abilities. In: Contemporary intellectual assessment: Theories, tests, and issues, 4th ed, 4th edn, The Guilford Press, New York, NY, US, pp 73–163

Shepperd M, Mair C, Jørgensen M (2018) An Experimental Evaluation of a De-biasing Intervention for Professional Software Developers. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, ACM, New York, NY, USA, SAC '18, pp 1510–1517, DOI 10.1145/3167132.3167293, URL http://doi.acm.org/10.1145/3167132.3167293, event-place: Pau, France

Simon HA (2000) Bounded rationality in social science: Today and tomorrow. Mind & Society 1(1):25–39, DOI 10.1007/BF02512227, URL https://doi.org/10.1007/BF02512227

Svedholm-Häkkinen AM, Lindeman M (2018) Actively open-minded thinking: development of a shortened scale and disentangling attitudes towards knowledge and people. Thinking & Reasoning 24(1):21–40, DOI 10.1080/13546783.2017.1378723, URL https://doi.org/10.1080/13546783.2017.1378723, publisher: Routledge _eprint: https://doi.org/10.1080/13546783.2017.1378723

Tamrakar R, Jørgensen M (2012) Does the use of Fibonacci numbers in planning poker affect effort estimates? In: 16th International Conference on Evaluation Assessment in Software Engineering (EASE 2012), IET, Ciudad Real, pp 228–232, DOI 10.1049/ic.2012.0030

Thaler RH (2018) From Cashews to Nudges: The Evolution of Behavioral Economics. American Economic Review 108(6):1265–1287, DOI 10.1257/aer.108.6.1265, URL https://www.aeaweb.org/articles?id=10.1257/aer.108.6.1265

Thaler RH, Sunstein CR (2021) Nudge: The Final Edition, final edition edn. Penguin Books, New York

Timon CE (2020) Defining the New Behavioral Science(s). Signs and Society 8(3):472–496, DOI 10.1086/710840, URL https://www.journals.uchicago.edu/doi/full/10.1086/710840, publisher: The University of Chicago Press

Trendowicz A, Münch J, Jeffery R (2011) State of the Practice in Software Effort Estimation: A Survey and Literature Review. In: Huzar Z, Koci R, Meyer B, Walter B, Zendulka J (eds) Software Engineering Techniques, Springer Berlin Heidelberg, Berlin, Heidelberg, Lecture Notes in Computer Science, pp 232–245

Tversky A, Kahneman D (1973) Availability: A heuristic for judging frequency and probability. Cognitive Psychology 5(2):207–232, DOI 10.1016/0010-0285(73)90033-9, URL https://www.sciencedirect.com/science/article/pii/0010028573900339

Tversky A, Kahneman D (1974) Judgment under Uncertainty: Heuristics and Biases. Science 185(4157):1124–1131, DOI 10.1126/science.185.4157.1124, URL https://www.science.org/doi/10.1126/science.185.4157.1124, publisher: American Association for the Advancement of Science

Usman M, Mendes E, Börstler J (2015) Effort estimation in agile software development: a survey on the state of the practice. In: Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering, Association for Computing Machinery, Nanjing, China, EASE '15, pp 1–10, DOI 10.1145/2745802.2745813, URL

https://doi.org/10.1145/2745802.2745813
Usman M, Börstler J, Petersen K (2017) An Effort Estimation Taxonomy for Agile Software Development. International Journal of Software Engineering and Knowledge Engineering 27(04):641–674, DOI 10.1142/S0218194017500243, URL https://www.worldscientific.com/doi/10.1142/S0218194017500243, publisher: World Scientific Publishing Co.
Usman M, Britto R, Damm LO, Börstler J (2018a) Effort estimation in large-scale software development: An industrial case study. Information and Software Technology 99:21–40, DOI 10.1016/j.infsof.2018.02.009, URL http://www.sciencedirect.com/science/article/pii/S0950584918300326
Usman M, Petersen K, Börstler J, Santos Neto P (2018b) Developing and using checklists to improve software effort estimation: A multi-case study. Journal of Systems and Software 146:286–309, DOI 10.1016/j.jss.2018.09.054, URL http://www.sciencedirect.com/science/article/pii/S0164121218302073
VandenBos G (2015) Bandwagon effect. URL https://doi.org/10.1037/14646-000
Wohlin C, Rainer A (2021) Challenges and recommendations to publishing and using credible evidence in software engineering. Information and Software Technology 134:106555, DOI 10.1016/j.infsof.2021.106555, URL https://www.sciencedirect.com/science/article/pii/S0950584921000409
Yamagishi T, Li Y, Takagishi H, Matsumoto Y, Kiyonari T (2014) In Search of Homo economicus. Psychological Science 25(9):1699–1711, DOI 10.1177/0956797614538065, URL https://doi.org/10.1177/0956797614538065, publisher: SAGE Publications Inc
Yang D, Wang Q, Li M, Yang Y, Ye K, Du J (2008) A survey on software cost estimation in the chinese software industry. In: Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement - ESEM '08, ACM Press, Kaiserslautern, Germany, p 253, DOI 10.1145/1414004.1414045, URL http://portal.acm.org/citation.cfm?doid=1414004.1414045
Zapata AH, Chaudron MRV (2013) An empirical study into the accuracy of it estimations and its influencing factors. International Journal of Software Engineering and Knowledge Engineering 23(04):409–432, DOI 10.1142/S0218194013400081, URL https://www.worldscientific.com/doi/abs/10.1142/S0218194013400081, publisher: World Scientific Publishing Co.
Zarour A, Zein S (2019) Software Development Estimation Techniques in Industrial Contexts: An Exploratory Multiple Case-Study. International Journal of Technology in Education and Science 3(2):72–84, URL https://eric.ed.gov/?id=EJ1227141, publisher: International Journal of Technology in Education and Science
Łabędzki M, Promiński P, Rybicki A, Wolski M (2017) Agile effort estimation in software development projects – case study. The Central European Review of Economics and Management 1(3), DOI 10.29015/cerem.359, number: 3

A Selected Papers in Phase 2

Table 7 presents the list of selected papers in Phase 2 (as described in Section 3.2.2).

Table 7

Id	Paper
1	Gandomani et al. (2014)
2	Grimstad and Jørgensen (2009)
3	Altaleb et al. (2020)
4	Yang et al. (2008)
5	Fægri (2010)
6	Rahikkala et al. (2015)
7	Łabędzki et al. (2017)
8	Matsubara et al. (2021b)
9	Arnuphaptrairong (2021)
10	Hughes (1996)
11	Shepperd et al. (2018)
12	Alhamed and Storer (2021)
13	Jørgensen and Escott (2022)
14	Magazinovic and Pernstål (2008)
15	Lederer and Prasad (1995)
16	Jørgensen et al. (2007)
17	Jørgensen (2011)
18	Usman et al. (2018b)
19	Usman et al. (2015)
20	Usman et al. (2018a)
21	Furulund and Molkken-stvold (2007)
22	Mahnič and Hovelja (2012)
23	Matos et al. (2013)
24	Jorgensen and Molokken-Ostvold (2004)
25	Jorgensen and Grimstad (2012)
26	Grapenthin et al. (2016)
27	Jørgensen and Halkjelsvik (2010)
28	Jørgensen and Gruschke (2009)
29	Rahikkala et al. (2018)
30	Jørgensen (2016)
31	Altaleb and Gravell (2019)
32	Arifin et al. (2017)
33	Gandomani et al. (2019)
34	Jørgensen and Halkjelsvik (2020)
35	Jørgensen (2015)
36	Halkjelsvik and Jorgensen (2011)
37	Haugen (2006)
38	Keaveney and Conboy (2006)
39	Tamrakar and Jørgensen (2012)
40	Moløkken-Østvold and Jørgensen (2004)
41	Lederer and Mirani (1990)
42	Glass et al. (2008)
43	Jørgensen (2013)
44	Lederer and Prasad (1991)
45	Moløkken-Østvold et al. (2008)