Software Development Estimation Biases: The Role of Interdependence

Magne Jørgensen[1,2] & Stein Grimstad[1]
[1]Simula Research Laboratory
[2]University of Oslo

***Abstract**: Software development effort estimates are frequently too low, which may lead to poor project plans and project failures. One reason for this bias seems to be that the effort estimates produced by software developers are affected by information that has no relevance for the actual use of effort. We attempted to acquire a better understanding of the underlying mechanisms and the robustness of this type of estimation bias. For this purpose, we hired 374 software developers working in outsourcing companies to participate in a set of three experiments. The experiments examined the connection between estimation bias and developer dimensions: Self-construal (how one sees oneself), thinking style, nationality, experience, skill, education, sex, and organizational role. We found that estimation bias was present along most of the studied dimensions. The most interesting finding may be that the estimation bias increased significantly with higher levels of interdependence, i.e., with stronger emphasis connectedness, social context and relationships. We propose that this connection may be enabled by an activation of one's self-construal when engaging in effort estimation, and, a connection between a more interdependent self-construal and increased search for indirect messages, lower ability to ignore irrelevant context, and a stronger emphasis on socially desirable responses.*

1 INTRODUCTION

Earlier, we reported results from studies that demonstrate how software professionals' effort estimates are affected by the following: clients' expectations about cost, even when the professionals are explicitly instructed to disregard those expectations (Jørgensen and Sjøberg 2001; Jørgensen and Grimstad 2008); changes in the format of requests for estimation or bidding that are irrelevant to the use of effort (Jørgensen and Carelius 2004; Jørgensen 2006; Jørgensen and Halkjelsvik 2008); variation in wording when describing the task to be estimated with no change in the requirement specification (Jørgensen and Grimstad 2008); the suggestion that future opportunities will arise given high quality, high productivity work on the current task (Jørgensen and Grimstad 2008); and the addition of different kinds of information to the requirement specification that have no relevance to the use of effort (Grimstad and Jørgensen 2007; Jørgensen and Grimstad 2008). Moreover, we have found that software developers and managers typically are unaware that they are affected, or at least strongly underestimate how much they have been affected; see, for example (Jørgensen and Sjøberg 2001). Applying field settings, we have demonstrated that the biases are not only experimental artifacts, but relevant in real-world bidding and estimation processes (Jørgensen and Carelius 2004; Jørgensen 2006; Jørgensen and Grimstad 2009). The large set of studies that demonstrate biases in human judgment and unawareness of this in a variety of domains and contexts (see (Pronin 2007) for a review) provides further support for the existence of systematic estimation biases.

An unfortunate consequence of the presence of the estimation biases is that the realism of software project plans is reduced. When the biases are in the direction of lower effort estimates, which is the typical case in software development (Moløkken and Jørgensen 2003), this may lead to project planning problems and, in the worst case, project failures. It is essential to understand the mechanisms that lead to the biases and how to avoid them, because such an understanding would, among other factors, enable the software industry to substantially improve its use of software project resources.

The goal of the study reported herein was to understand the mechanisms that lead to biases by studying a multicountry population of developers. Studying such a population enabled us to examine how the size of the estimation bias is connected with variables that may differ in different countries, e.g., how people see themselves in a greater context and their thinking styles. It also gave us the opportunity to test the robustness of previous findings with respect to different cultures. In spite of the fact that a substantial number of studies have been conducted on human estimation biases, we do not know much about the validity of the results and the effect sizes outside the cultures in which the studies typically conducted, i.e.,

Western cultures. In a review of studies on the effect of culture on human judgment and decision making, Weber and Hsee (2000) conclude that the levels of attention to culture "*are not just low, but are inadequate, and that progress could be made on multiple fronts by comparative cross-cultural research.*" Cole (1996) criticizes the strong reliance on US students as populations in studies on human judgment biases. It is likely that these claims are just as valid for studies of biases in software development effort estimation. The studies of such biases, including our own, do not investigate potential cultural differences and most results are based on students and developers from Western countries. Given that culture influences how we think (see, for example (Oyserman and Lee 2008)), our unsatisfactory knowledge about the context robustness and culture-dependence of biases in effort estimation may lead to incorrect recommendations to the software industry. More knowledge about how culturally dependent variables are connected with estimation bias may in particular be useful in outsourcing (offshoring) contexts in which clients and providers frequently come from different countries and may have different ways of thinking and communicating. The culturally related characteristics that we suspected were connected to estimation biases, based on previous research on this topic (Aaker and Maheswaran 1997; Weber and Hsee 2000; Choi, Dalal et al. 2003; Lehman, Chiu et al. 2004; Oyserman and Lee 2008), were people's concept of self and their thinking style. We measured these characteristics by using three questionnaire-based instruments that are commonly used in studies on culture. These instruments are designed to measure individuals' self-construal (Singelis 1994), degree of holistic/analytic thinking (Choi, Koo et al. 2007), and degree of need for cognition (Cacioppi and Petty 1982).

Our study also included analyses of the connection between estimation bias and different kinds of information about software developers, i.e., their length of experience, self-assessed development skill, education, sex, and organizational role. However, the main focus was the connection between culturally related characteristics and estimation biases.

The remainder of this paper is organized as follows: Section 2 describes the study design, the measures and the questionnaire-based instruments. Section 3 presents the results. Section 4 discusses the results and their implications. Section 5 concludes.

2 STUDY DESIGN

2.1 PARTICIPANTS

We searched the internet for outsourcing companies in different countries (applying the search string: *outsourcing* AND *software development*) and as a result contracted five Eastern European and seven Asian outsourcing companies to participate in our experiments. A total number of 374 software developers participated in the experiment, which lasted about 1.5 hours per participant. It is important to be extremely cautious when generalizing from this limited sample of developers in a few companies to the characteristics of all the developers in that country, and even more so for the corresponding region, i.e., the Asian or Eastern European region. For example, we doubt that it is meaningful to speak of an Indian (software development) culture, the existence of which would seem to be implied by Hofstader's measures and results (Hofstede 1980; Ford, Connelly et al. 2003). The population of the Indian subcontinent, for example, manifests a high number of distinct cultural differences (Majumdar 1961). Bearing in mind the difficulty of generalizing from the results of a study on a small sample to the culture of an entire country, our main purpose in using a multicountry population was to shed light on the robustness of previous findings and better understand the mechanisms of biases in effort estimation, rather than to characterize development cultures from different countries and regions.

The companies were paid normal hourly rates for their work. We required that all developers had at least six months of experience as software developers and were able to read and understand requirement specifications and instructions in English. Table 1 displays the number and the characteristics of participants from each country. One developer did not provide country information, leaving 373 responses.

Table 1: Participant Characteristics per Country

Country	N[1]	Compl. time[2]	Age[3]	Exp.[4]	Dev. skill[5]	Est. skill[6]	Engl. skill[7]	Fem.[8]	MSc[9]	Manager[10]
India	70	104	24	20	1.8	2.2	2.4	13%	42%	9%
Nepal	62	92	27	38	1.9	2.4	1.9	19%	13%	17%
Poland	50	100	29	56	1.9	3.0	2.4	10%	88%	12%
Romania	49	86	26	37	2.3	2.8	2.1	14%	45%	10%
Ukraine	78	91	27	52	2.1	2.9	2.7	10%	61%	21%
Vietnam	65	115	26	34	2.2	2.8	2.7	15%	11%	18%
Total	373	98	26	39	2.0	2.7	2.4	14%	41%	15%

[1] Number of participants from each country
[2] Mean number of minutes to complete the experiment. This time includes four additional tasks not analyzed in this paper. These four tasks took on average about 40% of the total time, leaving on average about 60 minutes for the tasks and questionnaires described in this paper.
[3] Mean age
[4] Mean number of months with experience as a software professional
[5] Mean self-assessed development skill value. Responses were made using the skill scale 1="Very good", 2="Good"; 3="Average"; 4="Acceptable"; 5="Poor"
[6] Mean self-assessed estimation skill value. Responses were made on the same skill scale as that used for development
[7] Mean self-assessed English skill value. Responses were made on the same skill scale as that used for development
[8] Proportion of female developers
[9] Proportion of developers with a Master's degree or higher
[10] Proportion of developers in the role of a manager (project manager, group leader, or general manager)

As can be seen in Table 1, the average values are about the same for all countries for most variables. Exceptions that are potentially of interest for the analysis of the collected estimation data, include:

- The lower mean experience level of the developers from India.
- The lower proportion of developers with a Master's degree from Nepal and Vietnam.
- The higher proportion of developers with a Master's degree from Poland and Ukraine.
- The lower proportion of developers in managerial roles for India, Romania, and Poland.

The skill data (development, estimation, and English skill) in Table 1 may illustrate the typical above-average effect and the limitations of the self-assessment of development skill, as implemented in the scale applied in our study. An unbiased self-assessment of skill would lead to a mean value of about 3 ("average"), given that the selection of developers is representative for the reference population of

developers and the interpretation of "average" is related to the same reference population, e.g., a national or company-specific population. An additional complicating factor of our type of skill scale is that some developers may have interpreted the values as absolutes rather than being relative to the attributes of others in their company, e.g., that it is reasonable to assess oneself as a "good" developer even when it is typical ("average") to be that good in their company. However, in spite of this limitation, the self-assessments may be interesting. For example, we found that the mean self-assessed skill was closer to "average" (3.0) when the developers were assessing their estimation skills (total mean of 2.7), than when assessing their English skills (total mean of 2.4) or their development skills (total mean of 2.0). As many as 77% of the developers selected one of the two categories better than "average", i.e., "good" or "very good", and only 2% one of the two categories worse than "average", i.e., "acceptable" or "poor", when assessing their own development skill. Interestingly, the most confident self-assessments were found among the Indian developers, where 90% assessed themselves to be better than average. Given that the Indian developers had, on average, less experience than the other developers in our study, this suggests that there are severe problems with the interpretation of self-assessed skill when using such scales, and hence in using the results to compare developers from different companies. The least confident skill assessments were found among Romanian developers, 61% of whom believed they were better than average. The size of the better-than-average effect may not only be dependent on culture (see, for example (Kurman 2002)), but also on sex. Seventy-nine percent of male developers and 69% of female developers assessed their development skill to be above average ($p=0.06$, Chi-square test of independence between sex and skill-category). There were only small and statistically non-significant differences in mean length of experience between the sexes. A sex difference with respect the better-than-average phenomenon is supported by several other studies; see (Visser, Ashton et al. 2008) for an overview and discussion. Frequently, there is an above-average effect when the skill is general or the problem to be solved is easy and a worse-than-average effect when the task is complex; see (Moore and Small 20007). This could explain our finding that more software developers felt more "average" when assessing their estimation skills, because they consider estimating to be a difficult task.

The participants in our study come from different countries, but had a similar educational background and a similar work environment, e.g., they have similar types of client, are given similar problem descriptions, use similar tools, and follow similar procedures when working. Our interviews with the managers of the companies revealed that the developers had experience in communicating with clients in English. For the most part, this communication consisted of reading and writing documents, and sending emails and instant messages. However, it did also include oral communication with the clients for some developers. The similarities in work and educational context may, on the one hand, reduce the effect of cultural differences, but may on the other hand allow deeply rooted culturally dependent differences to be isolated more effectively and result in useful information being generated for clients who are considering outsourcing their development work.

2.2 Experiment Material

The experiment material contained three parts. The first provided instructions for the completion of the experiment and asked for background information about the developer (Section 2.2.1). The second included the estimation tasks (Section 2.2.2). The third included the three questionnaires: one on self-construal, one on holistic/analytic thinking, and one on need-for-cognition (Section 2.2.3).

2.2.1 Instructions and Background Information

The instructions about the completion of the experiment included the following: i) the general purpose of the study was acquire information that would enable the improvement of effort estimation processes, ii) the participants should complete the tasks in the sequence they were presented and not look at the next page before all tasks on the current page were completed, iii) the participants should ask the person monitoring the experiment if a question was not clear or there were English words they did not understand (at least one of the authors was present at all times during the experiments), iv) all answers were strictly confidential and nobody would be able identify them or their company from the research

reports, v) they should *not* write their names on the responses, vi) the tasks (including four tasks not analyzed in this paper) would typically take 90-120 minutes to complete, but that they were allowed to use more time and would get paid for this if needed, vii) they could take a short break (5 minutes) if needed, and, viii) they should try to provide answers of as high quality as possible, because high quality answers were needed for our research. We monitored the work of the participants in the different locations. We observed that for the most part, they followed the instructions, with the exception that they sometimes asked each other or consulted a dictionary on their computer for support in understanding difficult English words instead of asking one of us. After reading the instructions, the participants completed the answers that provided the input to the summary presented in Table 1, i.e., answers related to nationality, age, length of experience, development skill, estimation skill, English skill, sex, education, and current role in the organization.

2.2.2 THE ESTIMATION TASKS

There were three estimation tasks to be completed, each of which had two variants. Two of the tasks had two estimation bias treatments and one of the tasks had one control group and one estimation bias treatment. The types of irrelevant and misleading information that were included as estimation bias treatments were based on results from our previous studies on the actual presence of such information in software requirement specifications (Jørgensen and Grimstad 2008).

Each developer was randomly assigned one of these variants. The sequence in which the estimation tasks were to be completed was also random. Using a randomized sequence for the estimation tasks had at least two advantages. 1) The developers, who sometimes sat close to each others, would be less tempted to look at their neighbours' estimates, because they most likely would be working on different estimation task. 2) There would be no systematic effect generated by completing the tasks in the same sequence, i.e., there would be no systematic effect from the task estimated immediately before on the following tasks. Sequence effects on effort estimates are demonstrated in (Grimstad and Jørgensen 2009).

Estimation Task 1 had one treatment group that received a low (1 line of code per work-hour) and one that received a high (200 lines of code per work-hour) development productivity "anchor", before estimating their own productivity in lines of code per work-hour on their last project. Anchoring is a cognitive bias that describes the common human tendency to rely too heavily, or "anchor," on one trait or piece of information when making decisions,; see, for example (Kahneman, Slovic et al. 1982). Except for the numerical anchors, all information and formulations were the same. It was expected that the developers who received the low anchor would produce lower estimates of productivity than those who received the high anchor. The anchor was presented to the participants as follows (The text inside the [ ] states the groups that received the information):

> [Both groups] *Lines of code (LOC) written per work-hour is used as an important variable in some effort estimation models. Consider your last completed project. Assume that the number of lines of code is interpreted as the number of lines of text written by you as programmer (not generated code), and that the number of work-hours includes all programming and unit testing related to the code you wrote, but not system or acceptance testing.*
> [Group Low Anchor] *Did you on average write more or less than 1 Line of Code per work-hours in your last project?*
> [Group High Anchor] *Did you on average write more or less than 200 Lines of Code per work-hours in your last project?*

Then, on the next page, the developers from both groups were asked to estimate the number of lines of code they wrote in their last project.

Estimation Task 2 had one treatment group that received the description that the development task to be estimated was a "minor extension" and one group the description that the task to be estimated constituted

"new functionality". The development task to be estimated and the rest of the instructions were exactly the same. The expected effect was that those informed that the development work consisted of a minor extension would estimate the required effort to be lower than those who were informed that the work consisted of the development of new functionality. The instruction was as follows:

[Both groups] *A few years ago MOSS Fotballklubb (MFK) – a Norwegian football club – developed a new ticket booking system. The ticket system has been a great financial success and saved MFK a lot of manual work. The web-based ticked system is written in Java and runs on a Weblogic platform.*
[Group Minor extension] *The management of MFK now wants a minor extension of the existing ticket system.*
[Group New functionality] *The management of MFK now wants to develop new functionality to improve the ticket system.*
[Both groups] *The management wants to offer their partners (sponsors etc.) an opportunity to sell football match tickets on the partners' own web pages. The partners who sell the most tickets will be rewarded by MFK with more free football match tickets. For this purpose, the following software needs to be developed: 1) a web service that the partners can use to book tickets, and 2) a reporting tool that shows how many tickets each partner has sold. The requirements are listed in the table below.* [Here a list of requirements was provided].

The list of requirements given to both groups included a mixture of requirements that were clearly relevant to effort and requirements that were completely irrelevant to the amount of development effort. The developers were asked to estimate the effort they would need to develop the web service and the reporting tool. The requirements will be sent to interested readers upon request to one of the authors.

Estimation Task 3 had one treatment group that received a specification of software with mainly effort-relevant information and one group that received the same effort-relevant information, but in addition received much information that had no intended relevance for the development effort. The actual irrelevance of the information for the purpose of effort estimation was confirmed by an independent, experienced software developer. The expected effect was that those who received more information would provide higher effort estimates, even though the extra information was not relevant for the use of effort. To give a couple of examples, a higher estimate might be a result of reading-between-the-lines (finding relevant information where none was intended), or the effect of an estimation strategy (judgment heuristic) that takes the length of the specification as an input variable to the resulting effort estimate. The instructions were as follows:

[Group Irrelevant] *The e-dating company sugar-date.com specializes in matching e-daters (people looking for a friend/partner/etc.) based on an extensive personal profile with 70 dimensions. The profile is based on questions that are carefully formulated and selected to establish and enable the matching of the preferences of both young and old. The matching process is performed by a sophisticated algorithm that has been developed by leading researchers in psychology. The matching process results in, for each of the relation to other members of a data base of people, a score between 0 and 100. This unique system has received world-wide attention. In fact, many of the features in their matching processes have led other major e-dating companies to change how they do their matching of e-daters. The e-dating system on sugar-date.com is also used for e-dater parties – these are large dating party events, held at up-class restaurants and clubs. At the premises, PCs, digital cameras and printers provide each e-dater with a card showing the photo of the 18 other e-daters present who are their best e-dating match (highest scores). As members arrive at the party, they are guided to one of many locations inside the premises where they can have their photo taken. The photo is attached to their profile and printed on the cards of those who have them as one of their 18 best matches. Many of the members are concerned that they look good on the photo (naturally), so several shots are often necessary. At present, the photographing process is quite slow, due to the many manual steps involved in taking, picking and storing the photos. The managers of sugar-date.com are as always eager to*

improve their business processes and are not satisfied with the current photo capturing.

[Both groups] *Assume that you have been asked to write a software application that captures photos of people using a web camera. The application should take one picture every time "ENTER" (the Return key) is pressed. New pictures are taken until the person is satisfied and selects one of them. During the picture taking and selection process, the last 20 pictures should be displayed on the screen. The selected picture should be stored on the hard disc as .jpg file with proper naming. The application should run on a Microsoft Windows XP platform and work with an Apple iSight web cam that features auto focus. This camera comes with a Java interface, and is connected to a Dell Latitude D800 laptop. The laptop is connected to the local area network available at the premises (10Mbit/s).*

2.2.3 QUESTIONNAIRES ON SELF-CONSTRUAL AND THINKING STYLE

After they had completed the estimation tasks, the developers completed three questionnaire-based tests measuring self-construal (interdependence and independence) and thinking style (holistic/analytic, Need-for-cognition). Typically, when applying these instruments, the text is translated into the respondents' native language. We did not do this, because we expected the developers of the outsourcing companies to be able to understand and use the English versions of the questionnaires. We evaluated this assumption by measuring the consistency of the responses, i.e., by use of the Cronbach alpha. If the English skill were too low, this would be represented by low Cronbach alpha values. The use of English as the questionnaire language for all countries may have some advantages. For example, it may not be possible to translate some words from one language to another without losing information and/or changing the meaning. Interestingly, the use of English may influence people whose native language is other than English to make slightly more individualistic responses (Oyserman and Lee 2008). This means that the cultural differences in responses may get slightly diluted by the use of English. However, it is expected that the effect of this would be small.

Although the instruments we used have been validated in several studies, they have many limitations. For example, Heine, Lehman et al. (2002) point out that people in some cultures may be more modest and play down their abilities, in some cultures it may be the norm to answer more towards the middle of a scale, in some cultures it may be the norm to disguise one's actual opinions or behavior, and that people from different cultures may use different reference groups in their self-reported values. These limitations are real. However, as reported in (Grimm and Church 1999), controlling for response biases of the above types typically seems to lead to trivial changes in effect sizes and no changes in conclusions.

The instrument that we used to measure self-construal (interdependence and independence) was the widely used questionnaire presented in (Singelis 1994). That instrument is based on work by Markus and Kitayama (1991), who proposed that people in the West hold an independent view of the self that emphasizes the separateness, internal attributes, and uniqueness of individuals and that many nonWestern peoples hold an interdependent image of self that stresses connectedness, social context, and relationships. Recent studies suggest that interdependence and independence are not mutually exclusive. In fact, several studies find a positive correlation between these two constructs; see, for example (Kolstad and Horpestad 2009), in which it is reported that Chilean students were found, on average, to be both more independent and more interdependent than Norwegian students. The self-construal questionnaire includes 24 statements, 12 of which are related to interdependence and 12 to independence. The respondents are asked to indicate the extent to which they agree or disagree with each of these 24 statements on a seven-point Likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree). An example of a statement (related to interdependence) is "*It is important for me to maintain harmony within my group*". Originally, the questionnaire was developed for studies on university students, which influenced the formulation of a few of the questions. To fit the context of work professionals, we replaced all (four) references to school contexts with the corresponding references to a work place, e.g., "*I would*

offer my seat in a bus to my professor." is replaced with "*I would offer my seat in a bus to my boss*". We thought that an increased level of interdependence could lead to greater awareness of the context and social desirable behaviour. It is possible that this could lead to larger estimation biases and lower effort estimates.

The instrument that we used to measure holistic/analytic thinking is described in (Choi, Koo et al. 2007). The instrument is based on 24 statements on people's thinking style. The respondents indicate the degree to which they agree or disagree with each of these statements on a seven-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). An example of a statement is: "*It is more desirable to take the middle ground than go to extremes.*" The instrument is based on the idea that some cultures (particularly in Asia) have a stronger tendency to view every element in the world as somehow interconnected, whereas other (particularly some Western) have a stronger tendency to view the universe as composed of independent objects. It is possible that those with a stronger holistic view will evaluate more information as relevant and as a consequence provide more biased effort estimates; see, for example, Study 6 in (Choi, Koo et al. 2007).

The instrument that we used to measure need-for-cognition is supposed to measure individual differences in the engagement in and enjoyment of effortful information processing. The 18-statement questionnaire that we used is described in (Cacioppo, Petty et al. 1984). The respondents are asked to indicate the extent to which they agree or disagree with the statements on a nine-point Likert scale ranging from -4 (very strong disagreement) to +4 (very strong agreement). An example of a statement is: "*I only think as hard as I have to.*" The effect of higher need-for-cognition on judgmental biases (priming effects) has recently been studied by Petty, Demarree et al. (2008). They found that as the need-for-cognition increases, the magnitude of the bias increases with "subtle" primes but decreases with "blatant" primes. Transferring these results to the context of our study, it may be expected that people who enjoy and engage in more effortful thinking may be less affected when the misleading or irrelevant information is obvious (blatant), as in Estimation Task 1, and more affected when it is less obvious (subtle), as in Estimation Tasks 2 and 3.

2.3 Completion of the Experiment

The experiments were conducted at the premises of the participating companies. The developers either sat in their ordinary work place or in meeting rooms arranged for this purpose. Typically, a manager from the company introduced us to the developers and explained the nature of the study (paid work, important to do serious work to support the research on improvement of estimation processes, they would get feedback on the results, etc.) Then, the experimenter(s) introduced himself/themselves and went through the instructions described in Section 2.2.1. The developers were requested to ask questions or request clarification on points that they were unsure of before they started. After this introduction, the estimation material was handed out and the experiment started. One or both of the authors of this article were present at all times during the experiment and were available to answer questions. When a developer had completed all the tasks, he/she gave the material to the experimenters, who ensured that all answers were complete. If there were questions that were not answered, the developer had to go back to answer them. As far as we observed, all participants (except for two, which we excluded) took the tasks seriously and had acceptable quality of the answers.

2.4 Measures and Analyses

On the basis of our previous experience, we expected the distribution of estimates to be non-normal, i.e., skewed towards high effort estimates, and that there would be several very high effort estimates that would have a large effect on the mean values. For these reasons, we decided to use the median effort estimation as the main comparison measure and the rank-based, non-parametric Kruskal-Wallis test to test the statistical significance of group differences in effort estimates in our analyses. We used the two-sided Kruskal-Wallis test (even when we had some expectations about the direction of the effect) and did not adjust the p-value for the number of statistical tests. Reasons for this include that we consider the p-

values to be only one out of several indicators of relationship (finding the same relationship in more than one estimation task and correspondence with previous findings are examples of others) and the difficulty we have in determining an appropriate level of p-value adjustment for several tests (some of the tests can be considered to be families of tests). However, the robustness and significance of the results will be discussed with these limitations of analysis in mind.

In the setting of our study, it makes no sense to say that one individual developer is biased, because we do not know how much effort he or she would actually use. That being so, we based our measure of estimation bias on the difference in median estimates between the treatment groups. For example, if the group of developers who received the description of a task as "minor extension" systematically provided lower estimates than those who received the description of the same task as "new functionality", this would mean that the estimates of one or both groups had been biased by the description.
The extent to which information that is included in a requirement specification or a task description is relevant or irrelevant for the use of effort depends on whether the client (in this case us) intended to provide information this way. We tried to design the biasing information so that it is reasonable to assume that a client would not intend to convey effort relevant messages this way.

3 RESULTS

3.1 GENERAL TREATMENT EFFECTS

Table 2 shows that all three treatments were successful in that they yielded significant differences in the effort estimates in the expected directions. As described earlier, we used the difference in median estimates between the treatments as our measure of the "estimation bias", i.e., as an indication of how much the numerical anchor (Estimation Task 1), the textual anchor (Estimation Task 2) and the irrelevant information (Estimation Task 3) affected the responses. Boxplots of the distributions of the estimates are displayed in Figures 1-3.

Table 2: General Treatment Effects

Task	**Estimation Task 1**		**Estimation Task 2**		**Estimation Task 2**	
Treatment	**Low anch.**	**Low anch.**	**"Minor ext."**	**"New func."**	**Control**	**Irr. inf.**
Estimates (median)	15	100	85	107	66	90
Difference median	85		22		24	
Krukal Wallis-test	p<0.001		p=0.01		p=0.01	

Figure 1: Estimation Task 1

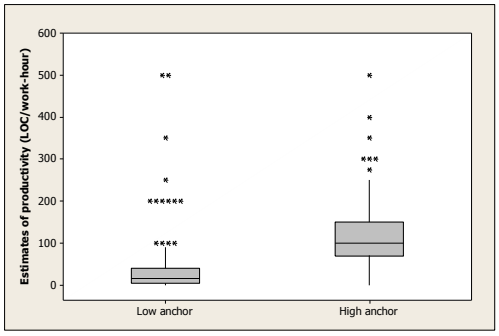


Figure 2: Estimation Task 2

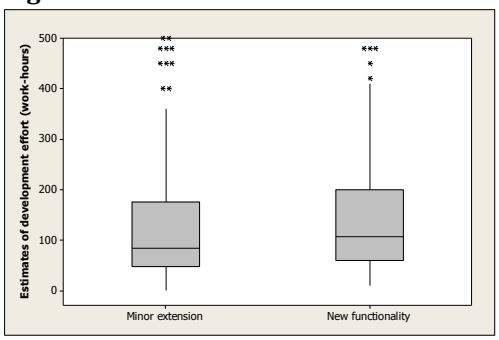
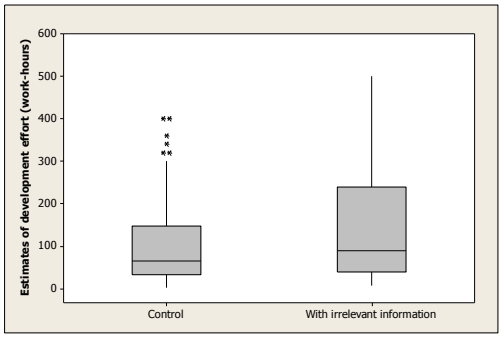


Figure 3: Estimation Task 3



As an indication of the effect sizes we log-transformed the estimates, which gave a close to normal distribution of them, and calculated Cohen's d on the log-transformed values. This gave d=0.7 for

Estimation Task 1, d=0.3 for Estimation Task 2, and d=0.4 for Estimation Task 3. This suggests that there was a medium/large treatment effect on Estimation Task 1, and a small/medium treatment effect on Tasks 2 and 3.

In the following sections, the treatment effects are analysed relative to the collected variables, i.e., self-construal and thinking style (Section 3.2), country (Section 3.3), experience, skill, education, sex, and organizational role (Section 3.4).

3.2 SELF-CONSTRUAL AND THINKING STYLE

3.2.1 THE INSTRUMENTS' VALIDITY AND CONNECTIONS WITH OTHER VARIABLES

It is recommended that scales used in basic research should have a Cronbach alpha of at least 0.7 (Nunnally 1978). A lower score would suggest that there was too much inconsistency in the responses. A Cronbach alpha analysis of the instruments of self-construal (interdependence and independence) and thinking styles (holistic/analytic thinking and need-for-cognition) of our study yielded values of 0.7-0.8 for all of them, except for the instrument measuring independence, which had a Cronbach alpha of only 0.6. Given the low Cronbach alpha level of the independence scale, we decided to remove it from further analyses.

Table 3 describes the mean values with standard deviation of the interdependence, holistic/analytic thinking, and need-for-cognition instruments for each country. The distributions of responses for each of the instruments were close to normal and we included an ANOVA test of differences in mean values between the countries.

Table 3: Mean Scores (with Standard Deviations) of the Instruments per Country

Country	Interdependence[1]	Holistic/Analytic	Need-for-cognition
India	5.1 (0.7)	4.7 (0.6)	0.7 (0,7)
Nepal	5.0 (0.9)	5.0 (0.6)	1.1 (0.8)
Vietnam	4.5 (0.8)	4.6 (0.6)	0.7 (0.6)
Poland	4.4 (0.7)	4.9 (0.5)	1.5 (0.7)
Romania	4.4 (0.6)	5.0 (0.7)	1.6 (0.9)
Ukraine	4.2 (0.7)	4.8 (0.6)	1.3 (0.9)
All countries	4.6 (0.8)	4.8 (0.6)	1.1 (0.9)
ANOVA	p<0.001	p=0.004	p<0.001

Table 3 shows that there were significant differences in scores between the different countries for all instruments, especially for interdependence and need-for-cognition. An indication of the validity of the interdependence values collected by us is provided by the study on Nepalese university students described in (Agrawal and Maheswaran 2005), which found a mean interdependence value of 5.1, similar to the one we found. Furthermore, the mean interdependence of the countries in Eastern Europe (Poland, Romania and Ukraine) is similar to those found among European Americans; see, for example (Coon and Kemmermeier 2001). We were unable to find relevant studies regarding the other countries and instruments.

The instruments are, to some degree, intercorrelated, as can be seen in Table 4.

Table 4: Correlation between the Instruments

	Interdependence	Holistic/Analytic
Holistic/Analytic	0.3 (p<0.001)	
Need-for-cognition	-0.07 (p=0.17)	0.3 (p<0.001)

[1] The Cronbach alpha improved considerably when Question 3 ("My happiness depends on the happiness of those around me") was removed from the instrument. The improvement was particularly evident in the case of Indian developers. Given the potential problems with question 3, we decided to remove it from the analysis.

While we would expect a correlation between more interdependence and more holistic thinking (because both measures correlate with stronger emphasis on the context), the correlation between more holistic thinking and higher need-for-cognition is more difficult to explain. Perhaps the attitude that everything is connected (higher degree of holistic thinking) somehow is connected with a stronger desire to reflect on and engage with problems that are not closely connected to solving daily problems, i.e., a higher need-for-cognition.

Table 5 displays the mean values of the instruments for different categories of experience, skill, education, sex, and roles. It also includes an ANOVA-test of differences in mean values for those values. The sum of observations per variable is sometimes lower than the total number of participants (374), due to incomplete or incomprehensible responses. In particular, the descriptions of level of education turned out to be difficult to interpret.

Table 5: Self-construal and Thinking Style vs Experience, Skill, Education, Sex and Role

Variable	Category	Interdependence	Holistic/Analytic	Need-for-cognition
Experience	High (>=35 months, n=185)	4.6	4.9	1.2
	Low (< 35 months, n=181)	4.7	4.8	1.0
	ANOVA-test	p=0.1	p=0.2	p=0.01
General development skill	Average or worse (n=86)	4.6	4.9	1.0
	Good or better (n=288)	4.7	4.8	1.2
	ANOVA-test	p=0.6	p=0.2	p=0.3
Education	Bachelor (n=197)	4.7	4.8	1.0
	Master or higher (n=139)	4.5	4.9	1.3
	ANOVA-test	p=0.03	p=0.3	p=0.004
Sex	Female (n=51)	4.7	4.9	0.9
	Male (n=322)	4.6	4.8	1.2
	ANOVA-test	p=0.5	p=0.5	p=0.06
Organizat-ional role	Developer (n=318)	4.6	4.8	1,1
	Manager (n=56)	4.6	5.0	1.2
	ANOVA-test	p=0.9	p=0.08	p=0.5

An emphasis of differences with p<=0.1 suggests that developers who have more experience are higher on need-for-cognition and slightly lower on interdependence, that those who have higher education (Master's degree or higher) were lower on interdependence and higher on need-for-cognition, that the male developers were higher on need-for-cognition than the female developers, and that those in the role of a manager had a more holistic thinking style. Of course, it is difficult to specify the degree to which this reflects actual attitudes and behaviour or just differences in answering behavior based on our study alone.

3.2.2 Estimation Biases Related to Self-construal and Thinking Style

To simplify the analysis, we categorized the developers as belonging to either a "High" or a "Low" group, according to their scores on the self-construal and thinking style instruments. The developers who had the 50% lowest scores on an instrument were included in the "Low" group, while those with the 50% highest scores were included in the "High" group. Table 6 shows the effect of the treatment (the estimation bias) per category, where the results of the two-sided Kruskal-Wallis test of differences in mean ranks are described through * (p<=0.1), ** (p<=0.01) and *** (p<=0.001).

Table 6: Estimation Bias vs Self-construal and Thinking Style

Instrument		Estimation Task 1			Estimation Task 2			Estimation Task 3		
		Low anch.	High anch.	Diff.	Min. ext.	New func.	Diff.	Contr.	Irr. inf.	Diff.
Interdep-endence	High	15	100	85***	70	100	30**	50	100	50**
	Low	15	105	85***	120	120	0	80	80	0
Holistic/ Ananalytic	High	14	100	86***	80	100	20*	60	90	30*
	Low	20	100	80***	93	120	27	80	90	10
Need-for-cognition	High	11	100	89***	80	107	27*	66	98	32*
	Low	20	100	80***	90	110	20	72	80	8*

The results from Table 6 suggest that:

- The numerical anchor in Estimation Task 1 had a strong effect on both categories for all instruments. The estimation bias (difference in median values) is about the same for all instrument categories.
- Those in the interdependence category "High" were affected much more by the treatments implemented in Estimation Tasks 2 and 3 than those in the category "Low", who may not have been affected at all.
- Those in the holistic/analytic and those in the need-for-cognition category "High" were affected more by the treatment implemented in Estimation Task 3 (effect of irrelevant information) than those in the category "Low". The main difference in estimates is, surprisingly, in the control group, not in the group that received the irrelevant information. As can be seen, those who are high on interdependence, holism or need-for-cognition gave systematically lower effort estimates on Estimation Tasks 2 and 3.
- Those in the interdependence category "High" gave much lower effort estimates in the control group of Estimation Task 3 than those in the interdependence category "Low" (Kruskal-Wallis, p=0.02). The difference in estimates related to Estimation Task 2 for the "minor extension" group is also statistically significant (Kruskal-Wallis, p=0.001). This suggests that the effort estimates of those high on interdependence may, in general, be lower[2]. We discuss the lower estimates of those in interdependency category "High" in more detail in Section 4.

To analyze possible interactions between the instruments, we conducted pair-wise analyses of their combined effects for Estimation Tasks 2-3, i.e., those tasks for which belonging to the low or high instrument category made a difference; see Tables 7-9. The purpose of this analysis was to determine whether there were one or two variables that explained the observed biases and whether the effect of the others was due to correlation with that variable. The values presented are the median difference between the treatment groups, e.g., Table 7 shows that the difference in median estimate between those who received the description "minor extension" and those who received the description "new functionality" was 20 work-hours for the subset of developers who had a low degree of interdependence and a high level of holistic thinking. The stars (*, **, ***) indicate, as before, the level of statistical significance (p<=0.1, p<=0.01, p<=0.001). Notice that the power of the tests is reduced with a lower number of developers in each subset of developers in the tests in Table 7, which would lead us to expect less statistically significant relationships.

Table 7: Combined Effect of Holistic/Analytic Thinking and Interdependence

	Estimation Task 2		**Estimation Task 3**	
	Interdep=Low	**Interdep=High**	**Interdep=Low**	**Interdep=High**
Holistic=High	20	27*	28	30*
Holistic=Low	0	25*	-18	90*

Table 8: Combined Effect of Holism and Need-for-Cognition (NFC)

	Estimation Task 2		**Estimation Task 3**	
	NFC=Low	**NFC=High**	**NFC=Low**	**NFC=High**
Holistic=High	5	30	4	27
Holistic=Low	18	16	32*	20*

Table 9: Combined Effect of Need-for-Cognition and Interdependence

	Estimation Task 2		**Estimation Task 3**	
	Interdep=Low	**Interdep=High**	**Interdep=Low**	**Interdep=High**
NFC=High	20	40*	8	34*
NFC=Low	0	25*	18	31*

[2] We also assigned a fourth estimation task, which is not included in this analysis because it was not related to the analysis of estimation biases. Those in the interdependence category "High" gave significantly (p=0.007) lower effort estimates than those in the category "Low" for that task, as well.

Tables 7-9 suggest that the estimation bias is connected mainly with a high level of interdependence and that the observed connections to need-for-cognition and, particularly, holistic/analytic thinking is strongly reduced in the group with a low level of interdependence. We interpret this as indicating that the dominating effect on the size and significance of estimation bias is the level of interdependence.

3.3 COUNTRY

Table 10 displays the estimation biases per country.

Table 10: Estimation Biases vs Country

	Estimation Task 1			Estimation Task 2			Estimation Task 3		
Group	**Low anch.**	**High anch.**	**Diff**	**"Minor ext."**	**"New func."**	**Diff**	**Control**	**Irr. Inf.**	**Diff**
India	25	150	125***	63	80	17	30	58	28*
Nepal	11	120	109***	50	152	102*	80	90	10
Poland	12	100	88***	102	110	8	80	100	20
Romania	10	70	60***	95	100	5	50	70	20
Ukraine	10	100	90***	120	120	0	60	200	140*
Vietnam	25	100	75***	90	120	30	100	100	0

The lack of more statistically significant differences may be due to the lower power of the statistical tests when splitting the developers into country-based groups. While most differences related to Estimation Tasks 2 and 3 are moderate or small, there are two exceptions: The estimates of the Nepali developers on Estimation Tasks 2 and those of the Ukrainian developers on Estimation Task 3. We were unable to find reasonable explanations for these exceptionally large deviations through further analysis of the data. Another interesting observation was that, in general, the developers from India had the lowest effort estimates. We find no clear connection between region (Asia vs Eastern Europe) and estimation bias. In sum, Table 10 does not document a strong, systematic effect of nationality or region on estimation bias.

3.4 EXPERIENCE, SKILL, EDUCATION, SEX AND ROLE

Table 11 displays values that indicate the estimation biases per variable (experience, skill, education, sex, and role) and estimation task.

Table 11: Estimation Bias vs Experience, Skill, Education, Sex, and Role

		Estimation Task 1			Estimation Task 2			Estimation Task 3		
Variable	**Category**	**Low anch.**	**High anch.**	**Diff.**	**Min. ext.**	**New func.**	**Diff.**	**Control**	**Irr. Inf.**	**Diff.**
Experience	High (>=35 months, n=185)	15	100	85***	80	120	40*	75	80	5*
	Low (<35 months, n=181)	11	100	89***	90	98	8	56	100	44*
General dev. skill	Av. or worse (n=86)	11	80	69***	120	105	-15	96	100	4
	Better than av. (n=288)	18	100	82***	80	104	24**	53	80	27**
Education	Bachelor (n=197)	15	100	85***	82	100	18	58	100	62*
	Master (n=139)	10	100	90***	90	120	30*	60	64	4
Sex	Female (n=51)	10	100	90***	90	120	30	80	80	0
	Male (n=323)	15	100	85***	84	100	16*	64	98	34**
Organizational role	Developer (n=318)	15	100	85***	88	102	14*	60	100	40**
	Manager (n=56)	11	90	79***	80	140	60*	80	80	0

Potentially interesting observations possible to derive from Table 11 include the following:

- The effect of the productivity anchor in Estimation Task 1 is about equally strong among all categories.
- Being a manager, having more experience, or having a Master's degree were each connected with the textual anchor ("minor extension" or "new functionality") having a stronger effect and adding irrelevant information having a lesser effect. These variables (role, experience, and skill) are far from independent (Chi-square Role and Experience, $p<0.001$, Chi-square Role and Education, $p=0.009$, Chi-square Education and Experience, $p=0.001$) and may, to some extent, represent the same underlying relationship. In other words, there may be some estimation biases that increase with education, experience, and managerial role and others that decrease.

- Those who assessed their development skill to be "better than average" were affected more by the biasing information than those who assessed it to be "average or less". The high proportion of developers who described themselves as being "better than average" (77%), may suggest that self-assessed skill is, to some extent, a measure of "over-confidence". That being so, it is possible to argue that the better-than-average group is over-populated with over-confident developers. This may explain why this group has a stronger estimation bias.
- The female developers were not affected by the addition of irrelevant information (Estimation Task 3), while the male developers were affected quite a lot. There are too few female developers to draw strong conclusions here, but there may be interesting sex-based differences among software developers that are worthy of further investigation.

The data displayed in Table 11 suggest that estimation biases are present for all categories of developers, but that the size and impact may vary with category and type of treatment. The degree of statistical significance depends on the power of the test. Note that some groups have few subjects, particularly the groups with female developers (n=51), self-assessed skill "average or worse" (n=86), and participants in managerial roles (n=56). The statistical power of these groups will be lower than the other groups.

4 Discussion

The two most interesting results from our studies may be summarized as follows:

- Estimation biases seem to be fairly robust and present in all categorie of developers and among developers from countries in both Asia and Eastern Europe. Previously, biases in software development effort estimation have been observed in Western Europe, e.g., (Jørgensen and Sjøberg 2004) and the North America, e.g., (Aranda and Easterbroook 2005). Similar, findings on bias in human judgment that are robust across cultures have been found in other fields, e.g., consumer research on persuasion (Aaker and Maheswaran 1997). In total, this supports the belief that bias in effort estimation is a global phenomenon.
- Although the presence of estimation biases seems to be global, the effect sizes may well depend on culturally-dependent variables and other contextual factors. In particular, our results suggest that a higher level of interdependence is connected with a higher level of estimation bias, at least for the types of treatments implemented in Estimation Task 2 (misleading, textual anchor) and Estimation Task 3 (addition of irrelevant information). Interestingly, not only were the estimation biases stronger for those who had a high interdependence score, but also the effort estimates were substantially lower

In the remaining part of this section, we discuss the finding that as interdependence increases, biases increase and estimates are lower. The discussion is divided into how higher levels of interdependence may be connected with the following: i) more search for indirect meaning, i.e., a difference in conversational indirectness (Section 4.1), ii) more attention to the context and/or less ability to ignore irrelevant context (Section 4.2), and, iii) stronger emphasis on socially desirable responses (Section 4.3). Our goal here is to understand why and how an increased level of interdependence is connected with an increase in estimation biases and lower, possibly more over-optimistic, effort estimates. This may produce information that is useful for software clients who produce material for bidding rounds and requests for cost or effort estimates, and for providers when designing their estimation processes and training their developers in effort estimation. We discuss limitations of our study in Section 4.4.

4.1 Search for Indirect meaning

There may be differences in the degree to which software developers are willing and able to search for indirect meanings when receiving and interpreting instructions and information. For example, some developers may, consciously or unconsciously, draw the conclusion that when a development task is described as a "minor extension", the client expects the developer to expend less effort on quality assurance than when the same task is described as "new functionality". In situations in which the

information sender (the client) and the receiver (the developer) have the same indirect style of communication, this may work well. In other situations, such as when the software is being developed offshore, in which case the producers and receivers of the requirement specification typically have different styles of communication, a stronger urge to find indirect meanings can lead to estimation biases. An example of such a situation is where the client did not mean to communicate that lower quality work would be accepted when using the term "minor extension" instead of "new functionality". Similarly, when a lot of effort irrelevant information is added to the requirement specification, those who are more prone to looking for indirect meanings may be more likely to find unintended requirements or functionality and, as a consequence, be more biased in their effort estimates.

Several studies, see for example (Gudykunst, Matsumoto et al. 1996; Hara and Kim 2004), find that those with highly developed interdependent self-construals are more likely to look for and find indirect messages in communication. This fits with our finding that those who have high scores on interdependence were affected to a greater extent by the misleading and irrelevant information in Estimation Tasks 2 and 3. However, it does not necessarily explain why those who have high scores on interdependence made systematically lower effort estimates. We try to explain this tendency towards lower effort estimates in Section 4.3, by appealing to an increase in attention to social desirability.

It is unclear to what degree an increase in search for indirect meanings with higher levels of interdependence fits with the lack of connection between interdependence and estimation bias for Estimation Task 1. Perhaps the numerical anchor in Estimation Task 1 was so much lower (in the LOW group) and higher (in the HIGH group) than the realistic productivity that the developers felt no inclination to search for indirect meanings. It is also possible that a simple numerical anchor in itself does not incline people to search for indirect meanings.

4.2 Attention to the Context

An increase in attention to the context is related to an increase in the search for indirect meanings, especially when this meaning can be derived from the immediate context, but is not exactly the same phenomenon. As it is understood in this paper, a tendency to pay more attention to the context is a tendency to focus less on the most important (the focal) aspects of the stimuli and more on the context. This tendency may vary among people and cultures. For example, Masuda and Nisbett (2001) found that Japanese people were more likely to report on background features (the context) when describing an animated scene of fish and other underwater life with which they had been presented than North American people. North American people, on the other hand, were more likely to observe and report on changes on the focal aspect, in this case the fish, of the animated scene. The results of several papers, e.g., (Lee, Hallahan et al. 1996; Choi, Nisbett et al. 1999), suggest that those who have higher scores for interdependence tend to have a higher attention to the context. For example, Korean participants (who have, on average, higher scores for interdependence) had a lower threshold for considering information to be relevant than North American participants (who have, on average, lower scores for interdependence) in a study of a criminal case (Choi, Dalal et al. 2003). The review by Lehman, Chiu and Schaller (2004) provides a review of the cultural differences with respect to assessment of the relevance of the context.

Increased attention to the context can both be an advantage and a disadvantage. If the included context is relevant it can be an advantage, but if the context contains irrelevant or misleading information, it can be a disadvantage. This relationship is demonstrated clearly in the paper by Krishna, Zhou and Zhang (2008), who found that biases in judgment were greater among those who paid more attention to the context when the context needed to be excluded in the mental processing, but lower when the context needed to be included.

From the above results, it may be concluded that the observed increase in estimation bias in Estimation Tasks 2 and 3 in those who have higher scores for interdependence is, at least partly, a result of a lower ability to ignore the context when it is irrelevant or misleading. Therefore, people who have higher scores for interdependence may, on average, be more prone to include the context in their judgment. This ability

is clearly useful when the context is relevant. On the other hand, this ability seems to go in hand with a lower ability to exclude the context when it is irrelevant or misleading. In our study, the context information was intentionally misleading and/or irrelevant and may have led to the stronger estimation biases of those who had higher scores for interdependence.

On first reflection, the described relationship may seem obvious: being more strongly interdependent leads people to pay more attention to the context, which in turn leads to more context impact. However, the observed relationships are not necessarily intuitive. To some extent, it is surprising that high-level cultural context characteristics, such as how people interact socially with family, college acquaintances and friends, predict how much they are affected by low-level context information, e.g., describing a task as a "minor extension" vs "new functionality" as in our Estimation Task 2. Even biases that are due to optical illusions, see (Krishna, Zhou et al. 2008), seems to be predicted by the high-level construct interdependence. It seems that somehow, estimation tasks activate how we see ourselves (our self-construal), which to some extent is culturally dependent and measured by our measure of interdependence.

4.3 Attention to Social Desirability

The two factors that were proposed above, i.e., a greater search for indirect meanings and paying more attention to the context, may be able to explain the stronger estimation bias of Estimation Tasks 2 and 3 for those developers who are more interdependent. However, they are not able to explain the tendency towards lower effort estimates of the same group of developers. This section suggests that this tendency may be due to a connection between paying more attention to social desirable responses and being more strongly interdependent.

Lalwani, Shavitt and Johnson (2006) give a summary of relevant research on this topic, together with a report of the results of their own studies. They summarize previous studies to suggest that collectivism (which is assumed to be linked to stronger interdependence and weaker independence (see, for example (Markus and Kitayama 1991))) is more likely than individualism to be associated with a tendency to dissemble in order to present oneself in a socially desirable manner. However, they also report from studies that suggest the opposite and state that: "… *both individualism and collectivism may be associated with socially desirable response style but in distinct ways. A primary goal associated with individualism is to view the self in unique and positive terms, but a primary concern associated with collectivism is to save face and maintain a good relationship with others.* (p. 166)". The authors suggest the cultural difference is related more to projecting an enhanced image of oneself via self-deception among those from individualistic cultures, and more to impression management (understood as appearing more normatively appropriate in the given context) among those from collectivistic cultures. Similar results, with respect how views on interdependence and independence are connected with different types of motivated reasoning, are reported in (Hannover, Pöhlmann et al. 2005). The tendency of people to wish to appear more normatively appropriate in collectivistic cultures may explain the tendency towards lower effort estimates for Tasks 2 and 3 for those developers who are more interdependent. Lower effort estimates may be perceived as being more in accordance with what clients or managers would like to receive, i.e., more socially appropriate. It is also possible that a stronger focus on what the client would perceive as a desirable response, e.g., through describing a task to be "minor", explains the larger estimation bias of those who are more strongly interdependent.

The three explanations provided in Sections 4.1-4.3 are not mutually exclusive. On the contrary, they support each other and enable an explanation of the finding that the largest difference between those who are more or less interdependent is to be found in the "Minor extension" group of Estimation Task 2. That task was the only one in which all three explanations are likely to have contributed in the same direction, i.e., describing the task as "minor" may affect the developers who are more interdependent to: i) be more engaged in searching for indirect meanings that support a low use of effort, ii) be affected to a greater extent by the context (of minor task), and, iii) respond in a socially desirable manner by providing low effort estimates. In cases in which the irrelevant or misleading information led to estimation biases in the

direction of higher effort estimates, the effect of searching for indirect meanings/more context dependency may act in the opposite direction of socially desirable responses, i.e., the bias will be diluted. Of course, the foregoing argument is based on the assumption that socially desirable responses in our setting will typically be in the direction of lower effort estimates. The finding that the estimates of those in the control group of Estimation Task 3 was lower for those who are more interdependent may provide some support for this assumption.

4.4 LIMITATIONS

Studies that analyze cultural variables typically have a lot of limitations. Ours is no exception. The following limitations are the most serious for the validity and generality of our results:

- We studied individual estimation and development work only. This is particularly limiting, because the phenomenon that turned out to be most interesting has to do with how a person sees himself/herself in a group. Although the effects that we observed suggest that the effect of culture is still there and can be substantial, it would be interesting to study the effect on larger tasks that involve people working in groups. This is particularly so in software development, because software projects typically involve people working in groups. We think that a group setting would activate/trigger/prime more interdependent thinking and lead to larger estimation biases (see also (Hannover, Pöhlmann et al. 2005)). However, the results of our study do not warrant us in drawing such a conclusion. The lack of a focus on group work does not mean that there will be no group effect or that our work is without relevance. Even when estimating individually, which is not unusual in practice, people are likely to see themselves as a member of one or more groups.
- Both authors have a Western background and may communicate more easily with people from a similar background, e.g., Eastern Europeans, than with Asians. That being so, it is possible that there are other differences in communication than those intended that led to the observed differences. The same East-West problem may be related to many concepts upon which we relied heavily in our study, e.g., the concept of difference in self-construal. In support of the view that there is some substance in the concept of self-construal is the finding discussed in Section 4.2, where self-construal predicted the degree of bias with respect to the quite basic phenomenon of optical illusions (Krishna, Zhou et al. 2008).
- Optimally, we would like to study the estimation biases in field settings with full realism. This was not possible for practical reasons; as a result, care should be exercised when considering generalizing the results. It is for example possible that effort when participating in an experiment, with higher time pressure and knowing that the projects will be not be started, are different from those in more typical field settings, even though we were close to ordinary clients and the work was paid.
- We conducted more than 100 statistical tests of differences. That being so, it is to be expected that some of the differences that we found to be statistically significant are “false positives”, i.e., spurious relationships. Therefore, the test results should be interpreted with care, with the greatest emphasis being placed on those that have the lowest p-values and largest effect sizes, that are replicated in more tasks, and that are supported by evidence from other studies.
- Some of the statistical tests have low power, due to the low number of observations and strong variation in effort estimates. This means that it is likely that there will be many “false negatives”, i.e., that there may be essential differences that were not identified in our study via the assessment of statistical significance.

5 CONCLUSION AND IMPLICATIONS

There are many variables that have the potential to affect biases in software development effort estimation. The study presented herein examines culturally related (self-construal and thinking style-related) and other (nationality, length of experience, self-assessed skill, education, sex, and organizational role) variables among developers in six outsourcing countries from Eastern Europe (Poland, Romania, Ukraine) and Asia (India, Nepal, Vietnam).

We consider the main contributions of our study to be the following:

- Bias in effort estimation seems to be present for developers from all the studied countries. We were unable to find strong and systematic differences between countries or regions (Eastern Europe and Asia).
- In spite of the lack of any strong and systematic difference between the countries and regions, there may be culturally related variables that are useful for understanding the mechanisms by which estimation biases occur. In particular, a developer's level of interdependence (emphasis on connectedness, social context, and relationship) seems to be connected systematically with how much he or she was affected by irrelevant and misleading information and with lower effort estimates. This finding seems to correspond with recent finding from other domains, e.g., findings that connect higher levels of interdependence with a lower ability to ignore irrelevant context and higher ability to include relevant context (Krishna, Zhou et al. 2008).

An increased level of interdependence seems to be connected with more engagement in search for indirect meanings, more awareness of the context, and a higher desire to act in a socially desirable manner. The mechanism that underlies these connections may be, as pointed out in (Markus and Kitayama 1991), *when* : "*If a cognitive activity implicates the self, the outcome of this activity will depend on the nature of the self-system.*" The estimation of software development effort may clearly implicate the self, e.g., it may implicate a developer's assessment of his skill or of his value for the employer, and this way determine how much misleading anchors and irrelevant information affect the estimates.

In earlier studies, see for example (Jørgensen and Grimstad 2008), we have documented that the only safe way to avoid the effect of irrelevant and misleading information is to avoid it completely when estimating effort. As soon as one has been exposed to this kind of information, it seems to be very difficult to return to the previous state. A major reason for this is that the effect is partly unconscious and the developers are unaware of the size of the effect. Frequently, this means that they will underestimate how much they have been affected. The results of the study presented herein further strengthen this previously formulated message. This is especially so for developers who score high on level of interdependence. While it may be more likely to find developers scoring higher on interdependence in some companies and countries than others, we found developers scoring high on this measure in all countries. This entails that it is essential to address the issue of potentially misleading and irrelevant information, regardless of whether the effort on software development is estimated by an East European or Asian outsourcing company. The methods and principles we propose in (Jørgensen and Grimstad 2008; Jørgensen 2009) describe ways to do this.

Acknowledgement: This research project was funded by The Research Council of Norway through the industry-project EVISOFT.

References:

1. Aaker, J. L. and Maheswaran, D. (1997). "The effect of cultural orientation on persuasion." Journal of Consumer Research 24(3): 315-328.
2. Agrawal, N. and Maheswaran, D. (2005). "The effects of self-construal and commitment on persuasion." Journal of Consumer Research 31(March): 841-849.
3. Aranda, J. and Easterbroook, S. M. (2005). Anchoring and Adjustment in Software Estimation. European Software Engineering Conference / ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE'05), Lisbon, Portugal.
4. Cacioppi, J. T. and Petty, R. E. (1982). "The need for cognition." Journal of Personality and Social Psychology 42(1): 116-131.
5. Cacioppo, J. T., Petty, R. E. and Kao, C. F. (1984). "The efficient assessment of need for cognition." Journal of Personality Assessment 48(3): 306-307.

6. Choi, I., Dalal, R., Kim-Preto, C. and Park, H. (2003). "Culture and judgment of causal relevance." Journal of Personality and Social Psychology 84(1): 46-59.
7. Choi, I., Koo, M. and Choi, J. A. (2007). "Individual differences in analytic versus holistic thinking." Personality and Social Psychology Bulletin 33: 691-705.
8. Choi, I., Nisbett, R. E. and Norenzayan, A. (1999). "Causal attribution acrosss cultures: Variation and universality." Psychological Bulletin 125(January): 47-63.
9. Cole, M. (1996). Cultural psychology: A once and and future discipline. Cambridge, MA, Harvard University Press.
10. Coon, H. M. and Kemmermeier, M. (2001). "Cultural orientation in the United States: (Re)examining differences among ethnic groups." Journal of Cross-cultural Psychology 32: 348-364.
11. Ford, D. P., Connelly, C. E. and Meister, D. B. (2003). "Information system research and Hofstede's Culture's consequences: An uneasy and incomplete partnership." IEEE Transactions on Engineering Management 50(1): 8-25.
12. Grimm, S. D. and Church, A. T. (1999). "A cross-cultural study of response biases in personality measures." Journal of Research on Personality 33: 415-441.
13. Grimstad, S. and Jørgensen, M. (2007). The Impact of Irrelevant Information on Estimates of Software Development Effort. The Australian Software Engineering Conference, Melbourne, IEEE Computer Society.
14. Grimstad, S. and Jørgensen, M. (2009). "A Preliminary Study of Sequence Effects in Judgment-based Software Development Work- Effort Estimation." To appear in: IET Software.
15. Gudykunst, W. B., Matsumoto, Y., ting-Toomey, S., Nishida, T., Kim, K. and Heyman, S. (1996). "The influence of cultural individualism-collectivism, self-construals, and individual values on communication styles across cultures." Human Communication Research 22: 510-543.
16. Hannover, B., Pöhlmann, C. and Springer, A. (2005). "Implications of independent versus independent self-knowledge for motivated social cognition: The semantic procedural interface model of the self." Self and Identity 4: 159-175.
17. Hara, K. and Kim, M.-S. (2004). "The effect of self-construal on conversational indirectness." International Journal of Intercultural Relations 28: 1-18.
18. Heine, S. J., Lehman, D. R., Peng, K. and Greenholtz, J. (2002). "What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect." Journal of Personality and Social Psychology 82(6): 903-918.
19. Hofstede, G. (1980). Culture's Consequences: International Differences in Work-related Values. Newbury Park, CA, Sage.
20. Jørgensen, M. (2006). "The Effects of the Format of Software Project Bidding Processes." International Journal of Project Management 24(6): 522-528.
21. Jørgensen, M. (2009). "How to Avoid Selecting Providers with Bids Based on Over-Optimistic Cost Estimates." To appear in IEEE Software May/June.
22. Jørgensen, M. and Carelius, G. (2004). "An empirical study of software project bidding." IEEE Transactions on Software Engineering 30(12): 953-969.
23. Jørgensen, M. and Grimstad, S. (2008). "Avoiding Irrelevant and Misleading Information when Estimating Development Effort." IEEE Software May/June: 78-83.
24. Jørgensen, M. and Grimstad, S. (2009). "The Impact of Irrelevant and Misleading Information on Software Development Effort Estimates: A Randomized Controlled Field Experiment." submitted to IEEE Transactions on Software Engineering (a preliminary version can be downloaded from: simula.no/research/engineering/publications/Simula.SE.299).
25. Jørgensen, M. and Halkjelsvik, T. (2008). "The Effects of Request Formats on Judgment-based Effort Estimation." To appear in Journal of Systems and Software.
26. Jørgensen, M. and Sjøberg, D. I. K. (2001). "Impact of effort estimates on software project work." Information and Software Technology 43(15): 939-948.
27. Jørgensen, M. and Sjøberg, D. I. K. (2004). "The impact of customer expectation on software development effort estimates." International Journal of Project Management 22: 317-325.

28. Kahneman, D., Slovic, P. and Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge, United Kingdom, Cambridge University Press.
29. Kolstad, A. and Horpestad, S. (2009). "Self-construal in Chile and Norway: Implications for cultural differences in individualism and collectivism." Journal of Cross-cultural Psychology 40: 274-281.
30. Krishna, A., Zhou, R. and Zhang, S. (2008). "The effect of self-construal on spatial judgments." Journal of Consumer Research 35: 337-348.
31. Kurman, J. (2002). "Measure to the Modesty Response Measured Cross-Cultural Differences in Self-Enhancement and the Sensitivity of the Self-Enhancement." Cross-Cultural Research 36(1): 73-95.
32. Lalwani, A. K., Shavitt, S. and Johnson, T. (2006). "What is the relation between cultural orientation and socially desirable responding? " Journal of Personality and Social Psychology 90(1): 165-178.
33. Lee, F., Hallahan, M. and Herzog, T. (1996). "Explaining real life events: How culture and domain shape attribution." Personality and Social Psychology 22(July): 732-741.
34. Lehman, D. R., Chiu, C. and Schaller, M. (2004). "Psychology and culture." Annual Review of Psychology 55: 690-714.
35. Majumdar, D. N. (1961). Races and cultures of India. New York, Asia Pub. House.
36. Markus, H. R. and Kitayama, S. (1991). "Culture and the self: Implications for cognition, emotions and motivations." Psychological Review 98: 224-253.
37. Masuda, T. and Nisbett, R. E. (2001). "Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans." Journal of Personality and Social Psychology 81(November): 922-934.
38. Moløkken, K. and Jørgensen, M. (2003). A review of software surveys on software effort estimation. International Symposium on Empirical Software Engineering, Rome, Italy, Simula Res. Lab. Lysaker Norway.
39. Moore, D. A. and Small, D. A. (20007). "Error and Bias in Comparative Judgment: On Being Both Better and Worse Than We Think We Are." Journal of Personality and Social Psychology 92(6): 972-989.
40. Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York, McGraw-Hill.
41. Oyserman, D. and Lee, S. W. S. (2008). "Does culture influence what and how we think? Effects of priming individualism and collectivism." Psychological Bulletin 134(2): 311-342.
42. Petty, R. E., DeMarree, K. G., Brinol, P., Horcajo, J. and Strathman, A. J. (2008). "Need for Cognition Can Magnify or Attenuate Priming Effects in Social Judgment." Personality and Social Psychology Bulletin 34(7): 900-912.
43. Pronin, E. (2007). "Perception and misperception of bias in human judgment." Trends in Cognitive Sciences 11(1): 37-43.
44. Singelis, T. M. (1994). "The measurement of independent and interdependent self-construals." Personality and Social Psychology Bulletin 20(5): 580-591.
45. Visser, B. A., Ashton, M. C. and Vernon, P. A. (2008). "What Makes You Think You're so Smart? Measured Abilities, Personality, and Sex Differences in Relation to Self-Estimates of Multiple Intelligences " Journal of Individual Differences 29(1): 35-44.
46. Weber, E. U. and Hsee, C. K. (2000). "Culture and Individual Judgement and Decision Making." Applied Psychology: An International Review49(1): 32-61.