Evidence of the Presence of Bias in Subjective Metrics: Analysis within a Family of Experiments

Alejandrina Aranda
Universidad Politécnica de Madrid
Campus de Montegancedo
28660 Boadilla del Monte (Spain)
am.aranda@alumnos.upm.es

Oscar Dieste
Universidad Politécnica de Madrid
Campus de Montegancedo
28660 Boadilla del Monte (Spain)
+34 91 336 5011
odieste@fi.upm.es

Natalia Juristo
Universidad Politécnica de Madrid
Campus de Montegancedo
28660 Boadilla del Monte (Spain)
+34 91 336 6922
natalia@fi.upm.es

ABSTRACT

Context: Measurement is crucial and important to empirical software engineering. Although reliability and validity are two important properties warranting consideration in measurement processes, they may be influenced by random or systematic error (bias) depending on which metric is used. **Aim:** Check whether, the simple subjective metrics used in empirical software engineering studies are prone to bias. **Method:** Comparison of the reliability of a family of empirical studies on requirements elicitation that explore the same phenomenon using different design types and objective and subjective metrics. **Results:** The objectively measured variables (experience and knowledge) tend to achieve more reliable results, whereas subjective metrics using Likert scales (expertise and familiarity) tend to be influenced by systematic error or bias. **Conclusions:** Studies that predominantly use variables measured subjectively, like opinion polls or expert opinion acquisition, must take every care to prevent bias that can result in incorrect results.

1. INTRODUCTION

There is well-grounded need in software engineering (SE) to determine the reliability and validity of any metric [1]. Reliability refers to the consistency of the metric. A metric is reliable when repeated measures of the same object yield similar results, differing only due to some random error. Validity refers to the metric actually representing the construct being investigated. A misrepresentation is determined by the existence of systematic error, which makes the metric incorrect [2].

Determining the reliability and validity of a metric is a relatively costly process, and therefore it is often only used for complex measuring instruments, such as lengthy questionnaires (e.g., [3]), especially questionnaires which are reusable across more than one investigation.

Fewer precautions are taken when the metrics are simple and well defined, target the construct being investigated and are specific to the ongoing empirical study. This type of metrics are often used to measure the dependent and independent variables in Requirements Engineering (e.g., [4-6]). They are frequently used in SE too (e.g., [7]).This is not surprising, because these measures appear to have all the essential properties: *content validity* (they are logically linked to the construct being investigated [8]), *construct validity* (internal consistency does not pose a problem [2]) and *reliability* (they have a precise definition which encourages consistent measurement [2]). We will address *criterion validity* below.

We have observed, however, that even simple metrics such as the above can be troublesome. We have measured the same underlying construct in a family of five empirical studies on requirements elicitation using simple metrics such as analyst experience (in years) or expertise (using a Likert scale). The result is that this second metric, which might be termed "subjective" appears to be prone to systematic error, i.e., bias, even in cases where *criterion validity* can be proven with respect to the first metric (which might be termed "objective"). We obviously do not claim that *all* subjective metrics are susceptible to this problem. But, considering how often they are used in the literature, we believe that this finding is worth reporting.

The paper is structured as follows. Section 2 describes the family of studies and the working methodology. Section 3 reports the calculations that suggest that there is bias. Finally, Section 4 justifies the existence of bias by identifying the source of the bias in each particular case.

2. METHODOLOGY

2.1 Research Question

The research question that we set out to answer in this paper is:

> **RQ:** What empirical support is there for our observation that simple subjective metrics are prone to bias?

2.2 Description of the Family of Studies

To answer the research question we will analyse the data generated by a family of five empirical studies, shown in Table 1. Those studies were conducted with different design and control degree, using objective and subjective metrics. In all cases the goal of the empirical studies was to study what analyst

characteristics influence elicitation process effectiveness. The experimental task was to perform an elicitation session for one (in the correlational studies), two (in the experimental study) or four (in the quasi-experimental study) software systems belonging to different problem domains. All the sessions were carried out using the interview technique.

The experimental population was composed of subjects with differing levels of experience, knowledge and proficiency in activities related to requirements elicitation. The subjects were Madrid Technical University postgraduate students and professionals with experience and knowledge of computing and similar fields, including researchers, faculty and PhD students at several universities and research groups.

In the elicitation sessions both students and professionals played the role of requirements analysts and the experimenter played the role of the client. At the end of the elicitation session, the subjects started the process of consolidation of the information brought up in the elicitation sessions.

2.3 Dependent and Independent Variables

The dependent variable is the effectiveness of the elicitation process. This variable was measured as a percentage of the information elicited in the elicitation sessions. It was possible to calculate this percentage because we had an exhaustive description of all the relevant information about the problem domain for each experimental task, which we used as a baseline for comparison.

The independent variables are related to the constructs EXPERIENCE and DOMAIN KNOWLEDGE. To avoid confusion, we will use SMALL CAPS to represent constructs and *italics* for independent variables. These constructs were measured differently:

- The EXPERIENCE construct was measured from two viewpoints: a) as years of subject *experience* and b) as the *expertise* that subjects personally believe that they have in requirements utilization. *Expertise* was measured on a five-point Likert scale (1=low and 5=high). These variables were studied in: C-2007, C-2011 and C-2012.
- Like EXPERIENCE, the DOMAIN KNOWLEDGE construct was measured: a) as the *familiarity* that subjects think they have with the problem domain, where familiarity was measured using a three-point Likert scale (1=low and 3=high). This variable has been tested in: C-2007, C-2011 and C-2012; and b) a factor termed *domain knowledge*, with two levels: known and unknown. As we know the experimental population beforehand, we can predict the problem domains of which they have knowledge and choose a software system belonging to the respective domains for the purposes of elicitation. This variable was measured in: E-2013 and Q-2012.

2.4 Analysis procedure

Pearson correlation was the statistical procedure used to evaluate the effects of the independent variables in correlation studies. Kendall's τ had to be used in the case of ordinal variables. However, the calculations yield similar results to Pearson's r. As there are no differences, we use r because it is the most common procedure. Mixed models and ANOVA were used in the quasi-experimental and experimental studies, but for the sake of comparability we have expressed those results in terms of correlations. All the calculations have been carried out using SPSS 21. Significance level is α=5%, 2-tailed.

These empirical studies have differences related to their experimental design (e.g., additional factors). These factors exert a further effect on the response variable. On this ground, we had to adjust the data in order to eliminate the effects of the additional factors to assure that the results of the different studies were fully comparable with each other. Otherwise, the calculation of some correlations (see below) could over or under estimate the underlying effect. We obtained some percentage adjustments that were subtracted from the total effectiveness achieved by the different experimental subjects, before the above-mentioned correlations were calculated. Table 1 merely lists the number of subjects actually used in the calculations, although there are more experimental units in both Q-2012 and E-2013.

2.5 Assessment Strategy

We assume, in accordance with the vast majority of the software engineering community, that the relationship between the constructs EXPERIENCE, DOMAIN KNOWLEDGE and the response variable effectiveness, is directly proportional, i.e.:

- Experienced analysts are more effective (/gather more information) than novice analysts.
- Analysts are more effective (/gather more information) when they analyse a known problem domain.

Consequently, we should note that the effect of the objective variables (*experience*, *domain knowledge*) and subjective variables (*expertise*, *familiarity*) in the family of studies completed tends to be positive, i.e., $r > 0$. We do not know the real value of r; we merely assume that it is positive. Nor do we expect $r > 0$ always, as the results of several experiments can differ due exclusively to natural variation. However, if a tendency for $r < 0$ were observed, this could be proof of the existence of bias.

2.6 Validity Threats

The research is affected by three validity threats:

- The precision of a statistical correlation is directly proportional to the sample size. The total sample size of all five studies is around 100 experimental units, and there are large differences from one study to another. As the sample size is small, possible adverse effects should be considered when discussing the results of the correlation analysis.
- There could be between-study differences in terms of both moderator variables and quality. Although this cannot be ruled out (and some studies do, in fact, yield results contrary to the global trends), we believe that the fact that the studies are five practically exact replications, which are executed internally, should be safeguard enough against this threat.
- There may be differences within the community about the expected effects of the EXPERIENCE and DOMAIN KNOWLEDGE constructs. Now, we believe that a positive correlation with effectiveness is the majority sentiment within the SE community.

3. RESULTS

This section examines the difference between taking an objective or subjective measure of the same underlying construct.

3.1 Experience vs. Expertise

Table 2 shows the relationship between subject experience, and effectiveness for each study. With the exception of C-2007, the

Table 1. Family of empirical studies about requirements elicitation

Studies	Empirical Study	No. Subjects	No. Experimental Units	Subject Type	Dependent Variable	Independent Variables		
						Factor	Obj.	Subj.
C-2007	Correlational study	7	-	Students	Effectiveness	-	Experience	Expertise / Familiarity
C-2011	Correlational study	16	-	Students		-	Experience	Expertise / Familiarity
C-2012	Correlational study	21	-	Professionals		-	Experience	Expertise / Familiarity
Q-2012	Quasi-experiment (repeated measures)	7	28	Students		Domain knowledge	-	-
E-2013	Experiment (totally randomized)	8	-	Students			-	-

correlation between *experience* and effectiveness is positive, that is, analysts with more years of experience tend to be more effective at eliciting requirements than novice analysts. This ties in with the hypothesis stated in Section 2.5 that *experience* has positive effects on effectiveness.

In the case of C-2007, the correlation is negative (r=-.348), i.e., novice analysts are better than experienced analysts at capturing relevant information about the domain of discourse. Although some empirical studies corroborate this result (e.g., [9]) the negative correlation is, in the light of the results of C-2011 and C-2012 (r=.004 and r=.338 respectively), mostly likely to be due to the small number of experimental subjects used (N=7). Another possible reason for the negative correlation of C-2007 is the type of subjects. The subjects in this study, unlike C-2011 and C-2012, had an eminently technical profile (e.g., programmer, architect). Our research suggests that experience acquired in this type of activities may turn out not to be useful for capturing requirements. An in-depth justification of this claim is beyond the scope of this paper.

Table 2. Relationship of analyst effectiveness to requirements *experience* and *expertise*

Study	Experience			Expertise		
	r	p-value	N	*r*	p-value	N
C-2007	-.348	.444	7	-	-	-
C-2011	.004	.988	16	-.010	.970	16
C-2012	.338	.134	21	-.047	.839	21
C-Total	.180	.242	44	-.012	.943	37

Separately, none of the studies really have a high number of experimental subjects (N=7, 16 and 21, respectively), and therefore the resulting correlation coefficients are not very precise. The significance levels corroborate this circumstance. One way of overcoming the limitations of the individual studies is to synthesize the results of all three correlational studies, which produces more precise and generalizable results. The calculations (C-Total) have yielded a result consistent with commonly accepted knowledge in RE maintaining that analyst experience plays an important role: the relationship between *experience* and effectiveness is positive (r=.180), and this result is nearer to statistical significance (p-value=.242, 2-tailed), which increases our confidence in its results.

The correlations between *expertise* and effectiveness are also shown in Table 2, matched against the correlations between *experience* and effectiveness. The calculations were similar, but the correlations were negative for all separate (*Expertise* was not measured in C-2007) and combined studies, that is, the results were consistently contrary to what generally accepted knowledge in SE suggests.

In our opinion, the conflicting values are due to metrics measuring the variables having a different bias propensity. The variables whose values can be established objectively (*experience)* result in more accurate results than those that are subjective (*expertise*) to the point that the observed effects may be contrary ($r_{experience} > 0$ but $r_{expertise} < 0$). This is noteworthy, bearing in mind that there is *criterion-related validity* between *expertise* and *experience* (r=0.441, p-value= 0.006), indicating that both variables are strongly related. We justify the presence of the bias to which we are referring in Section 4. But, first, let us look at another example of a simple variable that, depending on the type of measure (objective, subjective), yields opposite effects.

3.2 Domain knowledge vs. Familiarity

Table 3 shows the correlations between *domain knowledge* and subject effectiveness. The results are assorted. On the one hand, the value yielded by Q-2012 is as expected (r = .234). As regards E-2013, the results were surprising; the correlation coefficient (r = -.564) is, contrary to expected, negative. Both results are not statistically significant (p-value > 0.05).

Table 3. Relationship of analyst effectiveness to *domain knowledge*

Experiment	Knowledge		
	r	p-value	N
Q-2012	.234	.260	25
E-2013	-.564	.145	8
QE-Total	0.095	0.599	33

Table 4 shows the correlations between *familiarity* and subject effectiveness. The results show that, with the sole exception of C-2011, the effect of *familiarity* is contrary to generally accepted knowledge in SE, i.e., the correlation coefficients have negative values. Additionally, there is *criterion-related validity* between *familiarity* and *domain knowledge* (r=0.613 p-value=0.000). This is very likely to be the same phenomenon as observed in Section 3.1. We believe that the risk of bias caused by the subjective way

in which the *familiarity* variable is measured is the source of the negative correlation coefficients.

Table 4. Relationship of analyst effectiveness to *familiarity*

Experiment	Familiarity		
	r	p-value	N
C-2007	-.548	.203	7
C-2011	.164	.545	16
C-2012	-.090	.698	21
C-Total	-.105	.499	44

It remains to explain why the correlation coefficient is negative for E-2013. We believe that this is because of the small number of experimental subjects, which makes the calculations of the coefficients of correlation very sensitive to the differences between individual effectivenesses. Indeed, this was why we aggregated the correlational studies as C-Total in Section 3.1. Likewise, the aggregation of Q-2012 and E-2013 as QE-Total yields an estimate of r=0.095, p-value=0.599 for *domain knowledge* (see Table 3), which is in line with the expected value in SE.

4. SOURCE OF BIAS

The source of bias can differ depending on the particular variable that is being measured. In the case of the *expertise* variable, the source of bias appears to be fairly clear and therefore we will use it as an illustrative example.

As discussed in Section 2, we think that REQUIREMENTS EXPERIENCE does influence analyst effectiveness, as predicted by the *experience* variable. The reason why the *expertise* variable does not manage to predict effectiveness lies in the fact that people generally rate their expertise above what is warranted by their experience. Overrating is precisely the bias that is influencing *expertise*. **Figure 1** is illustrative in this respect.

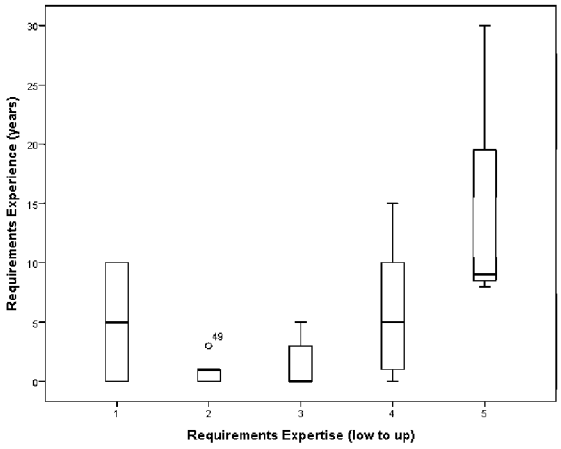

Figure 1. Box plot illustrating the relationship between requirements *expertise* and *experience*

Many of the subjects with fewer than 5 years' experience rate their expertise within the 1 to 3 or even 1 to 4 range. Only the more experienced subjects (over 5 years) rate their expertise as 4 to 5, but as they represent a relatively small share of sample data, they are unable to revert the negative overall trend. Additionally (see Table 2), *experience* seems to have small effect (r=.180), which means that the bias in the measurement of *expertise* renders the positive trends negative easily. Generally, as more data is collected, the effect of *expertise* should ostensibly be slightly positive or negative depending on the ratio among subjects with few and many years of experience.

5. CONCLUSIONS

We have shown that simple and reasonably well-defined metrics are susceptible to bias. The variables whose values can be established objectively (*experience, domain knowledge)* result in more accurate results than those that are subjective (*expertise, familiarity*). Although our data obviously represent only a minuscule step forward in the study of measurement validity, we believe that all studies that predominantly use variables measured subjectively, like opinion polls or expert opinion acquisition, must take every care to prevent bias that can result in incorrect results.

6. ACKNOWLEDGMENT

This research has been partially supported by the grants TIN-2011-23216 (Spanish Ministry of Economy and Competitiveness), FiDiPro (Finnish Funding Agency for Technology and Innovation) ESEIL and the Itaipu Binacional Postgraduate Grant / Fundación Parque Tecnológico Itaipu (Paraguay-Brazil).

7. REFERENCES

[1] B. Kitchenham, S. L. Pfleeger and N. Fenton, "Towards a framework for software measurement validation," *IEEE Transactions on Software Engineering,* vol. 21, pp. 929-944, 1995. DOI: http://dx.doi.org/10.1109/32.489070

[2] S. H. Kan, *Metrics and Models in Software Quality Engineering.* Addison-Wesley, 2003.

[3] T. Dyba, "An Instrument for Measuring the Key Factors of Success in Software Process Improvement," *Empirical Software Engineering,* vol. 5, pp. 357-390, 2000. DOI: http://dx.doi.org/10.1023/A:1009800404137

[4] A. Niknafs and D. M. Berry, "The impact of domain knowledge on the effectiveness of requirements idea generation during requirements elicitation," in *20th IEEE Int'l Requirements Engineering Conference (RE),* 2012, pp. 181-190. DOI: http://dx.doi.org/10.1109/RE.2012.6345802

[5] M. G. Pitts and G. J. Browne, "Stopping Behavior of Systems Analysts During Information Requirements Elicitation," *Journal of Management Information Systems,* vol. 21, pp. 203-226, 2004.

[6] Ö. Albayrak and J. Carver, "Investigation of individual factors impacting the effectiveness of requirements inspections: a replicated experiment," *Empirical Software Engineering,* vol. 19, pp. 241-266, 2014. DOI: http://dx.doi.org/10.1007/s10664-012-9221-0

[7] J. Jung, K. Hoefig, D. Domis, A. Jedlitschka and M. Hiller, "Experimental Comparison of Two Safety Analysis Methods and Its Replication," *7th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM),* 2013, pp. 223-232. DOI: http://dx.doi.org/10.1109/ESEM.2013.59

[8] A. Anastasi and S. Urbina, *Psychological Testing.* Prentice Hall, 1997.

[9] G. M. Marakas and J. J. Elam, "Semantic Structuring in Analyst Acquisition and Representation of Facts in Requirements Analysis," *Information Systems Research,* vol. 9, pp. 37-63, 1998. DOI: http://dx.doi.org/10.1287/isre.9.1.37