2011 International Symposium on Empirical Software Engineering and Measurement

Quantitative determination of the relationship between internal validity and bias in software engineering experiments: consequences for systematic literature reviews

Oscar Dieste
Facultad de Informática
Universidad Politécnica de Madrid
Madrid, Spain
odieste@fi.upm.es

Anna Grimán
Dept. Procesos y Sistemas
Universidad Simón Bolívar
Caracas, Venezuela
agriman@usb.ve

Natalia Juristo
Facultad de Informática
Universidad Politécnica de Madrid
Madrid, Spain
odieste@fi.upm.es

Himanshu Saxena
Facultad de Informática
Universidad Politécnica de Madrid
Madrid, Spain
himanshusaxena22@gmail.com

Abstract— **Quality assessment is one of the activities performed as part of systematic literature reviews. It is commonly accepted that a good quality experiment is bias free. Bias is considered to be related to internal validity (e.g., how adequately the experiment is planned, executed and analysed). Quality assessment is usually conducted using checklists and quality scales. It has not yet been proven; however, that quality is related to experimental bias. Aim: Identify whether there is a relationship between internal validity and bias in software engineering experiments. Method: We built a quality scale to determine the quality of the studies, which we applied to 28 experiments included in two systematic literature reviews. We proposed an objective indicator of experimental bias, which we applied to the same 28 experiments. Finally, we analysed the correlations between the quality scores and the proposed measure of bias. Results: We failed to find a relationship between the global quality score (resulting from the quality scale) and bias; however, we did identify interesting correlations between bias and some particular aspects of internal validity measured by the instrument. Conclusions: There is an empirically provable relationship between internal validity and bias. It is feasible to apply quality assessment in systematic literature reviews, subject to limits on the internal validity aspects for consideration.**

Keywords*- Systematic Literature Review (SLR); Quality Assessment (QA) of experiments; Checklist; Quality Scale*

I. INTRODUCTION

Systematic literature review (SLR) "is a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest" [1]. According to Kitchenham [1], the SLR process involves: (i) identifying experiments about a particular research topic, (ii) selecting the studies relevant to the research, (iii) including/excluding studies based on their quality assessment, (iv) extracting the data from the selected studies, and (v), eventually, aggregating the data to generate pieces of knowledge.

The SLR process can be quite challenging, especially if it has to account for a large number of experiments, because, in such cases, a lot of information usually has to be extracted and aggregated, too. But, not all the experiments in a SLR make a contribution to the final aggregation, and some of the experiments could even create bias in the aggregated result, as they have imprecise or unreliable outcomes.

Quality assessment (QA) can evaluate the importance of individual studies based on their internal validity and remove poor quality studies from a SLR, reducing the number of studies to be analysed and making the SLR process more efficient and less error prone. Generally, checklists and quality scales are used to do this, as they are able to score or assign a quality level to each experiment included in or excluded from the review process. This score will be used to include or exclude the experiment from the review process or to weight its contribution during the synthesis process.

Following the guidelines for other disciplines, Kitchenham [1] and Biolchini et al. [2] recommend a detailed QA of software engineering (SE) studies. These papers, which broke new ground in SE, were followed by Dybå and Dingsøyr's proposal [3], which they applied in later research [4] and was then adopted by other reviewers [5]. Recent research papers, like [6], have drawn attention not only to the importance of building QA into SLR but also to the practical problems of QA and the need to have access to the right number of assessors. In this work authors constructed and validated a quality scale that was mainly based on [3].

It is generally accepted that a good quality experiment is bias free. Freedom from bias is the result of careful planning, operation and control, which maximizes the experiment's internal validity [7]. As bias cannot normally be calculated, standard QA instruments are generally

978-0-7695-4604-9/11 $26.00 © 2011 IEEE
DOI 10.1109/ESEM.2011.37

aimed at assessing the internal validity of experiments and inferring the quality of the experiment from this assessment [7].

Unfortunately, the relationship between internal validity and bias is far from clear. There are studies both for [3, 8, 9] and against [10-14] this relationship. On both purely theoretical grounds (the question is interesting in itself) and for practical reasons (determining which aspects of internal quality, if any, are related to bias can provide valuable information for running better quality experiments, as well as improving QA in SLR), this led us to look at exactly what the situation in SE was. To do this, we assessed the quality and bias of a set of 28 experiments taken from two different SLRs. Then we correlated quality and bias. From the resulting correlations, we can say that there are some aspects of internal validity (e.g., allocation method) that have a clear impact on experimental bias.

These results led us to defend the inclusion of some aspects that turned out to be quality predictors (like the above-mentioned allocation method, but also the disclosure of statistical significance, hypotheses and threats to validity) in the QA instruments that are used in SLRs. At the same time, we also recommend the use of simple instruments, as many aspects universally associated with internal validity have not been shown to have any relationship to bias (obviously within the limited scope of this study), and their use could be misleading.

The paper is divided into eight main sections. Section 2 describes the background; Section 3 describes the research goal and methodology; Sections 4 to 6 describe the research results; Section 7 presents the threats to validity of this study; and Section 8 states the conclusions.

II. BACKGROUND

A. *Quality concept*

The quality of an experiment is generally construed as the degree to which a study employs methods to minimize biases, that is, tries to maximize its internal validity [15]. Kitchenham [9] defines bias as *a tendency to produce results that depart systematically from the 'true' results*. For example, if we find that testing technique A is 50% more effective than testing technique B in a specific experiment, but we know that the real difference actually is 30%, the remaining 20% would be considered as bias. Accordingly, a good quality experiment would be one that uses randomization to create homogeneous experimental groups, employs concealing to allocate subjects and researchers, follows up all results, etc., all with the aim of minimizing bias. On the other hand, a poor quality experiment would be one that uses few or none of the above methods [15].

Of course, internal validity is only one of many facets that determine the goodness of an experiment. External validity also plays a key role. However, this role is subordinated to internal validity. Fig. 1 clearly explains why [15]: internal validity reflects the degree to which the experimental results are accurate, and accuracy is determined by the minimization of the risk of bias.

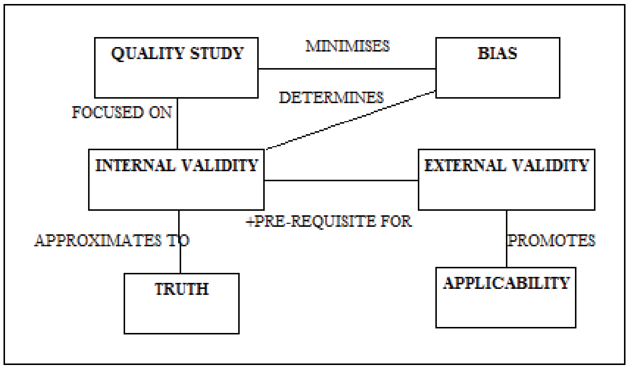

Figure 1. QA components

If we achieve accurate results, then we will be able to successfully extrapolate these results to other populations or settings. If the outcomes are not accurate (because of bias), extrapolation will be out of the question. In other words, quality is determined by a set of experimental design and execution parameters that ensures the validity of the outcome [1, 15-18]. Unbiased results are internally valid. Internal validity is a prerequisite for external validity [15].

The literature (essentially in the field of medicine [7, 18, 19]) identifies different types of bias, all of which apply in the field of SE:

- *Selection bias* (or *allocation bias*): Distortions in the treatment results due to how such treatments are allocated to experimental groups [7] (e.g., experienced subjects apply technique A, whereas novice subjects apply technique B). This can also occur because of how experimental subjects are recruited (e.g., the most brilliant students are allocated to one experimental group and the average students to the other). The mechanisms for preventing this bias are randomization and concealment.
- *Performance bias*: Variations in the results caused by inadvertently applying treatments other than those under evaluation to the subjects (e.g., experimental subjects learning structural testing techniques were taught statement coverage when they should have learned condition coverage). Another possibility is that the subjects do not apply the treatment properly (e.g., subjects do not follow the instructions on testing technique use) or receive external help (e.g., the researchers experimenting with a technique that they have invented inadvertently help the group to use that technique). The mechanisms for preventing this type of bias are protocol standardization and blinding.
- *Measurement bias* (or *detection bias*, or *ascertainment bias*): Variations caused by the subjective evaluation of the results of the experimental task (e.g., the rating of a design as *correct* or *incorrect*). The mechanism preventing

this kind of bias are blinding of outcome assessors and, especially, the definition of strict data collection and analysis protocols and goals.

- *Attrition bias* (or *exclusion bias*): Differences caused by the withdrawal or exclusion of experiment participants (e.g., some subjects do not submit the results of an experimental task because they think that they have done it *wrong)*. The mechanism for preventing this type of bias is to use *intention to treat analysis*.

Unlike random error, bias is not cancelled out among subjects, but has a tendency to grow. In the case of experimenter-induced performance bias, for example, all the subjects that received help (even if this assistance was provided inadvertently) will tend to perform better than the subjects that received no help. Bias, then, is a *systematic error* that is observed in the experimental results as a deviation of the experimental outcome from its true value. In principle, bias can be quantified if there are enough experiments, provided that such experiments are not affected by the same systematic error (which is a reasonable assumption for independent experiments).

B. Determining bias

The usual procedure for determining the bias of an experiment is to compare its outcome with the average of all the experiments in a set calculated by means of meta-analysis [10, 11, 20].

The bias, and therefore internal validity or quality, of a particular experiment would be straightforward to estimate if a sufficient number of experiments were always available: it would suffice to compare the results of the experiment against the respective meta-analysis. But this is not generally possible even in experimentally mature disciplines like medicine [15], and much less so in SE. On this ground, many researchers have tried to estimate bias (and therefore internal validity or quality) based on what is commonly accepted as the source of the bias: experimental design, operation and analysis weaknesses (i.e., the above-mentioned sources of bias). There are three approaches for estimating bias [7, 18]:

- First, we have the *simple approaches*, where a set of validity criteria are applied, which are answered qualitatively (e.g., *met/unmet*, *clear/unclear*), and a risk of bias is established (low, moderate, high) based on the number of satisfied criteria (e.g., *all criteria met*, *one or more criteria partly met*, *one or more criteria not met*).
- Second, *checklist*s are instruments based on quality items and are not scored numerically. This type of instruments are generally composed of a sizeable number of quality-related questions (items) with *yes*/*no* answers, e.g., Is there a well-defined question? Are the results generalizable to the setting of interest in the review?
- Third, *quality scales* are instruments based on a number of quality items, which are scored numerically to provide a quantitative estimate of overall study quality. All scoring systems tend to be subjective. Scores can be generated by weighting all items equally or by assigning them different weights in relation to their perceived importance. Table I shows an example of quality items in a quality scale [15].

Most checklists and quality scales were devised, logically, for the field of medicine [15, 16, 21-29], although there are proposals in other disciplines such as the social sciences [30], environment and public health and also SE [4, 9].

However, some objections have been raised concerning the use of checklists or quality scales for assessing the quality of experiments. According to Higgins and Green [18], the apparent simplicity of checklists and quality scales is not supported by empirical evidence, which has also been backed by some studies [11-14]. In fact, Higgins and Green [18] explicitly advise against the use of checklists or quality scales in Cochrane reviews and advocate the use of simple approaches for assessing validity, as they can report the values output by an experiment for each criterion.

C. Quality assessment in SE

In the particular setting of SE, SLR and QA are recent topics, and they started to receive attention after seminal work by Kitchenham in 2004 [9]. Since the publication of that report, multiple SLRs have been published in SE. Initially, not much importance was attached to QA. The field was not mature enough, and this subject was dealt with superficially [9], without proposing precisely how to run QA.

Things changed drastically as of 2007, when a new version of Kitchenham's report came out [1]. This new version contains a list of 50 questions aimed at assessing the quality of experiments.

TABLE I. EXAMPLES OF ITEMS IN A QUALITY SCALE

Quality item	Weight	Coding and explanation
Was the assignment to the treatment groups really random?	5	Adequate (random numbers table or computer and central office or coded packages) Partial (envelopes without further description or serially numbered opaque, sealed envelopes) Inadequate (alternation, case record number, birth date, or similar procedures) Unknown (just the term 'randomized' or 'randomly allocated' etc.)
Were outcome assessors blinded to the treatment allocation?	4	Adequate (independent person or panel or self-assessments in watertight double-blind conditions) Inadequate (clinician is assessor in trial on drugs with clear side effects or a different influence on lab results, ECGs etc.) Unknown (no statements on procedures and not deducible)

The questions were taken from different sources and disciplines and were organized according to the stage of the study that they assess. Kitchenham suggests that each researcher should review the list and select the questions best suited to the context of their own research questions. The questions summarized in [1] have to be assigned a measurement scale, as they do not all have a simple *yes/no* answer, that is, there is a mixture of both checklist-type and quality scale-type questions.

Dybå and Dingsøyr [3] also proposed a set of eleven criteria to assess the quality of studies and use them in their SLR on agile software development. These criteria were informed by the Critical Appraisal Skills Programme – CASP [31] and by principles of good practice for conducting empirical research in SE proposed by Kitchenham et al. [8].

III. Research goal and methodology

Our research goals are:

1. Determine whether there is really a relationship between internal quality and bias in SE experiments.
2. Find out what aspects of internal validity (or lack of internal validity) cause the greatest amount of bias.

Both objectives are important on three grounds:

- Identifying that there is such a relationship would signify that, for the first time in SE, the way experiments are designed, operated and analysed has been found to have an impact on their outcomes.
- Finding out which aspects of internal validity most influence bias would empower experimenters to take the necessary measures to minimize such bias.
- Finally, it would be possible to develop QA methods that precisely identify the quality of the experiments with a view to their aggregation in a SLR.

To achieve our goals, we used a similar methodology to the one used by other studies in the field of medicine [10, 11, 20]:

- First, we developed a QA instrument to determine the quality of experiments. Of the alternatives described in Section II (simple approaches, checklists or quality scales), we decided to use a quality scale because it outputs a quantitative result on quality that can be used to study the correlation between quality and bias. We applied this QA instrument to 28 experiments included in two SLRs [32, 33] and calculated quality scores for each one
- Second, we calculated the bias for each of the 28 experiments. As a measure of bias we used the absolute difference between the experimental effect size and effect size of the meta-analysis including all experiments. The number of SLRs used is low because meta-analysis is required to calculate bias, and statistical synthesis is still an open question in SE [34].
- The relationship between quality (understood as internal validity) and bias was established by means of an analysis of correlation of the score output by the QA instrument both globally (objective 1) and at the level of the individual items (objective 2).

Each of these steps is described in the following sections.

IV. Assessing quality in experiments

This activity aims to define a mechanism (or instrument) to estimate how good the quality of an experiment is with the purpose of later relating this value to bias and answering the research question that we set. The chosen mechanism was the quality scale, as it returns continuous values and can, therefore, be easily correlated to bias.

A critical aspect of this research is the quality scale to be used. The different quality scales available in the literature (see Section II) vary enormously in terms of the evaluated aspects. For example, [16] only contains 5 quality items, whereas [30] contains 33. This has a direct impact on the means of controlling the bias under consideration. For example, Dybå and Dingsøyr [3] do not account for aspects like randomization or blinding. Therefore, if we decided to use [3] as a quality scale, and randomization or blinding were related to bias, we would not detect this relationship. Note, however, that the checklist proposed by Dybå and Dingsøyr [3] was developed for empirical studies rather than specifically experimental studies, which explains the omission of the above criteria.

Alternatively, the quality scales containing many items (like the list of questions reported by Kitchenham [1] or [24]) increase the likelihood of finding relationships among experimental design, operation or analysis aspects and random bias (due to the accumulation of type I error). Kitchenham suggests analysing the list of questions in the light of each SLR to select the ones that best relate to the research question. However, this strategy is not easily applicable to this study, as we took experiments from two different SLRs to increase the sample size. This is an obstacle to selecting the best items for each SLR.

Finally, we opted for a pragmatic alternative. We used Kitchenham et al.'s seminal work [8] to identify the quality scale items for consideration. This way, the number of items was controllable (we only considered 10 different items), whereas the overlap was acceptable at [1-4]. In [8], the experimental quality assurance recommendations are organized according to several dimensions:

- *Experimental context*: for [8] the context of a study must include information about the industrial circumstances in which an empirical study takes place or in which a new software engineering technique is developed, a discussion of the research hypotheses and how they were derived, as well as information about related research.

- *Experimental design*: this refers to the products, resources and processes used in the experiment.
- *Analysis*: this evaluates the approach or approaches used to analyse the data output in the experiment.
- *Interpretation of results*: this assesses how the experiment presents its findings, and the relationship between the findings and the results.
- *Presentation of results*: this refers to how the experiment is reported so that the justification and results of the empirical study and their interpretation are clear to the reader.

We generated 10 questions based on these five dimensions, as shown in Appendix A. For example, a single question in our quality scale, *Does the introduction contain the industrial context (entities, attributes, and measures)?* ties in with recommendation C1 proposed by [8], '*be sure to specify as much of the industrial context as possible. In particular, clearly define the entities, attributes, and measures that are capturing the contextual information*'. To facilitate the use of the quality scale, the answers to each question are dichotomous, i.e., they have *yes* or *no* answers. The quality score is purely additive and is calculated as the percentage of *yes* answers obtained by an experiment after the application of the quality scale:

$$\text{Quality score} = \frac{\text{Number\_of\_yeses}}{\text{Number\_of\_yeses} + \text{Number\_of\_nos}} \quad (1)$$

The above quality scale has been applied to a set of 28 experiments that are part of two SLRs. Obviously, there is a much greater number of experiments [35] and SLRs [4] available in SE. Because we need to estimate bias, however, we can only use experiments that have been combined as part of a meta-analysis. To the best of our knowledge, this confines the set of possible SLRs to [4] and [36]. In fact, we cannot even use [36], as the between-study heterogeneity in that SLR could be confused with bias. Therefore we ran a specific meta-analysis [32] of the experiments on inspection considered in [36]. This way, we were able to get a reasonably reliable bias calculation. Table II summarizes the characteristics of the SLRs used in this research.

One of the authors (Saxena) applied the quality scale to 28 experiments. The results of the scale are shown in Table III. An experiment with a global score close to 1 (that is, 100%) would be acknowledged as a good quality experiment. An experiment with a global score close to 0 (0%) would we acknowledged as a poor quality experiment.

The mean global scores for both set of experiments were similar (0.78 for inspection and 0.73 for pair programming). The sub-scores of all papers are available at the following URL: www.grise.upm.es/sites/extras/6/.

TABLE II. THE TWO SLRS USED IN OUR RESEARCH

Domain	No of Exps.	Aggregation mechanism	Heterogeneity presence
SLR on inspection techniques [32]	13	Meta Analysis	No
SLR on pair programming [33]	15	Meta Analysis	??

TABLE III. QUALITY SCORES

Set	Study ID	Scale Scores
Inspection SLR	E05	0.8
	E01	0.8
	E02	0.7
	E06	0.8
	E03	0.7
	E04 - R1	0.8
	E04 - R2	0.8
	E13	0.8
	E14	0.8
	E15	0.8
	E19	0.8
	E20	0.8
Pair – Programming SLR	P98	0.7
	So0	0.6
	So1	0.5
	So2	0.6
	Po2	0.6
	So3	0.9
	So5a	0.7
	So5b	0.8
	So5c	0.8
	So6a	0.8
	So6b	0.8
	So6c	0.8
	So6d	0.6
	So7b	0.9
	So7a	0.8

The quality of the experiments included in the SLR on inspection is very uniform, probably because these experiments were developed in parallel with the definition of empirical SE as a discipline.

V. DETERMINING BIAS IN EXPERIMENTS

In the two SLRs used [32, 33], the synthesis procedure used was meta-analysis through weighted mean difference. Following Kitchenham [9] and similar studies in medicine, we used the deviation of the result of each experiment from the *mean* calculated by means of meta-analysis as a measure of bias. As the SLRs reported heterogeneous experiments, we considered the result of the fixed effects model as average.

For simplicity's sake, we will refer to this measure of bias as **proximity to mean value (PTMV).** To clarify the meaning of this measure, let us look at an example. Fig. 2 is a forest plot describing the aggregation of five experiments. The meta-analysis result (mean) is represented by means of a vertical line with a diamond in its tail. PTMV represents the distance d1…d5 between the results of experiments E01a, E01b, E06a, E06b, E08a respectively.

PTMV is a good measure of bias, as the farther an experiment is from the average value, the more biased the results of the experiment are. In Fig. 2, PTMV tells us that E06b is the most biased as it is farthest from the average value.

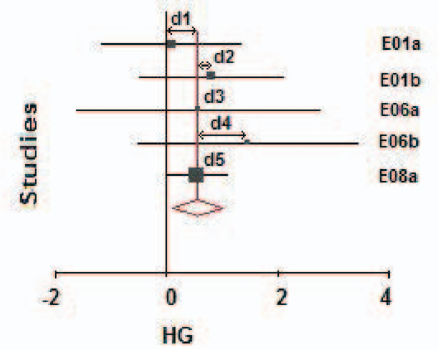

Figure 2. Aggregation forest plot of five experiments

E01b and E08a have the least bias. Therefore, according to the definition of quality used so far (lack of bias), E08a and E01b, being more precise, should be of better quality than E06b.

Two of the authors (Grimán and Saxena) calculated the bias indicator proposed in this study (see Table IV), although its objectivity means that the influence of the person who calculated it is negligible.

VI. Relating Quality and Bias

We identified the relationship between the quality scale (at the level of both the global score and the individual item) and bias. As this is a preliminary study, we thought it was best to use linear models rather than more complex approaches (like polynomial models, for example).

As the variables are continuous, the relationship between the quality score and bias was calculated by means of the Pearson correlation coefficient. The relationship between the items of the quality scale and bias was calculated by means of point biserial correlation [37] (as the items of the questionnaire are binary variables). All the calculations were made using SPSS® V.18 for Windows®.

A. *Correlation between quality score and bias*

The coefficients of correlation and statistical significance between the quality scale items and bias are shown in Table V.

TABLE IV. PTMV Scores

Set	Study ID	PTMV Scores
Inspection SLR	E05	0.34
	E01	0.13
	E02	0.36
	E06	0.23
	E03	0.21
	E04 - R1	0.09
	E04 - R2	0.63
	E13	0.32
	E14	0.26
	E15	0.37
	E19	0.33
	E20	0.14
Pair–Programming SLR	P98	2.11
	So0	0.66
	So1	0.24
	So2	0.265
	Po2	0.34
	So3	0.215
	So5a	0.335
	So5b	0.44
	So5c	0.08
	So6a	0.3
	So6b	0.26
	So6c	1.575
	So6d	1.2433
	So7b	0.65
	So7a	0.19

TABLE V. Correlation between global score and bias

		Bias
Score	***Pearson correlation***	-.140
	Sig. (bilateral)	.486
	N	27

Note that high bias values for an experiment indicate that it is of poor quality, which explains that the coefficient of correlation is negative ($r = -0.14$). However, the statistical significance indicates that it is clearly not significant (*p-value* = *0.486*). From this, linked to the low value of r, we can say that the quality score is a very poor predictor of the quality of a study, that is, the quality score and experimental bias are not related.

B. *Correlation between quality scale items and bias*

The coefficients of correlation between the quality scale items and bias are shown in Table VI. Items Q04 and Q10 evidently reveal no information whatsoever about the set of experiments under consideration, as their value is constant throughout. Of the other items, Q03, Q06 and Q09 show significant correlations with bias (in the cases of Q03 and Q06, the correlations are significant at the level of $\alpha = 0.01$). The values of the correlations are fairly high ($r < -0.4$ in all cases), indicating that the relationship between Q03, Q06 and Q09 and bias is quite strong, that is, Q03, Q06 and Q09 are fairly good predictors (each one is independent of the others) of the quality of an experiment.

TABLE VI. CORRELATION BETWEEN QUESTIONNAIRE ITEMS AND BIAS

		Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10
Bias	***Pearson correlation***	.144	.112	-.744**	.[a]	.129	-.694**	-.135	.250	-.406*	.[a]
	Sig. (bilateral)	.474	.578	.000	.	.520	.000	.501	.209	.035	.
	N	27	27	27	27	27	27	27	27	27	27

a. Correlation cannot be calculated as at least one variable is constant. *. The correlation is significant at the level 0.05 (bilateral). **. The correlation is significant at the level 0.01 (bilateral).

Items Q01, Q02, Q05 and Q07 do not appear to have any relationship whatsoever with bias; the p-values are very high, and the coefficients of correlation are positive (except for Q07), contrary to what we would have expected. Q08 is an item whose relationship to bias is hard to rate. On the one hand, the p-value associated with the correlation between Q08 and bias is clearly not significant (*p-value = 0.209*), but, on the other, its magnitude is much greater than the uncorrelated items (e.g., Q01). The high p-value could quite possibly be due to the small sample sizes used in this study.

Finally, note that the correlation between Q08 and bias is positive (*r = 0.250*), which is seemingly contrary to what was expected. We can decipher this apparent contradiction if we consider that Q08 refers to validity threats.

When we developed the questionnaire, we supposed that the disclosure of the validity threats was tantamount to a better control of experimental design and operation, which would lead to a better quality experiment. However, the positive correlation indicates just the opposite; that is, the existence of a high number of validity threats actually points to there being weaknesses in the experimental design and operation and, therefore, indicates a risk of bias.

C. *Multiple correlation between quality scale items and bias*

Table VII shows that, in some cases, the correlations are positive (specifically, Q01, Q02, Q05 and Q08), whereas, in others (Q03, Q06, Q07 and Q09), they are negative. This mixture of signs means that the decision to use the percentage sum of *yes* answers as the quality score was completely wrong. The items that correlate positively partially cancel out others that correlate negatively, leading to a poor correlation of the quality score with bias.

TABLE VII. MODEL COEFFICIENTS

Model		Typified coefficients		
		β_i	t	Sig.
	Q01	.035	.258	.799
	Q02	-.021	-.166	.870
	Q03	-.254	-.981	.340
	Q05	.071	.537	.598
	Q06	-.543	-2.830	.011
	Q07	.049	.354	.727
	Q08	.235	1.699	.107
	Q09	-.312	-1.495	.152

While it is true that, despite the problems of calculating the quality score, we were able to determine a clear relationship between some questionnaire items (Q03, Q06, Q09 and perhaps Q08) and bias, these are bivariate relationships, which can be misleading. For example, suppose that an experiment was evaluated *yes* for Q03 and *no* for Q06, would this be a good or poor quality experiment?

Obviously, quality construed as bias is a continuous, not a dichotomous, variable. In other words, the experiment in the above example would have some quality *level* (alternatively, a bias of some magnitude) induced by Q03 and Q06. Therefore, we opted to run an analysis of the data using multiple linear correlation to be able to analyse the joint relationship between *all* the quality scale items and bias simultaneously.

Essentially, multiple correlation is similar to bivariate correlation, save that the model to be fitted takes the form:

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon_i, \tag{2}$$

where y is the quality of the experiments, x_i is the different questionnaire items (Q01-Q10) and β_i is the relative weight of each quality scale item in the overall determination of quality. ε_i is the model error.

The β_i coefficients are shown in Table VII. Before going on to study these coefficients, note that the linear model (1) fits the data rather well. The corrected coefficient R^2 (percentage of explained variance) is *0.649*, resulting in a significant model ($F_{8,18}$ = *6.998*, *p-value < 0.001*). Fig. 3 shows that the residuals ε_i of the model are normally distributed (as expected).

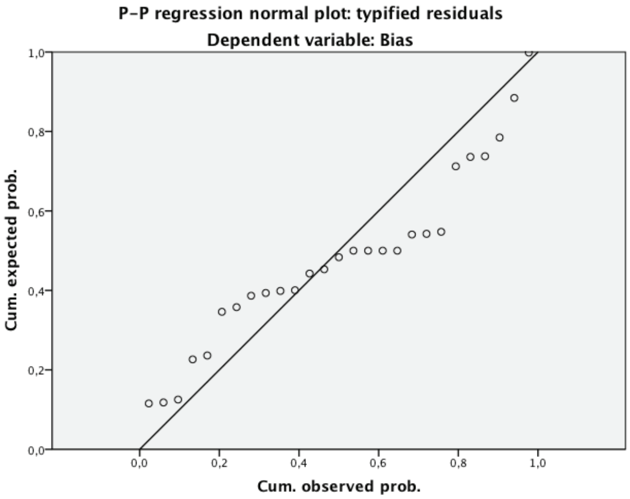

Figure 3. Normality of residuals

Now, with respect to items Q01-Q10, things have changed slightly compared to the above observations, although the overall trend is similar. We find that the p-values associated with the β coefficients are smaller than in the bivariate case. This is not surprising, as model (2) is more demanding in terms of sample size. However, the questionnaire items related to bias are unchanged:

- The relationship between Q06 and bias is well determined (*β = -0.543*, *p-value = 0.011*).
- Q08 and Q09 come close to statistical significance (*p-values* of *0.107* and *0.152*, respectively) with very similar coefficients (βs of *0.235* and *-0.312* respectively).
- Q03 is far from being significant (*p-value = 0.340)*, but its weight is greater than that of any of the other items (*β = -0.254*). It does, at any rate, show some tendency toward significance.
- Q01, Q02, Q05 and Q07 are far from achieving statistical significance.
- Q04 and Q10 do not participate in the model as they are variables whose value is constant in all instances, as they were in the case of the bivariate correlations.

In summary, we can conclude that, of all the studied aspects of internal validity, only Q03, Q06, Q08 and Q09 are related to bias and, therefore, are predictors of the quality of an experiment.

The item that best predicts quality (has the greatest β value) is Q06, followed by Q09, Q03 and Q08 (the latter with a negative β, that is, it should not be present in good quality experiments). Note, however, that the weight of Q06 is almost the same as Q03, Q08 and Q09 together (exactly 68%), meaning that it is the key factor for determining whether experiment quality is good or bad.

VII. Threats to validity

During the development of this research we identified several threats that could compromise the reliability of the outcomes. These threats were analysed and processed with the aim of promoting the applicability and generalizability of our findings.

- The indicator used to measure bias can return imprecise values because not many studies are used for its calculation. Typically we have a small number of experiments, and, at the same time, the sample size of most experiments is small. On the other hand, the indicator PTMV can be affected for the heteroginity of the set of studies. Heteroginity is low in the studies at [32] but it is higher for the set of studies at [33].
- The quality assessment was conducted by only one person - a master student with solid knowledge on quality of experiments. However, we realize the need of incorporating additional evaluators for the next iteration of this study in order to guarantee the reliabity of the quality scores.
- We have not considered all the elements of internal validity that could affect quality. While it is true that other disciplines (e.g., medicine) account for multiple characteristics of the experiments related to internal validity (e.g., drop-outs, concealment, secular changes, etc.), we have used only aspects that recur in the literature and are applicable to SE in this research.
- One of the SLRs included in this study was run by researchers of the Universidad Politécnica de Madrid's (UPM) Experimental Software Engineering Research Group. However, this association should not bias this study, as, on the one hand, the SLR was run for a different purpose than this study and, on the other, we do not have any hidden agenda whatsoever related to the QA of experiments in SLR.
- The researchers that calculated both the quality score and the bias of the experiments used in this study are members of the UPM's Experimental Software Engineering Research Group. We already mentioned that we have no agenda whatsoever regarding the QA of experiments in SLRs, but we do acknowledge that there is a need for this study to be replicated and extended by independent researchers.

VIII. Conclusions

We determined the quality of 28 experiments belonging to two SLRs on software inspection [32] and pair programming [33], using a quality scale based on the methodological recommendations reported in [8]. The quality of the experiments was correlated to an objective measure of experimental bias at the level of both the global quality score and each of the quality scale items (questions).

Of the analysed aspects of internal quality, the only strong relationship we found was between the allocation method and bias. Also we identified some relationship between the disclosure of the statistical significance, hypothesis, and threats to validity and bias.

Even though these are preliminary results and require an analysis based on a greater number of studies, we can list some findings based on the trends that we observed:

- Based on the outcomes we can state that there is a relationship between aspects of study quality, defined in terms of internal validity, and experimental bias.
- As it was only possible to identify a relationship between some aspects of internal quality (e.g., allocation) and bias, the QA instruments used in SLRs should not contain a multitude of items. Preferably, they should use just a few whose relationship with bias is reasonably well defined.
- As a corollary of the above, the output quality scores can be misleading especially in the case of quality scales with many items. Until we get a better understanding of the relationships between internal quality and bias, we advise against the use of quality scores and believe that it is better to use simple approaches as suggested by Khan et al. [15] (see Section II).

Finally, it is worth mentioning that the findings of our research are relevant not only to meta-analysers. Experimenters should also pay attention to the aspects of internal validity that are related to bias with the aim of controlling these aspects and outputting better quality experimental results.

APPENDIX A – DIMENSIONS, QUESTIONS AND FOUNDATIONS

Dimension	Question	Recommendation in [8]
Experimental Context	Does the introduction contain the industrial context (entities, attributes, and measures) and description of the techniques to be reviewed? For experiments that evaluate techniques developed in industry. (**Q1**)	In experiments evaluating techniques developed in industry, experimenters should understand how the technique works in the industrial setting before developing a version of the technique for experimental purposes. This is due to the fact that techniques developed in industrial settings are highly complex, and such complexity is difficult to reproduce in academia. The treatments that are tested in an experiment must be well defined in the report for the experiment to be able to be replicated or simply for the results to be able to be transferred to industry.
	Does the report summarize and discuss earlier similar experiments that have been conducted? (**Q2**)	Describing earlier research that is similar to this study and how they are related can help to build an integrated body of knowledge about a phenomenon in SE.
	Are the hypotheses being laid and are they synonymous with the goal discussed before in introduction? (**Q3**)	Specific hypotheses that are being tested in the study should be clearly established beforehand based on a theory.
Experimental Design	Does the researcher define the population from which objects and subjects are drawn? (**Q4**)	It is necessary to define the population from which the subjects and objects have been extracted to be able to extract inferences from the experimental results.
	Does the researcher define the process by which he applies the treatment to objects and subjects (e.g. randomization)? (**Q6**)	The subjects and objects should be allocated to the treatments in an unbiased manner so as not to compromise the experiment.
	Was randomization used for selecting the population and applying the treatment? (**Q7**)	The subjects and objects should be representative of the population to be able to extract conclusions from the experimental results.
	Does the researcher define the process from which the objects and subjects are selected (e.g. random sampling)? (**Q5**)	
	Is an appropriate blinding procedure used (e.g. blind allocation of materials, blind marking)? (**Q10**)	A double-blinding procedure, as run in medicine, is not possible in SE experiments, but other types of blinding are; these types of blinding can be applied to the allocation of materials, marking and analysis.
Analysis	Is an appropriate blinding procedure used (e.g. blind allocation of materials, blind marking)? (**Q10**)	The information on treatments should be somehow encoded to prevent analysts from knowing the treatment to which it corresponds and being able to introduce bias into the results of the analysis.
Presentation of results	Are the statistical significances mentioned with the results. (**Q9**)	The experiment should report the quantitative data including the effect size and the confidence limits.
Interpretation of results	Is mention made of the threats to validity and also how these threats affect the results and findings? (**Q8**)	Experimenters should discuss the limits of the study, at least threats related to internal and external validity.

ACKNOWLEDGEMENTS

This work was a part of a master thesis completed at the UPM Madrid, Spain, and TU Kaiserslautern, Germany, under Dieter Rombach, Natalia Juristo and Oscar Dieste as part of the European Masters Course in Software Engineering. This work was partially supported by Spanish Ministry of Science grants TIN 2008 – 00555 and HD 2008-0048.

REFERENCES

[1] B.A. Kitchenham. and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Tech. Rep. EBSE Technical Report EBSE-2007-01, 2007.

[2] J. Biolchini, P. Mian, A. Natali and G. H. Travassos, "Systematic review in software engineering," Rio de Janeiro, Tech. Rep. ES 679/05, 2005.

[3] T. Dybå and T. Dingsøyr, "Strength of evidence in systematic reviews in software engineering," Proc. 2nd ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM 2008), October 2008, pp. 178–187.

[4] T. Dybå and T. Dingsøyr, "Empirical studies of agile software development: A systematic review," Information and Software Technology*,* vol. 50, pp. 833–859, 2008.

[5] W. Afzal, R. Torkar and R. Feldt, " A systematic review of search-based testing for non-functional system properties. ," Information and Software Technology*,* vol. 51, pp. 959–976, 2009.

[6] B. Kitchenham, et al., "Can we evaluate the quality of software engineering experiments?," Proc. 4th ACM/IEEE International Symposium on Empirical Software Engineering and Measurements (ESEM 2010), September 2010.

[7] CDR, "Systematic reviews: Crd's guidance for undertaking reviews in health care," C. f. R. a. Dissemination, York, Tech., 2009.

[8] B. Kitchenham, et al., "Preliminary guidelines for empirical research in software engineering," IEEE Transactions on Software Engineering*,* vol. 28, pp. 721-734, 2002.

[9] B. Kitchenham, "Procedures for performing systematic reviews," Keele University, Tech. Rep. TR/SE-0401, NICTA 0400011T.1, 2004.

[10] P. Jüni, A. Witschi, R. Bloch and M. Egger, "The hazards of scoring the quality of clinical trials for meta-analysis," The Journal of the American Medical Association, vol. 282, pp. 1054-1060, 1999.

[11] E. M. Balk, et al., "Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials," JAMA, vol. 287, pp. 2973-2982, 2002.

[12] A. Britton, et al., "Choosing between randomised and non-randomised studies: A systematic review," Health Technology Assessment, vol. 2, pp. 3-4, 1998.

[13] J. J. Deeks, et al., "Evaluating non-randomised intervention studies," Health Technology Assessment, vol. 7, pp. 3-4, 2003.

[14] J. D. Emerson, E. Burdick, D. C. Hoaglin, F. Mosteller and T. C. Chalmers, "An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials," Controlled clinical trials vol. 11, pp. 339-352, 1990.

[15] K. Khan, et al., "Undertaking systematic review of research on effectiveness. Crd's guidance for those carrying out or commissioning reviews," NHS Centre for Reviews and Dissemination - University of York, Tech. Rep. CRD-4 (2nd edition), 2001.

[16] A. R. Jadad, et al., "Assessing the quality of reports of randomized clinical trials: Is blinding necessary? ," Controlled Clinical Trials, vol. 17, pp. 1-12, 1996.

[17] M. Petticrew and H. Roberts, Systematic reviews in the social sciences. A practical guide. Oxford: Blackwell Publishing, 2007.

[18] J. P. Higgins and S. Green, Cochrane handbook for systematic reviews of interventions 4.2.6 [updated september 2006] vol. 4. Chichester, UK: John Wiley & Sons, Ltd., 2006.

[19] National Health and Medical Research Council, How to review the evidence: Systematic identification and review of the scientific literature Canberra: Biotext, 2000.

[20] K. F. Schulz, I. Chalmers, R. J. Hayes and D. G. Altman, "Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials," JAMA, vol. 273, pp. 408-412, 1995.

[21] Alberta Research Centre for Health Evidence, "Conducting a systematic review." [online]. Available: http://www.ualberta.ca/ARCHE/sysreviewsproc.html#question

[22] National Institute for Health and Clinical Excellence, The guidelines manual. London: National Institute for Health and Clinical Excellence, 2009.

[23] D. Moher, K. F. Schulz and D. G. Altman, "The consort statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials," Lancet, vol. 357, pp. 1191–1194, 2001.

[24] S. H. Down and N. Black, "The feasibility of creating a checklist for assessment of the methodological quality both of the randomised and non-randomised studies of health care interventions," J Epidemiol Community Health, vol. 52, pp. 377-384, 1998.

[25] S. Zaza, et al., "Data collection instrument and procedure for systematic reviews in the guide to community preventive services," Am J Prev Med, vol. 18, pp. 44-74, 2000.

[26] Cochrane Effective Practice and Organisation of Care Review Group, "Data collection checklist," Institute of Population Health - University of Ottawa, Ottawa, Tech. Rep. (number not provided), June 2002.

[27] Effective Public Health Practice Project, "Quality assessment tool for quantitative studies." [online]. Available: http://www.ephpp.ca/PDF/Quality%20Assessment%20Tool_2010_2.pdf

[28] J. Reisch, J. Tyson and S. Mize, "Aid to evaluation of therapeutic studies," Pediatrics, vol. 84, pp. 815-824, November 1989.

[29] D. K. Owens, et al., "Grading the strength of a body of evidence when comparing medical interventions," Journal of Clinical Epidemiology, vol. 63, pp. 513-523, 2010.

[30] T. Cook and D. Campbell, Quasi-experimentation: Design and analys issues for field settings. Boston: Houghton Mifflin Company, 1979.

[31] PHRU, "Critical appraisal skills programme." [online]. Available: http://www.phru.nhs.uk/casp/casp.htm

[32] A. Grimán, O. Dieste and N. Juristo, "Systematic review on inspection techniques", Unpublished

[33] J. E. Hannay, T. Dybå, E. Arisholm and D. I. K. Sjøberg, "The effectiveness of pair programming: A meta-analysis," Information and Software Technology, vol. 51, pp. 1110-1122, 2009.

[34] D. Cruzes and T. Dyba, "Synthesizing evidence in software engineering research," Proc. 4th ACM/IEEE International Symposium on Empirical Software Engineering and Measurements (ESEM 2010), September 2010, pp. 1-10.

[35] D. I. Sjøberg, et al., "A survey of controlled experiments in software engineering," IEEE Transactions on Software Engineering, vol. 31, pp. 733-753, 2005.

[36] M. Ciolkowski, "What do we know about perspective-based reading? An approach for quantitative aggregation in software engineering," Proc. 3rd ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'09), October 2009, pp. 133-144.

Other references:

A list of experiments used in our study is available at the following URL: www.grise.upm.es/sites/extras/6/