Systematic Review of Effect Size in Software Engineering Experiments

Vigdis By Kampenes ^{a,b}, Tore Dybå ^{a,c}, Jo E. Hannay ^{a,b},
Dag I. K. Sjøberg ^{a,b}

[a] Department of Software Engineering, Simula Research Laboratory,P.O. Box 134, NO-1325 Lysaker, Norway
[b] Department of Informatics, University of Oslo, P.O. Box 1080 Blindern, NO-0316 Oslo, Norway
[c] SINTEF ICT, NO-7465 Trondheim, Norway
vigdis@simula.no, tore.dyba@sintef.no, {johannay, dagsj}@simula.no

Abstract

An effect size quantifies the effects of an experimental treatment. Conclusions drawn from hypothesis testing results might be erroneous if effect sizes are not judged in addition to statistical significance. This paper reports a systematic review of 92 controlled experiments published in twelve major software engineering journals and conference proceedings in the decade 1993-2002. The review investigates the practice of effect size reporting, summarizes standardized effect sizes detected in the experiments, discusses the results and gives advice for improvements. Standardized and/or unstandardized effect sizes were reported in 29% of the experiments. Interpretations of the effect sizes in terms of practical importance were not discussed beyond references to standard conventions. The standardized effect sizes computed from the reviewed experiments were equal to observations in psychology studies and slightly larger than standard conventions in behavioural science.

Keywords: Empirical software engineering; Controlled experiments; Effect size; Statistical significance; Practical importance.

1. Introduction

Software engineering experiments investigate the cause-effect relationships between treatments applied (process, method, technique, language, tool, etc.) and outcome variables measured (time, effectiveness, quality, efficiency, etc). An *effect size* is the magnitude of the relationship between treatment variables and outcome variables, and is computed on the basis of the sample data to make inferences about a population (analogously to the concept of hypothesis testing). An effect size tells us the degree to which the phenomenon under investigation is present in the population. There are several types of effect size measures[1], for example, correlations, odds ratios and differences between means.

Wrong or imprecise conclusions might be drawn from hypothesis testing results if effect sizes are not judged in addition to statistical significance. In particular, *p*-values are insufficient for decision-making; if an experiment includes a sufficient number of subjects, it is always possible to identify statistically significant differences, or if the

[1] We will refer to specific values as *effect sizes*, and ways (formulae) to compute effect sizes as *effect size measures*.

experiment includes too few subjects (insufficient power), *p*-values may also be misleading. So, whereas *p*-values reveal whether a finding is *statistically* significant, effect size indicates *practical* significance, importance or meaningfulness. Interpreting effect sizes is thus critical, because it is possible for a finding to be statistically significant but not meaningful, and *vice versa* [7, 27]. Hence, as also recommended by others [12, 23, 29], effect sizes should be part of experimental results in software engineering.

There is no unambiguous mapping from an effect size to a value of practical importance. Hence, observed effect sizes must be judged in context [2, 9, 18, 21, 35, 36, 41, 42, 45]. Even small effects might be of practical importance. For example, the optimization of a defect-detection method that yields only a one percent increase in error detection would be of little practical importance for most types of software, but might be of high practical importance for safety-critical software, particularly if the added one percent belongs to the most critical type of errors. This means that a contextual, subjective judgment of observed effect sizes must be made and a ritualized interpretation avoided. Hence, not only is the reporting of effect sizes important, but also a nuanced interpretation and discussion of those values.

Effect size estimation is not a new method. An approach to determining the magnitude of the effect of agricultural treatments was published seven decades ago [3], and reporting effect sizes in addition to statistical significance has been recommended for a long time in behavioural science [4, 45]. Reporting effect sizes is also urged in medical science. A group of scientists and editors have developed the CONSORT statement to improve the quality of reporting of randomized clinical trials. One recommendation is that one should report "for each primary and secondary outcome, a summary of results for each group and the estimated effect size and its precision (e.g., 95% confidence interval)" [1: p.682].

In addition to being meaningful in the analysis and reporting of experimental results, previously published effect sizes can be used in meta-analyses [17] and in statistical power analyses [5, 27], and for comparison purpose. Such use requires the reporting of either effect sizes, or sufficient data for effect size estimation.

This article reports on a systematic review of the literature on effect size issues in controlled experiments published in empirical software engineering. A total of 113 controlled experiments were reported in the decade from 1993-2002 in 12 leading journals and conference proceedings in software engineering [39]. Of these 113 experiments, this review investigates the 92 for which statistical hypothesis testing was performed and primary tests were identifiable. The aim of this review is to investigate the following:

- *The extent of effect size reporting and the interpretation of the effect sizes given by the authors of the reviewed experiments*, i.e., the extent to which effect sizes are used to describe the experimental result as a supplement to statistical significance, and when effect sizes are reported, how they are described and interpreted. This investigation is motivated by the belief that the use of effect sizes affects conclusions made from experiments.
- *The extent to which experimental results are reported in such a way that standardized effect sizes can be estimated.* This is an assessment of the completeness of the reporting of descriptive statistics. A complete reporting of descriptive statistics will

allow the reader to verify the reporting of test results and effect size estimates, and to estimate effect sizes other than those reported.

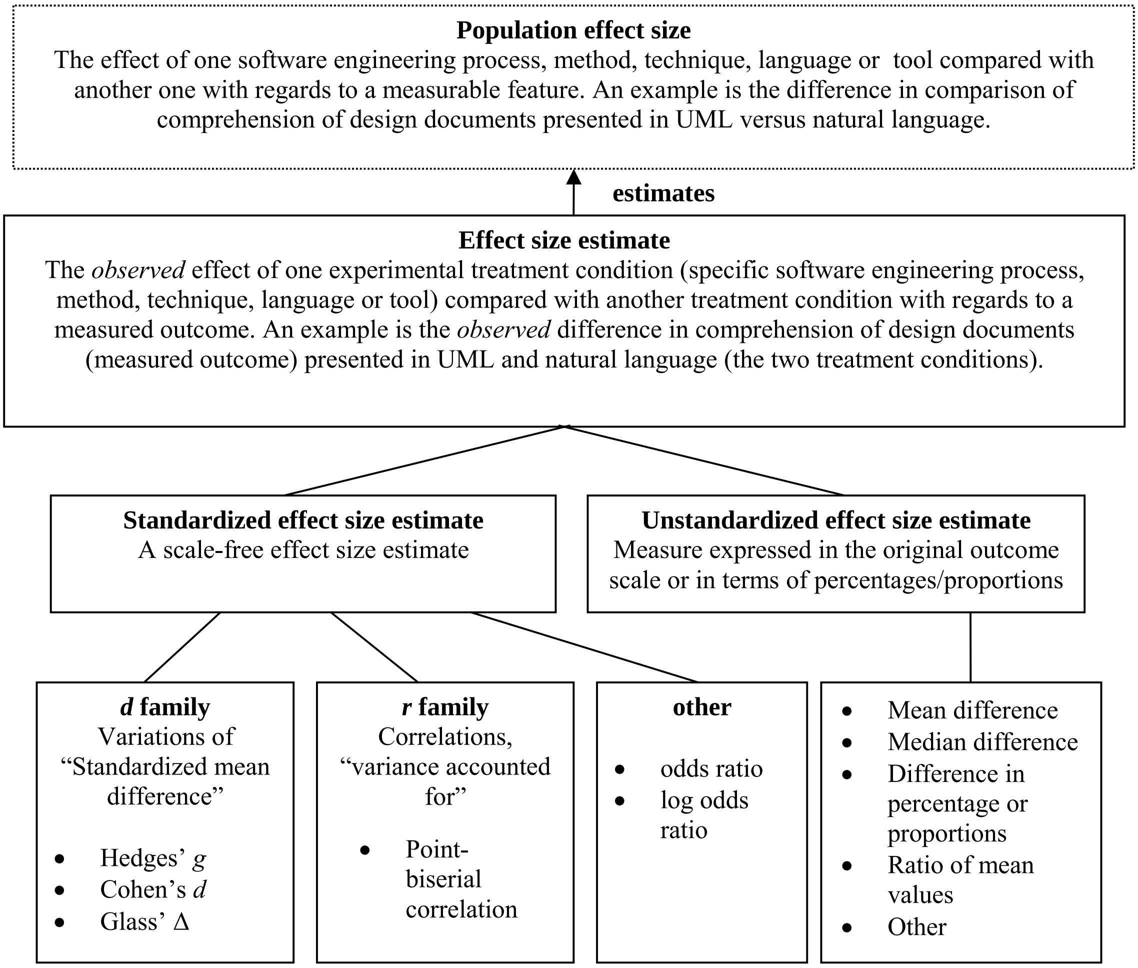
- *The standardized effect sizes detected in software engineering experiments.* The rationale for this investigation is to provide an overview of effect sizes detected in software engineering experiments so that researchers can make relative comparisons of observed effect size estimates.

The remainder of this paper is organized as follows. Section 2 summarizes relevant concepts and measures of effect size. Section 3 describes the research method applied in this review. Section 4 reports the results. Section 5 discusses the findings, the implications for power analysis, the limitations of the study, and presents guidelines for reporting effect sizes. Section 6 concludes.

2. Background: Effect size

The effect that one inspection method has on the number of defects detected compared with another inspection method is an example of an effect in software engineering that we wish to investigate by conducting experiments. This unknown effect is referred to as the *population effect size.* It cannot be computed directly as long as we do not have access to the total population of subjects that falls within the scope of the research questions of our investigation. However, the population effect size may be estimated from sample data from a single experiment. *Estimated effect sizes* from several experiments can further be aggregated and analyzed to provide even stronger foundations for inferences about the population effect size (meta-analysis).

Figure 1 gives an overview of the effect size concepts described in the next sections. Measures of effect size can be classified as *standardized* or *unstandardized*. Standardized measures are scale-free because they are defined in terms of the variability in the data. Types of standardized measures of effect size are presented in Section 2.1. Unstandardized measures encompass all other types of effect size measures and will be described in Section 2.2.



Population effect size
The effect of one software engineering process, method, technique, language or tool compared with another one with regards to a measurable feature. An example is the difference in comparison of comprehension of design documents presented in UML versus natural language.

Effect size estimate
The *observed* effect of one experimental treatment condition (specific software engineering process, method, technique, language or tool) compared with another treatment condition with regards to a measured outcome. An example is the *observed* difference in comprehension of design documents (measured outcome) presented in UML and natural language (the two treatment conditions).

Standardized effect size estimate
A scale-free effect size estimate

Unstandardized effect size estimate
Measure expressed in the original outcome scale or in terms of percentages/proportions

***d* family**
Variations of "Standardized mean difference"

- Hedges' *g*
- Cohen's *d*
- Glass' Δ

***r* family**
Correlations, "variance accounted for"

- Point-biserial correlation

other

- odds ratio
- log odds ratio

- Mean difference
- Median difference
- Difference in percentage or proportions
- Ratio of mean values
- Other

Fig. 1. Population and estimated effect size as defined for software engineering and examples of types of effect size measures for the comparison of two treatment conditions.

2.1. Standardized effect size

Two families of standardized effect size measures are often referred to in the literature: the *d family* and the *r family*. Below, we will emphasize Hedges' *g* in the *d family* and the point-biserial correlation in the *r family*, because these are the two types applied in this review.

2.1.1. The d family

The *d family* consists of variations over standardized mean difference. Assume that we have two groups, Group 1 and Group 2. Moreover, assume that the experimental observations in Group 1, $y_{11,}\ldots, y_{1n}$, are normally distributed with mean μ_1 and variance σ^2, and the observations in Group 2, $y_{21},\ldots,y_{2m}$, are normally distributed with mean μ_2 and variance σ^2. More specifically:

$$Y_1 \sim N(\mu_1, \sigma^2)$$

and

$$Y_2 \sim N(\mu_2, \sigma^2)$$

The population standardized mean difference effect size measure, which we will call d_{pop}, is defined as

$$\textit{Population standardized mean difference,} \quad d_{pop} = \frac{\mu_1 - \mu_2}{\sigma} \tag{1}$$

The population standardized mean difference takes positive or negative values, depending on the choice of μ_1 and μ_2. It is estimated by the difference between sample means ($\overline{X}_1, \overline{X}_2$) divided by an estimate of population standard deviation. Different estimators of the population standard deviation give different effect size estimators. The three estimators most often referred to in the literature are Hedges' *g*, Cohen's *d* and Glass' Δ [24, 34]. Hedges' *g* has the pooled standard deviation, S_p, as the standardizer:

$$\textit{Hedge's } g = \frac{\overline{X}_1 - \overline{X}_2}{s_p} \tag{2}$$

The pooled standard deviation is based on the standard deviations in both groups, s_1, s_2:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}, \tag{3}$$

Cohen's *d* also has the pooled standard deviation as its standardizer, but with n_i replacing $(n_i - 1)$ in Formula (3) and in the estimators of the single s_i. Glass' Δ applies the standard deviation in one group only; the one considered to be the control. According to [17], these three estimators have the same properties in large samples (i.e., they are equivalent in the limit $(n_1+n_2) \rightarrow \infty$), but Hedges' *g* has the best properties for small samples when multiplied by a correction factor that adjusts for small sample bias (Formula 4 below). Hence, we applied Hedges' *g* as the estimator for d_{pop} in our investigation and will not consider Cohen's *d* and Glass' Δ further.

$$\textit{correction factor for Hedge's } g = 1 - \frac{3}{4(N-2)-1}, \tag{4}$$

where *N* is the total sample size.

Hedges' *g* assumes homogeneity of variance in the two experimental groups. Kline [24] suggests that if the ratio of the largest standard deviation over the smallest standard deviation is larger than four, the effect sizes should be calculated twice using each standard deviation and the diverging results discussed. Other solutions are to replace s_p with an estimate of the standard deviation of whichever sample is the reasonable baseline comparison group [14], or to use the square root of the mean of s_1, s_2 [5].

Formulas (2) above are applicable for outcomes measured on the continuous scale. When aggregating study results from several studies and the standardized mean difference is to be estimated, there is a need for estimators that approximate a standardized mean difference effect size for variables that are measured on scales other than the continuous. When the outcome is dichotomous (binary), approximations to the standardized mean difference can be expressed in terms of an arcsine transformation [15] or an odds ratio [24, 37, 38]. When the outcome is ordinal (e.g., small, medium, large) a continuous scale might be assumed and formulas (2) applied, but note that when the number of categories is less then five, this approach will underestimate the population effect size [38]. When nominal outcomes are used, the standardized mean difference must be computed for pairs of categories applying the methods for dichotomous outcomes.

When raw data is unavailable, or means and standard deviations are not reported, effect size estimation can be based on various kinds of statistics. This is relevant for meta-analyses or statistical power analyses, or if a reader wants to judge published results in terms of effect sizes when these are not reported. Table 7 shows the set of formulas for computing Hedges' *g* that we applied in our investigation. Computation of Hedges' *g* in 40 different ways is provided by the ES software tool [37, 38]. Descriptions of computations of standardized mean difference effect size estimates for ANOVA designs are provided in [11].

2.1.2. The r family

The *r family* consists of the Pearson product-moment correlation in any of its combinations of continuous and dichotomous variables [33]. For two treatment conditions and a continuous outcome, the effect size is called the point-biserial correlation, which we will refer to as $r_{pb\text{-}pop}$. When $r_{pb\text{-}pop}$ is squared, it is also called η^2 and it can be interpreted to mean the proportion of variance accounted for by the population means. Hence, we can express the population point-biserial correlation as follows:

$$\textit{Population point-biserial correlation, } r_{pb\text{-}pop} = \sqrt{\frac{\sigma^2_{treatment}}{\sigma^2_{total}}}, \tag{5}$$

where the numerator is the variance of the population means around the grand mean, and the denominator is the variance of all scores around the grand mean. $r_{pb\text{-}pop}$ has the value range [0,1]. An estimator of, $r_{pb\text{-}pop}$, based on information from an ANOVA table, is obtained by taking the square root of the explained variance expressed in terms of the sum of squares of the treatments and the total sum of squares:

$$r_{pb} = \sqrt{\frac{SS_{Treatment}}{SS_{Total}}} \tag{6}$$

Formulas based on *t*-values and other statistics, as well as estimators that adjust for bias, are provided in [24, 28, 31, 32, 35].

The point-biserial correlation is affected by the proportion of subjects in each experimental group. It tends to be highest in a balanced design and approaches zero when the design becomes more unbalanced [24]. As a consequence, r_{pb} values from studies with different splits in the sample size will not be directly comparable. To counteract this, the following corrected r_{pb} is recommended [19]:

$$\textit{Corrected } r_{pb} = \frac{a\,r_{pb}}{\sqrt{(a^2-1)r_{pb}{}^2+1}}, \tag{7}$$

where $a = \sqrt{0.25/_{pq}}$, and p and q are the proportions of subjects in each experimental group (p+q=1).

Formula (6) above is applicable for outcomes measured on a continuous scale. When both variables are dichotomous, the population point-biserial correlation is called Φ and is expressed in terms of the proportions in a 2*2 table, [14]. When reporting results from a table larger than 2*2, an effect size estimator called Cramer's V can be applied [14]. When a categorical outcome is measured on an ordinal scale (e.g., small, medium, large), a continuous scale can be assumed and a point-biserial correlation calculated as for continuous outcome [14]. The population effect size will be underestimated if fewer than five categories are applied [38].

It is possible to compute r_{pb} from Hedges' g, and vice versa. Information might be unavailable for computing one or the other, or one may prefer to view the results in terms of a correlation coefficient when g, say, is reported in an article. The following formula maps g to r_{pb} [5, 35]:

$$r_{pb} = \frac{g}{\sqrt{g^2 + (1/pq)*((N-2)/N)}}, \tag{8}$$

where N is the total sample size. Note that the formula is simplified by the factor $1/pq=4$ for a balanced design, (p=q=0.5).

2.1.3. Interpretation of standardized effect sizes

It is not intuitively evident how to interpret standardized effect sizes. Some approaches are listed below and described further in this section.

- Standardized effect sizes can be interpreted in terms of the properties of the formula, for example, distributional overlap for the standardized mean difference and explained variance for the point-biserial correlation.
- Standardized effect sizes can be compared with
    - effect sizes reported in similar experiments,
    - effect sizes reported in the research field in question, for example, software engineering as a whole, and
    - standard conventions for small, medium and large effect sizes developed for research in behavioural science.

The population standardized mean difference, d_{pop}, is expressed in terms of mean difference divided by a measure of the variability in the data. We can interpret this formula as the degree of distributional overlap of values for two populations. A large degree of nonoverlap means a large effect size, and when the two distributions are perfectly superimposed, the effect size is zero [5], see Table 1. This is further visualized in Figure 2: The unstandardized effect sizes (represented by the differences between the full and dotted vertical line) are equal in (a) and (b). However, the standardized effect size in (a) is larger than the one in (b), because the degree of non-overlap is larger in (a) than in (b). The standardized mean difference reflects what is visualized in the figure: The effect size seems important in (a) but might be hardly noticeable in (b).

Table 1
Distributional nonoverlap percentages for values of d_{pop} [5]

d_{pop}	0.0	0.5	1.0	1.3	2.0	3.0	4.0
Degree of non-overlap	0%	33%	55%	65%	81%	93%	98%

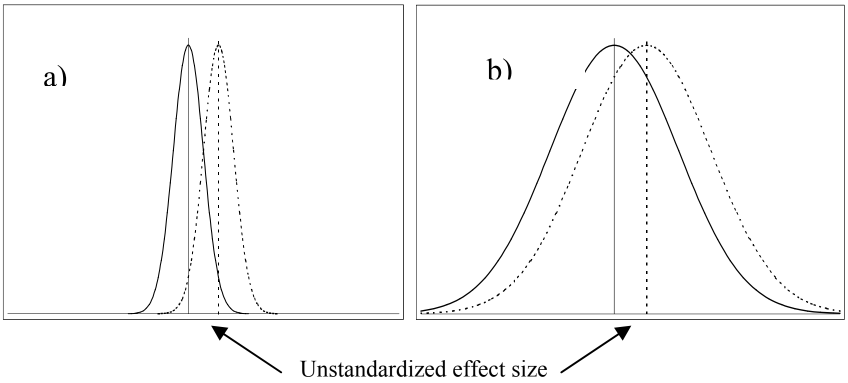


Fig. 2. Illustration of how the standardized mean difference effect size can be interpreted in terms of distributional overlap.

A point-biserial correlation can be interpreted in terms of the property of its square root (see Formula 5 and 6); the percentage of total variance that is explained by treatment.

The second possibility of interpretation of a standardized effect size is to take advantage of its standardized property, i.e., that it is comparable across measurement scales. The best interpretation arises from comparison with experiments that test the same hypothesis as the one in question [9]. In the absence of such experiments, an alternative is to compare the observed effect size to effect sizes reported in the field of interest. We present effect sizes observed in software engineering experiments in Section 4.2.2. A third alternative is to compare the observed effect size against standard conventions that have been developed in behavioural science. Values for small, medium and large population standardized effect sizes corresponding to various statistical tests and types of effect size measures are defined by Cohen (1988, 1992). His definitions are based on a combination of a subjective view of average effect sizes observed in behavioural science and a view of what small, medium and large effect sizes should mean. The definitions for d_{pop} and $r_{pb\text{-}pop}$ are shown in Table 2.

Table 2
Values for small, medium, and large d_{pop} and $r_{pb\text{-}pop}$ [5]

Effect size index		Effect size values		
		Small	Medium	Large
d_{pop}	Standardized mean difference	.20	.50	.80
$r_{pb\text{-}pop}$	Point-biserial correlation	.10	.24	.37

Cohen proposed his definitions for statistical power analyses, to help researchers guess on effect sizes when no other sources for effect size estimation existed, i.e., no similar experiments or pilot studies. His definitions are also used to interpret observed effect sizes, but this is also only advisable when no other sources for effect size estimation are available [43]. In later papers, Cohen recommends reporting effect size with a corresponding confidence interval, but does not himself recommend applying the small, medium and large categories in the evaluation of observed effect sizes [6, 8].

The interpretations described above do not include any contextual information. To evaluate whether an observed effect is of practical importance for a specific context, the effect size must be discussed in relation to each relevant contextual factor, for example, whether the size of efficiency improvement compensates for the effort needed for learning the new method.

2.2. Unstandardized effect size

Unstandardized effect size measures are expressed in terms of raw units of whatever is being measured. This may make the effect sizes easier to interpret, but in contrast to standardized effect sizes, they are not independent of measurement scale. Examples are these: (i) the difference between mean values (e.g., the difference in time taken to perform a given task when using two different methods), (ii) percentage mean difference, and (iii) the difference in proportion of subjects (e.g., the difference between experimental groups with respect to the proportion of subjects viewing a script as correct). The concept of population effect size applies here as well, for example, the effect size measure for population mean difference is expressed as follows:

$$Population\,mean\ difference = \mu_1 - \mu_2, \tag{9}$$

where μ_i is the mean value in population i, which is estimated by the mean $\bar{x}_i$, The standardized counterpart is the standardized mean difference (Formula 1).

Unstandardized effect sizes lend themselves more directly to interpretations of practical importance than do standardized values. For example, an unstandardized effect size of eight hours difference in development effectiveness between two methods used for the same task serves as a better basis for judging the practical importance of the result than a standardized effect size of g=0.5.

2.3. Nonparametric effect size

The standardized effect size measures described in the preceding sections assume parametric models for the outcome variable. Most of the standardized effect size measures developed are parametric. However, assuming parametric models may be inappropriate in many instances, and standardized nonparametric effect size measures based on median values have been suggested in the literature [16, 25, 26]. Computation of these measures requires raw data that is seldom available in articles presenting experimental results. Hence, these nonparametric effect size measures are appropriate for reporting effect sizes, but not always useful in meta-analyses.

Alternatives or supplements to the standardized nonparametric effect size measure are the unstandardized difference in median values or graphical presentations, for example, two box plots within the same figure for easy comparison.

3. Research method

This section describes how we identified the controlled experiments and primary tests, what kind of information we gathered, and how effect size estimates were computed.

3.1. Identification of controlled experiments and primary tests

We assessed all the 103 papers on controlled experiments (of a total of 5453 papers), identified by Sjøberg *et al.* [39]. Table 3 shows the actual journals and conference proceedings, which were chosen because they were considered to be representative of empirical software engineering research. Furthermore, since controlled experiments are empirical studies that employ inferential statistics, they were considered a relevant sample in this study. The 103 articles reported 113 controlled experiments. The article selection process was determined from predefined criteria as suggested in [22], see [39] for full details.

Since the term "experiment" is used inconsistently in the software engineering community (often being used synonymously with empirical study), we use the term "controlled experiment". A study was defined as a controlled experiment if individuals or teams (the experimental units) conducted one or more software engineering tasks for the sake of comparing different populations, processes, methods, techniques, languages or tools (the treatments). We did not distinguish between randomized experiments and quasi-experiments in this study, because both designs are relevant to software engineering experimentation. In this article, we consistently use the term 'experiment' in the above-mentioned sense of "controlled experiment".

Table 3
Distribution of articles describing controlled experiments in the period Jan. 1993 – Dec. 2002

Journal/Conference Proceeding[2]	Number	Percent
Journal of Systems and Software (JSS)	24	23.3
Empirical Software Engineering (EMSE)	22	21.4
IEEE Transactions on Software Engineering (TSE)	17	16.5
International Conference on Software Engineering (ICSE)	12	11.7
IEEE International Symposium on Software Metrics (METRICS)	10	9.7
Information and Software Technology (IST)	8	7.8
IEEE Software	4	3.9
IEEE International Symposium on Empirical Software Engineering (ISESE)	3	2.9
Software Maintenance and Evolution (SME)	2	1.9
ACM Transactions on Software Engineering (TOSEM)	1	1.0
Software: Practice and Experience (SP&E)	-	-
IEEE Computer	-	-
TOTAL:	103	100%

[2] The conference *Empirical Assessment & Evaluation in Software Engineering* (EASE) is partially included in that ten selected articles from EASE appear in special issues of JSS, EMSE, and IST.

Results from several statistical tests were often reported in the reviewed articles; one article reported 74 tests. We therefore classified each statistical test as either *primary* or *secondary*. The *primary* test what the experiment is designed to evaluate. They were specified in the article by hypotheses or research questions. If no hypothesis or research question was stated, we classified as *primary* those tests that were described to address the main incentive of the investigation. *Secondary tests* comprised all other tests.

Two of the authors of this paper read all the 103 articles and made separate extractions of the primary tests. Then three of the authors reviewed these two data sets to reach a consensus on which experiments and tests to include. In 14 of the experiments, no statistical testing was performed, and the corresponding articles were thus excluded from the investigation. Seven experiments were excluded because it was impossible to track which result answered which hypothesis or research question. Four experiments were reported in more than one article. In these cases, we included the most recently published. We identified 459 statistical tests corresponding to the main hypotheses or research questions of 92 experiments. Of these tests, we excluded 25 tests of interaction effects, because no well-developed procedures exist for computing effect sizes for interactions [11]. In addition, five tests were excluded because they were regression analyses and involved no treatment. Thus, the final set comprised 429 primary tests, detected in 92 experiments and 78 articles (Figure 3).

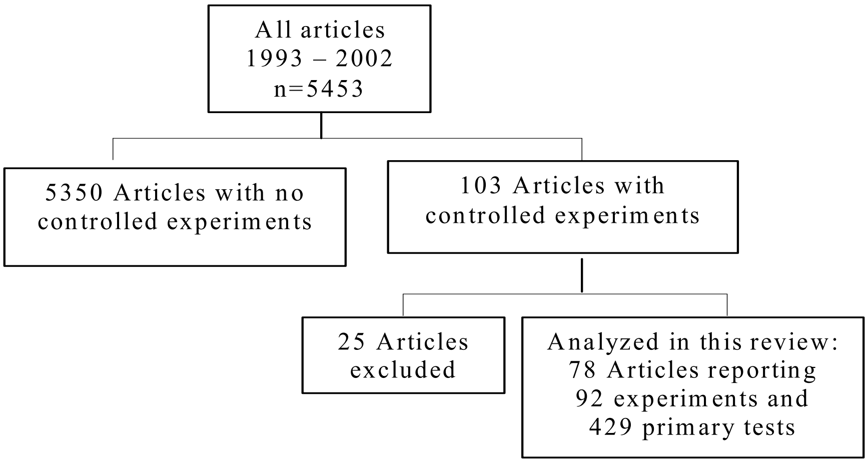


Fig. 3. Results of the literature review selection process.

3.2. Information extracted

For each primary test, we recorded

- whether a standardized and/or unstandardized effect size or a graphical visualization of the effect size was reported,
- when an effect size was reported, the interpretation of the effect size and whether practical importance was discussed, and
- sample size, level of significance, *p*-value or information about rejection or acceptation of the null hypothesis, and whether the test was one or two-sided.

In addition, we registered descriptive statistics and estimated the standardized mean difference effect size for those tests with sufficient information reported. Our aim with this computation was to investigate the range of effect sizes in software engineering experiments across experimental topic, treatment and outcome. We therefore estimated the same standardized mean difference population effect size, d_{pop}, for all tests, applying the absolute value for Hedges' g as the estimator. Each estimate was corrected for bias by Formula 4 in Section 2.1.1.

The primary tests included parametric tests that compare mean values, nonparametric tests that compare median values or ranks, and tests of the values of dichotomous variables. The applied estimation formulas are listed in Table 7.

We investigated the effect between *two* treatment conditions. Hence, when the primary test was an overall comparison of more than two treatment conditions, we looked at the pair-wise comparisons (contrasts) for our effect size estimation.

We wanted to present the effect sizes as point-biserial correlations as well as standardized mean differences. The g-values were transformed into r_{pb}, by applying Formula (8) in Section 2.1.2. Then the values were corrected for unbalanced design by Formula (7). This correction did not change the values to a great extent, since half of the tests had balanced design and the split in sample size was larger than 70-30 for eight tests only (see Section 2.1.2). For those primary tests for which g could not be computed, there was not sufficient information to compute r_{pb}, either.

As stated in Section 2.1.1, the pooled standard deviation assumes that the standard deviations are equal in both treatment groups. To check this assumption, we calculated the ratio of standard deviations, when these were reported. The ratio of the largest standard deviation over the smallest standard deviation exceeded four (Section 2.1.1) in seven tests. Consequently, we did not include effect sizes for these tests.

Ten tests were one-sided with results in the direction opposite to the alternative hypothesis. We regarded effect sizes for these tests as real effects and included them in our analysis.

4. Results

The findings comprise two main parts: (1) How effect sizes were reported in the surveyed experiments, with respect to the extent of reporting and interpretation of the reported values and (2) the result of our estimation of standardized effect sizes from information reported in the surveyed experiments.

4.1. The reporting of effect sizes in the surveyed experiments

4.1.1. Extent of effect size reporting

Only 29% of the experiments reported at least one effect size; see Table 4. Two of the 92 experiments reported both standardized and unstandardized effect sizes, eight reported standardized effect sizes only and 17 reported unstandardized effect sizes only. Standardized and unstandardized effect sizes were reported for, respectively, 55 and 46 of the 429 primary tests of the reviewed experiments.

Table 4
Extent of effect size reporting for experiments and primary tests, presented per type of statistical test method

Levels of effect size reporting	Experiments		Primary tests							
			Total		Parametric tests		Non-parametric tests		Tests of dichotomous variables	
	N	%	N	%	n	%	n	%	n	%
Both standardized and unstandardized	2	2.2	3	0.7	3	1.0	0	0	0	0
Standardized (only)	8	8.7	52	12.1	46	15.7	6	6.4	0	0
Unstandardized (only)	17	18.5	43	10.0	32	10.9	6	6.4	5	11.9
No effect size	65	70.7	331	77.2	212	72.4	82	87.2	37	88.1
Total	92	100	429	100	293	100	94	100	42	100

The different types of effect size measures are related to types of outcome and thereby to types of statistical test. Table 4 shows that standardized effect sizes were reported most frequently for parametric tests (46+3 of 293, that is, 17 percent), only a few for nonparametric tests (6 percent) and not for any tests of dichotomous variables. The corresponding parametric tests were ANOVA and *t*-tests; the nonparametric tests were Wilcoxon match pair tests. The standardized mean difference was reported for all but one test, for which the point-biserial correlation coefficient was reported (for an ANOVA test).

Unstandardized effect sizes were reported in equal proportions for parametric tests and tests of dichotomous variables (32+3 of 293 and 5 of 42, respectively, that is, 12 percent) and to a lesser extent for nonparametric tests (6 percent) see Table 4. Most of the 46 unstandardized effect sizes were reported as percentage mean difference (21 tests), but reported were also absolute mean difference (nine tests), difference in proportions or percentage (five tests), ratio of mean values (five tests), difference in average rank values (three tests) and confidence interval for the mean difference (three tests).

For most of the 331 primary tests for which no effect size was reported, mean values, frequencies or graphical presentations of results per experimental group were reported.

We compared the extent of effect size reporting according to whether the results were significant or not (as defined by the authors); see Table 5. For standardized effect sizes there was no difference, but unstandardized effect sizes were reported to a greater extent when significant results occurred than when non-significant results occurred (17.9 percent versus 3.7 percent).

Table 5
Reporting of effect size and significance of results

Levels of effect size reporting	Primary test results				
		Significant		Non-significant	
	N	n	%	n	%
Both standardized and unstandardized	3	3	1.42	0	0
Standardized effect size (only)	52	24	11.3	28	12.9
Unstandardized effect size (only)	43	35	16.5	8	3.7
No effect size	331	150	70.8	181	83.4
Total	429	212	100	217	100

Another factor that seems to influence the extent of effect size reporting is the number of treatment conditions tested in the experiment. None of the 51 primary tests that compared more than two treatment conditions reported the standardized effect size for the pair wise comparisons of treatments. Only four of these 51 tests reported the unstandardized effect size.

4.1.2. The interpretation of the effect sizes given by the authors of the reviewed experiments

Possible ways of interpreting the standardized effect size was presented in Section 2.1.3. In one of the surveyed experiment, the point-biserial correlation was interpreted as the percentage of explained variance, but the standardized mean difference effect size was not interpreted in terms of distributional overlap for any of the experiments.

One article reported and compared the standardized effect sizes from three related experiments. For the other experiments, standardized effect sizes were not compared with related research. In two experiments, effect sizes were reported to aid future researchers in planning their experiments, but the sizes were not discussed as part of the result. For the other experiments, standardized mean difference effect sizes were compared with Cohen's conventions from behavioral science [5], for example:

> We intend to discuss all practically significant results and not constrain ourselves to discussing only statistically significant results. For this exploratory study we consider effects where $\gamma \geq 0.6$ to be of practical significance (the unit is one standard deviation). We make this decision on the basis of effect size indices proposed by Cohen (1969).

This author judged sizes above 0.6 to be of practical importance. Two authors considered sizes above 0.5 to be of practical importance and one author regarded observed sizes of 0.77 as large. The unstandardized effect sizes were reported with no interpretations or references to practical importance, for example, "Procedural roles reduced the loss of only singular defects by about 30%."

4.2. Our computation of standardized effect sizes from information provided in the surveyed experiments

To identify the sizes of treatment effects found in software engineering experiments, we estimated standardized effect sizes for the primary tests in the reviewed experiments.

4.2.1. Extent of information available for effect size estimation

We managed to estimate standardized mean difference effect sizes for a total of 284 primary tests based on information provided in the reviewed articles. These tests were located in 64 (70%) of the 92 reviewed experiments.

The numbers of effect sizes that were estimated for the various statistical tests are shown in Table 6. Tests comparing two treatment conditions had sufficient information for effect size estimation to be reported for 64% of the parametric tests of continuous variables. The results for nonparametric tests and tests of dichotomous variables were 33% and 79%, respectively. The corresponding results for tests comparing more than two treatment conditions were lower; respectively, 33%, 25% and 25%. Hence, when more than two treatment conditions were compared in a test, information for effect size estimation for the corresponding pair-wise tests was, overall, sparsely reported in the reviewed articles.

Table 7 shows the formulas applied in our effect size estimation. Formula 2 was applied for the majority of tests, including 33 nonparametric tests. We considered mean values to be an appropriate measure of distributional location for nonparametric tests, as long as they were reported in the paper. In those cases where means and standard deviations were not reported, Formulas 3, 4, 5, 6 and 7, which are based on *t*-value, *F*-values, *p*-value, mean square error and/or sample sizes, respectively, were applied for parametric tests. Formula 8 was applied for tests of dichotomous variables when frequencies and sample sizes were reported.

Table 6
Extent of effect size estimation per type of statistical test method

Statistical test method	Total number of primary tests	Primary tests comparing two treatment conditions*			Primary tests comparing more than two treatment conditions*				**Total number of effect sizes computed**
		N	#ES	%	N	n	%	#ES	#ES
Parametric test of continuous dependent variable	**293**	**250**	**160**	**64**	**43**	**14**	**33**	**55**	**215**
ANOVA	116	78	50	64	38	12	32	40	
t-test	79	79	67	85	0				
Paired *t*-test	39	39	35	90	0				
ANCOVA	28	28	0	0	0				
Tukey’s pair wise comparisons	18	18	0	0	0				
Repeated ANOVA	8	5	5	100	3	1	33	6	
Poisson regression	3	3	3	100	0				
Duncan posttest ANOVA	1	0			1	0	0		
Repeated MANOVA	1	0			1	1	100	9	
Nonparametric test of continuous dependent variable	**94**	**90**	**30**	**33**	**4**	**1**	**25**	**3**	**33**
Wilcoxon	41	41	22	54	0				
Mann-Whitney	39	39	2	5	0				
Kruskal-Wallis	8	4	0	0	4	1	25	3	
Rank-sum test	6	6	6	100	0				
Dichotomous dependent variable	**42**	**38**	**30**	**79**	**4**	**1**	**25**	**6**	**36**
Chi-square	25	21	16	76	4	1	25	6	
Fisher’s exact test	15	15	12	80	0				
Proportion test	2	2	2	100	0				
Total	**429**	**378**	**220**	**58%**	**51**	**16**	**31%**	**64**	**284**

* **N**: total number of primary tests
n: number of primary tests for which effect sizes could be estimated for the pair-wise comparisons, for tests comparing more than two treatments
#ES: number of effect sizes estimated

Table 7
The estimation formulas for Hedges' *g* that were applied in this investigation

No	Data needed and definition of terms	Estimation formulas	References	Number of *g* estimated
1	Hedges' *g*	*g* reported in the paper		18
2	Mean values, standard deviations and group sample sizes	$g = \frac{\overline{X}_1 - \overline{X}_2}{s_p}$	[28]	190
3	Independent t-test value and sample size (n) for each group	$g = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$	[28]	16
4	F-ratio from two groups, one way ANOVA	$g = \sqrt{\frac{F\,(n_1 + n_2)}{n_1 n_2}}$	[28]	13
5	*P*-value and sample size/degrees of freedom	Find t-value based on the p-value and sample sizes, and use Formula 1.	[28]	1
6	Repeated measure design. One between-subject factor and one within-subject factor, t is the number of time points, MS_{bse} is the between-subject mean square error and MS_{wse} is the within-subject mean square error.	Formula (2) in the text using the following estimate for standard deviation $S = \sqrt{\frac{MS_{bse} + (t-1)MS_{wse}}{t}}$	[38], where also estimators for MS_{bse} and MS_{wse} are provided.	4
7	Factorial design.	Formula based on means, sample sizes, standard deviations, corrected for the other factors.	[11, 30]	6
8	Dichotomous outcome, 2*2 table of frequencies.	$g = \frac{\ln(odds\ outcome\ A) - \ln(odds\ outcome\ B)}{\pi/\sqrt{3}}$	[15, 28, 38]	36
Total				284

4.2.2. Standardized effect size values

The values for the 284 estimates of Hedges' *g* range from 0 to 3.40 with a median value of 0.60; see Table 8. The cumulative percentages in the table are, for each *g*, the percentage of effect sizes equal to or below that value. For example, 68% of the effect sizes in our review are equal to or below *g*=1.00. For readers who prefer to view standardized effect sizes in terms of correlations, the r_{pb} values are also presented in Table 8. The range of values is (0, 0.87) with a median value of 0.3 and represents effect sizes that can be expected in studies with balanced design. When the design is unbalanced, the effect sizes tend to decrease with increased split in experimental group sizes and the researcher should be aware of this when comparing r_{pb} values from different experiments.

We defined size categories of the estimated *g* and r_{pb} values by viewing the lower 33% of the effect sizes, the middle 34%, and the largest 33%. In Table 9, we present these categories, and we let the median value in these categories represent small, medium and large effect sizes..

Table 8
Cumulative percentages for estimated values for Hedges' *g* and the point-biserial correlation

Hedge's *g*	Cumulative percentages for 284 *g* effect size estimates in software engineering experiments	Point-biserial correlation	Cumulative percentages for 284 r_{pb} effect size estimates in software engineering experiments
0.00	7	0.00	7
.10	11	0.10	19
.20	19	0.20	35
.30	28	**0.30**	**50 median**
.40	35	0.40	62
.50	42	0.50	70
.60	**50 median**	0.60	84
.70	56	0.70	92
.80	60	0.80	97
.90	64	0.90	100
1.00	68		
1.10	71		
1.20	73		
1.30	77		
1.40	83		
1.50	86		
1.60	88		
1.70	90		
1.80	90		
1.90	93		
2.00	95		
2.30	97		
2.50	97		
3.00	99		
3.40	100		
Mean *g*	0.81	Mean r_{pb}	0.34
Std *g*	0.69	Std r_{pb}	0.23

Table 9
Small, Medium and Large categories for 284 estimated values for Hedges' *g* and the point-biserial correlation

Size category	Hedges*'* *g*		Point-biserial correlation*,* r_{pb}	
	Effect sizes	Median	Effect sizes	Median
Small (lower 33%)	0.00 to 0.376	0.17	0.00 to 0.193	0.09
Medium (middle 34%)	0.378 to 1.000	0.60	0.193 to 0.456	0.30
Large (upper 33%)	1.002 to 3.40	1.40	0.456 to 0.868	0.60

5. Discussion

This section discusses the findings, their implications, and the limitations to this review.

5.1. Comparison with research in behavioural science

It is only in the psychological and educational sciences that we have found similar investigations of effect size reporting, and these assessed only the reporting of *standardized* effect sizes. An assessment of 226 articles on educational and psychology research in 17 journals published in 1994-1995 revealed that standardized effect sizes were reported in 16 articles (7.1%) [20]. Both univariate and multivariate tests, analyzed by several different statistical methods, were included in these 226 articles. This is similar to the proportion of articles reporting standardized effect sizes found in our review (7.7%).

A study by Fidler *et al.* [13] investigated 239 articles published in 1993-2001 that reported new empirical data in the Journal of Consulting and Clinical Psychology. They found that standardized effect size was reported to a greater degree in articles that reported ANOVA tests and Chi-square tests, compared with our review; 32% and 13% compared with 3% and 0, respectively; see Table 10. The extent to which standardized effect sizes were reported in articles that reported *t*-tests was similar in our and Fidler *et al.'s* investigation (15% and 16%, respectively).

Table 10
Number of articles reporting effect size. Comparison of published experiments in software engineering and studies in psychology.

Source	Type of statistical test method applied *		
	ANOVA	*t*-test	Chi-square
Articles reporting controlled experiments in software engineering (This review)	3% (1 of 32)	16% (5 of 32)	0% (0 of 9)
Articles reporting psychology studies [13]	32% (38 of 120)	15% (16 of 108)	13% (16 of 126)

*In our review,116 ANOVA tests were reported in 32 articles, 118 *t*-tests were reported in 32 articles and 25 chi-square tests were reported in nine articles.

Considering the maturity of psychological and educational research compared with the relative young field of empirical software engineering, the sparse reporting of effect sizes in our field may be expected. It was more surprising to find similar results to those of Keselman *et al.* and Fidler *et al*. Still, this is a poor consolation, because the extent of effect size reporting in the field of psychological and educational research is regarded as too low, [13, 20].

The sparse reporting of standardized effect sizes in software engineering might be due to effect size estimation's being little known. It is not a topic in standard research methods courses, and formulas for the calculation of effect sizes do not appear in many statistical text books (other than those devoted to meta-analysis). This may improve, as recent literature in empirical software engineering recommends the reporting of effect sizes [12, 23, 29].

However, encouragements for the reporting of effect sizes do not seem to suffice. In the behavioural sciences, it has been suggested that changes in editorial policies will be required before reporting effect sizes will become a matter of routine [13, 44]. Trusty *et al.* [42] report that 23 journals in the social sciences now require that effect sizes be reported, and in their paper, they provide practical information for studies submitted to the Journal of Counseling & Development on generating, reporting and interpreting effect sizes for various types of statistical analysis.

We found one study in the behavioural sciences on the aggregation of standardized effect sizes that was comparable with ours; 1766 effect sizes (standardized mean differences) were estimated from 475 psychotherapy studies [10, 40]. This study found the same distribution of effect sizes as we obtained. Hence, the treatment effects observed in software engineering experiments are of the same magnitude as effects found in a large number of psychotherapy studies; the same average and nearly the same spread of values.

As shown in Table 9, we categorized the effect sizes in our review into the 33% smallest, the 34% middle and 33% largest values and let the median values in these categories represent small, medium and large values in the data. In Table 11, we compare the standardized mean difference effect sizes with corresponding results from an aggregation of average effect sizes from meta-analyses of psychological, educational and behavioural treatments effectiveness [27] (including the study of psychology studies by Smith *et al.*) and the conventions for small, medium and large effect sizes in the behavioural sciences [5].

The medium and large effect sizes in our review are larger than those observed in the meta-analyses and the conventions from the behavioural sciences. (Note that when we considered the median value as appropriate measure of the middle of the categories, the middle point values were even larger: (small: 0.19, medium: 0.69 and large: 2.2). The discrepancies between the aggregated effect sizes on a study level and the aggregated effect sizes on a meta-analysis level can be explained by the fact that the smallest and largest values on a study level disappear in the overview of average values on the meta-level. The standard conventions in the behavioural sciences seek to represent average values, which seems to be confirmed by the results from the aggregation of meta-analyses. Hence, as our results are the same as those from the aggregation of psychology studies, this might indicate that the conventions from the behavioural sciences (i.e. Cohen's definitions) are appropriate comparators for average effect sizes in software engineering experiments as well (when relevant related research is not present). The effect sizes obtained in our review provide additional information about the range of values in our field for Hedges' *g* and the point-biserial correlation.

Table 11
Small, medium and large standardized mean difference effect sizes as observed in this review, in an aggregation of meta-analyses in the social sciences and the conventions in the behavioural sciences

Source	N	Standardized mean difference values		
		Small	Medium	Large
Software engineering experiments (this review)*	284 effect sizes	0.17	0.60	1.40
Meta-analyses of psychological, educational and behavioural studies, [27]†	102 average effect sizes	0.15	0.45	0.90
Conventions from the behavioural sciences, [5]	Not empirically based	0.20	0.50	0.80

* The effect sizes were obtained as the median values for the 33% smallest, the 34% medium and the 33% largest values.

† The effect sizes were obtained as the middle point among the 33% smallest, the 34% medium and the 33% largest values.

5.2. Guidelines for reporting effect sizes

This section offers guidelines on how to report effect sizes.

5.2.1. Always report effect size

We recommend always reporting effect sizes as part of the experimental results, because there is a risk of making poor inferences when effect sizes are not assessed: (A) nonsignificant results might erroneously be judged to be of no practical importance, and (B) statistical significance might be mistaken for practical importance; see Table 12.

Table 12
Potential problems of inference, when the effect size is not reported,, as a function of statistical significance and effect size [35]

		Effect size	
		Acceptably large	Unacceptably small
Statistical Significance	*p*-values low enough	No inferential problem	(B) Mistaking statistical significance for practical importance
	p-values too high	(A) Failure to perceive practical importance of "non-significant" results	No inferential problem

The advantage of assessing both effect sizes and statistical significance when making inferences is illustrated by one of the reviewed experiments in which object-oriented design was compared with structured design with respect to the percentage of task-related questions that were answered correctly. The results of statistical tests were nonsignificant at the 0.1 level. The standardized effect size was reported as 0.7, which was regarded as practically important according to Cohen’s definitions. The sample size was 13, whereas 56 subjects were needed to achieve a power of 80% at the 0.1 level of significance. If only statistical significance had been reported, the result would have seemed less important than the effect size suggested it to be.

5.2.2. Discuss practical importance

The evaluation of effect sizes based on average values or standard conventions is a first step on the road to assessing the practical importance of the result. For a complete evaluation of practical importance, the effect sizes must be judged in context. Since judging the practical importance of one's experiment is nearly impossible without the relevant situational context and since the experimental results may be applicable in a wide range of contexts, it may be unrealistic to expect researchers to grade their results in terms of practical importance in their research papers. Nevertheless, we believe that the relevance of software engineering studies would be increased if researchers discussed this issue, possibly through illustrative examples.

Moreover, when an appropriate effect size is reported, the reader can assess practical importance by applying it in their context-specific cost-benefit analysis, as also suggested by [36].

5.2.3. Report both standardized and unstandardized effect size

We recommend reporting both standardized and unstandardized effect sizes, because these two types are supplementary. A standardized effect size includes the variability in the data and gives a complete “average” based on all the data in the sample. There are several approaches to interpreting standardized effect sizes as described in Section 2.1.3. Apply each of them if they bring more information to bear regarding discussion of the result. Moreover, reporting standardized effect sizes aids researchers in planning new experiments (power analysis) and enables comparisons with their own findings.

An unstandardized effect size is easier to interpret than a standardized one and serves as a good basis for discussing practical importance. We place particular emphasis on the value of measures in percentages, which makes the measure applicable to larger-scale projects.

5.2.4. Use the tool box of effect size measures

Many types of standardized effect size measures have been developed, 40 of which are presented in Kirk [21]. However, only two types were reported in the reviewed experiments: the standardized mean difference and the point-biserial correlation. Both of these are parametric. No standardized nonparametric effect size measures were used for the 22% of tests that were analysed by nonparametric methods, neither were any unstandardized effect size measures based on median values used.

When reporting experimental results, we will urge researchers to apply the effect size measure that best suites the data, e.g., nonparametric effect size measures for observations that cannot be assumed to have any known distribution. When aggregating results from different measurement scales, the choices are limited; the standardized mean difference effect size and the point-biserial correlation are most commonly used, because they provide good approximation formulas for variables that are not continuous.

5.2.5. Report confidence intervals

When reporting an effect size, the accuracy of the estimate, measured in terms of a confidence interval, should be reported as well. Although the exact calculation of confidence intervals for a standardized effect size is complicated, good approximations exist for small effect sizes and sample sizes that exceed 10 per group. Descriptions of both exact methods and approximations are found in [14, 17, 24]. Calculating a confidence interval for an unstandardized effect size is simpler and is provided by most statistical reporting tools.

5.2.6. Report descriptive statistics

We recommend always reporting, for each experimental group, results as mean values, standard deviations, frequencies and sample sizes. When performing analysis of variance, report standard ANOVA table results. Such information enables the reader to estimate effect sizes. Even if you report the effect size measure you find most appropriate, the reader might wish to compute a different one, to aggregate results or for purposes of comparison. For factorial design, there might be different views on how to include the effect of different factors; hence, descriptive statistics for subgroups might be useful.

5.3. Implication for power analysis

For statistical power analysis, Dybå *et al.* [12] recommend applying a medium effect size, as defined by Cohen, (for example, *g*=0.5) when no other information about the population standardized effect size is available. Table 8 can be used as a guide to assess the likelihood of obtaining specific values for Hedges' *g* and the point-biserial correlation. For example, there is a likelihood of 58% (100% - 42%) that Hedges' *g* will be larger then 0.5 in software engineering experiments.

If only large effects are interesting to detect, a large effect size is appropriate to apply in the power-analysis. Moreover, if sufficient power is seen as difficult to achieve, we recommend abstaining from hypothesis testing, and recommend instead reporting effect sizes and confidence intervals when investigating hypotheses. Note that confidence intervals contain all the information to be found in significance tests and much more [8].

5.4. Limitations of this study

The main limitations to this investigation are selection bias regarding articles and tests, and possible inaccuracy in data extraction. The limitations regarding selection of articles and tests are described in, respectively, [39] and [12].

The coding of effect size reporting has two limitations: it was performed by one person only, and the quantitative categorization represents a simplification of the complex matter of reporting experimental results. Important nuances might have been lost and some experiments treated "unfairly". However, the categorization was checked, rechecked and discussed among all authors.

The effect size calculations were also performed by one person only. Moreover, those tests for which an effect size was not calculated, due to lack of sufficient information reported in the article, represent a limitation to the completeness of the presentation of effect sizes. Possible effect size calculation formulas and data that may have been used for effect size calculation might have been overlooked in the reported experiments. Finally, the calculated effects might be biased by any methodological inadequacies of the original studies.

6. Conclusion

This review investigated the extent of effect size reporting in selected journals and conference proceedings in the decade 1993-2002, the interpretation of the effect sizes given by the authors of the reviewed experiments, the extent to which experimental results are reported in such a way that standardized effect sizes can be estimated, and the standardized effect sizes detected in software engineering experiments.

We found that effect sizes were sparsely reported in the reviewed experiments. Only 29% of the 92 experiments reported at least one standardized and/or unstandardized effect size, and only two experiments reported both. The extent to which standardized effect size was reported was equal to or below what is observed in research in psychology.

The standardized effect sizes were compared mainly with the standard conventions for small, medium and large values defined by Jacob Cohen for the behavioural sciences. The practical importance of the effect size in context was not discussed in any of the experiments.

We found sufficient information in the reviewed experiments to compute standardized effect sizes for 25% to 79% of the primary tests, depending on the type of test.

The effect sizes computed in this investigation were similar to what is observed in individual studies in research in psychology. These values are slightly larger than the standard conventions for small, medium and large effect sizes in the behavioural sciences.

Based on our experiences with working with this review, we have three main recommendations to make regarding effect size reporting. (1) Always report effect size in addition to statistical significance, to avoid erroneous inferences. (2) Avoid allowing the effect size interpretation to become rigorous and a matter of routine. Apply the unstandardized effect size for discussions of practical importance in context. (3) Always report basic descriptive statistics, such as means, standard deviations, frequencies and sample size, for each experimental group. This will enable researchers to estimate their own choice of effect sizes.

Acknowledgements

This research is funded by the Research Council of Norway through the INCO project. The authors are grateful to Lionel Briand, Magne Jørgensen and Amela Karahasanovic for useful discussions and to the anonymous referees for valuable comments. Thanks to Chris Wright for proofreading the paper.

References

[1] D.G. Altman, K.F. Schulz, D. Moher, M. Egger, F. Davidoff, D. Elbourne, P.C. Gøtzsche, and T. Lang, The revised CONSORT statement for reporting randomized trials: Explanation and elaboration, Annals of Internal Medicine 134 (8) (2001) 663-694.

[2] J.A. Breaugh, Effect size estimation: Factors to consider and mistakes to avoid, Journal of Management 29 (1) (2003) 79-97.

[3] W.G. Cochran, Problems arising in the analysis of a series of similar experiments, Journal of the Royal Statistical Society (Suppl.) 4 (1937) 102-118.

[4] J. Cohen, Some statistical issues in psychological research. In B.B. Wolman (Ed.) Handbook of Cinical Psychology, Academic Press, New York, 1965.

[5] J. Cohen, Statistical Power Analysis for the Behavioral Science, Lawrence Erlbaum Hillsdale, New Jersey, 1988.

[6] J. Cohen, Things I have learned (so far), American Psychologist 45 (12) (1990) 1304-1312.

[7] J. Cohen, A power primer, Psychological Bulletin 112 (1) (1992) 155-159.

[8] J. Cohen, The earth is round (p< .05), American Psychologist 49 (12) (1994) 997-1003.

[9] H.M. Cooper, On the significance of effects and the effects of significance, Journal of Personality and Social Psychology 41 (5) (1981) 1013-1018.

[10] D.S. Cordray and R.G. Orwin, Improving the quality of evidence: Interconnections among primary evaluations, secondary analysis, and quantitative synthesis, Evaluation Studies Review Annual 8 (1983) 91-119.

[11] J.M. Cortina and H. Nouri, Effect Size for Anova Designs, Sage, Thousand Oaks, CA, 2000.

[12] T. Dybå, V.B. Kampenes, and D.I.K. Sjøberg, A systematic review of statistical power in software engineering experiments, Information and Software Technology 48 (8) (2006) 745-755.

[13] F. Fidler, G. Cumming, N. Thomason, D. Pannuzzo, J. Smith, P. Fyffe, H. Edmonds, C. Harrington, and R. Schmitt, Toward improved statistical reporting in the journal of consulting and clinical psychology, American Psychological Association 73 (1) (2005) 136-143.

[14] R.J. Grissom and J.J. Kim, Effect Size for Research. A Broad Practical Approach, Lawrence Erlbaum Associates, Inc., Publishers, 2005.

[15] V. Hasselblad and L.V. Hedges, Meta-analysis of screening and diagnostic tests, Psychological Bulletin 117 (1995) 167-178.

[16] L.V. Hedges and I. Olkin, Nonparametric estimators of effect size in meta-analysis, Psychological Bulletin 96 (3) (1984) 573-580.

[17] L.V. Hedges and I. Olkin, Statistical Methods for Meta-Analysis, Academic Press, Inc., 1985.

[18] C.R. Hill and B. Thompson, Computing and interpreting effect sizes, in: J.C. Smart (Ed.), Higher Education: Handbook of Theory and Research, Kluwer Academic Publishers, 2004 175-196.

[19] J.E. Hunter and F.L. Smith, Methods for Meta-Analysis, Sage, Thousand Oaks , CA, 2004.

[20] H.J. Keselman, C.J. Huberty, L.M. Lix, S. Olejnik, R.A. Cribbie, B. Donahue, R.K. Kowalchuk, L.L. Lowman, M.D. Petosky, J.C. Keselman, and J.R. Levin, Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses, Review of Educational Research 68 (3) (1998) 350-386.

[21] R.E. Kirk, Practical significance: A concept whose time has come, Educational and Psychological Measurement 56 (5) (1996) 746-759.

[22] B. Kitchenham, Procedures for performing systematic reviews, Keele University Technical Report TR/SE-0401 and NICTA Technical Report 0400011T.1. (2004)

[23] B.A. Kitchenham, B.A. Pfleeger, S.L. Pickard, L.M. Jones, P.W. Hoaglin, D.C. El Emam, and K. Rosenberg, Preliminary guidelines for empirical research in software engineering, IEEE Transactions on Software Engineering 28 (8) (2002) 721-734.

[24] R.B. Kline, Beyond Significance Testing. Reforming Data Analysis Methods in Behavioral Research, American Psychological Association, Washington DC, 2004.

[25] H.C. Kraemer and G. Andrews, A nonparametric technique for meta-analysis effect size calculation, Psychological Bulletin 91 (2) (1982) 404-412.

[26] J. Krauth, Nonparametric effect size estimation: A comment on Kraemer and Andrews, Psychological Bulletin 94 (1) (1983) 190-192.

[27] M.W. Lipsey, Design Sensitivity: Statistical Power for Experimental Research, Sage, Newbury Park, CA, 1990.

[28] M.W. Lipsey and D.B. Wilson, Practical Meta-Analysis, Sage, Thousand Oaks, 2001.

[29] J. Miller, Applying meta-analytical procedures to software engineering experiments, Journal of Systems and Software 54 (2000) 29-39.

[30] H. Nouri and R.H. Greenberg, Meta-analytic procedures for estimation of effect sizes in experiments using complex analysis of variance, Journal of Management 21 (4) (1995) 801-812.

[31] S. Olejnik and J. Algina, Measures of effect size for comparative studies: Application, interpretations, and limitations, Contemporary Educational Psychology 25 (2000) 241-286.

[32] S. Olejnik and J. Algina, Generalized eta and omega squared statistics: Measures of effect size for some common research designs, Psychological Methods 8 (4) (2003) 434-447.

[33] R. Rosenthal, Effect sizes in behavioral and biomedical research, in: L. Bickman (Ed.), Validity & Social Experimentation: Donald Campbell's Legacy, Sage, Thousand Oaks, CA, 2000 121-139.

[34] R. Rosenthal and M.R. DiMatteo, Meta-analysis: Recent development in quantitative methods for literature reviews, Annual Review of Psychology 52 (2001) 59-82.

[35] R. Rosenthal, R.L. Rosnow, and D.B. Rubin, Contrasts and Effect Sizes in Behavioural Research. A Correlational Approach, Cambridge University Press, 2000.

[36] L. Sechrest and W.H. Yeaton, Empirical bases for estimating effect sizes, in: R.F. Boruch, P.M. Wortman, and D.S. Cordray (Ed.), Reanalyzing Program Evaluations, Jossey-Bass, San Francisco, 1981

[37] W.R. Shadish, L. Robinson, and C. Lu, ES: A computer program for effect size calculation, Assessment System Corporation, St. Paul, USA, 1999.

[38] W.R. Shadish, L. Robinson, and C. Lu, Manual for ES: A computer program for effect size calculation, Assessment System Corporation, St. Paul, USA, 1999.

[39] D.I.K. Sjøberg, J.E. Hannay, O. Hansen, V.B. Kampenes, A. Karahasanovic, N.-K. Liborg, and A.C. Rekdal, A survey of controlled experiments in software engineering, IEEE Transactions on Software Engineering 31 (9) (2005) 733-753.

[40] M.L. Smith, G.V. Glass, and T.I. Miller, The Benefits of Psychotherapy, The Johns Hopkins University Press, USA, 1980.

[41] B. Thompson, "Statistical", "Practical", and "Clinical": How many kinds of significance do counselors need to consider?, Journal of Counseling & Development 80 (2002) 64-71.

[42] J. Trusty, B. Thompson, and J.V. Petrocelli, Practical guide for reporting effect size in quantitative research in the Journal of Counseling & Development, Journal of Counseling & Development 82 (2004) 107-110.

[43] T. Vacha-Haase and B. Thompson, How to estimate and interpret various effect sizes, Journal of Counseling Psychology 51 (4) (2004) 473-481.

[44] T. Vacha-Haase, J.E. Nilsson, D.R. Reetz, T.S. Lance, and B. Thompson, Reporting practices and APA editorial policies regarding statistical significance and effect size, Theory & Psychology 10 (3) (2000) 413-425.

[45] L. Wilkinson and the Task Force on Statistical Inference, Statistical methods in psychology Journals, Guidelines and explanations, American Psychologist 54 (8) (1999) 594-604.