A 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

Received 23 October 2006; received in revised form 19 January 2007; accepted 8 February 2007
Available online 17 February 2007

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 12 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 behavioral science.

© 2007 Elsevier B.V. All rights reserved.

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, 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 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 to 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).

Fig. 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.

## 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, \(y_{11}, \ldots, y_{1m}\), are normally distributed with mean \(\mu_1\) and variance \(\sigma^2\), and the observations in Group 2, \(y_{21}, \ldots, y_{2m}\), are normally distributed with mean \(\mu_2\) and variance \(\sigma^2\). More specifically:

\[Y_1 \sim N(\mu_1, \sigma^2)\]
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 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

\[ d = \frac{\mu_1 - \mu_2}{\sigma} \]

where \( \sigma \) is the pooled standard deviation, \( \sigma_p \), as the standardizer:

\[ \sigma_p = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}} \]

(1)

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

\[ d_{pop} = \frac{\mu_1 - \mu_2}{\sigma} \]

The population standardized mean difference takes positive or negative values, depending on the choice of \( \mu_1 \) and \( \mu_2 \). It is estimated by the difference between sample means (\( \bar{X}_1, \bar{X}_2 \)) 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’ \( \Delta \). Hedges’ \( g \) has the pooled standard deviation, \( s_p \), as the standardizer:

\[ \text{Hedges’ } g = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \]

(2)

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

\[ s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \]

(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 \( s_p \) Glass’ \( \Delta \) 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 \( n_1 + n_2 \to \infty \)), but Hedges’ \( g \) has the best properties for small samples when multiplied by a correction factor that adjusts for small sample bias (Formula (4)). Hence, we applied Hedges’ \( g \) as the estimator for \( d_{pop} \) in our investigation and will not consider Cohen’s \( d \) and Glass’ \( \Delta \) further.

Correction factor for Hedges’ \( g = 1 - \frac{3}{4(N-2)-1}. \)(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 Formula (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*<sub>pb-pop</sub>. When *r*<sub>pb-pop</sub> is squared, it is also called η<sup>2</sup> 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:

\[
r_{pb-pop} = \sqrt{\frac{\sigma^2_{treatment}}{\sigma^2_{total}}},
\]

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*<sub>pb-pop</sub> has the value range [0, 1]. An estimator of *r*<sub>pb-pop</sub>, 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:

\[
r_{pb} = \sqrt{\frac{SS_{treatment}}{SS_{total}}}
\]

Formulas based on *r*-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*<sub>pb</sub> values from studies with different splits in the sample size will not be directly comparable. To counteract this, the following corrected *r*<sub>pb</sub> is recommended [19]:

\[
Corrected \, r_{pb} = \frac{ar_{pb}}{\sqrt{(a^2 - 1)r^2_{pb} + 1}},
\]

where \(a = \sqrt{0.25/\pi q}\), 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*<sub>pb</sub> 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*<sub>pb</sub> [5, 35]:

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

where *N* is the total sample size. Note that the formula is simplified by the factor 1/4 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*<sub>pop</sub>, 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 non-overlap 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 Fig. 2: The unstandardized effect
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:

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

where $\mu_i$ is the mean value in population $i$, which is estimated by the mean $\bar{y}_i$. 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 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

<table>
<thead>
<tr>
<th>Effect size index</th>
<th>Effect size values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small</td>
</tr>
<tr>
<td>$d_{pop}$</td>
<td>.20</td>
</tr>
<tr>
<td>$r_{pb-pop}$</td>
<td>.10</td>
</tr>
</tbody>
</table>

![Table 2](image)

Table 2

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

![Fig. 2](image)

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

Distributional non-overlap percentages for values of $d_{pop}$ [5]

<table>
<thead>
<tr>
<th>$d_{pop}$</th>
<th>0.0</th>
<th>0.5</th>
<th>1.0</th>
<th>1.3</th>
<th>2.0</th>
<th>3.0</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree of non-overlap (%)</td>
<td>0</td>
<td>33</td>
<td>55</td>
<td>65</td>
<td>81</td>
<td>93</td>
<td>98</td>
</tr>
</tbody>
</table>

![Table 1](image)

Table 1

Distributional non-overlap percentages for values of $d_{pop}$ [5]
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”.

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 tests 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 (Fig. 3).

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,

![Fig. 3. Results of the literature review selection process.](image-url)
• when an effect size was reported, the interpretation of
the effect size and whether practical importance was dis-
cussed, and
• sample size, level of significance, p-value or information
about rejection or acceptance of the null hypothesis,
and whether the test was one or two-sided.

In addition, we registered descriptive statistics and esti-
mated 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_{\text{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 com-
pare mean values, nonparametric tests that compare med-
ian 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 condi-
tions. Hence, when the primary test was an overall compar-
ison 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 unstandard-
ized effect sizes only. Standardized and unstandardized
effect sizes were reported for, respectively, 55 and 46 of
the 429 primary tests of the reviewed experiments.

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%), only a few for nonparametric tests (6%) and not
for any tests of dichotomous variables. The corresponding
parametric tests were ANOVA and t-tests; the nonpara-
metric tests were Wilcoxon match pair tests. The standard-
ized 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 propor-
tions for parametric tests and tests of dichotomous vari-
ables (32+3 of 293 and 5 of 42, respectively, that is, 12%)
and to a lesser extent for nonparametric tests (6%) 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 (9 tests), differ-
ence in proportions or percentage (5 tests), ratio of mean
values (5 tests), difference in average rank values (3 tests)
and confidence interval for the mean difference (3 tests).

<table>
<thead>
<tr>
<th>Levels of effect size reporting</th>
<th>Experiments</th>
<th>Primary tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>Both standardized and unstandardized</td>
<td>2</td>
<td>2.2</td>
</tr>
<tr>
<td>Standardized (only)</td>
<td>8</td>
<td>8.7</td>
</tr>
<tr>
<td>Unstandardized (only)</td>
<td>17</td>
<td>18.5</td>
</tr>
<tr>
<td>No effect size</td>
<td>65</td>
<td>70.7</td>
</tr>
<tr>
<td>Total</td>
<td>92</td>
<td>100.0</td>
</tr>
</tbody>
</table>
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% versus 3.7%).

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.

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 non-parametric 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)–(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.

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 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.
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 $\chi^2$ 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).

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 and Development on generating, reporting and interpreting effect sizes for various types of statistical analysis.

<table>
<thead>
<tr>
<th>Hedges’ $g$</th>
<th>Cumulative percentages for 284 $g$ effect size estimates in software engineering experiments</th>
<th>Point-biserial correlation</th>
<th>Cumulative percentages for 284 $r_{pb}$ effect size estimates in software engineering experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>7</td>
<td>0.00</td>
<td>7</td>
</tr>
<tr>
<td>.10</td>
<td>11</td>
<td>0.10</td>
<td>19</td>
</tr>
<tr>
<td>.20</td>
<td>19</td>
<td>0.20</td>
<td>35</td>
</tr>
<tr>
<td>.30</td>
<td>28</td>
<td>0.30</td>
<td>50 median</td>
</tr>
<tr>
<td>.40</td>
<td>35</td>
<td>0.40</td>
<td>62</td>
</tr>
<tr>
<td>.50</td>
<td>42</td>
<td>0.50</td>
<td>70</td>
</tr>
<tr>
<td>.60</td>
<td>50 median</td>
<td>0.60</td>
<td>84</td>
</tr>
<tr>
<td>.70</td>
<td>56</td>
<td>0.70</td>
<td>92</td>
</tr>
<tr>
<td>.80</td>
<td>60</td>
<td>0.80</td>
<td>97</td>
</tr>
<tr>
<td>.90</td>
<td>64</td>
<td>0.90</td>
<td>100</td>
</tr>
<tr>
<td>1.00</td>
<td>68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.10</td>
<td>71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.20</td>
<td>73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.30</td>
<td>77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.40</td>
<td>83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.50</td>
<td>86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.60</td>
<td>88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.70</td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.80</td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.90</td>
<td>93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.00</td>
<td>95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.30</td>
<td>97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.50</td>
<td>97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.00</td>
<td>99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.40</td>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Size category</th>
<th>Hedges’ $g$ effect sizes</th>
<th>Median</th>
<th>Point-biserial correlation, $r_{pb}$ effect sizes</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small (lower 33%)</td>
<td>0.00–0.376</td>
<td>0.17</td>
<td>0.00–0.193</td>
<td>0.09</td>
</tr>
<tr>
<td>Medium (middle 34%)</td>
<td>0.378–1.000</td>
<td>0.60</td>
<td>0.193–0.456</td>
<td>0.30</td>
</tr>
<tr>
<td>Large (upper 33%)</td>
<td>1.002–3.40</td>
<td>1.40</td>
<td>0.456–0.868</td>
<td>0.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Type of statistical test method applied$^a$</th>
<th>ANOVA</th>
<th>$t$-test</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articles reporting controlled experiments in software engineering (This review)</td>
<td>3%</td>
<td>16%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>Articles reporting psychology studies [13]</td>
<td>32%</td>
<td>15%</td>
<td>13%</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ In our review, 116 ANOVA tests were reported in 32 articles, 118 $t$-tests were reported in 32 articles and 25 $\chi^2$ tests were reported in nine articles.
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.

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) non-significant results might erroneously be judged to be of no practical importance, and (B) statistical significance might be mistaken for practical importance; see Table 12.

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” basis 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, Dyba 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%–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
