Empir Software Eng (2018) 23:1982–2017
https://doi.org/10.1007/s10664-017-9574-5

Effect sizes and their variance for AB/BA crossover design studies

Lech Madeyski[1] · **Barbara Kitchenham**[2]

Published online: 6 December 2017

Abstract Vegas et al. IEEE Trans Softw Eng 42(2):120:135 (2016) raised concerns about the use of AB/BA crossover designs in empirical software engineering studies. This paper addresses issues related to calculating standardized effect sizes and their variances that were not addressed by the Vegas et al.'s paper. In a repeated measures design such as an AB/BA crossover design each participant uses each method. There are two major implication of this that have not been discussed in the software engineering literature. Firstly, there are potentially two different standardized mean difference effect sizes that can be calculated, depending on whether the mean difference is standardized by the pooled within groups variance or the within-participants variance. Secondly, as for any estimated parameters and also for the purposes of undertaking meta-analysis, it is necessary to calculate the variance of the standardized mean difference effect sizes (which is not the same as the variance of the study). We present the model underlying the AB/BA crossover design and provide two examples to demonstrate how to construct the two standardized mean difference effect sizes and their variances, both from standard descriptive statistics and from the outputs of statistical software. Finally, we discuss the implication of these issues for reporting and planning software engineering experiments. In particular we consider how researchers should choose between a crossover design or a between groups design.

Keywords Empirical software engineering · Crossover designs · Effect size estimation · Effect size variance estimation · Meta-analysis

Communicated by: Natalia Juristo

✉ Lech Madeyski
Lech.Madeyski@pwr.edu.pl

Barbara Kitchenham
b.a.kitchenham@keele.ac.uk

[1] Faculty of Computer Science and Management, Wroclaw University of Science and Technology, Wyb.Wyspianskiego 27, 50-370 Wroclaw, Poland http://madeyski.e-informatyka.pl/

[2] School of Computing and Mathematics, Keele University, Keele, Staffordshire ST5 5BG, UK

1 Introduction

Vegas et al. (2016) reported that many software engineering experiments had used AB/BA crossover designs but the reports of the experiments did not use the correct terminology. In their literature review, they found a total of 82 papers that reported human-participant based experiments. 33 of those papers used crossover designs in a total of 68 experiments. Only five of papers employing a crossover design used the term *crossover*, other papers used terms that were incorrect or not specific enough. Furthermore, 17 papers did not take account of participant variability in their analysis (which is the main rationale for using a repeated measures design such as a crossover).

In their paper, Vegas et al. explain both the terminology used to describe a crossover design, and how to analyze a crossover design correctly. However, except for warning readers to "Beware of effect size" and to only calculate effect sizes when the main factor is the only significant variable, Vegas et al. did not discuss effect sizes for crossover designs. In this paper we explain how to construct effect sizes and their variances for crossover designs. We provide an overview of the crossover design, as well as its advantages and limitations in Section 2.

Effect size is a name given to indicators that measure the strength of the investigated phenomenon, in other words, the magnitude of a treatment effect. Effect sizes are much less affected by sample size than statistical significance. Hence, they are better indicators of practical significance (Madeyski 2010; Urdan 2005; Stout and Ruble 1995). Effect sizes are also essential in meta-analyzes (Kitchenham and Madeyski 2016), which in turn allow us to summarize results of empirical studies, even those with contradictory results, that address the same (or closely related) research questions.

Thus the objectives of this paper are as follows:

1. To present the formulas needed to calculate both non-standardized mean difference effect sizes and standardized mean difference effect sizes[1] for AB/BA crossover designs (see Sections 4 and 5).
2. To present the formulas needed to estimate the *variances* of the non-standardized and standardized effect sizes which in the later cases need to be appropriate for the small to medium sample sizes commonly used in SE crossover designs (see Section 5).
3. To explain how to calculate the effect sizes and their variances both from the descriptive statistics that should be reported from crossover experiments and from the raw data (see Section 6).

We discuss why these goals are important and how we address them in Section 3. In Section 7, we discuss the implications of the issues presented in this paper from the viewpoint of researchers trying to decide whether to undertake a crossover study or an independent groups study, particularly in the context of families of experiments. We present our conclusions in Section 8.

It is also worth mentioning that in order to streamline the uptake of the research results of this paper, the `reproducer` `R` package (Madeyski 2017) complements this paper, as well as Kitchenham et al. (2017a), Madeyski and Jureczko (2015), Jureczko and Madeyski (2015), with the aim of making our work easier to reproduce by others. We have embedded a number of the R functions (used to make the statistical analyzes and simulations in the

[1] For simplicity, we shall refer to these simply as the *standardized effect sizes* and will not continually repeat the terms *mean difference*.

paper) in the `reproducer` `R` package we developed and made available from CRAN (the official repository of `R` packages)[2]. The use of the functions (R commands and outputs) is presented throughout the paper (see Outputs 3 and 4), as well as Appendix A (see Outputs Outputs 5, and 6).

2 Background

A crossover design is a form of repeated measures design. A repeated measures design is one where an individual participant contributes more than a single outcome value.

In the case of a simple AB/BA crossover design, *A* refers to one software engineering *technique*, *B* refers to another and the goal of the design is to determine which technique delivers the better outcome. The difference between the outcomes of using technique A and using technique B is called the *technique effect*. Participants are split into two groups, and each participant in one group uses technique A first (on a software engineering task with materials related to a specific software system or component) and subsequently uses technique B to perform the same task using materials related to a different software system or component. Participants in the other group use technique B first, and technique A second. The group that a participant is assigned to determines the *sequence* in which a participant uses the techniques. The first set of outcomes are referred to as the *Period 1* outcomes, the second set of outcomes are referred to as the *Period 2* outcomes. The difference between the outcomes in *Period 1* and the outcomes in *Period 2* is called the *period effect*. A full mathematical definition of the crossover model[3] is shown in Table 1 and explained in Section 4.

The benefit of the crossover design (and other repeated measures designs) is that each individual acts as his/her own control. The impact of this is that if the resulting data are correctly analyzed:

1. The effect of individual differences related to innate ability is removed, (i.e., systematic between participant variation is removed). Thus, the effect of different techniques are assessed in terms of the individual improvement for each participant.
2. The removal of between participant variation, allows tests of significance to be based on smaller variances (i.e., significance tests are based on the within-participant variation).
3. Since the variance used to test the technique difference is reduced, it is possible either to reduce sample sizes and maintain statistical power, or to maintain sample sizes and increase power.

Since sample sizes are often relatively low in Software Engineering (SE) experiments, crossover designs have the potential to be very useful. There are obviously disadvantages as well. The correct analysis of crossover data is more complicated than analysis of data from simple experiments where participants are randomly allocated to two different treatment groups[4].

[2]Our package should not be confused with the `knitr` package we used to embed R code chunks in the paper.

[3]For readability, we sometimes omit the term *AB/BA* when referring to the crossover design, but any reference to a crossover design or model in this paper, refers to an AB/BA crossover, which is based on two techniques and two time periods.

[4]This is referred to as a between groups design or an independent groups design. We prefer the term independent groups in this paper to contrast with repeated measures designs

Table 1 Expected Outcome for Participants in AB/BA Crossover

Sequence Group	Participant ID	Period 1	Period 2
SG_1	j	$y_{1,1,j} = \mu_j + \tau_A$ (technique A)	$y_{2,2,j} = \mu_j + \pi + \tau_B + \lambda_A$ (technique B)
SG_2	k	$y_{2,1,k} = \mu_k + \tau_B$ (technique B)	$y_{1,2,k} = \mu_k + \tau_A + \pi + \lambda_B$ (technique A)

Perhaps more importantly, although the crossover design can cope with time period effects that are consistent across all the participants, crossover designs are vulnerable to interaction effects including period by technique interaction, where the performance of participants is affected by which technique they used first. For example, if a technique involves providing additional materials to participants, it may be easier first to understand the task using less (or simpler) documentation, and then perform the subsequent task with the additional (more complex) information, rather than try to perform the first task with too much information. The crossover design is also vulnerable to participant by technique interaction where individual participants behave differently depending on which technique they used. For example, one technique might improve the performance of less able participants but have no effect on more able participants, which would reduce the repeated measures correlation. If researchers expect either of these conditions to hold, they should avoid using a crossover design.

3 Goals and Methodology

Our first goal is to present the formulas needed to estimate the effect sizes used in crossover designs. This goal is important because researchers in all empirical disciplines are increasingly being encouraged to adopt the use of effect sizes rather than just report the results of t or F tests (see APA 2010; Kampenes et al. 2007; Cumming and Finch 2001; Cumming 2012).

To address this goal, we begin by presenting a detailed discussion of the AB/BA crossover model in Section 4, from which the means and variances needed to calculate both standardized and non-standardized effect sizes are derived.

In Section 5, we specify two different standardized effect sizes suitable for crossover designs depending on whether researchers are interested only in the personal improvement offered by a software engineering technique, or are more interested in the effect of the technique, and want an effect size comparable to that of a standard independent groups design.

Our second (but equally important) goal is to present formulas needed to calculate the variance of both non-standardized and standardized effect sizes. This goal is important because without knowing the variance of effect sizes, it is impossible to derive their confidence intervals (CIs). Researchers are advised to report CIs (see APA 2010; Cumming and Finch 2001) because they provide a direct link to null hypothesis testing and support meta-analysis.

To obtain the variances of the two standardized effect sizes, we reviewed the literature and found one paper that proposed formulas for the standardized effect size variances (see

Curtin et al. 2002). This paper proposed a formula suitable for small sample sizes and a simpler approximate formula suitable for larger sample sizes. However, we could not verify the proposed formulas. For this reason, we derived our equations from first principles based on the non-central t distribution (Johnson and Welch 1940), as explained in Section 5.3.1. After discussions with Dr. Curtin, we have, together, agreed revised versions of his equations (see Kitchenham et al. 2017b).

Our third goal is to explain how to calculate the standardized and non-standardized effect sizes and their variances both from the descriptive statistics that should be reported from crossover experiments and from the raw data. To address this goal, we present two examples in Section 6. This goal is important because researchers need to understand how the outcome from statistical analysis tools map to the parameters of the crossover model. Therefore, we include in Section 6.3 an explanation of how standardized effect sizes and their variances can be calculated from analyzes undertaken using `R` (R Core Team 2016) with the linear mixed model `lme4` package (Bates et al. 2015). In addition, researchers who replicate crossover studies and want to aggregate their results with previous studies may not have access to the raw data from previous studies. Therefore, they may need to estimate effect sizes and their estimates from descriptive data. Furthermore, if appropriate descriptive statistics are reported in studies using a crossover design, even if the studies used an inappropriate analysis, the results could easily be reassessed, if researchers know how the descriptive statistics map to the parameters of the crossover model.

4 The Non-Standardized Effect Sizes for Crossover Studies and their Variances

This section explains the AB/BA crossover model and how to calculate the non-standardized effect sizes and their variances.

4.1 Non-Standardized Effect Sizes of the AB/BA Crossover Model

Senn (2002) provides an extensive discussion of the AB/BA crossover design and we follow his analysis procedures throughout this section. Following his approach, the most straightforward way to represent the design is to model the outcomes for individuals in each sequence. If we assume:

- τ_A is the effect of technique A.
- τ_B is the effect of technique B.
- $\tau_{AB} = \tau_A - \tau_B$ is the difference between the effect of technique A and technique B. It is the non-standardized mean technique effect size.
- τ_{BA} is the difference between the effect of technique B and technique A where $\tau_{BA} = -\tau_{AB}$.
- π is the period effect size which is the difference between the outcome of using a technique in the first time period and the second time period.
- λ_A is the period by technique interaction due to using technique B after using technique A[5].

[5]In medical experiments, the period by technique interaction term is often referred to as *carry-over*. This is because crossover designs are often used for testing drugs and the effect of the first drug taken may interact with the second drug in an adverse way. Medical experiments therefore leave an appropriate *washout period*

Table 2 Expected Differences and Sums for Participants in an AB/BA Crossover

Sequence Group	Participant	Cross-over Difference	Period Difference	Participant Total
SG_1	j	$\tau_{AB} - \pi - \lambda_A$	$\pi + \lambda_A - \tau_{AB}$	$2\mu_j + \pi + \tau_A + \tau_B + \lambda_A$
SG_2	k	$\tau_{AB} + \pi + \lambda_B$	$\tau_{AB} + \pi + \lambda_B$	$2\mu_k + \pi + \tau_A + \tau_B + \lambda_B$

- λ_B is the period by technique interaction due to using technique A after technique B.
- $\lambda_{AB} = \lambda_A - \lambda_B = -\lambda_{BA}$ is the mean period by technique interaction effect size.
- μ_i is the average outcome for participant i.
- The group of participants that use technique A first is called *sequence* group SG_1, the group of participants that use technique B first are called sequence group SG_2.

The *expected* outcome in each period for a typical participant in each sequence group is shown in Table 1[6]. The observations in each cell (i.e., period and technique combination) referred to as $y_{t,p,s}$ are identified by the technique (where $t = 1$ equates to technique A, and $t = 2$ equates to technique B), the period (where $p = 1$ equates to time period 1, and $p = 2$ equates to time period 2), and participant (where $s = 1, ..., n_1$ corresponds to the participants in the group that used technique A in time period 1, i.e., group SG_1, and $s = 1, ..., n_2$ corresponds to the participants in the group that used technique B in time period 1, i.e., group SG_2)[7].

Senn (2002) demonstrates how a crossover analysis is based on summing and differencing outcomes for each participant as shown in Table 2.

The *crossover difference* for each participant in Table 2 is obtained by subtracting the outcome obtained using technique A from the outcome obtained using technique B in each time period. Thus, the crossover difference for participants in group SG_1 is:

$$CODiff_{1,j} = y_{1,1,j} - y_{2,2,j} \tag{1}$$

and the expected value for each participant is:

$$CODiff_{1,j} = \tau_A - \tau_B - \pi - \lambda_A = \tau_{AB} - \pi - \lambda_A \tag{2}$$

The crossover difference for participants in group SG_2 is:

$$CODiff_{2,k} = y_{1,2,k} - y_{2,1,k} \tag{3}$$

and the expected value for each participant is:

$$CODiff_{2,k} = \tau_A - \tau_B + \pi + \lambda_B = \tau_{AB} + \pi + \lambda_B \tag{4}$$

Calculating the crossover difference means the effect of the individual participant is removed.

to allow the effect of the first drug to be eliminated from participants before they are given a second drug. We use the term *period by technique interaction* because carry-over and a washout period are not really appropriate concepts for SE experiments. In fact, in the context of training, it might be argued that we want to encourage 'carry-over' of acquired skills and minimize their 'washout'.

[6]By expected outcome, we mean the outcome based on the model excluding any error term. We explain error terms and variances in Section 4.2.

[7]Table 1 is equivalent to TABLE 2 in Vegas et al. (2016), except we also specify the model of the data obtained from individual participants.

The *period difference* for each participant is obtained by *subtracting* the task outcome for period two from the task outcome for period one, as shown in Table 2. Thus, the period effect for participants in group SG_1 is:

$$PDiff_{1,j} = y_{2,2,j} - y_{1,1,j} \tag{5}$$

and the expected value for each participant is:

$$PDiff_{1,j} = \pi + \tau_B - \tau_A + \lambda_A = \pi - \tau_{AB} + \lambda_A \tag{6}$$

The period effect for participants in group SG_2 is

$$PDiff_{2,k} = y_{1,2,k} - y_{2,1,k} \tag{7}$$

and the expected effect for each participant is:

$$PDiff_{2,k} = \pi + \tau_A - \tau_B + \lambda_B = \pi + \tau_{AB} + \lambda_B \tag{8}$$

Again, calculating the period difference means that the effect of the individual participant is removed.

The *participant total* for each participant is obtained by *adding* the task outcome for period two to the task outcome for period one, as shown in Table 2. Thus, the participant total for a participant in group SG_1 is:

$$SG_{1,j} = y_{1,1,j} + y_{2,2,j} \tag{9}$$

and the expected value for each participant is:

$$SG_{1,j} = 2\mu_j + \pi + \tau_A + \tau_B + \lambda_A \tag{10}$$

The participant total for a participant in group SG_2 is

$$SG_{2,k} = y_{2,1,k} + y_{1,2,k} \tag{11}$$

and the expected value for each participants is:

$$SG_{2,k} = 2\mu_k + \pi + \tau_A + \tau_B + \lambda_B \tag{12}$$

It is important to note that the participant total includes the mean task outcome of the individual participant.

In order to estimate the model parameters, we average the sum of the crossover differences, the sum of the period differences and the sum of the participant totals over the participants in the same group. The expected value for groups are shown in Table 3. To emphasize that Table 3 provide estimates of the model parameters, each parameter is shown with a *hat* symbol over its Greek character. It is important to note that averaging the participant totals leads to replacing the individual participant outcomes with the average participant outcome.

Table 3 Expected value of groups means for the crossover design

Sequence Group	Mean crossover Difference	Mean period Difference	Mean participant Total
SG_1	$\hat{\tau}_{AB} - \hat{\pi} - \hat{\lambda}_A$	$\hat{\pi} + \hat{\lambda}_A - \hat{\tau}_{AB}$	$2\hat{\mu} + \hat{\tau}_{AB} + \hat{\pi} + \hat{\lambda}_A$
SG_2	$\hat{\tau}_{AB} + \hat{\pi} + \hat{\lambda}_B$	$\hat{\tau}_{AB} + \hat{\pi} + \hat{\lambda}_B$	$2\hat{\mu} + \hat{\tau}_{AB} + \hat{\pi} + \hat{\lambda}_B$

The mean crossover difference for SG_1, MCO_1, is obtained by averaging the crossover difference of the n_1 participants in SG_1:

$$MCO_1 = \frac{\Sigma_j CODiff_{1,j}}{n_1} = \hat{\tau}_{AB} - \hat{\pi} - \hat{\lambda}_A \tag{13}$$

The mean crossover difference for SG_2, MCO_2, is obtained by averaging the crossover difference of the n_2 participants in SG_2:

$$MCO_2 = \frac{\Sigma_k CODiff_{1,k}}{n_2} = \hat{\tau}_{AB} + \hat{\pi} + \hat{\lambda}_B \tag{14}$$

This means that:

$$MCO_1 + MCO_2 = 2\hat{\tau}_{AB} - (\hat{\lambda}_A - \hat{\lambda}_B) \tag{15}$$

and

$$\frac{MCO_1 + MCO_2}{2} = \hat{\tau}_{AB} - \frac{(\hat{\lambda}_A - \hat{\lambda}_B)}{2} = \hat{\tau}_{AB} - \frac{\hat{\lambda}_{AB}}{2} \tag{16}$$

A critical assumption underlying a crossover design is that:

$$\lambda_{AB} = 0 \tag{17}$$

so, if the assumption holds, the average of the mean crossover difference estimates the non-standardized technique effect size, $\hat{\tau}_{AB}$, for a crossover design:

$$\hat{\tau}_{AB} = \frac{MCO_1 + MCO_2}{2} \tag{18}$$

Thus, we assume that any effect caused by undertaking one technique followed by another is fully modeled by the period effect. We consider this issue further in Section 4.2.2.

The period effect can be calculated as:

$$\frac{-(MCO_1 - MCO_2)}{2} = \hat{\pi} + \frac{(\hat{\lambda}_A + \hat{\lambda}_B)}{2} \tag{19}$$

Assuming that $\hat{\lambda}_A = \hat{\lambda}_B = 0$, minus the average of the difference of the mean crossover differences estimates the period effect:

$$\hat{\pi} = \frac{-(MCO_1 - MCO_2)}{2} \tag{20}$$

Similar equations can be used to calculate the technique effect and the period effect using the mean period differences.

If the assumption that the period by technique interaction term is zero is true, then it will not be significantly different from zero. Nonetheless, to estimate the period by technique interaction term, we use the mean of the participant totals for sequence SG_1 and sequence SG_2 where:

$$MSG_1 = \frac{\Sigma_j ST_{1,j}}{n_1} = 2\hat{\mu} + \hat{\tau}_A + \hat{\tau}_B + \hat{\pi} + \hat{\lambda}_A \tag{21}$$

and

$$MSG_2 = \frac{\Sigma_k ST_{2,k}}{n_2} = 2\hat{\mu} + \hat{\tau}_A + \hat{\tau}_B + \hat{\pi} + \hat{\lambda}_B \tag{22}$$

Thus the difference between the mean participant totals of the two sequence groups estimates the period by technique interaction effect size:

$$MSG_1 - MSG_2 = \hat{\lambda}_A - \hat{\lambda}_B = \hat{\lambda}_{AB} \tag{23}$$

This means that the period by technique interaction effect can also be called the *sequence* effect.

4.2 Non-Standardized Effect Size Variances and *t*-tests

In this section we explain how to calculate the variance of the non-standardized effect sizes, and how these statistics relate to the *t*-test of the non-standardized effect size. The relationship between effect sizes and *t*-variables is also important for estimating the references of standardized effect sizes (see Section 5.3.1). We also discuss the problems introduced both by tests of the period by technique interaction effect, and by non-normally distributed data.

4.2.1 The technique effect size variance

In order to identify the variance of the estimated effects, we need to consider the error term in crossover designs. Senn (2002) points out that the error term associated with the outcome of a specific individual using a specific technique in a specific period is made up of two parts:

- $\beta_{s,i}$ which is the effect due to participant i in sequence group s where $s = SG_1$ or $s = SG_2$.
- $\zeta_{i,s,t}$ which is the within participant error.

The expected value of $\beta_{s,i}$ is zero and the variance of $\beta_{s,i}$ is σ_b^2. The expected value of $\zeta_{i,s,t}$ is zero and the variance of $\zeta_{i,s,t}$ is $\sigma^2 - w$. The simplest model assumes that $\beta_{s,i}$ and $\zeta_{i,s,t}$ are independent, so their covariance is zero, although Senn points out that other models are possible. The simplest model also assumes that all $\zeta_{i,s,t}$ are independent of each other.

If we calculate the pooled within period and within technique variance in a crossover study, we obtain a variance, that is an estimate of the sum of the between-participant variance and the within-participant variance. So if:

$$\sigma_{IG}^2 = \sigma_b^2 + \sigma_w^2 \tag{24}$$

We can estimate σ_{IG}^2 as follows:

$$s_{IG}^2 = \frac{\Sigma_{t,p}(n_t - 1)\left(y_{t,p,j} - \hat{y}_{t,p}\right)^2}{2n_1 + 2n_2 - 4} \tag{25}$$

where n_t equal n_1 for sequence group SG_1 and n_2 for sequence group SG_2. This calculation is exactly the same variance calculation we would use if we were analyzing a study based on four independent groups. For this reason, in the context of repeated measures analysis, it is labeled σ_{IG}^2 and its estimate is labeled s_{IG}^2, see, for example, Morris and DeShon (2002).

It is important to note that s_{IG}^2 *should never* be used as the basis for the standard error in a *t*-test because the repeated measures violate the assumption that all the individual values are independent.

In a simple independent groups study we are unable to separate the two components of σ_{IG}^2. In contrast, with a repeated measures design such as an A/B crossover we are able to estimate the separate components of variance. However, in order to estimate the variance components we need to consider the variance of the crossover difference scores, σ_{diff}^2.

Unlike the error term associated with an individual measurement, the error term associated with the crossover difference (or period difference), removes the participant effect and leaves only the within-participant variation. In simple before-after repeated measures

designs[8], the differences between before and after outcomes lead to a single group of differences scores, and the variance of the difference scores is an unbiased estimate of the within-participant variance (see, for example, Cumming 2012). However, the added complication of the crossover design means that the variance we obtain from the difference values is the pooled within sequence group variance (again assuming that the variance of the difference values in each sequence groups estimate the same underlying variance). Thus estimate of the difference score variance is calculated as:

$$s^2_{diff} = \frac{(n_1 - 1)\Sigma_j(CODiff_j - MCO_1)^2 + (n_2 - 1)\Sigma_k(CODiff_k - MCO_2)^2}{(n_1 + n_2 - 2)} \quad (26)$$

In simple repeated measures before-after design σ^2_{diff} and σ^2_w are equal, however, Freeman (1989) points out that in crossover designs:

$$\sigma^2_w = \frac{\sigma^2_{diff}}{2} \quad (27)$$

Furthermore, the correlation between the outcomes for an individual in both periods is:

$$\rho = \frac{\left(\sigma^2_{IG} - \frac{\sigma^2_{diff}}{2}\right)}{\sigma^2_{IG}} = \frac{\left(\sigma^2_{IG} - \sigma^2_w\right)}{\sigma^2_{IG}} \quad (28)$$

so

$$\sigma^2_{diff} = 2\sigma^2_{IG}(1 - \rho) \quad (29)$$

and

$$\sigma^2_w = \sigma^2_{IG}(1 - \rho) \quad (30)$$

From (18), and the fact that the variance of the mean difference in each sequence group is assumed to be the same, $var(MCO_i) = \frac{s^2_{diff}}{n_i}$ and we can calculate the variance of $\hat{\tau}$[9] since:

$$var(\hat{\tau}) = \frac{var(MCO_1) + var(MCO_2)}{4} = \frac{s^2_{diff}}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \quad (31)$$

Since, $\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{(n_1+n_2)}{n_1 n_2}$, the square root of the variance of $\hat{\tau}$ which is also called the standard error of $\hat{\tau}$ is:

$$se_{\hat{\tau}} = \frac{s_{diff}}{2}\sqrt{\frac{(n_1 + n_2)}{n_1 n_2}} = \frac{s_w}{\sqrt{2}}\sqrt{\frac{(n_1 + n_2)}{n_1 n_2}} = s_w\sqrt{\frac{(n_1 + n_2)}{2n_1 n_2}} \quad (32)$$

Thus, the non-standardized technique effect size for a crossover design is obtained from (18), while its variance is obtained from (31).

Then, the t-test for the significance of $\hat{\tau}$ is:

$$t = \frac{\hat{\tau}}{se_{\hat{\tau}}} \quad (33)$$

with degrees of freedom $df = n_1 + n_2 - 2$.

[8] In other disciplines, these are also referred to as pretest-posttest designs.

[9] The variance of $\hat{\tau}_{AB}$ is exactly the same as the variance of $\hat{\tau}_{BA}$, so for variances and standard deviations we refer to $\hat{\tau}$ without any additional subscript.

To see whether the period effect is significant, the t-test is based on the same standard error:

$$t = \frac{\hat{\pi}}{se_{\hat{\tau}}} \tag{34}$$

with $df = n_1 + n_2 - 2$.

4.2.2 The period by technique interaction effect

We have not yet considered what to do about the period by technique interaction effect. One approach is to test the interaction term for statistical significance. A t-test for the interaction is based on the variance of the sums for each individual (σ^2_{sum}) and is estimated from pooled variance of the individual totals within each sequence group (s^2_{sum}):

$$s^2_{sum} = \frac{\left((n_1 - 1)\Sigma_j (ST_{1,j} - MST_1)^2 + (n_2 - 1)\Sigma_k (ST_{2,k} - MST_2)^2\right)}{n_1 + n_2 - 2} \tag{35}$$

Referring the components of (35) shown in Tables 2 and 3, in terms of the variances we have already introduced:

$$\sigma^2_{sum} = 4\sigma^2_{IG} \tag{36}$$

However, although relationship between the parameters is exact, it may not be an exact relationship between the estimates s^2_{sum} and s^2_{IG} because the variances are estimated in different ways. Using s^2_{sum}, the t-test is:

$$t = \frac{\hat{\lambda}_{AB}}{2\sqrt{s^2_{sum}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{37}$$

with $df = n_1 + n_2 - 2$. Since the power of this test is usually low[10], an alpha level of 0.1 is usually adopted (see Senn 2002, Section 3.1.4). However, if the crossover design has been used to reduce sample size, even an alpha level of 0.1 may be insufficient to detect a genuine period by technique interaction.

If we find a statistically significant period by technique interaction, it might seem tempting to use (16) to calculate the non-standardized effect size by removing the estimate of $0.5\hat{\lambda}_{AB}$ from the mean difference. This appears to be mathematically sound, but it is not statistically sound. The reason is that the variance of $0.5\hat{\lambda}_{AB}$ is $0.25s^2_{sum} \approx s^2_{IG}$. If the crossover design is used in order to reduce the variance for statistical tests, adjusting the estimate of τ by half the estimate of λ_{AB} reintroduces the between participant variance into any statistical tests of the revised estimate. This negates any possible benefit of a crossover design compared with a standard between groups design.

The practical implication of these considerations is that a crossover design should not be used if a significant period by technique interaction is anticipated. Furthermore, if a period by technique interaction is not expected, there is no point testing for one[11]. Thus, we do not include the period by technique interaction term (which corresponds to the sequence order) in our data analyzes. However, as Vegas et al. point out, the possibility of an interaction remains a threat to the validity of the experiment. We return to the issue of what can be done to address the interaction problem in Section 7.

[10] Power is the probability of rejecting the null hypothesis when the null hypothesis is false.

[11] See also the discussion in Senn (2002), Section 3.1.4 that presents the arguments against a two-stage analysis, where analysts first check for a significant period by treatment interaction. Then, if there is one, they analyze only the data from the first period, and if there is not they perform a standard crossover analysis.

4.2.3 Handling non-stable variances and non-normal data

Equation (24) for σ^2_{IG} assumes that the within-subjects (participants) variance and the between-subjects (participants) variance are independent and not affected by the different techniques. This is not necessarily the case. For example, a new technique might improve the capability of less able participants thus reducing the difference among individuals. Alternatively, a new technique might be more difficult to apply than other techniques and might improve the performance of the most able participants and reduce the performance of the less able participants making the difference between individuals greater.

If the stability of the variances is in question, there are at least three possible approaches to consider:

- The least useful option is to base the estimate of s^2_{IG} solely on the n_1 participants in the first period control condition. This is not really useful because for crossover designs n_1 is likely to be relatively small, so the estimate is likely to be inaccurate.
- Estimate s^2_{IG} and s^2_{diff} allowing the within cells variance to be different. However, the implications of this approach, such as the relationship between s^2_{diff}, s^2_{IG} and $\hat{\rho}$ are not clear.
- Use a robust, ranked-based analysis. This is the most straightforward option and also protects against non-normal data, such as skewed data and/or data with outliers.

A robust analysis compares the *period differences*. This is because the expected value of the period differences are $\pi - \tau$ for sequence group 1 and $\pi + \tau$ for sequence group 2. Thus any significant difference between the period differences in the two sequence groups is due to a significant technique effect.

Thus, using a rank-based analysis, if the rank sum of the period differences in sequence 1 is significantly different from the rank sum of the period differences in sequence 2, you can reject the hypothesis that the technique effect size is zero (see Senn 2002, Section 4.3.9). However, if you use the Wilcoxon-Mann-Whitney test it is essential to use the *exact* test, which is the default in `R`. The probability of superiority[12] can be used as a non-parametric effect size constructed from the Mann-Whitney U statistic (see Wilcox 2012; Kitchenham et al. 2017a).

5 Standardized Effect Sizes for Crossover Studies and their Variances

In this section we discuss standardized effect sizes that can be calculated for crossover designs and their variances.

5.1 Formulas for the Standardized Effect Sizes

For purposes of meta-analysis, it is important that standardized effect sizes from crossover designs are comparable with effect sizes obtained from other designs.

[12] Also known as Varga and Delaney's $\hat{A}$ metric (see Vargha and Delaney 2000; Arcuri and Briand 2014; Madeyski et al. 2014)

A crossover standardized effect size comparable to before-after repeated measures designs is:

$$\delta_{RM} = \frac{\tau}{\sigma_w} \tag{38}$$

In contrast, a crossover standardized effect size comparable to independent groups designs is:

$$\delta_{IG} = \frac{\tau}{\sigma_{IG}} \tag{39}$$

The estimates of δ_{RM} and δ_{IG} which we refer to as d_{RM} and d_{IG} are obtained by substituting the sample estimates $\hat{\tau}$ for τ, s_w for σ_w and s_{IG} for σ_{IG}. These are similar to Cohen's d, although originally Cohen's d was developed for independent groups studies and used a variance based only data from the control group.

The relationship between σ_w^2 and σ_{IG}^2 in (30) means that there is a functional relationship between the two standardized effect sizes, such that:

$$\delta_{IG} = \delta_{RM}\sqrt{(1-\rho)} \tag{40}$$

This relationship is important for calculating the variance of δ_{IG} which we discuss later.

However, the estimates d_{IG} and d_{RM} are known to be biased for small to medium sample sizes and are usually adjusted to remove bias (see Hedges and Olkin 1985; Borenstein et al. 2009; Ciolkowski 1999; Laitenberger et al. 2001). The adjustment factor is:

$$c(m) = \sqrt{\frac{2}{m}}\left(\frac{\Gamma[m/2]}{\Gamma[(m-1)/2]}\right) \tag{41}$$

where Γ is the gamma function, which is an extension of the factorial function, and m is the number of degrees of freedom, i.e., $m = n_1 + n_2 - 2$. This function is approximated by the function:

$$c(m) \approx 1 - \frac{3}{4m-1} \tag{42}$$

Hedges and Olkin (1985) reported the exact values of $c(m)$ for values from $m = 2$ to $m = 50$, but even with $m = 2$, the difference between the exact value (i.e., 0.5642) and the approximate value (i.e., 0.5714) is only 1.28%, while for $m = 10$ the difference is less that 0.04%.

Thus, the unbiased estimate of δ_{RM} is:

$$g_{RM} = c(df)d_{RM} = \frac{c(df)\hat{\tau}}{s_w} \tag{43}$$

Since, $s_w = s_{IG}\sqrt{1-\hat{\rho}}$, this is the same formula reported by Laitenberger et al. (2001).

The unbiased estimate δ_{IG} is:

$$g_{IG} = c(df)d_{IG} = \frac{c(df)\hat{\tau}}{s_{IG}} \tag{44}$$

The statistics g_{RM} and g_{IG} are often referred to as Hedges g statistics.[13]

[13] It should be noted that Hedges and Olkin (1985) refer to the small sample size adjusted estimate as d and the unadjusted estimate as g.

5.2 Choosing the Appropriate Standardized Effect Size

In the past, researchers have proposed standardizing repeated measures studies using the independent groups variance (see Becker 1988; Dunlap et al. 1996; Borenstein et al. 2009). The reason for this is to make the results of repeated measures studies comparable with the results of independent group studies. This is particularly important in the context of meta-analysis.

Repeated measures designs are intended to remove the potentially large variation between participants and test the difference between techniques based on the potentially much smaller within-subject (participant) variation. However, this means that independent groups experiments standardized by s^2_{IG} would have a smaller effect size than the repeated measures experiments standardized by s^2_w *even if the non-standardized mean differences were the same*.

Morris and DeShon (2002), however, make the point that the choice of effect size should depend on the goal of the meta-analysis. If the goal is to assess the likely improvement in individual performance then δ_{RM} is appropriate. If the goal is to assess the difference between techniques then δ_{IG} is likely to be more appropriate. Nonetheless, whichever goal a meta-analyst has, it should be clearly stated and the method for calculating the appropriate variance explained. The need for both effect sizes is also supported by Lakens (2013).

It should be noted that none of the above sources discuss effect sizes in the context of AB/BA crossover designs. Dunlap et al. (1996), Becker (1988) and Lakens (2013) were concerned solely with within-subjects before-after experiments. Morris and DeShon (2002) discuss effect sizes of independent groups and two repeated measures designs: the before-after design and the independent groups before-after design, which measures all participants using the same technique prior to splitting the participants into two groups and performing an independent groups experiment.

5.3 Standardized Effect Size Variances

For standardized effect sizes to be useful we need to calculate their variances. However, with the exception of Kitchenham and Madeyski (2016), we are not aware of any software engineering studies that have identified the need to estimate the variance of standardized mean different effect sizes. In this section, we provide formulas to estimate the variance of δ_{IG} and δ_{RM} for small, moderate and large sample sizes.

5.3.1 The basic principle

The variance estimate most suitable for small samples (up to $\approx$ 30 participants) for any standardized mean difference effect size is derived from the distribution of Student's t (see Morris and DeShon 2002; Morris 2000). The distribution of a t-variable with mean θ and variance $V(\theta)$ is known to be the non-central t distribution. Johnson and Welch (1940) report the variance of a t variable to be:

$$V(\theta) = \frac{df}{df - 2}\left(1 + \theta^2\right) - \frac{\theta^2}{[c(df)]^2} \tag{45}$$

Where θ is estimated by the t-value, $df = (n_1 + n_2 - 2)$ are the degrees of freedom associated with the t-test, and $c(df)$ is the same function reported in (41) which is approximated by the formula given in (42).

If we can estimate the variance of a variable θ, and the relationship between θ and a standardized effect size δ is given by the equation:

$$\theta = A \times \delta \tag{46}$$

where A is a constant term, then[14], the variance of δ is:

$$var(\delta) = \frac{1}{A^2} var(\theta). \tag{47}$$

which expands to:

$$var(\delta) = \frac{df}{df-2}\left(\frac{n_1+n_2}{2n_1n_2} + \delta^2\right) - \frac{\delta^2}{[c(df)]^2} \tag{48}$$

This is true for *any* standardized effect size that can be calculated from a t-value, including those obtained from crossover designs, repeated measures before-after designs, and independent group designs.

Since g_{RM} is an unbiased estimate of δ, Kitchenham et al. (2017b) show that this leads to:

$$var(d_{RM}) = \frac{df}{df-2}\left(\frac{n_1+n_2}{2n_1n_2} + g_{RM}^2\right) - \frac{g_{RM}^2}{[c(df)]^2} \tag{49}$$

and

$$var(g_{RM}) = [c(df)]^2 \frac{df}{df-2}\left(\frac{(n_1+n_2)}{2n_1n_2} + g_{RM}^2\right) - g_{RM}^2 \tag{50}$$

Since $d_{RM} = \frac{d_{IG}}{sqrt(1-\hat{\rho})}$

$$var(d_{IG}) = \frac{df}{df-2}\left(\frac{(1-\hat{\rho})(n_1+n_2)}{2n_1n_2} + g_{IG}^2\right) - \frac{g_{IG}^2}{[c(df)]^2} \tag{51}$$

and

$$var(g_{IG}) = [c(df)]^2 \frac{df}{df-2}\left(\frac{(1-\hat{\rho})(n_1+n_2)}{2n_1n_2} + g_{IG}^2\right) - g_{IG}^2 \tag{52}$$

It is important to appreciate that the value of the constant A defined in (46) depends on study design type. For crossover designs $A = \sqrt{\frac{2n_1n_2}{(n_1+n_2)}}$. However, for repeated measures before-after designs $A = \sqrt{n}$, while for independent groups designs $A = \sqrt{\frac{n_1n_2}{(n_1+n_2)}}$. Thus, the construction of mean difference effect sizes and their variances depends on the specific design type.

5.3.2 Formulas to estimate the medium sample size variance of standardized effect sizes

For larger samples sizes, approximate equations for the variance of effect sizes are available. Based on an equation presented by Hedges and Olkin (1985), Kitchenham et al. (2017b) show that:

$$var(d_{RM})_{approx} = \frac{(n_1+n_2)}{2n_1n_2} + \frac{d_{RM}^2}{2(n_1+n_2-3.94)} \tag{53}$$

[14]Since the variance s^2 of a variable x multiplied by a constant b is $var(b \times x) = b^2 s^2$

Hedges and Olkin (1985) recommend a slightly different equation for the approximate variance of g_{RM}:

$$var(g_{RM})_{approx} = \frac{[c(df)]^2(n_1+n_2)}{2n_1n_2} + \frac{g_{RM}^2}{2(n_1+n_2)} \tag{54}$$

Based on the relationship between d_{RM} and d_{IG}:

$$var(d_{IG})_{approx} = \frac{(1-\hat{\rho})(n_1+n_2)}{2n_1n_2} + \frac{d_{IG}^2}{2(n_1+n_2-3.94)} \tag{55}$$

and

$$var(g_{IG})_{approx} = \frac{[c(df)]^2(1-\hat{\rho})(n_1+n_2)}{2n_1n_2} + \frac{g_{IG}^2}{2(n_1+n_2)} \tag{56}$$

5.3.3 The approximate variance for large sample sizes

Looking at (49), we can see that if the effect size is close to zero making $d_{RM}^2 \approx 0$, and the sample size is very large, so that $df \approx df - 2$ and $c(m) \approx 1$, then:

$$var(d_{RM}) \approx \frac{(n_1+n_2)}{2n_1n_2}$$

Furthermore, if $n_1 = n_2$, the variance is approximately half the inverse of the sample size. As would be expected, under the same conditions $var(g_{RM})$ converges on the same value. In addition, the variances of d_{IG} and g_{IG} also converge on the same value:

$$var(d_{IG}) \approx \frac{(1-\hat{\rho})(n_1+n_2)}{2n_1n_2}$$

6 Calculating Effect Sizes and their Variances

In this section, we present two small examples illustrating how to calculate crossover study effect sizes and their variances. One example is based on real software engineering data to illustrate the complexity of software engineering data. The other is based on simulated data to illustrate how the AB/BA crossover model is intended to work given that all the basic assumptions underlying the model are true.

It is useful to know how to calculate effect sizes (both non-standardized and standardized) and their variances both from descriptive data, as well as by using statistical packages to analyze the raw data. If authors report appropriate descriptive statistics, then other researchers (including reviewers and meta-analysts) can, without access to the raw data, construct the model parameters, effect sizes and their variances from the results reported. Therefore, we identify the descriptive statistics that are necessary to calculate the various crossover model parameters and effect sizes. In addition, we show how to analyze raw crossover data using the `R` language and the `lme4` package, and explain how to extract the crossover model parameters from the outcomes of the R package.

We, also, demonstrate two graphical methods of presenting the results of crossover studies. We suggest that they provide a more accurate graphical representation than box plots of the technique outcomes. In particular, they provide visual indications both of the outcomes of the experiment, and of the extent to which data conforms with the crossover model.

Table 4 Scanniello crossover data (labeled as `EUBASwideSelected` in Output 1)

ID	Comp_Level.AM	Comp_Level.SC	Comp_Diff	Comp_Sum	SequenceGroup
P3	0.82	0.77	0.05	1.59	SG1
P4	0.60	0.70	−0.10	1.30	SG2
P7	0.80	0.93	−0.13	1.73	SG1
P8	0.93	0.90	0.03	1.83	SG2
P11	0.70	0.83	−0.13	1.53	SG1
P12	0.90	0.96	−0.06	1.86	SG2
P15	0.67	0.83	−0.16	1.50	SG1
P16	0.77	0.66	0.11	1.43	SG2
P19	0.80	0.70	0.10	1.50	SG1
P20	1.00	0.85	0.15	1.85	SG2
P23	0.76	0.57	0.19	1.33	SG1
P24	0.87	0.66	0.21	1.53	SG2

This section, thus, provides some advice to authors about how to report the outcomes of their studies that should make their studies more useful to their readers. It also provides two worked examples that novice researchers can try out to help them better understand the crossover design.

6.1 Example 1: Scaniello's Data

The dataset in Table 4 comprises a subset of the data reported by Scanniello et al. (2014) to support their paper.

The study investigated the impact of UML analysis models on source code comprehensibility (measured with the Comp_Level metric) and modifiability. The two techniques being compared are AM (analysis model plus source code) and SC (source code only). The techniques were trialed on two software systems S1 (a system to sell and manage CDs/DVDs in a music shop) and S2 (a software system to book and buy theater tickets. One feature from each system from each system was used as the object of study. The data relates to two groups in the dataset from the *EUBAS* experiment which itself was one of a family of four experiments. We chose that experiment rather than one of the others, because when we analyzed the EUBAS data, we found a non-zero repeated measures correlation which is an important pre-requisite for a crossover design to be of any value in decreasing the variance. It was also the experiment with the largest number of participants.

The full EUBAS experiment was a four-group crossover with Group 1 and Group 2 comprising one AB/BA crossover and Group 3 and Group 4 comprising another. The difference was that Group 1 and Group 2 used S1 and then S2, while Group 3 and Group 4 used S2 and then S1 (see Scanniello et al. 2014, Table II). We used the data from participants in the Group 3 and Group 4, as an example of an AB/BA crossover, since we found an anomaly in the reported data for Group 2[15]. We selected only a subset of Scanniello's data because we wanted to explain the AB/BA crossover rather than discuss the more complicated four-group

[15]The data for participant 2, who is labeled as being in Group 2, are inconsistent. That is, for participant 2, the labels identifying the system and the time period are the same as for participants in Group 1.

crossover which can be analyzed as a pair of AB/BA crossovers. Thus, the small balanced dataset provides an example of how the two-group crossover experimental results can be reported and the relevant analysis statistics are calculated.

Figure 1 shows two ways to represent crossover data graphically, while the code used to produce the figure is presented in Output 1.

Output 1 Code to Produce Example of Graphical Methods to Represent Crossover Data using Data from Scanniello

```
par(mfrow=c(1,2), cex=0.75)
boxplot(Comp_Diff~SequenceGroup, data=EUBASwideSelected, xlab="Sequence Group", cex.main=0.85, main=
      "(a) Comp_Level CO Difference (AM-SC Values)")
EUBASwideSelectedSort<-EUBASwideSelected[order(EUBASwideSelected$SequenceGroup),]
plot(EUBASwideSelectedSort$Comp_Level.AM,type="b",lty=1,pch=3,ylab="Comp_Level",ylim=c(0.4,1.0),xlab
      ="Participant Index",main="(b) Individual Comp_Level Values",cex.main=0.85)
lines(EUBASwideSelectedSort$Comp_Level.SC, typ="b", lty=2, pch=8)
abline(v=6.5)
text(1,0.55, "AM before SC", pos=4)
text(7,0.55, "SC before AM", pos=4)
legend("bottomright", inset=0.05, title="Technique", c("AM","SC"), lty=c(1,2), pch=c(3,8))
```

Panel (a) of Fig. 1 shows a box plot of the cross-over difference for each sequence group. The box plots show that the median value of the differences for individuals is below zero for individuals in sequence group 1 and above zero for individuals in sequence group 2. However, a large part of each box spans zero suggesting that there is no significant technique effect. Panel (b) shows the outcomes for each individual for technique. Seven of the individuals performed better using AM compared with five who performed better using SC. Again

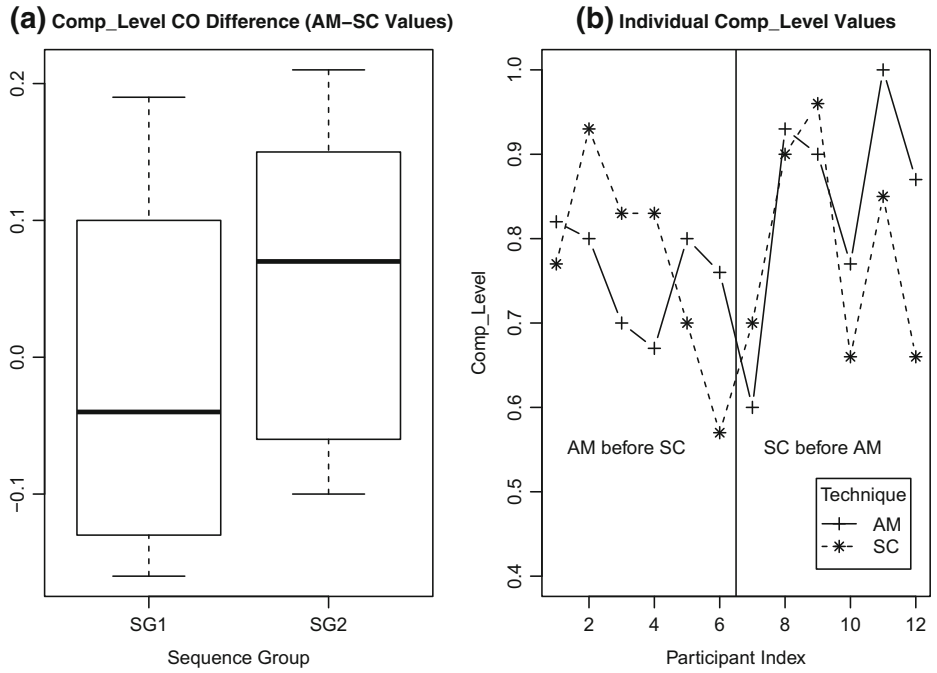

Fig. 1 Example of graphical methods to represent crossover data using data from scanniello

this gives no indication of any major difference between the techniques. An important issue to note is that participants that used AM before SC did *not* show the expected association between participant outcomes, i.e., participants who performed well using AM did not seem to perform well using SC and vice-versa. In contrast, participants who performed well using SC first generally performed well when subsequently using AM. The lack of a correlation between individual participant outcomes in SG_1 group means that overall the correlation between participants may be quite low. The graphical display in panel (b) is useful for small data sets since the results of box plots based on very few observations may be misleading, but for larger data sets, the box plots in panel (a) are usually more helpful.

The appropriate descriptive statistics for a cross-over study are shown in Table 5.

From these descriptive statistics all the effect sizes, their t-tests, and the effect size variances shown in Table 6 can be calculated (even if sample sizes in each sequence group are unbalanced). For example, using (18), $\hat{\tau} = \frac{-0.0133+0.0567}{2} = 0.0217$. Given that the Comp_Level metric varies between 0 and 1, the effect size is extremely small. Using (20), the period effect size is $\hat{\pi} = \frac{-((-0.0133)-(0.0567))}{2} = 0.035$. Using (23), the period by treatment interaction effect is $\hat{\lambda}_{AB} = 1.53 - 1.6333 = -0.1033$. Thus, in this case, it appears that interaction is large compared with the technique effect.

Table 6 reports that the t value for testing the significance of $\hat{\tau}$ is 0.5581 which, at an alpha level of 0.05, is not significantly different from zero. This outcome is consistent with the inferences we drew from panel (a) in Fig. 1. The estimate of ρ is 0.3613. Thus, the correlation between repeated measures in the EUBAS data set is rather low compared with that reported by Dunlap et al. (1996) for test-retest measurements. The low correlation was indicated by the lack of correlation between individual values for participants in SG_1 visible in panel (b) of Fig. 1. The correlation between an individual's performance indicates the extent to which the crossover design has decreased the variance compared with a standard independent groups design. In extreme cases, if $s^2_{IG} \approx s^2_w$, then ρ is assumed to be equal to zero and the crossover design has not reduced the variance at all.

6.2 Simulated Data Example

It is often helpful to use simulated data to understand the behavior of statistical tests and graphical representations. It allows us to check the accuracy of model parameter estimates against known values. It can also be used to check how sample size affects the accuracy of estimates or how violations of model assumptions affect analysis results. Examples of the use of simulation in software engineering include (Shepperd and Kadoda 2001) who used simulation to compare prediction techniques, Dieste et al. (2011) who investigated the use

Table 5 Descriptive statistics for the scanniello data

Sequence Group	Statistic	AM	SC	CODiff	Participant Total
SG_1	Mean	0.7583	0.7717	−0.0133	1.53
	Variance	0.0037	0.0155	0.0214	0.0171
	Num Obs	6	6	6	6
SG_2	Mean	0.845	0.7883	0.0567	1.6333
	Variance	0.0201	0.0173	0.0148	0.06
	Num Obs	6	6	6	6

Table 6 Statistics calculated from scaniello's data

Statistic	Equation Number	Value
$\hat{\tau}$	18	0.0217
$\hat{\pi}$	20	0.035
$\hat{\lambda}_{AB}$	23	−0.1033
s^2_{IG}	25	0.0142
s^2_{diff}	26	0.0181
s^2_w	27	0.009
$\hat{\rho}$	28	0.3613
$var(\hat{\tau})$	31	0.001508
$se_{\hat{\tau}}$	32	0.03884
t	33	0.5581

of the Q heterogeneity estimator for meta-analysis, and Foss et al. (2003) who investigated the properties of the MMRE statistic.

In this section we present a simulation study to illustrate the relationships between the graphical representations and descriptive statistics, in ideal circumstances (i.e., equal numbers of participants in each sequence group, stable variances, a large between participant correlation, no significant period by treatment interaction, and normal distributions). This dataset will also be used to allow the comparison of model parameter estimates with the known values of those parameters.

We simulated a data set such that:

- There are 15 participants in each sequence group.
- The average outcome across different participants is $\mu = 50$. We note that many of the papers used effectiveness measures based on a scale from 0 to 1 based on the proportion of questions answered correctly (see, for example, Scanniello et al. 2014; Abrahao et al. 2013). We chose a value of 50 which is equivalent to 50% of correct answers rather than a value between 0 and 1, so the effects would be clearer in the analysis.
- Users of technique 1 achieve an average of 10 units more than users of technique 2, that is $\tau = 10$. For a metric scale based on the number of correct answers to 10 questions, this would be equivalent to increasing the number of correct answers by one.
- Users achieve an average of 5 units more in period 2 than in period 1, that is, $\pi = 5$.
- There is no period by technique interaction effect built into the simulation (i.e. $\lambda_{AB} = 0$).
- The variance among participants using a specific technique in a specific time period is $\sigma^2 = 25$. This means the variance is unaffected by period or technique.
- The correlation between outcomes for an individual participant is $\rho = 0.75$. We chose the value 0.75 because (Dunlap et al. 1996) reported that such values are to be expected for test-retest reliabilities of psycho-metrically sound values. In the software engineering literature, Laitenberger et al. (2001) reported values of r varying from 0.78 to −0.02[16]

[16] According to Laitenberger et al. (2001) the results from one team had a large impact on this correlation coefficient. When removing this observation the correlation coefficient changes from −0.02 to 0.47.

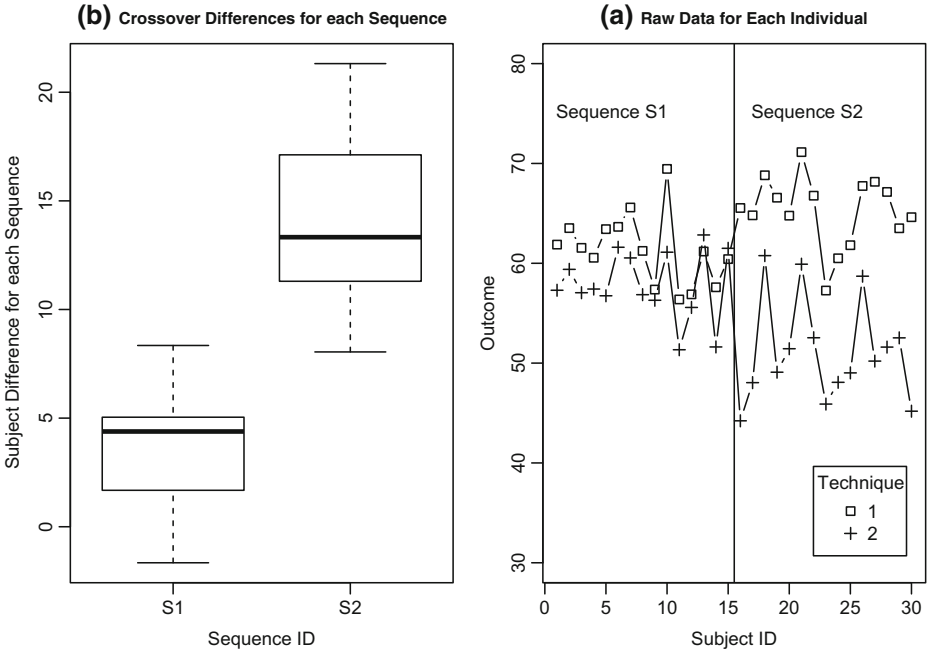

Fig. 2 Example of graphical methods to represent crossover data using simulated data

for the correlation between outcomes from teams. However, it would be reasonable to expect correlations based on individuals to be greater than those based on teams.

We simulated data from two different bivariate normal distributions, corresponding to the two different sequence groups. The first set of simulate data corresponding to sequence group SG_1 came from a bivariate distribution with means: $\mu_1 = 60$ corresponding to the simulated participants using technique 1 in time period 1 and $\mu_2 = 55$ corresponding to the simulated participants using technique 2 in time period 2. The covariance matrix was symmetric, with the variance of the simulated participants using a specific technique in a specific time period being set to 25 and the covariance being set to $25 * (1 - \rho) = 18.75$. Observations from sequence group SG_2 were simulated from a bivariate normal distribution with the same variance-covariance matrix and means $\mu_1 = 50$ corresponding to the simulated participants using technique 2 in time period 1, and $\mu_2 = 65$ corresponding to the simulated participants using technique 1 in time period 2. After allowing for the common mean effect of 50, the simulated data come from a population where the effect of technique 1 is 10 units and the effect of technique 2 is zero units.

The simulated data set, as well as how the data can be generated using the `reproducer` package are presented in Output 5 in Appendix A. The results of this simulation are shown in Fig. 2, while the code used to produce the figure is presented in Output 2.

Output 2 Code to Produce Example of Graphical Methods to Represent Crossover Data using Simulated Data

```
data<-reproducer::getSimulationData(25, 18.75, 50, 10, 5, 15)
dataWide=reshape(data,idvar="pid",timevar="technique",direction="wide")
dataWide$Difference=dataWide$result.T1-dataWide$result.T2
graphics::par(mfrow=c(1,2),cex=.75)
graphics::boxplot(Difference ~ sequence.T2,data=dataWide,ylab="Subject Difference for each Sequence"
      , xlab="Sequence ID",main="(b) Crossover Differences for each Sequence",cex.main=0.85)
graphics::plot(dataWide$pid,dataWide$result.T1,ylim=c(30,80),xlim=c(1,30),ylab="Outcome",xlab="
      Subject ID",pch=0,main="(a) Raw Data for Each Individual",typ="b",cex.main=0.85)
graphics::points(dataWide$pid,dataWide$result.T2,pch=3,typ="b")
graphics::abline(v=15.5)
graphics::text(0,75,"Sequence S1", pos=4)
graphics::text(16,75, "Sequence S2", pos=4)
graphics::legend("bottomright",title="Technique", inset=0.05, c("1","2"), pch=c(0,3))
```

The first thing to notice is that even with 15 data points in each sequence group, the box plots deviate from what we expect from a normal distribution (i.e., the median for each sequence group is not in the center of the box). Looking at the box plot, we see the difference between the medians of the boxes is approximately $(13-4)=9$ units. In general, since the average of the difference values for participants in sequence SG_1 are estimates of $\hat{\tau}-\hat{\pi}$ and the average of the difference values for participants in sequence SG_2 are estimates of $\hat{\tau}+\hat{\pi}$, the difference between the medians will be approximately twice the period effect, which for our simulation was 5. Also the sum of the medians $(13+4)=17$ will be approximately equal to twice the technique effect, which for our simulation was 10.

Looking at the raw data for each individual, we see that the simulated participants in sequence group $S2$, that used technique 2 first and subsequently used technique 1, show a strong difference between their outcomes. This is because the impact of using technique 1 is increased by the period effect. The simulated participants in sequence group $S1$ that used technique 1 first however, showed less of a clear advantage when they used technique 1 compared with their results using technique 2. The individual outcomes for simulated participants in period 2 was greater than the outcome for period 1, for 13 of the 15 simulated participants but the differences were quite small. This is because in the second time period, individual results were increased because of the positive impact of the period effect[17].

The descriptive statistics for the simulated data are shown in Table 7.

These statistics can be used to calculate the estimates of the sample parameters. For example, in this case $\hat{\tau}=\frac{3.5653+14.1282}{2}=8.84675$, which is a reasonable estimate of $\tau=10$ and confirms that the effect of the relatively large period effect has been removed.

Table 8 compares the simulation sample estimates to the values of the parameters we used to simulate the data set[18].

The relative error of an estimate is calculated as

$$PercentRelativeError=100*\frac{(TheoreticalValue-SampleEstimate)}{TheoreticalValue} \tag{57}$$

[17]Of course, it is also possible for a period effect to be negative.

[18]We omit an estimate of the period by technique interaction term λ_{AB} because we omitted any such term from our simulation model.

Table 7 Descriptive statistics for the simulated data crossover design

Sequence	Statistic	Technique	Technique	CODiff	Participant
Group		1	2		Total
SG_1	Mean	61.3772	57.8119	3.5653	119.1891
	Variance	12.4561	11.7601	7.7316	40.7007
	Num Obs	15	15	15	15
SG_2	Mean	65.2768	51.1486	14.1282	116.4254
	Variance	12.2649	26.4595	14.9214	62.5274
	Num Obs	15	15	15	15

There are some substantial differences between the theoretical values and the estimates from our sample, particularly for the estimate of σ^2_{IG}. Thus, even under ideal conditions, samples based on only 30 observations in all (i.e., 15 in each sequence group) may not give very reliable results. Nonetheless the value of the t-statistic is 14.40 which is statistically significant at $\alpha = 0.05$.

6.3 Using R to Calculate Non-Standardized Effect Sizes and their Variances

Vegas et al. (2016) proposed the use of linear mixed models to analyze crossover data. They did not recommend a specific statistical package, but in this study we used the R linear modeling package `lme4` which can correctly analyze crossover designs with unequal sample sizes. This package assumes that the data is in the *long* format, i.e., there are two rows for each participant, identifying the participant period, treatment, and results.[19]

6.3.1 Analyzing Scaniello's data

Using the long format with variable names: ID for the participant identifier (with values P1, P2, ..., P12), Time Period (with values R1 and R2) and Technique (with values AM and SC), and Comp_Level for the result, the first few lines of the SE example data would need to have the structure illustrated in Table 9[20].

The results of using the *lme4* package to analyze the Scanniello data are shown in Output 3.[21] ID is treated as a random effects term, whereas Time Period and Technique are treated as fixed effects terms. Unlike Vegas et al. who include a *Sequence* effect (which (23) confirms is a means of testing the period by treatment interaction) as well as a Time Period and Technique effect, we adopted Senn's approach, as discussed in Section 4.2.2, and did not include a parameter related to the period by treatment interaction term (λ_{AB}).

Output 3 Linear Mixed Model Analysis of the Scaniello Crossover data

[19] The data in Table 4 is in the *wide* format where there is one entry of each participant and the outcomes for each treatment, the sequence order, and the participant identifier are recorded.

[20] The data we used for the analysis in this section is exactly the same as the data we used in Section 6.1

[21] The term " + (1 — ID)" in the formula identifies the factor ID as a random effects term.

```
#Preliminary steps to install the required packages in proper versions locally
#install.packages("devtools",dependencies=T,repos="http://cran.rstudio.com/")
#library(devtools)
#install_version("reproducer",version="0.1.9",dependencies=T,type="source",repos="http://
      cran.rstudio.com/")
#install_version("lme4",version="1.1-12",dependencies=T,type="source",repos="http://cran.rstudio.com
      /")
ScanielloData<-reproducer::MadeyskiKitchenham.EUBASdata
ScanielloG3G4=ScanielloData[which(ScanielloData$SequenceGroup=="G3" | ScanielloData$SequenceGroup=="
      G4"),]
library(lme4)
EUBASfit=lme4::lmer(Comp_Level~TimePeriod+Technique+(1|ID),data=ScanielloG3G4)
summary(EUBASfit)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Comp_Level ~ TimePeriod + Technique + (1 | ID)
   Data: ScanielloG3G4

REML criterion at convergence: -24.3

Scaled residuals:
   Min     1Q Median     3Q    Max
-1.700 -0.681  0.150  0.505  1.311

Random effects:
 Groups   Name        Variance Std.Dev.
 ID       (Intercept) 0.00497  0.0705
 Residual             0.00904  0.0951
Number of obs: 24, groups:  ID, 12

Fixed effects:
             Estimate Std. Error t value
(Intercept)    0.7842     0.0393   19.95
TimePeriodR2   0.0350     0.0388    0.90
TechniqueSC   -0.0217     0.0388   -0.56

Correlation of Fixed Effects:
            (Intr) TmPrR2
TimePeridR2 -0.494
TechniqueSC -0.494  0.000
```

The effect size related to *TechniqueSC* is -0.0217 and the effect size related to *TimePeriodR2* is 0.035. The non-standardized effect size variance is the square of technique effect standard error (i.e., $0.388^2 = 0.1505$). The value for the period effect is the same as we found in our manual analysis, but the value of the technique effect is minus the value we found in our analysis. This is because the package calculated $\hat{\tau}_{BA}$ rather than $\hat{\tau}_{AB}$. We treated AM as the experimental effect and associated it with the sequence SG_1 as defined in Table 1. However, the `lme4::lmer` function in `R` treats the labels given to different categorical variables as arbitrary, and uses the category corresponding to the larger alphanumeric label as the one for which it will calculate the effect size[22]. Since SC is greater than

[22]This is the same for all `R` ANOVA-like functions.

Table 8 Parameter and variance estimates for the simulated data

Parameter	Sample Estimate	Theoretical Value	Percent relative Error
$\hat{\tau}$	8.8467	10	11.5325
$\hat{\pi}$	5.2814	5	−5.6284
s^2_{IG}	15.7351	25	37.0595
s^2_{diff}	11.3265	12.5	9.388
s^2_w	5.6632	6.25	9.388
$\hat{\rho}$	0.6401	0.75	14.6548
$var(\hat{\tau})$	0.3775	0.4167	9.4072
$se_{\hat{\tau}}$	0.6145	0.6455	3.1046
t	14.3978	15.4919	7.5992

AM alphabetically, it calculates the effect size for $SC - AM$ and labels the effect size $TechniqueSC$.

The variance term associated with the *Residual* is s^2_w, and the variance term associated with ID is s^2_b giving $s^2_{IG} = s^2_w + s^2_b = 0.014012$ compared with our manual estimate of 0.01416. Also, $\hat{\rho} = \frac{s^2_b}{s^2_{IG}} = 0.3546$ compared with our manual estimate of 0.36135. Minor differences between estimates of the variances and the correlation are to be expected when comparing a mixed effects analysis based on maximum likelihood estimation with a manual analysis.

6.3.2 Analyzing the simulated data

The results of the analysis of the simulated data set are shown in Output 4 and can be compared with the values shown in Table 8. The estimates of the period effect sizes are the same, but, again the estimate of the technique effect is negative. This occurs for the same reason that the sign of the technique effect changed for Scaniello's data. From the variances reported in Output 4, $s^2_{IG} = 10.12 + 5.66 = 15.78$ which compares with the manual estimate of 15.7351 and $\hat{\rho} = 10.12/15.78 = 0.6413$ which compares with the manual estimate 0.6401.

Table 9 Example of the long data format using the scanniello data

ID	TimePeriod	Technique	Comp_Level
P3	R1	AM	0.82
P3	R2	SC	0.77
P4	R1	SC	0.70
P4	R2	AM	0.60

Output 4 Linear Mixed Model Analysis of the Simulated Crossover data

```
simdata=reproducer::getSimulationData(25, 18.75, 50, 10, 5, 15)
simdatafit=lme4::lmer(result ~ technique+period+(1|pid), data=simdata)
summary(simdatafit)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: result ~ technique + period + (1 | pid)
   Data: simdata

REML criterion at convergence: 314.2

Scaled residuals:
   Min     1Q Median     3Q    Max
-1.880 -0.683  0.010  0.344  1.819

Random effects:
 Groups   Name        Variance Std.Dev.
 pid      (Intercept) 10.12    3.18
 Residual              5.66    2.38
Number of obs: 60, groups:  pid, 30

Fixed effects:
            Estimate Std. Error t value
(Intercept)   60.686      0.788    77.0
techniqueT2   -8.847      0.614   -14.4
periodP2       5.281      0.614     8.6

Correlation of Fixed Effects:
            (Intr) tchnT2
techniqueT2 -0.390
periodP2    -0.390  0.000
```

6.4 Calculating Standardized Effect Sizes and their Variances

The standardized effect sizes based on the linear mixed model analysis are shown in Table 10. d_{RM} is calculated from (38) and d_{IG} is calculated from (39). Since d_{IG} is standardized with s_{IG}, its absolute value is smaller than d_{RM} which is standardized with s_w. Only when there is no discernible correlation between the repeated measures will $s_{IG} = s_w$, and the effect sizes will be the same.

The adjusted standardized effect sizes are shown in Table 11. The values of $c(df)$ are derived from (42), with df replaced by the appropriate degrees of freedom, (i.e., 10 for Scaniello's data and 28 for the simulated data). g_{RM} is calculated from (43) and g_{IG} is calculated from (44).

The estimated variance of the effect sizes, the variance approximations and the percentage relative error (PRE) of the approximations are shown in Table 12 for each of the datasets. All the values reported in this table were obtained from values calculated from the `lmer` analyzes shown in Output 3 and Output 4. For the simulated data, we can compare the standardized effect size variances with the theoretical variances obtained by using the variance formulas with the values used to generate the simulated data. The theoretical variance of δ_{RM} for a data set of 30 observations with 15 in each group is $var(\delta_{RM}) = 0.4013$ and the theoretical variance of δ_{IG} is $var(\delta_{IG}) = 0.1003$. In comparison with the theoretical

Table 10 Example standardized effect sizes

Effect Size	Scaniello data Estimate SC-AM	Simulation data T2-T1	Theoretical Value	Percent Relative Error
d_{RM}	−0.2278	−3.7175	−4	7.0625
d_{IG}	−0.183	−2.2268	−2	−11.3384

values, $var(d_{RM})$ is the best estimate of $var(\delta_{RM})$ and $var(g_{RM})_{approx}$ is the worst, but, in contrast, $var(d_{IG})$ is the worst estimate of $var(\delta_{IG})$ and $var(g_{IG})_{approx}$ is the best. A more extended simulation study would be needed to determine which estimates were most likely, on average, to be the best.

The percentage relative accuracy is the same for $var(d_{RM})_{approx}$ and $var(d_{IG})_{approx}$. This occurs because the small sample and medium sample variance of d_{RM} and d_{IG} are simply a function of the variance of t multiplied by a constant which cancels out when the relative error is calculated.This is the same for $var(g_{RM})_{approx}$ and $var(g_{IG})_{approx}$.

7 Discussion

This paper is intended to follow-up some additional issues arising from Vegas et al. (2016)'s recent paper identifying problems with the analysis of crossover experiments. Vegas et al. discussed four repeated measures designs other than the simple AB/BA crossover. However, all those designs are extensions of the AB/BA crossover, including either additional sequences and/or additional periods and/or repeating the same techniques, so in order to understand these extensions it is important to understand the basic crossover design. We provide a discussion of the model underlying the AB/BA crossover design, so that issues connected with the construction of effect sizes and effect size variances can be properly understood.

7.1 Impact of Incorrect Analysis on Effect Sizes and their Variances

Vegas et al. (2016) reported that many researchers using crossover designs did not account for the repeated measures in their analysis. For an AB/BA crossover, analyzing the data, without including a factor relating to individual participants effects, would lead to an over-

Table 11 Example standardized effect size adjustment

Effect Size	Adjustment Scaniello Data c(10)	Scaniello data Estimate Revised	Adjustment Sim Data c(28)	Simulation Estimate Revised
g_{RM}	0.9231	−0.2103	0.973	−3.617
g_{IG}	0.9231	−0.169	0.973	−2.1666

Table 12 Standardized effect size variances and their approximations

Statistic	Equation number	Scanniello data	Simulation data
$var(d_{RM})$	49	0.2117	0.3412
$var(d_{IG})$	51	0.1366	0.1224
$var(g_{RM})$	50	0.1804	0.3231
$var(g_{IG})$	52	0.1164	0.1159
$var(d_{RM})_{Approx}$	53	0.1699	0.3318
$PREvar(d_{RM})_{Approx}$	57	19.7557%	2.763%
$var(d_{IG})_{Approx}$	55	0.1096	0.1191
$PREvar(d_{IG})_{Approx}$	57	19.7557%	2.763%
$var(g_{RM})_{Approx}$	54	0.1689	0.3003
$PREvar(g_{RM})_{Approx}$	57	6.3835%	7.0461%
$var(g_{IG})_{Approx}$	56	0.109	0.1077
$PREvar(g_{IG})_{Approx}$	57	6.3835%	7.0461%

estimate the degrees of freedom available for statistical tests by using $df = 2(n_1 + n_2 - 2)$ instead of $df = (n_1 + n_2 - 2)$. In this section, we consider, hypothetically, what the impact on the effect sizes and their variances would be if the subset of the Scaniello data and the simulated data analyzed in Section 6 were analyzed ignoring the repeated measures and period effects.

From the viewpoint of constructing effect sizes, the variances used to standardize the technique effect would be based on the pooled within technique group data. However, in the presence of a significant period effect, this would be a biased estimate of the σ^2_{IG} because the period effect would systematically inflate the variance. This is the inverse of removing a significant blocking effect in an analysis of variance in order to significantly reduce the residual error term. If period is a significant blocking effect, failing to remove its effect from the variance leave an inflated variance. In contrast ignoring the repeated measures might deflate the variance because, if there is a strong correlation between the repeated measures there will be less variability among the observations in each technique group than if the the observations in each group were completely independent.

The value of the technique effect will be correctly estimated by the difference between the mean of the values in each technique group. It is only likely to be biased if the number of observations in each sequence group is unequal (i.e., $n_1 \neq n_2$). Thus, the effect size calculated by using the treatment effect divided by the pooled within group standard deviation will lead to a slightly biased estimate of δ_{IG}.

To convert to Hedges'g the estimate of δ_{IG} is multiplied by $c(df)$. If the analysis has ignored the repeated measures, df will be $2n_1 + 2n_2 - 2$ rather than $n_1 + n_2 - 2$, and $c(df)$ will be slightly closer to one than it should be.

Based on the datasets introduced in Section 6, the effects of incorrectly calculating sample statistics are shown in Table 13. As can be seen the bias in the calculation of s^2_{IG} is very small for the analyzed subset of the Scaniello's data, which has both a small period effect and a small technique effect. However, the bias is much larger for the simulated data which has a relatively large standardized effect size and included a substantial period effect. As a result the bias in the estimate of δ_{IG} is negligible for Scanniello's data but more substantial for the simulated data. The impact on c(df) is more pronounced for Scanniello's data than

Table 13 Effect of incorrect analyzes

Statistic	Scaniello's data	Simulated data
s^2_{IG}	0.0142	15.7351
$s^2_{IGbiased}$	0.01393	22.9002
d_{IG}	−0.0183	−2.2268
$d_{IGbiased}$	−0.01835	−1.849
$c(df)$	0.9231	0.973
$c(dfWrong)$	0.9655	0.9870
g_{IG}	−0.169	−2.1666
$g_{IGbiased}$	−0.177	−1.8247
$var(g_{IG})$	0.1164	0.1159
$var(g_{IGBiased})$	0.1709	0.09720

the simulated data because the sample size is smaller. In each case the impact on Hedges' g is that the small sample adjustment factor is underestimated.

If researchers do not analyze their crossover data as a repeated measures study, they are likely to estimate the variance of their biased estimate of g_{IG} as if the study was an independent groups study. In Table 13, we compare the correct estimate of $var(g_{IG})$ with $var(g_{IGbiased})$ based on the formula for the variance of an adjusted effect size estimate of an independent groups study (see Hedges and Olkin 1985). For the Scaniello data, $abs(g_{IGbiased})$ is greater than $abs(g_{IG})$, and the estimate of $var(g_{IGBiased})$ is greater than $var(g_{IG})$. For the simulated data, $abs(g_{IGbiased})$ is less than $abs(g_{IG})$ and $var(g_{IGBiased})$ is less than $var(g_{IG})$. This happens because the formula for $var(g_{IG})$ includes the term g^2_{IG}. So, the larger the effect size, the larger the effect size variance.

Overall, we can say that if $\delta_{RM} \approx 0$ and $\rho \approx 0$, analyzing a crossover study incorrectly is unlikely to lead to an incorrect assessment of the significance of the technique effect. Furthermore, if the effect size is very large, we are likely to find that the effect is statistically significant. That is, for very small effects and very large effects, the incorrect analysis will lead to accidentally correct assessments of significance. However, for small to medium effects it is quite possible that real effects will be considered chance effects, or chance effects considered significant. In addition, in all cases where the non-standardized effect size, or ρ, or the period effect are non-zero, any estimates of the standardized effect sizes and their variances will be unreliable.

7.2 Standardized Effect Sizes and their Variances

Our presentation of the crossover model raises several issues that have not been fully discussed in the software engineering literature. In particular, we point out that for crossover designs, there are two different standardized effect sizes that can be calculated. Furthermore, each standarized effect size has a different formula for its variance. We also point out that standardized effect sizes and their variances are different for different design types. These issues have implications for meta-analysis in software engineering, where as far as we are aware, only (Madeyski 2010) has explicitly discussed the fact that experimental design type impacts the calculation of standardized effect sizes.

The results of our study have implications for the descriptive data from crossover designs. Our examples in Section 6 show what sample statistics need to be reported to allow effect sizes and their variances to be easily calculated from descriptive statistics. Specifically,

researchers should report the mean, sample size and standard deviation (or variance) for all four technique and period groups, *as well as* either the crossover difference mean and standard deviation for each sequence[23]. We also suggest graphical representations of crossover data that allow readers to easily visualize the results of the study.

7.3 Implications for Planning Experiments

The crossover model has limitations, and in particular, we have not identified any method to properly address the risk of a significant period by technique interaction biasing any analysis of crossover data. The specific effect of the bias is uncertain because the direction of the period by technique effect can be positive or negative. Assuming t_{AB} is positive, (16) confirms that if λ_{AB} is positive, the estimate of the non-standardized effect size will be decreased, if it is negative, the estimate of the non-standardized effect size will be increased.

In the case of software engineering techniques, it is difficult to provide convincing *a priori* arguments that techniques do not interact. Indeed, the subset of Scaniello's software engineering data that we show in Fig. 1b seems to suggest an interaction term is present since the results of participants who used AM first did not seem to be correlated while the results of participants who use SC first did seem to be correlated.

Another concern is that unless the repeated measures correlation, ρ, is relatively large, the reduction in the variance used for statistical testing will be relatively small. Equation (30) confirms that the reduction in variance is $100 \times (\sigma_{IG}^2 - \sigma_w^2)/\sigma_{IG}^2 = \rho$. Thus, for Scanniello's data, the percentage reduction in the variance given the repeated measure correlation of 0.3613 is approximately 36%. In contrast, for the simulated data, the percentage reduction in the variance is approximately 64%. Thus, unless we know in advance the likely value of the repeated measure correlation, we may radically under- or over-estimate the impact of the crossover design and could adopt an inappropriate sample size.

This suggests that before we could rely on an AB/BA crossover design to investigate some new topic, we would need to undertake an experiment in order to investigate both the nature of the period by technique interaction and the repeated measures correlation. We might envisage an investigatory crossover experiment aimed at providing such information, where the information from the first period could be used to test the difference between techniques and estimate effect sizes, using a between groups design, and the information from the second period used to investigate the interaction term and the correlation parameter. The problem is that to provide reliable information concerning the interaction, an experiment would have to have a sample size as large as a between groups experiment.

Another option is to consider an alternative to a crossover design that allows the impacts of skill levels to be removed. This can done using what Morris and DeShon (2002) refer to as an *independent-groups pretest-posttest design* [24]. In such a design all subjects do the same task using technique A and the same materials M1, then the participants are split into two groups and each group learns a new technique (i.e., techniques B and C) to perform the experimental task. In practice, one of the new techniques might simply be extra coaching for technique A, but it is likely that the design would be fairer if A was a control method and B and C were different competing methods. The difference scores can be used to estimate the difference between technique B and technique C with the effects of individual differences removed. The disadvantage of this design is that it assumes the time effect is equal for both groups.

[23] It is also acceptable to report the the value of $\hat{\rho}$ instead of the crossover difference statistics

[24] Morris and DeShon also report the formulas for the effect sizes and effect size variances for this design.

Vegas et al. (2016) mentioned four other possible repeated measures designs, but, unlike the independent-groups pretest-posttest design, those designs have not been discussed in the statistical literature. In our opinion, before such designs should be considered for adaption, the full model underlying the design needs to be articulated, as we did for the simple crossover in Section 4. This should include defining how to calculate appropriate effect sizes and their variances, as well as specifying the theoretical assumptions and practical limitations of the design.

Our best advice is to avoid over-complex designs that are not fully understood and always aim for the largest sample size. If large sample sizes are not possible, consider planning a distributed experiment as proposed by Budgen et al. (2013). In a distributed experiment, all related experiments must use the same protocol and results are aggregated as a single nested experiment. Kitchenham et al. (2017a) report the analysis of the data that was collected from this distributed experiment.

7.4 Implications for Meta-Analysis

The results in this paper make it clear that it is possible to aggregate experiments that used independent groups designs with experiments that used crossover designs. The crossover designs can be aggregated using d_{IG}, since this is comparable with the usual standardized effect size for independent groups experiments. The experiments that used an independent groups design, should use the usual standardized mean difference. It is, however, important to use the correct effect size variance, which is based on the variance of the related t-variable. The appropriate formula for the standardized difference of independent group studies and its variance can be found in Hedges and Olkin (1985). We note that the g_{RM} values from three studies reported by Laitenberger et al. (2001) were used without adjustment in a meta-analysis that involved crossover studies, independent groups studies and repeated measures before/after studies (see Ciolkowski 2009). For example, one of Laitenberger's studies reported $g_{RM} = 1.46$[25] but with $\hat{\rho} = 0.77$, the value of $g_{IG} = 0.70$. The results reported in this paper will, in the future, allow meta-analysts to select the most appropriate effect size for cross-over studies, before-after studies and independent groups studies.

7.5 Non-Normality and Unstable Variances

We discuss the issue of non-normality briefly in Section 4.2.3, but a more detailed investigation is needed to determine the most appropriate non-parametric effect sizes and the variance of non-parametric effect sizes. Also the issue of meta-analysis of non-parametric effect sizes needs to be investigated, not only when all effect sizes are non-parametric but also when there is a mixture of non-parametric and parametric effect sizes.

8 Conclusions

This paper provides a discussion of standardized effect size calculations and their variances for crossover designs. This is becoming important because as Vegas et al. (2016) point out many software engineering researchers are employing crossover designs and analyzing them incorrectly. Furthermore, crossover designs are often used in families of experiments, where researchers attempt to aggregate their results using meta-analysis.

[25] This was referred to as d in Laitenberger et al. (2001).

The contribution of this paper is:

- To provide equations for non-standardized and standardized effect sizes. We explain the need for two different types of standardized effect size, one for the repeated measures design and one that would be equivalent to an independent groups design.
- To provide formulas for both the small sample size effect size variance and the medium sample size approximation to the effect size variance, for both types of standardized effect size.
- To explain how the different effect sizes can be obtained either from standard descriptive statistics or from information provided by the linear mixed model package `lme4` in `R`.

We conclude that crossover designs should be considered only if:

- Previous research has suggested that ρ is greater than zero and preferably greater than 0.25.
- There is either strong theoretical argument, or empirical evidence from a well-powered study, that the period by technique interaction is negligible.

Having reproducible research in Empirical Software Engineering in mind we would be happy (after acceptance of the paper and obtaining a permission from the Editor) to make available the source version of the paper with embedded R code (in addition to the `reproducer` R package already available from CRAN) along with the article in the PDF format.

Acknowledgments We thank Prof. Scanniello and his co-authors (Scanniello et al. 2014) for sharing their data set. We also thank reviewers for their help in improving our manuscript.

Appendix

A Reproducibility of the presented research findings

In order to document the research process and to allow other researchers to check and reproduce the presented research findings the `reproducer` R package (Madeyski 2017) supports the paper. Usage of the functions of the `reproducer` package, which are closely related to this paper, is illustrated in the main body of paper. Furthermore, call to Function `getSimulationData()` and the first few rows of the simulated data used in Section 6.2 is illustrated in Output 5.

Output 5 R Commands and the first few rows of the Output of Function `getSimulationData()` from the reproducer R package

```
simulationData<-reproducer::getSimulationData(25, 18.75, 50, 10, 5, 15)
head(simulationData) #shows the first few rows of the simulated data
```

```
  pid period sequence result technique
1   1     P1       S1  61.87        T1
2   2     P1       S1  63.52        T1
3   3     P1       S1  61.54        T1
4   4     P1       S1  60.55        T1
5   5     P1       S1  63.42        T1
6   6     P1       S1  63.65        T1
```

A key part of documenting the research process with R is recording the R session info, which makes it easy for future researchers to recreate what was done in the past and which versions of the R packages were used. The information from the session we used to create this research paper is shown in Output 6:

Output 6 R session info (R command and related output)

```
utils::sessionInfo()
```

```
R version 3.4.1 (2017-06-30)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.2 LTS

Matrix products: default
BLAS: /usr/lib/libblas/libblas.so.3.6.0
LAPACK: /usr/lib/lapack/liblapack.so.3.6.0

locale:
 [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
 [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
 [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
[10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] methods   stats     graphics  grDevices utils     datasets  base

other attached packages:
[1] reshape_0.8.6    MASS_7.3-47      lme4_1.1-13      Matrix_1.2-10
[5] xtable_1.8-2     reproducer_0.1.9 knitr_1.16

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.11    lattice_0.20-35 plyr_1.8.4      grid_3.4.1
 [5] nlme_3.1-131    magrittr_1.5    evaluate_0.10.1 highr_0.6
 [9] stringi_1.1.5   minqa_1.2.4     nloptr_1.0.4    splines_3.4.1
[13] tools_3.4.1     stringr_1.2.0   compiler_3.4.1
```

References

APA (2010) Publication manual of the American Psychological Association, 6th edn. American Psychological Association, Washington

Abrahao S, Gravino C, Insfran Pelozo E, Scanniello G, Tortora G (2013) Assessing the effectiveness of sequence diagrams in the comprehension of functional requirements: results from a family of five experiments. IEEE Trans Softw Eng 39(3):327–342. https://doi.org/10.1109/TSE.2012.27

Arcuri A, Briand L (2014) A Hitchhiker's guide to statistical tests for assessing randomized algorithms in software engineering. Softw Test Verification Reliab 24(3):219–250. https://doi.org/10.1002/stvr.1486
Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48. https://doi.org/10.18637/jss.v067.i01
Becker BJ (1988) Synthesizing standardized mean-change measures. Br J Math Stat Psychol 41:257–278
Borenstein M, Hedges LV, Higgins JPT, Rothstein HT (2009) Introduction to meta-analysis. Wiley, NY
Budgen D, Kitchenham B, Charters S, Gibbs S, Pothong A, Keung J, Brereton P (2013) Lessons from conducting a distributed quasi-experiment. In: Proceedings $ACM/IEEE$ International Symposium on Empirical Software Engineering and Measurement
Ciolkowski M (1999) Evaluating the effectiveness of different inspection techniques on informal requirements documents. PhD thesis, University of Kaiserslautern, Kaiserslautern
Ciolkowski M (2009) What do we know about perspective-based reading? an approach for quantitative aggregation in software engineering. In: Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement. IEEE Computer Society, Washington, ESEM '09, pp 133–144. https://doi.org/10.1109/ESEM.2009.5316026
Cumming G, Finch S (2001) A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. Educ Psychol Meas 61(4):532–574
Cumming G (2012) Understanding the new statistics. Effect Sizes, Confidence Intervals and Meta-Analysis. Routledge Taylor & Francis Group, New York
Curtin F, Altman DG, Elbourne D (2002) Meta-analysis combining parallel and cross-over clinical trials. I: Continuous outcomes. Stat Med 21:2132–2144. https://doi.org/10.1002/sim.1205
Dieste O, Fernández E, Garcia-Martinez R, Juristo N (2011) The risk of using the Q heterogeneity estimator for software engineering experiments. In: Proceedings of the 2011 International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society, ESEM '11, pp 68–76. https://doi.org/10.1109/ESEM.2011.15
Dunlap WP, Cortina JM, Vaslow JB, Burke MJ (1996) Meta-analysis of Experiments with Matched Groups or Repeated Measures Designs. Psychol Methods 1(2):170–177
Foss T, Stensrud E, Myrtveit I, Kitchenham B (2003) A simulation study of the model evaluation criterion mmre. IEEE Trans Softw Eng 29(11):985–995
Freeman P (1989) The performance of the two-stage analysis of two-treatment, two-period cross-over trials. Stat Med 8:1421–1432
Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Academic Press Orlando, Florida
Johnson NL, Welch BL (1940) Applications of the non-central t-distribution. Biometrika 31(3-4):362–389
Jureczko M, Madeyski L (2015) Cross–project defect prediction with respect to code ownership model: an empirical study. e-Informatica Softw Eng J 9(1):21–35. https://doi.org/10.5277/e-Inf150102
Kampenes VB, Dybå T, Hannay JE, Sjøberg DIK (2007) Systematic review: A systematic review of effect size in software engineering experiments. Inf Softw Technol 49(11-12):1073–1086. https://doi.org/10.1016/j.infsof.2007.02.015
Kitchenham BA, Madeyski L (2016) Meta-analysis. In: Kitchenham BA, Budgen D, Brereton P (eds) Evidence-Based Software Engineering and Systematic Reviews. CRC Press, chap 11, pp 133–154
Kitchenham B, Madeyski L, Budgen D, Keung J, Brereton P, Charters S, Gibbs S, Pohthong A (2017a) Robust statistical methods for empirical software engineering. Empir Softw Eng 22(2):579–630. https://doi.org/10.1007/s10664-016-9437-5
Kitchenham B, Madeyski L, Curtin F (2017b) Corrections to effect size variances for continuous outcomes of cross-over clinical trials. Statistics in Medicine https://doi.org/10.1002/sim.7379, http://madeyski.e-informatyka.pl/download/KitchenhamMadeyskiCurtinSIM.pdf, (accepted)
Laitenberger O, Emam KE, Harbich TG (2001) An internally replicated quasi-experimental comparison of checklist and perspective-based reading of code documents, vol 27, pp 387–421. https://doi.org/10.1109/32.922713
Lakens D (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 4(Article 863):1–12. https://doi.org/10.3389/fpsyg.2013.00863
Madeyski L (2010) Test-driven development: an empirical evaluation of agile practice. Springer, Heidelberg. http://www.springer.com/978-3-642-04287-4, Foreword by Prof. Claes Wohlin
Madeyski L, Orzeszyna W, Torkar R, Józala M (2014) Overcoming the equivalent mutant problem: a systematic literature review and a comparative experiment of second order mutation. IEEE Trans Softw Eng 40(1):23–42. https://doi.org/10.1109/TSE.2013.44
Madeyski L, Jureczko M (2015) Which process metrics can significantly improve defect prediction models? an empirical study. Softw Qual J 23(3):393–422. https://doi.org/10.1007/s11219-014-9241-7

Madeyski L (2017) reproducer: Reproduce Statistical Analyses and Meta-Analyses. http://madeyski.e-informatyka.pl/reproducible-research/, R package version 0.1.9 http://CRAN.R-project.org/package=reproducer
Morris SB (2000) Distribution of the standardized mean change effect size for meta-analysis on repeated measures. Br J Math Stat Psychol 53:17–29
Morris SB, DeShon RP (2002) Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychol Methods 7(1):105–125. https://doi.org/10.1037//1082-989X.7.1.1.105
R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/
Scanniello G, Gravino C, Genero M, Cruz-Lemus JA, Tortora G (2014) On the impact of UML analysis models on source-code comprehensibility and modifiability. ACM Trans Softw Eng Methodol 23(2):13:1–13:26. https://doi.org/10.1145/2491912
Senn S (2002) Cross-over trials in clinical research, 2nd edn. Wiley, NY
Shepperd M, Kadoda G (2001) Comparing software prediction techniques using simulation. IEEE Trans Softw Eng 27(11):1014–1022
Stout DE, Ruble TL (1995) Assessing the practical significance of empirical results in accounting education research: the use of effect size information. J Account Educ 13(3):281–298
Urdan TC (2005) Statistics in plain english, 2nd edn. Routledge, UK
Vargha A, Delaney HD (2000) A critique and improvement of the CL common language effect size statistics of McGraw and Wong. J Educ Behav Stat 25(2):101–132. https://doi.org/10.3102/1076998602500.2101
Vegas S, Apa C, Juristo N (2016) Crossover designs in software engineering experiments: benefits and perils. IEEE Trans Softw Eng 42(2):120–135. https://doi.org/10.1109/TSE.2015.2467378
Wilcox RR (2012) Introduction to robust estimation and hypothesis testing, 3rd edn. Elsevier Inc., Amsterdam

Lech Madeyski is an Associate Professor and Acting Head of the Department of Software Engineering at Wroclaw University of Science and Technology, Poland. He has been a Visiting Researcher at Keele University and Brunel University London, both in UK, and a Visiting Professor at Blekinge Institute of Technology, Sweden. His main research focus is on empirical (evidence-based) software engineering, data science in software engineering, reproducible research, robust statistical methods, software quality, mutation testing, agile methodologies and practices in software engineering. He is a founder of e-Informatica Software Engineering Journal and one of the founders of the International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE) series. He has published papers in prestigious journals including, e.g., IEEE Transactions on Software Engineering, Empirical Software Engineering, Information and Software Technology, Software Quality Journal, IET Software, Software Process: Improvement and Practice, Journal of Intelligent & Fuzzy Systems, Cybernetics and Systems, Foundations of Computing and Decision Sciences, and Statistics in Medicine. He has published (in Springer) a book "Test-Driven Development An Empirical Evaluation of Agile Practice" including statistical analyses and meta-analysis of several designed and conducted experiments. He is a member of ACM and a Senior Member of IEEE.

Barbara Kitchenham is Professor of Quantitative Software Engineering at Keele University in the UK. She has worked in software engineering for over 40 years both in industry and academia. She has published over 150 software engineering journal and conference papers. Her main research interest is software measurement and experimentation in the context of project management, quality control, risk management, and evaluation of software technologies. Her most recent research has focused on the application of evidence-based practice to software engineering.