Test Statistics for the Comparison of Means for Two Samples Which Include Both Paired Observations and Independent Observations.

Introduction

Hypothesis tests for the comparison of two population means, μ_1 and μ_2, with two samples of either independent observations or paired observations are well established. When the assumptions of the test are met, the independent samples t-test is the most powerful test for comparing means between two independent samples (Sawilowsky & Blair, 1992). Similarly, when the assumptions of the test are met, the paired samples t-test is the most powerful test for the comparison of means between two dependent samples (Zimmerman, 1997). If a paired design can avoid extraneous systematic bias, then paired designs are generally considered to be advantageous when contrasted with independent designs.

There are scenarios where, in a paired design, some observations may be missing. In the literature, this scenario is referred to as paired samples that are either "incomplete" (Ekbohm, 1976) or with "missing observations" (Bhoj, 1978). There are designs that do not have completely balanced pairings. Occasions where there may be two samples with both paired observations and independent observations include:

i) Two groups with some common element between both groups. For example, in education when comparing the average exam marks for two optional subjects, where some students take one of the two subjects and some students take both.

ii) Observations taken at two points in time, where the population membership changes over time but retains some common members. For example, an annual survey of employee satisfaction may include new employees that were unable to respond at time point one, employees that left after time point one, and employees that remained in employment throughout.

iii) When some natural pairing occurs. For example, in a survey taken comparing views of males and females, there will be some matched pairs "couples" and some independent samples "single".

The examples given above can be seen as part of the wider missing data framework. There is much literature on methods for dealing with missing data and the proposals in this paper do not detract from extensive research into the area. The simulations and discussion in this paper are done in the context of data missing completely at random (MCAR).

Two samples which include both paired and independent observations is referred to using varied terminology in the literature. The example scenarios outlined can be referred to as "partially paired data" (Samawi & Vogel, 2011). However, this terminology has connotations suggesting that the pairs themselves are not directly matched. Derrick et.al. (2015) suggest that appropriate terminology for the scenarios outlined gives reference to "partially overlapping samples". For work that has previously been done on a comparison of means when partially overlapping samples are present, "the partially overlapping samples framework….has been treated poorly in the literature" (Martínez-Camblor, Corral, & María de la Hera, 2012, p.77). In this paper, the term partially overlapping samples will be used to refer to scenarios where there are two samples with both paired and independent observations.

When partially overlapping samples exist, the goal remains to test the null hypothesis $H_0 : \mu_1 = \mu_2$. Standard approaches when faced with such a situation, are to perform the paired samples t-test, discarding the unpaired data, or alternatively perform the independent samples t-test, discarding the paired data (Looney & Jones, 2003). These approaches are wasteful and can result in a loss of power. The bias created with these approaches may be of concern. Other solutions proposed in a similar context are to perform the independent samples t-test on all observations ignoring the fact that there may be some pairs, or alternatively randomly pairing unpaired observations and performing the paired samples t-test (Bedeian & Feild, 2002). These methods distort Type I error rates (Zumbo, 2002) and fail to adequately reflect the design. This emphasises the need for research into a statistically valid

approach. A method of analysis which takes into account any pairing but does not lose the unpaired information would be beneficial.

One analytical approach is to separately perform both the paired samples t-test on the paired observations and the independent samples t-test on the independent observations. The results are then combined using Fisher's (1925) Chi-square method, or Stouffer's (1949) weighted z-test. These methods have issues with respect to the interpretation of the results. Other procedures weighting the paired and independent samples t-tests, for the partially overlapping samples scenario, have been proposed by Bhoj, (1978), Kim et. al. (2005), Martínez-Camblor, Corral, & María de la Hera (2012), and Samawi & Vogel (2011).

Looney & Jones (2003) proposed a statistic making reference to the z-distribution that uses all of the available data, without a complex weighting structure. Their corrected z-statistic is simple to compute and it directly tests the hypothesis $H_0 : \mu_1 = \mu_2$. They suggest that their test statistic is generally Type I error robust across the scenarios that they simulated. However, they only consider normally distributed data with a common variance of 1 and a total sample size of 50 observations. Therefore their simulation results are relatively limited, simulations across a wider range of parameters would help provide stronger conclusions. Mehrotra (2004) indicates that the solution provided by Looney & Jones (2003) may not be Type I error robust for small sample sizes.

Early literature for the partially overlapping samples framework focused on maximum likelihood estimates, when data are missing by accident rather than by design. Lin (1973) use maximum likelihood estimates for the specific case where data is missing from one of the two groups. Lin (1973) uses assumptions such as the variance ratio is known. Lin & Strivers (1974) apply maximum likelihood solutions to the more general case, but find that no single solution is applicable.

For normally distributed data, Ekbohm (1976) compared Lin & Strivers (1974) tests with similar proposals based on maximum likelihood estimators. Ekbohm (1976) found that maximum likelihood solutions do not always maintain Bradley's liberal Type I error robustness criteria. The results suggest that the maximum likelihood approaches are of little added value compared to standard methods. Furthermore the proposals by Ekbohm (1976) are complex mathematical procedures and are unlikely to be considered as a first choice solution in a practical environment.

A solution available in most standard software is to perform a mixed model using all of the available data. In a mixed model, effects are assessed using Restricted Maximum Likelihood estimators

"REML". Mehrotra (2004) indicates that for positive correlation, REML is Type I error robust and more powerful approach than that proposed by Looney & Jones (2003).

For small sample sizes, an intuitive solution to the comparison of means with partially overlapping samples, would be a test statistic derived using concepts similar to that of Zumbo (2002) so that all available data are used making reference to the t-distribution.

In this paper, two test statistics are proposed. The proposed solution for equal variances acts as a linear interpolation between the paired samples t-test and the independent samples t-test. The consensus in the literature is that Welch's test is more Type I error robust than the independent samples t-test, particularly with unequal variances and unequal samples sizes (Derrick, Toher & White, 2016; Fay & Proschan, 2010; Zimmerman & Zumbo, 2009). The proposed solution for unequal variances is a test which acts as a linear interpolation between the paired samples t-test and Welch's test.

Standard tests and the proposal by Looney & Jones (2003) are given below. This is followed by the definition of the presently proposed test statistics. A worked example provided using each of these test statistics and REML is provided. The Type I error rate and power for the test statistics and REML is then explored using simulation, for partially overlapping samples simulated from a Normal distribution.

Notation

Notation used in the definition of the test statistics is given in Table 1.

Table 1. Notation used in this paper.

$n_a =$	number of observations exclusive to Sample 1
$n_b =$	number of observations exclusive to Sample 2
$n_c =$	number of pairs
$n_1 =$	total number of observations in Sample 1 (i.e. $n_1 = n_a + n_c$)
$n_2 =$	total number of observations in Sample 2 (i.e. $n_2 = n_b + n_c$)
$\overline{X}_1 =$	mean of all observations in Sample 1
$\overline{X}_2 =$	mean of all observations in Sample 2
$\overline{X}_{\mathrm{a}} =$	mean of the independent observations in Sample 1
$\overline{X}_{\mathrm{b}} =$	mean of the independent observations in Sample 2
$\overline{X}_{1\mathrm{c}} =$	mean of the paired observations in Sample 1
$\overline{X}_{2\mathrm{c}} =$	mean of the paired observations in Sample 2
$S_1^2 =$	variance of all observations in Sample 1
$S_2^2 =$	variance of all observations in Sample 2
$S_a^2 =$	variance of the independent observations in Sample 1
$S_b^2 =$	variance of the independent observations in Sample 2
$S_{1c}^2 =$	variance of the paired observations in Sample 1
$S_{2c}^2 =$	variance of the paired observations in Sample 2
$S_{12} =$	covariance between the paired observations
$r =$	Pearson's correlation coefficient for the paired observations

All variances above are calculated using Bessel's correction, i.e. the sample variance with $n_i - 1$ degrees of freedom (see Kenney & Keeping 1951, p.161).

As standard notation, random variables are shown in upper case, and derived sample values are shown are in lower case.

Definition of Existing Test Statistics

Standard approaches for comparing two means making reference to the t-distribution are given below. These definitions follow the structural form given by Fradette et.al. (2003), adapted to the context of partially overlapping samples.

To perform the paired samples t-test, the independent observations are discarded so that

$$T_1 = \frac{\overline{X}_{1c} - \overline{X}_{2c}}{\sqrt{\frac{S_{1c}^2}{n_c} + \frac{S_{2c}^2}{n_c} - 2r\left(\frac{S_{1c}S_{2c}}{n_c}\right)}}$$

The statistic T_1 is referenced against the t-distribution with $\nu_1 = n_c - 1$ degrees of freedom.

To perform the independent samples t-test, the paired observations are discarded so that

$$T_2 = \frac{\overline{X}_a - \overline{X}_b}{S_p\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} \text{ where } S_p = \sqrt{\frac{(n_a - 1)S_a^2 + (n_b - 1)S_b^2}{(n_a - 1) + (n_b - 1)}}$$

The statistic T_2 is referenced against the t-distribution with $\nu_2 = n_a + n_b - 2$ degrees of freedom.

To perform Welch's test, the paired observations are discarded so that

$$T_3 = \frac{\overline{X}_a - \overline{X}_b}{\sqrt{\frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}}}$$

The statistic T_3 is referenced against the t-distribution with degrees of freedom approximated by

$$\nu_3 = \frac{\left(\frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}\right)^2}{\left(\frac{S_a^2}{n_a}\right)^2 / (n_a - 1) + \left(\frac{S_b^2}{n_b}\right)^2 / (n_b - 1)}$$

For large sample sizes, the test statistic for partially overlapping samples proposed by Looney & Jones (2003) is

$$Z_{\text{corrected}} = \frac{\overline{\text{X}}_1 - \overline{\text{X}}_2}{\sqrt{\frac{S_1^2}{n_a + n_c} + \frac{S_2^2}{n_b + n_c} - \frac{(2n_c)S_{12}}{(n_a + n_c)(n_b + n_c)}}}$$

The statistic $Z_{\text{corrected}}$ is referenced against the standard Normal distribution. In the extremes of $n_a = n_b = 0$, or $n_c = 0$, $Z_{\text{corrected}}$ defaults to the paired samples z-statistic and the independent samples z-statistic respectively.

Definition of Proposed Test Statistics

Two new t-statistics are proposed; T_{new1}, assuming equal variances, and T_{new2}, when equal variances cannot be assumed. The test statistics are constructed as the difference between two means taking into account the covariance structure. The numerator is the difference between the means of the two samples and the denominator is a measure of the standard error of this difference. Thus the test statistics proposed here are directly testing the hypothesis $H_0 : \mu_1 = \mu_2$.

The test statistic T_{new1} is derived so that in the extremes of $n_a = n_b = 0$ or $n_c = 0$, T_{new1} defaults to T_1 or T_2 respectively, thus

$$T_{\text{new1}} = \frac{\overline{\text{X}}_1 - \overline{\text{X}}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2r\left(\frac{n_c}{n_1 n_2}\right)}} \text{ where } S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

The test statistic T_{new1} is referenced against the t-distribution with degrees of freedom derived by linear interpolation between v_1 and v_2 so that: $v_{\text{new1}} = (n_c - 1) + \left(\frac{n_a + n_b + n_c - 1}{n_a + n_b + 2n_c}\right)(n_a + n_b)$.

In the extremes, when $n_a = n_b = 0$, v_{new1} defaults to v_1; or when $n_c = 0$, v_{new1} defaults to v_2.

Given the superior Type I error robustness of Welch's test when variances are not equal, a test statistic is derived making reference to Welch's approximate degrees of freedom. This test statistic makes use of the sample variances, S_1^2 and S_2^2. The test statistic T_{new2} is derived so that in the extremes of $n_a = n_b = 0$ or $n_c = 0$, T_{new2} defaults to T_1 or T_3 respectively, thus

$$T_{\text{new2}} = \frac{\overline{\text{X}}_1 - \overline{\text{X}}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2r\left(\frac{S_1 S_2 n_c}{n_1 n_2}\right)}}$$

The test statistic T_{new2} is referenced against the t-distribution with degrees of freedom derived as a linear interpolation between v_1 and v_3 so that

$$v_{\text{new2}} = (n_c - 1) + \left(\frac{\gamma - n_c + 1}{n_a + n_b + 2n_c}\right)(n_a + n_b) \text{ where } \gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 / (n_1 - 1) + \left(\frac{S_2^2}{n_2}\right)^2 / (n_2 - 1)}$$

In the extremes, when $n_a = n_b = 0$, v_{new2} defaults to v_1; or when $n_c = 0$, v_{new2} defaults to v_3.

Note that the proposed statistics, T_{new1} and T_{new2}, use all available observations in the respective variance calculations. The statistic $Z_{\text{corrected}}$ only uses the paired observations in the calculation of covariance.

Worked Example

An applied example is given to demonstrate the calculation of each of the test statistics defined. In education, for credit towards an undergraduate Statistics course, students may take optional modules in either Mathematical Statistics, or Operational Research, or both. The programme leader is interested whether the exam marks for the two optional modules differ. The exam marks attained for a single semester are given in Table 2.

Table 2. Exam marks for Students studying on an undergraduate Statistics course.

Student	Mathematical Statistics	Operational Research
1	73	72
2	82	
3	74	89
4	59	78
5	49	64
6		83
7	42	42
8	71	76
9		79
10	39	89
11		67
12		82
13		85
14		92
15	59	63
16	85	

As per standard notion, the derived sample values are given in lower case. In the calculation of the test statistics, $\bar{x}_1 = 63.300,\ \bar{x}_2 = 75.786,\ s_1^2 = 263.789,\ s_2^2 = 179.874,\ n_a = 2,\ n_b = 6,\ n_c = 8,\ n_1 = 10,\ n_2 = 14,\ v_1 = 7,\ v_2 = 6,\ v_3 = 6,\ \gamma = 17.095,\ v_{\text{new1}} = 12,\ v_{\text{new2}} = 10.365,\ r = 0.366,\ s_{12} = 78.679.$

For the REML analysis, a mixed model is performed with “Module” as a repeated measures fixed effect and “Student” as a random effect. Table 3 gives the calculated test statistics, degrees of freedom and corresponding p-values.

Table 3. Test statistic values and resulting p-values (two-sided test).

	T_1	T_2	T_3	$Z_{\text{corrected}}$	REML	T_{new1}	T_{new2}
estimate of mean difference	-13.375	2.167	2.167	-12.486	-12.517	-12.486	-12.486
t-value	-2.283	0.350	0.582	-2.271	-2.520	-2.370	-2.276
degrees of freedom	7.000	6.000	6.000		11.765	12.000	10.365
p-value	0.056	0.739	0.579	0.023	0.027	0.035	0.045

With the exception of REML, the estimates of the mean difference are simply the difference in the means of the two samples, based on the observations used in the calculation. It can quickly be seen that the conclusions differ depending on the test used. It is of note that only the tests using all of the available data result in the rejection of the null hypothesis at $\alpha_{\text{nominal}} = 0.05$. Also note that the results of the paired samples t-test and the independent samples t-test have sample effects in different directions. This is only one specific example given for illustrative purposes, investigation is required into the power of the test statistics over a wide range of scenarios. Conclusions based on the proposed tests cannot be made without a thorough investigation into their Type I error robustness.

Simulation Design

Under normality, Monte-Carlo methods are used to investigate the Type I error robustness of the defined test statistics and REML. Power should only be used to compare tests when their Type I error rates are equal (Zimmerman & Zumbo, 1993). Monte-Carlo methods are used to explore the power for the tests that are Type I error robust under normality.

Unbalanced designs are frequent in psychology (Sawilowski & Hillman, 1982), thus a comprehensive range of values for n_a, n_b and n_c are simulated. These values offer an extension to the work done by Looney & Jones (2003). Given the identification of separate test statistics for equal and unequal variances, multiple population variance parameters $\{\sigma_1^2,\ \sigma_2^2\}$ are considered. Correlation has an impact on Type I error and power for the paired samples t-test (Fradette et. al., 2003), hence a range of correlations $\{\rho\}$ between two normal populations are considered. Correlated normal variates are obtained as per Kenney & Keeping (1951). A total of 10,000 replicates of each of the scenarios in Table 4 are performed in a factorial design.

All simulations are performed in R version 3.1.2. For the mixed model approach utilising REML, the R package lme4 is used. Corresponding p-values are calculated using the Satterthwaite approximation adopted by SAS using the R package lmerTest (Goodnight, 1976).

For each set of 10,000 p-values, the proportion of times the null hypothesis is rejected, for a two sided test with $\alpha_{\text{nominal}} = 0.05$ is calculated.

Table 4. Summary of simulation parameters

Parameter	Values
μ_1	0
μ_2	0 (under H_0)
	0.5 (under H_1)
σ_1^2	1, 2, 4, 8
σ_2^2	1, 2, 4, 8
n_a	5, 10, 30, 50, 100, 500
n_b	5, 10, 30, 50, 100, 500
n_c	5, 10, 30, 50, 100, 500
ρ	-0.75, -0.50, -0.25, 0.00, 0.25, 0.50, 0.75

Type I Error Robustness

For each of the test statistics, Type I error robustness is assessed against Bradley's (1978) liberal criteria. This criteria it is widely used in many studies analysing the validity of t-tests and their adaptions. Bradley's (1978) liberal criteria states that the Type I error rate α should be within $\alpha_{\text{nominal}} \pm 0.5\alpha_{\text{nominal}}$. For $\alpha_{\text{nominal}} = 0.05$, Bradley's liberal interval is [0.025, 0.075].

Type I error robustness is firstly assessed under the condition of equal variances. Under the null hypothesis, 10,000 replicates are obtained for the $4 \times 6 \times 6 \times 6 \times 7 = 6{,}048$ scenarios where $\sigma_1^2 = \sigma_2^2$. Figure 1 shows the Type I error rates for each of the test statistics under equal variances for normally distributed data.

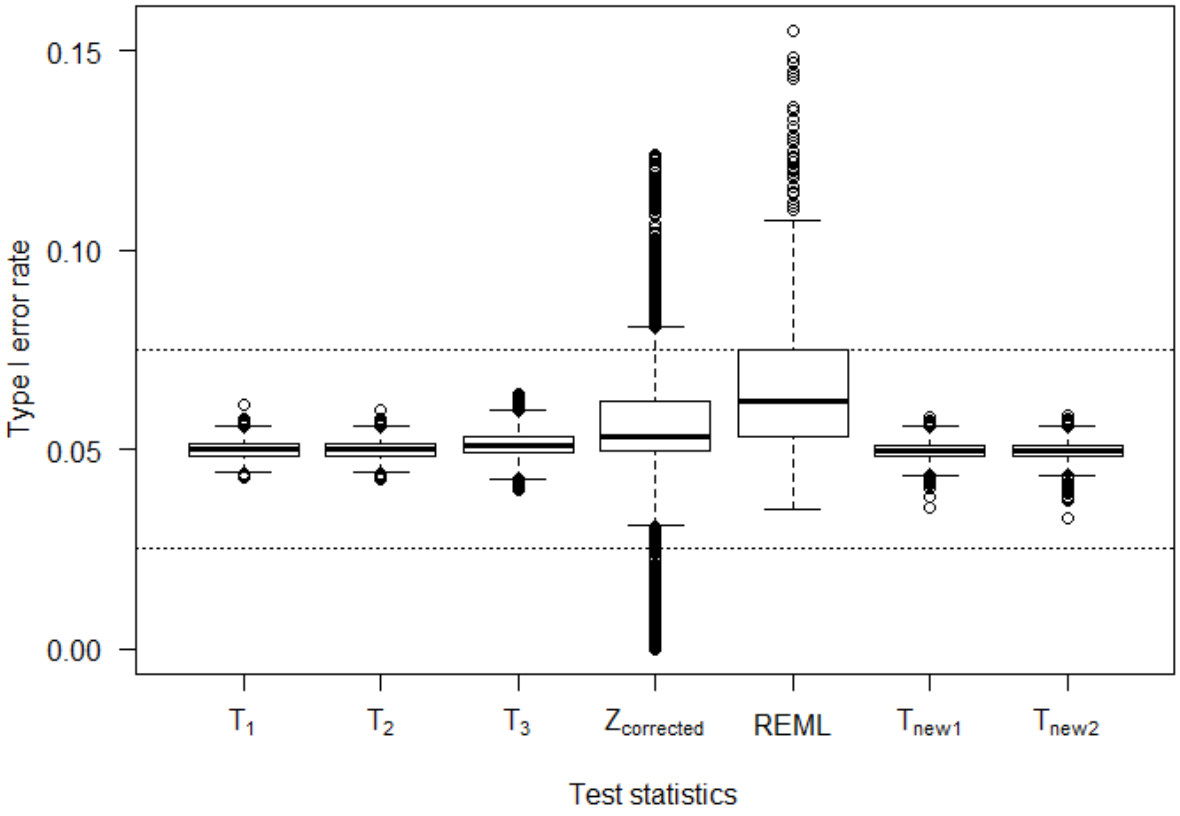

Figure 1. Type I error rates where $\sigma_1^2 = \sigma_2^2$, reference lines show Bradley's (1978) liberal criteria.

Figure 1 indicates that when variances are equal, the statistics T_1, T_2, T_3, T_{new1} and T_{new2} remain within Bradley's liberal Type I error robustness criteria throughout the entire simulation design. The statistic $Z_{\text{corrected}}$ is not Type I error robust, thus confirming the smaller simulation findings of Mehotra (2004). Figure 1 also shows that REML is not Type I error robust throughout the entire simulation design. A review of our results shows that for REML the scenarios that are outside the range of liberal Type I error robustness are predominantly those that have negative correlation, and some where zero correlation is specified. Given that negative correlation is rare in a practical environment, the REML procedure is not necessarily unjustified.

Type I error robustness is assessed under the condition of unequal variances. Under the null hypothesis, 10,000 replicates were obtained for the $4 \times 3 \times 6 \times 6 \times 6 \times 7 = 18{,}144$ scenarios where

$\sigma_1^2 \neq \sigma_2^2$. For assessment against Bradley's (1978) liberal criteria, Figure 2 shows the Type I error rates for unequal variances for normally distributed data.

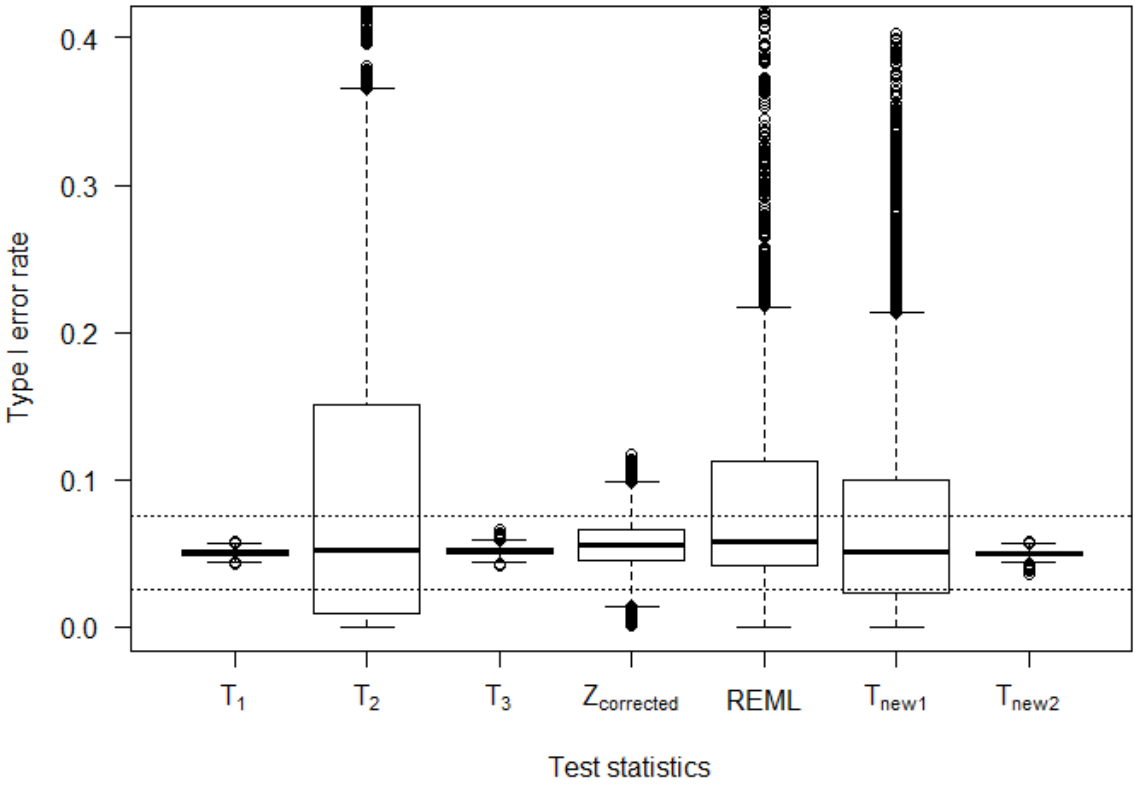

Figure 2. Type I error rates when $\sigma_1^2 \neq \sigma_2^2$, reference lines show Bradley's (1978) liberal criteria.

It can be seen from Figure 2 that the statistics defined using a pooled standard deviation T_2 and T_{new1}, do not provide Type I error robust solutions when equal variances cannot be assumed. The statistics T_1, T_3 and T_{new2} retain their Type I error robustness under unequal variances throughout all conditions simulated.

The statistic $Z_{\text{corrected}}$ maintains similar Type I error rates under equal and unequal variances. The statistic $Z_{\text{corrected}}$ was only designed to be used in the case of equal variances. For unequal variances, we observe that the statistic $Z_{\text{corrected}}$ results in an unacceptable amount of false positives when $\rho \leq 0.25$ or max $\{ n_a, n_b, n_c \}$ - min$\{ n_a, n_b, n_c \}$ is large. In addition, the statistic $Z_{\text{corrected}}$ is

conservative when ρ is large and positive. The largest observed deviations from Type I error robustness for REML are when $\rho \leq 0$ or max $\{ n_a, n_b, n_c \}$ - min$\{ n_a, n_b, n_c \}$ is large. Further insight to the Type I error rates for REML can be seen in Figure 3 showing observed p-values against expected p-values from a uniform distribution.

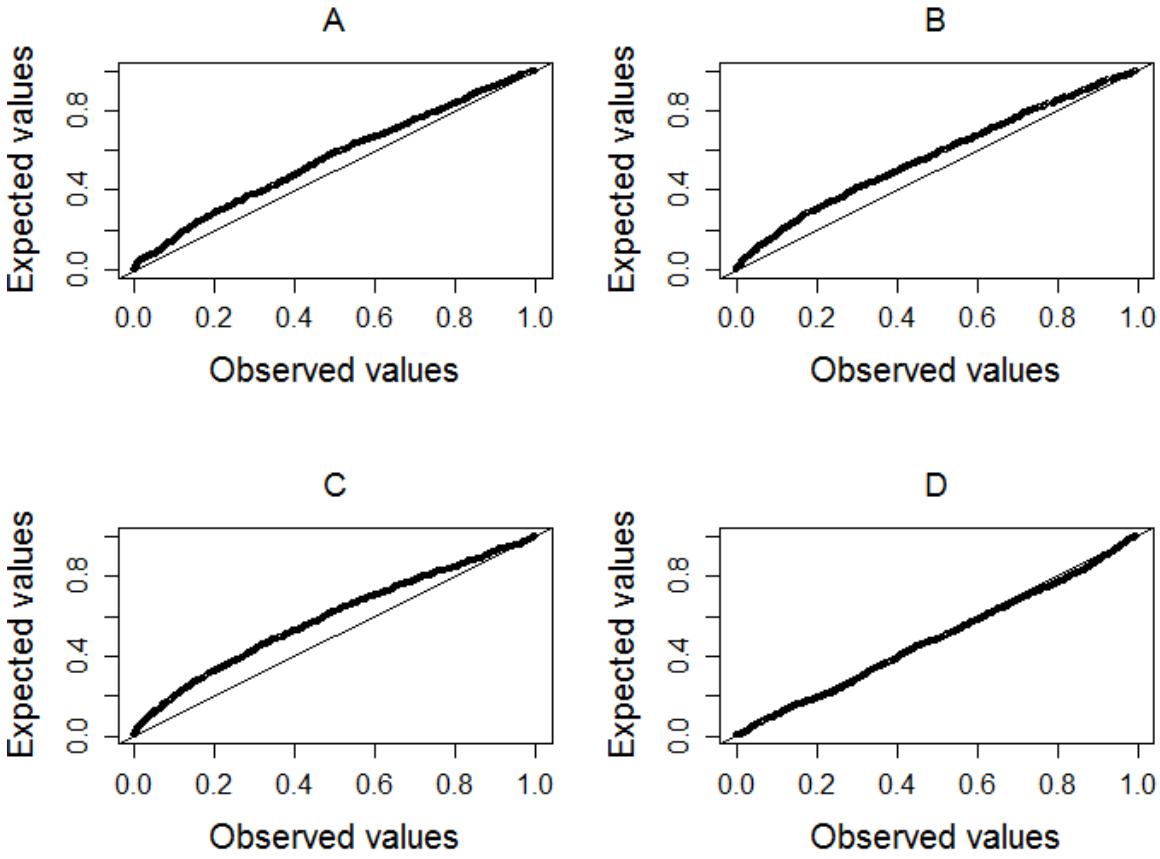

Figure 3. P-P plots for simulated p-values using REML procedure. Selected parameter combinations ($n_a, n_b, n_c, \sigma_1^2, \sigma_2^2, \rho$) are as follows; A = (5,5,5,1,1,-0.75), B = (5,10,5,8,1,0), C = (5,10,5,8,1,0.5), D = (10,5,5,8,1,0.5).

If the null hypothesis is true, for any given set of parameters the p-values should be uniformly distributed. Figure 3 gives indicative parameter combinations where the p-values are not uniformly distributed when applying a mixed model assessed using REML. It can be seen that REML is not Type I error robust when the correlation is negative. In addition, caution should be exercised if using REML when the larger variance is associated with the smaller sample size. REML maintains Type I

error robustness for positive correlation and equal variances or when the larger sample size is associated with the larger variance.

Power of Type I Error Robust Tests under Equal Variances

The test statistics that do not fail to maintain Bradley's Type I error liberal robustness criteria are assessed under H_1. REML is included in the comparisons for $\rho \geq 0$. The power of the test statistics are assessed where $\sigma_1^2 = \sigma_2^2 = 1$, followed by an assessment of the power of the test statistics where $\sigma_1^2 > 1$ and $\sigma_2^2 = 1$.

Table 5 shows the power of T_1, T_2, T_3, T_{new1}, T_{new2} and REML, averaged over all sample size combinations where $\sigma_1^2 = \sigma_2^2 = 1$.

Table 5. Power of Type I error robust test statistics, $\sigma_1^2 = \sigma_2^2 = 1$, $\alpha = 0.05$, $\mu_2 - \mu_1 = 0.5$.

	ρ	T_1	T_2	T_3	T_{new1}	T_{new2}	REML
$n_a = n_b$	0.75	0.785	0.567	0.565	0.887	0.886	0.922
	0.50	0.687	0.567	0.565	0.865	0.864	0.880
	0.25	0.614	0.567	0.565	0.842	0.841	0.851
	0	0.558	0.567	0.565	0.818	0.818	0.829
	< 0	0.481	0.567	0.565	0.778	0.778	-
$n_a \neq n_b$	0.75	0.784	0.455	0.433	0.855	0.847	0.907
	0.50	0.687	0.455	0.433	0.840	0.832	0.861
	0.25	0.615	0.455	0.433	0.823	0.816	0.832
	0	0.559	0.455	0.433	0.806	0.799	0.816
	< 0	0.482	0.455	0.433	0.774	0.766	-

Table 5 shows that REML and the test statistics proposed in this paper, T_{new1} and T_{new2}, are more powerful than standard approaches, T_1, T_2 and T_3, when variances are equal. Consistent with the paired samples t-test, T_1, the power of T_{new1} and T_{new2} is relatively lower when there is zero or negative correlation between the two populations. Similar to contrasts of the independent samples t-test, T_2, with Welch's test, T_3, for equal variances but unequal sample sizes, T_{new1} is marginally more

powerful than T_{new2}, but not to any practical extent. For each of the tests statistics making use of paired data, as the correlation between the paired samples increases, the power increases.

As the correlation between the paired samples increases, the power advantage of the proposed test statistics relative to the paired samples t-test becomes smaller. Therefore the proposed statistics T_{new1} and T_{new2} may be especially useful when the correlation between the two populations is small.

To show the relative increase in power for varying sample sizes, Figure 4 shows the power for selected test statistics for small-medium sample sizes, averaged across the simulation design for equal variances.

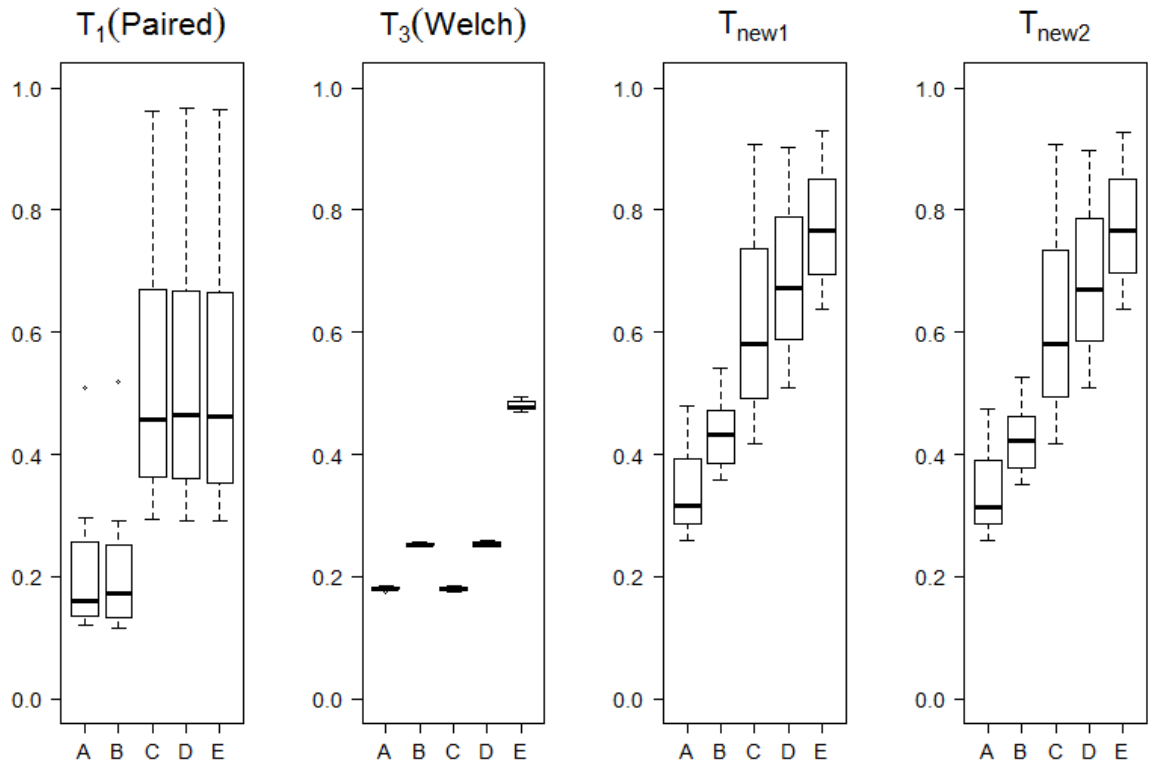

Figure 4. Power for Type I error robust test statistics, averaged across all values of ρ where $\sigma_1^2 = \sigma_2^2$ and $\mu_2 - \mu_1 = 0.5$. The sample sizes (n_a, n_b, n_c) are as follows; $A = (10,10,10)$, $B = (10,30,10)$, $C = (10,10,30)$, $D = (10,30,30)$, $E = (30,30,30)$.

From Figure 4 it can be seen that for small–medium sample sizes, the power of the proposed test statistics T_{new1} and T_{new2} is superior to standard test statistics.

Power of Type I Error Robust Rests under Unequal Variances

For the Type I error robust test statistics under unequal variances, Table 6 shows the power of T_1, T_3, T_{new2} and REML, averaged over the simulation design where $\mu_2 - \mu_1 = 0.5$.

Table 6. Power of Type I error robust test statistics where $\sigma_1^2 > 1$, $\sigma_2^2 = 1, \alpha = 0.05$, $\mu_2 - \mu_1 = 0.5$. Within this table, $n_a > n_b$ represents the larger variance associated with the larger sample size, and $n_a < n_b$ represents the larger variance associated with the smaller sample size.

	ρ	T_1	T_3	T_{new2}	REML
$n_a = n_b$	0.75	0.555	0.393	0.692	0.645
	0.50	0.481	0.393	0.665	0.588
	0.25	0.429	0.393	0.640	0.545
	0	0.391	0.393	0.619	0.515
	< 0	0.341	0.393	0.582	-
$n_a > n_b$	0.75	0.555	0.351	0.715	0.589
	0.50	0.481	0.351	0.688	0.508
	0.25	0.429	0.351	0.665	0.459
	0	0.391	0.351	0.642	0.422
	< 0	0.341	0.351	0.604	-
$n_a < n_b$	0.75	0.555	0.213	0.559	0.693
	0.50	0.481	0.213	0.539	0.649
	0.25	0.429	0.213	0.522	0.620
	0	0.391	0.213	0.507	0.603
	< 0	0.341	0.213	0.480	-

Table 6 shows that T_{new2} has superior power properties to both T_1 and T_3 when variances are not equal. In common with the performance of Welch's test for independent samples, T_3, the power of T_{new2} is higher when the larger variance is associated with the larger sample size. In common with the performance of the paired samples t-test, T_1, the power of T_{new2} is relatively lower when there is zero or negative correlation between the two populations.

The apparent power gain for REML when the larger variance is associated with the larger sample size, can be explained by the pattern in the Type I error rates. REML follows a similar pattern to the independent samples t-test, which is liberal when the larger variance is associated with the larger sample size, thus giving the perception of higher power.

To show the relative increase in power for varying sample sizes, Figure 5 shows the power for selected test statistics for small-medium sample sizes, averaged across the simulation design for unequal variances.

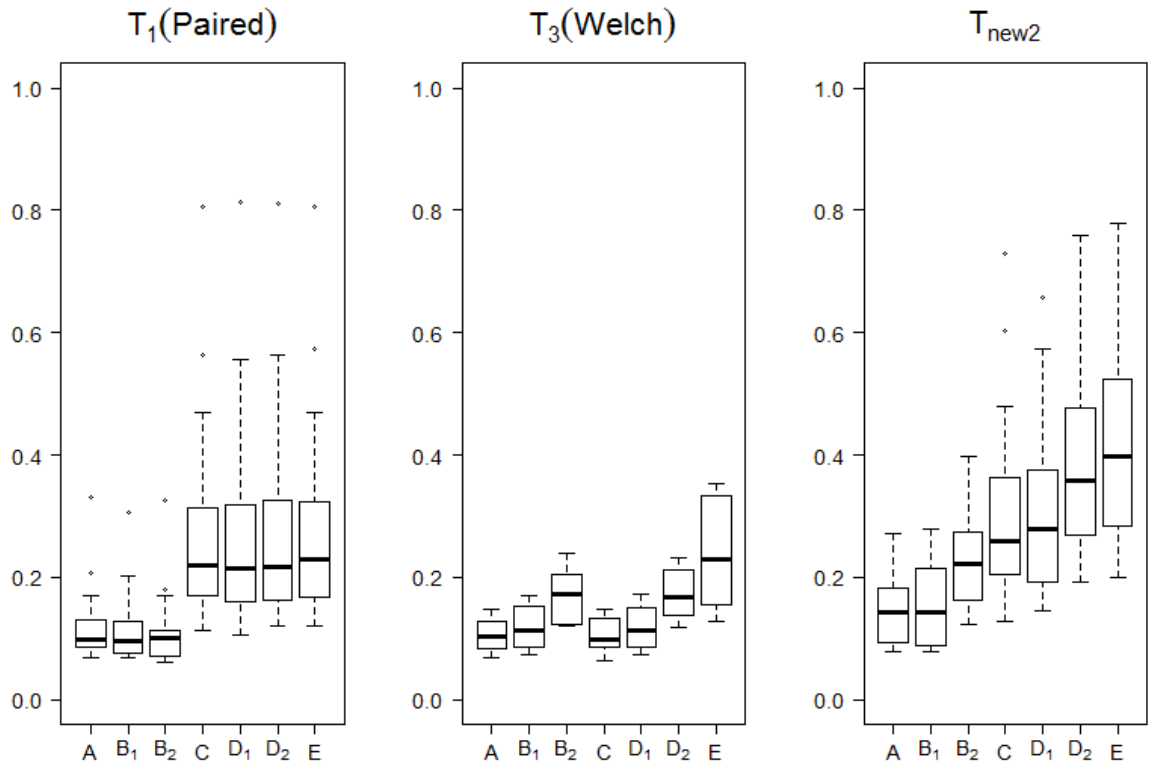

Figure 5. Power for Type I error robust test statistics, $\sigma_1^2 > \sigma_2^2$ and $\mu_2 - \mu_1 = 0.5$. The sample sizes (n_a, n_b, n_c) are as follows; $A = (10,10,10)$, $B_1 = (10,30,10)$, $B_2 = (30,10,10)$, $C = (10,10,30)$, $D_1 = (10,30,30)$, $D_2 = (30,10,30)$, $E = (30,30,30)$.

Figure 5 shows a relative power advantage when the larger variance is associated with the larger sample size, as per B_2 and D_2. A comparison of Figure 4 and Figure 5 shows that for small-medium sample sizes, power is adversely effected for all test statistics when variances are not equal.

Discussion

The statistic T_{new2} is Type I error robust across all conditions simulated under normality. The greater power observed for T_{new1}, compared to T_{new2}, under equal variances, is likely to be of negligible consequence in a practical environment. This is in line with empirical evidence for the performance of Welch's test, when only independent samples are present, which leads to many observers recommending the routine use of Welch's test under normality (e.g. Ruxton, 2006).

The Type I error rates and power of T_{new2} follow the properties of its counterparts, T_1 and T_3. Thus T_{new2} can be seen as a trade-off between the paired sample t-test and Welch's test, with the advantage of increased power across all conditions, due to using all available data.

The partially overlapping samples scenarios identified in this paper could be considered as part of the missing data framework and all simulations have been performed under the assumption of MCAR.

The statistics proposed in this paper form less computationally intensive competitors to REML. The REML procedure does not directly calculate the difference between the two sample means, in a practical environment this makes its results hard to interpret. The statistics proposed in this paper will also far more easily lend themselves to the development of non-parametric tests.

Conclusion

A commonly occurring scenario when comparing two means is a combination of paired observations and independent observations in both samples, this scenario is referred to as partially overlapping

samples. Standard procedures for analysing partially overlapping samples involve discarding observations and performing either the paired samples t-test, or the independent samples t-test, or Welch's test. These approaches are less than desirable. In this paper, two new test statistics making reference to the t-distribution are introduced and explored under a comprehensive set of parameters, for normally distributed data. Under equal variances, T_{new1} and T_{new2} are Type I error robust. In addition they are more powerful than standard Type I error robust approaches considered in this paper. When variances are equal, there is a slight power advantage of using T_{new1} over T_{new2}, particularly when sample sizes are not equal. Under unequal variances, T_{new2} is the most powerful Type I error robust statistic considered in this paper. We recommend that when faced with a research problem involving partially overlapping samples and MCAR can be reasonably assumed, the statistic T_{new1} could be used when it is known that variances are equal. Otherwise under the same conditions when equal variances cannot be assumed the statistic T_{new2} could be used.

A mixed model procedure using REML is not fully Type I error robust. In those scenarios in which this procedure is Type I error robust, the power is similar to that of T_{new1} and T_{new2}.

The proposed test statistics for partially overlapping samples provide a real alternative method for analysis for normally distributed data, which could also be used for the formation of confidence intervals for the true difference in two means.

References

Bedeian, A. G., & Feild, H. S. (2002). Assessing group change under conditions of anonymity and overlapping samples. *Nursing research, 51*(1), 63-65.

Bhoj, D. (1978). Testing equality of means of correlated variates with missing observations on both responses. *Biometrika, 65*(1), 225-228. doi: 10.1093/biomet/65.1.225

Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Derrick, B., Dobson-McKittrick, A., Toher, D., & White P. (2015). Test statistics for comparing two proportions with partially overlapping samples. *Journal of Applied Quantitative Methods, 10*(3)

Derrick, B., Toher, D., & White, P. (2016). Why Welch's test is Type I error robust. *The Quantitative Methods for Psychology, 12*(1), 30-38. doi: 10.20982/tqmp.12.1.p030

Ekbohm, G. (1976). Comparing Means in the Paired Case with Incomplete Data on Both Responses, *Biometrika, 63*(2), 299-304. doi: 10.1093/biomet/63.2.299

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys, 4,* 1. doi: 10.1214/09-SS051

Fisher, R. A. (1925). Statistical methods for research workers. Genesis Publishing Pvt Ltd.

Fradette, K., Keselman, H. J., Lix, L., Algina, J., & Wilcox, R. (2003). Conventional and Robust Paired and Independent Samples t-tests: Type I Error and Power Rates, *Journal of Modern Applied Statistical Methods, 2*(2), 481-496.

Kenney, J. F., & Keeping, E. S. (1951). *Mathematics of statistics*, Pt. 2, 2nd ed. Princeton, NJ: Van Nostrand.

Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y., & Chung, H. C. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics, 21*(4), 517-528. doi: 10.1093/bioinformatics/bti029

Lin, P. E. (1973). Procedures for testing the difference of means with incomplete data. *Journal of the American Statistical Association, 68*(343), 699-703.

Lin, P. E., & Strivers L. (1974). Difference of Means with Incomplete Data. *Biometrika, 61*(2), 325-334. doi: 10.1093/biomet/61.2.325

Looney, S., & Jones, P. (2003). A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in medicine, 22*, 1601-1610. doi:10.1002/sim.1514

Martínez-Camblor, P., Corral, N., & María de la Hera, J. (2013). Hypothesis test for paired samples in the presence of missing data. *Journal of Applied Statistics*, *40*(1), 76-87. doi: 10.1080/02664763.2012.734795

Mehrotra, D. (2004). Letter to the editor, a method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in medicine*, *23*(7), 1179–1180. doi: 10.1002/sim.1693

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org. 2014; version 3.1.2.

Ruxton., G. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology, 17*(4), 688. doi: 10.1093/beheco/ark016

Goodnight, J. H. (1976) General Linear Models Procedure. S.A.S. Institute. Inc.

Samawi, H. M., & Vogel, R. (2011). Tests of homogeneity for partially matched-pairs data. *Statistical Methodology, 8*(3), 304-313. doi: 10.1016/j.stamet.2011.01.002

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological bulletin*, *111*(2), 352. doi: 10.1037/0033-2909.111.2.352

Sawilowsky, S. S., & Hillman, S. B. (1992). Power of the independent samples t-test under a prevalent psychometric measure distribution, *Journal of Consulting and Clinical Psychology, 60*(2), 240-243. doi: 10.1037/0022-006X.60.2.240

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). The American soldier: adjustment during army life. *Studies in social psychology in World War II*, 1.

Zimmerman, D . W. (1997). A note on the interpretation of the paired samples, *Journal of educational and behavioral statistics, 22*(3), 349 – 360. doi:10.3102/10769986022003349

Zimmerman, D. W., & Zumbo, B. D. (1993). Significance testing of correlation using scores, ranks, and modified ranks. *Educational and psychological measurement, 53*(4), 897-904.

Zimmerman, D. W., & Zumbo, B. D. (2009). Hazards in choosing between pooled and separate-variances t tests. *Psicológica: Revista de metodología y psicología experimental, 30*(2), 371-390.

Zumbo, B. D. (2002). An adaptive inference strategy: The case of auditory data. *Journal of Modern Applied Statistical Methods, 1*(1), 60-68. doi: 10.22237/jmasm/1020255000