Psychologica Belgica
2012, 52/4, 387-406
DOI: http://dx.doi.org/10.5334/pb-52-4-387

RT4WIN: A WINDOWS-BASED PROGRAM FOR RANDOMIZATION TESTS

Ming Huo & Patrick Onghena*
KU Leuven

A randomization test (RndT) is a statistical significance test for which the validity is based on the random assignment of experimental units in a designed experiment. In a random sampling setting, it can also be applied in a very general way because its validity does not rely on distributional assumptions, homogeneity of variances, or independence of errors. However, the use of RndTs did not receive much attention in applied research because RndTs rely on computationally intensive algorithms and most popular and common statistical software packages do not provide facilities to easily perform randomization tests. In order to fill this gap, we present a software package, RT4Win. Unlike most stand-alone software and programs for RndTs, RT4Win is a fast Windows-based program with a user-friendly interface. It provides a facility to carry out RndTs in a series of experimental designs, for both systematic and Monte Carlo data partition methods. The program is free of charge and available upon request from the authors.

In most experiments in psychology, as in other scientific disciplines, random assignment rather than random sampling is the norm, which makes the use of parametric statistical procedures, such as standard *t* or *F* tests, problematic. A review of 252 studies published in five high impact journals revealed that experimental groups were constructed by random assignment in 96% of cases and by random sampling in only 4% (Ludbrook & Dudley, 1998). Ludbrook and Dudley (1998) also disturbingly confirm that in 84% of the randomized experiments common parametric *t* or *F* tests were applied, while more appropriate randomization tests (RndTs) should be the first choice for the analysis of experiments in the absence of random sampling (Edgington, 1966; Hunter & May, 2003; Pitman, 1937a; Pitman, 1937b; Pitman, 1938). Recent results suggest that the situation is only slowly improving (Anderson, 2001; Zieffler, Harring, & Long, 2011).

We can only guess at the reasons for this negligence or preference, but it is very plausible that critical events in the history of statistics and computer science, editorial policy, the teaching of statistics, and the availability of soft-

* Ming Huo and Patrick Onghena, Faculty of Psychology and Educational Sciences, Methodology of Educational Sciences Research Group, KU Leuven.
This study was partly funded by Grant G.0624.07 from the Research Foundation-Flanders. The authors would like to thank two anonymous reviewers for their valuable comments and suggestions, many of which were incorporated in the revised manuscript.
Correspondence concerning this article should be addressed to Ming Huo, Andreas Vesaliusstraat 2 box 3762, B-3000 Leuven. E-mail: hoffmanhm@gmail.com

ware shaped the behaviour of our fellow scientists (Cobb, 2007; Noreen, 1989; Ramsey & Schafer, 2002; Stigler, 1992). In this article we will mainly focus on the software issue and present a user-friendly and free program to remove at least one obstacle to provide to RndTs the popularity they deserve.

RndTs are formally defined as statistical significance tests for which the validity is based on the random assignment of experimental units in a designed experiment (Edgington & Onghena, 2007). This means that their *p* value can be derived in a valid way, just by taking into account the random assignment procedure that was actually used in the experiment. It also means that statistical tests that are not based on random assignment (e.g., in nonexperimental studies or nonrandomized experiments) are not called RndTs. The way to derive the *p* value in a valid way, just by taking into account the random assignment procedure that was actually used in the experiment, will be explained with an example below.

While there is some divergence in the literature on the terminology regarding RndTs and *permutation tests*, we follow the terminological distinction emphasised by Cox and Hinkley (1974), Kempthorne and Doerfler (1969), and Zieffler et al. (2011). RndTs are tests based on random assignment and do not have to involve data permutations in the combinatorial sense of the term (see e.g., Onghena, 1992; Onghena & Edgington, 1994; Onghena & Edgington, 2005) and permutation tests are distribution-free statistical tests for which the validity can be based on other arguments (e.g., random sampling from identical distributions) and for which the computation of the *p* value involves repeated data permutations (Good, 2000; Manly, 1997; Mielke & Berry, 2007). In many applications, however, the terminological distinction does not matter for the actual computations, and algorithms for RndTs can be used to compute permutation test *p* values, and vice versa (Edgington & Onghena, 2007).

Another important distinction is between randomization tests, parametric tests, and the traditional nonparametric rank tests. Unlike the parametric tests, which rely on random sampling from a population following a specified distribution or on a large sample approximation for samples of a broader class of distributions, RndTs do not assume any specific error distribution or large samples, and, unlike nonparametric rank tests, which transform the original data into ranks, RndTs can be applied on the original data and therefore can use all the information in the data (Edgington & Onghena, 2007; Lehmann & D'Abrera, 2006).

RndTs were proposed in the early twentieth century, but were not considered practical until much later with the emergence of cheap and fast computers. However, fast and user-friendly software is not commonly available to the average researcher. It is this gap that we want to fill with the presentation of RT4Win, a free Windows-based computer program for randomization

tests. This paper begins with an introduction of the rationale of RndTs, clarifying their difference with classical statistical methods, followed by two examples involving a one-way ANOVA design and one example involving a repeated-measures design. Afterwards, we discuss the advantages and disadvantages of RndTs, other available RndTs software, and the assets of RT4Win. Finally, we give some recommendations for future software development.

The rationale of randomization tests

Randomization tests are applicable for experimental designs involving the random assignment of available experimental units (e.g., human participants) into treatment groups. This random assignment is the only stochastic process that needs to be present for the valid application of RndTs. To help illustrate the rationale of RndTs, a practical example is provided.

Example

Imagine that a new treatment for back pain is being compared to a standard treatment by observing the recovery times (in days) of the patients on each treatment. The researcher randomly assigned three patients to take the new treatment and three others to take the standard treatment. After receiving the treatments, the recovery times of the six patients have been measured, which are shown in the first row of Table 1. Our question is whether the lower mean recovery times for the new treatment group indicating the new treatment is more effective or the difference of recovery times observed is just due to the random assignment of the six patients into the two treatments?

In order to answer the above question, we can first set up a null hypothesis which says that any difference of recovery times between the two treatments is purely attributed to chance. If there is sufficient evidence against the null hypothesis, then we should reject the null hypothesis in favour of the alternative hypothesis that the new treatment will lead to smaller recovery times. The null and the alternative hypothesis can be expressed as

H_0 : The new and standard treatment will lead to the same recovery times (in days) for the given patients;

H_1 : The recovery times would decrease if the patients would take the new treatment (for a one-tailed test).

To explore the difference between the two treatments, we need to choose an appropriate test statistic. There are several options for comparing the two treatments (e.g., the mean or median difference of recovery times between the two groups, or a two-sample *t*-statistic). In this case, we choose the mean difference of recovery times between the two groups, D, as the test statistic of

Table 1

All possible data partitions of six recovery times (days) to two treatment groups of sizes n=3 in each group

	Data Partitions of Recovery Times (days)						
No.	**New Treatment**			**Standard Treatment**			**Difference in means (D)**
1*	3	14	15	16	21	58	-21
2	3	14	16	15	21	58	-20.33333
3	3	14	21	15	16	58	-17
4	3	14	58	15	16	21	7.666667
5	3	15	16	14	21	58	-19.66667
6	3	15	21	14	16	58	-16.33333
7	3	15	58	14	16	21	-8.333333
8	3	16	21	14	15	58	-15.66667
9	3	16	58	14	15	21	9
10	3	21	58	14	15	16	12.33333
11	14	15	16	3	21	58	-12.33333
12	14	15	21	3	16	58	-9
13	14	15	58	3	16	21	15.66667
14	14	16	21	3	15	58	8.333333
15	14	16	58	3	15	21	16.33333
16	14	21	58	3	15	16	19.66667
17	15	16	21	3	14	58	-7.666667
18	15	16	58	3	14	21	17
19	15	21	58	3	14	16	-20.33333
20	16	21	58	3	14	15	21

Note. Each row represents one of the 20 data partitions, with the data partition on the observed difference marked with an asterisk (*)

interest. The effect of the new treatment can be manifested if the observed mean difference is considered extreme, as compared to an appropriate reference distribution.

The difference in the mean recovery times for the observed data is $d = 10.66667 - 31.66667 = -21$ days, indicating a decrease in recovery times in favour of the new treatment.

If the null hypothesis is true and the two treatments do not affect the recovery times differentially, then the recovery times for each patient would stay exactly the same if he/she would have been assigned to a different treatment group. Thus, the patient who recovered in 15 days on the new treatment is just as likely to recover in 15 days on the standard treatment because there is no difference between the two treatments.

As a result, we consider the recovery times as fixed, and the assignment of those recovery times to the two treatments as random. Therefore, the random assignment of the available patients to the treatments gives us the prob-

abilistic justification for considering reshuffled recovery times to derive the reference distribution. Under the null hypothesis of no treatment effect, the assignment of recovery times to the new and the standard treatments is arbitrary and the recovery times obtained under the two conditions are *exchangeable*. Given exchangeability under the null hypothesis, the obtained recovery times are equally likely to have arisen from any possible assignment. Hence, the mean differences of recovery times associated with each of the possible assignment are also equally likely. A probability distribution for the mean difference can be constructed by calculating the mean difference for each of the possible data divisions. Table 1 lists all $\binom{6}{3} = 20$ possible combinations of the six obtained recovery times into two groups of size 3, and the mean difference for each combination. Since each combination is equally likely, the probability of any of these combinations is .05.

The *p* value of the randomization test under the null hypothesis is the proportion of the values of the test statistic D in the reference distribution as small as, or smaller than the obtained test statistic d (for a two-tailed test this is formulated as "as extreme as or more extreme than"). Thus,

$$p = P(D \leq d | H_0) = \frac{\sum_{i=1}^{\binom{N}{n}} I(D_i \leq d)}{\binom{N}{n}}$$

where D_i is the value of the test statistic for the ith data partition and $I(\cdot)$ is the indicator function. In Table 1, the obtained mean difference is $d = -21$, which is the most extreme negative mean difference among all the data partitions, so the p value is $p = P(D \leq -21) = \frac{1}{20} = .05$ (for a one-tailed test).

This *p* value should be interpreted cautiously like any *p* value resulting from a statistical significance test. It is a measure of evidence against the null hypothesis, with smaller values indicating more evidence against the null hypothesis than larger values. It should not be interpreted as the probability that the null hypothesis is true; it is the probability to observe the given data or even more extreme data if the null hypothesis were true. If there was a preset significance level α, like the commonly used 5% significance level, then the *p* value can be compared with this level α. If the *p* value is smaller than or equal to the significance level α, as for the data in Table 1, the null hypothesis is said to be "rejected". In the example, it means that the null hypothesis that

the new and standard treatment lead to the same recovery times (in days) for the given patients is rejected. There is statistical evidence that recovery times would decrease if the patients would take the new treatment.

Comparison with a parametric t test

How does the randomization test compare with the parametric *t* test? The theory and procedure underlying the parametric *t* test is well known from most introductory statistics texts (see e.g., Moore, McCabe, & Craig, 2010). Suppose that we have a random sample of n_1 patients taking the new treatment and a random sample of n_2 patients taking the standard treatment, with corresponding sample means and variances of $\bar{x}_1$ and $\bar{x}_2$, and s_1^2 and s_2^2, respectively. We assume that the recovery times of patients taking the new treatment follow a Gaussian distribution at the population level with mean μ_1 and variance σ^2 and that the recovery times of patients taking the standard treatment follow a Gaussian distribution at the population level with mean μ_2 and variance σ^2. The null hypothesis is that $\mu_1 = \mu_2$ while the alternative hypothesis is $\mu_1 < \mu_2$ if one expects that the new treatment will be more effective in reducing the recovery times. The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with a pooled estimate of the common within-group standard deviation

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

If the null hypothesis of no difference between the two population means is true, *t* will be a random value from Student's *t* distribution with $n_1 + n_2 - 2$ degrees of freedom (Fisher, 1925; Student, 1908).

For the observed data in Table 1, with $n_1 = n_2 = 3$, the means and variances for the two samples are $\bar{x}_1 = 10.67$, $\bar{x}_2 = 31.67$, $s_1^2 = 44.33$, and $s_2^2 = 526.33$. The test statistic *t* equals -1.5226 with 4 degrees of freedom. The *p* value for the one-tailed test is .1013, which is more than twice the randomization test *p* value, and would not result in a rejection of the null hypothesis at any of the conventional significance levels. Furthermore, the validity of this *p* value relies on the plausibility of the parametric assumptions:

1. random sampling of patients from certain populations
2. Gaussian distributions for the values of recovery times
3. equal population variances for the values of recovery times.

Assumption 1 is obviously problematic since the patients were not a random sample from a specified population. Assumptions 2 and 3 may be true, but are difficult to check when sample sizes are small. In fact, assumption 3 is not needed for the more generally applicable Welch version of the two-sample *t* test (Welch, 1947). In this version of the *t* test, the degrees of freedom are adjusted, and for the data in Table 1 this results in 2.335 degrees of freedom, and a *p* value of .1249, an even larger value than for the parametric *t* test that assumes equal population variances.

Comparison with a nonparametric rank test

Some researchers who are reluctant to make the Gaussian assumption, or more in general are reluctant to apply a parametric *t* test with such small samples, might consider a nonparametric rank test to analyse the back pain data in Table 1. In this case, the Wilcoxon-Mann-Whitney test is most popular. It tests the null hypothesis that the population distributions are identical. The one-tailed *p* value for this test on the data in Table 1 is .05.

This *p* value of the Wilcoxon-Mann-Whitney test is identical to the *p* value of the randomization test on the original data, and this is no coincidence. The Wilcoxon-Mann-Whitney test was originally developed as a randomization test on rank-transformed data (see Mann & Whitney, 1947; Wilcoxon, 1945), and because the data in Table 1 show no overlap in the original recovery times between the two treatment groups, there will also be no overlap in the ranks. Because absence of overlap always results in the smallest possible randomization test *p* value, there will be no difference between a randomization test on the original recovery times and a randomization test on the rank-transformed recovery times.

Table 2 illustrates the derivation of the Wilcoxon-Mann-Whitney test as a randomization test based on ranks, with the sum of ranks of the new treatment as the test statistic (therefore the test is also sometimes called the "Wilcoxon rank-sum test"). All data partitions of ranks on six recovery times are shown in Table 2.

Thus, if the two treatments will lead to the same recovery times, the probability that the three patients having the new treatment show fastest recovery times (ranked 1, 2, 3) is 1/20 or .05. In this example, the rank test and the randomization test yield the same *p* value. However, due to the fact that ranks instead of the original data are employed, in other applications a loss of information may occur and eventually, reduce the statistical power and efficiency (Lehmann & D'Abrera, 2006).

Table 2

All possible data partitions of ranks on six recovery times (days) to two treatment groups of sizes n=3 in each group

	Data Partitions of ranks						
No.	**New Treatment**			**Standard Treatment**			**Sum of ranks of the new treatment**
1*	1	2	3	4	5	6	6
2	1	2	4	3	5	6	7
3	1	2	5	3	4	6	8
4	1	2	6	3	4	5	9
5	1	3	4	2	5	6	8
6	1	3	5	2	4	6	9
7	1	3	6	2	4	5	10
8	1	4	5	2	3	6	10
9	1	4	6	2	3	5	11
10	1	5	6	2	3	4	12
11	2	3	4	1	5	6	9
12	2	3	5	1	4	6	10
13	2	3	6	1	4	5	11
14	2	4	5	1	3	6	11
15	2	4	6	1	3	5	12
16	2	5	6	1	3	4	13
17	3	4	5	1	2	6	12
18	3	4	6	1	2	5	13
19	3	5	6	1	2	4	14
20	4	5	6	1	2	3	15

Note. Each row represents one of the 20 data partitions, with the data partition of ranks on the original data marked with a star (*)

Systematic versus Monte Carlo randomization tests

When all possible data partitions are exhaustively listed, the relevant RndTs are often called *systematic* RndTs. However, the systematic RndTs are not always practical. Noreen (1989, p. 14) stated that "exact randomization is feasible, however, with present computer technology only for very small data sets". For instance, a systematic RndT for an experiment with 30 participants randomly assigned to 3 treatments with 10, 10, and 10 participants in each treatment requires 5.55 trillion arrangements of the data and corresponding calculations of the test statistic. Even specialised statistical software and special-purpose algorithms would have problems with this large number of computations. When the total number of data partitions is too large, *Monte Carlo* RndTs are usually implemented. Monte Carlo RndTs are not limited by the size of the sample because they use only a subset of all possible data partitions to derive a valid *p* value (Dwass, 1957; Edgington & Onghena, 2007; Manly, 1997). A few

thousand data partitions can yield an accurate estimate of the exact *p* value and a valid (i.e., conservative) significance test (Edgington, 1969).

Equivalent test statistics

One of the advantages of RndTs is that they allow researchers to use the test statistic that is most sensitive to the effect that the researchers are interested in. In other words, researchers are not restricted to the conventional test statistics, like *t* or *F*, but can also consider medians, ranges, quartiles, or any other statistical measure. In the back pain example, the difference between means of recovery times, *D*, was used as the test statistic. People might wonder whether the same *p* value would be obtained if another test statistic had been employed. If two test statistics always give the same *p* value for an RndT, they are called *equivalent test statistics* (Edgington & Onghena, 2007; Manly, 1997). For instance, one might wonder what result would be obtained if a two-sample *t* statistic had been used for the example data in Table 1. In fact, the same *p* value would result, and it can be demonstrated that for a completely randomized design with two treatments, the difference between means and the two-sample *t* statistic will always give the same *p* values (Edgington & Onghena, 2007).

The reason why the equivalent test statistics are useful is that they take less time to finish the calculation by using a simpler test statistic (Edgington & Onghena, 2007; Manly, 1997). Manly (1997, pp. 15-16) mentioned that "minor differences such as the multiplication by a constant may become important when the statistic used has to be evaluated thousands of times". At the same time, equivalent statistics also have an important theoretical role to play. They show what part of the statistic is crucial for showing the tested effect.

One aspect we need to remember is that not all the test statistics are equivalent test statistics. To the question that under what kind of circumstances two test statistics are equivalent, Edgington and Onghena (2007, p. 43) state that "two test statistics are equivalent if and only if they are perfectly monotonically correlated over all data permutations in the set". In the back pain example, we want to know whether *D* and the two-sample *t* statistic are equivalent test statistics. The formula of *t* statistic is composed of a numerator and a denominator. The numerator of *t*, which is the difference between two means, manifests the difference of treatment effects; the denominator is an estimate of the variability of the numerator for parametric tests. Due to the fact that the RndT procedure in this example generates its own distribution of *D*s, this makes the denominator irrelevant. Consequently, as *t* decreases *D* must decrease, providing the same order of *t* and *D* over the data partitions. Therefore the two test statistics will give the same *p* value and *D* is an equivalent test statistic to *t*.

RT4Win: a new software package for RndTs

When applying RndTs in practice, researchers will inevitably encounter a computational burden. Although examples with very small sample sizes can be easily calculated by hand, the number of data partitions increases sharply with an increasing number of observations. By means of computer programs, several thousands or even millions of data partitions can be generated within a short period. Even if the number of data partitions is gigantic, Monte Carlo algorithms can still make RndTs practical. Therefore, suitable programs and software packages are needed to execute RndTs.

In this article, we present a stand-alone RndTs software package, RT4Win. RT4Win is a software package for the Windows platform with a user-friendly interface. RT4Win provides a set of RndTs for analysing data in many experimental designs described in Table 3.

Table 3

List of experimental designs and corresponding programs in RT4Win

Experimental Designs	Programs
Between-subjects Design	One-way ANOVA: Systematic Data Partition
	One-way ANOVA - Equal N: Systematic Data Partition
	One-way ANOVA: Monte Carlo Data Partition
	Independent *t* Test: Systematic Data Partition
	Independent *t* Test: Monte Carlo Data Partition
Factorial Design	Test of Main Effects: Monte Carlo Data Partition
Repeated-measures Design	Repeated-measures ANOVA: Systematic Data Partition
	Repeated-measures ANOVA: Monte Carlo Data Partition
	Correlated *t* Test: Systematic Data Partition
	Correlated *t* Test: Monte Carlo Data Partition
Multivariate Design	Multivariate tests based on composite *z* scores
Correlation	Product-moment Correlation: Systematic Data Partition
	Product-moment Correlation: Monte Carlo Data Partition
Trend Test	Correlation Trend Test: Monte Carlo Data Partition
Matching Test	Matching Test: Monte Carlo Data Partition

Both systematic and Monte Carlo RndTs are available for most of the listed designs. As to Monte Carlo RndTs, millions of data partitions can be executed in a limited time frame, which can greatly increase the precision of the *p* values. In addition, RT4Win allows the users to save and load data in a common Windows environment. The output also provides the value of conventional test statistics such as *t* or *F* for the obtained data, total number of data partitions and the number of data partitions that gives a test statistic value equal to or more extreme than the obtained one. RT4Win is free of charge and is available upon request from the authors.

Examples of RndTs via RT4Win

To demonstrate our software package, we illustrate its use with three sets of data.

Example 1

The following example is from Kutner, Nachtsheim, Neter, and Li (2005, pp. 723-724). A psychologist is interested in the effect of colour of paper (blue, green, orange) on response rates for questionnaires distributed by the "wind-shield method" in supermarket parking lots, 15 representative supermarket parking lots were chosen in a metropolitan area and each colour was assigned at random to five of the lots. The response rates (in per cent) are listed in Table 4.

Table 4

Response rates (in percentages) by colours for each of the 15 parking lots

Colour	Response Rates (in per cent)
Blue	28
Blue	26
Blue	31
Blue	27
Blue	35
Green	34
Green	29
Green	25
Green	31
Green	29
Orange	31
Orange	25
Orange	27
Orange	29
Orange	28

The calculation procedure should be carried out in the following steps:

1. First, the user should select the program that matches the design of the experiment. Since this experiment intends to test the null hypothesis of no differential effect among the three colours on response rates, the program on RndTs for one-way ANOVA should be selected. Meanwhile, with only $\frac{15!}{5!5!5!}$ = 756756 data partitions, a systematic RndT is feasible to enumerate all partitions in seconds. As a result, the program "One-way ANOVA: Systematic Data Partition" should be selected to execute the calculation (see Figure 1).

2. After opening the program window, the user can start the data input. Data should be input from the keyboard following the instructions at the "instruction region" on the top of the window. First, input the number of groups, then the number of subjects in each group, lastly the measurements (response rates in this example) for each group.

3. After finishing the data input, one can simply click the "Calculation" button to execute the required RndT. It is noted that the chosen statistic in this program is

$$\sum_{i=1}^{k} T_i^2$$

where k is the number of groups, T_i is the sum of all the measurements in the ith group. As an equivalent test statistic of F, the test statistic $\Sigma_{i=1}^{k} T_i^2$ will yield the same p value. For this case, $\Sigma_{i=1}^{k} T_i^2 = 12739$ for the obtained data. This RndT yields $p = .683724$, with 517412 out of the total 756756 test statistics equal to or greater than 12739. Figure 2 gives the calculation results from RT4Win.

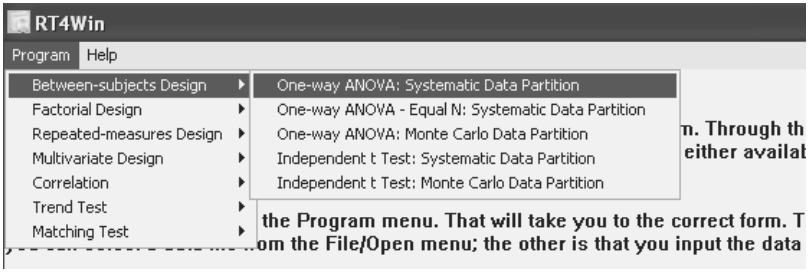

Figure 1

Screen shot on how to choose appropriate program from RT4Win

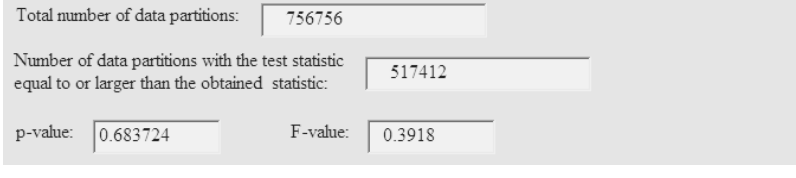

Figure 2

Screen shot on calculation results of Example 1

As mentioned above, systematic RndTs are not always feasible. A Monte Carlo RndT should be used instead when the number of data partitions is too

large. The following example will show how to perform Monte Carlo RndTs via RT4Win.

Example 2

The data of this example are from Kutner et al. (2005, pp. 685-686). The Kenton Food Company wished to test four different package designs for a new breakfast cereal. Nineteen stores, with approximately equal sales volumes, were selected as the experimental units. Each store was randomly assigned one of the package designs, with each package design assigned to five stores. An accident occurred in one store during the study period, so this store had to be dropped from the study. Hence, one of the designs was tested in only four stores. This drop-out problem does not affect the validity of RndTs because the store dropped out independently of the treatment assignment of that store. In other words, the random assignment does not give rise to the drop-out and RndTs are still valid. All conditions that could affect sales were kept the same for all of the stores in the experiment. Sales, in number of cases, were observed for the study period, and the results are recorded in Table 5.

Table 5

Number of cases sold by stores for each of four package designs

Package Design	Sales (cases sold)
A	11
A	17
A	16
A	14
A	15
B	12
B	10
B	15
B	19
B	11
C	23
C	20
C	18
C	17
D	27
D	33
D	22
D	26
D	28

For this example, there are

$$\frac{19!}{5!5!4!5!} = 2{,}933{,}186{,}256$$

possible data partitions to be considered. By using the same program as for Example 1, we obtained an exact *p* value of

$$\frac{118464}{2{,}933{,}186{,}256} = .0004$$

It took us 4 minutes and 38 seconds to finish this computation on RT4Win (PC used was a DELL Optiplex 760, Intel(R) Core(TM)2 Duo E8500 processor, 3.16GHz, 3.21GB RAM). If a Monte Carlo RndT is desired, the user should take the following steps:

1. First, the user should select the program "One-way ANOVA: Monte Carlo Data Partition" from the menu "program" on the main window of RT4Win.

2. After opening the program window, the user can start the data input. The process will be identical to that in Example 1.

3. After inputting the data, the user should fill in the desired number of data partitions. The specific number is arbitrary, but should be as large as possible, taking into account the computational speed of your computer.

4. Afterwards, one can simply click the "Calculation" button to execute the required RndT (one can get different *p* values on every click of the "Calculation" button since each time a different set of partitions is chosen).

Table 6 summarises the results for six different numbers of data partitions and the associated *p* values.

Table 6

P values of six sets of data partitions with the exact p value given in last line – Kenton Food Company Example

Data partitions	*p* value
100	0.01
1000	0.001
10,000	0.0001
100,000	0.0005
1,000,000	0.00035
10,000,000	0.000302
2,933,186,256	0.00040

Example 3

The data of this example are from Kutner et al. (2005, p. 1132). In a wine-judging competition, four Chardonnay wines of the same vintage were judged by six experienced judges. Each judge tasted the wines in a blind fashion, i.e., without knowing their identities. The order of the wine presentation was randomized independently for each judge. Each wine was scored on a 40-point scale; the higher the score, the greater is the excellence of the wine. The data are listed in Table 7.

Table 7
Wine judging scores data for example 3

Judge	Wine 1	Wine 2	Wine 3	Wine 4
1	20	24	28	28
2	15	18	23	24
3	18	19	24	23
4	26	26	30	30
5	22	24	28	26
6	19	21	27	25

The calculation procedure should be carried out in the following steps:

1. First, the user should select the program that matches the design of the experiment. Since this experiment intends to determine the p value in a repeated-measures design, one of the two RndTs programs for repeated-measures design should be selected. For a repeated-measures experiment, each of the n subjects takes all the k treatments and the total number of possible data partitions is $(k!)^n$ (for details, see Edgington & Onghena, 2007, pp. 116-117). In this example, the total number of data partitions is $(4!)^6 = 191102976$. Program "Repeated-measures ANOVA: Systematic Data Partition" can be selected to execute the calculation.

2. After opening the program window, the user can start the data input. First, the user needs to input the number of participants (number of judges), then number of treatments (number of wines), lastly the measurements under all the treatments of each participant (wines scores).

3. When the data input is done, the user can simply click the "Calculation" button to execute the program. It is noted that the chosen statistic in this program is

$$\sum_{i=1}^{k} T_i^2$$

where k is the number of treatments, T_i is the sum of all the measurements on the ith treatment. As an equivalent test statistic of F, the test statistic

$\Sigma_{i=1}^{k} T_i^2$ will yield the same p value. In this example, with $k = 4$, $T_1 = 20 + 15 + 18 + 26 + 22 + 19 = 120$, $T_2 = 24 + 18 + 19 + 26 + 24 + 21 = 132$, $T_3 = 28 + 23 + 24 + 30 + 28 + 27 = 160$, and $T_4 = 28 + 24 + 23 + 30 + 26 + 25 = 156$, the test statistic $\Sigma_{i=1}^{k} T_i^2 = (120)^2 + (132)^2 + (160)^2 + (156)^2 = 81760$. For all the 191102976 data partitions performed, there are 576 data partitions that can produce a value of test statistic equal to or larger than the obtained one. As a result, the p value is $576 / 191102976 = .000003$.

Besides between-subject designs and repeated-measures designs, RT4Win also allows users to perform RndTs for factorial designs, multivariate designs, correlation, trend tests, and matching and proximity experiments. For further information on each program, users are referred to the help file embedded in the software.

Discussion

Within the family of statistical significance tests, RndTs are not frequently used by researchers. One of the reasons for this is probably that most of the introductions to inferential statistics focus on classical parametric statistical tests (t or F tests) and rarely include randomization tests and their rationale. Another reason for the underutilisation of RndTs may be that RndTs rely on computationally intensive algorithms and that most popular and common statistical software packages do not provide facilities to easily perform randomization tests. In this article, we have presented an efficient software package, RT4Win, to fill this gap. In this section, we will discuss the advantages and disadvantages of RndTs, other available RndTs software, the assets of our software and suggestions for future software development.

The most important advantage of RndTs is that they do not make any assumption regarding the probability distribution underlying the data at hand. The validity of RndTs is only based on the random assignments of experimental units to treatments. As a result, RndTs are free from the assumption of random sampling, an assumption that lies at the heart of parametric statistical inference, but that is unrealistic in many practical situations (Anderson, 2001; Hunter & May, 2003; Zieffler et al., 2011). Also, RndTs are very easy to apply and versatile, so that researchers can develop an RndT for their own particular design (Edgington & Onghena, 2007).

It has been widely recognised that there are three important drawbacks of RndTs: (a) they were too computationally intensive, (b) their applicability was limited to simple scenarios, and (c) they could be replaced by the available classical nonparametric tests based on ranks (Welch, 1990). However, it is fortunate to see that the above mentioned problems of RndTs have been resolved to a large extent by the efforts of researchers from many disciplines. Gill (2007), for example, invented a clever algorithm by using a Fourier

expansion to count extreme cases, which decreases the computing load to affordable proportions. Based on Gill's algorithm, Mewhort, Johns, and Kelly (2010) introduced an RndT for a 2 by 2 factorial design, which greatly limits the computational load. Moreover, it is reassuring to see that RndTs are more and more frequently applied in new and exciting research domains, such as ERP- (event-related potential) (e.g., Fehr, Wiedenmann, & Herrmann, 2007; Kayser, Tenke, Gates, & Bruder, 2007; Maris, 2004) and EEG- (electroencephalogram) research (e.g., Henderson, Yoder, Yale, & McDuffie, 2002; Keil, Mussweiler, & Epstude, 2006). Finally, as was mentioned earlier in this paper, nonparametric rank methods are not the first choice if the numerical information in the data is valid and should not be wasted (Edgington & Onghena, 2007).

Before the appearance of RT4Win, three types of computer software were available to perform RndTs. First, there are a number of computer languages available that can be used to produce specific RndT programs. These vary from basic programming languages such as FORTRAN (e.g., Berry & Mielke, 1996; Berry & Mielke, 1999) to higher level languages such as SAS (e.g., Chen & Dunlap, 1993), SPSS (e.g., Hayes, 1998), and R (Bulté & Onghena, 2008; Bulté & Onghena, 2009). Second, there are some software packages solely for carrying out RndTs, such as RANDIBM (Edgington, 1995) or SCRT (Onghena & Van Damme, 1994). However, most of this older kind of software was designed for DOS machines and does not have state-of-the-art interfaces. Finally, RndTs can be performed under some commercial software packages (e.g., StatXact). Unlike the first type of computer programs that perform RndTs under some specific experimental contexts, RT4Win, as an integrated environment, can perform RndTs to test statistical significance in large variety of experimental contexts and is extremely fast. Unlike the second type of software packages that perform RndTs for DOS machines, RT4Win has a user-friendly interface that is compatible with the popular Windows platform. A final advantage is that RT4Win is free of charge.

Although RT4Win can perform RndTs in a variety of experimental contexts, the facility to deal with multivariate designs is still limited. In the present version of the software, the test statistic for multivariate ANOVA is limited to composite z scores. It might be worthwhile to add options for Wilks' Lambda, Pillai-Bartlett's Trace, Hotelling-Lawley's Trace, and Roy's Greatest Root. Furthermore, extensions for discriminant analysis and canonical correlation might be useful. It would also be interesting to develop more tools so that researchers can perform simulation studies on the performance of RndTs, which in turn give further guidance to the applications of RndTs in specific designs. Finally, an integrated R package would be helpful. R could have a speed disadvantage for computer-intensive statistical tests as compared to RT4Win, but it has the advantage of gaining popularity among psy-

chologists and of offering a comprehensive data-analytic environment, including powerful visualisation and graphical tools (Kelley, 2007).

References

Anderson, M.J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences, 58*, 626-639.

Berry, K.J., & Mielke, P.W. (1996). Analysis of multivariate matched-paired data: A FORTRAN 77 program. *Perceptual and Motor Skills, 83*, 788-790.

Berry, K.J., & Mielke, P.W. (1999). A FORTRAN program for permutation covariate analyses of residuals based on Euclidean distance. *Psychological Reports, 82*, 371-375.

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40*, 467-478.

Bulté, I., & Onghena, P. (2009). Randomization tests for multiple baseline designs: An extension of the SCRT-R package. *Behavior Research Methods, 41*, 477-485.

Chen, R.S., & Dunlap, W.P. (1993). SAS procedures for approximate randomization tests. *Behavior Research Methods, Instruments, & Computers, 25*, 406-409.

Cobb, G.W. (2007). The introductory statistics course: A Ptolemaic curriculum. *Technology Innovations in Statistics Education, 1*. Retrieved from http://escholarship.org/uc/uclastat_cts_tise?volume=1;issue=1.

Cox, D.R., & Hinkley, D.V. (1974). *Theoretical statistics*. London: Chapman and Hall.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics, 28*, 181-187.

Edgington, E.S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin, 66*, 485-487.

Edgington, E.S. (1969). Approximate randomization tests. *Journal of Psychology: Interdisciplinary and Applied, 72*, 143-149.

Edgington, E.S. (1995). *Randomization tests* (3rd ed.). New York: Marcel Dekker.

Edgington, E.S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Fehr, T., Wiedenmann, P., & Herrmann, M. (2007). Differences in ERP topographies during color matching of smoking-related and neural pictures in smoker and non-smokers. *International Journal of Psychophysiology,* 65, 284-293.

Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh and London: Oliver & Boyd.

Gill, P.M.W. (2007). Efficient calculation of *p*-values in linear-statistic permutation significance tests. *Journal of Statistical Computation & Simulation, 77*, 55-61.

Good, P.I. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses* (2nd ed.). New York: Springer-Verlag.

Hayes, A.F. (1998). SPSS procedures for approximate randomization tests. *Behavior Research Methods, Instruments, & Computers, 30*, 536-543.

Henderson, L.M., Yoder, P.J., Yale, M.E., & McDuffie, A. (2002). Getting the point: Electrophysiological correlates of protodeclarative pointing. *International Journal of Developmental Neuroscience, 20*, 449-458.

Hunter, M.A., & May, R.B. (2003). Statistical testing and null distributions: What to do when samples are not random. *Canadian Journal of Experimental Psychology, 57*, 176-188.

Kayser, J., Tenke, C.E., Gates, N.A., & Bruder, G.E. (2007). Reference-independent ERP old/new effects of auditory and visual word recognition memory: Joint extraction of stimulus- and response-locked neuronal generator patterns. *Psychophysiology, 44*, 949-967.

Keil, A., Mussweiler, T., & Epstude, K. (2006). Alpha-band activity reflects reduction of mental effort in a comparison task: A source space analysis. *Brain Research, 1121*, 117-127.

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods, 39*, 979-984.

Kempthorne, O., & Doerfler, T.E. (1969). The behaviour of some significance tests under experimental randomization. *Biometrika, 56*, 231-247.

Kutner, M.H., Nachtsheim, C.J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). Boston, MA: McGraw-Hill/Irwin.

Lehmann, E.L., & D'Abrera, H.J. (2006). *Nonparametrics: Statistical methods based on ranks.* New York, NY: Springer.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to *t* and *F* tests in biomedical research. *The American Statistician, 52*, 127-132.

Manly, B.F.J. (1997). *Randomization, bootstrap, and Monte Carlo methods in biology* (2nd ed.). New York: Chapman & Hall.

Mann, H.B., & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*, 50-60.

Maris, E. (2004). Randomization tests for ERP topographies and whole spatiotemporal data matrices. *Psychophysiology, 41*, 142-151.

Mewhort, D.J.K., Johns, B.T., & Kelly, M. (2010). Applying the permutation test to factorial designs. *Behavior Research Methods, 42*, 366-372.

Mielke, P.W., & Berry, K.J. (2007). *Permutation methods: A distance function approach* (2nd ed.). New York: Springer-Verlag.

Moore, D.S., McCabe, G.P., & Craig, B. (2010). *Introduction to the practice of statistics* (7th ed.). New York: W.H. Freeman.

Noreen, E.W. (1989). *Computer-intensive methods for testing hypotheses: An introduction.* New York: John Wiley & Sons.

Onghena, P. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: A rejoinder. *Behavioral Assessment, 14*, 153-171.

Onghena, P., & Edgington, E.S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research and Therapy, 32*, 783-786.

Onghena, P., & Edgington, E.S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain, 21*, 56-68.

Onghena, P., & Van Damme, G. (1994). SCRT 1.1: Single-case randomization tests. *Behavior Research Methods, Instruments, & Computers, 26*, 369.

Pitman, E.J.G. (1937a). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society Series B, 4*, 119-130.

Pitman, E.J.G. (1937b). Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society Series B, 4*, 225-232.

Pitman, E.J.G. (1938). Significance tests which may be applied to samples from any populations. III. The analysis of variance test. *Biometrika, 29*, 322-335.

Ramsey, F.L., & Schafer, D.W. (2002). *The statistical sleuth: A course in methods of data analysis* (2nd ed.). Belmont, CA: Duxbury Press.

Stigler, S.M. (1992). A historical view of statistical concepts in psychology and educational research. *American Journal of Education, 101*, 60-70.

Student (1908). The probable error of a mean. *Biometrika, 6*, 1-25.

Welch, B.L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika, 34*, 28-35.

Welch, W.J. (1990). Construction of permutation tests. *Journal of the American Statistical Association, 85*, 693-698.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*, 80-83.

Zieffler, A.S., Harring, J.R., & Long, J.D. (2011). *Comparing groups: Randomization and bootstrap methods using R.* Hoboken, NJ: Wiley.

Received September 1, 2011
Revision received December 25, 2011
Accepted February 6, 2012