Behavior Research Methods, Instruments, & Computers
1995, 27 (3), 408–411

Pitfalls in computing and interpreting randomization test *p* values: A commentary on Chen and Dunlap

PATRICK ONGHENA
Katholieke Universiteit Leuven, Leuven, Belgium
and
RICHARD B. MAY
University of Victoria, Victoria, British Columbia, Canada

Chen and Dunlap (1993) added to the growing list of papers promoting the use of randomization tests in statistical testing. Their particular contribution was an SAS program that could bring computation of these tests to a wider audience. The present paper points to several problems with the presentation of Chen and Dunlap and provides solutions to these problems. It is concluded that randomization tests deserve more attention, but that they are best computed by programs written in a low-level programming language or, if using SAS on a mainframe, by using the MULTTEST procedure.

Chen and Dunlap (1993) presented an SAS program that could serve as a template for testing hypotheses using an approximate randomization test (ART). They included SAS code for an ART, testing the equality of two means, testing the significance of correlation, and testing the equality of more than two means.

We acknowledge that Chen and Dunlap (1993) brought these powerful and versatile statistical tests to the fore, realizing that their use would remain limited as long as the popular statistical computer packages do not provide user-friendly routines, but we also want to comment on several features of their approach that may be problematic in practice. Our comments concern (1) the computation of *p* values, (2) one-tailed versus two-tailed randomization tests, (3) the number of pseudosamples, (4) computing time and memory requirements, (5) the SAS MULTTEST procedure, (6) approximate versus exact randomization tests, and (7) power and Type I error rate. In each comment, a potential problem is identified and a solution is suggested.

1. The Computation of *p* Values

A small technical problem could be that the Chen and Dunlap (1993) SAS implementations do not perform tests that provide exact Type I error rate control. Their programs generate 1,000 pseudosamples (number of pseudosamples, or NOP), compute 1,000 pseudostatistics in addition to the original statistic, count the number of pseudostatistics that are equal to or more extreme than the original statistic (number of significant cases, or NOS), and compute the ratio NOS/NOP to get the p value. In order to have a valid ART, however, the original statistic has to be among the pseudostatistics—that is, the p value can never be smaller than 1/NOP (Edgington, 1987, pp. 43–45; Manly, 1991, pp. 15–16; Noreen, 1989, p. 17).

This small deviation of the Chen and Dunlap (1993) programs can be inferred from scrutinizing their algorithm or can be demonstrated by running a simulation study to determine the Type I error rate. More directly, the implications of their approach can be shown with a dataset where the scores for Treatment A are all larger than for Treatment B and running their Program 1. If the dataset is large enough, it is very likely that all pseudostatistics are smaller than the original statistic. Consequently a p value of zero, which should be an impossible value, is obtained.

With a modification of the Chen and Dunlap (1993) program, however, valid tests can be obtained. In this modification, one should ensure that 999 pseudosamples are generated and that the p value is computed as (NOS+1)/(NOP+1).

2. One-Tailed ART With Unequal Group Sizes

Chen and Dunlap (1993) recommended their programs for equal as well as unequal group sizes and computed the p value of a one-tailed ART by dividing the two-tailed p value by two. The division of the two-tailed p value by two, however, provides a valid one-tailed ART only if the group sizes are equal (Edgington, 1987, p. 82). If the group sizes are unequal, the randomization distribution of the t statistic (or an equivalent) may be nonsymmetric and the absolute value statistic $|t|$ cannot capture this nonsymmetry.

For example, suppose the following data are observed: 7, 5, 5 for Treatment A and 4, 4, 3, 3, 2 for Treatment B. In this case, a one-tailed exact randomization test (with t as a test statistic) gives a p value that is *equal* to the p value given by a two-tailed exact randomization test (with $|t|$ as a test statistic)—namely, $p = 1/56$. The Chen and Dunlap (1993) method would yield a value around 1/112 for a one-tailed ART.

Therefore, it is necessary to restrict the use of their programs to two-tailed ARTs or to one-tailed ARTs with equal group sizes. One-tailed tests with unequal group sizes can be performed after changing the test statistic in the program. A straightforward change is the removal of the absolute value SAS function *ABS* on the 11th and 41st line of their Program 1 to obtain a signed difference between means.

The authors wish to thank Marc Brysbaert, Stef Decoene, Luc Delbeke, Eugene S. Edgington, and William P. Dunlap for their helpful comments on an earlier version of the manuscript. The first author is Research Assistant of the National Fund for Scientific Research of Belgium. Correspondence should be addressed to Patrick Onghena, K. U. Leuven, Department of Psychology, Center for Mathematical Psychology and Psychological Methodology, Tiensestraat 102, B-3000 Leuven, Belgium (e-mail: patrick.onghena@psy.kuleuven.ac.be).

3. The Number of Pseudosamples

In the Chen and Dunlap (1993) programs, 1,000 pseudosamples are generated. Although 1,000 is a reasonable number to demonstrate the procedure (and efficient at the 5% level of significance), it is important to add (1) that the reliability of the p value increases with increasing NOP, and (2) that the reasonable minimum NOP depends on the level of significance. For example, Efron and Tibshirani (1993, p. 211), Manly (1991, p. 35), and Westfall and Young (1993, p. 39) recommended use of 5,000 to 10,000 pseudosamples, especially if the test is performed at a level of significance smaller than 5%. This is important to realize because the Chen and Dunlap (1993) programs print out confidence levels for the 1% significance level.

The influence of NOP on the reliability of the p value can be demonstrated with the sample data for Program 1 used by Chen and Dunlap (1993, p. 408). They obtained a p value of 33/1,000 = .033 with their program, while an ART with NOP = 10,000 gave a p value of 256/10,000 = .0256 and the exact randomization test p value is 4,790/184,756 = .02593 (calculated with NPStat 3.7; May, Hunter, & Masson, 1993; May, Masson, & Hunter, 1989).

Again, a small modification of the Chen and Dunlap (1993) programs, increasing the NOP value, performs more reliable tests. This increase, however, may make the computing time and memory requirements prohibitive. Therefore, for several applications it might be recommended to use another SAS program or another programming language for ARTs (see next two problems).

4. Computing Time and Memory Requirements on a PC

We checked the computing time of the SAS programs on two PCs, five numbers of observations, and two values of NOP. The PCs were an IBM PC/AT with 640K RAM running at 8 MHz under DOS 3.2 and an IBM PC/80486 with 4 Mb RAM running at 50 MHz under DOS 5.0. The numbers of observations (N) were 10, 20, 50, 100, and 200, with equal group sizes. The NOPs were 1,000 and 10,000. Table 1 shows the CPU time for each of the combinations, compared with the CPU time for the ART that is included in the Single-Case Randomization Test (SCRT) software package, which was programmed in Pascal (Onghena & Van Damme, 1994; Van Damme & Onghena, 1993).

The computing times for the Chen and Dunlap (1993) programs are very large relative to the computing times for the SCRT program. Furthermore, insufficient memory was available on the PCs to work with 10,000 pseudosamples. This outcome follows from saving all the pseudosamples and all the pseudostatistics: NOP * (N−1) * 3 values have to be stored in the SHUFFLE dataset, and NOP * 3 * 5 values have to be stored in the PSEUDO dataset. The programs appear to be feasible only on a mainframe (see Section 5 below).

Although it is true that the SAS tools, such as the PROBBETA function, make the user-written code simple, the price to pay in computing time for using a high-level programming language appears to be too high. Therefore, we would prefer to use a low-level programming language and compiler for computer-intensive tests, even on a mainframe. A software library, such as NAG's (Numerical Algorithms Group, 1990), could be used to provide the ready-made statistical functions.

5. The Use of the SAS MULTTEST Procedure to Perform ARTs

As a means of reducing computing time and memory requirements, it may also be interesting to consider using the MULTTEST procedure, which is included in SAS from Version 6.07 on and which has PERMUTATION as one of its options (SAS Institute Inc., 1992; Westfall & Young, 1993). The MULTTEST procedure is designed to address the multiple testing problem by adjusting the p values from a family of hypothesis tests, but it can easily be put at the service of the ART. The following SAS program is a simple alternative to Chen

Table 1
CPU Time Needed to Perform an Approximate Randomization Test Using the SAS (Version 6.04) Program 1 of Chen and Dunlap (1993) and the SCRT (Version 1.1) Program of Van Damme and Onghena (1993)

	SAS 6.04				SCRT 1.1*			
	AT286 (8 MHz)		80486 (50 MHz)		AT286 (8 MHz)		80486 (50 MHz)	
N	1,000	10,000†	1,000	10,000†	1,000	10,000	1,000	10,000
10	12m13s	—	1m02s	—	24s	4m03s	1s	9s
20	18m09s	—	1m23s	—	33s	5m37s	2s	23s
50	29m49s	—	2m28s	—	1m02s	10m26s	3s	29s
100	52m08s	—	4m21s	—	1m50s	18m30s	5s	45s
200	96m15s	—	7m57s	—	3m25s	34m21s	9s	1m28s

Note—The CPU time was assessed on an IBM PC/AT with 640K RAM running at 8 MHz under DOS 3.2 and on an IBM PC/80486 with 4 Mb RAM running at 50 MHz under DOS 5.0, with different numbers of observations (N) equally divided among two groups, for 1,000 and 10,000 pseudosamples. *CPU time calculated in fast mode (updating only timer). †Insufficient memory for 10,000 pseudosamples.

and Dunlap's (1993) Program 1 using the MULTTEST procedure:

```
DATA edging95;
  INPUT x y @@;
CARDS;
1 .33 1 .27 1 .44 1 .28 1 .45 1 .55 1 .44 1 .76 1 .59 1 .01
2 .28 2 .80 2 3.72 2 1.16 2 1.00 2 .63 2 1.14 2 .33 2 .26 2 .63
RUN;
PROC MULTTEST PERM NSAMPLE=10000;
  CLASS x;
  TEST MEAN(y);
  CONTRAST 'art' 1 -1;
RUN;
```

This alternative SAS program has the advantage of being simpler and, as shown in Table 2, more than 10 times faster. An SAS data step with the PROBBETA function may be added to obtain the confidence levels, as in the Chen and Dunlap (1993) program.

Unfortunately, with the MULTTEST procedure, it is also easy to elicit an adjusted p value (ADJ_P) of zero, indicating that the observed raw p value is not counted among the pseudovalues (see Section 1). Consequently, to have a valid ART, the final p value should be computed as $[(ADJ\_P * NSAMPLE) + 1]/(NSAMPLE + 1)$. For reasons of efficiency, it is recommended that $\alpha(NSAMPLE + 1)$ be an integer (Noreen, 1989, pp. 50–53). For example, $NSAMPLE = 9{,}999$ satisfies this condition for $\alpha = .01$, .05, and .10, and is large enough to have reliable p values (see Section 3). It should be remarked that the MULTTEST procedure is not available in the PC-DOS version of SAS.

6. The Exact Randomization Test

Chen and Dunlap (1993, p. 408) asserted that their program can be used for all group sizes. However, one should not encourage the use of the program for small numbers of observations (N not larger than 12 for equal group sizes, or, more general, if the number of possible permutations is smaller than NOP). The exact randomization test is not only more efficient in terms of number of permutations to be generated (as Chen and Dunlap, 1993, acknowledge) but, in general, also in terms of power (Onghena, 1994). Furthermore, exact randomization tests are easy to perform with NPSTAT (May et al., 1993; May et al., 1989) or SCRT (Onghena & Van Damme, 1994; Van Damme & Onghena, 1993).

Table 2
CPU Time Needed to Perform an Approximate Randomization Test Using the SAS (Version 6.08) Program 1 of Chen and Dunlap (1993) and the SAS (Version 6.08) MULTTEST Procedure on an IBM 3090/600e VF Mainframe Running Under the TSO Operating System

	Chen & Dunlap		MULTTEST	
N	1,000	10,000	1,000	10,000
10	3.15s	28.60s	0.28s	2.08s
20	4.17s	30.34s	0.34s	2.89s
50	6.98s	48.45s	0.65s	5.84s
100	12.52s	1m27.61s	1.44s	13.74s
200	24.73s	2m47.19s	4.13s	40.39s

Note—The CPU time was assessed with different numbers of observations (N) equally divided among two groups, for 1,000 and 10,000 pseudosamples.

7. Power and Type I Error Rates of the ART

After comparing randomization tests with their counterpart parametric and nonparametric methods in a random sampling model, Chen and Dunlap (1993, p. 407) concluded: "In summary, the ART procedures can be considered for hypothesis testing whenever the normality and/or the homoscedasticity assumptions appear to be violated." Although it is true that the ART procedures perform well without invoking the assumption of normality, they are not robust to violations of the assumption of equal variances. This is because nonparametric tests, in general, do not test the null hypothesis of identical population means, but they do test the null hypothesis of identical distributions. The nonrobustness of the ART to heterogeneity of variances was shown empirically by Boik (1987), and the general rationale, applying to all nonparametric tests, was pointed out explicitly by Edgington (1965).

Furthermore, the power superiority of the ART to the parametric and nonparametric competitors is not unequivocal. For example, Rasmussen (1986), Keller-McNulty and Higgins (1987), and van den Brink and van den Brink (1989) have shown that, under various non-normality conditions, the power curves of the ART and the t test were similar and that the Wilcoxon rank sum test had superior power.

In a random assignment model, power comparisons might give different results (see Lehmann, 1975, 1986, and May, Masson, & Hunter, 1990, for the distinction between random sampling and random assignment models). For example, the results of Kempthorne and Doerfler (1969), mentioned by Chen and Dunlap (1993, p. 407), were obtained under a random assignment model, and Edgington (1987) discussed randomization tests mainly from this perspective.

Conclusion

Several problems with performing ARTs using the SAS programs of Chen and Dunlap (1993) were identified and solutions were offered. On the technical side, it was pointed out that (1) the p value should be computed as $(NOS+1)/(NOP+1)$, (2) the test statistic should be changed to the signed difference between means if one wants to perform one-tailed ARTs in the case of unequal group sizes, and (3) the number of pseudosamples should be increased to at least 5,000 if one wants to test at the 1% level of significance. On the practical side, the programs take much time and resources in most applications, and some fast and economical alternatives (a lower level programming language, ready-made PC software packages for ART, and the SAS MULTTEST procedure) were suggested. Furthermore, for small group

sizes, the exact randomization test was proposed as a superior test, and it was pointed out that, in random sampling models with unequal population variances, the ART should not be used to test the null hypothesis of identical population means.

Our commentary on the Chen and Dunlap (1993) article does not imply that we question the ART procedure. On the contrary, we believe that the ART is the method of choice in many applied research situations, and we appreciate the contribution made by Chen and Dunlap (1993) to demonstrate how modern software can be used to perform these tests. The strength of exact and approximate randomization tests is particularly evident in situations where the random sampling assumption is not appropriate, with uncommon randomized designs, or with test statistics whose sampling distribution is unknown (Onghena, 1992; Onghena & Edgington, 1994). Furthermore, teaching statistical hypothesis testing from the resampling perspective with a clear distinction between random sampling and random assignment models may give students a better insight into the subject (May & Hunter, 1988; Simon & Bruce, 1991).

Finally, it should be acknowledged that Chen and Dunlap's (1993) intention was probably not to develop maximally efficient software. However, as access to these procedures grows through efforts such as theirs, there may be increasing concern about efficiency and precision. We hope that our refinements and suggestions make the ART appealing to a wider audience and that further discussion of the merits and demerits of ARTs is stimulated.

REFERENCES

BOIK, R. J. (1987). The Fisher-Pitman permutation test: A non-robust alternative to the normal theory F test when variances are heterogeneous. *British Journal of Mathematical & Statistical Psychology*, **40**, 26-42.

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

EDGINGTON, E. S. (1965). The assumption of homogeneity of variance for the t test and nonparametric tests. *Journal of Psychology*, **59**, 177-179.

EDGINGTON, E. S. (1987). *Randomization tests* (2nd ed.). New York: Marcel Dekker.

EFRON, B., & TIBSHIRANI, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

KELLER-MCNULTY, S., & HIGGINS, J. J. (1987). Effect of tail weight and outliers on power and type-I error of robust permutation tests for location. *Communications in Statistics: Simulation & Computation*, **16**, 17-35.

KEMPTHORNE, O., & DOERFLER, T. E. (1969). The behaviour of some significance tests under experimental randomization. *Biometrika*, **56**, 231-248.

LEHMANN, E. L. (1975). *Nonparametrics, statistical methods based on ranks*. San Francisco: Holden-Day.

LEHMANN, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York: Wiley.

MANLY, B. F. J. (1991). *Randomization and Monte Carlo methods in biology*. London: Chapman & Hall.

MAY, R. B., & HUNTER, M. A. (1988). Interpreting students' interpretations of research. *Teaching of Psychology*, **15**, 156-158.

MAY, R. B., HUNTER, M. A., & MASSON, M. E. J. (1993). *NPStat* (Version 3.7) [Computer program]. Department of Psychology, University of Victoria (Canada).

MAY, R. B., MASSON, M. E. J., & HUNTER, M. A. (1989). Randomization tests: Viable alternatives to normal curve tests. *Behavior Research Methods, Instruments, & Computers*, **21**, 482-483.

MAY, R. B., MASSON, M. E. J., & HUNTER, M. A. (1990). *Applications of statistics in behavioral research*. New York: Harper & Row.

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

NUMERICAL ALGORITHMS GROUP (1990). *The NAG Fortran Library* (Mark 14) [Computer Program]. Oxford, U.K.: Author.

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

ONGHENA, P. (1994). *The power of randomization tests for single-case designs*. Unpublished doctoral dissertation, Department of Psychology, Katholieke Universiteit, Leuven, Belgium.

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

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

RASMUSSEN, J. L. (1986). An evaluation of parametric and nonparametric tests on modified and non-modified data. *British Journal of Mathematical & Statistical Psychology*, **39**, 213-220.

SAS INSTITUTE INC. (1992). *SAS technical report P-229, SAS/STAT software: Changes and enhancements, release 6.07*. Cary, NC: Author.

SIMON, J. L., & BRUCE, P. (1991). Resampling: A tool for everyday statistical work. *Chance*, **4**, 22-32.

VAN DAMME, G., & ONGHENA, P. (1993). *Single case randomization tests* (Version 1.1) [Computer program]. Department of Psychology, Katholieke Universiteit, Leuven, Belgium.

VAN DEN BRINK, W. P., & VAN DEN BRINK, S. G. J. (1989). A comparison of the power of the t test, Wilcoxon's test, and the approximate permutation test for the two-sample location problem. *British Journal of Mathematical & Statistical Psychology*, **42**, 183-189.

WESTFALL, P. H., & YOUNG, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: Wiley.

(Manuscript received October 18, 1993;
revision accepted for publication May 16, 1994.)