An Aligned Rank Transform Procedure for Multifactor Contrast Tests

Lisa A. Elkin
Paul G. Allen School of Computer Science & Engineering,
University of Washington
Seattle, WA, USA
lelkin@cs.washington.edu

Matthew Kay
School of Communication,
Northwestern University
Evanston, IL, USA
mjskay@northwestern.edu

James J. Higgins
Department of Statistics,
Kansas State University
Manhattan, KS, USA
jhiggins@ksu.edu

Jacob O. Wobbrock
The Information School,
University of Washington
Seattle, WA, USA
wobbrock@uw.edu

ABSTRACT

Data from multifactor HCI experiments often violates the assumptions of parametric tests (i.e., *nonconforming data*). *The Aligned Rank Transform (ART)* has become a popular nonparametric analysis in HCI that can find main and interaction effects in nonconforming data, but leads to incorrect results when used to conduct *post hoc* contrast tests. We created a new algorithm called *ART-C* for conducting contrast tests within the ART paradigm and validated it on 72,000 synthetic data sets. Our results indicate that ART-C does not inflate Type I error rates, unlike contrasts based on ART, and that ART-C has more statistical power than a *t*-test, Mann-Whitney *U* test, Wilcoxon signed-rank test, and ART. We also extended an open-source tool called *ARTool* with our ART-C algorithm for both Windows and R. Our validation had some limitations (e.g., only six distribution types, no mixed factorial designs, no random slopes), and data drawn from Cauchy distributions should not be analyzed with ART-C.

KEYWORDS

Statistical methods; data analysis; experiments; quantitative methods; nonparametric statistics; aligned rank transform.

ACM Reference Format:
Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21), October 10–14, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3472749.3474784

UIST '21, October 10–14, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8635-7/21/10…$15.00
https://doi.org/10.1145/3472749.3474784

1 INTRODUCTION

Statistical procedures are a mainstay of quantitative human-computer interaction (HCI) research, particularly in the analysis of human performance data, like task times and error rates; subjective response data, like ordinal ratings and preference indications; and count data, like counts of participants, behaviors, or choices. To improve the soundness of conclusions drawn from HCI experiments, many in the HCI community have tried to improve upon the tools and methods we use to conduct our statistical analyses. For example, Jun et al. [21] and Wobbrock et al. [51] created new software tools for aiding statistical practice, Kay et al. [25] and Robertson and Kaptein [38] introduced more modern statistical methods, and Wobbrock et al. [51] and Kaptein et al. [23] developed statistical techniques for analyzing data commonly arising in HCI.

Parametric tests such as *t*-tests and ANOVAs are widely used in HCI, but when experiments give rise to data with residuals that are not normally distributed (i.e., *nonconforming data*), researchers and practitioners alike often turn to less familiar nonparametric tests. The *Aligned Rank Transform* (*ART*) [9, 16, 17, 40–42] is a nonparametric procedure that can properly assess both main and interaction effects in factorial designs. The ART pre-processes data with an "alignment step" [19, 35] and then applies midranks to that aligned data [8]. The resulting aligned-and-ranked data then can be analyzed with an omnibus test, typically an ANOVA. Since its introduction to HCI by Wobbrock et al. [51] in 2011, the ART procedure has quickly become a popular technique within HCI, and many HCI venues have published papers that use the ART in their analyses (e.g., *CHI* [2, 14, 15], *ASSETS* [3], *UIST* [22, 39]). Wobbrock et al.'s *ARTool* [51] has also been used in several fields outside HCI (e.g., cellular biology [7], dentistry [36], zoology [10], and cardiology [13]), and has been cited over 1000 times as of this writing.

Although Wobbrock et al. [51] mention in passing that the original ART's aligning and ranking procedure can be followed by contrast tests, a subsequent R package vignette by Kay [24] indicated that contrasts involving combinations of levels *across multiple factors* cannot be conducted on ART's aligned-and-ranked data without exploding Type I errors. As it turns out, the data after aligning-and-ranking are *not* properly aligned-and-ranked for

the multifactor contrasts. Rather, different alignment-and-ranking procedures must be carried out to enable correct contrast tests. Our work here contributes an algorithm for proper alignment-and-ranking for contrast tests. Further, our work completes and extends the open-source ART tools begun by Wobbrock et al. [51] for Windows, and Kay [24] for R.

In addition to these contributions, we formally validated that the original ART was incorrect for use with multifactor contrasts, and we validated the correctness of our new algorithm and tools. Inspired by the procedure presented in the aforementioned R package vignette [24], we devised a new procedure for nonparametric multifactor contrasts within the ART paradigm, *Aligned Rank Transform Contrasts*, or *ART-C*. ART-C uses an aligning-and-ranking procedure, similar to ART before it, but specific to contrast tests. To validate ART-C, we created 72,000 synthetic data sets using established statistical simulation procedures [1, 5, 27, 34]. A range of experimental designs, sample sizes, and distribution families were simulated. We then compared the Type I error rates of ART-C and *t*-tests [44]; we also compared the statistical power of ART-C to *t*-tests [44], Wilcoxon signed-rank tests [48], Mann-Whitney *U* tests [28], and the original ART procedure [16, 17, 41].

Our key findings are that when used to conduct contrast tests involving levels across multiple factors, the original ART's Type I error rates are generally far from their expected values. Further, ART's statistical power is low. By comparison, ART-C's Type I error rates are at their expected values and are generally not inflated. Also, for contrasts, ART-C has more statistical power than a *t*-test, Wilcoxon signed-rank test, Mann-Whitney *U* test, and the original ART procedure. Indeed, ART-C seems to be both correct and powerful by comparison.

That said, we also discovered that ART-C should not be used in cases where data appears to have been drawn from a Cauchy distribution.[1] Additionally, the 72,000 synthetic data sets used in our validation cover a wide range of experimental designs, but were not exhaustive. Our synthetic data were limited to two factors of up to three levels each, six types of population distributions, condition sample sizes between 8 and 40, and fully between-subjects or within-subjects designs, not mixed factorial (i.e., split-plot) designs. Moreover, like ART before it, ART-C is an alignment-and-ranking procedure; it is followed by a statistical test. Since ART is often followed by an ANOVA, for ART-C, we chose the *t*-test, but did not evaluate other tests.

As mentioned, to facilitate the use of our new ART-C procedure, we extended both the open-source Windows[2] and R[3,4] versions of *ARTool* [51]. Both tools are already in widespread use, and our extended versions integrate our new ART-C procedure for multifactor contrasts. Thus, HCI researchers and others who already use either tool can easily use our new versions. They neither have to risk incorrectly running multifactor contrasts on their aligned-and-ranked data, nor do they have to break from the ART paradigm to conduct *post hoc* contrast tests after a statistically significant omnibus test with ART.

The work reported here contributes: (1) a careful elucidation of the problem of multifactor contrast testing using the original ART procedure, now in regular use in HCI; (2) an algorithm, ART-C, for correctly aligning-and-ranking data for multifactor contrasts within the ART paradigm; (3) validation results from simulation studies showing the correctness and power of ART-C; and (4) significant extensions to the widely used ARTool Windows application and ARTool R package.

2 RELATED WORK

We created a multifactor contrast testing procedure within the ART paradigm, called ART-C, to enable multifactor contrasts for nonconforming data. Therefore, relevant prior research includes the ART procedure itself, the lack of a multifactor contrast testing method within the ART paradigm, and prior statistical contributions, including tools, directed at the HCI community.

2.1 The Aligned Rank Transform

Rank transforms have been explored in statistics for decades as a basis for nonparametric analyses (e.g., [12, 48]). Conover and Iman's [8] popular rank transform (RT) procedure applies midranks to responses and then conducts an ANOVA on ranks. However, although Type I error rates for main effects were reasonable under RT, they were drastically inflated for interactions. The aligned rank transform (ART) procedure was developed in response to this problem [9, 16, 17, 40–42]. With ART, responses are first "aligned" [19, 35] with respect to a specific main effect or interaction before midranks are applied. The upshot is that both main effects and interactions can be safely analyzed on aligned ranks using ANOVA-type procedures without inflating Type I error rates. Owing to (1) the prevalence of multifactor experiments in HCI, (2) the likelihood of data arising that do not conform to the assumptions of parametric analysis, and (3) the dearth of common statistical procedures to analyze such data, the need for the ART was evident. So in 2011, a paper at CHI was published [51] that offered *ARTool*, a Windows application capable of performing data alignment-and-ranking that would be otherwise tedious and error prone. In the decade since, this CHI paper has garnered over 1000 citations according to Google Scholar,[5] indicating the utility of ARTool. However, to the best of our knowledge, no prior publication (or tool) has offered a method for conducting contrast tests in the ART paradigm, an essential missing piece, particularly after detecting a statistically significant interaction. In this work, we supply this missing piece by devising ART-C and extending the open-source ARTool tools for both Windows and R.

2.2 Multifactor Contrasts

Using a single example, Kay [24] demonstrated that using aligned-and-ranked data from the original ART procedure [9, 16, 17, 40, 41] to conduct multifactor contrasts leads to incorrect results, a claim we validate below. A thorough search of the statistics literature did not uncover a suitable solution to the problem of multifactor

[1] Cauchy distributions are often used to "stress test" statistical analyses. As Wikipedia puts it, "The Cauchy distribution is often used in statistics as the canonical example of a 'pathological' distribution since both its expected value and its variance are undefined." http://en.wikipedia.org/wiki/Cauchy_distribution

[2] Windows code: http://depts.washington.edu/acelab/proj/art/

[3] R package: https://cran.r-project.org/package=ARTool

[4] R code: https://dx.doi.org/10.5281/zenodo.594511

[5] https://scholar.google.com/scholar?cites=16254127723353600671

contrast testing within the ART paradigm. Here we discuss the most closely related statistics work.

Indeed, some ART contrast methods have been presented in the literature [1, 4, 29–31], but the authors did not address how or whether their methods can be used across multiple factors, showing only examples of single-factor contrasts even in the presence of significant interaction effects. Simulation studies analyzing the effectiveness of ART contrasts also only included data with a single factor [1, 4, 34]. Furthermore, none of these prior works provided software tool support, as we do here.

Mansouri et al. [29] developed ART analogues of well-known contrast procedures (Tukey's HSD, Scheffé's method, Fisher's least significant difference procedure) for data with two factors. However, they did not specify or demonstrate whether their methods are applicable to multifactor contrasts; nonetheless, devising ART analogues to complex contrast procedures is not our objective here. Rather, we sought to find an aligning-and-ranking procedure that can be followed by a common contrast test, especially the familiar t-test.

Peterson et al. [34] compared the effectiveness of the ART using six different statistics in the alignment process (sample mean, sample median, lightly trimmed Winsorized mean, heavily trimmed Winsorized mean, Huber M-estimator, and Harrel-Davis estimator of the median). Rather than changing the statistic used for alignment, our method changes the alignment process itself, tailoring it to specified contrast tests. (Also, Peterson et al. did not test whether their methods can be used to analyze multifactor contrasts.)

2.3 Statistics-Related Research in HCI

Statistical analyses are foundational, but only meaningful when statistical tests are used correctly. HCI researchers are well positioned to improve the quality of results drawn from statistical analyses by looking at them through a usability lens. Tools that aid researchers in using statistical tests correctly can improve the quality of our quantitative practices. Wobbrock [50] and Wobbrock and Kay [52] have argued for the importance of nonparametric statistics in HCI. To that end, Wobbrock et al. [51] extended the ART procedure to multiple factors and provided a tool for carrying it out. Their *ARTool* Windows application made the ART easy to use in HCI and beyond. Kay subsequently created ARTool as an R package. (See footnote 3.)

Other researchers in HCI have also recognized the value of providing useful tools for statistical analyses. Jun et al. [21] provided *Tea*, a system in which users specify their study design and hypotheses at a high level, and then Tea figures out which tests to run, runs them, and returns the results, lowering the barrier to performing valid statistical tests. Kay et al. [25] took a different approach to user-centered statistics and looked at how using Bayesian analysis can help the HCI community accrue knowledge without having to conduct replication studies or meta-analyses, which conflict with the priority the HCI community places on novelty.

Another important aspect of usable statistics is their visibility and framing. Kay et al. [25], Wobbrock et al. [51], and Kaptein et al. [23] introduced methods from other fields into the HCI literature. Robertson and Kaptein's [38] book, *Modern Statistical Methods for HCI*, introduced the HCI community to modern statistical methods. Although none of these methods were wholly new, framing them for an HCI audience and curating them in an HCI book made them accessible to the HCI community, which might not otherwise discover them.

This work extends a method already embraced by the HCI community—the Aligned Rank Transform—making it more complete with ART-C for multifactor contrasts. This work also provides an updated version of *ARTool* for both Windows and R that make multifactor contrasts easy to conduct, lowering the barrier to performing correct statistical analyses within HCI and beyond.

3 THE PROBLEM: MULTIFACTOR CONTRASTS IN ART

The problem we address in this work is best explained with an example. We refer to this example as our "running example" throughout this paper. Let us consider a within-subjects experiment with three factors having two levels each: $A : \{A1, A2\}$, $B : \{B1, B2\}$, $C : \{C1, C2\}$, and response Y. There are 40 subjects and data for each condition is drawn from a log-normal distribution. Table 1 shows the log-scale true population means for each condition, and Figure 1 shows the resulting sample data.

Table 1: Log-scale population means for each condition in our running example.

Condition	A	B	C	Log-scale Population Mean
1	A1	B1	C1	0.00
2	A1	B1	C2	0.50
3	A1	B2	C1	0.00
4	A1	B2	C2	0.50
5	A2	B1	C1	0.75
6	A2	B1	C2	1.25
7	A2	B2	C1	1.00
8	A2	B2	C2	0.50

Suppose we analyze this data using the original ART procedure and an ANOVA. A significant main effect of A would tell us that the levels of A (i.e., $A1$ *vs.* $A2$) correspond to significantly different mean aligned-ranks of Y. A significant $A \times B$ interaction would tell us that the effect A has on aligned-and-ranked Y is significantly different for different levels of B, and vice-versa. Indeed, the original ART procedure works well for detecting main effects and interactions. But it lacks a suitable method for contrast tests. Contrast tests can tell us which levels of each factor cause these effects; they are commonly used to conduct *post hoc* pairwise comparisons following a statistically significant main effect or interaction. Contrast tests can also be used to compare levels of factors directly when warranted by the research question (i.e., "planned contrasts"). We use the term "single-factor contrasts" to refer to comparisons between levels within a single factor (e.g., $A1$ *vs.* $A2$), which is common after significant main effects, and "multifactor contrasts" to refer to comparisons between combinations of levels from multiple factors (e.g., $(A1, B2)$ *vs.* $(A2, B1)$), which is common after significant interactions.

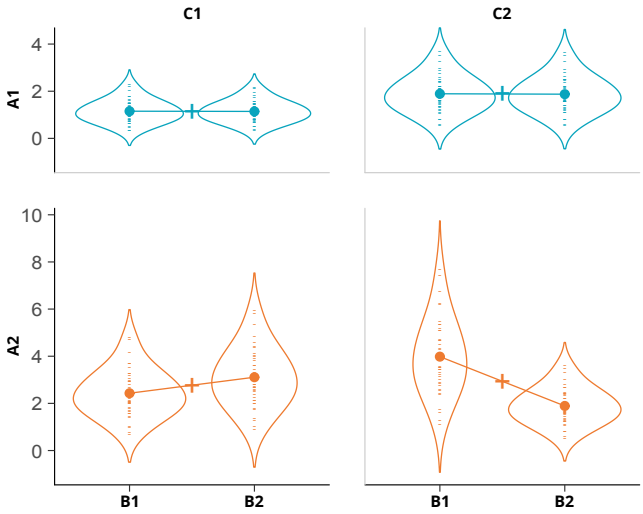

Figure 1: Sample data for each condition in our running example. Dots indicate condition means, lines connect condition means for visual comparison, and a plus indicates the mean of both connected conditions.

Single-factor contrasts *can* be conducted safely on data that has been aligned-and-ranked with the original ART procedure. However, conducting multifactor contrasts on data that has been aligned-and-ranked with the original ART procedure produces incorrect results. Our ART-C procedure for multifactor contrasts remedies this problem.

We demonstrate these issues using our running example. Since we *know* the data in our running example is drawn from a log-normal distribution, we fit a linear mixed model (LMM) [11, 46] to log-transformed data as a baseline, and fit an ART model to the original (not log-transformed) data. Specifically, we wish to compare levels in A and B, averaging over the levels of C. That is, C is not directly involved in the contrasts. This is achieved in R using the following code. (Note that S is the subject identifier, which is given a random intercept to account for repeated measures.)

```
# Fit LMM to log−transformed response
m.lmm = lmer(log(Y) ~ A*B*C + (1|S), data=df)

# Fit ART model to raw response
m.art = art(Y ~ A*B*C + (1|S), data=df)

# Conduct A x B contrasts on LMM
contrast(emmeans(m.lmm, ~ A:B), method="pairwise",
    adjust="holm")

# Conduct A x B contrasts on ART model
contrast(emmeans(artlm(m.art, "A:B"), ~ A:B),
    method="pairwise", adjust="holm")
```

We created the data such that there is no difference between $(A1, B1)$ and $(A1, B2)$, and there *is* a difference between $(A1, B1)$ and $(A2, B2)$ (Table 1). Comparing the two tests above, LMM contrasts produce results that match the true effects ($A1, B1 - A1, B2: p = .1792$, i.e., no significant difference) and ($A1, B1 - A2, B2: p < .0001$, i.e., a significant difference) (Table 2), but ART contrasts result in a *Type I error* (i.e., finding a significant difference when there is no true difference) ($A1, B1 - A1, B2: p < .0001$), and a *Type II error* (i.e., not finding a significant difference when there *is* a true difference) ($A1, B1 - A2, B2: p = .9144$) (Table 3, Figure 2).

Table 2: Highlighted results of contrasts conducted on an LMM of log-transformed responses, comparing levels of A and B in our running example. In the top row, a significant difference was correctly not detected between $(A1, B1)$ and $(A1, B2)$ ($p = .1792$), and indeed, there is no true difference. In the bottom row, a difference was correctly detected between $(A1, B1)$ and $(A2, B2)$ ($p < .0001$), and indeed, there *is* a true difference.

contrast	estimate	SE	df	t.ratio	p.value
(A1, B1) – (A1, B2)	0.0	0.0	273	1.3	0.1792
(A1, B1) – (A2, B2)	-0.5	0.0	273	-165.9	<.0001

Table 3: Highlighted results of contrasts conducted on ART data, comparing levels of A and B in our running example. In the top row, a difference was incorrectly detected between $(A1, B1)$ and $(A1, B2)$ ($p < .0001$), but there is no true difference. In the bottom row, a difference was incorrectly *not* detected between $(A1, B1)$ and $(A2, B2)$ ($p = .9144$), but there *is* a true difference.

contrast	estimate	SE	df	t.ratio	p.value
(A1, B1) – (A1, B2)	-43.2	3.8	273	-11.3	<.0001
(A1, B1) – (A2, B2)	0.4	3.8	273	0.1	0.9144

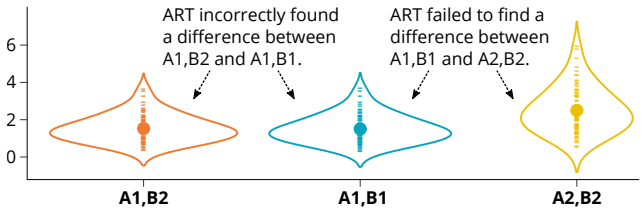

Figure 2: Sample data for $(A1, B2)$, $(A1, B1)$, and $(A2, B2)$. Dots indicate condition means. Contrasts on data from the original ART procedure found a difference between $(A1, B2)$ ad $(A1, B1)$ even though there was no true difference. Furthermore, no difference was found between $(A1, B1)$ and $(A2, B2)$ even though there *was* a true difference.

Obviously, we cannot judge the validity of a statistical procedure from just one example! Therefore, we assessed the correctness of the original ART procedure when used in multifactor contrasts on 72,000 synthetic data sets representing several different experimental designs and data distributions, confirming the issues illustrated

by our running example. (Further details on our simulation procedure are given below.) Our results show: (1) using the the original ART procedure to conduct multifactor contrasts on data drawn from log-normal, Cauchy, or exponential distributions produces inflated Type I error rates (Figure 3); (2) conducting contrasts with the original ART procedure on data drawn from *any* distribution produces Type I error rates that are far from their expected value ($\alpha = .05$), either too high or too low (Figure 3); and (3) for multifactor contrasts, the original ART procedure has low statistical power (high Type II error rates) (Figure 4). Single-factor contrasts conducted with the the original ART procedure are, in fact, correct; indeed, our new method ART-C reduces to the same mathematical formula as the original ART in the single-factor case. ART-C results are included in Figures 3 and 4 for comparison. We show ART-C's derivation and validation below.

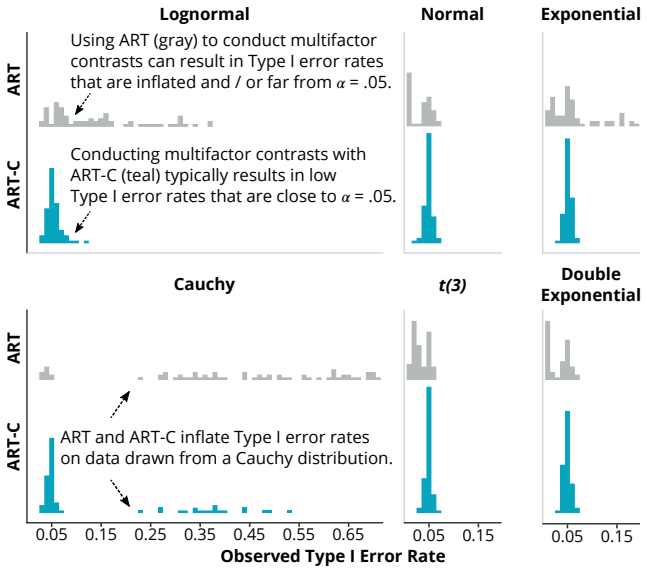

Figure 3: ART Type I error rates (gray) compared to ART-C Type I error rates (teal). Each data point represents the observed Type I error rate of one "design," explained below. Values closer to $\alpha = .05$ are better, indicating greater correctness. ART-C Type I error rates are closer to .05 for all distributions.

4 THE SOLUTION: THE ART-C PROCEDURE FOR ART CONTRASTS

To address the limitations of ART described above, and to extend ART so as to make it more complete as an analysis, we developed *ART-C*, a procedure to conduct nonparametric multifactor contrasts within the ART paradigm. ART-C offers an alignment process specific to contrast tests involving one or more factors. Like ART before it for main effects or interactions, ART-C first aligns data and then ranks it with ascending midranks. Then, multifactor contrast tests can be conducted to compare combinations of factors' levels for which the data was aligned-and-ranked. Thus, the process is much like the original ART procedure, but the data is aligned not for main effects and interactions, but for intended contrast tests.

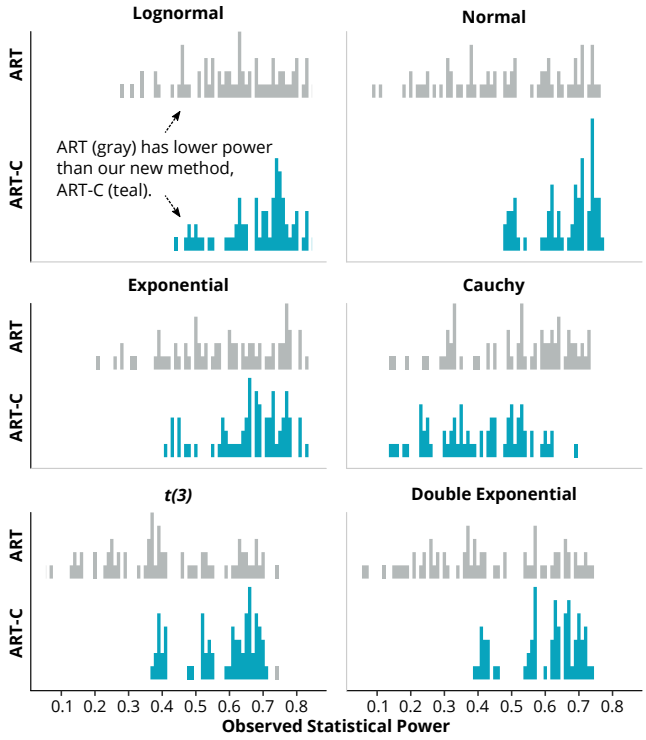

Figure 4: ART statistical power (gray) compared to ART-C statistical power (teal). Each data point represents the observed statistical power of one "design," explained below. Larger values are better, indicating greater power. ART-C has greater power for all distributions except Cauchy.

With ART-C, data must be aligned-and-ranked for each set of factors whose levels will be compared. In our running example, we found an $A \times B$ interaction effect. Response Y must be aligned-and-ranked to compare combinations of levels of A and B. And, for example, had we found an $A \times C$ interaction effect and wanted to conduct contrasts tests thereof, response Y would have to be aligned-and-ranked separately to compare combinations of levels of A and C.

Like the original ART procedure, ART-C can be used on nonconforming data, i.e., data that do not meet the parametric ANOVA assumptions. For example, responses that are not conditionally normal, or, equivalently, residuals that are not normally distributed, can be handled correctly by both ART and ART-C. Note, however, that the original ART procedure has been shown to inflate Type I error rates on heteroscedastic data [37]. We did not examine heteroscedasticity specifically for ART-C, but given ART-C's underlying similarity to ART, the same observation probably holds.

4.1 ART-C Procedure for Multifactor Contrasts

In this section, we walk through the ART-C procedure with an example, similar to our running example, with three factors: A with levels $A_i, i = 1...a$, B with levels $B_j, j = 1...b$, and C with levels $C_k, k = 1...c$, and response Y. We present the ART-C procedure in four steps:

Step 1. Prepare data. To prepare data for ART-C:

1. "Concatenate" the factors of interest to create a new factor. For example, when conducting contrasts on A and B, we concatenate A and B and create a new factor labeled AB. For any response Y for which A has level A_i and B has level B_j, AB has level AB_{ij}.
2. Remove original copies of the factors involved (here, A and B).
3. Keep unchanged any factors not concatenated in the contrasts (here, C).

Step 2. Compute aligned response Y'. Regardless of which factors were concatenated in Step 1.1 and which original factors were removed in Step 1.2, Y_{ijk} denotes all responses Y where A had level A_i, B had level B_j, and C had level C_k *before* Step 1.1 was completed. As we will see, sometimes, the levels of all factors are taken into account when aligning Y_{ijk}. For example, $\overline{AB_{ij}C_k}$ denotes the mean of all responses where the new concatenated factor AB has level AB_{ij} and factor C has level C_k. Other times, we only care about the levels of the concatenated factor. For example, $\overline{AB_{ij}}$ denotes the mean of all responses where AB has level AB_{ij}, regardless of the level of C. In all cases, μ denotes the grand mean (i.e., the mean of all responses Y_{ijk}).

In our running example, there are three possible types of contrasts: three-factor contrasts, two-factor contrasts, and single-factor contrasts. We present the ART-C alignment formula for all three types of contrasts and then present the general case. As an example, Table 4 shows a small subset of sample calculations for two-factor contrasts ($A \times B$) in a three-factor model ($A \times B \times C$).

Three-factor contrasts in a three-factor model. To align response Y_{ijk} for contrasts between levels of factors A, B, and C, compute:

$$\begin{aligned} Y'_{ijk} &= Y_{ijk} - \overline{ABC_{ijk}} + \overline{ABC_{ijk}} - \mu \\ &= Y_{ijk} - \mu. \end{aligned}$$

Two-factor contrasts in a three-factor model. To align response Y_{ijk} for contrasts between levels of factors A and B, compute:

$$Y'_{ijk} = Y_{ijk} - \overline{AB_{ij}C_k} + \overline{AB_{ij}} - \mu$$

Single-factor contrasts in a three-factor model. The focus of this work is not on single-factor contrasts since the original ART alignment procedure can be used for single-factor contrasts, but it is worth noting that our method is mathematically equivalent to the ART in the single-factor case. To align response Y_{ijk} for contrasts between levels of factor A compute:

$$Y'_{ijk} = Y_{ijk} - \overline{A_iB_jC_k} + \overline{A_i} - \mu$$

General Case: M-factor contrasts in an N-factor model. We need more complex notation to describe the general case. In the general case, we align response $Y_{ij\ldots n}$ for contrasts between levels of M factors in an N-factor model. In the example above, we named our factors A, B, and C. Here, we name them X_1, X_2, ..., X_N and denote level j of factor X_i as $X_{i,j}$ (e.g., level 2 of factor X_1 is denoted as $X_{1,2}$). In Step 1.1, we concatenated the M factors for which we were aligning the response to create a new factor $X_1X_2\ldots X_M$. The level of factor $X_1X_2\ldots X_M$ that was created by concatenating $X_{1,i}$, $X_{2,j}$, ..., $X_{M,m}$ is denoted as $(X_1X_2\ldots X_M)_{ij\ldots m}$. In Step 1.2, we removed the original copies of the M factors concatenated in Step 1.1. So, after Step 1.3, there are $N - M$ other non-concatenated factors in the model denoted $X_{M+1}, X_{M+2}, \ldots, X_N$. Thus, $X_{M+1,m+1}$ denotes a level of factor X_{M+1}; $X_{M+2,m+2}$ denotes a level of factor X_{M+2}; and $X_{N,n}$ denotes a level of factor X_N. With this notation in hand, to align the data in the general case, we compute:

$$Y'_{ij\ldots n} = Y_{ij\ldots n} - \overline{(X_1X_2\ldots X_M)_{ij\ldots m}X_{M+1,m+1}X_{M+2,m+2}\ldots X_{N,n}} + \overline{(X_1X_2\ldots X_M)_{ij\ldots m}} - \mu$$

For example, with this notation, our "two-factor contrasts in a three-factor model" formula would be:

$$Y'_{ijk} = Y_{ijk} - \overline{(X_1X_2)_{ij}X_{3,k}} + \overline{(X_1X_2)_{ij}} - \mu$$

Step 3. Compute ranked response Y''. Apply midranks to all aligned values Y' in ascending order to create aligned-and-ranked responses Y'' (see example in Table 4). That is, the smallest Y' is given rank Y'' = 1, the next smallest Y' is given rank Y'' = 2, until all Y' values have been assigned a rank. If there is a tie among k values, the mean of the next k ranks that would have been assigned is used as the rank for all k values (i.e., midranks). For example, if there is a tie between the third and fourth smallest Y', they would both be assigned rank Y'' = (3 + 4) / 2 = 3.5. This is a standard application of applying ascending midranks to data.

Table 4: Sample calculations to compute aligned response Y' and aligned-and-ranked response Y'' for two-factor contrasts ($A\times B$) in a three-factor model ($A\times B\times C$) using the ART-C procedure. The concatenated factor made from $A_i \times B_j$ is AB_{ij}. Factor C is not involved in the intended contrast tests, so is left unchanged. The grand mean is $\mu = 5$. Only 4 of 8 conditions are shown here, where $A = A_1$ only, for considerations of space.

A	B	AB	C	Y	Y'	Y''
A_1	B_1	AB_{11}	C_1	7	$7 - \frac{7+5}{2} + \frac{7+5+2+2}{4} - 5 = 0$	5.5
A_1	B_1	AB_{11}	C_1	5	$5 - \frac{7+5}{2} + \frac{7+5+2+2}{4} - 5 = -2$	1
A_1	B_1	AB_{11}	C_2	2	$2 - \frac{2+2}{2} + \frac{7+5+2+2}{4} - 5 = -1$	3
A_1	B_1	AB_{11}	C_2	2	$2 - \frac{2+2}{2} + \frac{7+5+2+2}{4} - 5 = -1$	3
A_1	B_2	AB_{12}	C_1	10	$10 - \frac{10+8}{2} + \frac{10+8+5+1}{4} - 5 = 2$	7
A_1	B_2	AB_{12}	C_1	8	$8 - \frac{10+8}{2} + \frac{10+8+5+1}{4} - 5 = 0$	5.5
A_1	B_2	AB_{12}	C_2	5	$5 - \frac{5+1}{2} + \frac{10+8+5+1}{4} - 5 = 3$	8
A_1	B_2	AB_{12}	C_2	1	$1 - \frac{5+1}{2} + \frac{10+8+5+1}{4} - 5 = -1$	3

Step 4. Conduct contrast tests on Y''. As with the original ART procedure, ART-C is an alignment-and-ranking procedure that is meant to be followed by statistical tests on aligned ranks. The ART procedure is commonly followed by an ANOVA [51]; similarly, our ART-C procedure can be followed by a t-test, since two means are being compared in the case of contrasts. Despite ANOVAs and t-tests being parametric, owing to their application to aligned ranks, the ART and ART-C procedures can be considered nonparametric in nature.

Post hoc contrast tests with ART-C are justified when the original ART procedure results in significant main effects or interactions. However, as stated above, contrasts do not need to follow significant omnibus tests if warranted by the research question (i.e., "planned

contrasts"). Also, note that conducting multiple *post hoc* tests should be accompanied by a p-value correction for multiple comparisons (e.g., with the Bonferroni correction [47], Holm's sequential Bonferroni procedure [20], or Tukey's HSD test [45], to name a few). In the ART-C procedure, contrasts should be conducted using the full factorial model comprising all factors that remain after Step 1.3, but only the results of comparisons between levels of the concatenated factor created in Step 1.1 can be interpreted; comparisons between levels of any non-concatenated factors are meaningless.

Returning to our running example of conducting contrasts to compare levels of A and B, we have factors AB and C, and have computed Y'' as aligned ranks for factor AB. We would therefore conduct contrasts using a full-factorial model with factors AB and C (e.g., $Y \sim AB \times C$). We ignore the omnibus test results for this model, but we follow it with contrasts among desired levels of AB. Contrasts that would involve C are meaningless.

5 VALIDATING OUR APPROACH

In this section, we describe how we validated our ART-C procedure for multifactor contrasts. As is customary, we examined Type I error rates and statistical power. We conducted our validation consistent with simulation-based validations from the statistics literature [1, 5, 27, 34].[6]

5.1 Generating Synthetic Data

To create our 72,000 synthetic data sets, we drew responses as random samples from known populations. We use the term "condition" to refer to combinations of levels from any number of factors. Each sample was drawn for conditions comprising one level of each factor. Our synthetic data sets varied according to the following four properties:

- Layout: The number of factors and number of levels per factor in the data set. Values: two factors with two levels each (2×2), two factors with three levels each (3×3), and three factors with two levels each ($2 \times 2 \times 2$). We chose these layouts owing to their commonality in HCI studies.
- Population Distribution: The type of distribution from which samples in the data set were drawn. Specific distributions (see Table 5) were chosen because they represent data frequently found in HCI studies (e.g., normal, log-normal, exponential), or because they are commonly used in simulation studies in statistics due to their heavy tails [1, 5] (e.g., Cauchy, t with 3 degrees of freedom, double exponential). Note that the mean is a type of location and the standard deviation is a type of scale; for consistency, we use the general terms "location" and "scale."
- Condition Sample Size: The number of data points randomly sampled from a population for each condition. Values: 8, 16, 24, 32, and 40, selected because they represented typical sample sizes in HCI.
- Between- or Within-Subjects: In a between-subjects design, each subject contributes one response to the data set, and the number of responses is equal to the number of subjects. In a within-subjects design, each subject contributes one response in each condition, and the number of subjects is equal to the condition sample size. Values: "between" or "within." (Mixed factorial designs, i.e., "split-plots," were left for future work.)

[6] In statistics, validation of tests or procedures is generally achieved via either asymptotic theory or simulation studies. For the latter, myriad data sets whose properties are known are synthesized. Then, the results of the tests or procedures in question are compared to expected results, revealing Type I error rates and statistical power.

Table 5: Population Distributions and their parameters.

Distribution	Parameters
Normal	Mean, standard deviation
Log-normal	Log mean, log standard deviation
Exponential	Rate
Cauchy	Location, scale
$t(3)$	Location, scale
Double Exponential	Location, scale

Our running example has a $2 \times 2 \times 2$ layout, condition sample size of 40, log-normal population distribution, and is within-subjects. For each of the $3 \times 6 \times 5 \times 2 = 180$ combinations of property values, we generated approximately 200 data sets in which all conditions in one data set were sampled from population distributions with equal locations, and approximately another 200 data sets in which all population distributions' locations were randomly chosen. Population distributions' scales were equal to 1. We describe our synthetic data-generating process in the following four steps:

Step 1. Determine latent location. We begin by determining a *latent location* for each condition (μ_c^*), which will undergo several transformations before being used as a parameter value in a population distribution. When conditions have equal population locations, the latent location is fixed at 0 (Equation (1a)). Otherwise, its value is sampled from a standard normal distribution (Equation (1b)). Scale is always equal to 1 in our analyses (Equation (2)).

$$\mu_c^* = 0 \tag{1a}$$

Used when creating data to measure Type I error rate.

$$\mu_c^* \sim \mathcal{N}(0, 1) \tag{1b}$$

Used when creating data to measure statistical power.

$$\sigma_c = 1 \tag{2}$$

In Equations (1a) and (1b), μ_c^* is the latent location for condition c, and in Equation (2), σ_c is the scale for condition c.

Step 2. Add random intercepts per subject. When generating within-subjects data, each subject is assigned a unique random offset (β_s) sampled from a normal distribution with mean 0, and standard deviation SD (Equation (3b)), where SD is randomly chosen from $\{0.1, 0.5, 0.9\}$ (Equation (3a)) and is the same value for the entire data set. These values were chosen to represent a reasonable ratio between within-subject variance and between-subject variance. We now update our latent mean notation to ($\mu_{c,s}^*$) to represent the latent mean for each combination of condition and subject, and a subject's random offset is added to all of its associated latent locations (Equation (3c)). For consistency, we use this notation for

between-subjects data as well, but with a random per-subject offset of 0 (Equation (3d)).

$$SD \sim Random(0.1, 0.5, 0.9) \tag{3a}$$
$$\beta_s \sim \mathcal{N}(0, SD) \tag{3b}$$
$$\mu^*_{c,s} = \mu^*_c + \beta_s \tag{3c}$$

Used when generating within-subjects data

$$\mu^*_{c,s} = \mu^*_c + 0 \tag{3d}$$

Used when generating between-subjects data.

Equations (3a), (3b), and (3c) are used for within-subjects data. In Equation (3b), β_s is the random offset for subject s, and in Equation (3c), β_s is added to all latent locations μ^*_c associated with subject s, resulting in a new latent location $\mu^*_{c,s}$ for condition c and subject s. In Equation (3d), μ^*_c is simply relabeled $\mu^*_{c,s}$ for consistency; however, each subject still has the same latent location μ^*_c for condition c.

Step 3. Transform latent location with an inverse link function. Latent location is currently expressed as a linear model, but some distributions' parameters must be expressed as a *function* of a linear model. This function, customarily termed the "inverse link function" (g^{-1}), transforms the latent location (μ^*) into the appropriate location for the distribution (μ) (Equation (4)).

$$\mu_{c,s} = g^{-1}(\mu^*_{c,s}) \tag{4}$$

In Equation (4), g^{-1} is the inverse link function and it transforms a latent location $\mu^*_{c,s}$ into a location $\mu_{c,s}$.

All POPULATION DISTRIBUTIONS (Table 5) use the identity inverse link function (Equations 5a) except for the exponential distribution (Equation 5b).

$$g^{-1}_{id}(x) = x \tag{5a}$$
$$g^{-1}_{exp}(x) = exp(x) \tag{5b}$$

In Equation (5a), g^{-1}_{id} is the identity inverse link function. In Equation (5b), g^{-1}_{exp} is the inverse link function used by the exponential distribution.

Step 4. Generate data. Response $Y_{c,s}$ is sampled from the relevant distribution, represented here by the generic function $Distribution(x, y)$ (Equation (6)). The exponential distribution only has a single parameter ($rate = 1/location$) and follows Equation (7):

$$Y_{c,s} \sim Distribution(\mu_{c,s}, \sigma_{c,s}) \tag{6}$$
$$Y_{c,s} \sim Exp(1/\mu_{c,s}) \tag{7}$$

5.2 Example of Generating Synthetic Data

In this section, we illustrate the four steps of the above data-generating process by generating response $Y_{5,2}$ for *condition 5* and *subject 2* in our running example.

Step 1. Since our example does not have equal population locations, we use Equations 1b and 2.

$$\mu^*_5 \sim \mathcal{N}(0, 1) = 0.75 \tag{1b}$$
$$\sigma_5 = 1 \tag{2}$$

Step 2. Our example uses a within-subjects design, so we add per-participant offsets. Note that *SD* would have already been chosen for *condition 1*. The same value would be used here.

$$SD \sim Random(0.1, 0.5, 0.9) = 0.5 \tag{3a}$$
$$\beta_2 \sim \mathcal{N}(0, 0.5) = 0.1 \tag{3b}$$
$$\mu^*_{5,2} = 0.75 + 0.1 = 0.85 \tag{3c}$$

Step 3. We use the inverse link function for the log-normal distribution, which is the identity function.

$$\mu_{5,2} = g^{-1}_{id}(0.85) = 0.85 \tag{5a}$$

Step 4. Finally, we sample a log-normal distribution with log mean $\mu_{5,2}$ and log standard deviation $\sigma_{5,2}$ to get response $Y_{5,2}$.

$$Y_{5,2} \sim Lognormal(0.85, 1) = 3.27 \tag{6}$$

5.3 Testing Procedure

To explain our testing procedure, we first introduce some definitions:

- An *x-factor contrast* is a contrast between two conditions composed of one level each from x factors.
- CONTRAST SIZE is the x in x-factor contrast.
- A *design* as a unique combination of a LAYOUT, POPULATION DISTRIBUTION, CONDITION SAMPLE SIZE, BETWEEN- or WITHIN-SUBJECTS, and CONTRAST SIZE.
- A *trial* consists of one contrast test result, and all possible contrasts were conducted. There were:
  - 8 trials in a data set with a 2×2 LAYOUT
  - 42 trials in a data set with a 3×3 LAYOUT
  - 49 trials in a $2 \times 2 \times 2$ LAYOUT.

There were 72,000 total data sets split evenly among designs with each LAYOUT (24,000 data sets each). Thus, there were a total of $24,000 \times (8 + 42 + 49) = 2{,}376{,}000$ trials. There were 1,094 data sets out of 72,000 data sets (1.5%) with at least one trial for which ART-C did not converge. All of these data sets were within-subjects and were modeled as linear mixed models using the *lmer* method in the R package *lme4*; it is not uncommon for *lmer* to fail to converge. In any case, these data sets were removed.

By definition, ART-C is an aligning-and-ranking procedure followed by a contrast testing method—our validation used a *t*-test. Since we were validating a contrast testing method and not investigating the cause of a significant omnibus test, we did not correct for multiple comparisons. The R programming language was used to generate all data sets, conduct all contrasts, and analyze the results. All R code is included as supplementary material and is available online for replication and extension.[7]

Following the common approach in the statistics literature [1, 5, 27, 32–34], we validated our method on two metrics: Type I error rate and power.

[7] http://www.doi.org/10.5281/zenodo.4536432

5.4 Type I Error Rate

A significance level (α) represents the probability of a Type I error (false positive) and is used as a threshold to reject a null hypothesis ($p < \alpha$). Many readers will recognize that typically, α is set to .05, although other values may be used. A large-scale simulation such as ours shows that a method is correct when the proportion of tests in which a true null hypothesis was rejected (the observed Type I error rate) is close to the significance level α. That is, the proportion of tests in which p was less than α should be close to α.

For example, 5,516 trials were conducted on data from a $2 \times 2 \times 2$ LAYOUT, log-normal POPULATION DISTRIBUTION, CONDITION SAMPLE SIZE 40, WITHIN-SUBJECTS, CONTRAST SIZE of three, and no differences between conditions' population locations. Using a significance level of .05, ART-C found a significant difference in 265 trials, resulting in a $265 \div 5516 = .048$ observed Type I error rate, which is very close to the $\alpha = .05$ significance threshold, indicating the correctness of ART-C for this design.

Each data point in the following results represents the observed Type I error rate for one design. All population locations were set to 0 (Equation (1a)), and thus, the null hypothesis that there is no true difference between conditions' population locations is true for all trials. As is common practice in statistics, we include observed Type I error rates for the t-test as a baseline [32, 33].

Table 6: Mean Type I error rates (and standard deviations) for ART-C and, for comparison, the t-test, grouped by CONTRAST SIZE and LAYOUT over all designs, excluding designs with a Cauchy POPULATION DISTRIBUTION. Recall that CONTRAST SIZE refers to the number of factors whose levels are involved in the contrast test. Results for ART-C and the t-test that are closer to .05 indicate greater correctness. ART-C has comparable Type I error rates to the t-test, but as our additional results show, much greater power.

CONTRAST SIZE	LAYOUT	ART-C	t-test
1	2×2	.046 (.011)	.040 (.014)
1	3×3	.053 (.009)	.048 (.007)
1	$2 \times 2 \times 2$	.057 (.016)	.048 (.009)
2	2×2	.048 (.008)	.042 (.009)
2	3×3	.050 (.004)	.058 (.016)
2	$2 \times 2 \times 2$	.048 (.005)	.045 (.007)
3	$2 \times 2 \times 2$	.049 (.004)	.054 (.011)

Contrasts conducted with ART-C on designs with a CONTRAST SIZE of one, Cauchy POPULATION DISTRIBUTION, and 3×3 or $2 \times 2 \times 2$ LAYOUT had inflated Type I error rates ($M = .373$, $SD = .076$), while t-test contrasts did not ($M = .025$, $SD = .004$). Those Cauchy designs were considered outliers and were not included in the remainder of our analysis of Type I error rates; we address this further in our discussion.

Results show that observed Type I error rates for contrasts conducted with both ART-C and the t-test on remaining designs were clustered around .05: ART-C ($M = .050$, $SD = .009$) and t-test ($M = .048$, $SD = .012$), and design properties do not appear to have an effect on observed Type I error rates, confirming the robustness of the ART-C procedure. Observed Type I error rates for all designs are included as supplementary material. Table 6 and Figure 5 illustrate both methods' observed Type I error rates, closely clustered around .05, and show high Type I error rates for ART-C with a Cauchy distribution.

Additional results are included in supplementary material and available online (see Footnote 7).

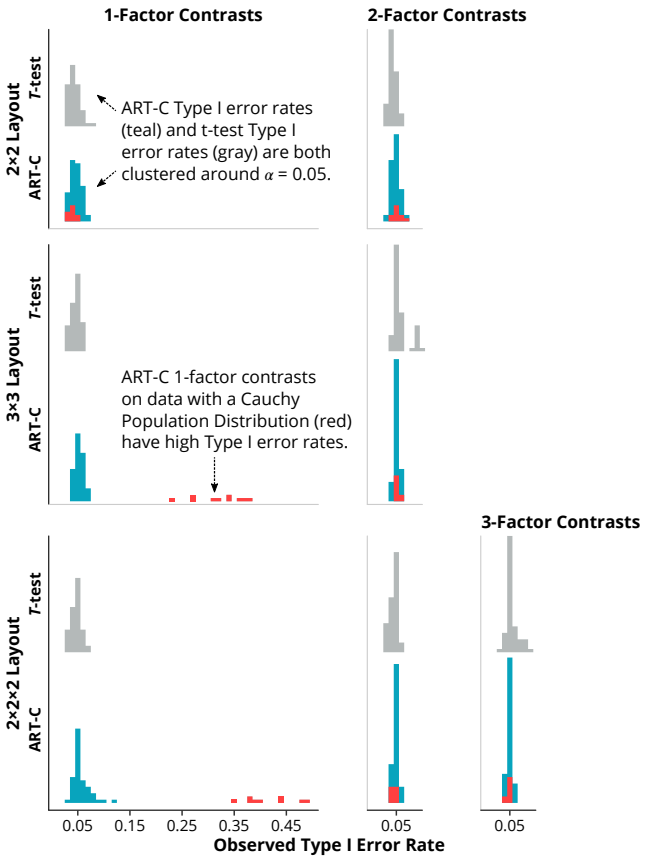

Figure 5: ART-C (teal) and t-test (gray) observed Type I error rates by CONTRAST SIZE and LAYOUT. Designs with a Cauchy POPULATION DISTRIBUTION are shown in red. Each point represents an observed Type I error rate for one design. Values closer to .05 indicate greater correctness. ART-C has comparable Type I error rates to the t-test, but as our results below show, much greater power.

5.5 Power

Statistical power is the probability of rejecting a false null hypothesis (detecting a true difference) given a particular significance level. Observed power is the proportion of tests in which a false null hypothesis was rejected. Unlike for Type I errors, there is no expected value to compare observed power to; instead, we followed common practice in statistics [1, 5, 27] and compared to other methods, specifically the t-test [44], Mann-Whitney U test [28] for between-subjects designs, and Wilcoxon signed-rank test [48] for within-subjects designs.

For example, ART-C contrast tests conducted on data with a $2 \times 2 \times 2$ LAYOUT, log-normal POPULATION DISTRIBUTION, CONDITION SAMPLE SIZE 40, WITHIN-SUBJECTS, CONTRAST SIZE of three, and different population locations for each condition detected a true significant difference in 4408 out of 5824 trials, therefore having $4408 \div 5824 = .76$ observed power.

POPULATION DISTRIBUTION latent locations for each condition were randomly sampled from a standard normal distribution (Equation (1b)). Although we cannot guarantee these locations were different, there is an infinitely small chance they were the same, and we therefore assume that the null hypothesis of no difference between condition population locations is false. In the following results, a significance level of $\alpha = .05$ was used, and each data point represents the observed power of one design.

When averaged over all designs, our results show that ART-C had the highest observed power ($M = .598$, $SD = .143$), followed by Mann Whitney U test / Wilcoxon signed-rank test ($M = .521$, $SD = .149$), and finally the t-test ($M = .461$, $SD = .149$). Observed powers for all designs are included as supplementary materials.

POPULATION DISTRIBUTION and CONDITION SAMPLE SIZE were the only design properties that had a large impact on observed power. ART-C had higher observed power than the t-test for all POPULATION DISTRIBUTIONS other than the normal distribution, for which it was the same, and had higher observed power than the Mann-Whitney U test and Wilcoxon signed-rank test for all POPULATION DISTRIBUTIONS (Table 7, Figure 6).

ART-C had higher observed power than the t-test and Mann-Whitney U test / Wilcoxon signed-rank test regardless of CONDITION SAMPLE SIZE, but all tests' power increased as CONDITION SAMPLE SIZE increased, which is expected.

Additional results are included in supplementary material and available online (see Footnote 7).

Table 7: Mean statistical power (and standard deviations) for ART-C, *t*-test, Mann-Whitney *U* test (M-W) / Wilcoxon signed-rank test (WSR), and ART, grouped by POPULATION DISTRIBUTION. Higher values indicate more statistical power. ART-C has the highest power of all tests on all POPULATION DISTRIBUTIONS except Cauchy.

	Normal	Log-normal	Exponential
ART-C	.66 (.09)	.69 (.10)	.66 (.11)
t-test	.66 (.09)	.46 (.11)	.52 (.12)
M-W/WSR	.59 (.12)	.58 (.12)	.56 (.12)
ART	.49 (.18)	.62 (.14)	.58 (.16)
	Cauchy	***t*(3)**	**Double Exponential**
ART-C	.41 (.13)	.58 (.11)	.60 (.10)
t-test	.07 (.02)	.50 (.10)	.56 (.10)
M-W/WSR	.34 (.12)	.52 (.13)	.54 (.12)
ART	.51 (.16)	.42 (.18)	.44 (.19)

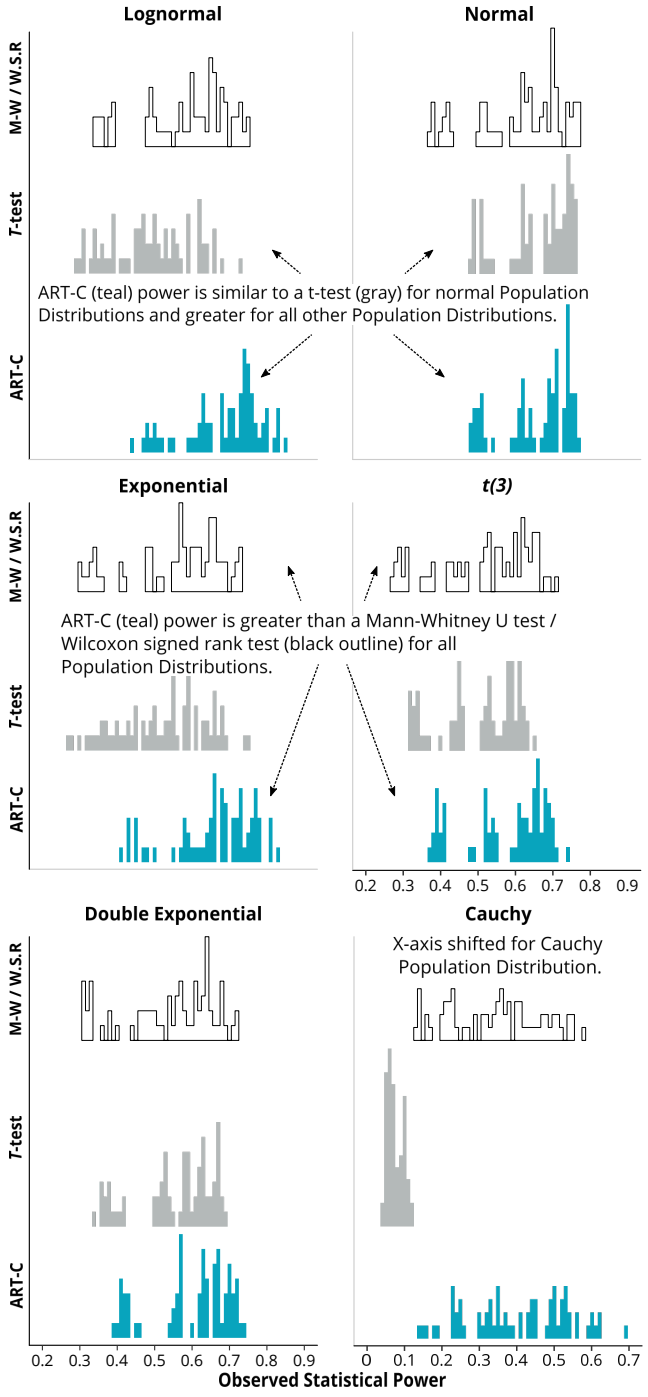

Figure 6: Mean statistical power by POPULATION DISTRIBUTION for ART-C (teal), *t*-test (gray), and Mann-Whitney *U* test / Wilcoxon signed rank-test (black outline). Higher values indicate greater power. ART-C has the greatest power overall. Each point represents observed statistical power from one design.

5.6 Comparison to Original ART

Contrasts conducted with ART-C had lower observed Type I error (M = .067, SD = .072) than contrasts conducted on data aligned-and-ranked using the original ART procedure (M = .121, SD = .174). ART-C also had higher observed power (M = .598, SD = .143) *vs.* ART (M = .511, SD = .182).

When separated by POPULATION DISTRIBUTION, ART-C had lower observed Type I error rates than ART for the log-normal, exponential, and Cauchy distributions (Table 8, Figure 3), but ART-C's observed Type I error rates were closer to the significance level (α = .05) than ART's for all POPULATION DISTRIBUTIONS, indicating that ART-C is more correct. ART-C also had higher observed power than ART for all POPULATION DISTRIBUTIONS except Cauchy (Table 7, Figure 4).

Table 8: Mean Type I error rates (and standard deviations) for ART-C and ART, grouped by POPULATION DISTRIBUTION. Values closer to α = .05 are better, indicating greater correctness.

	Normal	Log-normal	Exponential
ART-C	.049 (.008)	.054 (.015)	.051 (.007)
ART	.024 (.023)	.141 (.096)	.065 (.048)
	Cauchy	***t*(3)**	**Double Exponential**
ART-C	.140 (.154)	.049 (.006)	.049 (.007)
ART	.425 (.209)	.033 (.015)	.026 (.023)

CONTRAST SIZE also had an interesting effect on power. Observed power with ART-C was highest for single-factor contrasts, followed by three-factor contrasts, and then two-factor contrasts, but the differences were small. However, ART's power decreased as CONTRAST SIZE increased, and the differences were much larger (Table 9, Figure 7). Recall that the alignment formulas for ART and ART-C become mathematically equivalent in the single factor case.

Table 9: Mean statistical power (and standard deviations) for ART-C and ART, grouped by CONTRAST SIZE. Higher values indicate greater power.

	1-Factor Contrasts	2-Factor Contrasts	3-Factor Contrasts
ART-C	.620 (.150)	.580 (.140)	.590 (.130)
ART	.620 (.150)	.460 (.150)	.340 (.180)

Recall in our running example that we conducted multifactor contrasts on levels of factors A and B, and contrasts conducted with ART produced a Type I error and a Type II error (see Table 3, Figure 2), but contrasts conducted with a linear mixed model (LMM) on log-transformed data resulted in correct conclusions (see Table 2) that agreed with ground truth (see Table 1). Now we can say that contrasts conducted on the same data with ART-C agree with the LMM results and ground truth in finding a difference between

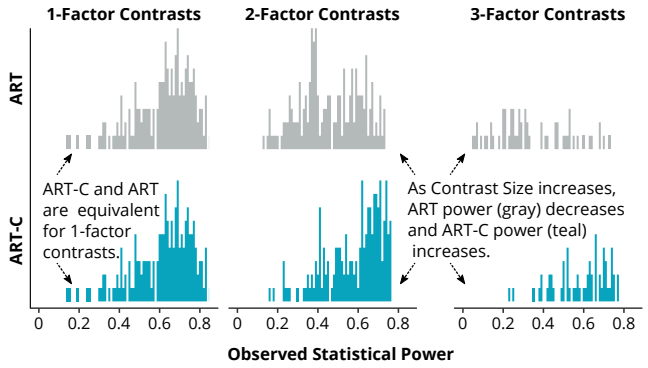

Figure 7: ART-C (teal) and ART (gray) observed statistical power by CONTRAST SIZE. Each point represents observed statistical power for one design. Higher values indicate greater power. ART-C power is greater for all CONTRAST SIZES, and increases with CONTRAST SIZE, compared to ART, which decreases. Both methods are equivalent when conducting single-factor contrasts because ART-C mathematically reduces to ART in such cases.

($A1$, $B1$) and ($A2$, $B2$), and correctly *not* finding a difference between ($A1$, $B1$) and ($A1$, $B2$) (Table 10).

Thus, taken as a whole, our results show that ART-C has appropriate Type I error rates clustered around α = .05, except for data sampled from Cauchy distributions, for which ART-C should not be used. Furthermore, ART-C has high statistical power, outperforming the t-test, Mann-Whitney U test, Wilcoxon signed-rank test, and original ART. These results show that ART-C is a correct and powerful procedure for use within the overall ART paradigm for conducting nonparametric contrast tests within or across levels of multiple factors.

Given ART-C's impressive performance as an analysis method, we sought to make it available to researchers and easy to employ. Toward that end, we extended existing open-source tools that already provide the original ART procedure for main effects and interactions to also include ART-C for multifactor contrast tests. We describe our tool extensions in the next section.

Table 10: Highlighted results of contrasts conducted using ART-C, comparing levels of A and B in our running example. In the top row, a difference was correctly not detected between ($A1$, $B1$) and ($A1$, $B2$) (p = .6758), and indeed, there is not a true difference. In the bottom row, a difference was detected correctly between ($A1$, $B1$) and ($A2$, $B2$) (p < .05), and indeed, there *is* a true difference.

contrast	estimate	SE	df	t.ratio	p.value
($A1$, $B1$) – ($A1$, $B2$)	1.5	3.7	273	0.4	0.6758
($A1$, $B1$) – ($A2$, $B2$)	-89.9	3.7	273	-24.5	<.0001

6 WINDOWS ARTOOL AND R PACKAGE "ARTOOL"

To make ART-C available to the HCI community (and beyond), as ART was made available previously [51], we extended the existing open-source tools ARTool.exe for Windows (see footnote 2) and the R package "ARTool" (see footnotes 3 and 4). This section briefly describes our new tools.

6.1 ARTool.exe Windows Application

The ARTool.exe Windows application was released as an open-source tool in 2011 [51] to facilitate aligning-and-ranking data for analysis using the ART procedure. We extended this open-source tool to include our ART-C procedure for multifactor contrasts. Users can now indicate that they want contrasts with a checkbox (Figure 8, Top), which then offers them a separate dialog box (Figure 8, Bottom) from which they can select the factors whose levels are involved in their desired contrast test. ARTool then uses our ART-C procedure to produce aligned-and-ranked output suitable for statistical analysis (e.g., with a t-test).

6.2 R Package "ARTool"

The open-source R package "ARTool" makes it easy to conduct nonparametric tests of main effects and interactions using the original ART procedure. A single function aligns-and-ranks data for each fixed effect in a formula f provided by the user. The result is an ART statistical model m that retains a copy of formula f and the data. Given m, another function in ARTool, *anova*, runs multiple ANOVAs behind the scenes, one for each fixed effect in f, and returns the results of each test. In this work, we have added a new function, *art.con*, that uses our ART-C procedure to conduct multifactor contrast tests. Given the same model m and a contrast formula f_c, the ART-C procedure is used to align-and-rank the data saved in m for the contrasts specified in f_c. It then parses the formula f saved in m, conducts the contrasts, and returns the results.

In our running example, we first conducted $A \times B$ contrasts with ART, which, of course, is incorrect given ART's propensity for Type I errors. Now, we can correctly use ART-C to perform these contrasts. Figure 9 shows how we would use ART-C to conduct contrasts correctly in R.

7 DISCUSSION

In this work, we have presented a new procedure and software tools for aligning-and-ranking data for multifactor contrast tests in the Aligned Rank Transform (ART) paradigm, now widely used in HCI (and beyond) for nonparametric analyses. Owing to the popularity of ART in HCI, the ART-C procedure promises to fill an extant gap in current statistical practice.

Our results showed that ART-C's Type I error rate is clustered around $\alpha = .05$, offering strong evidence for ART-C's correctness. Our results also showed that ART-C's statistical power is greater than the t-test, Mann-Whitney U test, Wilcoxon signed-rank test, and original ART, except for when used on data drawn from a Cauchy distribution. In fact, ART-C's greater power was largest for data drawn from log-normal and exponential distributions.

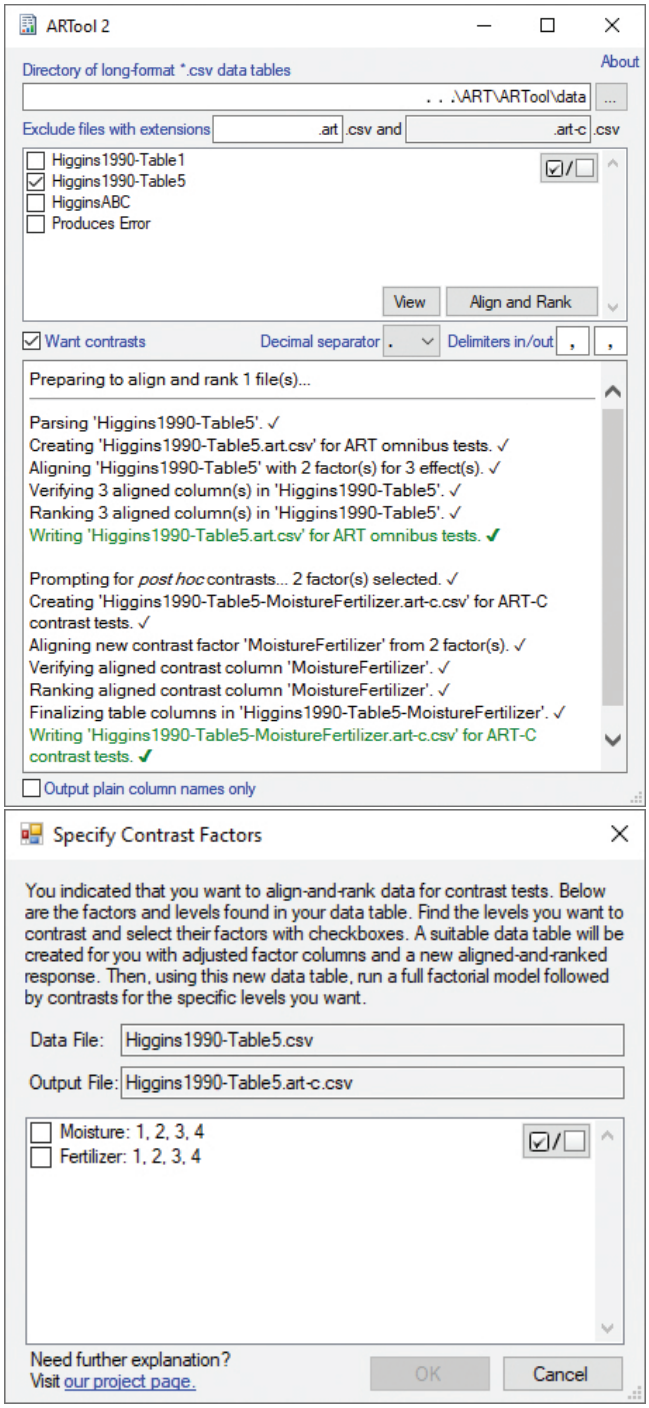

Figure 8: Top: ARTool.exe with "Want contrasts" checked. Higgins et al.'s Table 5 [16] is being aligned-and-ranked. Bottom: Our new tool for specifying multifactor contrasts. Two factors, *Moisture* and *Fertilizer*, each have levels 1-4. Selecting both factors would allow, e.g., a comparison of (Moisture 2, Fertilizer 3) *vs.* (Moisture 4, Fertilizer 1), etc.

```
# Align and rank data in df
# for all fixed effects in Y ~ A*B*C
# and random effect S.
# Create ART model m.
m = art(Y ~ A*B*C + (1|S), data=df)

# Run ANOVAs: one for each
# fixed effect in Y ~ A*B*C.
anova(m)

# Conduct contrasts on AxB interaction.
art.con(m, "A:B", adjust="holm")
```

```
 contrast      estimate   SE  df t.ratio p.value
 A1,B1 - A1,B2     1.54 3.67 273   0.419 0.6758
 A1,B1 - A2,B1  -130.54 3.67 273 -35.540 <.0001
 A1,B1 - A2,B2   -89.90 3.67 273 -24.476 <.0001
 A1,B2 - A2,B1  -132.07 3.67 273 -35.959 <.0001
 A1,B2 - A2,B2   -91.44 3.67 273 -24.895 <.0001
 A2,B1 - A2,B2    40.64 3.67 273  11.064 <.0001
```

Figure 9: Screenshot using ART-C to conduct $A \times B$ contrasts in our running example in R. The *anova* call first would produce omnibus test results for any A, B, and C main effects and interactions; if, for example, the $A \times B$ interaction were statistically significant, *art.con* could be used to conduct *post hoc* pairwise comparisons as shown here.

This finding is particularly satisfying because the log-normal and exponential distributions were included due to their frequent emergence in HCI studies.

As noted, our results showed that single-factor ART-C contrasts conducted on data drawn from a Cauchy distribution had high observed Type I error rates. This is not unique to ART-C; the Cauchy distribution is known to be "pathological" and many well-known statistics concepts do not hold for Cauchy-distributed data (e.g., the Central Limit Theorem [26]). This situation occurs because Cauchy distributions have tails that are so fat that neither their mean nor variance is well defined. In practice, this concern can arise in data with extreme outliers. Thus, we encourage users to avoid using ART-C if they have theoretical reasons to suspect the data is drawn from a Cauchy distribution or if the data has extreme outliers.

In HCI, nonparametric tests are typically used as a catch-all when parametric tests are not appropriate. The particular Cauchy result above illuminates that this practice can be problematic. In fact, ART-C is mathematically equivalent to ART in the single-factor case, and ART was thought to be appropriate for single-factor contrasts, but would also be ill-suited in the Cauchy case. A disclaimer to not use a method to analyze data drawn from a particular distribution is not useful unless researchers investigate experimental data distributions beyond checking for normality. The American Psychological Association's Taskforce on Statistical Inference encourages researchers to take a closer look at their data by saying:

> As soon as you have collected your data, before you compute *any* statistics, *look at your data*. Data screening is not data snooping. It is not an opportunity to discard data or change values to favor your hypotheses. However, if you assess hypotheses without examining your data, you risk publishing nonsense. [49] (*emphasis in original*)

Even nonparametric tests are have certain assumptions. There are many tried-and-tested visualizations for model diagnostics that can be applied to assess assumptions relevant to ART: quantile-quantile (Q-Q) plots [18], for example, allow one to check for the presence of fat tails in the distribution of residuals (i.e., *excess kurtosis*, which in extreme cases could indicate the presence of Cauchy-distributed data). Modern visualizations like worm plots [6] can make it even easier to diagnose fat tails. The point, though, is that there is no all-encompassing solution in statistical analysis: model fit and assumptions cannot be assumed and must be checked, and nonparametric approaches are no exception to this maxim.

8 LIMITATIONS AND FUTURE WORK

There are infinitely many combinations of layouts, population distributions, and condition sample sizes one could examine in a study like ours, but we could only analyze a finite amount of data and had to be selective. These decisions were carefully made, considering the needs of the HCI community and statistical norms—but they were certainly not exhaustive.

Our validation only investigated data in which all conditions' populations had the same location or all conditions' populations had different locations. Additionally, even when parameter values were varied, conditions in the same data set were always drawn from the same distribution. Data in which there are differences between some conditions and not others arises frequently in HCI, but we chose our validation process because it is commonly used in statistics [1, 5, 27, 43].

We included models with random intercepts, which represented the impact each subject had on the response. However, we did not include models with random slopes, which allow, for some types of responses, better-fitting models where subjects' responses vary differentially across another variable (e.g., time). Random intercept models generally handle the kinds of repeated measures experiments that occur often in HCI, whereas models with random slopes are often relevant for longitudinal studies. Although random slope models would certainly be valuable, fixed-effects models and models with random intercepts are used more frequently in HCI, so we chose to focus our validation on such models, leaving other models for future work.

ART-C is an alignment procedure, followed by a ranking procedure, and then a contrast test like the t-test. We chose to use the t-test in our analyses because it is the most familiar to the HCI community and therefore how we anticipate most researchers will use our new method. Still, it would be worthwhile to examine how ART-C performs when a different contrast test is used.

In addition to our extensions to the open-source Windows ARTool application and "ARTool" R package, we envision a platform-agnostic tool that does not require programming experience, and even an ART and ART-C package for other common statistics software packages (e.g., SAS, SPSS, Stata). With the algorithmic and

validation work we have done here, it should be relatively straightforward to create additional add-ons for common statistical packages.

9 CONCLUSION

The Aligned Rank Transform (ART) procedure has enabled anyone familiar with an ANOVA to conduct nonparametric statistical analyses on data arising from factorial experiments in HCI, and to correctly obtain results for not only main effects, but interactions as well [51]. Unfortunately, however, until now, multifactor *post hoc* contrast tests could not be appropriately conducted within the ART paradigm, and required using different tests altogether. This state of affairs not only creates an inconsistency in statistical analyses, but opens the door to widespread errors in statistical practice if original ART data is used in multifactor contrast tests. Specifically, we have shown that the original ART procedure results in inflated Type I error rates and poor statistical power when used to conduct multifactor contrast tests.

To remedy these problems, we have developed, presented, and validated the ART-C procedure for aligning-and-ranking data for nonparametric multifactor contrasts within the ART paradigm, giving researchers a procedure and tools to analyze data correctly from factorial experiments. We have validated our method's Type I error rate and statistical power on 72,000 synthetic data sets whose properties represent data commonly arising within HCI experiments. Our results show that ART-C does not inflate Type I error rates, and has greater statistical power than a t-test, Mann-Whitney U test, Wilcoxon signed-rank test, and the original ART. To facilitate the widespread use of ART-C, we have extended existing open-source tools, namely the Windows ARTool.exe application and the "ARTool" R package.

It is our hope that by providing ART-C to supplement analyses conducted in the ART paradigm, we will positively impact the HCI community by enabling researchers to correctly and consistently conduct nonparametric statistical analyses, especially for factorial experiments. Owing to ART's evident popularity, we believe ART-C, and our tools that provide it, can be immediately useful to many researchers in HCI and beyond.

REFERENCES

[1] Marisela Abundis. 2001. *Multiple comparison procedures in factorial designs using the aligned rank transformation.* Thesis. Texas Tech University. https://ttu-ir.tdl.org/handle/2346/22569

[2] Saleema Amershi, James Fogarty, and Daniel Weld. 2012. Regroup: interactive machine learning for on-demand group creation in social networks. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems - CHI '12.* ACM, Austin, Texas, USA, 21–30. https://doi.org/10.1145/2207676.2207680

[3] Shiri Azenkot, Kyle Rector, Richard Ladner, and Jacob Wobbrock. 2012. PassChords: secure multi-touch authentication for blind people. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '12.* ACM, Boulder, Colorado, USA, 159–166. https://doi.org/10.1145/2384916.2384945

[4] Eric Barefield and H. Mansouri. 2001. An empirical study of nonparametric multiple comparison procedures in randomized blocks. *Journal of Nonparametric Statistics* 13, 4 (2001), 591–604. https://doi.org/10.1080/10485250108832867

[5] R. Clifford Blair and James J. Higgins. 1980. A Comparison of the Power of Wilcoxon's Rank-Sum Statistic to that of Student's t Statistic under Various Nonnormal Distributions. *Journal of Educational Statistics* 5, 4 (1980), 309–335. https://doi.org/10.2307/1164905

[6] Stef van Buuren and Miranda Fredriks. 2001. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine* 20, 8 (2001), 1259–1277. https://doi.org/10.1002/sim.746

[7] D. Ciavardelli, C. Rossi, D. Barcaroli, S. Volpe, A. Consalvo, M. Zucchelli, A. De Cola, E. Scavo, R. Carollo, D. D'Agostino, F. Forlì, S. D'Aguanno, M. Todaro, G. Stassi, C. Di Ilio, V. De Laurenzi, and A. Urbani. 2014. Breast cancer stem cells rely on fermentative glycolysis and are sensitive to 2-deoxyglucose treatment. *Cell Death & Disease* 5, 7 (2014), e1336–e1336. https://doi.org/10.1038/cddis.2014.285

[8] W. J. Conover and Ronald L. Iman. 1981. Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics. *The American Statistician* 35, 3 (1981), 124–129. http://www.jstor.org/stable/2683975

[9] R. F. Fawcett and K. C. Salter. 1984. A Monte Carlo study of the F test and three tests based on ranks of treatment effects in randomized block designs. *Communications in Statistics: Simulation and Computation* 13, 2 (1984), 213–225. https://doi.org/10.1080/03610918408812368

[10] Kara L. Feilich and George V. Lauder. 2015. Passive mechanical models of fish caudal fins: effects of shape and stiffness on self-propulsion. *Bioinspiration & Biomimetics* 10, 3 (2015), 036002. https://doi.org/10.1088/1748-3190/10/3/036002

[11] Brigitte N Frederick. 1999. Fixed-, Random-, and Mixed-Effects ANOVA Models: A User-Friendly Guide for Increasing the Generalizability of ANOVA Results. (1999). https://eric.ed.gov/?id=ED426098

[12] Milton Friedman. 1937. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Amer. Statist. Assoc.* 32, 200 (1937), 675–701. https://doi.org/10.2307/2279372

[13] António Gaspar, André P Lourenço, Miguel Álvares Pereira, Pedro Azevedo, Roberto Roncon-Albuquerque, Jorge Marques, and Adelino F Leite-Moreira. 2018. Randomized controlled trial of remote ischaemic conditioning in ST-elevation myocardial infarction as adjuvant to primary angioplasty (RIC-STEMI). *Basic research in cardiology* 113, 3 (2018), 1–10. https://doi.org/10.1007/s00395-018-0672-3

[14] Jan Gugenheimer, Evgeny Stemasov, Julian Frommel, and Enrico Rukzio. 2017. ShareVR: Enabling Co-Located Experiences for Virtual Reality between HMD and Non-HMD Users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17.* ACM, Denver, Colorado, USA, 4021–4033. https://doi.org/10.1145/3025453.3025683

[15] Nur Al-huda Hamdan, Adrian Wagner, Simon Voelker, Jürgen Steimle, and Jan Borchers. 2019. Springlets: Expressive, Flexible and Silent On-Skin Tactile Interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19.* ACM, Glasgow, Scotland, UK, 1–14. https://doi.org/10.1145/3290605.3300718

[16] James J. Higgins, R. Clifford Blair, and Suleiman Tashtoush. 1990. The Aligned Rank Transform Procedure. *Conference on Applied Statistics in Agriculture* (1990), 185–195. https://doi.org/10.4148/2475-7772.1443

[17] James J. Higgins and Suleiman Tashtoush. 1994. An aligned rank transform test for interaction. *Nonlinear World* 1, 2 (1994), 201–211.

[18] David C Hoaglin. 2011. Using quantiles to study shape. *Exploring data tables, trends, and shapes* 101 (2011), 417–460.

[19] J. L. Hodges and E. L. Lehmann. 1962. Rank Methods for Combination of Independent Experiments in Analysis of Variance. *The Annals of Mathematical Statistics* 33, 2 (1962), 482–497. https://doi.org/10.1214/aoms/1177704575

[20] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* 6, 2 (1979), 65–70. https://www.jstor.org/stable/4615733

[21] Eunice Jun, Maureen Daum, Jared Roesch, Sarah Chasins, Emery Berger, Rene Just, and Katharina Reinecke. 2019. Tea: A High-level Language and Runtime System for Automating Statistical Analysis. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology - UIST '19.* ACM, New Orleans, LA, USA, 591–603. https://doi.org/10.1145/3332165.3347940

[22] Shaun K. Kane, Meredith Ringel Morris, Annuska Z. Perkins, Daniel Wigdor, Richard E. Ladner, and Jacob O. Wobbrock. 2011. Access overlays: improving non-visual access to large touch screens for blind users. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology - UIST '11.* ACM, Santa Barbara, California, USA, 273–282. https://doi.org/10.1145/2047196.2047232

[23] Maurits Clemens Kaptein, Clifford Nass, and Panos Markopoulos. 2010. Powerful and consistent analysis of likert-type ratingscales. In *Proceedings of the 2010 CHI Conference on Human Factors in Computing Systems - CHI '10.* ACM, Atlanta, Georgia, USA, 2391–2394. https://doi.org/10.1145/1753326.1753686

[24] Matthew Kay. 2020. Contrast tests with ART. https://cran.r-project.org/web/packages/ARTool/vignettes/art-contrasts.html

[25] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. 2016. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16.* ACM, San Jose, California, USA, 4521–4532. https://doi.org/10.1145/2858036.2858465

[26] Kalimuthu Krishnamoorthy. 2006. *Handbook of statistical distributions with applications.* CRC Press. https://doi.org/10.1201/9781420011371

[27] Dong Li. 2017. *Robustness And Power Of The Student T, Welch-Aspin, Yuen, Tukey Quick, And Haga Tests.* Thesis. Wayne State University. https://digitalcommons.wayne.edu/oa_dissertations/1722

[28] H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1 (1947), 50–60. https://www.jstor.org/stable/2236101

[29] H. Mansouri. 1999. Multifactor analysis of variance based on the aligned rank transform technique. *Computational Statistics & Data Analysis* 29, 2 (1999), 177–189. https://doi.org/10.1016/S0167-9473(98)00077-2

[30] H. Mansouri. 2015. Simultaneous inference based on rank statistics in linear models. *Journal of Statistical Computation and Simulation* 85, 4 (2015), 660–674. https://doi.org/10.1080/00949655.2013.836292

[31] H. Mansouri, R. L. Paige, and J. G. Surles. 2004. Aligned Rank Transform Techniques for Analysis of Variance and Multiple Comparisons. *Communications in Statistics - Theory and Methods* 33, 9 (2004), 2217–2232. https://doi.org/10.1081/STA-200026599

[32] H. R. Neave and C. W. J. Granger. 1968. A Monte Carlo Study Comparing Various Two-Sample Tests for Differences in Mean. *Technometrics* 10, 3 (1968), 509–522. https://doi.org/10.2307/1267105

[33] Daryle Alan Olson. 2013. *The Efficacy Of Select Nonparametric And Distribution-Free Research Methods: Examining The Case Of Concomitant Heteroscedasticity And Effect Of Treatment.* Thesis. Wayne State University. https://digitalcommons.wayne.edu/oa_dissertations/684

[34] Kathleen Peterson. 2002. Six Modifications Of The Aligned Rank Transform Test For Interaction. *Journal of Modern Applied Statistical Methods* 1, 1 (2002), 100–109. https://doi.org/10.22237/jmasm/1020255240

[35] M. L. Puri and P. K. Sen. 1969. A class of rank order tests for a general linear model. *Annals of Mathematical Statistics* 40, 4 (1969), 1325–1343. https://www.jstor.org/stable/2239597

[36] Marta Revilla-León, Peng Jiang, Mehrad Sadeghpour, Wenceslao Piedra-Cascón, Amirali Zandinejad, Mutlu Özcan, and Vinayak R. Krishnamurthy. 2019. Intraoral digital scans—Part 1: Influence of ambient scanning light conditions on the accuracy (trueness and precision) of different intraoral scanners. *The Journal of Prosthetic Dentistry* 124, 3 (2019), 372–378. https://doi.org/10.1016/j.prosdent.2019.06.003

[37] Scott J. Richter. 1999. Nearly exact tests in factorial experiments using the aligned rank transform. *Journal of Applied Statistics* 26, 2 (1999), 203–217. https://doi.org/10.1080/02664769922548

[38] Judy Robertson and Maurits Kaptein. 2016. *Modern Statistical Methods for HCI.* Springer International Publishing. https://doi.org/10.1007/978-3-319-26633-6

[39] Joan Sol Roo and Martin Hachet. 2017. One Reality: Augmenting How the Physical World is Experienced by combining Multiple Mixed Reality Modalities. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology - UIST '17.* ACM, Québec City, Québec, Canada, 787–795. https://doi.org/10.1145/3126594.3126638

[40] K.C. Salter and R.F. Fawcett. 1985. A robust and powerful rank test of treatment effects in balanced incomplete block designs. *Communications in Statistics: Simulation and Computation* 14, 4 (1985), 807–828. https://doi.org/10.1080/03610918508812475

[41] K. C. Salter and R. F. Fawcett. 1993. The ART test of interaction: a robust and powerful rank test of interaction in factorial models. *Communications in Statistics: Simulation and Computation* 22, 1 (1993), 137–153. https://doi.org/10.1080/03610919308813085

[42] Shlomo S. Sawilowsky. 1990. Nonparametric tests of interaction in experimental design. *Review of Educational Research* 60, 1 (1990), 91–126. https://doi.org/10.2307/1170226

[43] Shlomo S. Sawilowsky and R. Clifford Blair. 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 (1992), 352–360. https://doi.org/10.1037/0033-2909.111.2.352

[44] Student. 1908. The Probable Error of a Mean. *Biometrika* 6, 1 (1908), 1–25. https://doi.org/10.2307/2331554

[45] John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics* 5, 2 (1949), 99–114. https://doi.org/10.2307/3001913

[46] James H. Ware. 1985. Linear Models for the Analysis of Longitudinal Studies. *The American Statistician* 39, 2 (1985), 95–101. https://doi.org/10.2307/2682803

[47] Eric W Weisstein. 2004. Bonferroni correction. *MathWorld–A Wolfram Web Resource* (2004). https://mathworld.wolfram.com/BonferroniCorrection.html

[48] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. https://doi.org/10.2307/3001968

[49] Leland Wilkinson, The Task Force on Statistical Inference, American Psychological Association, and Science Directorate. 1999. Statistical methods in psychology journals: Guidelines and explanations. *American psychologist* 54, 8 (1999), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

[50] Jacob O. Wobbrock. 2017. The relevance of nonparametric and semi-parametric statistics to HCI. In *Workshop on "Moving Transparent Statistics Forward." 2017 CHI Conference on Human Factors in Computing Systems - CHI '17.* ACM, Denver, Colorado, USA, Paper No. 2. http://faculty.washington.edu/wobbrock/pubs/chi-17.06.pdf

[51] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems - CHI '11.* ACM, Vancouver, British Columbia, Canada, 143–146. https://doi.org/10.1145/1978942.1978963

[52] Jacob O. Wobbrock and M. Kay. 2016. Nonparametric statistics in human-computer interaction. In *Modern Statistical Methods for HCI,* M. C. Robertson, J. & Kaptein (Ed.). Springer, Switzerland, Chapter 7, 135–170. https://doi.org/10.1007/978-3-319-26633-6_7