USING DYADIC PROPORTION TESTS WITH INDIVIDUAL PAIRWISE COMPARISONS TO OVERCOME POWER LIMITATIONS WITH SMALL-SAMPLE EXPERIMENTS

RAYMOND D. COLLINGS
LESLIE G. EATON
JOHN T. FOLEY
STATE UNIVERSITY OF NEW YORK COLLEGE AT CORTLAND

ANTHONY J. NELSON
PENNSYLVANIA STATE UNIVERSITY

The dyadic proportions test (DPT) is an analytic approach intended for small sample research (e.g., clinical and exploratory studies). The procedure involves using proportion tests to compare outcomes from dyads (one participant in each condition) associated with the actual experimental assignment with outcomes from all possible dyadic assignments. Monte Carlo analyses in Study 1 revealed that sampling distributions generated by DPTs were normally-distributed, as expected with traditional approaches. In Study 2, data from eight small-sample experiments ($n = 8$) were analyzed with DPTs and t tests. DPTs provided more power and resulted in more significant effects ($p < .05$) than did t tests. In Study 3, data from an intervention study conducted with a small clinical sample ($n = 12$) were analyzed with DPTs and t tests. The experimental effect was significant with the DPT, but not the t test. The DPT is an appropriate approach when large samples are not feasible.

Key words: Methodology quantitative; Dyadic proportion tests; Pairwise comparisons; Small samples; Statistical power.

Correspondence concerning this article should be addressed to Raymond D. Collings, Department of Psychology, State University of New York College at Cortland, P.O. Box 2000, Cortland, NY 13045, USA. Email: Raymond.Collings@cortland.edu

Over the last several decades, statistical power has become an increasingly large concern for researchers. Technically, power is defined as the probability of correctly rejecting a false null hypothesis (1–β), that is, the probability of avoiding a false negative conclusion (Type II error, β). However, power is influenced by several factors, including the researcher's tolerance for a false positive result (Type I error, α); the magnitude of the effect (e.g., d, r^2, *etc.)*; and the sample size used in the statistical test. Since α is dictated by convention ($p < .05$), and since the effect size is a product of the experimental manipulation, the most expedient and often used method for maximizing power is to increase sample size. Researchers are typically expected to conduct power analysis during the design phase to estimate the appropriate sample size, using estimated effect sizes from prior published research or pilot studies, and to recruit sufficiently large samples to avoid Type II error (Cohen, 1977; 1992; Maxwell, 2004).

There are times, however, when recruiting large samples is not practical, possible, or ethically desirable. Clinical researchers often face a formidable challenge recruiting large samples for intervention studies involving low-base rate disorders. For example, recruiting a large sample of individuals diagnosed

TPM Vol. 26, No. 2, June 2019 – 221-236 – doi:10.4473/TPM26.2.4 – © 2019 Cises

with Capgras Syndrome (less than 2 in 1,000) or amyotrophic lateral sclerosis (ALS; less than 3.9 in 1,000) would be challenging (Mehta et al., 2014; Tamam, Karatas, Zeren, & Ozpoyraz, 2003). Recruitment can also be difficult even with more prevalent disorders (e.g., ADHD, learning disabilities, etc.), when the study involves smaller subgroups (e.g., individuals with comorbid diagnoses, diverse populations, rural settings, etc.). Large multisite collaborations allow teams of researchers to recruit large clinical samples. However, when exploratory studies involve innovative or high-risk methodologies that entail large expenditures of time, labor, and financial resources, researchers may be limited to small samples that provide inadequate power for traditional statistical approaches. Unfortunately, this has contributed to *the file drawer problem* (Bradley & Gupta, 1997; Rosenthal, 1979). This term refers to the overrepresentation of significant results (i.e., $p < .05$) and the underrepresentation of nonsignificant results (i.e., $p \geq .05$). Although researchers should strive to recruit sufficiently large samples to meet the power demands of the design, Etz and Arroyo (2015) argued that science also would benefit from the development of more powerful statistical approaches when large samples are *not* obtainable. This paper outlines one viable strategy for such situations.

The Logic of Traditional Data Analysis

It is helpful to consider the computational logic of traditional analytic approaches in the context of a simple experimental between groups design. Participants are randomly assigned to one of two experimental conditions (independent variable), and subsequently measured on some outcome (dependent variable). Under the null hypothesis, the measures on the dependent variable for the two groups typically are expected to be roughly equivalent. When an independent groups *t* test is used to test the null hypothesis, the difference between the two group means are compared with the estimated standard error, which is a function of both sample variability (s^2_{pooled}) and sample size (see Equation 1). The resulting *t* value is used to determine whether or not the null hypothesis should be rejected ($p < .05$).

$$t = \frac{M_1 - M_2}{\sqrt{\frac{s^2{}_{\text{pooled}}}{N}}} \tag{1}$$

Ironically, *t* tests were originally proposed by Gossett (1908 as cited in Lehmann, 1999) as an approach for working with small samples in agricultural applications. However, *t* tests often do not provide sufficient power to detect the relatively small effects frequently observed in certain areas of educational and social science research (De Winter, 2013). This problem then is compounded when researchers are limited to small samples.

Resampling Approaches

Several *resampling* strategies have been proposed as a solution for problems associated with small sample research. For a more comprehensive review of these, we refer the reader to Hesterberg, Moore, Monaghan, Clipson, and Epstein (2005). Fisher's (1935) *permutation test* provides one such alternative (Good, 2005; LaFleur & Greevy, 2009). Permutation tests, sometimes referred to as *randomization tests*, involve resampling *without* replacement from the original sample to create a permutation distribution of outcomes from all possible experimental assignments (Hesterberg et al. 2005; LaFleur & Greevy, 2009).

The outcome from the actual assignment is then compared with this permutation distribution, to determine if it is significantly different from what one would expect under the null.

Many of the assumptions inherent in traditional parametric tests do not apply to permutation tests, most notably the assumption regarding the normality of the population (LaFleur & Greevy, 2009). These tests also are less biased by outliers than are traditional parametric tests, and they frequently provide more statistical power than do other resampling approaches (i.e., bootstrapping; Good, 2005). However, the somewhat labor-intensive computations has undoubtedly slowed the widespread use of permutations tests by researchers. For example, suppose a researcher conducts a between-groups design with eight participants randomly assigned to one of two experimental conditions. In this case, 70 unique permutations are possible, including the original assignment used by the experimenter (see Equation 2). Additionally, permutation tests frequently provide no more statistical power than do traditional parametric *t* tests (Berger, 2000). In the following section, we propose an alternative approach derived from the basic idea of permutation tests that offers increased power with small sample between groups research designs.

$$\text{number of unique permutations} = \frac{N!}{n_1!n_2!} = \frac{8!}{4!4!} = 70 \tag{2}$$

Dyadic Proportions Test

Rather than creating a distribution of permutations associated with *all* scores being assigned to the experimental conditions, our approach involves creating a distribution of outcomes associated with pairs of scores, in which each score comes from one of the two experimental conditions. For example, suppose a researcher had recruited four participants for a between-groups experiment, with Subjects 1 and 2 randomly assigned to the experimental condition and Subjects 3 and 4 randomly assigned to the control condition. Table 1 depicts the 12 possible dyads in which each score is paired with each of the other scores (example scores in parentheses). Only four pairs (marked with an asterisk) involve dyads associated with the original random assignment. The remaining eight dyads involve all other pairings that *might have* occurred under a different random assignment. Next, each dyad is marked as either a success (i.e., subject in experimental condition greater than subject in control condition) or nonsuccess (i.e., subject in experimental condition not greater than subject in control condition). This allowed us to calculate the proportion of the 12 possible dyads that were identified as a success (p_{total}). Finally, the proportion of successes among the four dyads associated with the original random assignment would be compared with the proportion of successes among all 12 possible dyads (p_{obt}). A one-sample proportions (*Z*) test (see Equation 3) is used to test the null hypothesis (i.e., the two proportions are roughly equivalent). We refer to this overall approach as the dyadic proportions test (DPT).

$$z = \frac{p_{\text{obt}} - p_{\text{total}}}{se} \tag{3}$$

$$se = \sqrt{p_{\text{total}} * (1 - p_{\text{total}}) * (\frac{1}{n})}$$

p_{obt} = proportion of success among dyads associated with original random assignment;
p_{total} = proportion of success among all possible dyadic pairings;
n = number of dyads associated with original random assignment.

TABLE 1
Scheme for using all possible dyads with one member assigned to experimental condition and one member assigned to control condition for pairwise contrasts ($N = 4$)

	Experimental		Control	
Dyad 1	S1 score	(8)	S2 score	(5)
Dyad 2*	S1 score	(8)	S3 score	(4)
Dyad 3*	S1 score	(8)	S4 score	(3)
Dyad 4	S2 score	(5)	S1 score	(8)
Dyad 5*	S2 score	(5)	S3 score	(4)
Dyad 6*	S2 score	(5)	S4 score	(3)
Dyad 7	S3 score	(4)	S1 score	(8)
Dyad 8	S3 score	(4)	S2 score	(5)
Dyad 9	S3 score	(4)	S4 score	(3)
Dyad 10	S4 score	(3)	S1 score	(8)
Dyad 11	S4 score	(3)	S2 score	(5)
Dyad 12	S4 score	(3)	S3 score	(4)

Note. S = subject. Example scores in parentheses. Dyads marked with an * involve pairings associated with original random assignment (Subjects 1 and 2 in experimental condition and Subjects 3 and 4 in control condition).

Using the example scores depicted in Table 1, six of the 12 possible dyads ($p_{\text{total}} = .50$) resulted in successes (experimental score greater than control score). All of the four dyads that resulted from the original assignment were successes ($p_{\text{obt}} = 1.0$). When we calculated a one sample proportions test, the difference between p_{obt} and p_{total} was found to be significant (Equation 4).

$$z = \frac{p_{\text{obt}} - p_{\text{total}}}{\sqrt{p_{\text{total}} * (1 - p_{\text{total}}) * (\frac{1}{n})}} = \frac{1.0 - .50}{\sqrt{.5 * (1 - .5) * (\frac{1}{4})}} = 2.00, p = .023 \tag{4}$$

Our proposed strategy differs from traditional tests on two important points. First, the DPT can be viewed as a series of internal replications of the same experiment, using single scores from each condition. This contrasts sharply with the traditional approach of relying on a single comparison between the two group means, with no regard for the number of times that the research hypothesis correctly predicted outcomes among individuals assigned to each condition. Second, the dyadic approach is similar to Fisher's permutations test, in that it involves resampling. Consequently, rules regarding independence associated with parametric tests (i.e., *t* tests) do not apply. This again contrasts with traditional *t* tests that demand that each score is used only once.

THREE SMALL SAMPLE STUDIES COMPARING THE DPT WITH TRADITIONAL *T*-TESTS

The aim of Study 1 was to use Monte Carlo analyses to compare the distribution of test results produced by our DPT approach with the distribution of results produced by traditional independent groups *t* tests. We anticipated that under the null hypothesis, the two distributions would be roughly equivalent. In Study 2, we applied the two approaches to empirical data from eight independent small-sample replications (N = 8) of a widely-used memory task. This approach allowed us to compare the effectiveness of the two statistical approaches in detecting significant effects (p < .05) with a phenomenon that has been widely-replicated in larger sample studies. We anticipated that the DPTs would result in greater statistical power than would traditional tests. Finally, Study 3 tested the DPT approach in a small-sample clinical application. For this, we analyzed data from a small quasiclinical study involving an effect of mindfulness training on working memory. Our aim was to determine if the DPT would detect an effect that had been previously observed in a larger nonclinical study (Frank, Adams, Leverton, Delmuro, & Collings, 2018). We anticipated that the DPT would be more likely to detect a significant effect than would an independent groups *t* test.

STUDY 1: MONTE CARLO STUDY OF EFFECTS ASSOCIATED WITH DYADIC PROPORTIONS

In Study 1, we used a Monte Carlo method to create sampling distributions of effects that would be associated with DPTs and traditional *t* tests. Under probability theory, we would expect both distributions to be roughly normal. However, it should be noted that this expectation is based on the *rule of proportions* for one-sample proportions tests and the *central limit theorem* for traditional *t* tests, both of which impose minimum sample sizes (N = 10 and N = 30, respectively). We purposely violated these sample size restrictions, in order to test the utility of the DPT with smaller samples (N = 8).

Method

We conducted our Monte Carlo analyses in Microsoft Excel, using 100,000 small samples consisting of eight random numbers. The DPT was used to test the null hypothesis for each sample. For example, suppose a sample of eight random numbers was created, with four assigned to the experimental condition (56, 38, 79, 75) and four assigned to the control condition (44, 29, 23, 98). A total of 56 dyads, with one score assigned to each condition would be possible (see Table 2). Of these, 28 would be considered successes (p_{total} = .50). Of the 16 dyads created from the original assignment, 11 were successes (p_{obs} = .69). A one-sample proportions tests would then be calculated to compare p_{obs} with p_{total} (Equation 5). As expected, p_{obs} was not significantly different from p_{total} when a DPT was conducted.

$$z = \frac{.69 - .50}{\sqrt{.50 * (1 - .50) * \frac{1}{16}}} = 1.5, p = .067 \qquad (5)$$

For comparison purposes, we also conducted independent groups *t* tests for each of the 100,000 samples. This allowed us to create and compare the sampling distributions for both tests (Z and t).

Results and Discussion

The distribution of 100,000 Z scores associated with our DPT approach depicted in Figure 1A appeared to be somewhat platykurtic. However, when statistics were calculated for skewness (–.004) and excess kurtosis (–.468), both were within the range of normality. The distribution of test scores associated

TABLE 2
Scores from all possible dyads with one score assigned to each condition in example data

Experimental	Control		Experimental	Control		Experimental	Control		Experimental	Control	
44	29	Success	29	44		23	44		98	44	Success
44	23	Success	29	29	Success	23	29		98	29	Success
44	98		29	23		23	98		98	23	Success
44	56		29	98		23	56		98	56	Success
44	38	Success	29	56		23	38		98	38	Success
44	79		29	38		23	79		98	79	Success
44	75		29	79		23	75		98	75	Success
56*	44*	Success	38*	44*		79*	44*	Success	75*	44*	Success
56*	29*	Success	38*	29*	Success	79*	29*	Success	75*	29*	Success
56*	23*	Success	38*	23*	Success	79*	23*	Success	75*	23*	Success
56*	98*		38*	98*		79*	98*		75*	98*	
56	38	Success	38	56		79	56	Success	75	56	Success
56	79		38	79		79	38	Success	75	38	Success
56	75		38	75		79	75	Success	75	79	

Note. * Corresponds with original assignment: experimental condition (56, 38, 79, 75) and control condition (44, 29, 23, 98).

with independent groups *t* tests with the same 100,000 samples, depicted in Figure 1B, was roughly symmetrical, as reflected by the skewness (.081). However, as expected with *t* distributions for only 6 degrees of freedom, the resulting sampling distribution was rather leptokurtic (excess kurtosis = 6.136).

These findings provide mathematical support for the notion that DPTs come closer to producing a normal distribution of test results under the null hypothesis than do traditional *t* tests when conducted with extremely small samples. It is important to note, however, these results may tell us little about empirical observations in messy "real world" conditions. In the following section, we compare the two approaches using data collected from a series of small-sample experiments.

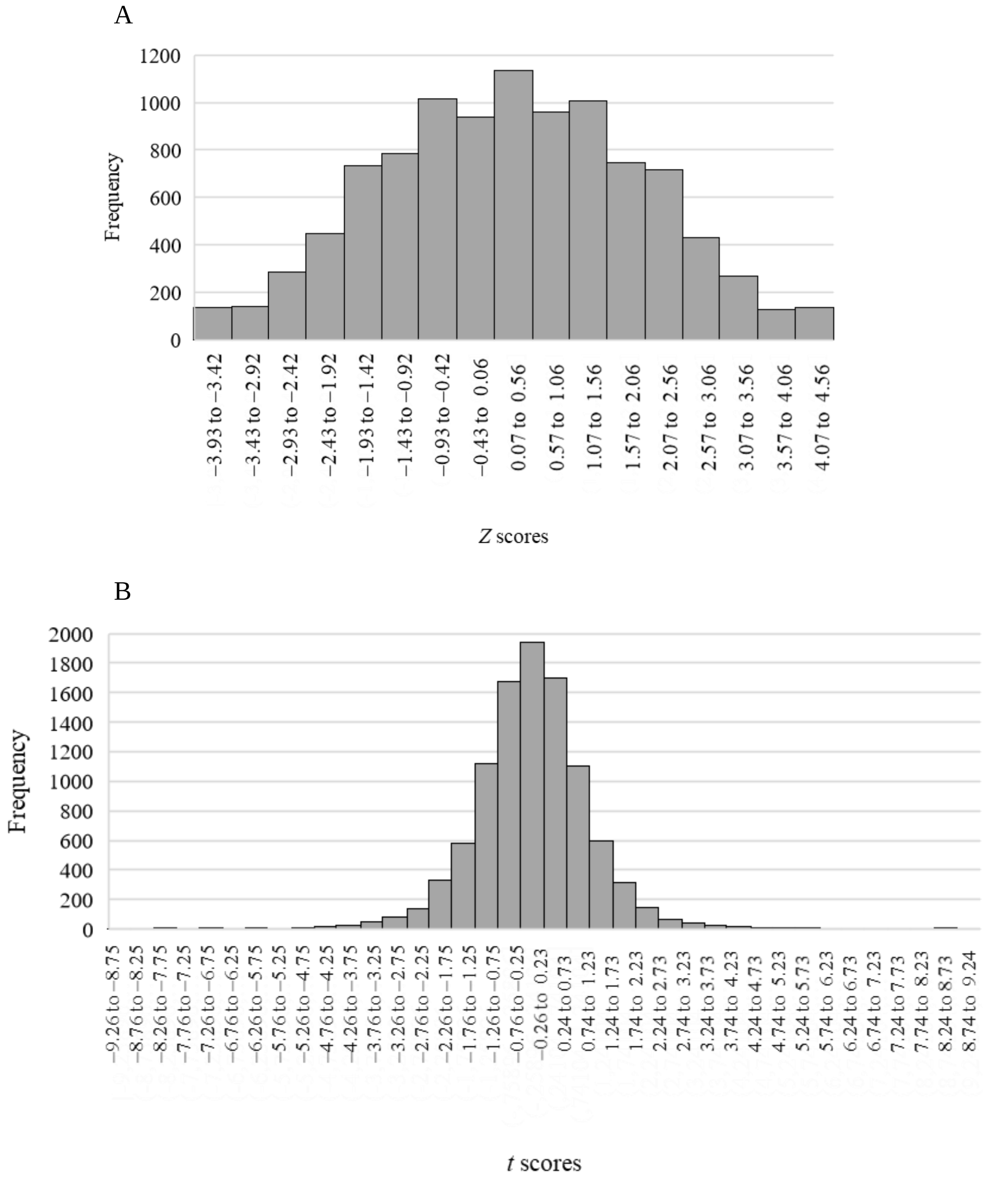

FIGURE 1
Sampling distributions of effects for 100,000 samples.
Panel A: Distribution of *Z* scores associated with one-sample proportions tests (DPT).
Panel B: Distribution of *t* scores associated with independent groups *t* tests.

STUDY 2: COMPARISON BETWEEN DPTS AND TRADITIONAL ANALYSES WITH DATA FROM EIGHT REPLICATIONS OF A CLASSIC MEMORY EXPERIMENT

In the following study, we analyzed data from eight small-sample replications of Craik and Lockart's *levels of processing* experimental paradigm (1972), using the DPTs and traditional *t* tests. We selected this experimental task because it is known to produce a reliable, theoretically-sound outcome. In this experimental paradigm, participants are shown a list of words and asked to identify either a superficial feature of the word (shallow processing) or a more complex semantic quality of the word (deep processing). Later, the participants are asked to recall as many of the words as possible. Participants assigned to the deep processing condition generally recall more words than do participants in the shallow processing condition. However, with extremely small samples ($N = 8$), traditional *t* tests are unlikely to provide sufficient power to detect the effect. This allowed us to examine the utility of using DPTs to detect an expected effect. We anticipated that the increased power of the DPT approach would result in fewer Type II errors than would conventional *t* tests.

Method

Participants

Students enrolled in an introductory honors statistics class served as experimenters, each recruiting small samples ($N = 8$) of friends, roommates, and in some cases, family members to participate in a replication of the aforementioned level of processing experiment. In total, eight samples, each with eight participants were recruited to perform independent experiments. This allowed us to compare the results of traditional tests and DPTs across several heterogeneous small-sample studies.

Procedure and Design

We used a variation of Craik and Lockhart's experimental paradigm (1972, as cited in Goldstein, 2004; see Figure 2). Participants were shown a list of eight words, one at a time, with a two-second lag between words.[1] The participants were instructed either to count the vowels (shallow processing condition) or to imagine how useful the item would be if stranded on a desert island (deep processing condition) when they saw each word. After the presentation of the list, the experimenter asked the participant to count backwards from 100 by three's for 15 seconds. Finally, the participants were instructed to use free-recall to write down as many of the words as they could remember, in any order. Each replication experiment represented a simple between-group design, with deep versus shallow as the two levels of the independent variable. The number of words correctly remembered served as the dependent variable.

Results and Discussion

After completing data collection, the experimenter counted and recorded the number of stimuli correctly recalled by each participant. This allowed for the calculation of a mean number of correct responses for the two conditions (shallow and deep processing) for each experimenter's replication study.

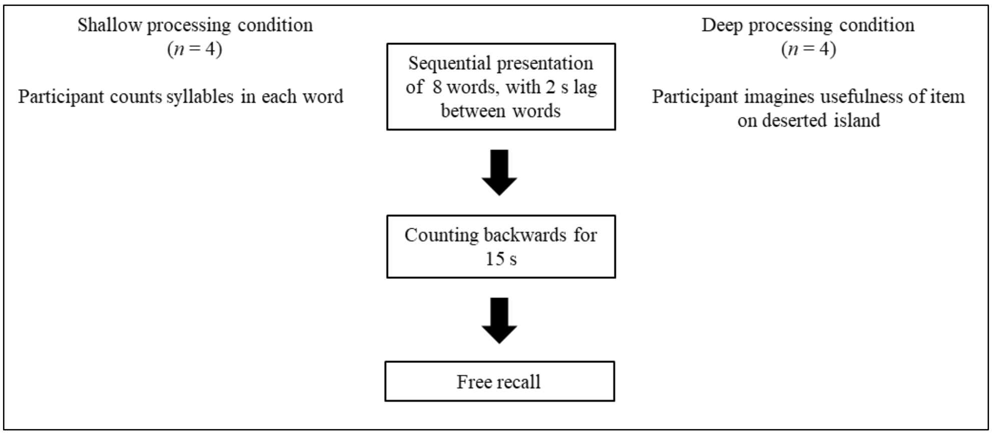
Shallow processing condition
($n = 4$)
Participant counts syllables in each word
Sequential presentation of 8 words, with 2 s lag between words
Deep processing condition
($n = 4$)
Participant imagines usefulness of item on deserted island
Counting backwards for 15 s
Free recall

FIGURE 2
Levels of processing experimental paradigm.

As illustrated in Figure 3, the mean scores for the deep processing condition appeared to be greater than the mean scores from the shallow processing condition in seven of the eight experiments. However, when a series of independent groups *t* tests were conducted, only four significant effects were detected at conventional levels ($p < .05$; see Table 3).

We then conducted a meta-analysis to test the significance of the effect across all eight experiments, to determine if an effect was present when the sample size was increased through aggregation. To accomplish this, we used the *method of testing the mean p* (see Rosenthal, 1991). The results of this analysis revealed a significant and large effect across the eight experiments, $Z = 3.67$, $p < .001$, $r_{mean} = .60$. This finding is in line with both theory and prior research (see Craik & Lockhart, 1972).[2] However, these results demonstrate that traditional *t* tests provided insufficient power to detect the effects in half of the small sample studies, even when the effect sizes were relatively large.

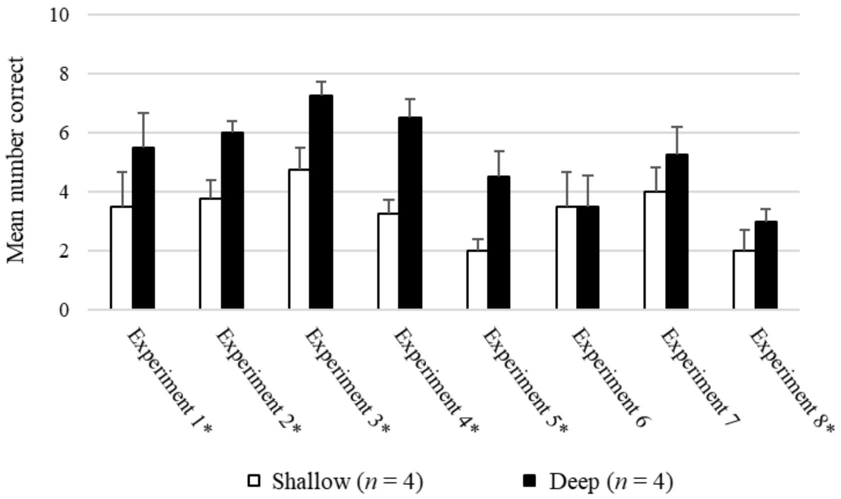

FIGURE 3
Mean number of words correctly recalled for each experimental condition in eight experiments.
* indicates experiment resulted in *success*, in which mean number correct for deep condition was greater than the mean number correct for the shallow condition. Error bars reflect standard error.

TABLE 3
Results from independent groups t tests for each of the eight experiments

	Independent groups t-tests		
	t	p	r_{effect}
Experiment 1	1.19	.140	.44
Experiment 2	3.00	.012*	.78
Experiment 3	2.81	.015*	.75
Experiment 4	4.04	.003*	.86
Experiment 5	2.61	.020*	.73
Experiment 6	0.00	.500	.00
Experiment 7	1.00	.178	.38
Experiment 8	1.23	.133	.45

* $p < .05$, one-tailed ($df = 6$).

We next conducted DPTs for each of the eight replications. An example of this procedure is illustrated in Table 4. Each individual score is paired with the remaining scores, creating a total of 56 unique dyads. The 16 dyads in which one score from the deep processing condition was paired with one score from the shallow processing condition reflect the original random assignment. The remaining 40 dyads reflect all other possible dyads that might have occurred with different random assignments. When the score labelled "deep" was greater than the score labelled "shallow" the dyad was considered a success, regardless of whether the dyad corresponded with the actual random assignment. In the example depicted in Table 4, the individuals originally assigned to the deep processing condition recalled three, four, seven, and eight words respectively, and the participants assigned to the shallow processing condition correctly recalled one, two, five, and six words respectively. Consequently, 12 of the 16 dyads associated with the random assignment resulted in "successes" (p_{obt} = .75). This contrasts with 28 successes out of the 56 possible dyads (p_{total} = .50).

The final step in our DPT analyses involved conducting one-sample proportion tests to test the difference between p_{obt}, and p_{total} for each of the eight replications. These tests revealed significant effects for six of the eight replications ($p < .05$; see Table 5). One additional replication resulted in a nonsignificant trend ($p < .10$). Because the DPT is a nonparametric test, we wanted to compare the results of our approach to three widely used nonparametric tests. The Mann-Whitney U test is a nonparametric equivalent to the independent groups t test. When we conducted this test with each of our eight replications, only two were found to be statistically significant (see Table 6). When we conducted independent samples Wald-Wolfowitz runs tests, and Kolmogorov-Smirnov goodness-of-fit tests with the eight replications, both performed even more poorly. Only one replication experiment resulting in significant effects with either approach.

These findings suggest that DPTs provided substantially more power than did traditional t tests, and were likely to have reduced Type II errors resulting from the small sample sizes. The same was true to an even greater extent with other nonparametric tests. It should be noted that our demonstration involved replications of a simple experimental paradigm with a normative sample, so collecting larger samples would not be a problem. In the next section we turn to a situation more typical of clinical studies, in which collecting larger samples would not have been feasible.

TABLE 4
All pairwise contrasts for one replication of the levels of processing experiment

Original random assignments S_{1-4} assigned to shallow condition; S_{5-8} assigned to deep condition

Shallow	Deep	Shallow	Deep	Shallow	Deep	Shallow	Deep
S_1 (1)	S_5 (3)*	S_2 (2)	S_5 (3)*	S_3 (5)	S_5 (3)	S_4 (6)	S_5 (3)
S_1 (1)	S_6 (4)*	S_2 (2)	S_6 (4)*	S_3 (5)	S_6 (4)	S_4 (6)	S_6 (4)
S_1 (1)	S_7 (7)*	S_2 (2)	S_7 (7)*	S_3 (5)	S_7 (7)*	S_4 (6)	S_7 (7)*
S_1 (1)	S_8 (8)*	S_2 (2)	S_8 (8)*	S_3 (5)	S_8 (8)*	S_4 (6)	S_8 (8)*

Other possible random assignments

Shallow	Deep	Shallow	Deep	Shallow	Deep	Shallow	Deep
S_1 (1)	S_2 (2)*	S_2 (2)	S_1 (1)	S_3 (5)	S_1 (1)	S_4 (6)	S_1 (1)
S_1 (1)	S_3 (5)*	S_2 (2)	S_3 (5)*	S_3 (5)	S_2 (2)	S_4 (6)	S_2 (2)
S_1 (1)	S_4 (6)*	S_2 (2)	S_4 (6)*	S_3 (5)	S_4 (6)*	S_4 (6)	S_3 (5)
S_5 (3)	S_1 (1)	S_6 (4)	S_1 (1)	S_7 (7)	S_1 (1)	S_8 (8)	S_1 (1)
S_5 (3)	S_2 (2)	S_6 (4)	S_2 (2)	S_7 (7)	S_2 (2)	S_8 (8)	S_2 (2)
S_5 (3)	S_3 (5)*	S_6 (4)	S_3 (5)*	S_7 (7)	S_3 (5)	S_8 (8)	S_3 (5)
S_5 (3)	S_4 (6)*	S_6 (4)	S_4 (6)*	S_7 (7)	S_4 (6)	S_8 (8)	S_4 (6)
S_5 (3)	S_6 (4)*	S_6 (4)	S_5 (3)	S_7 (7)	S_5 (3)	S_8 (8)	S_5 (3)
S_5 (3)	S_7 (7)*	S_6 (4)	S_7 (7)*	S_7 (7)	S_6 (4)	S_8 (8)	S_6 (4)
S_5 (3)	S_8 (8)*	S_6 (4)	S_8 (8)*	S_7 (7)	S_8 (8)*	S_8 (8)	S_7 (7)

Note. S = Subject; # of items recalled in parentheses; * Success (i.e., deep > shallow).

TABLE 5
Pairwise contrasts with one sample proportion tests and effect size indicators (r_{effect})
for eight replications in Study 2

	Number of successes	z	p	r_{effect}
Experiment 1	12	2.00	.023*	.45
Experiment 2	15	3.95	.001*	.70
Experiment 3	14	3.77	.001*	.69
Experiment 4	16	4.15	.001*	.72
Experiment 5	13	3.10	.001*	.61
Experiment 6	8	.29	.387	.07
Experiment 7	9	1.39	.082$^!$	.33
Experiment 8	11	2.09	.018*	.46

Note. $^!p < .10$. $^*p < .05$.

$$r_{effect} = \sqrt{\frac{z^2}{z^2 + n_{actual}}}$$

TABLE 6
Test results (*p* values) for three widely used nonparametric tests for eight replications in Study 2

	Mann-Whitney *U* test (*p*)	Kolmogorov-Smirnov goodness-of-fit test (*p*)	Wald-Wolfowitz runs test (*p*)
Study 2			
Experiment 1	.343	.699	.371
Experiment 2	.029*	.211	.371
Experiment 3	.057	.211	.371
Experiment 4	.029*	.037*	.029*
Experiment 5	.114	.211	.371
Experiment 6	.999	.999	.883
Experiment 7	.486	.999	.999
Experiment 8	.343	.699	.886

$^{*}p < .05$.

STUDY 3

The following analyses were conducted on data collected from a small sample of college students (*N* = 12). These individuals had participated in a larger between-groups experiment involving the effectiveness of mindfulness training on working memory (Frank et al., 2018). However, because these individuals had met diagnostic criteria for adult attention deficit/hyperactivity disorder (ADHD), their data were excluded from the original study. Frank et al. found that among the larger normative sample, a single mindfulness training session resulted in a significant increase in response accuracy during a computer-based working memory test. In the following section, we report the results of our analyses of the data from the small sample of excluded participants. We again compare the results of DPTs and traditional tests.

Method

Participants

The participants in the original study had been administered a self-report questionnaire to screen for symptoms of ADHD (Current Symptom Scale; CSS, Barkley & Murphy, 1998). In the original sample, 16 participants had scored above published diagnostic cutoffs for the ADHD-inattentive subscale. All of these individuals had participated in the original between-groups experiment (see the Subsection "Procedure"). Six of them had been assigned to the mindfulness condition and 10 had been assigned to the control condition. In order to create two groups of equal size (*n* = 6), we selected six of the control participants, based on matching their CSS scores to those of the mindfulness participants (see the Subsection "Instruments"). It should be noted that although the participants' high scores on the CSS would not be sufficient to warrant an actual clinical diagnoses, they do serve the purposes of this demonstration.

Instruments

The Current Symptom Scale (CSS; Barkley & Murphy, 1998) is a 4-point Likert-type scale derived from ADHD symptoms from the Diagnostic and Statistical Manual, 4th edition (American Psychiatric Association, 1994). This scale includes nine items related to inattention and nine items related to hyperactivity or impulsivity, although only the scores on the inattention items were used for the current study. Scores above 13.4 are considered to be above the clinical threshold for inattention for young adults (17-29 years; Barkley & Murphy, 1998). The mean CSS-inattention scores for the two mindfulness and control groups were roughly equivalent ($M_{mindfulness}$ = 21.5, $SD_{mindfulness}$ = 2.3, and $M_{control}$ = 22.7, $SD_{control}$ = 1.0).

A computer-based version of Kirchner's (1958) *n*-back task was used to assess working memory capacity. In this task, a random series of numbers (0-9) appeared one at a time in the center of the screen. Participants were instructed to compare each number with the number appearing two spaces prior in the sequence. Participants were instructed to press the red key if the two numbers matched, and to press the blue key if the number did not match. The percentage of correct responses for when the two numbers matched was used as the dependent variable, and as a measure of their working memory capacity.

Procedure

Participants first completed the CSS, as part of a battery of other assessments not included in the current study. They then participated in one of two training sessions. Participants assigned to the mindfulness condition participated in a ten-minute guided mindful breathing exercise. Participants assigned to the control condition participated in a 10-minute free-recall exercise in which they were asked reflect on a recent experience in their lives. All other conditions were kept constant across the two conditions (e.g., lighting, chair, voice on the instructional tape, etc.). After the training session, participants were administered the *n*-back task.

Results and Discussion

As expected, the percentage of matches correctly reported for the participants assigned to the mindfulness condition was greater than the percentage of matches correctly reported by the control participants (see Figure 4). When we conducted an independent groups *t* test, the effect was not found to be significant, despite a large observed effect size, t (10) = 1.64, p = .132, r_{effect} = .46. Similarly, none of the three nonparametric tests used in Study 2 (i.e., Mann-Whitney *U* test, Kolmogorov-Smirnov goodness-of-fit test, and Wald-Wolfowitz runs test) resulted in significant effects (all *p*'s > .05). However, when we conducted a DPT, the proportion of successes among the dyads associated with the original assignment (p_{obs} = .64) was significantly greater than the proportion of successes among all possible dyads (p_{total} = .49), Z = 2.42, p = .008, r_{effect} = .37.

As in Study 2, the DPT provided sufficient power to detect an experimental effect that was missed with an independent groups *t* tests or any of the other three nonparametric tests. It is important to note that the divergent results from the two approaches may be the product of either Type I error with the DPTs or a Type II error with the traditional *t* test. We suspect the latter is more likely, given that we essentially replicated an effect found with the original study. The magnitude of the effect size provided further support for the validity of the results from the DPT. We would encourage researchers to include such ancillary information in their evaluation of the results from small sample studies, regardless of the statistical tests used.

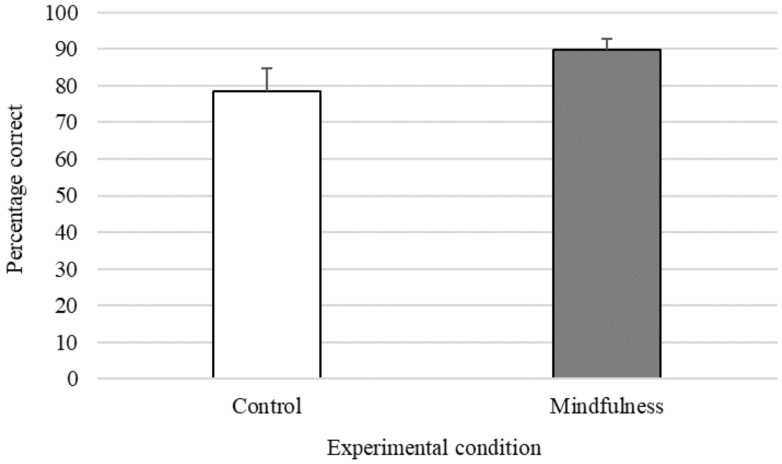

FIGURE 4
Percentage correct responses when target letter matched numbers two spaces prior in the sequence during the *n*-back task. Error bars reflect standard error.

GENERAL DISCUSSION

Our aim was to examine the utility of the DPT approach to analyze data from small sample experiments. The results from our Monte Carlo analyses in Study 1 suggest that the distribution of effects from our dyadic comparisons are normally distributed, similar to the distribution of effects produced with traditional *t* tests used with larger samples. In Study 2, we found that DPTs were better able to detect significant differences between the two experimental conditions than were the traditional *t* tests, even with as few as eight participants. Finally, Study 3 provided an example of how DPTs could be used with small clinical samples, again providing more power than traditional analyses.

Several points should be made about the appropriate application of DPTs. Whenever possible, researchers would be well-advised to recruit sufficiently large samples, based on a priori power analyses. Recruiting larger samples from the general population for a simple memory task used in Study 2 should not be an insurmountable obstacle. In that case, traditional analyses (e.g., *t* tests, chi square, etc.*)* would be easier to conduct, less subject to spurious results (i.e., Type I error), and more accessible for most readers. DPTs should be reserved for circumstances when recruiting larger samples is not feasible. For example, in Study 3, we were able to make use of data that had been excluded from an earlier study. It would not have been practical to recruit a larger sample. The dyadic proportions approach allowed us to make use of those data, rather than relegating them to the "file drawer."

DPTs provide a viable analytic tool for exploratory work. Researchers might use this approach with pilot studies and Phase 1 of Clinical Trials. In addition to providing a test of a hypothesized prediction, the results of this might be used for power analyses, for calculating prior probabilities for Bayasian analyses, and for subsequent meta-analyses across small samples. We would discourage researchers from overly relying on the results of a single study. However, as others have pointed out, replication is an integral aspect of all science, regardless of the analytic approach. In keeping with the recommendations of the American Psychological Association (APA) taskforce on statistical inference (see Wilkinson, 1999), we recommend that researchers report sufficient statistics to allow for meaningful interpretation (i.e., descriptive statistics and effect size indicators). We would also encourage authors to include or make available the results from traditional analyses (i.e., t or χ^2 tests), even when the DPTs are used for the primary analyses.

Finally, an important distinction should be made between the DPTs and traditional approaches. Traditional parametric tests (such as t tests) examine the likelihood of a result given the sample. DPTs examine the likelihood of a result given the random assignment. They are not appropriate for nonexperimental approaches (e.g., testing group differences based on diagnoses, age, or gender, or other nonrandomized conditions). Although subsequent applications may be developed for designs involving more than two groups or for repeated measures, we currently recommend the use of this procedure only with simple, two-group designs.

Conclusion

Although recruiting large samples offer clear advantages in terms of statistical power, this often is not feasible. By capitalizing on pairwise comparisons between individual scores in two independent groups designs, statistical power can be increased, even with extremely small samples. When used appropriately, the DPTs allow researchers to conduct experiments when larger samples are not possible. When this strategy is limited to testing theory-driven hypotheses, and when traditional effect size indicators are included, readers have an ability to evaluate the risk of Type I errors. This approach has the potential to add to the literature in meaningful ways.

Notes

1. The following two lists of stimuli were used: List 1# *chair, mathematics, elephant, lamp, car, elevator, thoughtful,* and *cactus*; List 2# *umbrella, exercise, forgiveness, rock, hamburger, sunlight, coffee,* and *bottle*. Each word was written on a single index card, and shown to the participant for approximately two seconds. Our experimenters were trained to use one list for the control condition and the other list with the experimental condition, and to alternate which lists were used for each condition, to avoid bias.
2. It should be noted, however, than in actual research the nonsignificant results ($p > .05$) would typically go unpublished (i.e., file draw problem), hence biasing such meta-analyses.

Acknowledgements

Anthony J. Nelson contributed to data collection for Study 2 while he was in residence at the State University of New York College at Cortland. He has contributed as an author while at the Pennsylvania State University.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Barkley, R. A., & Murphy, K. R. (1998). *Attention-Deficit Hyperactivity Disorder: A clinical workbook* (2nd ed.). New York, NY: Guilford Press.

Berger, V. W. (2000). Pros and cons of permutation tests in clinical trials. *Statistics in Medicine*, *19*(10), 1319-1328.

Bradley, M. T., & Gupta, D. (1997) Estimating the effect of the file drawer problem in meta-analysis. *Perceptual and Motor Skills*, *85*, 719-722.
doi:10.2466/PMS.85.6.719-722

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*, 98-101.
doi:10.1111/1467-8721.ep10768783

Craik, F., & Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior*, *11*, 671-684.
doi:10.1016/S0022-5371(72)80001-X

De Winter, J. C. F. (2013). Using the Student's *t*-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*, *18*(10), 1-12.

Etz, K. E., & Arroyo, J. A. (2015). Small sample research: Considerations beyond statistical power. *Prevention Science*, *16*, 1033-1036.
doi:10.1007/s11121-015-0585-4

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, Scotland: Oliver and Boyd.

Frank, J., Adams, A., Leverton, A., Delmuro, K., & Collings, R. D. (2018). *The effects of a single mindfulness exercise on attention and working memory*. Unpublished manuscript.

Goldstein, E. B. (2004). *Cognitive psychology: Connecting mind, research and everyday experience.* Belmont, CA: Thomson/Wadsworth.

Good, P. I. (2005). *Permutation, parametric and bootstrap tests of hypotheses* (3rd ed.). New York, NY: Springer Science & Business Media, Inc.

Hesterberg, T., Monaghan, S., Moore, D. S., Clipson, A., & Epstein, R. (2005). Bootstrap methods and permutation tests. In D. S. Moore & G. P. McCabe (Eds.), *Introduction to the practice of statistics* (5th ed., pp. 14.1-14.70). New York, NY: W. H. Freeman and Company New York.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352-358.
doi:10.1037/h0043688

LaFleur, B. J., & Greevy, R. A. (2009). Introduction to permutation and resampling-based hypothesis tests. *Journal of Clinical Child & Adolescent Psychology*, *38*, 286-294.
doi:10.1080/15374410902740411

Lehmann, E. L. (1999). "Student" and small-sample theory. *Statistical Science*, *14*(4), 418-426.

Maxwell, S. E. (2004) The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147-163.
doi:10.1037/1082-989X.9.2.147

Mehta, P., Antao, V., Kaye, W., Sanchez, M., Williamson, D., Bryan, L., . . . & Horton, K. (2014). Prevalence of amyotrophic lateral sclerosis - United States, 2010–2011. *Morbidity and Mortality Weekly Report: Surveillance Summaries*, *63*(7), 1-13.

Rosenthal, R. (1979) The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 683-641.
doi:10.1037/0033-2909.86.3.638

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage Publications.

Tamam, L., Karatas, G., Zeren, T., & Ozpoyraz, N. (2003). The prevalence of Capgras syndrome in a university hospital setting. *Acta Neuropsychiatrica*, *15*, 290-295.
doi:10.1034/j.1601-5215.2003.00039.x

Wilkinson, L. (1999) Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.
doi:10.1037/0003-066X.54.8.594