

Comparing Different Methods for Multiple Testing in Reaction Time Data

Massimiliano Pastore Massimo Nucci Giovanni Galfano
Università di Padova

Reaction times were simulated for examining the power of six methods for multiple testing, as a function of sample size and departures from normality. Power estimates were low for all methods for non-normal distributions. With normal distributions, even for small sample sizes, satisfactory power estimates were observed, especially for FDR-based procedures.

Keywords: multiple testing, reaction times, power, False Discovery Rate, Type I error

Introduction

Scientific research often deals with the problem of performing many tests of significance. However, this practice results in an increase of the likelihood of committing one or more Type I errors, which grows as the number of tests increases (e.g., Keselman, Cribbie, & Holland, 1999). In the most common approach, error rate is familywise controlled (Familywise Error Rate, FWER) by reducing the α value as a direct function of the number of comparisons to be computed. In the classic Bonferroni method (1936), the threshold probability (α_{FW} , usually set at .05) is divided by the total number of comparisons. This approach to controlling errors in multiple-testing contexts ensures that the probability of committing Type I error at least

once is $\leq \alpha$. The intrinsic limit of multiple testing with FWER control is that such approach becomes more conservative as the number of tests rises: Indeed, a major criticism frequently levelled at multiple testing is their lack of power.

A different perspective to controlling Type I error when performing many tests of significance is represented by the False Discovery Rate (FDR). This statistical procedure, introduced by Benjamini and Hochberg (1995), can be implemented in all those experimental contexts in which the computation of a large number of comparisons is required. The FDR is focused on the proportion of errors committed when H_0 is rejected, which results in both keeping Type I error under control and in an increase of power. Further advantages characterizing FDR are represented by its easy and quick implementation (Thissen, Steinberg, & Kuang, 2002), and by its wide applicability, as proved by the fact that FDR can be adopted when multiple comparisons involve either independent or correlated test statistics (Benjamini & Yekutieli, 2001).

A third possible way for dealing with multiple testing is represented by resampling-based procedures (Westfall & Young, 1993). Following this approach, the values of observed variables are randomly re-assigned to the experimental groups, and then the test statistics are re-computed. Thus, the resampling-based p-value is the proportion of resampled data sets

Massimiliano Pastore is Assistant Professor of Psychometrics in the Faculty of Statistical Science, Dipartimento di Psicologia dello Sviluppo e della Socializzazione, at the University of Padova. E-mail: massimiliano.pastore@unipd.it. Massimo Nucci is Assistant Professor of Psychometrics in the Faculty of Psychology, Dipartimento di Psicologia Generale at the University of Padova. Giovanni Galfano is Assistant Professor of Experimental Psychology in the Faculty of Psychology, Dipartimento di Psicologia dello Sviluppo e della Socializzazione at the University of Padova.

MULTIPLE TESTING IN REACTION TIME DATA

yielding a statistic as extreme as the original statistic.

In this article, a Monte Carlo study is illustrated aimed at comparing the performance of six different procedures for treating multiple testing. The analysis has been conducted in the context of multiple comparisons among means resulting from nonnormally-distributed and correlated variables. Specifically, the classic Bonferroni method, two single-step FDR methods, two resampling-based methods, and a combined resampling-based FDR procedure were examined.

These methods were used for adjusting p-value and then comparing their power. Because in multiple comparison testing more than one definition of power and Type I error rate is available, three different definitions associated to these measures were considered. It should be noted that FDR-based methods do not control for FWER (e.g., Wilcox, 2003). However, it is important to remark that comparing power of methods that do not have similar control over Type I errors can provide critical information as to the choice of a particular test in light of the associated costs (in terms of Type I error) and benefits (in terms of power; e.g., Horn & Dunnett, 2004).

Reaction time (RT) data were simulated for this research. The present study focused on this particular type of variable for two main reasons. First, RTs represent the dominant dependent measure in cognitive psychology (e.g., Van Zandt, 2002). Second, RTs possess critical features that make them hard to be analyzed with classical statistical procedures (Heathcote, 1996).

In the most common experimental paradigms using RTs, participants are submitted to a series of stimuli that have to be responded to as fast as possible. Therefore, measurements can hardly be considered as independent from each other. In addition, it is well known that RTs are not distributed according to a normal function (e.g., Schwarz, 2001; Van Zandt, 2000). McGill (1963) and Hohle (1965) proposed as a descriptive model of RTs, a theoretical distribution obtained through the convolution of a normal distribution and an exponential distribution, subsequently known as ex-Gaussian (Burbeck & Luce, 1982). Although other

descriptive models are available such as the ex-Wald, the Weibull and the Gamma distributions (see, e.g., Schwarz, 2001; Van Zandt, 2000), to date the ex-Gaussian distribution is among the most representative models for describing RTs (Ratcliff, 1978; 1979; Ratcliff & Murdock, 1976). In addition, it is worth noting that, using the ex-Gaussian model, the usefulness of decomposing the normal and exponential components has been consistently demonstrated (e.g., Heathcote, 1996; Heathcote, Popiel, & Mewhort, 1991). For example, the simple arithmetic mean cannot be considered a satisfactory statistic within this context, given the skewness characterizing RTs. By contrast, there is wide agreement that ex-Gaussian parameters are more appropriate for describing (and interpreting) RTs (Heathcote, 1996). In the present paper, the ex-Gaussian distribution was adopted as a plausible model for RT data.

An experimental setting with three stimuli requiring a response of some sort was simulated. Each stimulus was repeated three times. Multiple comparisons among the observed RT means, obtained in this hypothetical task were then performed. Both sample size and the magnitude of the RT exponential component, were manipulated. The estimated power of the six procedures was then compared. Before illustrating the methods and results of the Monte Carlo study, the basics of p-value adjustment in the examined procedures will be outlined, and the features of the ex-Gaussian distribution and analysis will be briefly summarized.

p-Value Adjustment

Suppose there is interest in testing m hypotheses simultaneously. For each hypothesis H_i , $i=1,2,\dots,m$, m test statistics and the relative p-values will be computed. It is possible to compute an adjusted p-value \tilde{p}_i for each test. Thus, the decision to reject H_i at FWER = α is obtained by merely checking whether $\tilde{p}_i \leq \alpha$. According to Westfall and Young (1993, p. 11), the mathematical definition of an adjusted p-value is as follows:

$$\tilde{p}_i = \inf\{\alpha : H_i \text{ is rejected at FWER} = \alpha\} \quad (1)$$

That is, \tilde{p}_i is the smallest significance level for which one still rejects H_i , given a particular simultaneous test procedure. Adjusted p-values for FDR controlling procedures are defined similarly (Yekutieli & Benjamini, 1999):

$$\tilde{p}_i = \inf\{\alpha : H_i \text{ is rejected at FDR} = \alpha\} \quad (2)$$

In the present study, the following p-value adjustment procedures were considered: Bonferroni adjustment (B), two single-step FDR-type adjustments, that is Benjamini-Hochberg (BH; the basic FDR method) and Benjamini-Yekutieli (BY), and three resampling-based adjustments, that is the method described by Reiner, Yekutieli and Benjamini (2003; RYB) and two methods proposed by Ge, Dudoit and Speed (2003), called maxT and minP. Whereas B, minP, and maxT control FWER, BH, BY, and RYB control FDR. B adjustment. This adjustment by Bonferroni (1936) consists of multiplying each observed probability, p_i , by the number of comparisons that have been performed. In case the value resulting from this computation exceeds 1, then probability is set at 1:

$${}_B\tilde{p}_i = \min(p_i m, 1) \quad (3)$$

BH adjustment. This method has been introduced by Benjamini and Hochberg (1995) for independent and positive regression dependent test statistics. Let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ be the observed probabilities arranged in increasing order, then:

$${}_{BH}\tilde{p}_i = \min\left\{p_{(j)} \frac{m}{j} : i \leq j\right\}; \quad j = 1, \dots, m \quad (4)$$

BY adjustment has been proposed by Benjamini and Yekutieli (2001) for controlling general dependency structures.

Let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ be the observed probabilities arranged in increasing order, then:

$${}_{BY}\tilde{p}_i = \min\left\{p_{(j)} \frac{m \sum_{k=1}^m \frac{1}{k}}{j} : i \leq j\right\}; \quad j = 1, \dots, m \quad (5)$$

RYB adjustment. This is a resampling-based FDR adjustment. In particular the method described by Reiner et al. (2003) was considered, which can be summarized as follows: First, the data are repeatedly resampled under complete null hypothesis (meaning that all H_i are true) and a vector of resampling-based p-values is computed for each H_i . For the k-th hypothesis, with an observed test statistics t_k , the estimated p-value is:

$$p_k^{est} = \frac{\#\{t_{ij}^* : |t_{ij}^*| \geq |t_k|\}}{mN} \quad (6)$$

where $i = 1, \dots, m$, m are the number of hypotheses, $j = 1, \dots, N$, N the number of resampling, and t_{ij}^* are the resampling-based test statistics.

The adjusted p-values using the BH adjustment is obtained as follows:

$${}_{RYB}\tilde{p}_i = \min\left\{p_{(k)}^{est} \frac{m}{k} : i \leq k\right\} \quad (7)$$

Resampling maxT adjustment

This algorithm, originally proposed by Westfall and Young (1993), has been further examined by Ge et al. (2003). The step-down maxT adjusted p-values are defined by:

$${}_{maxT}\tilde{p}_i = \max_{k=1, \dots, i} \{\Pr(\max_{l=k, \dots, m} |T_l| \geq |t_{(k)}|) / \text{under complete null hypothesis}\} \quad (8)$$

MULTIPLE TESTING IN REACTION TIME DATA

where $|T_l|$ is the random variable associated to the statistical test, and $|t_{(1)}| \geq |t_{(2)}| \geq \dots \geq |t_{(m)}|$ denote the ordered observed test statistics.

Resampling minP adjustment

This algorithm was also put forward by Westfall and Young (1993). However, the version considered in the present study is based on a modified adaptation (see Ge et al., 2003). The step-down minP adjusted p-values are defined by

$$\min_P \tilde{p}_i = \max_{k=1, \dots, i} \{ \Pr(\min_{l=k, \dots, m} P_l \leq p_{(k)} / \text{under complete null hypothesis}) \} \quad (9)$$

where P_l denotes the random variable for the unadjusted p-value of the l -th hypothesis and $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ denote the ordered observed p-values.

Ex-Gaussian Distribution

The ex-Gaussian function is identified as a good theoretical approximation of RT distribution (e.g., Heathcote, 1996; Heathcote et al., 1991; Van Zandt, 2000) and its shape can be formally described as follows:

$$f(t) = \frac{1}{t} e^{-\frac{t}{\tau} + \frac{\mu}{\tau} + \frac{\sigma^2}{2\tau^2}} \Phi \left(\frac{\tau - \mu - \frac{\sigma^2}{\tau}}{\sigma} \right) \quad (10)$$

where

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \quad (11)$$

This density function depends on three parameters: μ and σ , corresponding to mean and standard deviation of the Gaussian component of the distribution respectively, and τ , corresponding to the exponential component of the distribution. Ratcliff (1979) showed that mean (${}_{RT}\mu$) and standard deviation (${}_{RT}\sigma$) of the ex-Gaussian can be rewritten as a function of these three parameters. In particular:

$${}_{RT}\mu = \mu + \tau \quad (12)$$

and

$${}_{RT}\sigma = \sqrt{\sigma^2 + \tau^2} \quad (13)$$

Examples of ex-Gaussian density functions are depicted in Figure 1, where the influence of the exponential component on the shape of the distribution function is illustrated. The curves have $\mu = 550$ and $\sigma = 50$ as fixed parameters, whereas the τ value is varied. It is worth noticing that the exponential component determines an increase of the positive skew.

As briefly anticipated earlier, Heathcote (1996; Heathcote et al., 1991; Mewhort, Braun, & Heathcote, 1992) has proposed an RT analysis method based on the properties highlighted above. In particular, Heathcote (1996) has developed a statistical package, RTSYS, that allows researchers to easily obtain values for μ , σ , and τ by means of RT decomposition.

Several arguments support the need of using an RT decomposition technique prior to statistical analysis. First, RT data can contain extreme values (i.e., outliers) that do not reflect the effects of the independent variables and can be problematic for interpreting the results. Solutions to the problem of outliers usually rely on trimming observations (e.g., Ratcliff, 1993; Van Selst & Jolicoeur, 1994; Wilcox, 2005). However, finding a general criterion for removing data is problematic because real data are almost inevitably rejected along with spurious data. Second, as discussed above, skew in RT distribution can cause serious problems of interpretation for descriptive statistics. For instance, a given independent variable may influence the mean and median differently by modifying the degree of skew. It should also be stressed that significantly skewed data violate the assumption underlying most parametric tests, that variability in data is normal. Whereas the common approach in research practice is to ignore skew, several studies have shown that the magnitude of skew often contains information about the effect of experimental manipulations (Ratcliff & Murdock, 1976; Heathcote et al., 1991; Campbell & Penner-Wilger, 2006). It

follows that even if one circumvents the problem of violating the normality assumption of parametric tests by transforming RTs, the risk of losing information and missing potentially important effects is still present.

Turkington, Provenzale, Hawk, & Coleman, 1999; Mewhort, & Johns, 2000; Penner-Wilger, Leth-Steensen, & LeFevre, 2002; Spieler, Balota, & Faust, 2000; West, Murphy, Armilio, Craik, & Stuss, 2002).

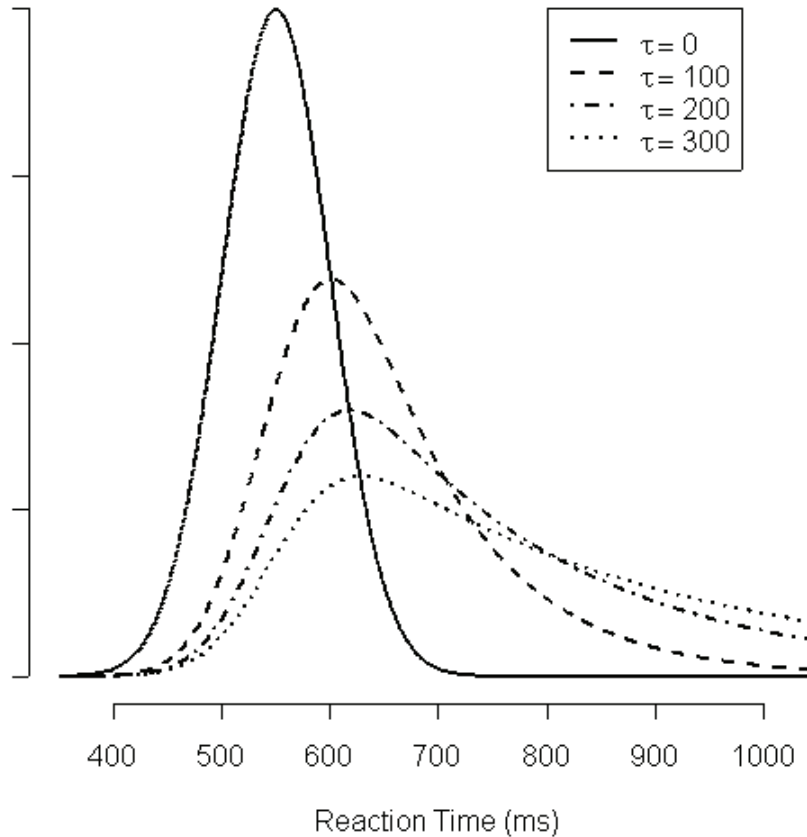


Figure 1. Ex-Gaussian density functions with $\mu = 550$, $\sigma = 50$ and $\tau \in \{0,100,200,300\}$

In summary, through quantifying RT distribution shape, ex-Gaussian decomposition can reveal structure within RT data not revealed by conventional analyses. Such procedure has successfully been adopted in a variety of studies dealing with RTs in several research fields related to cognitive psychology (e.g., Andrews & Heathcote, 2001; Armstrong & Munoz, 2003; Balota & Spieler, 1999; Dell'Acqua, Job, & Grainger, 2001; Leth-Steensen, Elbaz, & Douglas, 2000; Madden, Gottlob, Denny,

Methodology

Data Generation

Data were generated and analyzed by means of a custom-made program written in R (R Development Core Team, 2003). Random number generation was achieved by using the Mersenne-Twister method (Matsumoto & Nishimura, 1998). This generator guarantees far longer period and far higher order of

MULTIPLE TESTING IN REACTION TIME DATA

equidistribution than any other implemented generators.

RTs were generated through the application of the *rnorm* function concerning the normal component (with μ and σ as mean and standard deviation, respectively) and the *rexp* function concerning the exponential component (with τ as parameter), as follows:

$$RT = rnorm(n, \mu, \sigma) + r \exp(n, 1/\tau) \quad (14)$$

Clearly, with $\tau = 0$, the exponential component is set to 0. As a result, the ex-Gaussian function reduces to a normal distribution with mean μ and standard deviation σ . Note that this very same values would be obtained after performing the RT decomposition algorithm (Heathcote, 1996) in any generated RT raw data set.

In order to generate correlated data, the method described by Jöreskog and Sörbom (1996, p. 189-190) was used. Such method is based on the adoption of a lower triangular matrix T such that $\Sigma = TT'$, where Σ is the population correlation matrix. Application of such procedure ensures the generation of correlated ex-Gaussian distributions.

Experimental Design

An experimental situation with three stimuli (e.g., pictures) requiring a speeded response of some sort in a given task (e.g., picture naming) was simulated. Each stimulus was repeated three times. Multiple comparisons were then performed among all the observed RT means. In such a context, differences may be expected for comparisons between different stimuli. Conversely, no differences should be expected in comparisons between repetitions of the same stimulus. Whereas in an empirical setting this latter type of comparisons may be relevant to test the consistency of a given stimulus (or participant), in the present study it was critical for evaluating Type I error.

The parameters for the simulation were chosen after an extensive review and analysis of the studies employing the ex-Gaussian decomposition technique cited above. Specifically, RT means ranged from 446 (Spieler et al., 2000) to 1199 milliseconds (Leth-

Steensen et al., 2000). Using the ex-Gaussian decomposition, the mean value of μ was about 522 milliseconds, ranging from 286 (Dell'Acqua et al., 2001) to 865 (Leth-Steensen et al., 2000). σ varies between 32 (Spieler et al., 2000) and 175 (Leth-Steensen et al., 2000), with mode 50. The estimated values of τ ranged from 41 (Spieler et al., 2000) to 414 (Leth-Steensen et al., 2000). Consequently, three distributions were considered (one for each of the three stimuli) with mean $\mu_1 = 595$, $\mu_2 = 550$, and $\mu_3 = 535$, all of which had a standard deviation of $\sigma = 50$, and four values of $\tau : 0, 100, 200$, and 300. In addition, the correlation value across distributions was set to $\rho = .6$, with the purpose of simulating a setting with a medium-to-high correlation level, and the correlation value within distributions was set to $\rho = 1$.

To summarize, the notation $ExG(\mu, \sigma, \tau)$ was used to indicate a generic ex-Gaussian distribution with μ , σ , and τ as parameters. Consequently, the resulting three distributions were defined as follows:

$$D_1 \approx ExG(595, 50, \tau)$$

$$D_2 \approx ExG(550, 50, \tau)$$

$$D_3 \approx ExG(535, 50, \tau)$$

$$\text{where } \tau \in \{0, 100, 200, 300\}.$$

The manipulation of τ was aimed to evaluate the performance of the six p-value adjustment methods as a function of departures from normality.

For each of the three distributions (D_1, D_2 and D_3) three repetitions were performed, thus producing nine RTs in total. A scheme representing the procedure adopted is depicted in Figure 2. The sample size was varied in four different sizes (n): 12, 20, 40, and 80. These particular values were chosen because they are representative of those generally adopted in empirical research (e.g., Andrews & Heathcote, 2001; Dell'Acqua et al., 2001).

By combining the four chosen τ values with the four different sample sizes, sixteen different scenarios were obtained. For each scenario, the sampling was replicated five

thousand times. Therefore, the total number of generated samples was $4 \times 4 \times 5000 = 80000$.

Pairwise Comparisons

For each sample, after computing mean RTs, all the possible paired comparisons were

5. maxT: following definition (8). For each of the 5000-generated raw data sets, data were resampled 1000 times.
6. minP: following definition (9).

For each of the 5000-generated raw data sets, data were resampled 1000 times.

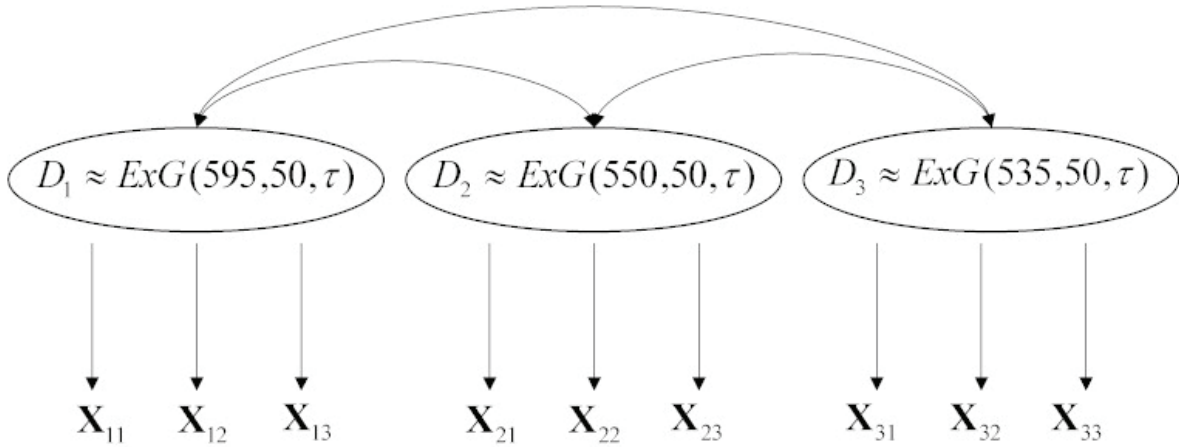


Figure 2. Schematic summary of data generation. X_{ij} with $i, j \in \{1, 2, 3\}$, is the j -th variable obtained from the D_i distribution.

performed by means of paired samples t -tests, equals to $\binom{9}{2} = 36$. In order to determine whether the difference was statistically significant, the p -value adjustments described earlier were used:

1. B: following the procedure specified in (3).
2. BH: following the procedure specified in (4).
3. BY: following the procedure specified in (5).
4. RYB: following Resampling FDR Adjustment definition described in (6) and (7); a modified version of the R program by A. Reiner available over the internet at <http://www.math.tau.ac.il/~ybenja> was used. For each of the 5000-generated raw-data sets, data were resampled 1000 times.

For both maxT and minP, the R-package Multtest by Dudoit and Ge was used. This may be downloaded from the Bioconductor website <http://www.bioconductor.org/>.

The tested hypothesis was the following:

$$H_{0:RT} : \mu_{ij-RT} - \mu_{hk} = 0$$

where

$$i, j, h, k \in \{1, 2, 3\}$$

and

$$(i, j) \neq (h, k) \quad (15)$$

This hypothesis is true when the comparison is made between two variables belonging to the same distribution, and false when the variables belong to different distributions. The Null Hypothesis status for the considered comparisons is shown in Table 1.

MULTIPLE TESTING IN REACTION TIME DATA

Table 1: Null Hypothesis status in the examined comparisons

	μ_{11}	μ_{12}	μ_{13}	μ_{21}	μ_{22}	μ_{23}	μ_{31}	μ_{32}	μ_{33}
μ_{12}	true								
μ_{13}	true	true							
μ_{21}	false	false	false						
μ_{22}	false	false	false	true					
μ_{23}	false	false	false	true	true				
μ_{31}	false	false	false	false	false	false			
μ_{32}	false	false	false	false	false	false	true		
μ_{33}	false	false	false	false	false	false	true	true	

Table 2. True value of the differences between means

Comparisons	θ	
$\mu_{ij} - \mu_{ik}$	0	with $i, j, k \in \{1,2,3\}$ for $j \neq k$
$\mu_{2j} - \mu_{3k}$	15	with $i, j, k \in \{1,2,3\}$
$\mu_{1j} - \mu_{2k}$	45	with $i, j, k \in \{1,2,3\}$
$\mu_{1j} - \mu_{3k}$	60	with $i, j, k \in \{1,2,3\}$

The true values of the differences between means ($\theta = \mu_{ij} - \mu_{hk}$), are represented in Table 2. As a result, nine comparisons for each of the θ values were considered. Note that when $\theta=0$, H_0 is true, being false in all the other cases.

Empirical Evaluation of Power and Type I Error Rate

Because the present study was aimed at evaluating the power of each adjustment procedure, defining power represents a critical

issue. Crucially, in multiple testing situations, power is not univocally characterized. In the H_0 was rejected at least once was computed.

This value was then divided by the total number of replications (i.e., 5000). In the experimental practice, the any-pair definition is generally chosen for dealing with exploratory scenarios, because of a higher discriminatory capability.

All-pair power was the probability of correctly rejecting all hypotheses for each level of $\theta > 0$ (Ramsey, 1978; Westfall & Young, 1993, p. 205). Consequently, the number of

times, for each level of θ , in which all H_0 were rejected was computed. This value was then divided by the total number of replications (i.e., 5000). In the experimental practice, the all-pair power definition is generally chosen when missing the rejection of even a single false H_0 has particularly dramatic consequences.

Per-pair power was the rejection probability for a given pair of hypotheses, for each level of $\theta > 0$ (Ramsey, 1978). Consequently, the number of rejected H_0 was counted and then divided by the total number of hypotheses for each level of $\theta > 0$ (i.e., $9 \times 5000 = 45000$). In the experimental practice, the per-pair power definition is generally adopted in meta-analytic contexts (Westfall & Young, 1993), and can be interpreted as an intermediate solution between any-pair and all-pair definitions.

When $\theta = 0$, all H_0 are true. Hence, the number of times in which H_0 were rejected was evaluated for estimating Type I error rate. Three types of Type I error rate were considered:

FWER was the probability of rejecting at least one true null hypothesis. Consequently, the number of times in which H_0 was rejected at least once was counted. This value was then divided by the total number of replications (i.e., 5000).

FDR was the expectation of the proportion of the rejected null hypotheses which are erroneously rejected. Consequently, the proportion of erroneously rejected H_0 was counted. This value was then divided by the total number of replications (i.e., 5000).

Per-Comparison error rate (PCER) was the rejection probability for a given pair of true null hypotheses. Consequently, the number of rejected H_0 was counted and then divided by the total number of hypotheses in which $\theta = 0$ (i.e., $9 \times 5000 = 45000$).

Because the computed values associated to the different power and Type I error definitions vary as a function of the proportion of true null hypotheses (cfr. Dudoit, Shaffer, & Boldrick, 2003), it is worth noting that, in the present context, this proportion was .25.

Results

For each of the sixteen considered scenarios, before estimating power, the mean number of significant tests for all the considered values of the θ parameter was computed. It must be stressed once again that sampling was replicated five thousand times.

Type I error rates

Type I error estimates are illustrated in Table 3. Given that the different methods control different kinds of Type I error, following Dudoit et al. (2003), FWER estimates are reported for B, maxT and minP, whereas FDR estimates are reported for BH, BY and RYB. In addition, PCER estimates are reported for unadjusted p-values (rawp). Inspection of Table 3 shows that B always succeeded in keeping Type I error under .05. The performance of all the remaining methods was modulated by both sample size (n) and the magnitude of the exponential component (τ). More specifically, all methods were weakened as τ increased, whereas increasing sample size resulted in a more efficient control. Crucially, however, when sample size was sufficiently large ($n = 80$), all the FDR-based methods (BH, BY, and RYB) were effective in controlling Type I error adequately even when the magnitude of the exponential component was highest ($\tau = 300$).

Any-Pair Power

Figures 3 and 4 represent the power estimates obtained with n set at 12 and 80, respectively. The four graphs in each figure represent the functions obtained for each specific τ value (0, 100, 200, 300) with the six different methods. In abscissa the value of the θ parameter (i.e., the real difference between means) is represented.

As a general trend, an expected increase of significant results as both θ and n increased can be observed. However, it is worth remarking that the number of significant tests dramatically decreased as τ increased, thus showing that departures from normality directly result in a loss of power.

For $\tau > 0$, RYB showed the best performance when sample size was small ($n =$

MULTIPLE TESTING IN REACTION TIME DATA

12). As sample size increased, however, RYB performance was equivalent to BH performance in all conditions. When sample size was sufficiently large ($n = 80$), all methods seemed to achieve a good performance in terms of power even for moderate values of the exponential component ($\tau = 100$). Finally, for $\tau > 100$, RYB and BH showed the best performance, followed by maxT and minP (showing overlapping functions), and BY and B (showing overlapping functions).

In general, the results seem to suggest that for small sample sizes (e.g., $n = 12$, Figures 3, 5 and 7) the power of all methods tended to lower as the value of τ increases, meaning that the likelihood of committing a Type II error tends to rise as the distribution progressively departs from normality. The performance of RYB and BH always proved the best. Also, a general order relationship emerged, for every power definition, so that

$$RYB \cong BH \geq \max T \cong \min P \geq BY \cong B \quad (16)$$

where $X \cong Y$ denotes that X is approximately equivalent to Y, and \geq denotes that X is equivalent or more powerful than Y.

For $n = 80$ (Figures 4, 6 and 8), all methods achieved acceptable power estimates even when $\tau = 100$, provided that $\theta > 15$. This seems to suggest that, with a large sample size, departures from normality do not strongly affect power. When $\tau = 200$, neither RYB nor BH revealed a fully satisfactory performance even for $\theta > 45$. These two methods tended to produce an equivalent performance in all the different scenarios. The order relationship emerged in the situations with lower sample sizes was confirmed, with RYB and BH being the most powerful methods, and B the least.

Conclusion

The present article was aimed at comparing the power of six different p-value adjustment procedures for treating multiple testing. In particular, RTs, which are the main dependent variable in many experimental contexts related to cognitive psychology (Van Zandt, 2002), were considered. Because it is well known that

RTs are not distributed normally, the six p-value adjustment procedures were evaluated by manipulating the parameters related to the Ex-Gaussian distribution. This distribution was chosen because it is one of the most prominent descriptive models for RTs in the literature (Van Zandt, 2000). In order to maintain a close reference with empirical research, the values of the different parameters were chosen based on a series of studies that have employed an RT decomposition technique. This allowed for the examination of the effects of departures from normality on the power estimate associated to each different p-value adjustment procedure. In addition, sample size was manipulated, whose values were selected following the same studies that used the RT decomposition technique. Because sample size is often quite small, the present study tested whether this factor played a major role in modulating the shape of the power function.

As a general comment, two main results emerged in the present investigation. First, the power of the different adjustment procedures was substantially influenced by both sample size and the shape of the distribution. Second, the adjustment procedures included in the present study can be ordered in a constant relationship. In particular, RYB always resulted the most powerful method, although closely followed by BH, whereas B, as expected, appeared very conservative in all the different scenarios. The difference between the most powerful methods (i.e., RYB and BH) and the remaining adjustment procedures was more pronounced for $\theta = 15$. This result is important, because the phenomena investigated in cognitive psychology and mental chronometry are often inferred on the grounds of mean differences in similar orders of magnitude. Relevant examples are offered by the Simon effect (see Lu & Proctor, 1995, for a review), the inhibition of return effect (see Klein, 2000, for a review), and the semantic priming effect (see Neely, 1991, for a review).

Table 3. Type I error estimates as a function of sample size (n) and departures from normality (τ). FWER estimates are reported for B, maxT and minP, FDR estimates are reported for BH, BY and RYB, and PCER estimates are reported for unadjusted p-values (rawp).

n	τ	PCER		FWER			FDR		
		rawp	B	maxT	minP	BH	BY	RYB	
12	0	.048	.011	.035	.033	.012	.004	.012	
	100	.062	.026	.068	.051	.054	.025	.060	
	200	.069	.036	.100	.073	.142	.105	.200	
	300	.070	.045	.112	.081	.255	.193	.327	
20	0	.049	.011	.038	.038	.012	.003	.012	
	100	.056	.029	.067	.061	.031	.017	.032	
	200	.065	.045	.099	.083	.112	.096	.138	
	300	.067	.047	.099	.083	.216	.206	.248	
40	0	.048	.010	.043	.044	.012	.003	.012	
	100	.055	.026	.060	.056	.016	.006	.016	
	200	.057	.033	.073	.065	.046	.037	.051	
	300	.060	.037	.072	.063	.106	.094	.108	
80	0	.050	.010	.042	.042	.013	.003	.013	
	100	.052	.020	.054	.052	.013	.004	.013	
	200	.053	.027	.060	.058	.023	.012	.023	
	300	.054	.025	.056	.054	.048	.029	.050	

MULTIPLE TESTING IN REACTION TIME DATA

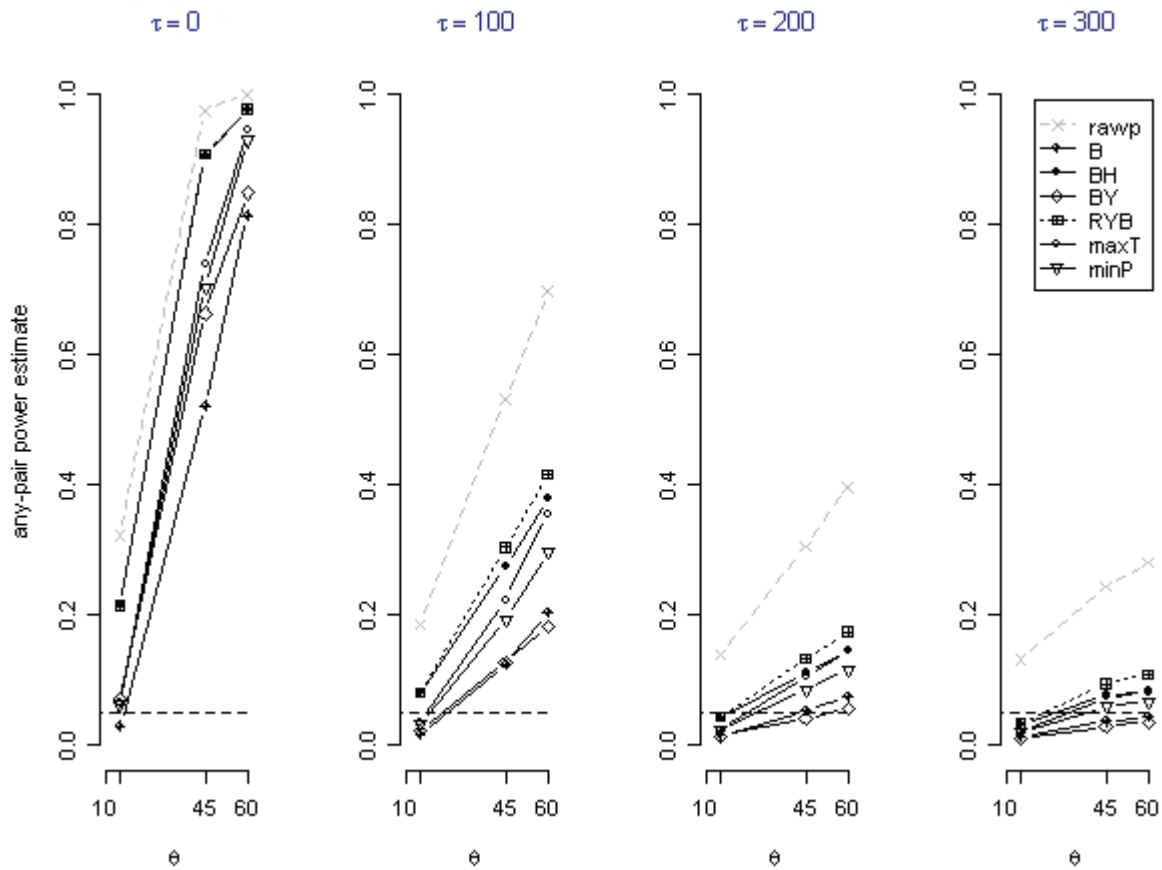


Figure 3. Any-pair power estimates for the different p-value adjustment methods as a function of the true difference between means (θ) for $n = 12$. Each graph refers to a different τ value (0 to 300, from left to right). B = Bonferroni method (FWER); BH = Benjamini-Hochberg (FDR); BY = Benjamini-Yekutieli (FDR); RYB = Reiner-Yekutieli-Benjamini (resampling-based FDR); minP and maxT (resampling). The horizontal line refers to .05.

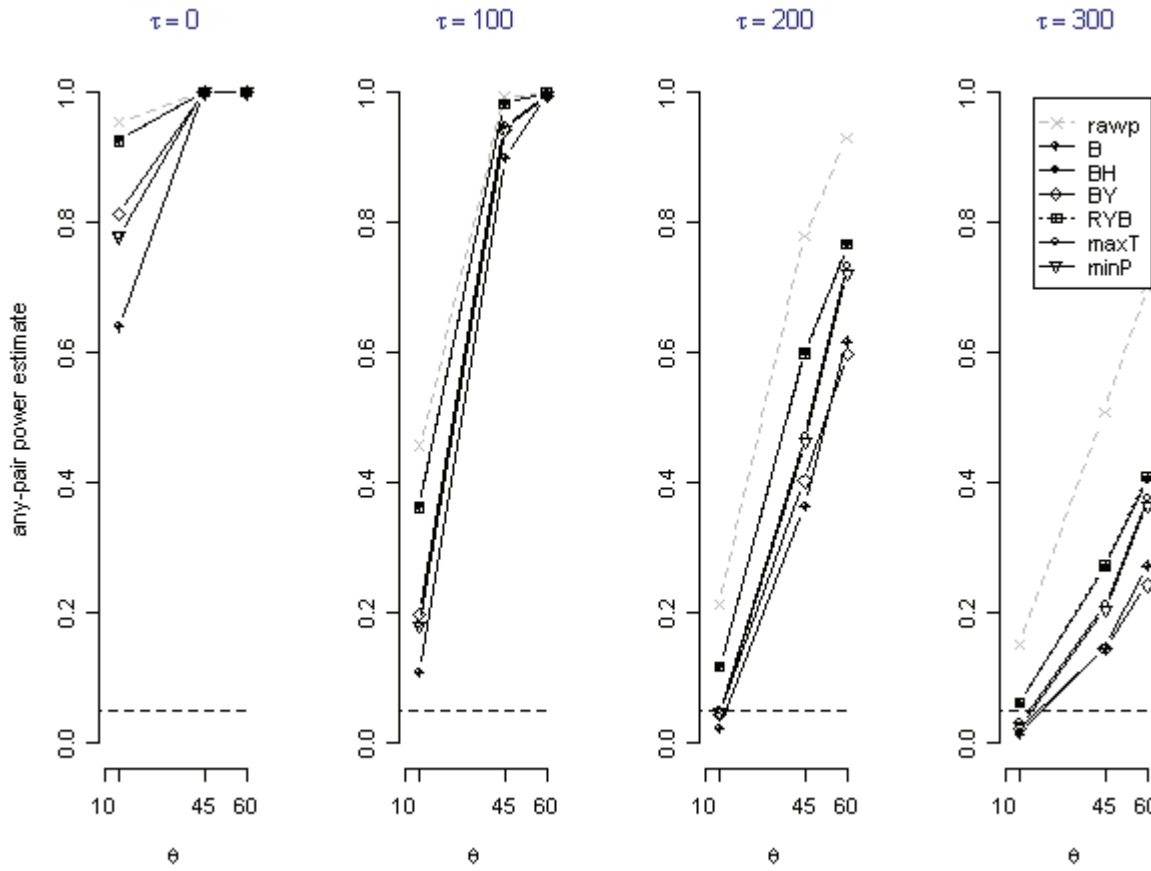


Figure 4. Any-pair power estimates for the different p-value adjustment methods as a function of the true difference between means (θ) for $n = 80$. Each graph refers to a different τ value (0 to 300, from left to right). Conventions as in Figure 3.

MULTIPLE TESTING IN REACTION TIME DATA

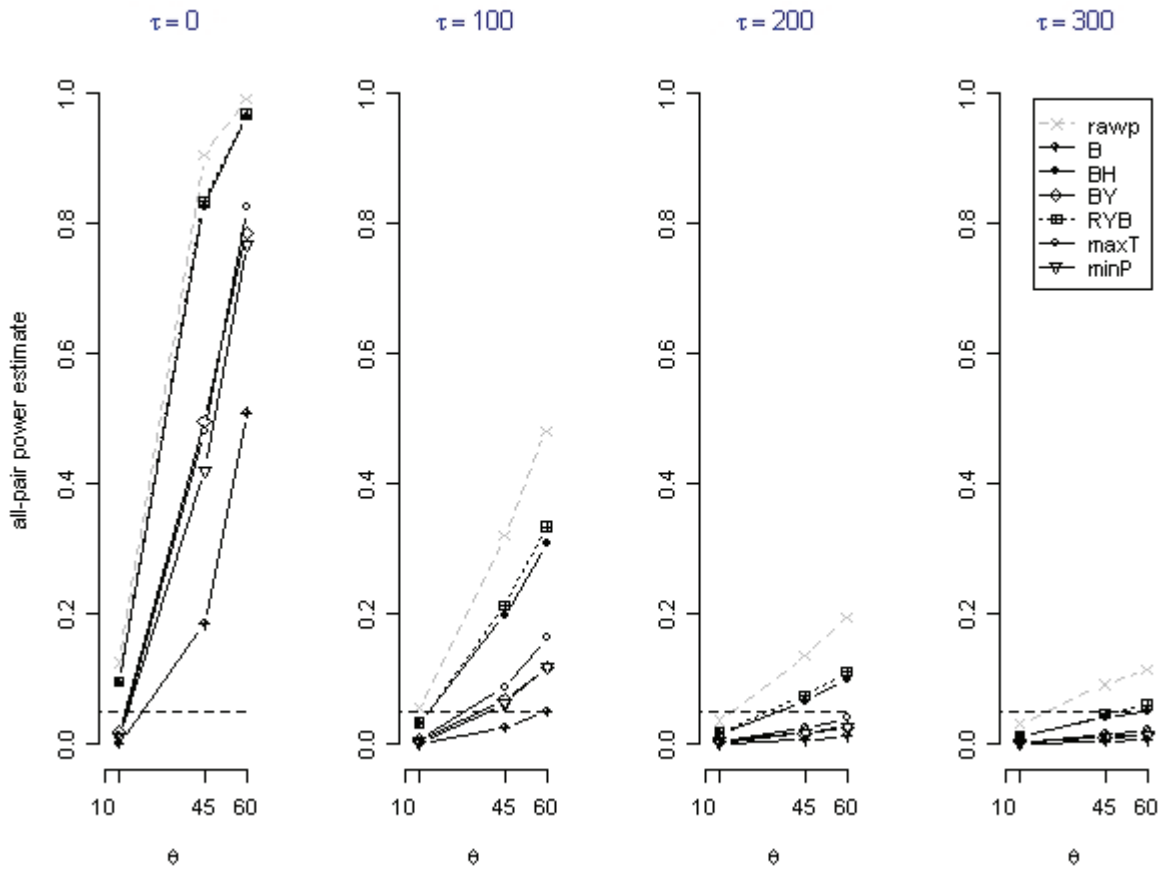


Figure 5. All-pair power estimates for the different p-value adjustment methods as a function of the true difference between means (θ) for $n = 12$. Each graph refers to a different τ value (0 to 300, from left to right). Conventions as in Figure 3.

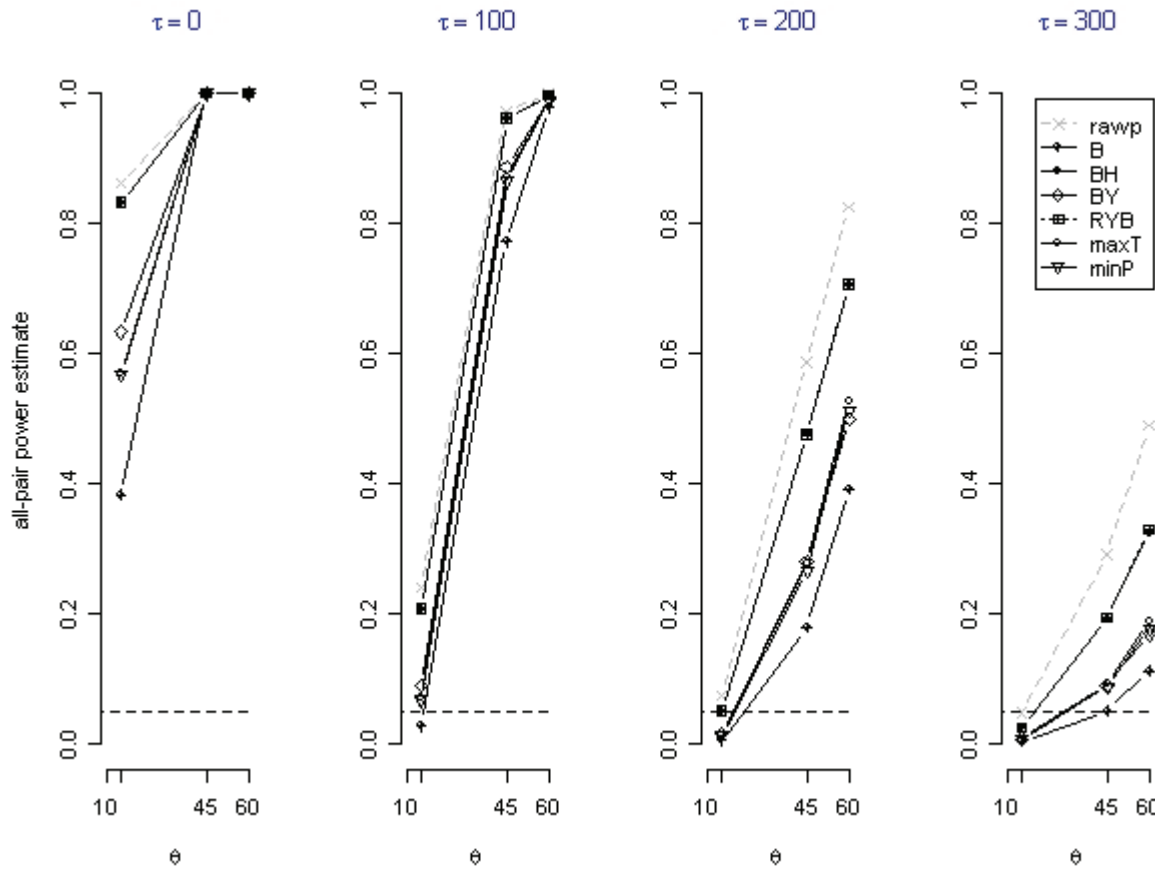


Figure 6 All-pair power estimates for the different p-value adjustment methods as a function of the true difference between means (θ) for $n = 80$. Each graph refers to a different τ value (0 to 300, from left to right). Conventions as in Figure 3.

MULTIPLE TESTING IN REACTION TIME DATA

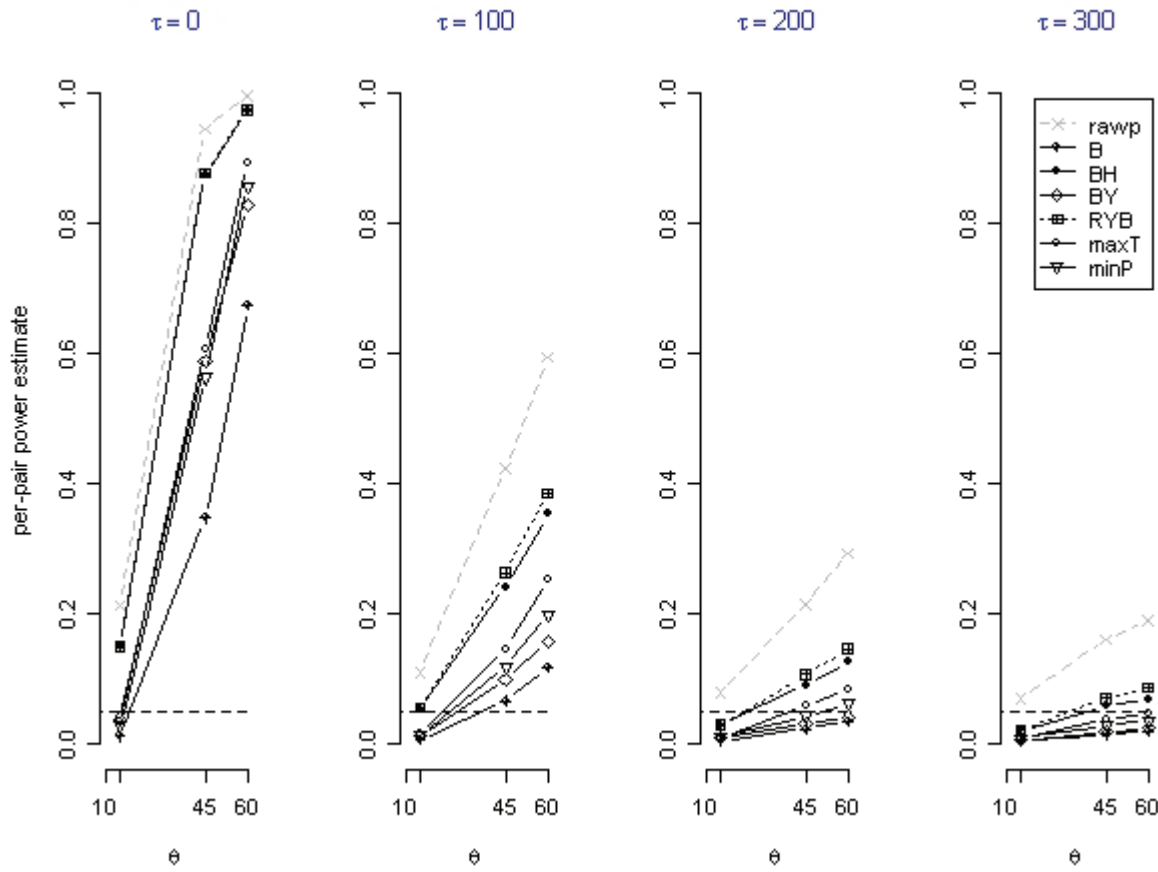


Figure 7. Per-pair power estimates for the different p-value adjustment methods as a function of the true difference between means (θ) for $n = 12$. Each graph refers to a different τ value (0 to 300, from left to right). Conventions as in Figure 3.

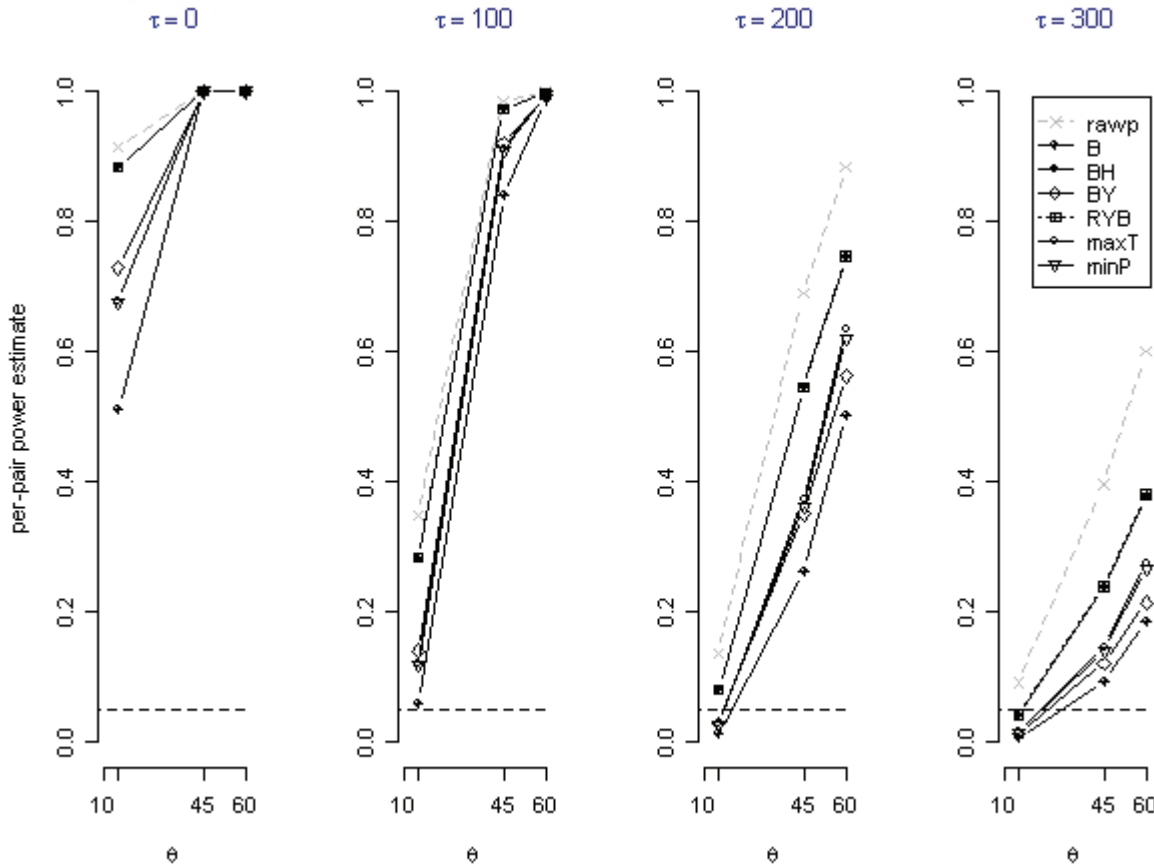


FIGURE 8 Per-pair power estimates for the different p-value adjustment methods as a function of the true difference between means (θ) for $n = 80$. Each graph refers to a different τ value (0 to 300, from left to right). Conventions as in Figure 3.

In more detail, several observations can be made related to the different controlling methods. Within the class of FWER controlling methods, as illustrated in Table 3, minP and maxT showed a good Type I error control only when $\tau = 0$. For $\tau > 0$, Type I error was not controlled anymore, although it can be observed that performance in this regard increased as n increased. On the other side, minP and maxT showed a clearly higher performance in terms of power, for small sample sizes, provided that $\tau \leq 100$ (see Figures 3, 5, and 7). With large sample sizes and $\tau = 0$, particularly when $\theta \geq 45$, minP, maxT and B showed overlapping power functions (see Figures 4, 6, and 8). In light of these arguments, minP and maxT may be

preferred in the former scenario, whereas B is certainly to be preferred in the latter scenario. Notably, these results hold for all the different power types. Within the class of FDR controlling methods, Table 3 inspection highlights that all methods showed a good Type I error control when $\tau = 0$. Surprisingly, some sort of linear relation seems to characterise Type I error control as a function of n and τ . In particular, when $n = 20$, all methods controlled Type I error for $\tau \leq 100$. When $n = 40$, Type I error control was extended to $\tau = 200$, and when $n = 80$, a good Type I error control was observed even for $\tau = 300$. In terms of power, BH and RYB consistently showed a better performance than BY, across all conditions and power types (see Figures 3-8). Only for $n = 12$ and $\tau \geq 100$, RYB behaved slightly better than BH,

MULTIPLE TESTING IN REACTION TIME DATA

independently of power type. In all the other conditions, the BH method is recommended, because of its quick and easy implementation (Thissen et al., 2002).

When comparing methods controlling a different kind of Type I error, several observations can be made. First, with $\tau = 0$, B should be preferred over FDR-based methods when $\theta \geq 45$ and $n = 80$. In fact, given that they show overlapping power estimates, it may seem more reasonable to choose the method providing the strongest Type I error control. By contrast, when BH and RYB show a clear power advantage over B (e.g., for $n = 12$ and $\tau = 0$), it may be more appropriate choosing either of these FDR-based control methods.

In general, the RT exponential component produced a conspicuous loss of power, especially when sample size was small. For $\tau = 300$, no method among those included in the present study showed power estimates higher than .4, even when the real distance among means was 60 and $n = 80$. Consequently, the results suggest that performing multiple comparisons with RT data is less than ideal when the data distribution is characterised by a strong exponential component. In light of the good performance with distributions in which $\tau = 0$, operating an RT decomposition technique such as put forward by Heathcote (1996) is strongly recommended. In fact, after performing the RT decomposition, the different adjustment methods appeared adequately powerful even with small sample sizes.

References

Andrews, S., & Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Learning, Memory and Cognition*, 27, 514-544.

Armstrong, I. T., & Munoz, D. P. (2003). Inhibitory control of eye movements during oculomotor countermanding in adults with attention-deficit hyperactivity disorder. *Experimental Brain Research*, 152, 444-452.

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, 128, 32-55.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the False Discovery Rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165-1188.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3-62.

Burbeck, S. L., & Luce, R. D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception and Psychophysics*, 32, 117-133.

Campbell, J. I. D., & Penner-Wilger, M. (2006). Calculation latency: The μ of memory and the τ of transformation. *Memory and Cognition*, 34, 217-226.

Dell'Acqua, R., Job, R., & Grainger, J. (2001). Is global shape sufficient for automatic object identification? *Visual Cognition*, 8, 801-821.

Dudoit, S., Popper Shaffer, J., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 1, 71-103.

Ge, Y., Dudoit, S., & Speed, T. (2003). Resampling-based multiple testing for microarray data analysis. Technical Report #633 of UCB Stat}. <http://www.stat.berkeley.edu/~gyc>.

Heathcote, A. (1996). RTSYS: A DOS application for the analysis of reaction time data. *Behavior Research Methods, Instruments and Computers*, 28, 427-445.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, 109, 340-347.

- Hohle, R.H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, *69*, 382-386.
- Horn, M., & Dunnett, C. W. (2004). Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. In Y. Benjamini, F. Bretz, & S. Sarkar (Eds.), *Recent developments in multiple comparison procedures*. Lecture Notes – Monograph Series. Vol. 47 (pp. 48-64). Beachwood, OH: Institute of Mathematical Statistics.
- Jöreskog, K. G., & Sörbom, D. (1996). *PRELIS 2: User's reference guide*. Scientific Software, Chicago, IL.
- Keselman, H. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to Familywise and Comparisonwise Type I error control. *Psychological Methods*, *4*, 58-69.
- Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*, 138-147.
- Leth-Steensen, C., Elbaz, Z. K., & Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of ADHD children: a response time distributional approach. *Acta Psychologica*, *104*, 167-190.
- Lu, C.-H., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychonomic Bulletin and Review*, *2*, 174-207.
- Madden, D. J., Gottlob, L. R., Denny, L. L., Turkington, T. G., Provenzale, J. M., Hawk, T. C. & Coleman, R. E. (1999). Aging and recognition memory: Changes in regional cerebral blood flow associated with components of reaction time distributions. *Journal of Cognitive Neuroscience*, *11*, 511-520.
- Matsumoto, M., & Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, *8*, 3-30.
- McGill, W. J. (1963). Stochastic latency mechanisms. In R. D. Luce, R. R. Busch, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 309-360). New York: John Wiley and Sons.
- Mewhort, D. J. K., & Johns, E. E. (2000). The extralist-feature effect: Evidence against item matching in short-term recognition memory. *Journal of Experimental Psychology: General*, *129*, 262-284.
- Mewhort, D. J. K., Braun, J. G., & Heathcote, A. (1992). Response time distributions and the Stroop task: A test of the Cohen, Dunbar, and McClelland (1990) model. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 872-882.
- Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., (p. 264-336).
- Penner-Wilger, M., Leth-Steensen, C., & LeFevre, J.-A. (2002). Decomposing the problem-size effect: A comparison of response time distributions across cultures. *Memory and Cognition*, *30*, 1160-1167.
- R Development Core Team (2003). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108.
- Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*, 446-461.
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510-532.
- Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*, 190-214.
- Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, *363*, 479-485.
- Reiner, A., Yekutieli, D., & Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, *19*, 368-375.
- Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments and Computers*, *33*, 457-469.

MULTIPLE TESTING IN REACTION TIME DATA

Spieler, D. H., Balota, D. A., & Faust, M. E. (2000). Levels of selective attention revealed through analyses of response time distributions. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 506-526.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of Benjamini-Hochberg procedure for controlling the False Discovery Rate in multiple comparison. *Journal of Educational and Behavioral Statistics*, 27, 77-83.

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *Quarterly Journal of Experimental Psychology*, 47A, 631-650.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin and Review*, 7, 424-465.

Van Zandt, T. (2002). Analysis of response time distributions. In J. T. Wixted (Vol. Ed.) & H. Pashler (Series Ed.) Stevens' *Handbook of Experimental Psychology* (3rd Edition), Volume 4: Methodology in Experimental Psychology (pp. 461-516). New York: Wiley Press.

West, R., Murphy, K. J., Armilio, M. L., Craik, F. I. M., & Stuss, D. T. (2002). Lapses of intention and performance variability reveal age-related increases in fluctuations of executive control. *Brain and Cognition*, 49, 402-419.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing*, Wiley, NY.

Wilcox, R. (2003). *Applying contemporary statistical techniques*. San Diego, Academic Press.

Wilcox, R. (2005). *Introduction to robust estimation and hypothesis testing*. Academic Press, MA.

Yekutieli, D., & Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82, 171-196.