Gaining power in multiple testing of interval hypotheses via conditionalization

Ellis, J.L.; Pecanka, J.; Goeman, J.J.

Biostatistics (2020) **21**, 2, *pp.* e65–e79
doi:10.1093/biostatistics/kxy042
Advance Access publication on September 21, 2018

Gaining power in multiple testing of interval hypotheses via conditionalization

JULES L. ELLIS*,†

Behavioral Science Institute, Radboud University Nijmegen, Postbus 9104, 6500 HE, Nijmegen, The Netherlands
j.ellis@psych.ru.nl

JAKUB PECANKA‡, JELLE J. GOEMAN‡

Biomedical Data Sciences, Leiden University Medical Center, Postbus 9600, 2300 RC, Leiden, The Netherlands

Summary

In this article, we introduce a novel procedure for improving power of multiple testing procedures (MTPs) of interval hypotheses. When testing interval hypotheses the null hypothesis *P*-values tend to be stochastically larger than standard uniform if the true parameter is in the interior of the null hypothesis. The new procedure starts with a set of *P*-values and discards those with values above a certain pre-selected threshold, while the rest are corrected (scaled-up) by the value of the threshold. Subsequently, a chosen family-wise error rate (FWER) or false discovery rate MTP is applied to the set of corrected *P*-values only. We prove the general validity of this procedure under independence of *P*-values, and for the special case of the Bonferroni method, we formulate several sufficient conditions for the control of the FWER. It is demonstrated that this "filtering" of *P*-values can yield considerable gains of power.

Keywords: Conditionalized test; False discovery rate; Family-wise error rate; Multiple testing; One-sided tests; Uniform conditional stochastic order.

1. Introduction

Data of questionnaires, psychological tests, and exams usually consist of scores on various items (questions). The relationships between items are often analyzed with monotone latent variable models (Holland and Rosenbaum, 1986), such as a linear factor analysis models or item response theory models (e.g. Rasch, 1960). Several authors have advocated models with only general assumptions, such as monotonicity, known as non-parametric item response theory (NIRT) (Sijtsma and Junker, 2006).

Checking the validity of NIRT models usually involves a large number of one-sided hypotheses tests. For example, as shown by Holland and Rosenbaum (1986), unidimensional monotone latent variable models imply that each pair of items has nonnegative covariances in each group defined by the other

*To whom correspondence should be addressed.
†Montessorilaan 3 6525 HR Nijmegen The Netherlands.
‡Einthovenweg 20 2333 ZC Leiden The Netherlands.

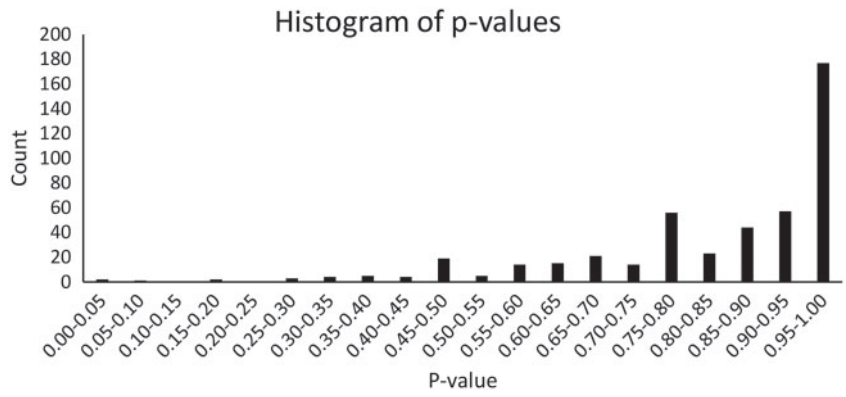

Fig. 1. *P*-values in a test of nonnegative covariances of subtest B of Raven's progressive matrices.

items. In NIRT scaling software (Van der Ark, 2007, 2012) based on the method of Mokken (1971), it is routinely tested whether the item covariances are nonnegative, whether the item-restscore regressions are monotonically increasing (Junker, 1993), and whether the order of item means is consistent across score groups defined by other items. For binary items, Ellis (2014) furthermore derived the prediction that the partial correlations are nonnegative. The multiple testing problem in this context can easily be very large since all possible pairs of items will be tested. Mokken (1971) and Rosenbaum (1984) suggest the Bonferroni correction but that risks a substantial loss of power.

We argue that multiple testing procedures (MTPs) can take advantage of the fact that many *P*-values are inflated due to the fact that the hypotheses are one-sided, or generally when interval hypotheses are tested. As an example, Figure 1 shows a histogram of the *P*-values for the covariances of the 12 items of subtest B of Raven's Progressive Matrices test, obtained in the data discussed by Van der Ven and Ellis (2000). For each pair, the null hypothesis that the covariance is nonnegative was tested in each score group of the 36 items of subtests A, C, and D. As can be seen, the *P*-values are not uniformly distributed, but rather are inflated in the sense that they are concentrated near 1, because the covariances are actually positive in many cells. This is to be expected when a unidimensional monotone latent variable model holds for most items.

In this article, we propose a procedure, which improves existing MTPs in the presence of inflated *P*-values by adding a simple conditionalization step at the onset of the analysis. For an a priori selected threshold $\lambda \in (0, 1]$ (e.g. $\lambda = 0.5$) we remove (i.e. do not reject) all hypotheses with *P*-value above λ. The remaining *P*-values are scaled up as $P'_i = P_i/\lambda$. The selected MTP is subsequently performed on the rescaled P'-values only. We refer to the altered procedure as the conditionalized version of the MTP, leading to procedures such as the conditionalized Bonferroni procedure (CBP). Applied to the above example and the Bonferroni procedure with $\lambda = 0.5$, this procedure means the *P*-values are multiplied by 2, but that only the *P*-values below 0.5 are counted. In this example, there were 466 *P*-values, with mean *P*-value of 0.836, and only 40 *P*-values were less than or equal to 0.5. Consequently, the Bonferroni correction would multiply each *P*-value by 466, while the CBP entails that each *P*-value is multiplied by 40/0.5 = 80. Thus the CBP may have greater power than the ordinary Bonferroni procedure. Moreover, this conditionalization step can also be used in combination with other MTPs.

We emphasize here that our approach specifically targets the situation that interval null hypotheses are tested, and this causes *P*-values to be potentially stochastically larger than uniform. If *P*-values are stochastically larger than uniform for other reasons, we do not recommend our method. For example, *P*-values stochastically larger than uniform may occur when using non-randomized tests with discrete test statistics. In this case, power may be gained with procedures that take the discrete nature into account, or

by filtering approaches (Ignatiadis and others, 2016; Zhu and Guo, 2017). In omics data, similar situations to Figure 1 may arise that the histogram of observed P-values is skewed with a peak near 1 even when point null hypotheses are tested. Such skewness may arise due to correlations between P-values even when marginally no P-values are stochastically larger than uniform. In such situations, power may be gained by incorporating covariates into the model to reduce the correlations between P-values (Listgarten and others, 2010; Risso and others, 2014).

Although we developed the procedure initially in the context of NIRT, it can be useful in other settings with interval hypotheses. For example, in large-scale survey evaluations of public organizations, such as schools or health care organizations, it can be interesting to test whether organizations score lower than a benchmark (Normand and Shahian, 2007; Ellis, 2013). If many organizations score well above the benchmark, which is usually the case in developed countries, then the P-values of true null hypotheses become inflated.

In the remainder of the article, we formally investigate the effects and benefits of conditionalization. We prove that for scenarios with inflated P-values conditionalized procedures retain type I error control whenever P-values are independent. Our result applies to both the family-wise error rate (FWER), the false discovery rate (FDR), and other error rates. We also show that if P-values are not independent, such control is not automatically guaranteed. However, we conjecture that the CBP is generally valid for positively correlated P-values. To support this conjecture, we formulate and prove several sufficient conditions for the control of the FWER by the CBP. We use simulations to investigate the power and the robustness of conditionalized procedures. Finally, we return to our motivating example of testing the validity of NIRT models, showing substantial power gain for the new procedure.

2. Definition of conditionalized tests

We define a MTP $\mathcal{P}$ as a mapping that transforms any finite vector of P-values into an equally long vector of binary decisions. If $\mathcal{P}(P_1, \ldots, P_m) = (d_1, \ldots, d_m)$, then d_i indicates whether the null hypothesis corresponding to P_i is rejected ($d_i = 1$) or not ($d_i = 0$). We define a decision rate as the expected value of a function of $\mathcal{P}(P_1, \ldots, P_m)$.

For $\lambda \in (0, 1]$ and an MTP $\mathcal{P}$ we define the corresponding conditionalized MTP $\mathcal{P}^\lambda$ as the MTP that, on input of a vector of P-values $(P_1, \ldots, P_m)$, applies $\mathcal{P}$ to the sub-vector consisting of only the rescaled P-values P_i/λ with $P_i \leq \lambda$, and that does not reject the null hypotheses of the P-values with $P_i > \lambda$. Throughout the paper we always assume that both the level of significance α and the conditionalization factor λ are fixed (independently of the data) prior to the analysis.

In this article, we pay special attention to the CBP and its control of the FWER. For $\lambda \in (0, 1]$ define $R_m(\lambda) = \sum_{i=1}^m \mathbf{1}\{P_i \leq \lambda\}$. Let $\mathcal{T} \subseteq \{1, \ldots, m\}$ be the index set of true null hypotheses. The FWER of the CBP is defined as

$$\mathrm{FWER}_{\mathrm{CB}}^{\lambda,\alpha} = P\Big(\bigcup_{i \in \mathcal{T}} \Big[P_i < \frac{\alpha\lambda}{R_m(\lambda) \vee 1}\Big]\Big).$$

If $\mathrm{FWER}_{\mathrm{CB}}^{\lambda,\alpha} \leq \alpha$ for given λ and α we say that the CBP controls the FWER for those λ and α.

Note that for the sake of simplicity in the rest of the article, we sometimes suppress one or both arguments and simply use R_m, $R(\lambda)$, or even R in the place of $R_m(\lambda)$. The proofs of all theorems and lemmas formulated below can be found in the supplementary material available at *Biostatistics* online.

We note that application of conditionalized procedures is extremely simple. For example, applying conditionalized Benjamini-Hochberg in R on a named vector of P-values `P` with $\lambda = 0.5$ requires simply

```
p.adjust(P[ P < 0.5 ] / 0.5, method="BH" )
```

3. FWER AND FDR OF INDEPENDENT TESTS

In this section, we state our main result: a conditionalized procedure controls FWER (or FDR) if the non-conditionalized procedure controls FWER (or FDR) and if the test statistics are independent and the marginal distributions satisfy a condition that we call supra-uniformity.

DEFINITION 3.1 (supra-uniformity) The distribution of P_i is supra-uniform if for all $\lambda, \gamma \in [0,1]$ with $\gamma \leq \lambda$ it holds $P(P_i < \gamma \mid P_i \leq \lambda) \leq \gamma/\lambda$. We say that P_i is supra-uniform if its distribution is supra-uniform.

Supra-uniformity is also known as the uniform conditional stochastic order (UCSO, defined by Whitt, 1980, 1982; Keilson and Sumita, 1982; Rüschendorf, 1991) relative to the standard uniform distribution $U(0,1)$. It is well-known that this condition is implied if P_i dominates $U(0,1)$ in likelihood ratio order (e.g. Whitt, 1980; Denuit and others, 2005). Whitt (1980) shows that when the sample space is a subset of the real line and the probability measures have densities, then UCSO is equivalent to the monotone likelihood ratio (MLR) property (i.e. for every $y > x$ it holds $f(y)/g(y) \geq f(x)/g(x)$). In the case of $U(0,1)$ (i.e. when g is a constant) it is immediately clear that MLR is equivalent to the P-values having densities that are increasing on $(0,1)$, which is further equivalent to having cumulative distribution functions that are convex on $(0,1)$.

THEOREM 3.2 Let $\mathcal{P}$ be an MTP and D be a decision rate (e.g. FWER or FDR) such that $D_{\mathcal{P}} \leq \alpha$ for $\alpha \in (0,1)$ whenever the P-values of the true hypotheses are independent and supra-uniformly distributed. Then, for the conditionalized MTP $\mathcal{P}^{\lambda}$ it holds that $D_{\mathcal{P}^{\lambda}} \leq \alpha$.

The proof of Theorem 3.2 can be found in the supplementary material available at *Biostatistics* online. The basic idea behind the proof is to divide the space of P-values into orthants partitioned by the events $[P_i \leq \lambda]$ versus $[P_i > \lambda]$ for all i. Conditionally on each of these orthants, the FWER (or FDR) of $\mathcal{P}^{\lambda}$ is at most α. Therefore, the total FWER (or FDR) of $\mathcal{P}^{\lambda}$ must also be at most α. A similar argument is used by Wollan and Dykstra (1986) in the context of order restricted inference.

Many popular MTPs satisfy the conditions of Theorem 3.2, since they only require a weaker condition $\mathrm{P}(p_i \leq c) \leq c$ in order to preserve type I error control. Consequently, for independent P-values the validity of the conditionalized versions of the methods by Holm (1979), Hommel (1988), Hochberg (1988) for FWER control and by Benjamini and Hochberg (1995) for FDR control follows by Theorem 3.2.

4. FWER INVESTIGATION: THEORY

Generalizing Theorem 3.2 to the setting with dependent P-values is not trivial. In the next sections of this article, we focus on the specific case of CBP, which is relevant for the motivating NIRT example. We present a clear case of lack of control with (extreme) negative correlations. Although we do not have a formal proof, we conjecture that conditionalized Bonferroni is valid when P-values are generally positively correlated, at least in the case of multivariate normally distributed test statistics.

We have found several sufficient conditions for the control of the FWER by the CBP, all of which fit in the overarching theme of positive correlations between P-values. Since these results are technical and specific we give them in the supplementary material available at *Biostatistics* online. Further justification for our conjecture is the extreme case when the P-values under the null are all identical at which point the proof of the control of the FWER by the CBP is trivial.

4.1. *Negative correlations: a counterexample*

To see that the requirement of independence of P-values in Theorem 3.2 cannot simply be dropped, consider a multiple testing problem with $m = 2$ where P_1 and P_2 both have a $U(0, 1)$ and $P_1 = 1 - P_2$. Assume $\lambda > 1/2$, since otherwise CBP is uniformly less powerful than the classical Bonferroni method. In this setting

$$\mathrm{FWER}_{\mathrm{CBP}}^{\lambda,\alpha} = \mathrm{P}(P_1 \leq \lambda\alpha) + \mathrm{P}(P_2 \leq \lambda\alpha) = 2\lambda\alpha > \alpha.$$

In other words, under the considered setting the CBP either fails to control FWER (with $\lambda > 1/2$) or is less powerful than the Bonferroni method (with $\lambda \leq 1/2$).

4.2. *The bivariate normal case*

Proposition 1 below guarantees FWER control by the CBP for all $\alpha, \lambda \in (0, 1)$ in the setting with two P-values corresponding to two bivariate zero-mean normally distributed test statistics with positive correlation. Denote as Φ the standard normal distribution function and as I_2 the 2×2 identity matrix.

Proposition 1 Let $m = 2$ and let $(X_1, X_2)' \sim N(0, \Sigma_\rho)$, where $\Sigma_\rho = \rho + (1 - \rho)I_2$. Set $P_1 = 1 - \Phi(X_1)$ and $P_2 = 1 - \Phi(X_2)$. If $\rho \geq 0$, then $\mathrm{FWER}_{\mathrm{CBP}}^{\lambda,\alpha} \leq \alpha$.

4.3. *The equicorrelated normal case*

In the supplementary material available at *Biostatistics* online, we study the case of P-values based on equicorrelated standard normal test statistics. We show that the CBP controls the FWER in this case if a certain unidimensional integral does not exceed λ^{-1}. Numerical evaluations suggest this to be the case whenever $\alpha \leq 0.368$.

4.4. *Large testing problems*

Finally, we give a sufficient conditions for FWER control by the CBP as the number of hypotheses m approaches infinity. Suppose for a moment that the expectation of $R(\lambda)$ (i.e. the number of P-values below λ) is known. In such case one could use the alternative to CBP that rejects hypothesis H_i whenever $P_i \leq \lambda\alpha/\mathrm{E}[R(\lambda)]$. If the P-values are supra-uniform then under arbitrary dependence this procedure controls FWER, since it holds

$$\begin{aligned}
\mathrm{FWER}_{\mathrm{CBP}'} &\leq \sum\nolimits_{i \in T} P(P_i < \alpha\lambda/E[R(\lambda)]) \\
&\leq \sum\nolimits_{i \in T} P(P_i \leq \lambda)\, P(P_i < \alpha\lambda/E[R(\lambda)] \mid P_i \leq \lambda) \\
&\leq \sum\nolimits_{i \in T} P(P_i \leq \lambda)\, \alpha/E[R(\lambda)] \leq \alpha.
\end{aligned}$$

This suggests that the CBP should also control FWER for $m \to \infty$ whenever $R(\lambda)$ is a consistent estimator of $\mathrm{E}[R(\lambda)]$. This heuristic argument is formalized in Proposition 2, where plim denotes convergence in probability.

Proposition 2 Let the P-values $P_1, \ldots, P_m$ have supra-uniform distributions and let $\mathrm{plim}_{m\to\infty} R/m = \eta$ and $\lim_{m\to\infty} \mathrm{E}(R/m) = \eta$ for some $\eta \in \mathbb{R}$. Then $\limsup_{m\to\infty} \mathrm{FWER}_{\mathrm{CBP}}^{\lambda,\alpha} \leq \alpha$.

An application of Proposition 2 in a situation where correlations between P-values vanish with $m \to \infty$ leads to Corollary 4.1.

COROLLARY 4.1 Denote $\rho_{ij} = \text{cor}(\mathbf{1}[P_i \leq \lambda], \mathbf{1}[P_j \leq \lambda])$ and put $\rho_{ij}^+ = \max\{0, \rho_{ij}\}$. Denote the average off-diagonal positive part of the correlations as

$$\bar{\rho}(m) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij}^+.$$

If the P-values are supra-uniform and $\lim_{m\to\infty} \text{E}(R/m) = \eta$ for some $\eta \in \mathbb{R}$ and $\lim_{m\to\infty} \bar{\rho}(m) = 0$, then $\limsup_{m\to\infty} \text{FWER}_{\text{CBP}}^{\lambda,\alpha} \leq \alpha$.

Corollary 4.1 will be important for the application of Section 8.2.

5. FWER INVESTIGATION: SIMULATIONS

5.1. *Models used in simulations*

Our conjecture is that the CBP controls the FWER in the case of positively correlated multivariate normal test statistics. To substantiate this, we conducted the following simulation studies, with positive correlations (models P1–P3), both positive and negative correlations (models N1 and N2), and negative correlations (model N3). In each study, we used a range of values of m between 2 and 100 000, but in some models the values over 1000 were omitted, because they were too time consuming. For each m we generated 100 correlation matrices $\Sigma = (\sigma_{ij})$, and for each Σ we conducted 10 000 simulations and computed the FWER for the CBP with $\alpha = 0.05$ and $\lambda \in \{0.1, 0.2, \ldots, 0.9\}$. Per Σ we report the maximum of the FWER across all chosen levels of λ, in order to check the worst case of FWER control over the entire range of λ. Each Σ was generated as $\Sigma = AA^T + L^2$, where the a_{ij} were drawn randomly and $L = (I_m - \text{diag}(AA^T))^{1/2}$. The generation of A will be described below. In each simulation, we generated first vectors $Z = (Z_1, \ldots, Z_m)' \sim N(0, I_m)$ and $E = (E_1, \ldots, E_m)' \sim N(0, I_m)$. Next, we obtained the "test statistics" $X = (X_1, \ldots, X_m)'$ as $X = AZ + LE$, and the P-values as $P_i = \Phi(X_i)$.

(P1) We used $a_{ij} = h_i * |b_{ij}| / \sqrt{\sum_j b_{ij}^2}$ with $b_{ij} \sim N(0,1)$ and $h_i^2 \sim U(0,1)$, independently drawn. For simulations with larger m we used a simpler model, because (P1) becomes very time consuming.

(P2) For each Σ we first obtained $a, b \sim U(0,1)$, and then a vector $A = (a_i)$ with $a_i \sim \text{Beta}(a, b)$. The test statistics were computed as $X_i = a_i X_{i-1} + \sqrt{1 - a_i^2} Z_i$ and $P_i = \Phi(X_i)$. Thus, each X_i has random correlation a_i with its precursor.

(P3) For each Σ, we obtained a random value $r \sim U(0,1)$, and computed $X_i = \sqrt{r} Z_1 + \sqrt{1-r} E_i$ and $P_i = \Phi(X_i)$. Thus, all X_i have the same correlation r with each other.

(N1) We used $a_{ij} = h_i * c_{ij} / \sqrt{\sum_j c_{ij}^2}$, where $c_{ij} = b_{ij} - \sum_j b_{ij}/m$, with $b_{ij} \sim N(0,1)$ and $h_i^2 \sim U(0,1)$, independently drawn.

(N2) For each Σ, we first obtained $a, b \sim U(0,1)$, and then a vector $A = (a_i)$ with $a_i \sim \text{Beta}(a, b)$. The test statistics were computed as $X_i = -a_i X_{i-1} + \sqrt{1 - a_i^2} Z_i$ and $P_i = \Phi(X_i)$. Thus, each X_i has random correlation $-a_i$ with its precursor.

(N3) The test statistics were computed as $X_i = (Z_i - \sum Z_i/m) / \sqrt{1 - 1/m}$ and $P_i = \Phi(X_i)$. This creates variables X_i with correlations $-1/(m-1)$, which is equal to the lowest possible average correlation of m variables.

Table 1. *Simulation results for models P1 through N3: mean FWER (standard deviation). Here, m is the number of hypotheses. These results were obtained with 100 correlation matrices per value of m, and 10 000 replications per correlation matrix.*

m	P1	P2	P3	N1	N2	N3
2	0.046 (0.003)	0.041 (0.009)	0.043 (0.006)	0.052 (0.010)	0.069 (0.018)	0.090 (0.003)
3	0.045 (0.004)	0.039 (0.011)	0.041 (0.007)	0.052 (0.006)	0.061 (0.007)	0.074 (0.003)
4	0.044 (0.003)	0.039 (0.010)	0.039 (0.009)	0.051 (0.004)	0.055 (0.005)	0.068 (0.002)
5	0.043 (0.004)	0.037 (0.010)	0.039 (0.009)	0.051 (0.004)	0.054 (0.005)	0.065 (0.002)
6	0.043 (0.003)	0.038 (0.011)	0.037 (0.010)	0.051 (0.003)	0.052 (0.006)	0.062 (0.002)
7	0.043 (0.003)	0.035 (0.012)	0.035 (0.010)	0.050 (0.003)	0.051 (0.006)	0.061 (0.002)
8	0.043 (0.003)	0.035 (0.011)	0.034 (0.012)	0.050 (0.002)	0.050 (0.006)	0.060 (0.003)
9	0.042 (0.003)	0.037 (0.009)	0.035 (0.011)	0.050 (0.002)	0.048 (0.007)	0.059 (0.002)
10	0.042 (0.003)	0.036 (0.012)	0.031 (0.013)	0.050 (0.002)	0.048 (0.006)	0.059 (0.002)
15	0.041 (0.003)	0.034 (0.011)	0.035 (0.013)	0.049 (0.002)	0.046 (0.009)	0.056 (0.002)
20	0.040 (0.003)	0.035 (0.013)	0.034 (0.013)	0.049 (0.002)	0.047 (0.006)	0.054 (0.002)
25	0.039 (0.003)	0.033 (0.012)	0.027 (0.013)	0.049 (0.002)	0.044 (0.010)	0.054 (0.002)
50	0.038 (0.002)	0.034 (0.012)	0.027 (0.015)	0.049 (0.002)	0.045 (0.010)	0.052 (0.002)
75	0.037 (0.002)	0.036 (0.011)	0.027 (0.016)	0.049 (0.002)	0.042 (0.011)	0.051 (0.002)
100	0.036 (0.002)	0.035 (0.011)	0.027 (0.015)	0.049 (0.002)	0.042 (0.012)	0.051 (0.002)
1000	0.031 (0.002)	0.036 (0.013)	0.017 (0.014)	0.049 (0.002)	0.044 (0.008)	0.049 (0.002)
10 000		0.036 (0.012)	0.018 (0.016)		0.042 (0.012)	0.049 (0.002)
100 000		0.037 (0.011)	0.014 (0.016)		0.044 (0.008)	0.049 (0.002)

5.2. *Simulation results per model*

For model (P1) the mean off-diagonal correlation per Σ ranged from 0.03 to 0.90 with mean 0.30. For model (N1) the mean off-diagonal correlation per Σ ranged from -0.91 to 0.91 with mean 0.00. The minimum correlation per Σ ranged from -0.96 to 0.91 with mean -0.32.

The averages and standard deviations of the FWER are shown in Table 1. In the models with positive correlations, the average FWER was well below the nominal α in all cells of Table 1. Additionally, the individual FWERs are displayed Figures 2 and 3 of the supplementary material available at *Biostatistics* online. We conclude that the FWER is under control in these models.

The models with negative correlations showed that the FWER is not necessarily controlled with negative correlations, especially for small m. In model N1, significant breaks of the FWER at 0.01 level occurred in 104 cases, but only in cases with negative minimum correlations and mostly with $m \leq 10$. In model N2, this occurred in 319 cases, again mostly with small m. In model N3, because the correlations depend only on m, we can consider the 100 correlation matrices with 10 000 replications each as one correlation matrix with 1 000 000 replications, resulting in a single FWER per value of m. All these FWERs were significantly above α if $\alpha = 0.05$ and $m \leq 100$, but dropped below α for $m \geq 1000$. For the cases with $m \geq 20$, the upperbound of the 99% confidence interval of the FWER with $\alpha = 0.05$ was less than 0.055. In sum, negative correlations can lead to a FWER above α, but even in the worst case that we considered, model N3, we found for $\alpha = 0.05$ that a FWER close to the asymptotic control implied by Corollary 4.1 was already obtained with $m \geq 20$ (implying correlations of -0.0417 or greater).

The simulation results support our conjecture that CBP controls FWER with positive, but not necessarily with negative correlations, and that lack of control tends to vanish when the size of the multiple testing problem increases.

6. POWER INVESTIGATION—ANALYTICAL

In this section, we consider how the power depends on the choice of λ and other parameters in a simple situation. Assume that $Z_1, \ldots, Z_m$ are independent standard normal variables, and that for some $\delta, \epsilon \geq 0$, the P-values are given by $P_i = \Phi(Z_i + \delta)$ if H_i is true, and $P_i = \Phi(Z_i - \epsilon)$ if H_i is false. Let $\mathcal{T} \subseteq \{1, \ldots, m\}$ be the index set of true null hypotheses, and define $S_0 = \sum_{i \in \mathcal{T}} \mathbf{1}\{P_i \leq \lambda\}$ and $S_1 = R - S_0$. Let π_0 be the proportion of true hypotheses and π_1 the proportion of false hypotheses, en denote $z_\lambda = \Phi^{-1}(\lambda)$. Now $R = S_0 + S_1$, where S_0 and S_1 are independent binomial variables with distributions $\text{Binom}(m\pi_0, \Phi(z_\lambda - \delta))$ and $\text{Binom}(m\pi_1, \Phi(z_\lambda + \epsilon))$, respectively. If we write $\xi = \pi_0 \Phi(z_\lambda - \delta) + \pi_1 \Phi(z_\lambda + \epsilon)$, then for large m, we have $R \approx m\xi$. For a single false hypothesis i, the power of the Bonferroni procedure is $P(P_i \leq \alpha/m)$, while the power of the CBP is approximately $P(P_i \leq \alpha\lambda/m\xi)$. We will proceed with this approximation, under which the power decreases with $\rho := \xi/\lambda$. The power of the CBP exceeds that of the Bonferroni procedure if $\rho < 1$.

We can make the following observations.

1. If $\delta = 0$ then $\rho = (\pi_0\lambda + \pi_1\Phi(z_\lambda + \epsilon))/\lambda \geq 1$. Thus, the power of the CBP will not exceed that of the BP if the true hypotheses have standard uniform P-values.
2. $\lim_{\delta \to \infty} \rho = \pi_1 \Phi(z_\lambda + \epsilon)/\lambda \leq \pi_1/\lambda$. Thus, if we take $\lambda > \pi_1$, then the power of the CBP exceeds that of the BP for large δ.
3. $\partial\rho/\partial\pi_0 = (\Phi(z_\lambda - \delta) - \Phi(z_\lambda + \epsilon))/\lambda \leq 0$. Consequently, the power increases with π_0.
4. If $\pi_0 \geq 0.5$, $\lambda \geq \pi_1$, and $\epsilon = \delta$, then $\rho \leq 1$. The proof is given in the supplementary material available at *Biostatistics* online.

Note that often we can increase δ and ϵ by increasing the sample size per hypothesis.

For a given set of $(\pi_0, \delta, \epsilon)$, we define λ_{min} as the value of λ for which $\rho(\pi_0, \delta, \epsilon, \lambda)$ is minimal. For any value of λ, we define the loss as $loss(\lambda) = \max_{\pi_0, \delta, \epsilon}\{\rho(\pi_0, \delta, \epsilon, \lambda)/\rho(\pi_0, \delta, \epsilon, \lambda_{min})\}$. We define the minimax solution of λ as the value of λ that minimizes $loss(\lambda)$. We computed this with a grid of $\pi_0 \in [0.5, 0.995]$ with steps of 0.005, $\delta \in [0.1, 10]$ and $\epsilon \in [0.1, 10]$ with steps of 0.1, and $\lambda \in [0.5, 0.999]$ with steps 0.001. The minimax solution of λ was 0.623 (with loss 1.60, attained with $\pi_0 = 0.995, \delta = 10, \epsilon = 8$, and resulting in $\rho = 0.00803$). This minimax solution is close to the "natural default value" of $\lambda = 0.5$.

Table 2 shows the values of ρ for $\lambda = 0.5$. Since ρ increases with ϵ, we displayed the worst case with $\epsilon \to \infty$ (which we approximated with $\epsilon = 50$). We see that the CBP could be beneficial if (1) the majority of hypotheses is true and (2) the noncentrality parameter of the true hypotheses is larger than 1 or larger than the absolute value of the noncentrality parameter of the false hypotheses.

Alternatively, one could prefer a relatively high value of $\lambda = 0.90$, because there are more situations in which its power exceeds that of the BP. A table with values of ρ for $\lambda = 0.90$ and $\epsilon = 2$ is given in the supplementary material available at *Biostatistics* online. Particularly, $\pi_0 \geq 0.85$ is sufficient to yield $\rho < 1$ for all studied values of δ in the table.

In Table 2, we have also computed the expected P-value of the true hypotheses. In the example of the introduction, the P-values had mean 0.835. The corresponding z-values, with $z = \Phi^{-1}(P)$, had mean 1.24 with standard deviation 1.10. Assuming these values as parameters for the true hypotheses, and assuming that $\pi_0 \geq 0.5$, similar computations show that it would be enough to pick $\lambda \geq 0.59$ in order to obtain $\rho < 1$.

7. POWER INVESTIGATION—SIMULATIONS

In this section, we investigate the power performance of conditionalized tests relative to their non-conditionalized versions through simulations in a broader range of MTPs. We consider the following procedures: Bonferroni; Šidák (attributed to Tippet by Davidov, 2011, p 2433); Fisher combination method

Table 2. *Values of* ρ *with* $\lambda = 0.5$ *and* $\epsilon \to \infty$

			δ			
	0.1	0.2	0.4	0.8	1.6	3.2
π_0	(0.528)	(0.556)	(0.611)	(0.714)	(0.871)	(0.988)
0.10	1.89	1.88	1.87	1.84	1.81	1.80
0.20	1.78	1.77	1.74	1.68	1.62	1.60
0.30	1.68	1.65	1.61	1.53	1.43	1.40
0.40	1.57	1.54	1.48	1.37	1.24	1.20
0.50	1.46	1.42	1.34	1.21	1.05	1.00
0.60	1.35	1.30	1.21	1.05	0.87	0.80
0.70	1.24	1.19	1.08	0.90	0.68	0.60
0.80	1.14	1.07	0.95	0.74	0.49	0.40
0.90	1.03	0.96	0.82	0.58	0.30	0.20

Values between parentheses are the expected P-values of the true hypotheses. Note that $\rho < 1$ implies that the CBP is more powerful than regular Bonferroni, and vice versa for $\rho > 1$. Since ρ increases with ϵ, more realistic values of ϵ are more favorable for CBP.

based on the transformation $F = -2\sum_{i=1}^{m} \log P_i$ (see Davidov, 2011, p 2433); the likelihood ratio (LR) procedure based on the theory of order restricted statistical inference of Robertson and others (1988), using the chi-bar distribution with binomial weights; the I_+ statistic, based on the empirical distribution function (Davidov, 2011, p 2433); the Bonferroni plug-in procedure as defined by Finner and Gontscharuk (2009) based on the work of Storey (2002) referred to as the FGS procedure; the Benjamini-Hochberg method of Benjamini and Hochberg (1995), referred to as the BH procedure.

7.1. *Power as the number of true hypotheses increases*

For the power investigation the P-values were generated based on m parallel z-tests of null hypotheses of type $H_0 : \mu_i \geq 0$, each based on a sample of size n. The P-values were calculated as $P_i = \Phi(X_i)$, with $\text{var}(X_i) = 1$ and noncentrality parameter $\text{E}(X_i) = \mu_i\sqrt{n}$. To each set of P-values, we applied the conditionalized and the ordinary versions of the considered testing procedures at the overall significance level of $\alpha = 0.05$. Conditionalizing was applied with $\lambda = 0.5$. A number of combinations in terms of noncentrality parameter, hypothesis count, and proportion of false hypotheses was considered. For each combination, we performed 10 000 replications.

Figure 2 shows the results of a simulation where the number of false hypotheses is fixed while the number of true hypotheses increases. The plot shows for most MTPs the minimal power, defined as the probability to reject at least one false hypothesis (Chen and others, 2011). Only for the BH procedure the plot shows the true positive rate (TPR), or average power, defined as the expectation of the number of true discoveries divided by the number of false hypotheses (Benjamini and Liu, 1999; Glueck and others, 2008). For the false hypotheses the value of the noncentrality parameter was set at -2, while for the true null hypotheses it was set at 2. The plot illustrates that the power decreases rapidly with the number of true hypotheses for most non-conditionalized procedures. The only exception to this is the LR procedure. In contrast, for all of the considered conditionalized procedures the power decreases much more slowly. This shows that, with the exception of the LR procedure, the conditionalization substantially improves the power performance of the considered procedures in this setting. The TPR of BH shows a similar improvement by conditionalization.

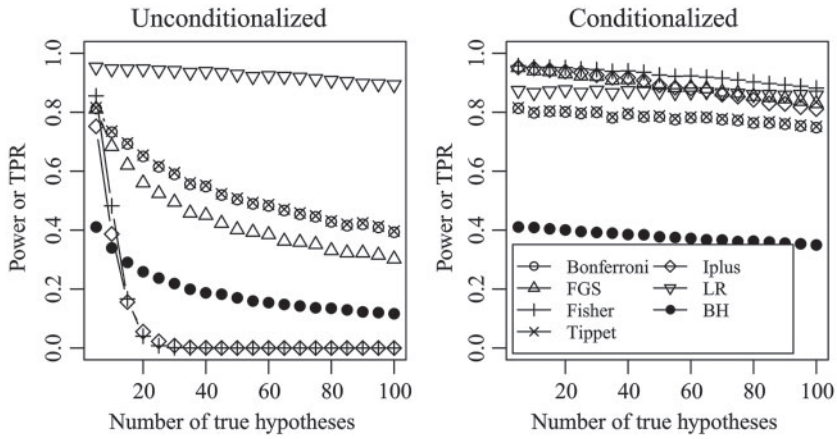

Fig. 2. Power as a function of the number of true hypothesis. The number of false hypotheses is fixed at five in all points. The number of true hypotheses increases from left to right. The considered noncentrality parameter is ±2. The plot displays the average power or TPR for BH, and the minimal power for the other MTPs.

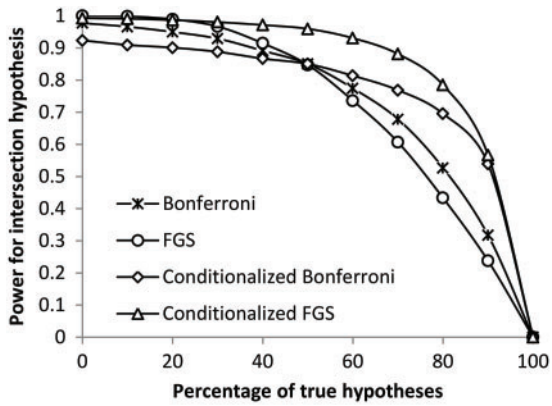

Fig. 3. Power as a function of the percentage of true hypothesis. The total number of hypotheses is fixed at 100 in all points. The percentage of true hypotheses increases from left to right. The considered noncentrality parameter is ±1.5.

Figure 3 illustrates the influence of conditionalization on the performance of the Bonferroni and FGS methods in a setting where the percentage of true hypotheses increases while the total number of hypotheses remains fixed. The figure shows that the conditionalized FGS procedure is the overall best performing procedure among the four. A more extensive simulation with up to 200 000 hypotheses, using the model of Section 6 is shown in Figure 4 of the supplementary material available at *Biostatistics* online.

7.2. *Power in pairwise comparisons of ordered means*

Consider a series of independent sample means $y_i \sim N(\mu_i, \sigma^2/n)$ with the compound hypothesis $H : \mu_1 \leq \mu_2 \leq \ldots \leq \mu_k$. An analysis method specifically designed for this setting is the isotonic regression (Robertson and others, 1988), although this method does not allow to deduce specifically which pairs (μ_i, μ_j) violate the ordering specified by the null hypothesis. Alternatively, the $k(k-1)/2$

individual hypotheses $H_{ij} : \mu_i \leq \mu_j$ with $i < j$ can be analyzed using one-sided t-tests, and the conditionalized Bonferroni or the conditionalized FGS procedures can be applied. The average correlation between the P-values vanishes as $m \to \infty$, thus the asymptotic control of the FWER by the CBP follows by Corollary 4.1. The simulations below indicate that the FWER is in fact controlled even for the small hypothesis counts.

The means in this simulation were modeled as $\mu_{i+1} = \mu_i + \delta$ for $i = 1, 2, \ldots, k-2$, and $\mu_k = \mu_1$. Thus, most means satisfy the ordering of the hypothesis, but the last mean violates it. We used $\sigma = 1$, $n = 10$ and set $\lambda = 0.5$. At $\delta = 0$, we observed that all four procedures exhibit FWER below α. For both $k = 5$ and $k = 20$, the two conditionalized procedures perform essentially as good or better than their non-conditionalized counterparts across the whole range of $\delta \in [0.1, 3]$. Note that the same results would be obtained with, for example, $n = 90$ and $\delta \in [1/30, 1]$. Figure 5 of the supplementary material available at *Biostatistics* online shows the results for $k = 20$ and $k = 5$ respectively.

8. Applications

8.1. *Example 1. Testing for nonnegative covariances in IRT*

We return to the example of Raven's Progressive Matrices Test discussed in the introduction. The unidimensionality of subtests B and E was disputed by Van der Ven and Ellis (2000), based on analyses with the Rasch model. A more general unidimensional latent variable model (Holland and Rosenbaum, 1986) may still hold, however. We analyzed the item covariances separately for subtests B and E, while subtests A, C, and D were combined them into a single test score per person. We conducted four analyses: (1) the covariances of the items of B, conditionally on the score of A+C+D; (2) the covariances of the items of E, conditionally on the score of A+C+D; (3) the covariances of each pair of items of B, conditionally on the sum score of the 10 other items of B; (4) the covariances of each pair of items of E, conditionally on the sum score of the 10 other items of E. If a unidimensional monotone latent variable model holds, then the tested covariances in (1)–(4) will be nonnegative and their sample estimates will have nonnegative correlations, and therefore the P-values will have nonnegative correlations. Thus, according to our conjecture, the CBP may be applied. The supplementary material available at *Biostatistics* online provides the unadjusted P-values that we will now analyze.

(1) Part of this example was already discussed in the introduction. The 466 uncorrected P-values are displayed in Figure 1. Furthermore, 40 P-values were less than or equal to 0.5. The five smallest P-values are 0.0370, 0.0468, 0.0928, 0.1648, and 0.2000. With Bonferroni correction and $\alpha = 0.05$, the critical value would be $\alpha/466 = 1.07 \times 10^{-4}$. Using the CBP with $\lambda = 0.5$, the critical value is $0.5\alpha/40 = 6.25 \times 10^{-4}$. With $\lambda = 0.90$ there are 232 P-values less than or equal to λ. Thus, after the CBP with $\lambda = 0.90$, the critical value is $0.90\alpha/232 = 1.94 \times 10^{-4}$. However, no P-value was smaller than any of these critical values.

(2) Similarly, for each item pair of subtest E, the null hypothesis that the covariance is nonnegative was tested in each score group of the 36 items of subtests A, C, and D. There were 711 P-values, of which 109 were less than or equal to 0.5. The three smallest P-values were 0.0320, 0.0343, and 0.0552. With Bonferroni correction and $\alpha = 0.05$, the critical value would be $\alpha/711 = 7.03 \times 10^{-5}$. With the CBP with $\lambda = 0.5$, the critical value would be $0.5\alpha/109 = 2.29 \times 10^{-4}$. With $\lambda = 0.90$ there are 379 P-values less than or equal to λ. Thus, after the CBP with $\lambda = 0.90$, the critical value is $0.90\alpha/379 = 1.19 \times 10^{-4}$.

(3) In this analysis, the smallest uncorrected P-value was 0.0873, making any multiple testing correction futile.

(4) There were 617 P-values, of which 137 were less than or equal to 0.5. There were six P-values below 0.05, namely 0.0044, 0.0050, 0.0284, 0.0290, 0.0329, and 0.0368. With Bonferroni correction, the

critical value would be $\alpha/617 = 8.10 \times 10^{-5}$. Using the CBP with $\lambda = 0.5$, the critical value would be $0.5\alpha/137 = 1.82 \times 10^{-4}$. With $\lambda = 0.90$ there are 402 P-values less than or equal to λ. Thus, after the CBP with $\lambda = 0.90$, the critical value is $0.90\alpha/402 = 1.12 \times 10^{-4}$.

In summary, these analyses make clear that the CBP was potentially more powerful than the Bonferroni correction, because the critical values are substantially smaller, but even with the CBP no hypotheses were rejected at 0.05 level.

8.2. *Example 2. Testing for manifest monotonicity in IRT*

In Mokken scale analysis, it is recommended to test manifest monotonicity (Van der Ark, 2007). With k items to be tested suppose that the variables $X_1, \ldots, X_k$ indicate correctness of response for the k items (with $X_i = 1/0$ indicating a correct/incorrect answer for the i-th item). Denote the *rest score* of the i-th item as $X_{-i} = (\sum_{j=1}^{k} X_j) - X_i$. A question of interest is whether $\pi_{ij} := P(X_i = 1 \mid X_{-i} = j)$ is a nondecreasing function of j within each item i. This leads to testing the $k(k-1)/2$ pairwise hypotheses $\pi_{ij'} \leq \pi_{ij}$ for $j' < j$ (Van der Ark, 2007). In this situation, although negative correlations between P-values cannot be excluded, we can apply 4.1 to motivate that the CBP is valid if the number of tests is large, as is the case in this application.

In the subtest E of the Raven Progressive Matrices test in the data set reported by Van der Ven and Ellis (2000) we obtained the following result with the `mokken` package of Van der Ark (2012) (the complete summary table is given in Table 2 of the supplementary material available at *Biostatistics* online). For item 11, there were 21 pairs of rest score groups that had to be compared—small adjacent groups were joined together by the program; there were five violations with a maximum z-statistic of 2.32, yielding an unmodified P-value $P = 0.010$. Now, if no multiplicity correction is performed the probability of false rejection for each item undesirably increases with the number of rest score groups. The classical Bonferroni correction yields the adjusted P-value of $P' = 0.010 \times 21 = 0.21$, while the CBP with $\lambda = 0.5$ yields the adjusted P-value of $P'' = 0.010 \times 5/0.5 = 0.10$. Thus, according to this analysis, manifest monotonicity of this item of subtest E remains undisputed.

8.3. *Example 3. Detecting substandard organizations in quality benchmarking*

Several countries have developed programs in which the quality of public organizations such as schools or hospitals is assessed. As stated by Ellis (2013), "such research can consist of large-scale studies where dozens [3], hundreds [4], or thousands [5, 6] of organizations are compared on one or more measures of performance or quality of care, on the basis of a sample of clients or patients from each organization." A goal of such programs is to identify under-performing organizations. For example, the Consumer Quality Index (CQI) program of the Netherlands developed questionnaires about the experiences of patients. In 2010, 85 natal care organizations were evaluated with eight performance indicators based on these questionnaires. Here, we consider only the four indicators based on multiple questionnaire items, for which a random effects model with normal errors was used. These indicators are labelled PI1, PI2, PI3, and PI4 here. These indicators were expressed on a scale from 1 to 4. A casemix-adjustement was based on age, education, and health. Now suppose that as a minimum standard, we require a mean rating of 3.5 after casemix-adjustment. For each combination of performance indicator and natal care organization we computed $P = \Phi((c - 3.5)/SE)$, where c and SE are the corrected mean and standard error of the organization on the performance indicator, after casemix-adjustment. A small P-value would signify that the organization performs substandard on the indicator. Since the organizations were independent, we can directly apply Theorem 3.2 to motivate the use of CBP in this application.

The supplementary material available at *Biostatistics* online provides the unadjusted P-values that we will now analyze. For PI1, PI2, and PI3, no organization had a P-value below 0.05. For PI4, there was one organization with a P-value below 0.05, and this P-value was 0.0002055. With the Bonferroni correction, since there are $4 \times 84 = 340$ tests conducted, the smallest corrected P-value would be 0.07. Applying the CBP with $\lambda = 0.5$, we found $R = 8$ P-values less than or equal to 0.5, so the smallest CBP-corrected P-value is $0.0002055 \times 8/0.5 = 0.0033$. Similarly, with $\lambda = 0.623$ ($R = 14$) and $\lambda = 0.90$ ($R = 31$) the CBP-corrected P-values (0.0046 and 0.0071) are below 0.05 level. The smallest P-values after BH-correction are the same (0.07, 0.0033, 0.0046, and 0.0071).

Thus, in this example the two conditionalized MTPs reveal a substandard organization, while the unconditionalized MTPs do not leave enough power. The advantage of using a conditionalized MTP in such settings is that the presence of organizations that score high above the minimum standard does not exacerbate the severity of the multiple testing problem and much of the power is preserved even with many high-performing hospitals included in the analysis.

9. Discussion

We have proposed a very simple and general method, called conditionalization, to gain power when testing interval null hypotheses. We suggest to discard all hypotheses with P-values above a pre-chosen constant λ (typically 0.5 or higher), and to divide the remaining P-values by λ before applying the MTP of choice. For independent P-values, we have proven that the conditionalized procedure controls the same error rate as the original procedure, provided null P-values are marginally supra-uniform (i.e. dominate the standard uniform distribution in likelihood ratio order). As a rule of thumb, conditionalized procedures can be expected to be more powerful than their ordinary, non-conditionalized counterparts if there are more true hypotheses with inflated P-values (i.e. with true parameter values deep inside the null hypothesis) than there are false null hypotheses. The power gain achieved by conditionalizing can be substantial, especially for adaptive procedures that incorporate an estimate of the proportion of true null hypotheses.

For the case of the conditionalized Bonferroni procedure (CBP) we conjecture that the CBP is valid when the P-values are positively correlated, while we have shown that it is not universally valid for negatively correlated variables. To support this conjecture we have given several sufficient conditions for FWER control by the CBP. We accompanied these results with an extensive simulation study and the results give supporting evidence for our conjecture. Nonetheless, a proof of our conjecture still eludes us and thus remains for future research. Also it is of interest to investigate further the option of choosing λ *post hoc*. For a priori chosen λ, $\lambda = 0.9$ is a safe chose that will often result in a power gain, while smaller values of λ can be more powerful if there are many null hypotheses with true values of the parameter that are deep inside the null hypothesis.

We emphasize that inflated P-values can arise for several reasons, and that the solution presented in this article is specific for the situation many P-values are marginally stochastically larger than uniform because interval hypotheses are tested. If instead inflated P-values arise due to discreteness of test statistics, the assumption of supra-uniformity is violated and our method cannot be used. Histograms of P-values may also show inflation due to correlations even when marginally no P-value is stochastically larger than uniform. In such situations, we do not expect a gain in power for our method.

We believe that this article makes a strong case for the usage of the conditionalized MTPs since they mitigate the loss of power typically associated with MTPs on inflated P-values and thus make it more attractive for researchers to formulate their scientific questions in terms of interval hypotheses. In light of the fact that shifting the focus towards interval hypotheses has been advocated as one of the solutions to get out of the current "P-value controversy" (Wellek, 2017) this likely makes conditionalization a very powerful method of analysis.

SUPPLEMENTARY MATERIALS

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

Jelle Goeman and Jakub Pecanka were supported by NWO VIDI grant 639.072.412.

REFERENCES

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* **57**, 289–300.

BENJAMINI, Y. AND LIU, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* **82**, 163–170.

CHEN, J., LUO, J., LIU, K. AND MEHROTRA, D. V. (2011). On power and sample size computation for multiple testing procedures. *Computational Statistics and Data Analysis* **55**, 110–122.

DAVIDOV, O. (2011). Combining *p*-values using order-based methods. *Computational Statistics and Data Analysis* **55**, 2433–2444.

DENUIT, M., DHAENE, J., GOOVAERTS, M. AND KAAS, R. (2005). *Actuarial Theory for Dependent Risks*. Chichester, England: Wiley.

ELLIS, J. L. (2013). Probability interpretations of intraclass correlations. *Statistics in Medicine* **32**, 4596–4608.

ELLIS, J. L. (2014). An inequality for correlations in unidimensional monotone latent variable models for binary variables. *Psychometrika* **79**, 303–316.

FINNER, H. AND GONTSCHARUK, V. (2009). Controlling the family-wise error rate with plug-in estimator for the proportion of true null hypotheses. *Journal of the Royal Statistical Society, Series B* **71**, 1031–1048.

GLUECK, D. H., MANDEL, J., KARIMPOUR-FARD, A., HUNTER, L. AND MULLER, K. E. (2008). Exact calculations of average power for the benjamini-hochberg procedure. *The International Journal of Biostatistics* **4**, 11.

HOCHBERG, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802.

HOLLAND, P. W. AND ROSENBAUM, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics* **14**, 1523–1543.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

HOMMEL, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383–386.

IGNATIADIS, N., KLAUS, B., ZAUGG, J. B. AND HUBER, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods* **13**, 577.

JUNKER, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *Annals of Statistics* **21**, 1359–1378.

KEILSON, J. AND SUMITA, U. (1982). Uniform stochastic ordering and related inequalities. *Canadian Journal of Statistics* **10**, 181–198.

LISTGARTEN, J., KADIE, C., SCHADT, E. E. AND HECKERMAN, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 16465–16470.

MOKKEN, R. J. (1971). *A Theory and Procedure of Scale-Analysis*. The Hague: Mouton.

NORMAND, S.-L. T. AND SHAHIAN, D. M. (2007). Statistical and clinical aspects of hospital outcomes profiling. *Statistical Science* **22**, 206–226.

RASCH, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Nielson and Lydiche.

RISSO, D., NGAI, J., SPEED, T. P. AND DUDOIT, S. (2014). Normalization of rna-seq data using factor analysis of control genes or samples. *Nature Biotechnology* **32**, 896.

ROBERTSON, T., WRIGHT, F. T. AND DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Chichester, England: Wiley.

RÜSCHENDORF, L. (1991). On conditional stochastic ordering of distributions. *Advances in Applied Probability* **23**, 46–63.

SIJTSMA, K. AND JUNKER, B. W. (2006). Item response theory: past performance, present developments, and future expectations. *Behaviormetrika* **33**, 75–102.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–498.

VAN DER ARK, L. A. (2007). Mokken scale analysis in r. *Journal of Statistical Software* **20**, 1–19.

VAN DER ARK, L. A. (2012). New developments in mokken scale analysis in r. *Journal of Statistical Software* **48**, 1–27.

VAN DER VEN, A. H. G. S. AND ELLIS, J. L. (2000). A rasch analysis of ravens standard progressive matrices. *Personality and Individual Differences* **29**, 45–64.

WELLEK, S. (2017). A critical evaluation of the current âŁœ*p*-value controversyâŁž. *Biometrical Journal* **59**, 854–872.

WHITT, W. (1980). Uniform conditional stochastic order. *Journal of Applied Probability* **17**, 112–123.

WHITT, W. (1982). Multivariate monotone likelihood ratio and uniform conditional stochastic order. *Journal of Applied Probability* **19**, 695–701.

WOLLAN, P. C. AND DYKSTRA, R. L. (1986). Conditional tests with an order restriction as a null hypothesis. In: Dykstra, R. L., Robsertson, T. and Wright, F. T. (editors), *Advances in Order Restricted Statistical Inference*. New Yorks: Springer-Verlag.

ZHU, Y. AND GUO, W. (2017). Familywise error rate controlling procedures for discrete data. Preprint arXiv:1711.08147.

[*Received November 7, 2017; revised July 12, 2018; accepted for publication August 4, 2018*]