Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 260

Avoiding Data Snooping in Multilevel and Mixed Effects Models

David Afshartous and Michael Wolf

December 2005

Avoiding Data Snooping in Multilevel and Mixed Effects Models

David Afshartous
School of Business
University of Miami
Coral Gables, FL 33214
U.S.A.

Michael Wolf
Institute for Empirical Research in Economics
University of Zurich
CH-8006 Zurich
Switzerland

December 2005

Abstract

Multilevel or mixed effects models are commonly applied to hierarchical data; for example, see Goldstein (2003), Raudenbush and Bryk (2002), and Laird and Ware (1982). Although there exist many outputs from such an analysis, the level-2 residuals, otherwise known as random effects, are often of both substantive and diagnostic interest. Substantively, they are frequently used for institutional comparisons or rankings. Diagnostically, they are used to assess the model assumptions at the group level. Current inference on the level-2 residuals, however, typically does not account for data snooping, that is, for the harmful effects of carrying out a multitude of hypothesis tests at the same time. We provide a very general framework that encompasses both of the following inference problems: (1) Inference on the 'absolute' level-2 residuals to determine which are significantly different from zero, and (2) Inference on any prespecified number of pairwise comparisons. Thus, the user has the choice of testing the comparisons of interest. As our methods are flexible with respect to the estimation method invoked, the user may choose the desired estimation method accordingly. We demonstrate the methods with the London Education Authority data used by Rasbash et al. (2004), the Wafer data used by Pinheiro and Bates (2000), and the NELS data used by Afshartous and de Leeuw (2004).

KEY WORDS: Data snooping, hierarchical linear models, hypothesis testing,
pairwise comparisons, random effects, rankings.

1 Introduction

Multilevel modeling is a popular statistical method for analyzing hierarchical data. As such data is commonplace in many disciplines, it naturally follows that multilevel models are employed by researchers in a wide array of subject areas, ranging from clinical trials to educational statistics. The foundation of this technique is the explicit modeling of variability at each level of the hierarchy. Moreover, regression coefficients for individual-level relationships are expressed as random variables, often a function of covariates at higher levels. Depending upon one's statistical allegiance, the multilevel model can be viewed from the perspective of a mixed effects model, a linear model with complex error structure, or a hierarchical Bayes model. Commonly cited motivations for performing a multilevel analysis include the desire to obtain more realistic gauges of estimation uncertainty (i.e., standard errors), the ability to explicitly model the relationship between information at different levels, and improved estimation and prediction via the seminal statistical principal of 'borrowing of strength' (James and Stein, 1961). For details on the history, estimation methods, and available software for multilevel models, see Raudenbush and Bryk (2002), Goldstein (2003), and de Leeuw and Kreft (1986, 1995).

Formally, say we have an outcome measure y_{ij} for the ith observation in the jth group, e.g., the ith student in the jth school. The sample size in the jth group is n_j and there are a total of J groups. The simplest multilevel model is a random intercept model:

$$y_{ij} = \beta_{0j} + \epsilon_{ij} \tag{1}$$

$$\beta_{0j} = \beta_0 + u_j, \tag{2}$$

where $u_j \sim N(0, {\sigma_u}^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$, and $\text{cov}(\epsilon_{ij}, u_j) = 0$. Substituting for β_{0j}, we have:

$$y_{ij} = \beta_0 + u_j + \epsilon_{ij}. \tag{3}$$

This is also recognizable as a random effects ANOVA. Covariate information can be introduced at both the individual and group level to create a more general multilevel model (in matrix notation):

$$Y_j = X_j\beta_j + r_j \tag{4}$$

$$\beta_j = Z_j\gamma + u_j, \tag{5}$$

where X_j is of dimension $n_j \times p$, β_j is a p-vector of random level-1 regression coefficients, r_j is the level-1 error term and may be taken as $N(0, \sigma^2 I)$ where I is dimension $p \times p$, Z_j is a $p \times q$ matrix that includes level-2 variables, γ is a q-vector that includes the level-2 coefficients or fixed effects, and the level-2 error u_j has dispersion matrix τ which expresses the between group variability and covariance of the level-1 coefficients. Substituting for β_j, we have the single equation format:

$$Y_j = X_j Z_j \gamma + X_j u_j + r_j. \tag{6}$$

This format is more similar to the general mixed effects model, where one commonly finds the following equation for longitudinal or repeated measures data:

$$Y_i = X_i \alpha + Z_i b_i + e_i, \tag{7}$$

where $b_i \sim N(0, D)$ and $e_i \sim N(0, \sigma^2 I)$. A subtle difference between multilevel and mixed effects models is that in the mixed effects literature level-2 variables, i.e., variables that are the same for each grouping unit, are rarely employed; Z_i in equation (7) is usually a subset of X_i that determines which of the individual or level-1 regression coefficients (α) are random.[1] Regardless, both equations have the same format in that there are both fixed effects and random effects. For the purpose of illustration, we shall focus on the simple multilevel or random effects ANOVA model of equation (3).

There exist many outputs from a multilevel model analysis. For instance, one may be interested in parameter estimates of the fixed effect β_0, the variance components σ^2 or ${\sigma_u}^2$, or the random effects u_j.

In this paper, we focus on inference for the random effects. Inference for random effects is important for a variety of reasons. Random effects are of substantive interest since they represent the effect or departure of the jth group from the grand mean. To be sure, as the 'true' random effects are unobserved, we base inference for random effects on the estimated random effects.[2] In applied research, it common to see rankings of these estimates, where the implication is that the groups at the top of the ranking perform better with respect to the given outcome measure, and vice versa for the groups at the lower end. Goldstein et al. (1993) argue against such a simplistic use of rankings with respect to educational league tables in the U.K. Instead, they strongly advocate the inclusion of confidence bands to reflect the uncertainty in the level-2 residual estimates.[3] Figure 1 is reproduced from Rasbash et al. (2004, page 39), where the data concerns school achievement in a sample of 65 London Local Education Area (LEA) schools. With the inclusion of the confidence bands, it becomes difficult to infer the rankings of the true unknown random effects u_{0j}. Nevertheless, it is likely that Figure 1 will be used by different individuals for different purposes; many of these purposes are likely to involve multiple hypothesis testing problems.

The multiple hypothesis tests for the random effects reflect many practical questions of interest.

[1] In the mixed effects literature, level-2 variables are referred to as "outer" variables by Pinheiro and Bates (2000).

[2] Indeed, there exists three aspects of the random effects: 1) the random effects probability distribution, 2) the realized values of the random effects that arise from this distribution, and 3) the estimates of these realized values given the data. We only observe 3).

[3] Also note that there exists uncertainty in the confidence bands themselves, since the estimated standard error of $\hat{u}_j$ is used instead of the true standard error when forming the confidence bands. See Longford (1999) for details.

For instance, one may investigate the absolute problem of which random effects are significantly different from zero, thereby identifying the groups which are particularly 'good' (above zero) or 'bad' (below zero); such an analysis may be viewed with respect to the diagnostics of checking the level-2 model assumptions. On the other hand, one may be interested in the set of all pairwise comparisons, where each group is compared to every other group. Given these multiple tests with different policy implications, it behooves the researcher to be clear with respect to the test being conducted. Regardless, when performing multiple hypothesis tests, there exists the potential for overly liberal results due to what is commonly referred to as *data snooping.* Take the example of making inference for the random effects and assume that there are 100 groups that are indeed all equally good (so that all true random effects are equal to zero). However, if the 100 individual p-values are compared to the cutoff point 0.05 (i.e. an individual Type I error of 0.05), then one expects $100 \times 0.05 = 5$ groups to be falsely 'detected' as different from the rest.[4] Clearly, such overly liberal analyses are worrisome, especially if they constitute the basis for policy making.

We develop a general framework for inference on the random effects, accounting for data snooping. The are several advantages to this method. By requiring the formal specification of the test of interest, it becomes more difficult to misuse the results for a different test. More importantly, the method extends an innovative stepwise procedure to account for data snooping by Romano and Wolf (2005b). As a result, our method is more powerful than traditional methods, such as Bonferroni, in the sense that it will often reject more false hypotheses, while still controlling the the *familywise error rate* (FWE), defined as the probability of rejecting at least one true null hypothesis. Given the political saliency of random effects estimates and institutional rankings, this is clearly of benefit to the researcher comparing many institutions. Moreover, our method may be invoked under any general estimation method for the multilevel model. Although many statisticians would argue against formalized hypothesis tests for random effects, testing is still common practice in many fields and thus we should at least provide methods that do the tests correctly.

When the number of individual tests is very large, then controlling the FWE can be too strict. In other words, by controlling the probability of rejecting even one true null hypothesis, it can become very difficult to detect false hypotheses. For example, this situation will typically arise when all pairwise comparisons of level-2 residuals are of interest. If the number of level-2 residuals is J, then there are a total of $\binom{J}{2}$ pairwise comparisons, and this number grows rapidly with J. Of course, this situation can also arise for absolute comparisons in case J itself is very large. In such instances we propose to control the *false discovery proportion* (FDP), defined as the number of false rejections divided by the total number of rejections; and defined to be 0 if there are no rejections at all.

The outline of the paper is as follows. Section 2 formally presents the multiple hypothesis testing problems of interest. Section 3 discusses how to avoid data snooping via the application of

[4] Figure 1 reproduced from Rasbash et al. (2004, page 39) corresponds to such an analysis based on individual p-values, not accounting for data snooping.

novel multiple testing procedures. Section 4 applies the various methods to real data sets. Section 5 concludes with a brief summary.

2 Inference for Level-2 Residuals

The general problem of interest concerns inference for random effects in a multiple hypothesis testing setting. First, we formally define the multiple hypothesis tests of interest. Second, we introduce a general, nonspecified method to arrive at an estimate of u_j, and a specific bootstrap method to arrive at an estimate of u_j, but based on bootstrap data instead of the real data. Finally, we derive the corresponding stepwise multiple testing procedure.

2.1 Absolute Comparisons

In the population, u_j from equation (3) is distributed normally with zero mean and variance ${\sigma_u}^2$. As $j = 1, \dots, J$, we would like to know the values of the J realizations from this distribution. Instead, given the data, we have J estimates $\hat{u}_j$. The first problem of interest is to test if the value of each u_j is significantly different from zero. Formally, for each j, we are testing:

$$H_j : u_j = 0 \quad \text{vs.} \quad H_j' : u_j \neq 0.$$

2.2 Prespecified Pairwise Comparisons

The next problem of interest concerns testing a prespecified number of pairwise comparisons. Formally, one is testing:

$$H_{j,k} : u_j = u_k \quad \text{vs.} \quad H_{j,k}' : u_j \neq u_k$$

for all $(j, k) \in A$ where $A \subset \{1, \dots, J\} \times \{1, \dots, J\}$ is a prespecified set indexing the pairs under comparison. For example, the user might be interested in all pairwise comparisons resulting in:

$$A = \{(j, k) : 1 \leq j \leq J - 1, j < k \leq J\} \quad \text{and} \quad |A| = \binom{J}{2}$$

Here $|A|$ denotes the cardinality of the set A. As another example, the user might want to compare a specific residual, say residual j, to all other residuals resulting in:

$$A = \{(j, k) : 1 \leq k \leq J, k \neq j\} \quad \text{and} \quad |A| = J - 1$$

Of course, other constellations are also possible, such as comparing each residual in a subset of $\{1, \dots, J\}$ to each residual in another (disjoint) subset of $\{1, \dots, J\}$.

2.3 Estimation

Various estimation methods exist for multilevel and mixed models; they manifest themselves in various software packages as well (MLwiN, HLM, Terrace-Two, PROC MIXED, S-Plus, R, etc.). These methods range from simple two-step methods (de Leeuw and Kreft, 1986), to iterative methods based on (full or restricted) maximum likelihood (Raudenbush and Bryk, 2002; Longford, 1987; Goldstein, 2003), to Bayesian Markov Chain Monte Carlo (MCMC) methods (Browne, 1998). Regardless of the estimation procedure of choice, our stepwise multiple testing method is defined in a general manner such that any estimation method may be employed.

Let $\hat{u}_j$ represent a generic estimator for the random effect u_j. Similarly, let $\hat{\sigma}(\hat{u}_j)$ represent the corresponding standard error; that is, $\hat{\sigma}(\hat{u}_j)$ estimates the unknown standard deviation of $\hat{u}_j$. Finally, given a pair of estimated residuals $\hat{u}_j$ and $\hat{u}_k$, let $\widehat{cov}(\hat{u}_j, \hat{u}_k)$ represent the corresponding estimated covariance between $\hat{u}_j$ and $\hat{u}_k$.[5] Regardless of the estimator or test statistic that is employed, we may formulate the multiple testing problem and our stepwise testing procedure.

One commonly employed option for the random effects estimator is the classic shrinkage estimator, which may be viewed as the posterior mode of the distribution of u_j given the data and estimators of the variance components. It is called a shrinkage estimator because the estimate for groups with few observations (n_j) is "shrunk" towards zero. For the classic mixed effects model format of equation (7), Laird and Ware (1982) and Robinson (1991) provide full details on the random effects estimator and corresponding standard error. Briefly, assuming that $\Sigma_i = \text{cov}(y_i) = \sigma^2 I + Z_i D Z_i{}'$ is known[6], for the fixed effects we have:

$$\hat{\alpha} = \left(\sum_i X_i{}' W_i X_i \right)^{-1} \sum_i X_i{}' W_i y_i \tag{8}$$

where $W_i = \Sigma_i{}^{-1}$, and

$$\text{var}(\hat{\alpha}) = \left(\sum_i X_i{}^{-1} W_i X_i \right)^{-1} . \tag{9}$$

Of course, in practice Σ_i is unknown and must be estimated; there exists various iterative methods for estimating these variance components, e.g., Fisher Scoring and the EM algorithm (Longford, 1987; Dempster et al., 1977). Given an estimate of the variance components and fixed effects, we have the well-known random effects estimator (Harville, 1976):

$$\hat{b}_i = \hat{D} Z_i{}' W_i (y_i - X_i \hat{\alpha}) \tag{10}$$

[5] It is assumed that the underlying generic estimation method allows for the computation of standard errors and estimated covariances.

[6] Recall that D is the covariance matrix for the random effects.

and

$$\hat{V}_i \quad = \quad \hat{\mathrm{var}}(\hat{b}_i - b_i) = \hat{D} - \hat{D}Z_i{}'\hat{W}_i Z_i \hat{D} + \hat{D}Z_i{}'\hat{W}_i X_i \left(\sum_i X_i{}'\hat{W}_i X_i \right)^{-1} X_i{}'\hat{W}_i Z_i \hat{D}. \qquad (11)$$

As stated earlier, the random effects estimator $\hat{b}_i$ is a shrinkage estimator, a linear transformation of the ordinary residuals $y_i - X_i\hat{\alpha}$. It may be viewed as a weighted combination of 0 and $\bar{b}_i$, where the latter is the OLS estimate obtained by treating b_i as a fixed effect (Laird and Ware, 1982). One may also examine the realized regression coefficients, i.e., the sum of the estimates for the fixed effects and random effects vectors. These may be compared to the regression coefficients obtained by performing OLS separately in the different groups.[7] As the random effects represent the difference between the two quantities, in a simple intercept-slope model they are called the *residual level* and *residual slope.*

2.4 Bootstrap Method

There exist several variants of bootstrapping for multilevel models, and they may be divided into three basic categories: (1) parametric bootstrap; (2) residual bootstrap; and (3) cases bootstrap. Categories (2) and (3) are both variants of the nonparametric bootstrap. The parametric bootstrap generates new data by keeping the explanatory variables fixed and simulating level-1 and level-2 residuals from an estimated model distribution (typically a normal distribution with mean zero); see Goldstein (2003, Section 3.5) and van der Leeden et al. (2005). The residual bootstrap generates new data by keeping the explanatory variables fixed and resampling the estimated level-1 and level-2 residuals; see Carpenter et al. (2003) and van der Leeden et al. (2005). The cases bootstrap generates new data by resampling entire 'cases' of response variables joint together with their explanatory variables. Depending on the context, only level-1 units are resampled, only level-2 units are resampled, or both level-1 and level-2 units are resampled; see van der Leeden et al. (2005).

Crucially, only a variant of the cases bootstrap is appropriate for our purposes. The reason is that the multiple testing procedure we present in the next section corresponds to (multiple) hypothesis testing via the inversion of joint confidence regions. Therefore, to achieve power, the bootstrap employed must not reflect the constraints of the individual null hypotheses.

To illustrate this important point, consider the simple problem of making inference for the (common) mean μ of a univariate i.i.d. sample $X_1, \ldots, X_n$, where it is assumed that the underlying distribution has a finite second moment. The hypotheses of interest are $H : \mu = 0$ vs. $H' : \mu \neq 0$. There are two ways to use the bootstrap here. First, one can use the bootstrap to carry out a 'direct' test; for example, by attaching a p-value to the observed test statistic $|\bar{X}|$. In this case,

[7] The OLS estimates will exhibit more variability.

one has to resample from a distribution that satisfies the null hypothesis. Most simply, this can be achieved by resampling from the centered data $X_1 - \bar{X}, \ldots, X_n - \bar{X}$. Second, one can use the bootstrap to construct a two-sided confidence interval for μ and then reject the null hypothesis if zero is not contained in the interval. That is, one inverts the confidence interval for μ to carry out an 'indirect' test. In this case, one must not resample from the centered data but from the original sample. Otherwise, one does not achieve any power; see Politis et al. (1999, page 34) for a general discussion.

Now return to the inference problems at hand, detailed in Subsections 2.1 and 2.2. If we use a bootstrap method where the expected values of the level-2 residuals in the bootstrap world are all equal to zero, then the individual null hypotheses of interest are all satisfied. Since our multiple testing method described below is based on the inversion of joint confidence regions, we would not achieve any power in this way. Therefore, the parametric bootstrap and residual bootstrap are ruled out, as for both of them all level-2 residuals have mean zero in the bootstrap world. Instead, we must employ an appropriate cases bootstrap which corresponds to resampling from the observed data. As our hypotheses of interest are about the expected values of the level-2 residuals, the level-2 units and their unit-specific (level-2) variables remain fixed and only the level-1 units are resampled; see Example 2 of van der Leeden et al. (2001, Section 3.3).

Given an estimation method to compute $\hat{u}_j$, the estimator of the random effect u_j, from the original data set, we employ the cases bootstrap, resampling the level-1 units only, to produce a sequence of B bootstrap data sets. Let $\hat{u}_j^{*,b}$ denote the estimator of the random effect u_j computed from the bth bootstrap sample and let $\hat{\sigma}(\hat{u}_j^{*,b})$ denote the standard error of $\hat{u}_j^{*,b}$. The chains $\{\hat{u}_1^{*,b}, \ldots, \hat{u}_B^{*,b}\}$ and $\{\hat{\sigma}(\hat{u}_1^{*,b}), \ldots, \hat{\sigma}(\hat{u}_B^{*,b})\}$ are then used in the stepwise multiple testing procedure described below. This provides the researcher with the option of employing his/her preferred estimation procedure.

3 Avoiding Data Snooping

Much of the current practice for inference on level-2 residuals fails to take the data snooping, which often arises naturally, into account. To motivate our methodology, we briefly mention two representative examples.

Example 3.1 (Data Snooping When Making Absolute Comparisons) Rasbash et al. (2004, page 39) line up confidence intervals for the level-2 residuals with *individual* coverage probability of 95% in a so-called caterpillar plot; this plot is reproduced in our Figure 1. Then residuals whose confidence intervals do not contain zero are identified and the conclusion regarding them is as follows: "Remembering that these residuals represent school departures from the overall average . . . , this means that these are the schools that differ significantly from the average at the 5% level."

However, for this conclusion to be valid, the confidence intervals should have been constructed in such a way that the *joint* coverage probability was given by 95%.

Example 3.2 (Data Snooping When Making Pairwise Comparisons) Figure 2 in Subsection 4.2 of Goldstein and Spiegelhalter (1996) presents school intercept residual estimates and their 95% overlap intervals based on the method of Goldstein and Healy (1995). The intervals are constructed in such a way that for a *single, prespecified* comparison of two residuals, the two can be distinguished (that is, declared significantly different) if their corresponding intervals do not overlap.[8] But Goldstein and Spiegelhalter (1996) conclude that, for example, the school with the smallest estimated residual can be distinguished from each of the highest six schools. These are *multiple, data-dependent* comparisons instead and so the method of Goldstein and Healy (1995) does not apply.

3.1 Problem Formulation

We proceed by presenting a general framework in which the data snooping problem can be formulated and addressed. The unknown probability mechanism generating the data is denoted by P.

Interest focuses on a parameter vector $\theta = \theta(P)$ of dimension S, that is, $\theta = (\theta_1, \ldots, \theta_S)'$. The individual hypotheses are about the elements of θ and of the form

$$H_s \colon \theta_s = 0 \quad \text{vs.} \quad H_s' \colon \theta_s \neq 0 \tag{12}$$

Example 3.3 (Absolute Comparisons) If the expected values of the level-2 residuals are under test, we have $S = J$ and $\theta_s = u_s$.

Example 3.4 (Prespecified Pairwise Comparisons) If pairwise comparisons of the expected values of the level-2 residuals are of interest, we have $S = |A|$ where A is the pre-specified set indexing the pairs under comparison. Denote the bivariate elements of A, ordered in any fashion, by $\{a_1, \ldots, a_S\}$ with typical element $a_s = (a_{s,1}, a_{s,2}) \in \{1, \ldots, J\} \times \{1, \ldots, J\}$. Then $\theta_s = u_{a_{s,1}} - u_{a_{s,2}}$.

A multiple testing method yields a decision concerning each individual testing problem by either rejecting H_s or not. Crucially, in doing so, it takes into account the multitude of the tests, that is, the data snooping.

[8] Under their method, the average type I error over all pairwise comparisons should be 0.05. Note that the validity of their method requires that the individual level-2 residual estimates be independent.

3.2 Problem Solution Based on the FWE

The common approach to this end is to control the *familywise error rate* (FWE), defined as the probability of making at least one false rejection:

$$\mathrm{FWE}_P = P\{\text{Reject at least one } H_s\text{: } \theta_s = 0\}$$

If the FWE is controlled at level α, then one can be $1-\alpha$ confident that all rejected hypotheses are indeed false. In other words, the joint confidence is equal to the FWE level. On the other hand, if the individual tests have each level α, then the confidence that all rejected hypotheses are indeed false is generally less than $1-\alpha$, and potentially much less.[9] In other words, the joint confidence is smaller than the individual confidence.

Strictly speaking, a multiple testing procedure controls the FWE if

$$\mathrm{FWE}_P \leq \alpha \quad \text{for all sample sizes } (n_1, \ldots, n_J) \text{ and for all } P$$

However this is only feasible in very special circumstances, for example in linear regression models under strict parametric assumptions. Realistically, we can only hope to achieve asymptotic control of the FWE defined as

$$\limsup_{\min_{1 \leq j \leq J} n_j \to \infty} \mathrm{FWE}_P \leq \alpha \quad \text{for all } P$$

In the remainder of the paper, when we speak of control of the FWE—and later of alternative criteria to account for data snooping—we always mean asymptotic control.

Traditional methods to control the FWE are based on individual p-values $\hat{p}_1, \ldots, \hat{p}_S$, where $\hat{p}_s$ tests the hypothesis H_s. The well-known Bonferroni method rejects H_s if $\hat{p}_s \leq \alpha/S$. It is a *single-step* method, since all p-values are compared to the same critical value. Its advantage is its simplicity, but it can result in low power. A perhaps less well-known improvement is the method of Holm (1979). The p-values are ordered from smallest to largest: $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \ldots \leq \hat{p}_{(S)}$. Then the hypothesis $H_{(s)}$ is rejected if $\hat{p}_{(j)} \leq \alpha/(S-j+1)$ for all $j = 1, \ldots, s$. This *stepwise* method is obviously more powerful than Bonferroni, since the (ordered) p-values are compared to critical values that are increasing in each step (and equal to the Bonferroni critical value in the first step).

Nevertheless, even the Holm method can be quite conservative. It shares with Bonferroni the disadvantage of being based on the individual p-values. Therefore, to guarantee control of the FWE in general, these methods must assume a 'worst-case' dependence structure of the p-values (or test-statistics). If the true dependence structure could be taken into account, power would

[9] The joint confidence depends on the number of tests, S, and the dependence structure of the individual test statistics. Hence, it can be computed explicitly only in special circumstances where this dependence structure is known. For example, if the test statistics are independent, then the joint confidence is given by $(1-\alpha)^S$.

increase.[10]

Romano and Wolf (2005b) develop a novel stepwise multiple testing procedure that (asymptotically) accounts for the dependence structure of the test statistics and therefore is more powerful than the Holm method. Their framework is that of comparing many strategies (such as investment strategies) to a common benchmark (such as a market index) and deciding which strategies outperform the benchmark.[11] Given this context, the individual tests are (all) one-sided. We therefore now detail how the procedure of Romano and Wolf (2005b) has to be modified when the individual tests are two-sided, which is the case for the applications we have in mind.

The test statistic for the null hypothesis H_s is of the form $|z_s| = |w_s|/\hat{\sigma}_s$, where w_s is a (consistent) estimator of the parameter θ_s and $\hat{\sigma}_s$ is a standard error of w_s.

Example 3.3 continued (Absolute Comparisons) We have $w_s = \hat{u}_s$ and $\hat{\sigma}_s = \hat{\sigma}(\hat{u}_s)$. (Recall that $S = J$ in this example, and so we can 'rewrite' the level-2 residuals as $u_1, \ldots, u_S$ here.)

Example 3.4 continued (Pre-specified Pairwise Comparisons) We have $w_s = \hat{u}_{a_{s,1}} - \hat{u}_{a_{s,2}}$ and $\hat{\sigma}_s = \sqrt{\hat{\sigma}^2(\hat{u}_{a_{s,1}}) + \hat{\sigma}^2(\hat{u}_{a_{s,2}}) - 2\widehat{cov}(\hat{u}_{a_{s,1}}, \hat{u}_{a_{s,2}})}$.

Our method starts out by relabeling the hypotheses in descending order of the test statistics. Hypothesis H_{r_1} corresponds to the largest test statistic and hypothesis H_{r_S} to the smallest one. The first step of the procedure computes a $1 - \alpha$ (asymptotic) joint confidence region for the parameter vector $(\theta_{r_1}, \ldots, \theta_{r_S})'$ of the form

$$[w_{r_1} \pm \hat{\sigma}_{r_1} \hat{d}_1] \times \ldots \times [w_{r_S} \pm \hat{\sigma}_{r_S} \hat{d}_1] \qquad (13)$$

Then, for $s = 1, \ldots, S$, the hypothesis H_{r_s} is rejected if zero is not contained in the interval $[w_{r_s} \pm \hat{\sigma}_{r_s} \hat{d}_1]$. Denote by R_1 the number of hypotheses rejected in this first step. Obviously, if $R_1 = 0$, we stop. Otherwise, in the second step, we construct a $1 - \alpha$ (asymptotic) joint confidence region for the 'remaining' parameter vector $(\theta_{r_{R_1+1}}, \ldots, \theta_S)'$ of the form

$$[w_{r_{R_1+1}} \pm \hat{\sigma}_{r_{R_1}+1} \hat{d}_2] \times \ldots \times [w_{r_S} \pm \hat{\sigma}_{r_S} \hat{d}_1] \qquad (14)$$

Then, for $s = R_1 + 1, \ldots, S$, the hypothesis H_{r_s} is rejected if zero is not contained in the interval $[w_{r_s} \pm \hat{\sigma}_{r_s} \hat{d}_2]$. Denote by R_2 the number of hypotheses rejected in this second step. If $R_2 = 0$, we stop and otherwise we continue in this stepwise fashion.

We are left to specify how to compute the constants $\hat{d}_1, \hat{d}_2, \ldots$. To this end, define

[10] To give an extreme example, if all p-values are equal, then the single-step critical value can be increased to α compared to the Bonferroni 'worst-case' critical value of α/S.

[11] So the multitude of tests arises via the multiple comparisons 'strategy versus benchmark'.

$$d(1-\alpha, P, R) = \inf\{x : \text{Prob}_P\{\max_{R+1 \le s \le S} |w_{r_s} - \theta_{r_s}|/\hat{\sigma}_{r_s} \le x\} \ge 1-\alpha\}$$

That is, $d(1-\alpha, P, R)$ is a $1-\alpha$ quantile of the sampling distribution under P of the random variable $\max_{R+1\le s\le S} |w_{r_s} - \theta_{r_s}|/\hat{\sigma}_{r_s}$. The ideal choices would then be given by $\hat{d}_1 = d(1-\alpha, P, 0), \hat{d}_2 = d(1-\alpha, P, R_1)$, and so on.[12] But since the true probability mechanism P is unknown, these choices are not feasible. Instead, a bootstrap approach yields feasible constants: P is replaced by an estimator $\hat{P}$ and then one takes $\hat{d}_1 = d(1-\alpha, \hat{P}, 0), \hat{d}_2 = d(1-\alpha, \hat{P}, R_1)$, and so on. For details on how to compute the constants $\hat{d}_j$ in Examples 3.3 and 3.4, see Appendix A.

We can now summarize our stepwise method by the following algorithm. The acronym StepM stands for 'Stepwise Multiple Testing'.[13]

Algorithm 3.1 (StepM Method)

1. Relabel the hypotheses in descending order of the test statistics $|z_s|$: strategy r_1 corresponds to the largest test statistic and strategy r_S to the smallest one.
2. Set $j = 1$ and $R_0 = 0$.
3. For $R_{j-1} + 1 \le s \le S$, if $0 \notin [w_s \pm \hat{\sigma}_s \hat{d}_j]$, reject the null hypothesis H_{r_s}.
4. (a) If no (further) null hypotheses are rejected, stop.

   (b) Otherwise, denote by R_j the total number of hypotheses rejected so far and, afterwards, let $j = j + 1$. Then return to step 3.

We briefly return to the motivating examples at the beginning of this section. Algorithm 3.1 applied to Example 3.3 would avoid the data snooping in Example 3.1. In particular, the joint confidence region (13) could be easily turned into an appropriate caterpillar plot which allows the user to identify school departures from the overall average without falling into the data snooping trap. Nevertheless, some further departures might be identified in subsequent steps. Therefore, the caterpillar plot 'adjusted for data snooping' is a useful and intuitive tool but should not be the end of the analysis (unless all intervals contain zero). Algorithm 3.1 applied to Example 3.4 would avoid the data snooping in Example 3.2. Note that comparing the lowest school to the highest school(s) requires an adjustment for data snooping based on all $S = \binom{J}{2}$ pairwise comparisons. Unfortunately, in this example, the first step of our method cannot be translated into a convenient plot.

[12] These choices would guarantee exact joint coverage probability $1-\alpha$ of the confidence regions (13) and (14).

[13] Both the Holm method and our StepM method are *stepdown* methods, that is, they start by examining the most significant hypothesis $H_(1)$ and then move 'down' to the less significant hypotheses.

3.3 Problem Solution Based on the FDP

If the number of hypotheses under consideration, S, is very large, controlling the FWE may be too strict. In such instances, one might be willing to tolerate a certain (small) proportion of false rejections out of the total rejections. This suggests to base error control on the *false discovery proportion* (FDP). Let F be the number of false rejections made by a multiple testing method and let R be the total number of rejections. Then the FDP is defined as follows:

$$\text{FDP} = \begin{cases} \frac{F}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

By control of the FDP, we mean control of the tail probability $P\{\text{FDP} > \gamma\}$ where $\gamma \in [0, 1)$ is a user-defined number:

$$\limsup_{\min_{1 \le j \le J} n_j \to \infty} P\{\text{FDP} > \gamma\} \le \alpha \quad \text{for all } P$$

Typical values are $\gamma = 0.05$ and $\gamma = 0.1$; the choice $\gamma = 0$ corresponds to control of the FWE.

Lehmann and Romano (2005) propose a stepwise method based on individual p-values. But similar to the Holm (1979) method for FWE control it often is overly conservative because it does not account for the dependence structure across the p-values. Romano and Wolf (2005a) develop a procedure that, similar to the StepM method for FWE control, takes into account the dependence structure across test statistics via the use of the bootstrap. They do this in the context of one-sided individual tests. Therefore, we detail how their procedure, coined the FDP-StepM method, has to be modified for the context of two-sided tests, which is what we are interested in.

The method is built upon a generalization of the StepM method introduced in Subsection 3.2. This generalization is called the k-StepM method and it controls the *generalized familywise error rate* (k-FWE), defined as the probability of making at least k false rejections, where $k \geq 1$ is a pre-specified integer:

$$k\text{-FWE}_P = P\{\text{Reject at least } k \text{ of the } H_s\text{: } \theta_s = 0\}$$

Obviously, for $k = 1$ we have 1-FWE$_P$ = FWE$_P$ and the 1-StepM method therefore is identical to the StepM method. But for $k > 1$ the k-FWE criterion is less strict than the FWE criterion and so in general a larger number of hypotheses are rejected. The description of the algorithm to achieve control of the k-FWE is necessarily terse to maintain the flow of the paper. Readers interested in motivation and background are referred to Romano and Wolf (2005a).

Some notation is required. Suppose $\{y_s : s \in K\}$ is a collection of real numbers indexed by a finite set K having $|K|$ elements. Then, for $k \leq |K|$, the $k\text{-}\max_{s \in K}(y_s)$ is used to denote the kth largest value of the y_s with $s \in K$. So, if the elements y_s, $s \in K$, are ordered as

$$y_{(1)} \leq \cdots \leq y_{(|K|)} \ ,$$

then

$$k\text{-max}_{s \in K}(y_s) = y_{(|K|-k+1)}$$

Further, for any $K \subset \{1, \ldots, S\}$, define

$$d_K(1-\alpha, k, P) = \inf\{x : P\{k\text{-max}_{s \in K}|w_{r_s} - \theta_{r_s}|/\hat{\sigma}_{r_s} \leq x\} \geq 1 - \alpha\} \tag{15}$$

That is, $d_K(1-\alpha, k, P)$ is the smallest $1-\alpha$ quantile of the sampling distribution under P of $k\text{-max}_{s \in K}|w_{r_s} - \theta_{r_s}|/\hat{\sigma}_{r_s}$. These quantiles would yield finite sample control of the k-FWE. But since the true probability mechanism P is unknown, these choices are not feasible. Instead, a bootstrap approach yields feasible constants, resulting in asymptotic control of the k-FWE: P is replaced by an estimator $\hat{P}$ and then one takes $\hat{d}_K(1-\alpha, k, P) = d_K(1-\alpha, k, \hat{P})$. For details on how to compute such constants via the bootstrap in Examples 3.3 and 3.4, see Appendix A.

Algorithm 3.2 (k-StepM Method)

1. Relabel the strategies in descending order of the test statistics $|z_s|$: strategy r_1 corresponds to the largest test statistic and strategy r_S to the smallest one.

2. For $1 \leq s \leq S$, if $\theta_{r_s} \notin [w_{r_s} \pm \hat{\sigma}_{r_s} \hat{d}_1]$, reject the null hypothesis H_{r_s}. Here

$$\hat{d}_1 = d_{\{1,\ldots,S\}}(1-\alpha, k, \hat{P})$$

3. Denote by R_1 the number of hypotheses rejected. If $R_1 < k$, stop; otherwise let $j = 2$.

4. For $R_{j-1} + 1 \leq s \leq S$, if $\theta_{r_s} \notin [w_{r_s} \pm \hat{\sigma}_{r_s} \hat{d}_j]$, reject the null hypothesis H_{r_s}. Here

$$\hat{d}_j = \max\{d_K(1-\alpha, k, \hat{P}) : K = I \cup \{R_{j-1}+1, \ldots, S\}, I \subset \{1, \ldots, R_{j-1}\}, |I| = k-1\} \tag{16}$$

5. (a) If no further hypotheses are rejected, stop.

   (b) Otherwise, denote by R_j the number of all hypotheses rejected so far and, afterwards, let $j = j + 1$. Then return to step 4.

Remark 3.1 (Operative Method) The computation of the constants $\hat{d}_j$ in (16) may be very expensive in case $\binom{R_{j-1}}{k-1}$ is large. In such cases, we suggest the following shortcut. Pick a user-defined number N_{max}, say $N_{max} = 50$ and let N^* be the largest integer for which $\binom{N^*}{k-1} \leq N_{max}$.

The constant $\hat{d}_j$ is then computed as follows

$$\hat{d}_j = \max\{\hat{d}_K(1-\alpha, k, \hat{P}_T) : K = I \cup \{R_{j-1}+1, \ldots, S\}, I \subset \{R_j - N^* + 1, \ldots, R_j\},\ |I| = k-1\}$$

That is, we maximize over subsets I not necessarily of the entire index set of previously rejected hypotheses but only of the index set corresponding to the N^* least significant hypotheses rejected so far. Note that this shortcut does not affect the asymptotic control of the k-FWE even if $N_{max} = 1$ is chosen, resulting in $N^* = k - 1$ and

$$\hat{d}_j = \hat{d}_{\{R_{j-1}-k+2,\ldots,S\}}(1-\alpha, k, \hat{P}_T)$$

Nevertheless, in the interest of better k-FWE control in finite samples, we suggest to choose N_{max} as large as possible.

Having defined the k-StepM method, we can now detail the algorithm for the method controlling the FDP. Basically, the method successively applies the k-StepM method, for increasing values of k, until a termination criterion is satisfied.

Algorithm 3.3 (FDP-StepM Method)

1. Let $j = 1$ and let $k_1 = 1$.
2. Apply the k_j-StepM method and denote by N_j the number of hypotheses rejected.
3. (a) If $N_j < k_j/\gamma - 1$, stop.
   (b) Otherwise, let $j = j + 1$ and, afterwards, let $k_j = k_{j-1} + 1$. Then return to step 2.

3.4 Problem Solution Based on the FDR

Benjamini and Hochberg (1995) propose a stepwise method for controlling the expected value of the FDR, E(FDR), which they coin the *false discovery rate*. The idea is to ensure $FDR \leq \gamma$, at least asymptotically, for some user-defined $\gamma \in (0, 1)$. The method is based on individual p-values and works as follows.

The p-values are ordered from smallest to largest: $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \ldots \leq \hat{p}_{(S)}$ with their corresponding null hypotheses labeled accordingly: $H_{(1)}$, $H_{(2)}$, ..., $H_{(S)}$. Then define

$$j^* = \max\left\{j : \hat{p}_{(j)} \leq \gamma_j\right\} \quad \text{where} \quad \gamma_j = \frac{j}{S}\gamma \tag{17}$$

and reject $H_{(1)}, \ldots, H_{(j^*)}$. If no such j exists, reject no hypotheses.[14]

[14] This is an example of a *stepup* method. It starts with examining the least significant hypothesis, $H_{(S)}$, and then moves 'up' to the more significant hypotheses.

Williams et al. (1999) endorse this approach for inference concerning all pairwise comparisons. But two problems need to be mentioned. First, the procedure of Benjamini and Hochberg (1995) does not work under arbitrary dependence structure of the individual p-values. There exist certain sufficient conditions on this dependence structure, but the scenario of all pairwise comparisons does not meet any of them. Yekutieli (2002) provides a more conservative FDR procedure that is shown to work for the scenario of all pairwise comparisons. Second, the FDR is the expected value of the FDP. Controlling the expected value says very little about the actual realization of the FDP in a given application. Indeed, the realized value of the FDP could be quite far away from the nominal upper bound γ on the FDR; see Korn et al. (2004) for some simulation evidence. To give an example, consider controlling FDR ≤ 0.1. This does not allow one to make any statement about the realized FDP in a given application.[15] On the other hand, if one controls $P\{\text{FDP} > 0.1\} \leq 0.05$, say, then one can be 95% confident that the realized FDP in a given application is at most 0.1.

3.5 Comparison of Problem Solutions

Which is the most appropriate of the multiple testing procedures we have presented so far? The answer is 'it depends'.

The StepM method has the advantage that it allows for the strongest conclusions. Since it controls the strict FWE criterion, one can be confident that indeed all rejected hypotheses are false ones. For example, such a 'joint confidence' may be very desirable in the context of policy making.

On the other hand, when the number of hypotheses under consideration is very large, controlling the FWE may be too strict and, as a result, the StepM method may reject only a (relatively) small number of hypotheses. In such cases, both the FDP-StepM method and the FDR method offer greater power, at the expense of tolerating a small (expected) fraction of true hypotheses rejected among all rejections. Of the two, the FDP-StepM method has the advantage that it allows for a statement about the realized FDP in any given application. Say, one can be 95% confident that the realized FDP is at most 10%. The FDR method, on the other hand, only controls the expected value of the FDP and in any given application it could be quite far away from the realized value. Though, by being less 'safe' in this sense, the FDR method often rejects some more hypotheses than the StepM-FDP method.

While, for these reasons, the (globally) most appropriate method does not exist, there clearly does exist an inappropriate method. Namely, the naive approach of basing inference on individual p-values without taking the data snooping into account.

[15] If one controlled FDR ≤ 0.1 in a large number of independent applications, then one could make certain claims concerning the average realized FDP over the many applications. However, most applied researchers will be interested in a single application at hand only.

4 Applications

We compare the various multiple testing methods for three data sets. Although we employ a specific estimation method, the user may implement the various multiple testing procedures with his/her estimation method of choice. Random effects models were estimated via the `nlme` package of Pinheiro and Bates (2000) which is contained in the statistical software `R`.[16] The default estimation method in `nlme` is restricted maximum likelihood (REML), and this is the estimation method we used; see Pinheiro and Bates (2000, Chapter 2) for further details. `R` extensions were written for the standard errors of the random effects estimates, the covariances between random effects estimates, the bootstrapping of the data, as well as the StepM and FDP-StepM methods themselves.

In all applications below, we use the significance level $\alpha = 0.05$ and the value $\gamma = 0.1$ (for the FDP-StepM and the FDR methods). The k-StepM building blocks for the FDP-StepM method use $N_{max} = 100$. All bootstraps use $B = 1,000$ repetitions.

4.1 Data Snooping When Making Absolute Comparisons

Consider the data set used in Rasbash et al. (2004), where the response variable is the score achieved by 16 year old students in an examination (exam score) and the predictor is the London Reading Test score (LRT score) obtained by the same students just before they entered secondary school at the age of 11 years. The data is from an English Local Education Authority (LEA) and consists of $4,059$ students in 65 schools. As in Rasbash et al. (2004), we fit a multilevel or random effects model with random intercept and constant slope across schools. Since there are 65 schools, there are $S = 65$ absolute comparisons, where the absolute comparison of a group's level-2 residual to zero is equivalent to examining whether the school's average exam score differs from the grand mean after accounting for LRT score. If one simply computes the separate test statistics for the random effects and their corresponding p-values, 28 null hypotheses are rejected, i.e., we conclude that 28 schools differ significantly from the grand mean. This method is equivalent to forming separate 95% confidence intervals and rejecting the hypotheses that correspond to intervals that do not include zero. Figure 1 illustrates such a plot for this data. Of course, this approach does not account for data snooping. The application of the StepM, FDP-StepM, and FDR methods yield 17, 27, and 27 rejections, respectively.

Consider the NELS data set used by Afshartous and de Leeuw (2004), where the base year sample from the National Educational Longitudinal Study of 1988 (NELS:88) is used. The base-year sample consists of 24,599 eighth grade students, distributed amongst 1,052 schools nationwide. The response variable is student mathematics score and the predictor is the socio-economic status (SES) of the student. As above, we fit a multilevel or random effects model with random intercept and constant slope across schools. If one simply computes the separate test statistics for the random

[16] This software can be freely downloaded from `http://cran.r-project.org/`.

effects and their corresponding p-values, 289 hypotheses are rejected, i.e., we conclude that 289 of 1052 schools differ significantly from the grand mean. However, as mentioned above, this approach does not account for data snooping. The application of the StepM, FDP-StepM, and FDR methods yield 38, 249, and 244 rejections, respectively. The results are summarized in Table 1.

4.2 Data Snooping When Making Pairwise Comparisons

Consider the Wafer data presented in Pinheiro and Bates (2000). The data was collected to study the variability in the manufacturing of analog MOS circuits and consists of 40 observations on each of 10 wafers; the response variable is the intensity of current and the predictor variable is voltage. Given that there are 10 wafers, there are $S = 45$ possible pairwise comparisons. If one simply examines the test statistics for the pairwise differences of random effects and their corresponding p-values[17], 30 hypotheses are rejected. The application of the StepM, FDP-StepM, and FDR methods yield 26, 30, and 32 rejections, respectively.

Remark 4.1 The graphical method of Goldstein and Healy (1995) can be interpreted as a 'visual shortcut' to an analysis based on individual p-values, ignoring the effects of data snooping. For a given pair of level-2 residuals, u_j and u_k, the null hypothesis $H_0 : u_j = u_k$ is rejected if the overlap intervals for u_j and u_k do not overlap. Crucially, the method of Goldstein and Healy (1995) assumes independence of the level-2 residuals estimates. However, for most estimation routines this assumption is violated because these estimates share common estimated parameters (in particular the estimates of the variances and covariances of residuals). Falsely assuming independence can therefore lead to faulty analyses. Figure 2 presents the method of Goldstein and Healy (1995) applied to the Wafer data of Pinheiro and Bates (2000). A total of 24 rejections is obtained. Obviously, it is counterintuitive that a method which does not account for data snooping should reject fewer hypotheses than even the StepM method. But this riddle is solved by incorporating the estimated covariances of the level-2 residual estimates in a modified Goldstein and Healy (1995) plot, which is presented in Figure 3. Now the lengths of the intervals are reduced and a total of 30 rejections are obtained, the same amount as for the above analysis based on individual p-values.

We also investigate all pairwise comparisons for the data set of Rasbash et al. (2004). Given that there are 65 schools, there are a total of $S = 2,016$ possible pairwise comparisons. If one examines only the individual p-values, a total of $1,027$ hypothesis are rejected. The application of the StepM, FDP-StepM, and FDR methods yield 348, 1,066, and 1,026 rejections, respectively. For the method of Goldstein and Healy (1995), falsely assuming independence of the level-2 residual estimates, there are 977 rejections; see Figure 4. If one accounts for the covariances, there are $1,031$ rejections; see Figure 5. As expected, with the covariances accounted for, this 'visual shortcut' number is now very close to the number of 1,027 rejections for the 'exact analysis' based on individual p-values.

[17] This must take into account the covariances between the corresponding level-2 residual estimates.

5 Conclusion

Level-2 residuals, also known as random effects, are of both substantive and diagnostic interest for multilevel and mixed effects models. A common example is the interpretation of level-2 residuals as school performance. Unfortunately, current inference on level-2 residuals typically ignores the pitfalls of data snooping which arises when multiple hypothesis tests are carried out at the same time. As a consequence, often too many findings are declared significant. This can have undesirable consequences, in particular if such analyses constitute the basis for policy making. Take the example when a particular school is unjustly declared an 'underperformer' with respect to the main body of schools.

In this paper, we have presented two novel multiple testing methods which account for data snooping. Our general framework encompasses both of the following inference problems: (1) Inference on the 'absolute' level-2 residuals to determine which are significantly different from zero, and (2) Inference on any prespecified number of pairwise comparisons. (Thus, the user has the choice of testing the comparisons of interest.)

Out first method controls the *familywise error rate* (FWE) which is defined as the probability of making even one false rejection. If the FWE is controlled at level 5%, say, then one can be 95% confident that all rejected hypotheses are indeed false. The advantage of the method we propose over traditional methods controlling the FWE, such as the methods of Bonferroni and Holm (1979), is an increase in power. This is because our method takes advantage of the dependence structure of the individual test statistics, while the methods of Bonferroni and Holm assume a 'worst-case' scenario.

When the number of hypotheses under test is very large—which can happen, for example, when all pairwise comparisons are of interest—then controlling the FWE may be too strict. In such cases, we propose to control the *false discovery proportion* (FDP) instead, which is defined as the proportion of false rejections divided by the total number of rejections. By allowing a small proportion of the 'discoveries' to be false ones, often a much larger number of hypotheses can be rejected. Our second method is related to the procedure of Benjamini and Hochberg (1995) that controls the *false discovery rate* (FDR), which is defined as the expected value of the false discovery proportion, that is, FDR $= E(\text{FDP})$. However, their method has the drawback that it does not allow for any probability statements concerning the realized FDP in a given application. This can be a problem if the analysis constitutes the basis for policy making.

The practical application of our methods is based on the bootstrap and so it is computationally expensive. However, given the fast computers of today, this no longer is a serious drawback.

References

Afshartous, D. and de Leeuw, J. (2004). An application of multilevel model prediction to NELS:88. *Behaviormetrika*, 31(1):43–66.

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(1):289–300.

Browne, W. J. (1998). *Applying MCMC methods to multilevel models.* PhD thesis, University of Bath, Bath, U.K.

Carpenter, J. R., Goldstein, H., and Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Applied Statistics*, 52(4):431–443.

de Leeuw, J. and Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics*, 11:57–85.

de Leeuw, J. and Kreft, I. (1995). Questioning multilevel models. *Journal of Educational and Behavioral Statistics*, 20:171–189.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–8.

Goldstein, H. (2003). *Multilevel Statistical Models.* Hodder Arnold, Hodder Arnold: London, third edition.

Goldstein, H. and Healy, M. J. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society: Series A*, 158:175–177.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., and Thomas, S. (1993). A multilevel analysis of school examination results. *Oxford Review of Education*, 19:425–433.

Goldstein, H. and Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A*, 159:385–443.

Harville, D. (1976). Extension of the Gauss-Markov theorem to include estimation of random effects. *Annals of Statistics*, 4:384–395.

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

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:361–380.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2):379–398.

Laird, N. M. and Ware, J. W. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963–974.

Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *Annals of Statistics*, 33(3):1138–1154.

Longford, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with random effects. *Biometrika*, 74:817–827.

Longford, N. (1999). Standard errors in multilevel analysis. *Multilevel Modeling Newsletter*, 11:10–13.

Pinheiro, J. and Bates, D. (2000). *Mixed Effects Models in S and S-Plus.* Springer, New York.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling.* Springer, New York.

Rasbash, J., Steele, F., Browne, W., and Prosser, B. (2004). *A User's Guide to MLwiN Version 2.0.* Institute of Education, London. Available at `http://multilevel.ioe.ac.uk/download/userman20.pdf`.

Raudenbush, S. and Bryk, A. (2002). *Hierarchical Linear Models.* Sage Publications, Sage Publications: Newbury Park.

Robinson, G. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6:15–51.

Romano, J. P. and Wolf, M. (2005a). Control of generalized error rates in multiple testing. Working Paper 245, IEW, University of Zurich. Available at `http://www.iew.unizh.ch/wp/index.php`.

Romano, J. P. and Wolf, M. (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

van der Leeden, R., Meijer, E., and Busing, F. M. T. A. (2005). Resampling multilevel models. In de Leeuw, J. and Kreft, I., editors, *Handbook of Quantitative Multilevel Analysis.* Kluwer, New York.

Williams, V. S. L., Jones, L. V., and Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, 24(1):42–69.

Yekutieli, D. (2002). A false discovery rate procedure for pairwise comparisons. Technical report, Department of Statistics and Operations Research, Tel Aviv University. Available at `http://www.math.tau.ac.il/%7Eyekutiel/work.html`.

A Use of the Bootstrap

A.1 Use of the Bootstrap for the StepM Method

We now detail how to compute the constants $\hat{d}_j$ in Examples 3.3 and 3.4 via the bootstrap for use in Algorithm 3.1. Again, the bootstrap method employed is the cases bootstrap resampling the level-1 units only; see van der Leeden et al. (2001, Section 3.3). Denote the observed data by V. The application of the cases bootstrap results in a (generic) bootstrap data set V^*.

Recall that $S = J$ in Example 3.3 and that we can therefore 'rewrite' the level-2 residuals as $u_1, \ldots, u_S$ for this example.

Algorithm A.1 (Computation of the $\hat{d}_j$ for Example 3.3)

1. The labels $r_1, \ldots, r_S$ and the numerical values of $R_0, R_1 \ldots$ are given in Algorithm 3.1.
2. Generate B bootstrap data sets $V^{*,1}, \ldots, V^{*,B}$.
   (One should use $B \geq 1,000$ in practice.)
3. From each bootstrap data set $V^{*,b}$, $1 \leq b \leq B$, compute the individual level-2 residual estimates $\hat{u}_1^{*,b}, \ldots, \hat{u}_S^{*,b}$. Also, compute the corresponding standard errors $\hat{\sigma}(\hat{u}_1^{*,b}), \ldots, \hat{\sigma}(\hat{u}_S^{*,b})$.
4. (a) For $1 \leq b \leq B$, compute $max_j^{*,b} = \max_{R_{j-1}+1 \leq s \leq S} |\hat{u}_{r_s}^{*,b} - \hat{u}_{r_s}| / \hat{\sigma}(\hat{u}_{r_s} s^{b,*})$.
   (b) Compute $\hat{d}_j$ as the $1 - \alpha$ empirical quantile of the B values $max_j^{*,1}, \ldots, max_j^{*,B}$.

Algorithm A.2 (Computation of the $\hat{d}_j$ for Example 3.4)

1. The labels $r_1, \ldots, r_S$ and the numerical values of $R_0, R_1 \ldots$ are given in Algorithm 3.1.
2. Generate B bootstrap data sets $V^{*,1}, \ldots, V^{*,B}$.
   (One should use $B \geq 1,000$ in practice.)
3. From each bootstrap data set $V^{*,b}$, $1 \leq b \leq B$, compute the individual level-2 residual estimates $\hat{u}_1^{*,b}, \ldots, \hat{u}_J^{*,b}$. Also, for a particular difference $w_s^* = \hat{u}_{a_{s,1}}^{b,*} - \hat{u}_{a_{s,2}}^{b,*}$ compute the corresponding standard error $\hat{\sigma}_s^{*,b} = \hat{\sigma}(\hat{u}_{a_{s,1}}^{b,*} - \hat{u}_{a_{s,2}}^{b,*})$.
4. (a) For $1 \leq b \leq B$, compute $max_j^{*,b} = \max_{R_{j-1}+1 \leq s \leq S} |w_{r_s}^* - w_{r_s}| / \hat{\sigma}_s^{*,b}$.
   (b) Compute $\hat{d}_j$ as the $1 - \alpha$ empirical quantile of the B values $max_j^{*,1}, \ldots, max_j^{*,B}$.

A.2 Use of the Bootstrap for the k-StepM Method

We next detail how to compute the constants $\hat{d}_j$ in Examples 3.3 and 3.4 via the bootstrap for use in Algorithm 3.2.

Algorithm A.3 (Computation of the $\hat{d}_j$ for Example 3.3)

1. The labels $r_1, \ldots, r_S$ and the numerical values of $R_0, R_1 \ldots$ are given in Algorithm 3.2.
2. Generate B bootstrap data sets $V^{*,1}, \ldots, V^{*,B}$.
   (One should use $B \geq 1,000$ in practice.)
3. From each bootstrap data set $V^{*,b}$, $1 \leq b \leq B$, compute the individual level-2 residual estimates $\hat{u}_1^{*,b}, \ldots, \hat{u}_S^{*,b}$. Also, compute the corresponding standard errors $\hat{\sigma}(\hat{u}_1^{*,b}), \ldots, \hat{\sigma}(\hat{u}_S^{*,b})$.
4. (a) For $1 \leq b \leq B$, and any needed K, compute $kmax_K^{*,b} = k\text{-max}_{s \in K}(|\hat{u}_{r_s}^{*,b} - \hat{u}_{r_s}|/\hat{\sigma}_{r_s}^{*,b})$.
   (b) Compute $d_K(1-\alpha, k, \hat{P}_T)$ as the $1-\alpha$ empirical quantile of the B values $kmax_K^{*,1}, \ldots, kmax_K^{*,B}$.
5. If $j = 1$, $\hat{d}_1 = d_{\{1,\ldots,S\}}(1-\alpha, k, \hat{P}_T)$
   If $j > 1$, $\hat{d}_j = \max\{d_K(1-\alpha, k, \hat{P}_T) : K = I \cup \{R_{j-1}+1, \ldots, S\}, I \subset \{1, \ldots, R_{j-1}\}, |I| = k-1\}$

Algorithm A.4 (Computation of the $\hat{d}_j$ for Example 3.4)

1. The labels $r_1, \ldots, r_S$ and the numerical values of $R_0, R_1 \ldots$ are given in Algorithm 3.2.
2. Generate B bootstrap data sets $V^{*,1}, \ldots, V^{*,B}$.
   (One should use $B \geq 1,000$ in practice.)
3. From each bootstrap data set $V^{*,b}$, $1 \leq b \leq B$, compute the individual level-2 residual estimates $\hat{u}_1^{*,b}, \ldots, \hat{u}_J^{*,b}$. Also, for a particular difference $w_s^* = \hat{u}_{a_{s,1}}^{b,*} - \hat{u}_{a_{s,2}}^{b,*}$ compute the corresponding standard error $\hat{\sigma}_s^{*,b} = \hat{\sigma}(\hat{u}_{a_{s,1}}^{b,*} - \hat{u}_{a_{s,2}}^{b,*})$.
4. (a) For $1 \leq b \leq B$, and any needed K, compute $kmax_K^{*,b} = k\text{-max}_{s \in K}(|w_{r_s}^{*,b} - w_{r_s}|/\hat{\sigma}_{r_s}^{*,b})$.
   (b) Compute $d_K(1-\alpha, k, \hat{P}_T)$ as the $1-\alpha$ empirical quantile of the B values $kmax_K^{*,1}, \ldots, kmax_K^{*,B}$.
5. If $j = 1$, $\hat{d}_1 = d_{\{1,\ldots,S\}}(1-\alpha, k, \hat{P}_T)$
   If $j > 1$, $\hat{d}_j = \max\{d_K(1-\alpha, k, \hat{P}_T) : K = I \cup \{R_{j-1}+1, \ldots, S\}, I \subset \{1, \ldots, R_{j-1}\}, |I| = k-1\}$

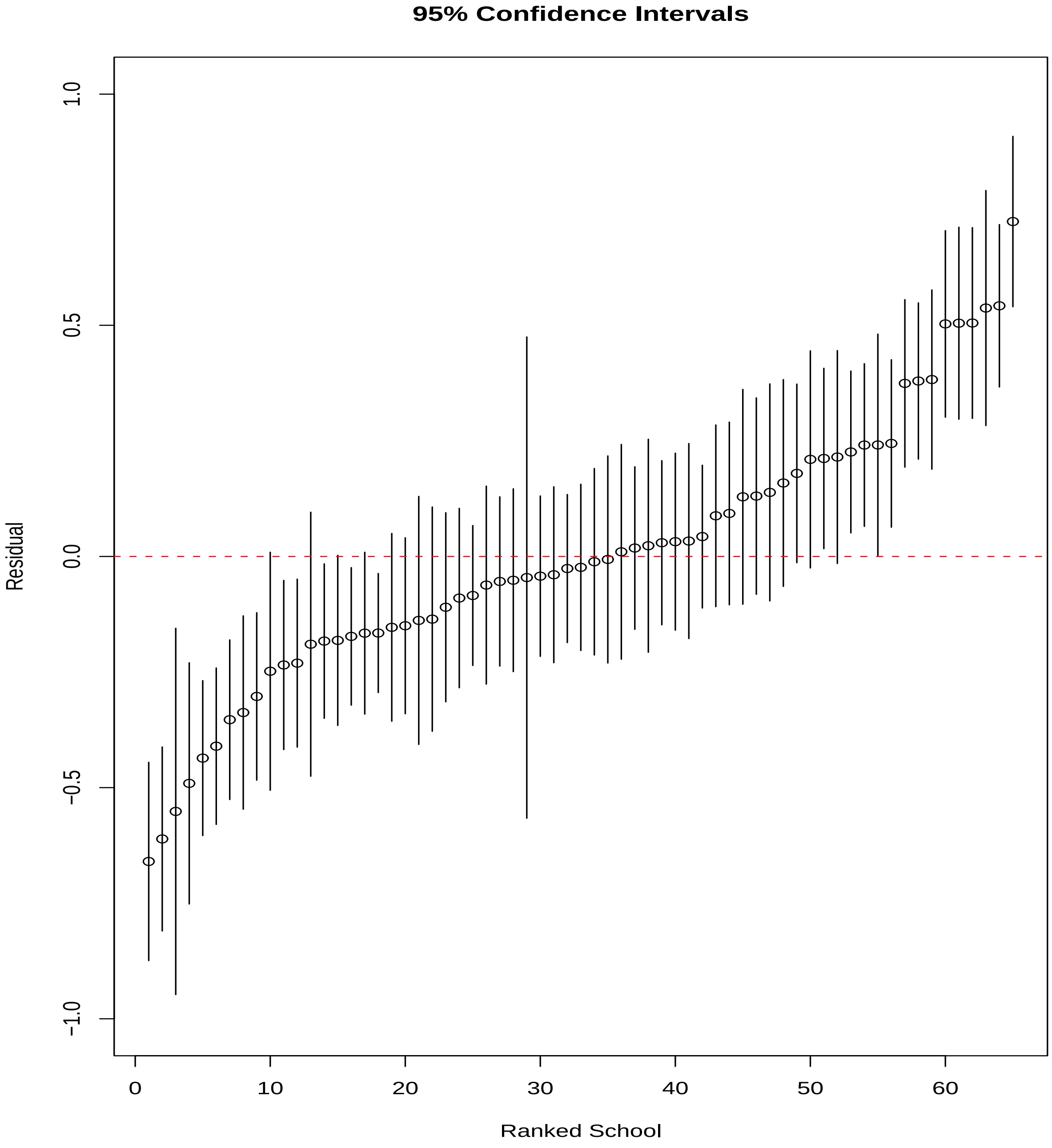

Figure 1: The 'caterpillar plot' of Rasbash et al. (2004, page 39): the level-2 residuals of the 65 schools in ascending order together with their respective 95% confidence intervals. 28 of the 65 intervals do not contain zero.

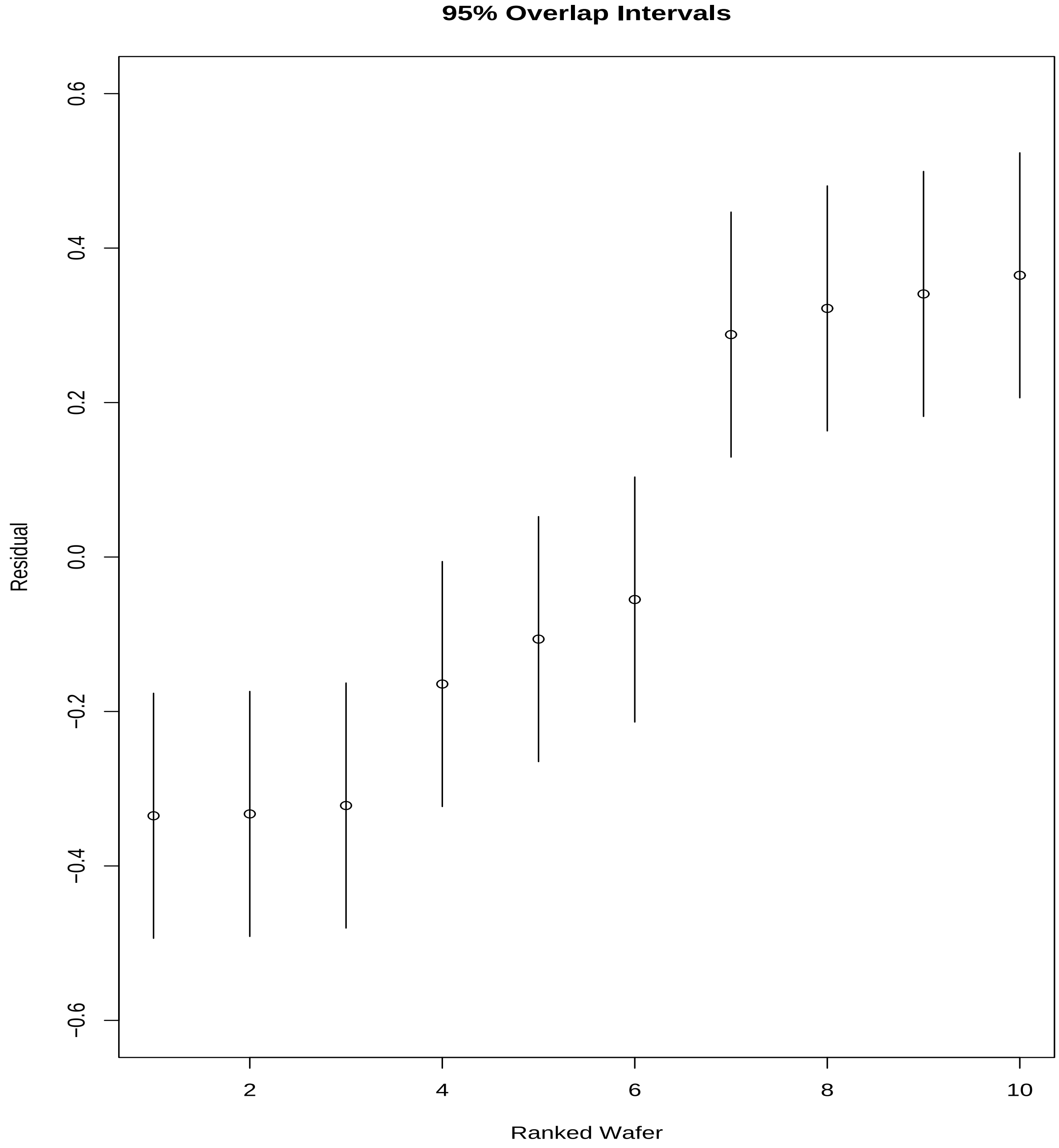

Figure 2: The graphical method of Goldstein and Healy (1995) applied to the Wafer data of Pinheiro and Bates (2000), falsely assuming independence of the level-2 residual estimates. There are 45 pairs of intervals, out of which 24 do not overlap.

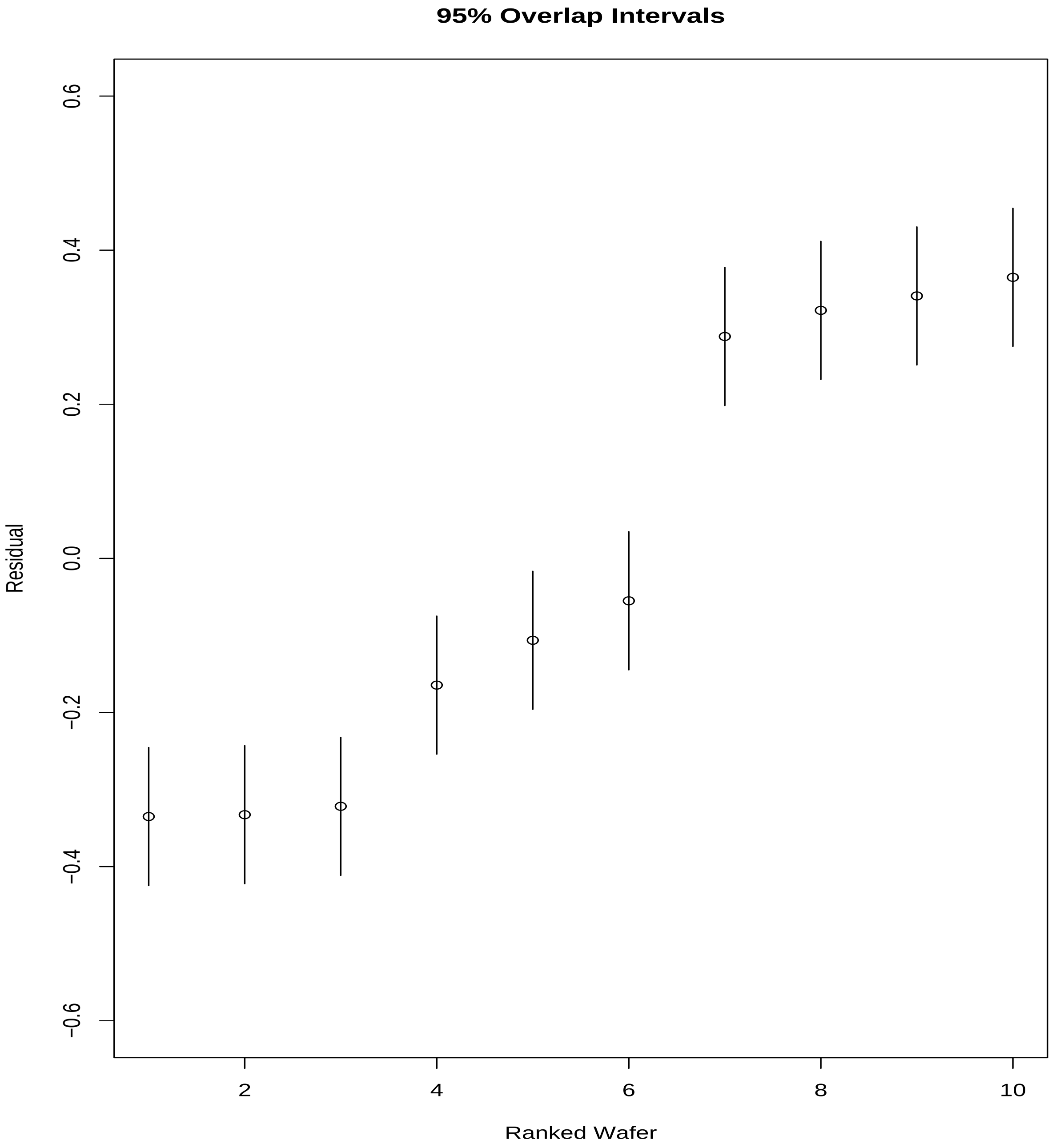

Figure 3: The modified graphical method of Goldstein and Healy (1995) applied to the Wafer data of Pinheiro and Bates (2000), accounting for the covariances of the level-2 residual estimates. There are 45 pairs of intervals, out of which 30 do not overlap.

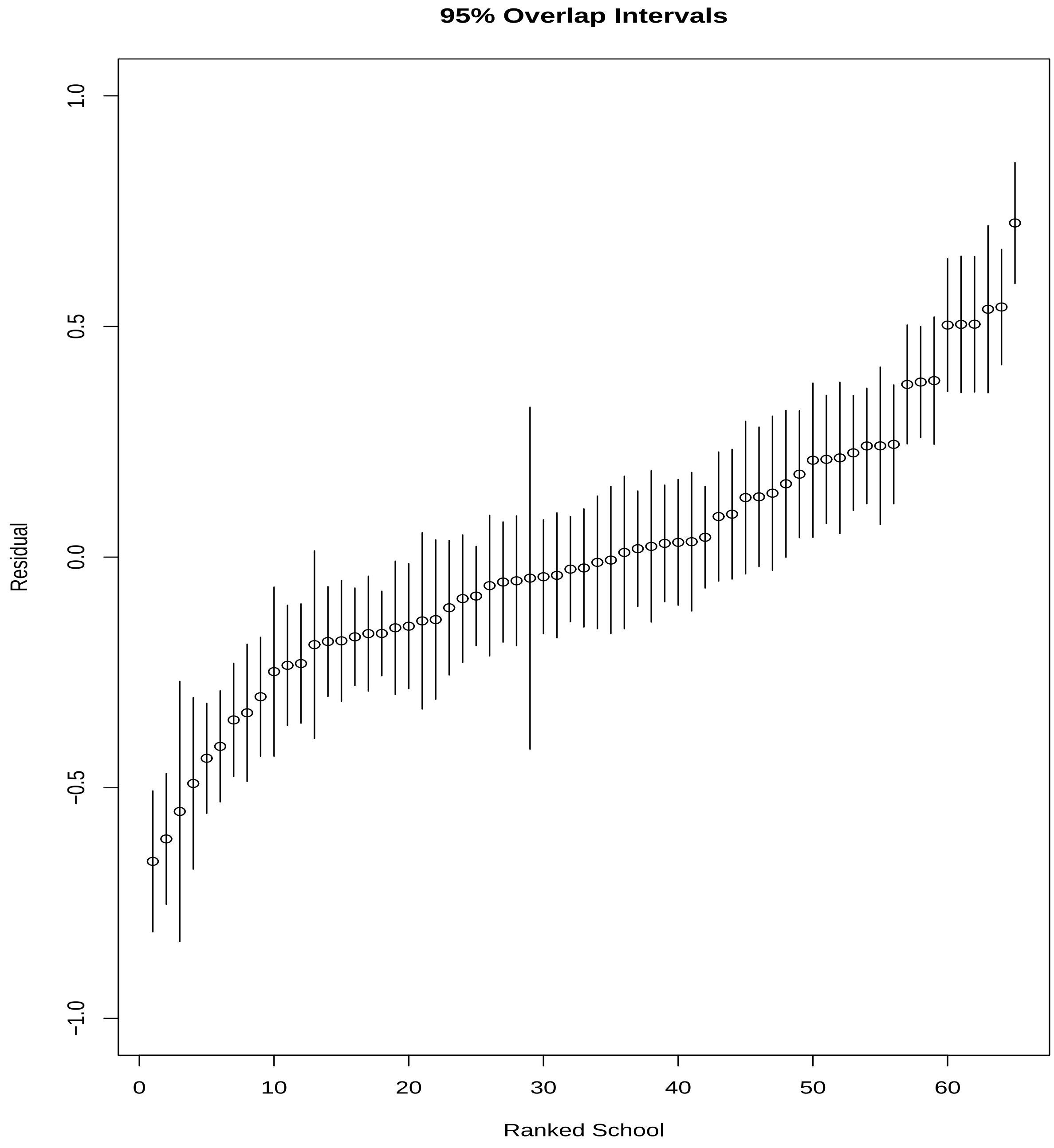

Figure 4: The graphical method of Goldstein and Healy (1995) applied to the LEA data of Rasbash et al. (2004), falsely assuming independence of the level-2 residual estimates. There are 2,016 pairs of intervals, out of which 977 do not overlap.

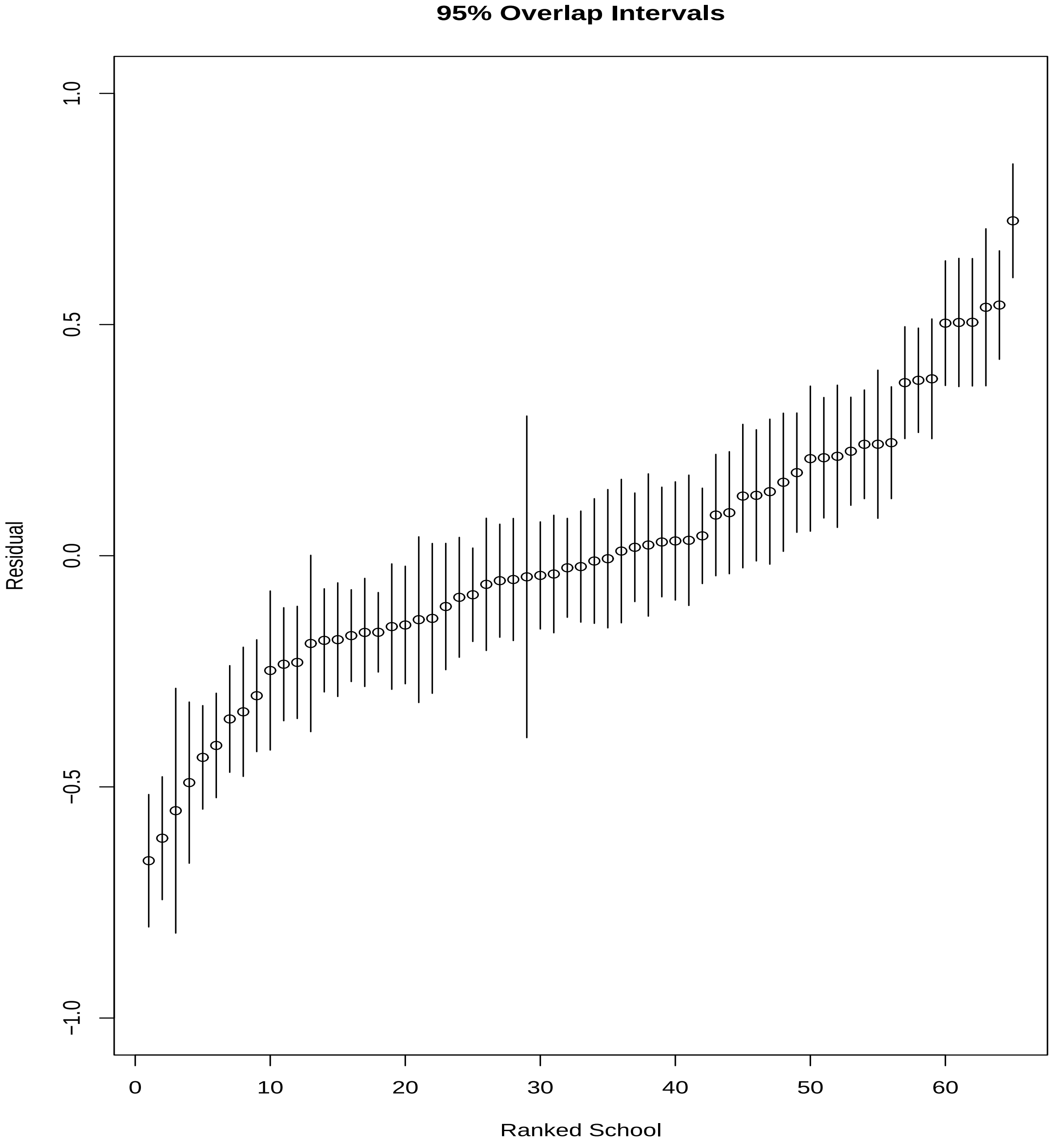

Figure 5: The modified graphical method of Goldstein and Healy (1995) applied to the LEA data of Rasbash et al. (2004), accounting for the covariances of the level-2 residual estimates. There are 2,016 pairs of intervals, out of which 1,031 do not overlap.

Table 1: Number of rejected hypotheses for various applications and methods.

LEA data, absolute comparisons, $S = 65$	
StepM	17
FDP-StepM	27
FDR	27
Naive	28
NELS data, absolute comparisons, $S = 981$	
StepM	38
FDP-StepM	249
FDR	244
Naive	289
Wafer data, pairwise comparisons, $S = 45$	
StepM	26
FDP-StepM	30
FDR	32
Naive	30
LEA data, pairwise comparisons, $S = 2,016$	
StepM	348
FDP-StepM	1,066
FDR	1,026
Naive	1,027