J. R. Statist. Soc. A (2007)
170, *Part* 4, *pp.* 1035–1059

Avoiding 'data snooping' in multilevel and mixed effects models

David Afshartous

University of Miami, Coral Gables, USA

and Michael Wolf

University of Zurich, Switzerland

[Received January 2006. Final revision January 2007]

Summary. Multilevel or mixed effects models are commonly applied to hierarchical data. The level 2 residuals, which are 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. Inference on the level 2 residuals, however, typically does not account for 'data snooping', i.e. 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: inference on the 'absolute' level 2 residuals to determine which are significantly different from 0, and 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 that is invoked, the user may choose the desired estimation method accordingly. We demonstrate the methods with the London education authority data, the wafer data and the National Educational Longitudinal Study data.

Keywords: Data snooping; Hierarchical linear models; Hypothesis testing; Pairwise comparisons; Random effects; Rankings

1. Introduction

Multilevel modelling is a popular statistical method for analysing hierarchical data. As such data are 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 modelling 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 on 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 model explicitly the relationship between information at different levels and improved estimation and prediction via the seminal statistical principle 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).

Address for correspondence: Michael Wolf, Institute for Empirical Research in Economics, University of Zurich, CH-8006 Zurich, Switzerland.
E-mail: mwolf@iew.uzh.ch

 0964–1998/07/1701035

Formally, say that we have an outcome measure y_{ij} for the ith observation in the jth group, e.g. the ith student (level 1) in the jth school (level 2). The sample size in the jth group is n_j and there are a total of J groups. The simplest multilevel model is a two-level variance components model:

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

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

Here the usual assumptions are that each of the u_j and ε_{ij} are sets of independent and identically distributed random variables with mean 0 and unknown variances σ_u^2 and σ_ε^2 respectively, and $\mathrm{cov}(\varepsilon_{ij}, u_j) = 0$. Often the distributions are assumed normal. However, we do not want to make any distributional assumptions in this paper, as they might be violated in practice.

Substituting for β_{0j}, we have

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

This is also recognizable as a random-effects analysis of variance. Covariate information can be introduced at both the individual and the group level to create a more general multilevel model. The general model in matrix notation is, for $j = 1, \ldots, J$,

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\gamma} + \mathbf{Z}_j\mathbf{u}_j + \boldsymbol{\varepsilon}_j, \tag{4}$$

where $\mathbf{Y}_j$ and $\boldsymbol{\varepsilon}_j$ are column vectors of length n_j, $\mathbf{X}_j$ is of dimension $n_j \times p$ and $\mathbf{Z}_j$ is of dimension $n_j \times q$. It might be noted that some of the level 1 variables in $\mathbf{Z}_j$ are often a subset of variables in $\mathbf{X}_j$ and represent random regression coefficients varying over groups; $\mathbf{X}_j$ may contain level 2 variables and cross-level interaction terms. The distribution of the level 1 errors can have the same assumptions as those considered in the basic model (1). The level 2 random error q-column vector $\mathbf{u}_j$ is usually assumed to be distributed multivariate $N(\mathbf{0}, \mathbf{D})$. The elements of $\mathbf{u}_j$, expressing the residual variability between level 2 units, may covary. Hence the matrix $\mathbf{D}$ is not generally diagonal. However, as stated earlier, we do not want to make any distributional assumptions in this paper. $\boldsymbol{\gamma}$ is a p-vector that includes the level 2 coefficients or fixed effects. In multilevel modelling, the first term of equation (4) is usually referred to as the fixed part of the model, and $\mathbf{Z}_j\mathbf{u}_j + \mathbf{e}_j$ is the random part. As equation (3) has no covariates, it would have $\mathbf{X}_j$ and $\mathbf{Z}_j$ as n_j-column vectors of 1s. For a model with a single covariate with non-random slope and random intercept, $\mathbf{X}_j$ would be $n_j \times 2$ with the first column consisting of 1s. An example of this is where the response $\mathbf{Y}_j$ is an educational outcome and $\mathbf{X}_j$ has observations on a prior ability control covariate that is introduced for adjusting the outcome. We briefly refer to such an example below.

In this paper, we focus on inference for the random effects, i.e. the level 2 residuals. 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. In applied research, it is 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) argued against such a simplistic use of rankings with respect to educational league tables in the UK. Instead, they strongly advocated the inclusion of confidence bands to reflect the uncertainty in the level 2 residual estimates.

Fig. 1 is reproduced from Rasbash *et al.* (2004), page 39, where a variance components model is fitted to school achievement data in a sample of 65 schools in six inner London local education

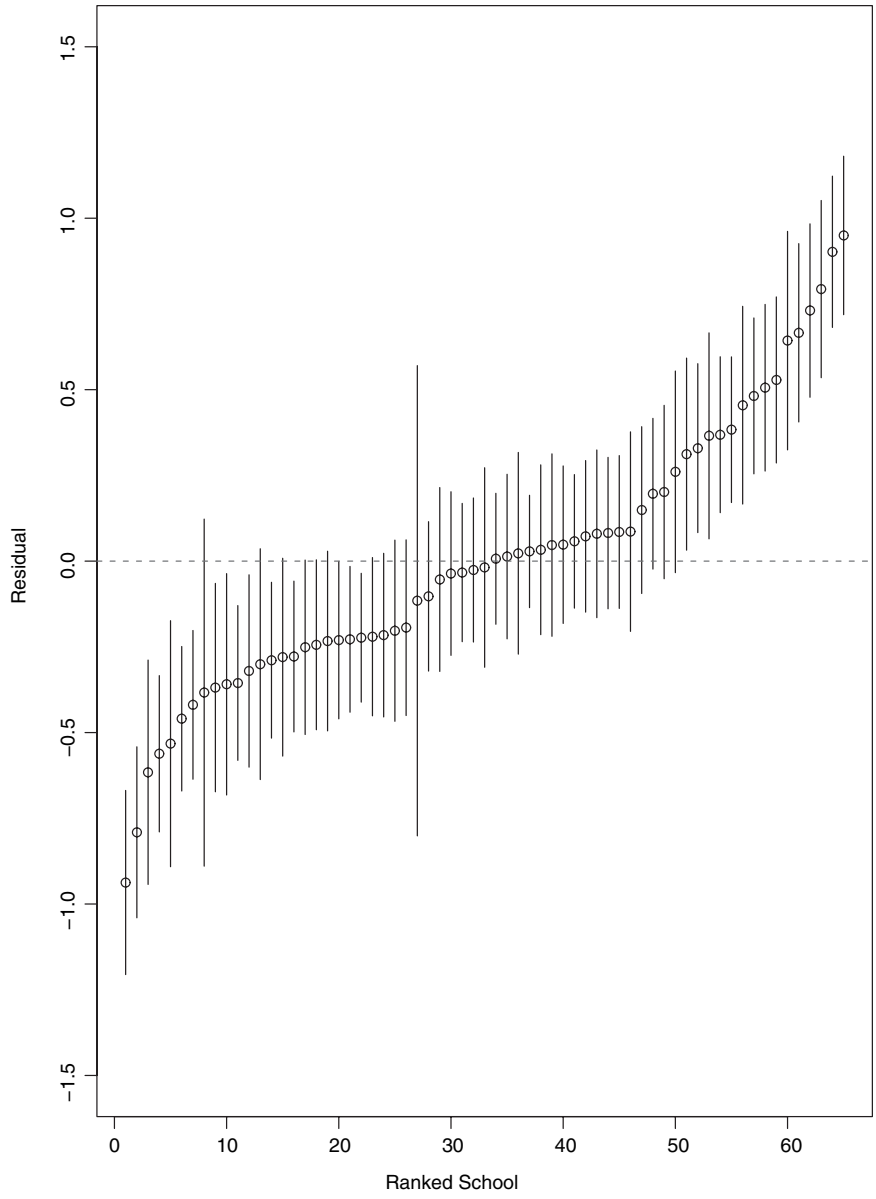

Fig. 1. 'Caterpillar plot' of Rasbash *et al.* (2004), page 39: the level 2 residuals of the 65 schools in ascending order together with their respective 95% confidence intervals

authorities. The response variable is the score that is achieved by 16-year-old students in an examination and the model does not adjust for prior ability measure. With the inclusion of the confidence bands, it becomes difficult to infer the rankings of the true unknown random effects u_j. Nevertheless, it is likely that Fig. 1 will be used by different individuals for different purposes; for instance, a school head may want to know whether his or her particular school differs from the average. Others may be interested in questions that involve inference on several schools at the same time; for instance, a governing board may be interested in finding out which are the schools that differ from the average. In such instances, joint testing of several hypotheses at

the same time occurs. We continue by briefly talking about the general problem of joint testing of several related hypotheses. Afterwards, we shall discuss how this problem comes up when making inference for random effects in multilevel models. Although the discussion is mainly with respect to the intercept-only variance components model, the arguments might just as well be made for 'adjusted outcomes' which have controlled for an intake measure.

Whenever several hypotheses are tested at the same time, a *multiple-testing* scenario arises. The challenge then becomes how to decide which hypotheses to reject, accounting for the multitude of tests. The naïve, yet common, approach of basing the decisions on the individual p-values or the individual confidence intervals, ignoring the multitude of tests, will typically result in a liberal analysis. Take the example of carrying out $S = 100$ hypotheses tests at the same time. The naïve approach rejects an individual null hypothesis if the corresponding p-value is less than or equal to $\alpha = 0.05$, say. Then, for any given true null hypothesis, the chance of a false rejection is equal to 0.05. But the 'combined' chance of making any false rejection at all can be much greater than that. To illustrate, assume that all 100 null hypotheses are true and that the individual test results are independent of each other. Then the expected number of false rejections is $100 \times 0.05 = 5$ and the chance of making at least one false rejection is $1 - 0.95^{100} = 0.994$. The naïve approach of basing the decisions on the individual p-values or the individual confidence intervals, ignoring the multitude of tests, is commonly referred to as *data snooping*.

The classical approach to multiple testing is to control the probability of making at least one false rejection. This probability is called the *familywise error rate* FWE. For example, the well-known Bonferroni method controls familywise error at joint level α by comparing the individual p-values with α/S, where S is the number of hypotheses that are tested at the same time; for example, see Lehmann and Romano (2005a), section 9.1. So, in the above example, an individual p-value would have to be less than or equal to $0.05/100 = 0.0005$ so that the corresponding null hypothesis could be rejected. Intuitively, what happens is that the bar for any individual analysis must be raised higher so that the overall probability of making a false rejection is controlled.

Of course, safeguards against false rejections are not the only concern of multiple-testing procedures. Corresponding to the power of a single test, we must also consider the ability of a procedure to detect false hypotheses, i.e. to make true rejections. In this sense, the Bonferroni method is suboptimal. There are alternative, albeit more complex, methods that also control FWE but often detect more false hypotheses; details are given in Section 3.

Another issue is that, if the number of hypotheses tested at the same time is very large, then FWE can be overly strict. In other words, by controlling the probability of making even one false rejection, the bar for any individual analysis can be raised so high that it becomes very difficult to make true rejections. For example, in the Bonferroni method the cut-off for the individual p-values, α/S, declines rapidly as the number of tests, S, increases. Hence, when many hypotheses are under test, we might be willing to tolerate a small number of false rejections if there is a large number of total rejections. In other words, we might be willing to tolerate a certain (small) proportion of false rejections out of the total rejections. This proportion is called the *false discovery proportion* FDP. By controlling FDP instead of FWE, often a substantial gain in power can be achieved; details are given in Section 3.

When making inference for random effects in multilevel models, there are two broad possibilities for multiple-testing scenarios to arise. First, there is the problem of absolute comparisons, i.e. one investigates which random effects are significantly different from 0, thereby identifying the groups which are particularly 'good' (above 0) or 'bad' (below 0). If one does not account for data snooping, then it can be quite likely that some groups will be either falsely identified as good or falsely identified as bad.

One might argue that if we are doing hypothesis tests for specific groups we should be employing a fixed effects rather than a random-effects model. However, these two aspects are independent of each other, i.e. we may specify a random-effects model and still be interested in specific random effects. Indeed, in addition to the above educational statistics example there is a strong and well-developed tradition of ranking of random effects in the context of animal breeding (Searle *et al*., 1992).

1.1. Example 1 (data snooping when making absolute comparisons)

Rasbash *et al*. (2004), page 39, lined up confidence intervals for the level 2 residuals, which correspond to schools, with *individual* coverage probability of 95% in a so-called caterpillar plot; this plot is reproduced in our Fig. 1.

What this plot allows us to do is to make inference on a *single* school that was chosen *a priori*. For example, a parent who considers sending their child to a particular school might wonder whether this school is different from the average. She can answer this question at the 5% level then by checking whether the confidence interval for the level 2 residual corresponding to this school contains 0 or not. However, the caterpillar plot cannot be used to make inference on several schools and/or on schools that are determined by looking at the plot first (e.g. exactly those schools whose confidence intervals do not contain 0). For example, Rasbash *et al*. (2004), page 39, state:

> 'Looking at the confidence intervals around them, we can see a group of about 20 schools at the lower and upper end of the plot where the confidence intervals for their [level 2] residuals do not overlap zero. 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.'

Although the main intention of Rasbash *et al*. (2004) was to use confidence bands to put uncertainty into perspective, the reader must be careful to avoid the pitfalls of data snooping. If we want to claim at level 5% that the entire group of about 20 schools that are identified in the above manner is different from the average, the confidence intervals should have been constructed in such a way that the *joint* coverage probability was given by 95%. In such a case we would typically obtain fewer rejections, as the intervals would naturally be wider.

Second, we may be interested in the set of all pairwise comparisons, where each group is compared with every other group. Again, if we do not account for data snooping, then it can be quite likely that some pairs will be falsely declared as different. In both instances, such false decisions are clearly worrisome, especially if they constitute the basis for policy making, as often happens in the evaluations of schools.

1.2. Example 2 (data snooping when making pairwise comparisons)

Fig. 2 in section 4.2 of Goldstein and Spiegelhalter (1996) presented school intercept residual estimates and their 95% overlap intervals based on the method of Goldstein and Healy (1995). Under their method, the average type I error over all pairwise comparisons should be 0.05. The intervals are constructed in such a way that, for a *single prespecified* comparison of two residuals, the two can be distinguished (i.e. declared significantly different) if their corresponding intervals do not overlap. Goldstein and Spiegelhalter (1996) concluded that, for example, the school with the smallest estimated residual can be distinguished from each of the highest six schools. We should be careful not to take this as a multiple comparison arising from the data themselves, but rather as separate comparisons of particular pairs of schools, as say would be done by six separate parents who had chosen their pairs of interest *a priori*. If we attempt the

former instead, these are *multiple, data-dependent* comparisons and so the method of Goldstein and Healy (1995) does not apply. We emphasize that the method itself is a correct method, but, when it is applied to inference problems for which it was not designed, misleading analyses can arise. In this light, it is important to point out an unfortunate mistake in Goldstein and Spiegelhalter (1996), page 395, concerning the use of the method of Goldstein and Healy (1995):

> 'We also note ... that where more than two institutions are compared, [overlap] diagrams such as Figs 2 and 3 present a conservative picture as they are designed only for pairwise comparisons'.

Instead of '*conservative* picture' it ought to be '*liberal* picture'.

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 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 non-specified 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. (Note that for our case u_j is a scalar and will only be boldface when discussing the general case.)

2.1. Absolute comparisons

In the population, u_j from equation (3) is distributed with mean 0 and unknown variance σ_u^2. As $j = 1, \ldots, J$, we would like to know the values of the J realizations u_j. Instead, given the data, we have J estimates $\hat{u}_j$. The first problem of interest is to test whether the value of each u_j is significantly different from 0. Formally, for each j, we are testing

$$H_j : u_j = 0 \quad \textit{versus} \quad H_j' : u_j \neq 0.$$

To illustrate the inference problem that we are interested in, consider again the caterpillar plot of Rasbash *et al.* (2004), page 39, which is reproduced in Fig. 1. On the one hand, there may be a school head, say of school 4, who wants to know whether his particular school differs from the average. In this case, he or she is interested only in the single comparison

$$H_4 : u_4 = 0 \quad \textit{versus} \quad H_4' : u_4 \neq 0.$$

An examination of the (uncorrected) caterpillar plot is entirely appropriate and allows him or her to decide whether to reject hypothesis H_4 or not. On the other hand, a school governing board examining all the schools in the district may want to know which schools differ from the average. In this case, the board must consider all hypotheses H_j simultaneously. An examination of the (uncorrected) caterpillar plot is no longer appropriate, as, owing to data snooping, typically too many schools will be identified as different from the average.

We should also like to stress that it would not be useful to test the 'global' hypothesis

$$H : u_1 = \ldots = u_J = 0 \quad \textit{versus} \quad H' : \text{some } u_j \neq 0.$$

If the global null H is rejected, we do not necessarily know which are the non-zero u_j. Therefore, our methodology focuses simultaneously on the individual nulls H_j so that the non-zero u_j can be identified.

2.2. Pairwise comparisons

The next problem of interest concerns making pairwise comparisons. We shall restrict attention to the two most common scenarios: all pairwise comparisons and comparing one residual with all the others.

Formally, we are testing

$$H_{j,k}: u_j = u_k \quad \textit{versus} \quad H'_{j,k}: u_j \neq u_k.$$

When all pairwise comparisons are considered, then the index set is $\{(j,k): 1 \leqslant j < k \leqslant J\}$ and there are a total of $\binom{J}{2}$ comparisons. When one residual is compared with all others, then the index set is $\{(j,k): 1 \leqslant k \leqslant J, k \neq j\}$ and there are a total of $J-1$ comparisons.

Of course, other hypotheses are also possible, such as comparing each residual in a subset of $\{1,\ldots,J\}$ with each residual in another (disjoint) subset of $\{1,\ldots,J\}$; the details are left to the reader.

2.3. Estimation

Various estimation methods exist for multilevel and mixed models; they manifest themselves in various software packages as well. 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 methods (Browne, 2003). 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 estimated standard error, i.e. $\hat{\sigma}(\hat{u}_j)$ estimates the unknown standard deviation of $\hat{u}_j$. Finally, given a pair of estimated residuals $\hat{u}_j$ and $\hat{u}_k$, let $\widehat{\mathrm{cov}}(\hat{u}_j, \hat{u}_k)$ represent the corresponding estimated covariance between $\hat{u}_j$ and $\hat{u}_k$. 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 0. For the classic mixed effects model format of equation (4), Laird and Ware (1982) and Robinson (1991) provided full details on the random-effects estimator and the corresponding estimated standard error. Below we summarize the general development. Assuming that $\mathbf{\Omega}_j = \mathrm{cov}(\mathbf{Y}_j) = \sigma^2\mathbf{I} + \mathbf{Z}_j\mathbf{D}\mathbf{Z}'_j$ is known, the fixed effects and their variances may be estimated by the standard generalized least squares estimators (Laird and Ware, 1982). Of course, in practice $\mathbf{\Omega}_j$ is unknown and must be estimated; there are 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):

$$\hat{\mathbf{u}}_j = \hat{\mathbf{D}}\mathbf{Z}'_j\hat{\mathbf{W}}_j(\mathbf{Y}_j - \mathbf{X}_j\hat{\gamma}) \tag{5}$$

and

$$\widehat{\mathrm{var}}(\hat{\mathbf{u}}_j - \mathbf{u}_j) = \hat{\mathbf{D}} - \hat{\mathbf{D}}\mathbf{Z}'_j\hat{\mathbf{W}}_j\mathbf{Z}_j\hat{\mathbf{D}} + \hat{\mathbf{D}}\mathbf{Z}'_j\hat{\mathbf{W}}_j\mathbf{X}_j\Big(\sum_i \mathbf{X}'_i\hat{\mathbf{W}}_i\mathbf{X}_i\Big)^{-1}\mathbf{X}'_j\hat{\mathbf{W}}_j\mathbf{Z}_j\hat{\mathbf{D}}, \tag{6}$$

where $\hat{\mathbf{W}}_j = \hat{\boldsymbol{\Omega}}_j^{-1}$. The set of estimates in $\hat{\mathbf{u}}_j$ will not be independent of a set of estimates $\hat{\mathbf{u}}_k$ for a different unit k since they are subject to the same sampling error in estimating γ and $\mathbf{D}$. The covariance of estimates from different units may be obtained by extracting the appropriate elements from the general variance–covariance matrix for the entire data (Goldstein, 2003):

$$E[(\hat{\mathbf{u}} - \mathbf{u})(\hat{\mathbf{u}} - \mathbf{u})'] = \mathbf{S} - \mathbf{R}'\boldsymbol{\Omega}^{-1}(\boldsymbol{\Omega} - \mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\Omega}^{-1}\mathbf{R}, \tag{7}$$

where $\mathbf{u}$ is the complete vector of random effects for all groups, $\mathbf{S}$ is block diagonal with blocks $\mathbf{D}$, $\mathbf{R}$ is block diagonal with blocks $\mathbf{Z}_j\mathbf{D}$, $\boldsymbol{\Omega}$ is block diagonal with blocks $\boldsymbol{\Omega}_j$ and $\mathbf{X}$ is obtained by stacking the $\mathbf{X}_j$. Estimates of the model parameters in equation (7) may then be substituted to provide the estimates of these covariances.

As stated earlier, the random-effects estimator $\hat{\mathbf{u}}_j$ is a shrinkage estimator, a linear transformation of the ordinary residuals $\mathbf{Y}_j - \mathbf{X}_j\hat{\gamma}$. It may be viewed as a weighted combination of 0 and $\bar{\mathbf{u}}_j$, where the latter is the ordinary least squares estimate that is obtained by treating $\mathbf{u}_j$ as a fixed effect (Laird and Ware, 1982).

2.4. Bootstrap method

There are several variants of the bootstrap for multilevel models, and they may be divided into three basic categories:

(a) parametric bootstrap,
(b) residual bootstrap and
(c) cases bootstrap.

Categories (b) and (c) are both variants of the non-parametric 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 0); see Goldstein (2003), section 3.5, and van der Leeden *et al.* (2007). The residual bootstrap generates new data by keeping the explanatory variables fixed and resampling the (centred) estimated level 1 and level 2 residuals; see Carpenter *et al.* (2003) and van der Leeden *et al.* (2007). The cases bootstrap generates new data by resampling entire 'cases' of response variables jointly 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.* (2007).

The crucial difference between categories (a) and (b), on the one hand, and (c), on the other hand, is as follows: by design, the level 2 residuals in the bootstrap world are resampled from a distribution with mean 0. As a result, they are random rather than identical to the level 2 residuals in the real world. In contrast, consider the cases bootstrap (c) when only the level 1 units are resampled but the level 2 units and their unit-specific variables remain fixed. In this way, the level 2 residuals in the bootstrap world are identical to those in the real world.

Our multiple-testing procedure is based on the construction of simultaneous confidence intervals, similar to the caterpillar plot of Rasbash *et al.* (2004), page 39, but with a *joint* coverage probability of 95%. When the bootstrap is used to construct confidence intervals, then it is necessary that the particular bootstrap method does *not* impose the constraints of the individual null hypotheses. The null hypotheses of interest in this paper involve the level 2 residuals, so we need a bootstrap method that 'inherits' exactly these level 2 residuals from the real world. Therefore, we employ the cases bootstrap (c) that resamples the level 1 units only. To provide some intuition, if the particular bootstrap method imposed the constraints of the individual null hypotheses, then all level 2 residuals in the bootstrap world would be equal to 0. As a result,

the bootstrap confidence intervals would tend to contain 0 rather than the true level 2 residuals. So, if a particular level 2 residual was indeed different from 0, we would have no power to detect this; more details are given in Section 3.2.

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 by the same estimation method from the bth bootstrap sample and let $\hat{\sigma}(\hat{u}_j^{*,b})$ denote its estimated standard error. The chains $\{\hat{u}_1^{*,b}, \ldots, \hat{u}_J^{*,b}\}$ and $\{\hat{\sigma}(\hat{u}_1^{*,b}), \ldots, \hat{\sigma}(\hat{u}_J^{*,b})\}$ are then used in the stepwise multiple-testing procedure that is described below. This provides the researcher with the option of employing his or her preferred estimation procedure.

3. Avoiding data snooping

As discussed in Section 1, much of the current practice for inference on level 2 residuals suffers from data snooping. In this section, we present novel stepwise multiple-testing methods to address this shortcoming.

3.1. Problem formulation

We proceed by presenting a unified framework in which a general multiple-testing problem can be formulated and addressed. This unified framework allows for a concise presentation of our multiple-testing methods later on, rather than having to develop the methods several times from scratch for each application.

The unknown probability distribution generating the data is denoted by P. Interest focuses on a parameter vector $\theta = \theta(P)$ of dimension S, i.e. $\theta = (\theta_1, \ldots, \theta_S)'$. The individual hypotheses are about the elements of θ and of the form

$$H_s : \theta_s = 0 \quad \textit{versus} \quad H_s' : \theta_s \neq 0. \tag{8}$$

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. In contrast with data snooping, the decisions are not based on the individual p-values or confidence intervals, ignoring the fact that S tests are carried out at the same time.

3.1.1. Example 3 (absolute comparisons)

If the values of the level 2 residuals u_j are under test, then $S = J$ and $\theta_s = u_s$.

3.1.2. Example 4 (pairwise comparisons)

If all pairwise comparisons of the level 2 residuals are of interest, then $S = \binom{J}{2}$; if we compare one residual with all others, then $S = J - 1$. In either case, an element θ_s is of the form $\theta_s = u_j - u_k$, where s can be taken as referring to the ordered pair (j, k).

3.2. Problem solution based on the familywise error rate

The traditional approach to account for the multitude of tests is to control the *familywise error rate* FWE, which is defined as the probability of making at least one false rejection:

$$\mathrm{FWE}_P = P(\text{reject at least one } H_s : \theta_s = 0).$$

The subscript P in FWE_P makes it clear that the FWE in any particular application depends on the underlying probability model P generating the data.

If FWE is controlled at level α, then the probability that at least one true null hypothesis will be rejected is less than or equal to α. Hence, we can be $1-\alpha$ confident that in a particular application no true null hypotheses have been rejected. In other words, we can be $1-\alpha$ confident that all rejected hypotheses are indeed false. However, if the individual tests have each level α, then the confidence that all rejected hypotheses are indeed false, which is also called the *joint confidence*, is generally less than $1-\alpha$, and potentially much less. To be more precise, 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 $(1-\alpha)^S$. In other words, the joint confidence is smaller than the individual confidence.

Strictly speaking, a multiple-testing procedure controls FWE if

$$\text{FWE}_P \leqslant \alpha \qquad \text{for all sample sizes } (n_1, \ldots, n_J) \text{ and for all } P.$$

However, this is only feasible in special circumstances, usually involving strong distributional assumptions. Since we do not want to make any distributional assumptions, we focus instead on asymptotic control of FWE defined as

$$\limsup_{\min_{1 \leqslant j \leqslant J}(n_j) \to \infty} (\text{FWE}_P) \leqslant \alpha \qquad \text{for all } P.$$

In words, we achieve control of the maximum FWE as the smallest n_j increases. In the remainder of the paper, when we speak of control of FWE—and later of alternative criteria to account for data snooping—we always mean asymptotic control.

Traditional methods to control FWE are based on the 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 at the (joint) level α if $\hat{p}_s \leqslant \alpha/S$. It is a *single-step* method, since all p-values are compared with the same critical value. Its advantage is its simplicity, but it can result in low power, as will now be explained.

First, a perhaps less-well-known improvement is the method of Holm (1979). The p-values are ordered from smallest to largest: $\hat{p}_{(1)} \leqslant \hat{p}_{(2)} \leqslant \ldots \leqslant \hat{p}_{(S)}$. Then the hypothesis $H_{(s)}$ is rejected if $\hat{p}_{(j)} \leqslant \alpha/(S-j+1)$ for all $j = 1, \ldots, s$. This is a *stepwise* method, since the p-values are compared with increasingly larger critical values. In contrast, the Bonferroni method is a *single-step* method, since all p-values are compared with the same critical value α/S. Since the first critical value for the Holm method is equal to the (common) critical value of the Bonferroni method, but all subsequent critical values are larger, it follows that all hypotheses that are rejected by the Bonferroni method will also be rejected by the Holm method, but it can happen that the Holm method will reject some further hypotheses in addition. Therefore, the Holm method is more powerful than the Bonferroni method.

Second, and nevertheless, even the Holm method can be quite conservative. It shares with the Bonferroni method the disadvantage of being based on the individual p-values. Therefore, to guarantee control of FWE in general, these methods must assume a 'worst case' dependence structure of the test statistics. If the true dependence structure could be taken into account, power would increase. To give an extreme example, if all test statistics are perfectly correlated with each other, then the single-step critical value can be increased to α compared with the Bonferroni worst case critical value of α/S.

Romano and Wolf (2005) developed a novel stepwise multiple-testing procedure that accounts for the dependence structure of the test statistics and therefore is more powerful than the Holm method. Their framework is that of comparing several strategies (such as investment strategies) with a common bench-mark (such as a market index) and deciding which strategies

outperform the bench-mark. Given this context, the individual tests are one sided. We therefore now detail how the procedure of Romano and Wolf (2005) must be modified when the individual tests are two sided, which is the case for the applications that 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 an estimate of the standard error of w_s.

3.2.1. Example 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.)

3.2.2. Example 4 continued (pairwise comparisons)
We have $w_s = \hat{u}_j - \hat{u}_k$ and $\hat{\sigma}_s = \sqrt{\{\hat{\sigma}^2(\hat{u}_j) + \hat{\sigma}^2(\hat{u}_k) - 2\,\widehat{\mathrm{cov}}(\hat{u}_j, \hat{u}_k)\}}$.

The modified method of Romano and Wolf (2005) starts out by sorting the test statistics from largest to smallest. Label r_1 corresponds to the largest test statistic and label r_S to the smallest, so $|z_{r_1}| \geqslant |z_{r_2}| \geqslant \ldots \geqslant |z_{r_S}|$. 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

$$[w_{r_1} \pm \hat{\sigma}_{r_1}\hat{d}_1] \times \ldots \times [w_{r_S} \pm \hat{\sigma}_{r_S}\hat{d}_1]. \tag{9}$$

Here, the common value $\hat{d}_1$ is chosen in such a fashion that the joint coverage by the region (9) is asymptotically equal to $1-\alpha$. In other words, the probability that all parameters $\theta_1, \ldots, \theta_S$ are contained in this region is asymptotically equal to $1-\alpha$. Of course, the task of finding such a value $\hat{d}_1$ is non-trivial and central to our proposed method.

Then, for $s = 1, \ldots, S$, the hypothesis H_{r_s} is rejected if 0 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 that are rejected in this first step. It should be clear from the order information that is contained in the labels r_s and the common multiplier $\hat{d}_1$ that the hypotheses that are rejected will then be H_{r_s} for $s = 1, \ldots, R_1$. 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

$$[w_{r_{R_1+1}} \pm \hat{\sigma}_{r_{R_1+1}}\hat{d}_2] \times \ldots \times [w_{r_S} \pm \hat{\sigma}_{r_S}\hat{d}_2]. \tag{10}$$

Then, for $s = R_1 + 1, \ldots, S$, the hypothesis H_{r_s} is rejected if 0 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 that are 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 $\hat{d}_1, \hat{d}_2, \ldots$. We start with the constant $\hat{d}_1$ and ask what its ideal value, called d_1, would be. In other words, which value would result in a finite sample joint coverage of exactly $1-\alpha$ by the region (9)? One can show that this ideal value is the $(1-\alpha)$-quantile of the sampling distribution under P of $\max_{1 \leqslant s \leqslant S}(|w_{r_s} - \theta_{r_s}|/\hat{\sigma}_{r_s})$, the maximum of the S centred statistics. The ideal constant, called d_2, in the second step is the $(1-\alpha)$-quantile of the sampling distribution under P of $\max_{R_1+1 \leqslant s \leqslant S}(|w_{r_s} - \theta_{r_s}|/\hat{\sigma}_{r_s})$, the maximum over only those statistics relating to hypotheses that were not rejected in the first step. As a result, $d_2 \leqslant d_1$ and it is possible that some hypotheses will be rejected in the second step that were not rejected in the first step; and so on for the subsequent steps.

The problem is that the ideal constants $d_1, d_2, \ldots$ depend on the unknown probability distribution P and are therefore not available. Instead, a bootstrap approach yields feasible constants: P is replaced by an estimator $\hat{P}$ and then the quantiles are computed under $\hat{P}$. The resulting bootstrap quantiles are then denoted $\hat{d}_1, \hat{d}_2, \ldots$ to make it clear that they come from an esti-

mated distribution. Our bootstrap approach is of asymptotic nature, since we resample from $\hat{P}$, which converges to the true P as the smallest sample size n_j increases, as opposed to from the true P. For details on how to compute the constants $\hat{d}_j$ in examples 3 and 4, see Appendix B. Importantly, as explained there, it holds also true for the bootstrap quantiles that $\hat{d}_2 \leqslant \hat{d}_1$, say, so further hypotheses can be rejected in subsequent steps.

We can now summarize our stepwise method by the following algorithm. The name StepM stands for 'stepwise multiple testing'.

3.2.3. Algorithm 1 (StepM method)

Step 1: relabel the hypotheses in descending order of the test statistics $|z_s|$; label r_1 corresponds to the largest test statistic and label r_S to the smallest.
Step 2: set $j = 1$ and $R_0 = 0$.
Step 3: for $R_{j-1} + 1 \leqslant s \leqslant S$, if $0 \notin [w_{r_s} \pm \hat{\sigma}_{r_s} \hat{d}_j]$, reject the null hypothesis H_{r_s}.
Step 4:

(a) if no (further) null hypotheses are rejected, stop;
(b) otherwise, denote by R_j the total number of hypotheses that have been rejected so far and, afterwards, let $j = j + 1$. Then return to step 3.

We now give some heuristics on why the StepM method provides asymptotic control of FWE. The formal proof for one-sided alternative hypotheses that is detailed in Romano and Wolf (2005) can be straightforwardly extended to two-sided hypotheses. Consider the asymptotic joint confidence region (9) in the first step. By construction, this region will contain the entire parameter vector with limiting probability $1 - \alpha$. In particular, the probability that at least one of the parameters $\theta_s = 0$ will not be contained in that region is asymptotically less than or equal to α. As a result, the first step asymptotically controls FWE. But even the 'full' stepwise method achieves this (while improving power). To see why, consider the second step. Assume that all rejections in the first step are true rejections; otherwise, the FWE criterion has been violated already and moving on to the second step can do no further damage. At this stage, we construct the asymptotic joint confidence region (10) for the remaining parameter vector $(\theta_{r_{R_1+1}}, \ldots, \theta_S)'$. By construction, this region will contain the remaining parameter vector with limiting probability $1 - \alpha$. In particular, the probability that at least one of the parameters $\theta_{r_s} = 0$, with $R_1 + 1 \leqslant s \leqslant S$, will not be contained in that region is asymptotically less than or equal to α: and so on.

The StepM method is a multiple-testing method based on the inversion of joint confidence regions, i.e., at any given stage, we construct a joint confidence region for the remaining parameter vector and then reject a particular null hypothesis $H_{r_s} : \theta_{r_s} = 0$ if the value 0 for θ_{r_s} is not contained in the joint confidence region. By the definition of this region, this happens if and only if 0 is not contained in the interval $[w_{r_s} \pm \hat{\sigma}_{r_s} \hat{d}_j]$; see step 3 of algorithm 1. Crucially, when the bootstrap is used to compute a confidence region, we must resample from an unrestricted estimate $\hat{P}$ of the underlying distribution P. Otherwise, when we resample from a restricted estimate of P, then the resulting region will tend to contain the null parameters instead of the true parameters (in our case the 0s instead of the true θ_s-values), so it no longer is a valid confidence region.

We briefly return to the motivating examples at the beginning of this section. Algorithm 1 applied to absolute comparisons with 0 would avoid the data snooping which the diagram in example 1 might unintentionally encourage. In particular, the joint confidence region (9) 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 0).

Algorithm 1 applied to pairwise comparisons would avoid the temptation of data snooping in situations such as example 2. Note that comparing the lowest school with the highest school(s) requires an adjustment for data snooping based on all $S = \binom{J}{2}$ pairwise comparisons, since, in each case, the two schools being compared have been selected in a data-dependent way (i.e. 'smallest' *versus* 'largest'). Unfortunately, in this example, the first step of our method cannot be translated into a convenient plot. Furthermore, if a particular school that had been selected *ex ante* was compared with some data-dependent schools (e.g. school 5 *versus* smallest or largest), then the method would require only an adjustment for data snooping based on all $S = J - 1$ comparisons of one school with all others.

3.3. Problem solution based on the false discovery proportion

As explained in Section 1, if the number of hypotheses under consideration, S, is very large, controlling FWE may be too strict. In such instances, we might be willing to tolerate a certain (small) proportion of false rejections out of the total rejections. This suggests basing error control on the false discovery proportion FDP. Let F be the number of false rejections that are made by a multiple-testing method and let R be the total number of rejections. Then FDP is defined as

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

By control of FDP, we mean control of the tail probability $P(\text{FDP} > \gamma)$ where $\gamma \in [0, 1)$ is a user-defined number. Similarly to Section 3.2, we shall focus on asymptotic control of FDP, i.e. we want $P(\text{FDP} > \gamma)$ bounded above by α as the smallest $n_j \to \infty$. Typical values of γ are $\gamma = 0.05$ and $\gamma = 0.1$; the choice $\gamma = 0$ corresponds to control of FWE.

Given the last observation, there are actually two alternative ways to improve the power of controlling FWE. To illustrate, consider controlling FWE at level $\alpha = 0.05$; this corresponds to controlling FDP at level $\alpha = 0.05$ and choosing $\gamma = 0$. The first possibility is simply to increase α, say to $\alpha = 0.1$, but to stick to control of FWE (i.e. to stick to $\gamma = 0$). The second possibility is to stick to $\alpha = 0.05$ but to switch to 'actual' FDP control by choosing a positive γ, say $\gamma = 0.1$. The two possibilities are philosophically different and comparing them is a little like comparing apples to oranges. Is it better to be 90% confident that all rejections are true rejections (i.e. that the realized FDP is 0) or is it better to be 95% confident that the realized FDP is at most 0.1? We do not know how to answer this question. However, we can say that, when the number of hypotheses that are under test is large, then the second possibility will typically reject more hypotheses. This will be illustrated with the empirical applications in Section 4.

Lehmann and Romano (2005b) proposed a stepwise method to control FDP that is based on the 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 test statistics. Romano and Wolf (2007) took such dependence into account in a bootstrap method which, again, we now need to extend to two-sided alternatives.

The goal is to control FDP in the end. However, it turns out that as a stepping-stone towards this we first need to control the *generalized familywise error rate* k-FWE, which is defined as the probability of making at least k false rejections, where $k \geqslant 1$ is a prespecified integer:

$$k\text{-FWE}_P = P(\text{reject at least } k \text{ of the } H_s : \theta_s = 0).$$

The method to control k-FWE is somewhat complex and so the details are deferred to Appendix A. For now, simply assume that we have a method to control k-FWE, which is called the k-StepM method. With this more general label the simpler 1-StepM method is what we have previously called simply the StepM method. We now describe how successive control of k-FWE, for increasing values of k, leads to control of FDP.

To develop the idea, consider controlling $P(\text{FDP} > 0.1)$. We start out by applying the 1-StepM method, i.e. by controlling FWE at level α. Denote by N_1 the number of hypotheses that are rejected. Owing to the FDP control, we can be confident that no false rejection has occurred and that, in turn, FDP has been controlled. Consider now rejecting $H_{r_{N_1+1}}$, the next 'most significant' hypothesis. Of course, if $H_{r_{N_1+1}}$ is false, there is nothing to worry about, so suppose that $H_{r_{N_1+1}}$ is true. Assuming FWE control in the first step, the FDP on rejection of $H_{r_{N_1+1}}$ then becomes $1/(N_1+1)$, which is greater than 0.1 if and only if $N_1 < 9$. So if $N_1 \geqslant 9$ we can reject one true hypothesis and still avoid $\text{FDP} > 0.1$. This suggests stopping if $N_1 < 9$ or otherwise applying the 2-StepM method at level α which, by design, should not reject more than one true hypotheses. (It does not matter at this point whether the 2-StepM method is applied to the full set of hypotheses or the remaining set of non-rejected hypotheses. The reason is that all hypotheses that are rejected by the StepM method will, by design, automatically be rejected by the 2-StepM method as well.) Denote the total number of hypotheses that are rejected by the 2-StepM method by N_2. Reasoning similarly to before, if $N_2 < 19$, we stop and otherwise we apply the 3-StepM method at level α. This procedure is continued until $N_j < 10j - 1$ at some point. The following algorithm summarizes the method for arbitrary γ.

3.3.1. Algorithm 2 (FDP–StepM method)

Step 1: let $k = 1$.
Step 2: apply the k-StepM method at level α and denote by N_k the number of hypotheses that are rejected.
Step 3:

(a) if $N_k < k/\gamma - 1$, stop;
(b) otherwise, let $k = k + 1$ and return to step 2.

3.4. Problem solution based on the false discovery rate

Benjamini and Hochberg (1995) proposed a stepwise method for controlling the expected value of FDP, $E(\text{FDP})$, which they coined the *false discovery rate* FDR. The idea is to ensure that $\text{FDR} \leqslant \gamma$, at least asymptotically, for some user-defined γ. The method is based on the individual p-values and works as follows.

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

$$j^* = \max\{j : \hat{p}_{(j)} \leqslant \gamma_j\}, \qquad \gamma_j = \frac{j}{S}\gamma, \tag{11}$$

and reject $H_{(1)}, \ldots, H_{(j^*)}$. If no such j^* exists, reject no hypotheses. This is an example of a *step-up* method. It starts with examining the least significant hypothesis, $H_{(S)}$, and then moves 'up' to the more significant hypotheses. Williams *et al.* (1999) endorsed 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 are certain sufficient conditions on this dependence structure,

but the scenario of all pairwise comparisons does not meet any of them. The underlying reasons are of rather technical nature; see Benjamini and Yekutieli (2001) and Yekutieli (2002). Yekutieli (2002) provided a more conservative FDR procedure that is shown to work for the scenario of all pairwise comparisons. Second, FDR is the expected value of FDP. Controlling the expected value says rather little about the actual realization of FDP in a given application. Indeed, the realized value of FDP could be quite far from the nominal upper bound γ on FDR; see Korn *et al*. (2004) for some simulation evidence.

To give an example, consider controlling FDR $\leqslant 0.1$. This does not allow us to make any informative statement about the realized FDP in a given application. (If we controlled FDR $\leqslant 0.1$ in a large number of independent applications, then, analogous to the central limit theorem, we could make certain claims concerning the average realized FDP over the many applications. However, most applied researchers will be interested in a particular single application only.) However, if we control $P(\text{FDP} > 0.1) \leqslant 0.05$, say, then we 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 that 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, we can be confident that indeed all rejected hypotheses are false hypotheses. For example, such a joint confidence may be very desirable in the context of policy making.

However, when the number of hypotheses that are under consideration is very large, controlling 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. This can be seen from the empirical applications in Section 4 as well as from the simulation study in Romano and Wolf (2007). 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 that we can be 95% confident that the realized FDP is at most 0.1. The FDR method, in contrast, only controls the expected value of FDP. So, in any given application, the realized FDP could be quite far from this expected value, say 0.1. This point is made clear in Korn *et al*. (2004). Therefore, controlling FDP can be considered 'safer' than controlling FDR.

Although, for these reasons, the (globally) most appropriate method does not exist, there clearly does exist an inappropriate method, namely data snooping by basing inference on the individual *p*-values or the individual confidence intervals, without taking the multitude of tests into account.

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 or her estimation method of choice. However, in practice the multiple-testing methods will have differing practical difficulties given different estimation methods; for example, using a Markov chain Monte Carlo estimation method in conjunction with bootstrapping may present difficulties with respect to time constraints. Random-effects models were estimated via the `nlme` package of Pinheiro and Bates (2000) which is contained in the statistical software R (R Project, 2006). The default estimation method in `nlme` is restricted maximum likelihood,

and this is the estimation method that we used; see Pinheiro and Bates (2000), chapter 2, for further details. R extensions were written for the estimated standard errors of the random-effects estimates, the covariances between random-effects estimates, and the bootstrapping of the data, as well as the StepM and FDP–StepM methods themselves. These are available at `http://moya.bus.miami.edu/~dafshartous/`.

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). An operationalization of k-StepM building-blocks for the FDP–StepM method requires a value of a user-set parameter $N_{\max}$ to be provided (see Appendix A). Here $N_{\max} = 100$ is used. The bootstrap procedures that are required for this, which were discussed previously in Section 2.4 and are elaborated in Appendix A, use $B = 1000$ repetitions, which are deemed sufficient for our purposes; see Efron and Tibshirani (1993), section 19.3.

4.1. Data snooping when making absolute comparisons

Consider the data set that was used in Rasbash *et al*. (2004) and was considered earlier in example 1. Here the response variable is the score that is achieved by 16-year-old students in an examination and the predictor is the London reading test score that was obtained by the same students just before they entered secondary school at the age of 11 years. The data are on 4059 students in 65 inner London schools. As in Rasbash *et al*. (2004), page 11, 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 with 0 is equivalent to examining whether the school's average examination score differs from the grand mean after accounting for London reading test score. If we simply compute 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 0. 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. The results are summarized in Table 1, which also summarizes the results of all subsequent empirical investigations that we shall make.

Consider the National Educational Longitudinal Study data set that was used by Afshartous and de Leeuw (2004), where the base year sample from the 1988 study is used (National Center for Educational Statistics, 2006). The base year sample consists of 24599 eighth-grade students, distributed among 1052 schools across the USA. The response variable is student mathematics score and the predictor is the socio-economic status of the student. As above, we fit a multilevel or random-effects model with random intercept and constant slope across schools. If we simply compute the separate test statistics for the random 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, again, 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.

4.2. Data snooping when making pairwise comparisons

Consider the wafer data that were presented in Pinheiro and Bates (2000). The data were collected to study the variability in the manufacturing of analogue metal oxide semiconductor circuits and consist 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 we simply examine the test statistics for the pairwise

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

Method	*Number of rejected hypotheses*
Local education authority data, absolute comparisons, $S = 65$	
StepM	17
FDP–StepM	27
FDR	27
Naïve	28
National Educational Longitudinal Study data, absolute comparisons, $S = 981$	
StepM	38
FDP–StepM	249
FDR	244
Naïve	289
Wafer data, pairwise comparisons, $S = 45$	
StepM	26
FDP–StepM	30
FDR	32
Naïve	30
Local education authority data, pairwise comparisons, $S = 2080$	
StepM	348
FDP–StepM	966
FDR	1026
Naïve	1027

differences of random effects and their corresponding p-values, 30 hypotheses are rejected. The application of the StepM, FDP–StepM and FDR methods yield 26, 30 and 32 rejections respectively. In this application, the FDR method rejects more hypotheses than the naïve method that does not account for data snooping. This can indeed happen when the γ-parameter of the FDR method is greater than the individual critical value α for the naïve method, as is the case here with $\gamma = 0.1$ and $\alpha = 0.05$. For the reader who wants to check: the ordered p-value numbers 30–33 are given by 0.0185, 0.0623, 0.0661 and 0.0856 respectively.

The graphical method of Goldstein and Healy (1995) can be interpreted as a 'visual short cut' to an analysis that is 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 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. In large sample applications this may not be unrealistic since it is tantamount to assuming that the fixed effects in the model are known or estimated precisely (see Goldstein (2003), appendix 2.2). However, for many applications 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 misleading analyses. If the method of Goldstein and Healy (1995) is applied to the wafer data of Pinheiro and Bates (2000), a total of 24 rejections are obtained. Obviously, it is counterintuitive that a method which does not account for data snooping should reject fewer hypotheses than

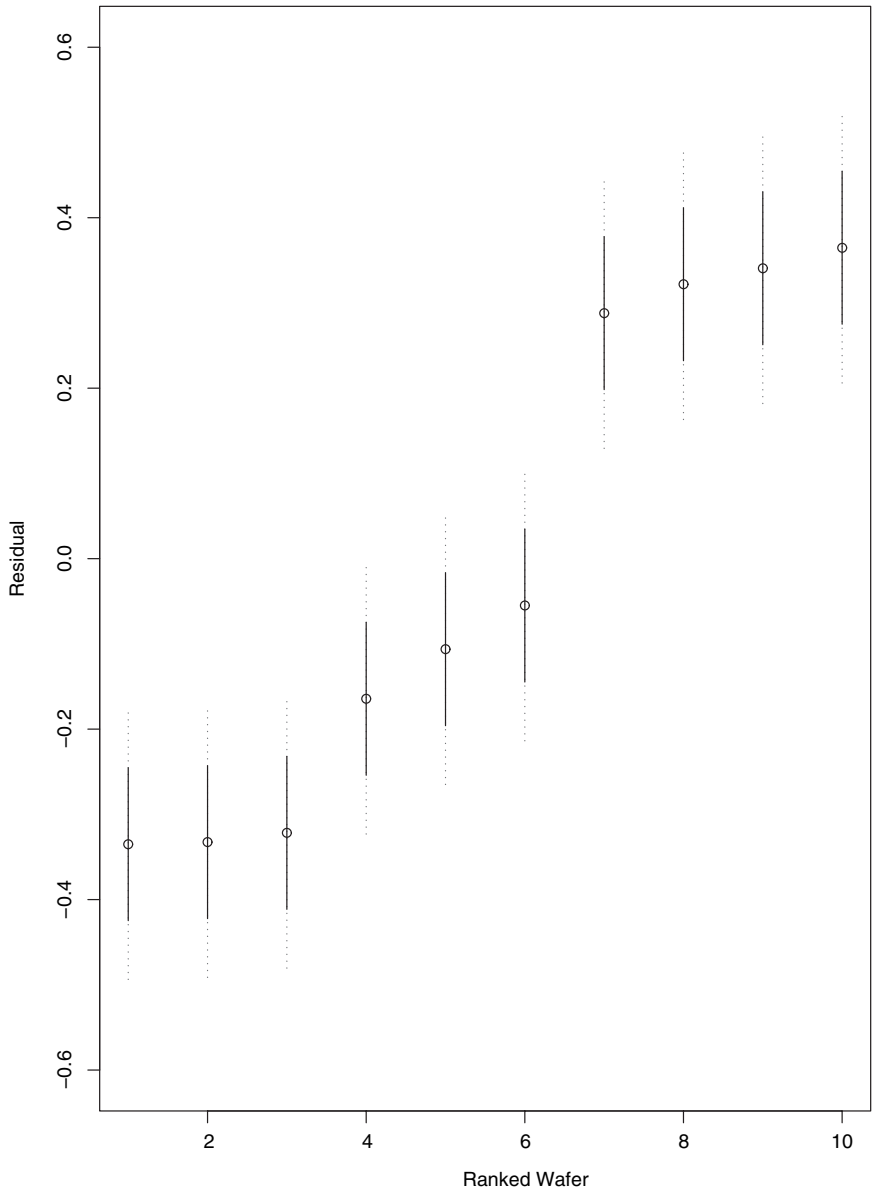

Fig. 2. Graphical method of Goldstein and Healy (1995) applied to the wafer data of Pinheiro and Bates (45 pairs of 95% intervals): |, accounting for the covariances of the level 2 residual estimates (in this case, 30 of the 45 pairs do not overlap); ⋮, falsely assuming independence (in this case, 24 of the 45 pairs do not overlap)

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; see Appendix C. Now the lengths of the intervals are reduced and a total of 30 rejections are obtained, the same number as for the above analysis based on individual p-values. This difference is illustrated in Fig. 2.

To be sure, we do not say that the method of Goldstein and Healy (1995) is incorrect. We only say that it can lead to misleading results when it is applied to situations for which it was

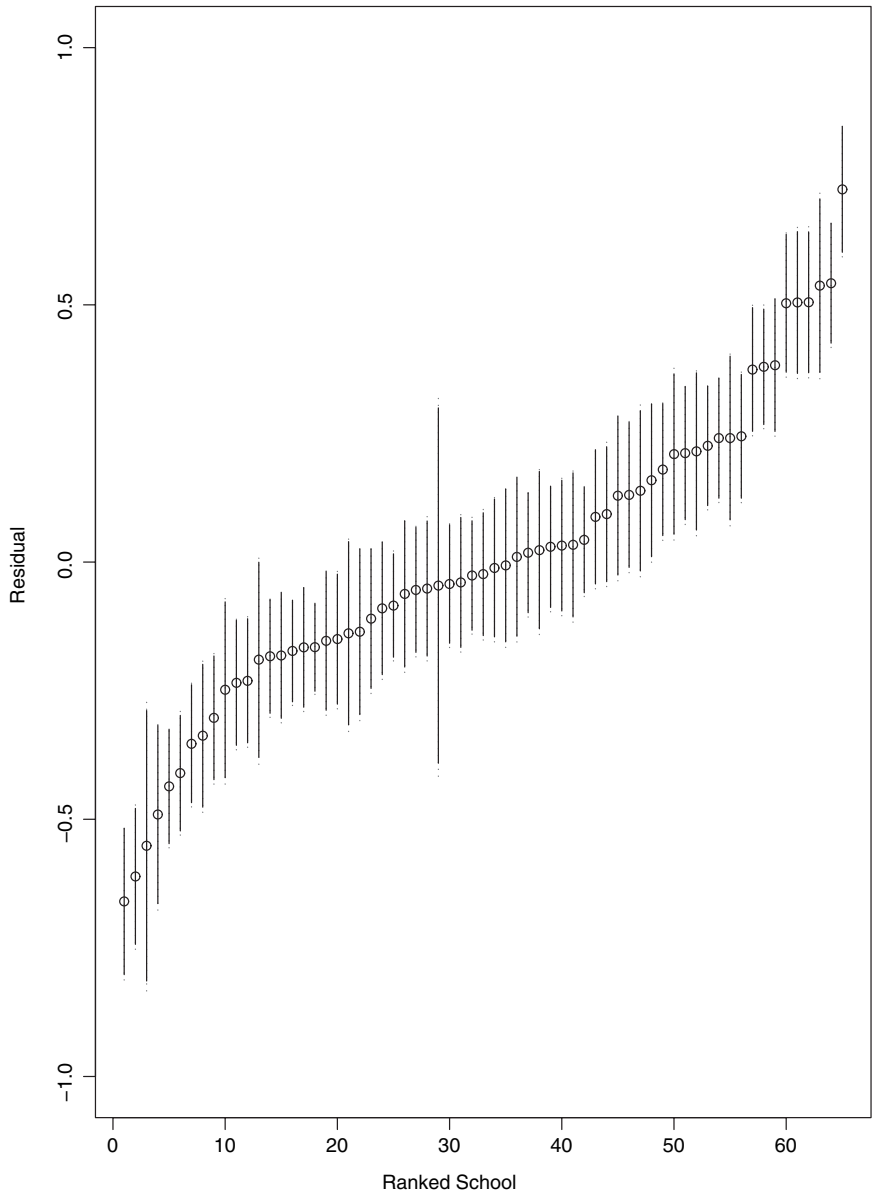

Fig. 3. Graphical method of Goldstein and Healy (1995) applied to the local education authority data of Rasbash *et al.* (2004) (2080 pairs of 95% invervals): |, accounting for the covariances of the level 2 residual estimates (in this case, 1031 of the 2080 pairs do not overlap); ⋮, falsely assuming independence (in this case, 977 of the 2080 pairs do not overlap; unfortunately, the differences are rather small and hardly show up on this plot)

not designed. We also should add that, by applying the method of Goldstein and Healy (1995) to situations for which it was not designed, the inference becomes typically conservative (i.e. too few rejections), as in the example here. Of course, this is preferable to the analysis becoming liberal (i.e. too many rejections).

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 $S = 2080$ possible pairwise comparisons. If we

Table 2. Number of rejected hypotheses for various applications and methods

Method	*Number of rejected hypotheses*
Local education authority data, absolute comparisons, $S=65$	
StepM with $\alpha=0.05$	17
StepM with $\alpha=0.1$	17
FDP–StepM with $\alpha=0.05$ and $\gamma=0.1$	27
National Educational Longitudinal Study data, absolute comparisons, $S=981$	
StepM with $\alpha=0.05$	38
StepM with $\alpha=0.1$	42
FDP–StepM with $\alpha=0.05$ and $\gamma=0.1$	249
Wafer data, pairwise comparisons, $S=45$	
StepM with $\alpha=0.05$	26
StepM with $\alpha=0.1$	27
FDP–StepM with $\alpha=0.05$ and $\gamma=0.1$	30
Local education authority data, pairwise comparisons, $S=2080$	
StepM with $\alpha=0.05$	348
StepM with $\alpha=0.1$	411
FDP–StepM with $\alpha=0.05$ and $\gamma=0.1$	966

examine only the individual p-values, a total of 1027 hypotheses are rejected. The application of the StepM, FDP–StepM and FDR methods yield 348, 966 and 1026 rejections respectively. For the Goldstein and Healy (1995) approximate method not using the covariance terms for residual estimates there are 977 rejections. If we account for the covariances, as described in Appendix C, there are 1031 rejections. This difference is illustrated in Fig. 3.

As expected, with the covariances accounted for, the visual short cut number 1031 is now very close to the number of 1027 rejections for the exact analysis based on individual p-values.

In view of the discussion at the beginning of Section 3.3, we now also apply the StepM method with nominal level $\alpha=0.1$ to all four applications. The numbers of rejections are displayed in Table 2. It can be seen that applying the FDP–StepM method using $\alpha=0.05$ and $\gamma=0.1$ typically rejects many more hypotheses.

4.3. Extension to random slopes

We may extend the methods that are proposed in this paper to random-slope models. Under such a scenario, the hypothesis testing problems for comparing schools (among each other or with a bench-mark) would be fundamentally different. Specifically, there would exist u_{0j} and u_{1j} corresponding to the intercept and slope respectively for each group. In the context of the examples that we considered the slope coefficient of the prior ability covariate instead of being fixed may be considered to vary across schools and u_{1j} is then the random departure of school j from the average slope. In much educational effectiveness research this has been shown to be a more promising and realistic type of model. Since there are now two random effects summarizing the school's position no single ranking applies. Indeed it may be that the values of u_{0j}

and u_{1j} are such that summary lines for different schools may cross. In this case, for example, a school which had a lower estimated outcome than another for one particular value of the covariate may be higher for another value.

This situation is rather more complex than the one that was directly considered in this paper since inferences about which schools could be said to differ from average or between themselves would now involve joint testing of the intercept and slope effects, but still within the multiple-testing framework across all schools. Extensions may, however, be possible in principle though they may be more difficult in implementation.

However, a referee of an earlier version of this paper has suggested a relatively simple solution to this situation which enables the methods that are considered in this paper to be applied directly as follows. We might think it sufficient in practice to consider school rankings at a few specific values across the range of the covariate x, say a low ranking, a middle ranking and a high ranking. In each case we centre the data at that value of $x = x_{\text{spec}}$ say, by defining a new covariate $x_{\text{new}} = x - x_{\text{spec}}$. We then fit the model with both random intercepts and slopes. Since the centred data now have the value $x_{\text{new}} = 0$ for $x = x_{\text{spec}}$ the school effects at that value are summarized by the intercept residual alone and the methods of our paper apply. We can repeat the exercise by fitting models with data centred at the different values of x. A range of conclusions may be reached which then enable us to address the real phenomenon of differential school effectiveness, which possibly different slopes as well as intercepts imply.

5. Conclusion

Level 2 residuals, which are 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. For some of the associated inference problems there may be a temptation to ignore the pitfall of data snooping. Data snooping occurs when multiple hypothesis tests are carried out at the same time and inference is based on the individual p-values or the individual confidence intervals, without taking the multitude of tests into account. 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:

(a) inference on the 'absolute' level 2 residuals to determine which are significantly different from 0 and
(b) inference on pairwise comparisons of level 2 residuals.

Our first method controls the familywise error rate FWE which is defined as the probability of making even one false rejection. If FWE is controlled at level 5%, say, then we can be 95% confident that all rejected hypotheses are indeed false. The advantage of the method that we propose over traditional methods controlling 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, whereas the methods of Bonferroni and Holm assume a worst case scenario.

When the number of hypotheses that are under test is very large—which can happen, for example, when all pairwise comparisons are of interest—then controlling 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 discoveries, 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, i.e. FDR $= E$(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.

Acknowledgements

We thank the Joint Editor and two referees for some very detailed and helpful comments that have greatly improved the presentation of the paper. We also thank Daniel Yadgar for many helpful comments. The research of the second author has been partially supported by the Spanish Ministry of Science and Technology and Federacion Espanola de Enfermedades Raras, grant BMF2003-03324.

Appendix A: Control of *k*-FWE

Some notation is required. Suppose that $\{y_s : s \in K\}$ is a collection of real numbers which are indexed by a finite set K having $|K|$ elements. Then, for $k \leqslant |K|$, k-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 in ascending order as $y_{(1)} \leqslant \ldots \leqslant y_{(|K|)}$, then k-max$_{s\in K}(y_s) = y_{(|K|-k+1)}$.

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

$$d_K(1-\alpha, k, P) = \inf\{x : P\{k\text{-}\max_{s\in K}(|w_{r_s} - \theta_{r_s}|/\hat{\sigma}_{r_s}) \leqslant x\} \geqslant 1-\alpha\}, \tag{12}$$

i.e. $d_K(1-\alpha, k, P)$ is the smallest $(1-\alpha)$-quantile of the sampling distribution under P of k-max$_{s\in K}(|w_{r_s} - \theta_{r_s}|/\hat{\sigma}_{r_s})$. Here, the random variable k-max$_{s\in K}(|w_{r_s} - \theta_{r_s}|/\hat{\sigma}_{r_s})$ is given by the kth largest centred test statistic where the maximization is taken over the specified index set K.

These quantiles would yield finite sample control of k-FWE. But, since the true probability distribution P is unknown, these choices are not feasible. Analogously to the StepM method for control of FWE, a bootstrap approach yields feasible constants: P is replaced by an estimator $\hat{P}$. The ideal constant $d_K(1-\alpha, k, P)$ is then estimated as $\hat{d}_K(1-\alpha, k, P) = d_K(1-\alpha, k, \hat{P})$. For details on how to compute such constants via the bootstrap in examples 1 and 2, see Appendix B.

A.1. Algorithm 3 (k-StepM method)

Step 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.
Step 2: for $1 \leqslant s \leqslant S$, if $0 \notin [w_{r_s} \pm \hat{\sigma}_{r_s}\hat{d}_1]$, reject the null hypothesis H_{r_s}. Here

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

Step 3: denote by R_1 the number of hypotheses rejected. If $R_1 < k$, stop; otherwise let $j = 2$.
Step 4: for $R_{j-1} + 1 \leqslant s \leqslant S$, if $0 \notin [w_{r_s} \pm \hat{\sigma}_{r_s}\hat{d}_j]$, reject the null hypothesis H_{r_s}. Here

$$\hat{d}_j = \max_K[d_K(1-\alpha, k, \hat{P}) : K = \{k-1 \text{ elements of } \{1, \ldots, R_{j-1}\}\} \cup \{R_{j-1}+1, \ldots, S\}] \tag{13}$$

Step 5:

(a) if no further hypotheses are rejected, stop;
(b) otherwise, denote by R_j the number of all hypotheses that have been rejected so far and, afterwards, let $j = j + 1$. Then return to step 4.

The intuitive reason for the complicated expression (13) is the following. To achieve control of k-FWE, in any given step, hypotheses that have been rejected previously cannot be simply ignored or 'forgotten about'. Instead, we must acknowledge that, when we consider a set of remaining hypotheses, we may have reached that stage by rejecting true null hypotheses, but presumably at most $k-1$ of them. Since we do not know which of the hypotheses that have been rejected thus far are true or false, we must maximize over all subsets K that include $k-1$ of the previously rejected hypotheses and all the hypotheses that have not been rejected so far. Note that a slightly more concise definition of $\hat{d}_j$, which is used later in Appendix B, is given as

$$\hat{d}_j = \max_{I \subset \{1,\ldots,R_{j-1}\}, |I|=k-1} [\hat{d}_K(1-\alpha, k, \hat{P}) : K = I \cup \{R_{j-1}+1, \ldots, S\}].$$

The computation of the constants $\hat{d}_j$ may be very expensive if the number of choices is very large. In such cases, we suggest the following short cut, which has been coined the operative method. Pick a user-defined number $N_{\max}$, say $N_{\max} = 50$, and let N^* be the largest integer for which $\binom{N^*}{k-1} \leqslant N_{\max}$. In finding $\hat{d}_j$ in each step, we now restrict maximization to the set of choices involving only the N^* least significant of hypotheses rejected so far instead of all R_{j-1} of them. Note that this short cut does not affect the asymptotic control. The reason is that, with probability tending to 1, the false hypotheses will be rejected first and so we can restrict attention to the less significant hypotheses that have been rejected so far. Nevertheless, in the interest of better k-FWE control in finite samples, we suggest choosing $N_{\max}$ as large as possible.

Appendix B: Use of the bootstrap

B.1. Use of the bootstrap for the StepM method

We now detail how to compute the constants $\hat{d}_j$ in examples 1 and 2 via the bootstrap for use in algorithm 1. Again, the bootstrap method that is employed is the cases bootstrap resampling the level 1 units only; see van der Leeden *et al.* (2007), section 3.3. Denote the observed data by V. From V we compute the individual level 2 residual estimates $\hat{u}_1, \ldots, \hat{u}_J$. The application of the cases bootstrap results in a (generic) bootstrap data set V^* by sampling from the estimated distribution $\hat{P}$.

Recall that $S = J$ in example 1 and that we can therefore 'rewrite' the level 2 residuals as $u_1, \ldots, u_S$ for this example.

B.1.1. Algorithm 4 (computation of the $\hat{d}_j$ for example 1)

Step 1: the labels $r_1, \ldots, r_S$ and the numerical values of $R_0, R_1 \ldots$ are given in algorithm 1.
Step 2: generate B bootstrap data sets $V^{*,1}, \ldots, V^{*,B}$. (One should use $B \geqslant 1000$ in practice.)
Step 3: from each bootstrap data set $V^{*,b}$, $1 \leqslant b \leqslant B$, compute the individual level 2 residual estimates $\hat{u}_1^{*,b}, \ldots, \hat{u}_S^{*,b}$. Also, compute the corresponding estimated standard errors $\hat{\sigma}(\hat{u}_1^{*,b}), \ldots, \hat{\sigma}(\hat{u}_S^{*,b})$.
Step 4:

(a) for $1 \leqslant b \leqslant B$, compute $\max_j^{*,b} = \max_{R_{j-1}+1 \leqslant s \leqslant S}\{|\hat{u}_{r_s}^{*,b} - \hat{u}_{r_s}|/\hat{\sigma}(\hat{u}_{r_s}^{b,*})\}$.
(b) compute $\hat{d}_j$ as the $1-\alpha$ empirical quantile of the B values $\max_j^{*,1}, \ldots, \max_j^{*,B}$.

B.1.2. Algorithm 5 (computation of the $\hat{d}_j$ for example 2)

Step 1: the labels $r_1, \ldots, r_S$ and the numerical values of $R_0, R_1 \ldots$ are given in algorithm 1.
Step 2: generate B bootstrap data sets $V^{*,1}, \ldots, V^{*,B}$. (One should use $B \geqslant 1000$ in practice.)
Step 3: from each bootstrap data set $V^{*,b}$, $1 \leqslant b \leqslant B$, compute the individual level 2 residual estimates $\hat{u}_1^{*,b}, \ldots, \hat{u}_J^{*,b}$. Also, for a particular difference $w_s^* = \hat{u}_j^{b,*} - \hat{u}_k^{b,*}$ compute the corresponding estimated standard error $\hat{\sigma}_s^{*,b} = \hat{\sigma}(\hat{u}_j^{b,*} - \hat{u}_k^{b,*})$.
Step 4:

(a) for $1 \leqslant b \leqslant B$, compute $\max_j^{*,b} = \max_{R_{j-1}+1 \leqslant s \leqslant S}(|w_{r_s}^* - w_{r_s}|/\hat{\sigma}_s^{*,b})$;
(b) compute $\hat{d}_j$ as the $1-\alpha$ empirical quantile of the B values $\max_j^{*,1}, \ldots, \max_j^{*,B}$.

B.2. Use of the bootstrap for the k-StepM method

We next detail how to compute the constants $\hat{d}_j$ in examples 1 and 2 via the bootstrap for use in algorithm 3, resampling from $\hat{P}$.

B.2.1. Algorithm 6 (computation of the $\hat{d}_j$ for example 1)

Step 1: the labels $r_1, \ldots, r_S$ and the numerical values of $R_0, R_1 \ldots$ are given in algorithm 3.
Step 2: generate B bootstrap data sets $V^{*,1}, \ldots, V^{*,B}$. (One should use $B \geqslant 1000$ in practice.)
Step 3: from each bootstrap data set $V^{*,b}$, $1 \leqslant b \leqslant B$, compute the individual level 2 residual estimates $\hat{u}_1^{*,b}, \ldots, \hat{u}_S^{*,b}$. Also, compute the corresponding estimated standard errors $\hat{\sigma}(\hat{u}_1^{*,b}), \ldots, \hat{\sigma}(\hat{u}_S^{*,b})$.
Step 4:

(a) for $1 \leqslant b \leqslant B$, and any needed K, compute $k\max_K^{*,b} = k\text{-max}_{s \in K}(|\hat{u}_{r_s}^{*,b} - \hat{u}_{r_s}|/\hat{\sigma}_{r_s}^{*,b})$;
(b) compute $d_K(1-\alpha, k, \hat{P})$ as the $1-\alpha$ empirical quantile of the B values $k\max_K^{*,1}, \ldots, k\max_K^{*,B}$.

Step 5:

(a) if $j=1$, $\hat{d}_1 = d_{\{1,\ldots,S\}}(1-\alpha, k, \hat{P})$;
(b) if $j>1$, $\hat{d}_j = \max_{I \subset \{1,\ldots,R_{j-1}\}, |I|=k-1}[d_K(1-\alpha, k, \hat{P}) : K = I \cup \{R_{j-1}+1, \ldots, S\}]$.

B.2.2. Algorithm 7 (computation of the $\hat{d}_j$ for example 2)

Step 1: the labels $r_1, \ldots, r_S$ and the numerical values of $R_0, R_1 \ldots$ are given in algorithm 3.
Step 2: generate B bootstrap data sets $V^{*,1}, \ldots, V^{*,B}$. (One should use $B \geqslant 1000$ in practice.)
Step 3: from each bootstrap data set $V^{*,b}$, $1 \leqslant b \leqslant B$, 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 estimated standard error $\hat{\sigma}_s^{*,b} = \hat{\sigma}(\hat{u}_{a_{s,1}}^{b,*} - \hat{u}_{a_{s,2}}^{b,*})$.
Step 4:

(a) for $1 \leqslant b \leqslant B$, and any needed K, compute $k\max_K^{*,b} = k\text{-max}_{s \in K}(|w_{r_s}^{*,b} - w_{r_s}|/\hat{\sigma}_{r_s}^{*,b})$;
(b) compute $d_K(1-\alpha, k, \hat{P}_T)$ as the $1-\alpha$ empirical quantile of the B values $k\max_K^{*,1}, \ldots, k\max_K^{*,B}$.

Step 5:

(a) if $j=1$, $\hat{d}_1 = d_{\{1,\ldots,S\}}(1-\alpha, k, \hat{P}_T)$;
(b) if $j>1$, $\hat{d}_j = \max_{I \subset \{1,\ldots,R_{j-1}\}, |I|=k-1}[d_K(1-\alpha, k, \hat{P}_T) : K = I \cup \{R_{j-1}+1, \ldots, S\}]$.

When applying one of these stepwise methods in practice, one first computes $\hat{d}_1$, then $\hat{d}_2$, and so on. It is important that for the computation of the various $\hat{d}_j$, $j = 1, 2, \ldots$, only a unique set of bootstrap data sets $V^{*,1}, \ldots, V^{*,B}$ be used, i.e. the bootstrap data sets are generated 'once and for all' at the outset rather than anew in each step. Only in this way is the monotonicity condition $\hat{d}_{j+1} \leqslant \hat{d}_j$ guaranteed. (This condition holds if P was known and so the ideal constants could be computed; so the condition should also hold for the feasible constants computed from the bootstrap.) Of course, this way also speeds up the computations.

Appendix C: Modified Goldstein and Healy plot

We briefly describe how to modify the graphical method of Goldstein and Healy (1995), to incorporate the covariances between estimates. In doing so, we adopt their notation (page 176). Let ϕ_{ij} denote the covariance between m_i and m_j. Then $\text{var}(m_i - m_j)$ is generalized to

$$\text{var}(m_i - m_j) = \sigma_i^2 + \sigma_j^2 - 2\phi_{ij} = \sigma_{ij}^2.$$

Equation (2) of Goldstein and Healy (1995) and the subsequent procedure remain unchanged, with the understanding that the generalized definition for σ_{ij} is used.

Of course, for our applications, m_i is replaced by $\hat{u}_i$, m_j is replaced by $\hat{u}_j$, σ_i is replaced by $\hat{\sigma}(\hat{u}_i)$, σ_j is replaced by $\hat{\sigma}(\hat{u}_j)$ and ϕ_{ij} is replaced by $\widehat{\text{cov}}(\hat{u}_i, \hat{u}_j)$.

References

Afshartous, D. and de Leeuw, J. (2004) An application of multilevel model prediction to NELS:88. *Behaviormetrika*, **31**, 43–66.
Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.
Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.

Browne, W. (2003) *MCMC Estimation in MLwiN (Version 2.0)*. London: Institute of Education.
Bryk, A. and Raudenbush, S. (1992) *Hierarchical Linear Models*. Newbury Park: Sage.
Carpenter, J. R., Goldstein, H. and Rasbash, J. (2003) A novel bootstrap procedure for assessing the relationship between class size and achievement. *Appl. Statist.*, **52**, 431–443.
Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.
Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.
Goldstein, H. (2003) *Multilevel Statistical Models*, 3rd edn. London: Hodder Arnold.
Goldstein, H. and Healy, M. J. R. (1995) The graphical presentation of a collection of means. *J. R. Statist. Soc.* 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. *Oxf. Rev. Educ.*, **19**, 425–433.
Goldstein, H. and Spiegelhalter, D. J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *J. R. Statist. Soc.* A, **159**, 385–443.
Harville, D. (1976) Extension of the Gauss-Markov theorem to include estimation of random effects. *Ann. Statist.*, **4**, 384–395.
Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, **6**, 65–70.
James, W. and Stein, C. (1961) Estimation with quadratic loss. In *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 361–380. Berkeley: University of California Press.
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. *J. Statist. Planng Inf.*, **124**, 379–398.
Laird, N. M. and Ware, J. W. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.
van der Leeden, R., Meijer, E. and Busing, F. M. T. A. (2007) Resampling multilevel models. In *Handbook of Multilevel Analysis* (eds J. de Leeuw and E. Meijer). New York: Springer.
de Leeuw, J. and Kreft, I. (1986) Random coefficient models for multilevel analysis. *J. Educ. Statist.*, **11**, 57–85.
de Leeuw, J. and Kreft, I. (1995) Questioning multilevel models. *J. Educ. Behav. Statist.*, **20**, 171–189.
Lehmann, E. L. and Romano, J. P. (2005a) *Testing Statistical Hypotheses*, 3rd edn. New York: Springer.
Lehmann, E. L. and Romano, J. P. (2005b) Generalizations of the familywise error rate. *Ann. Statist.*, **33**, 1138–1154.
Longford, N. (1987) A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with random effects. *Biometrika*, **74**, 817–827.
National Center for Educational Statistics (2006) National educational longitudinal study of 1988. National Center for Educational Statistics, Washington DC. (Available from `http://nces.ed.gov/surveys/nels88/`.)
Pinheiro, J. and Bates, D. (2000) *Mixed Effects Models in S and S-Plus*. New York: Springer.
Rasbash, J., Steele, F., Browne, W. and Prosser, B. (2004) *A User's Guide to MLwiN Version 2.0*. London: Institute of Education. (Available from `http://multilevel.ioe.ac.uk/download/userman20.pdf`.)
Raudenbush, S. and Bryk, A. (2002) *Hierarchical Linear Models*. Newbury Park: Sage.
Robinson, G. (1991) That BLUP is a good thing: the estimation of random effects. *Statist. Sci.*, **6**, 15–51.
Romano, J. P. and Wolf, M. (2005) Stepwise multiple testing as formalized data snooping. *Econometrica*, **73**, 1237–1282.
Romano, J. P. and Wolf, M. (2007) Control of generalized error rates in multiple testing. *Working Paper 245*. Institut für Empirische Wirtschaftsforschung, University of Zurich, Zurich. (Available from `http://www.iew.unizh.ch/wp/index.php`.)
R Project (2006) *R—Version 2.2.1*. Vienna: R Project. (Available from `http://www.r-project.org/`.)
Searle, S. R., Casella, G. and McCulloch, C. E. (1992) *Variance Components*. New York: Wiley.
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. *J. Educ. Behav. Statist.*, **24**, 42–69.
Yekutieli, D. (2002) A false discovery rate procedure for pairwise comparisons. *Technical Report*. Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv. (Available from `http://www.math.tau.ac.il/%7Eyekutiel/work.html`.)