Large-scale Assessments in Education

METHODOLOGY

Open Access

Defensible inferences from a nested sequence of logistic regressions: a guide for the perplexed

Gulsah Gurkan[1*], Yoav Benjamini[2] and Henry Braun[1]

*Correspondence: gurkan@bc.edu
[1] Boston College, Chestnut Hill, MA, USA
Full list of author information is available at the end of the article

Abstract

Employing nested sequences of models is a common practice when exploring the extent to which one set of variables mediates the impact of another set. Such an analysis in the context of logistic regression models confronts two challenges: (i) direct comparisons of coefficients across models are generally biased due to the changes in scale that accompany the changes in the set of explanatory variables, (ii) conducting a large number of tests induces a problem of multiplicity that can lead to spurious findings of significance if not heeded. This article aims to illustrate a practical strategy for conducting analyses in the face of these challenges. The challenges—and how to address them—are illustrated using a subset of the findings reported by Braun (Large-scale Assess Educ 6(4):1–52, 2018. 10.1186/s40536-018-0058-x), drawn from the Programme for the International Assessment of Adult Competencies (PIAAC), an international, large-scale assessment of adults. For each country in the dataset, a nested pair of logistic regression models was fit in order to investigate the role of Educational Attainment and Cognitive Skills in mediating the impact of family background and demographic characteristics on the location of an individual's annual income in the national income distribution. A modified version of the Karlson–Holm–Breen (KHB) method was employed to obtain an unbiased estimate of the true differences in the coefficients between nested logistic models. In order to address the issue of multiplicity, a recent generalization of the Benjamini–Hochberg (BH) False Discovery Rate (FDR)-controlling procedure to hierarchically structured hypotheses was employed and compared to two conventional methods. The differences between the changes in coefficients calculated conventionally and with the KHB adjustment varied from negligible to very substantial. When combined with the actual magnitudes of the coefficients, we concluded that the more proximal factors indeed act as strong mediators for the background factors, but less so for Age, and hardly at all for Gender. With respect to multiplicity, applying the FDR-controlling procedure yielded results very similar to those obtained by applying a standard per-comparison procedure, but quite a few more discoveries in comparison to the Bonferroni procedure. The KHB methodology illustrated here can be applied wherever there is interest in comparing nested logistic regressions. Modifications to account for probability sampling are practicable. The categorization of variables and the order of entry should be determined by substantive

considerations. On the other hand, the BH procedure is perfectly general and can be implemented to address multiplicity issues in a broad range of settings.

Keywords: PIAAC, Logistic regression, Nested model comparisons, KHB method, Multiplicity, False Discovery Rate, BH procedure, Hierarchical testing

Introduction

A common problem in the statistical analysis of observational data is to elucidate the relationships among various potential explanatory variables and a focal outcome. When the data arise in a social or behavioral context, the explanatory variables can often be categorized according to one or more criteria and, accordingly, the analysis can be organized to take advantage of the categorization. A common situation is one in which the categories are determined by their temporal ordering in relation to the focal outcome. In this case, the analysis can proceed by stages, with the set of variables in the most distal temporal category entered first *en masse*, followed by the set of variables in the second most distal category, and so on. Of course, purpose guided by theory should guide the analysis. One purpose might be to find the most parsimonious explanatory model, subject to certain constraints. Another is a form of mediation analysis; that is, tracking how the magnitudes of the coefficients of each set of variables change as additional sets of variables enter the model. This latter purpose is the subject of the present article.

As noted above, one context for mediation analysis occurs when there is a focal outcome and the pool of explanatory variables has a temporal sequence in relation to that outcome. For example, the focal outcome can be graduation from tertiary education or earning a salary above a certain threshold. One set of variables comprises demographic characteristics and a second set of variables comprises measures of human capital accumulated prior to determination of the outcome.

Typically, one finds strong statistical associations between the first set of variables and the outcome. By adding the second set of variables to the regression model, one can address the following question: to what extent do differences in the second set of variables account for those earlier statistical associations? If inclusion of the second set of variables appears to have substantial explanatory power (i.e., substantially reduce the coefficients of one or more of the variables in the first set), then subsequent analyses should examine more carefully the patterns of association between the two sets of explanatory variables as that can lead to substantively useful insights on how differences in the focal outcome develop and even how they might be mitigated (if need be). On the other hand, if one or more variables in the first set retain their association with the outcome, then a different investigation is called for. In order to make these determinations, one must compare the coefficients of the variables in the first set between the two models. An extended example of this kind of analysis is presented below. However, first a number of methodological challenges must be identified and addressed.

The investigator seeking to carry out a mediation analysis in the context of logistic regression (i.e., with a dichotomous focal outcome) confronts two challenges. The first is that to obtain an unbiased estimate of the difference between corresponding coefficients in two nested logistic regression models, one must take account of the fact that the scales of the two models are typically no longer the same. Consequently, simple

differences of the coefficients are generally biased, sometimes seriously so (Karlson et al., 2012), hereafter denoted KHB.

The second is that when making many such comparisons, one encounters the problem of multiplicity. Conducting a large number of tests at the usual threshold (e.g., 0.05 or even 0.01) can result in an unacceptably high simultaneous Type I error rate; that is, too many false rejections of the null hypothesis. This problem has long been recognized (Hochberg & Tamhane, 1987) and various strategies proposed to address it. The most commonly used is the Bonferroni procedure, which controls the Type I error rate simultaneously over all tests, but at the expense of a substantial loss of power. For example, multi-stage procedures have been shown to possess greater power than the Bonferroni procedure (Braun & Tukey, 1983). However, an increasingly popular alternative is to replace the Type I error rate with another criterion: the False Discovery Rate (FDR) described in Benjamini and Hochberg (1995). Essentially, the FDR is the expected value of the ratio of the number of 'false significances' to the 'total number of significances' declared by the hypothesis testing procedure. In that article, the authors also proposed a strategy (BH procedure) for controlling the FDR at a predetermined level. Since then, there have been a number of refinements of the BH procedure to extend its applicability (Benjamini & Yekutieli, 2001).

Although it is straightforward to apply BH in a simple situation, more complex settings require deeper consideration of both the substantive and the statistical aspects of the problem, as well as somewhat more involved calculations. Corresponding generalizations of the BH procedure have been developed. For example, Bogomolov et al. (2020) report results on controlling the FDR when the set of tested hypotheses conform to a hierarchical structure.

It bears mentioning that data from large-scale assessment surveys such as PIAAC introduce two further complexities. First, the analysis must take account of the sampling weights attached to each observation, as failure to do so could result in biased estimates, to the extent that the weights are related to the variables of interest. Second, the direct assessment of cognitive skills does not yield scores in the conventional sense. Rather, for each respondent and each domain, the analysis produces a set of imputed values from the (estimated) posterior distribution of proficiency. In this context, the imputed values are referred to as *plausible values* (PV) and they serve as the basis for computing aggregate score distributions (Braun & von Davier, 2017).

A practical strategy for conducting analyses that take into account these challenges is illustrated here with a subset of the findings reported by Braun (2018).[1] Braun employed data from an international assessment of adult literacy, the Programme for the International Assessment of Adult Competencies (PIAAC Cycle 1 - Round 1). The focal outcome that concerns us here is the individual's location in the national income distribution. Braun (2018) investigated the extent to which the strength of the associations between temporally earlier variables were attenuated as more proximal variables were entered. Specifically, interest centered on quantifying the 'long shadow' of demographic characteristics and family background on an individual's labor market success—both as

[1] Braun (2018) contained several sets of analysis. In this paper, attention focuses on one particular set.

the sole set of predictors and in the presence of more proximal measures of cognitive skills and educational attainment. The striking patterns revealed by the analyses were described and informally summarized. However, no formal tests of significance were conducted.

This article is organized as follows. The next section provides some background on the PIAAC data and the substantive issue that serves as the exemplar, as well as aspects of inference in logistic regression and of addressing problems of multiplicity. Next, we briefly present the original findings, followed by a description of the methodologies employed here. The penultimate section presents the results of applying these methodologies to the PIAAC data. The final section summarizes the insights gained and the general implications of employing these methodologies in empirical research.

Background

Data

The Programme for the International Assessment of Adult Competencies is a household survey conducted under the auspices of the OECD. It collects data from nationally representative samples of adults ages 16–65 (OECD, 2019). PIAAC provides direct measures of cognitive skills such as literacy and numeracy. More broadly, the full data set enables the examination of the relationships among adults' demographic and birth family characteristics, cognitive skills, educational attainment, work experiences and labor market outcomes. For policymakers and policy researchers, the value of PIAAC is that it offers a common framework for comparing patterns of relationships across countries and, in particular, the contribution(s) of differences in the extent to which family background, cognitive skills, and educational attainment account for variation in a wide range of outcomes.

For the purpose of this study, family background is represented by *Parental Education* and *Books in the Home*. Demographic variables are gender and age category in ten-year increments. The labor market outcomes studied comprised two dichotomous variables indicating whether the respondent's income was in the highest quartile (Q4) or in the lowest quartile (Q1) of the national annual income distribution. These two outcomes were chosen in preference to average income in order to examine potential differences in relationships at the two tails of the income distribution. For further information on PIAAC, consult (OECD, 2019).

Nested logistic regressions

Logistic regression is a particular case of a generalized linear model (GLM; Dobson & Barnett, 2008). A GLM is employed when interest centers on examining the relationship between an outcome variable with a distribution in the exponential family of distributions, and a set of explanatory variables. It is characterized by representing some function of the outcome variable as a linear combination of the explanatory variables. That function is referred to as the *link function*.

The simplest case is the normal distribution with the link function being the identity function. In this case, the structural component of the model (the linear combination of explanatory variables) and the stochastic component (the residual) are completely separate. For other distributions, the structural and stochastic components

are entangled. In logistic regression, for example, interest centers on accounting for the variability in the probabilities associated with a dichotomous outcome. The usual link function is the logit = log [p/(1 − p)]. Changing the set of explanatory variables simultaneously alters both the structural and stochastic components. Consequently, comparing coefficients of a particular variable that appears in both models is not as straightforward as in the normal case: the simple difference of coefficients between the two models is typically a biased estimate of the true difference because of the changes in scale that accompany the changes in the set of explanatory variables.

As KHB note, this difficulty had been pointed out by various authors in the past, but appears not to be widely appreciated. The important contribution of KHB was to develop a relatively straightforward procedure to obtain an unbiased estimate of the true difference, along with a software program to implement the procedure. Further developments were documented in Breen et al. (2018).

In the KHB formulation, variables are categorized as either predictors (denoted X) or mediators (denoted Z). The regression coefficients of the former are the focus of interest. The mediators are said to confound the "effect" of the predictors and the goal is to estimate the impact of the confounding on the coefficients of the predictors. Thus, to obtain an unbiased estimate of the impact of confounding, it is necessary to eliminate the impact of rescaling. As described in KHB (2012), this can be done by first isolating the confounding. That is, if we construct the X-residualized Z variables ($\tilde{Z}$) such that their correlation with X is zero, then we can assume that adding a variable uncorrelated with X will not alter the coefficient of X. Consequently, the observed changes in the coefficients of X must be due to rescaling. When the estimated coefficient of X in the model where it is affected by both confounding and rescaling ('full model') is compared to the estimate from the model where $\tilde{Z}$ is used ('reduced model'), the difference yields an estimate of the change in the coefficients of X due to confounding only.

The False Discovery Rate (FDR) and its control

Consider a typical situation in which a family of (null) hypotheses is tested. Denoting by R the number of rejected hypotheses, suppose that V of them are rejected in error (i.e., Type I errors). Identifying a rejection with a *statistical discovery*, the False Discovery Proportion is defined as V/R when R > 0, and 0 if none are rejected. The False Discovery Rate (FDR) is defined as the expectation of the false discovery proportion: FDR = E [V/R].

The BH procedure was originally designed to enable control of the FDR for a family of hypotheses investigated by a corresponding family of independent test statistics at a predetermined level (e.g., 0.05, 0.01, or 0.005). The procedure makes use of the observed p-values only. Suppose that there are m hypotheses to be tested and the desire is to bound the FDR by a value q (0 < q < 1). The BH procedure is conducted as follows:

> Order the m p-values in the family from the smallest $p_{(1)}$ (most extreme) to the largest $p_{(m)}$ (least extreme), so that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)} \leq \ldots \leq p_{(m)}$.

Calculate the largest k for which $p_{(k)} \leq$ qk/m, and reject the k hypotheses corresponding to $p_{(1)},\ldots,p_{(k)}$, rejecting none if no such k exists.

Equivalently, an FDR-adjusted p-value for the BH procedure can be computed and compared to the desired FDR level in the same way an observed p-value is compared to a pre-determined significance level. The FDR-adjusted p-values using BH (also called BH q-values) are constructed as follows:

Adjust each $p_{(i)}$ to $p_{(i)}^{BH} = min_{j \geq i}$ ($p_{(j)}$ m/j);
Reject the corresponding hypothesis if $p_{(i)}^{BH} \leq$ q.

For purposes of reporting, $p_{(i)}^{BH}$ can be used in the same way that the usual p-value is.

The general principle underlying the FDR criterion is that it measures the average Type I error rate over a set of hypotheses selected after viewing the data—and being rejected. Since the original paper, the BH procedure has been shown, both theoretically and empirically, to control the FDR in a wider range of settings, especially for dependent normally distributed test statistics (Benjamini & Yekutieli, 2001). The BH FDR-controlling procedure further controls the false directional rate, and has also been generalized to the false coverage rate for confidence intervals. In the present case, we make use of a more recent generalization to hierarchically structured hypotheses (Bogomolov et al., 2020). In this setting, a two-stage BH procedure is recommended. It is described below.

The notion of a family of hypotheses to be tested jointly is central to any treatment of simultaneous inference. In this context, we consider a family to be a set of hypotheses that can interchangeably serve a similar purpose in the research framework. Further, some of these hypotheses may be selected for highlighting after viewing the data; for example, by being tested and rejected. For a tested family of hypotheses, some measure of the error (e.g., the FDR) over the family should be controlled, otherwise the statistical assurance offered by a single hypothesis test or confidence interval may be severely compromised.

In the present setting, we have a collection of families of hypotheses (higher level), with each family itself comprising a collection of hypotheses (lower level). The error is defined as the false discovery proportion within each lower-level family, averaged over a subset of families selected at the higher level. The expected value of this average is the FDR for the ensemble. Equally important, the two-stage BH procedure also controls the FDR (at the chosen level) for each family of hypotheses. This is particularly useful when there is interest in reporting the results for each family, as well as overall.

Original method and findings

In Braun (2018), the target population comprised adults ages 25–55 in 21 OECD countries that participated in PIAAC. The primary goal was to quantify the "long shadow "of family background on an individual's labor market success as measured by an indicator derived from her income percentile in the national income distribution. The factors employed in the analyses are listed in Table 1. There were two dichotomous outcome variables. One was whether the individual's annual income placed them in the highest quartile (Q4) of the national income distribution. The other was

Table 1 Factors employed in analyses, the number of categories corresponding to each factor and the reference group used in regression analyses (if appropriate). Adapted from Braun (2018)

Factor name	Number of categories	Reference group
Age (set 1)	3	45–54
Gender (set 1)	2	Male
Parental education (set 1)	3	At least one parent with BA+
Books in the home (set 1)	3	100 books or more
Literacy & numeracy (set 2)	Scale scores	N/A
Educational attainment (set 2)	4	BA+

whether the individual's annual income placed them in the lowest quartile (Q1) of the national income distribution.

In Braun (2018), the logistic regressions were implemented with the IDB Analyzer (IEA, 2019), a program that accommodates both sampling weights and plausible values. For the present study, only the first plausible value of the cognitive measure was employed. Note that plausible values are random draws from the posterior distribution of proficiencies for the individuals. They are intended to be used as a set by following procedures described in the technical documentation; that is, an analysis is conducted separately for each set of plausible values, and then the results are combined for reporting purposes (OECD, 2019). Because the main focus of this study was to provide an illustrative example of how to address the abovementioned challenges (i.e., making inferences from a nested sequence of logistic regressions), for simplicity only the first plausible value of the cognitive measure was employed in the models. A full implementation of the procedures described in this paper would require conducting the analysis separately with each set of plausible values, reporting the average of the coefficients, and adding the imputation variance to the calculated sampling variance to obtain more comprehensive error estimates. Analyses were carried out in R (R Core Team, 2020) using the *svydesign* and *svyglm* functions in the *survey* package (Lumley, 2020), which accommodates sampling weights, and various functions in the *dplyr* package (Wickham et al., 2020) for data preparation and manipulation.[2]

In the analyses presented in the next section, we focus on Models 1 and 3 from Braun (2018). The Model 1 explanatory factors are age category, gender, family background, and books in the home. Model 3 adds two more factors: a composite measure of cognitive skills and levels of educational attainment. The former is represented by a continuous variate that combines standardized scores on the literacy and numeracy scales. The latter is represented by four ordered categories. Equivalent analyses were carried out for the Q1 and Q4 outcomes for each country. It is important to note that the national samples for the Q4 analyses were restricted to individuals working full-time. To illustrate, equations representing the models for the Q1 outcome are given below.

Model 1 (M1) :

$$\widehat{logit(Q1)} = \beta_0 + \beta_1(AGE_1) + \beta_2(AGE_2) + \beta_3(GENDER) + \beta_4(PARED_1) + \beta_5(PARED_2) + \beta_6(BOOKS_1) + \beta_7(BOOKS_2)$$

[2] R code is available upon request.

Model 3 (M3):

$$\widehat{logit(Q1)} = \beta_0 + \beta_1(AGE_1) + \beta_2(AGE_2) + \beta_3(GENDER) + \beta_4(PARED_1) + \beta_5(PARED_2) + \beta_6(BOOKS_1) + \beta_7(BOOKS_2) + \beta_8(COGN) + \beta_9(EDCAT_1) + \beta_{10}(EDCAT_2) + \beta_{11}(EDCAT_3)$$

where Q1 denotes the probability of an individual's annual income falling in Q1 of the national income distribution.

For present purposes, the main findings in Braun (2018) were that: (i) Family Background (Parental Education and/or Books in the Home) was strongly associated with both wage-related labor market outcomes. The magnitudes varied across countries and were largely mediated by Educational Attainment and Cognitive Skills; (ii) Age and Gender were strongly associated with both wage-related labor market outcomes. The magnitudes of the associations varied considerably across countries. However, these associations were apparently NOT mediated by Educational Attainment and Cognitive Skills. (Hence, the title employed the phrase "the long shadow".)

Methodology

In the current study, analyses were carried out for 19 of the original 21 countries, as Canada and Slovak Republic were excluded. Canada was dropped from the Q4 analyses because Canada did not collect information on full-time work. The data collected in Slovak Republic did not have any observations for the 3rd category of Educational Attainment factor so it was dropped from the analysis. In practice, it is possible to carry out the BH procedure with different numbers of hypotheses in each lower-level family. However, in the interest of simplicity, we chose to have the same number in each family. See Appendix for a list of the countries comprising this subset, as well as final sample sizes and the abbreviations used in the presentation of the results.

KHB analyses

Following the procedure described in KHB, residualized variables were calculated from the model M3. Specifically, the mediators Cognitive Skills and Educational Attainment were regressed on the predictors Age, Gender, Books in the Home, and Parental Education. These residualized mediators, along with the predictors, were then used to fit a new model, the so-called reduced M3. The differences in the estimates of the coefficients of the predictors gathered from the full M3 and the reduced M3 estimate the impact of the confounding, net of the impact of scale changes.

To proceed further, estimated standard errors of the difference statistics for each of the coefficients are required. These are generated by the KHB program for simple random samples. However, the KHB program does not allow for incorporating the complex sampling design used in PIAAC. Consequently, the jackknife resampling technique with replicate weights was employed (OECD, 2019). The procedure is as follows:

1. A difference statistic is calculated by using the full sample weights in the original M3 and reduced M3 models, yielding ($\widehat{D}_0$).

2. Step 1 is replicated 80 times, each time using a different replicate weight. This yields $\left(\widehat{D}_r = \left\{\widehat{D}_1, \widehat{D}_2, \ldots, \widehat{D}_{80}\right\}\right)$.
3. The sampling variance of the difference statistic for a coefficient is calculated using the formula:

$$\widehat{\mathrm{Var}}\left(\hat{D}\right) = \mathrm{c}^{*} \sum_{r=1}^{80} \left(\hat{D}_r - \hat{D}_0\right)^2,$$

where $c = 1$, for countries using the paired jackknife (JK2).

$\mathrm{c} = (\mathrm{g} - 1)/\mathrm{g}$ where g is the number of replicates, for countries using the random groups (delete-one) approach (JK1).

Standard errors were obtained by taking the square root of the estimated variances. These were used subsequently for calculating z-statistics and conducting (two-tailed) tests of significance.

Hierarchical FDR analyses

Applying any simultaneous inference procedure requires not only a choice of the type of error rate and its level, but also careful consideration of the set of results over which one wishes to control the error rate. As noted earlier, in the present context interest centered on the changes in coefficients between M1 and M3. However, from a substantive point of view, the magnitudes and significance of the coefficients in M3 were also of interest as they indicated the strengths of the (partial) associations between the explanatory variables and the two focal outcomes. Inasmuch as differences in regression coefficients and regression coefficients themselves are fundamentally different, and since significance of a coefficient's difference between models is not exchangeable in meaning with significance of the coefficient itself, it was decided that they would be treated as two different "super-families". Accordingly, the FDR was controlled separately for each super-family and this point is addressed later in this section. We denote the two super-families as SF/Diff and SF/Coef, respectively.

Consider first the structure of SF/Diff. There are seven variables representing the four background factors for the Q1 analysis, and seven again for the Q4 analysis. For each variable there is a family of null hypotheses that states: *The differences in the regression coefficients for this variable equal 0 in all 19 countries*. The collection of these 14 families (regarding the differences in coefficients from M1 to M3) comprise the SF/Diff superfamily. Note that there is a substantive interest in the results for each variable in both Q1 and Q4 (i.e., in each of the 14 families).

The structure of SF/Coef is similar; however, in this case there are 11 variables of interest, representing the four background factors, a cognitive score, and Educational Attainment. (Since testing the significance of the intercept is of no interest, it is not included in the analysis.) As before, combining the analyses for Q1 and Q4, we have a super-family comprising 22 families. The null hypothesis for each family states: *The regression coefficients for this variable equal 0 in all 19 countries*. Again, there is substantive interest in the results for each family of hypotheses. Figure 1 displays the hierarchical structure of these hypotheses.

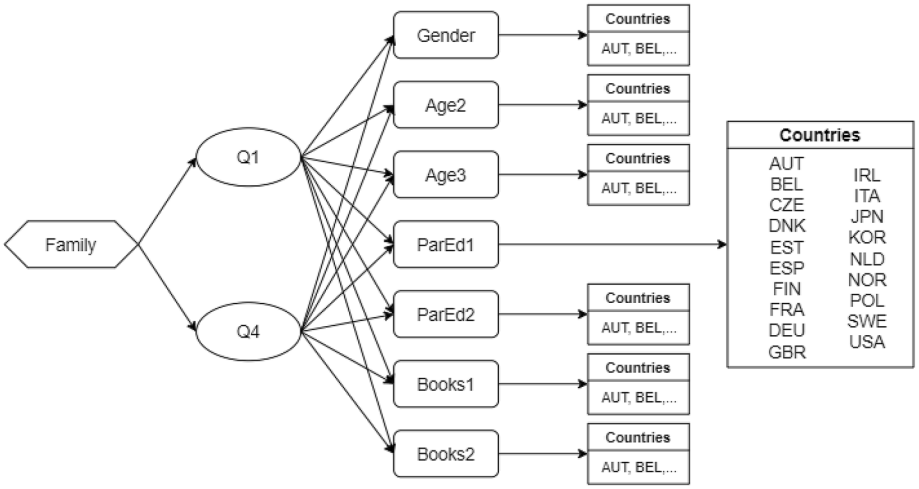

Fig. 1 The hierarchical structure of the hypotheses tested

As noted above, the FDR-controlling procedure for each super-family comprises two stages (Bogomolov et al., 2020). First, for each family (variable) at the lower level of countries, we compute the family-level p-values for the differences in coefficients. This is accomplished as follows: In each family, the 19 p-values are adjusted using the BH procedure as described above.[3] The minimum of the 19 FDR-adjusted p-values yields the FDR-adjusted p-value for the family. For the 14 higher level family of coefficients' differences in the SF/Diff super-family, this yields a set of p-values denoted as $P = \{p_1^*, \ldots, p_{14}^*\}$. The 14 p-values in P are then ordered from smallest to largest and a new set of BH-adjusted p-values are computed. For illustrative purposes, we first chose to test at the 0.05 level.

In general, suppose that at the first stage r of the corresponding fourteen null hypotheses (higher level) are rejected. If $r = 0$, the procedure halts and no discoveries are declared. If $r > 0$, then these r families are designated to enter the second stage of analysis. For each designated family (and only those families), the set of 19 differences are tested using the BH procedure at level 0.05 (r/14). That is, a discovery is declared if a country's adjusted p-value $\leq 0.05(r/14)$. Note that these are the adjusted p-values computed at the first stage of the analysis.

Now consider the super family SF/Coef. Testing is done in exactly the same way as for SF/Diff. In this case, there are 22 families of hypotheses. For the i^{th} family, we calculate the 19 corresponding BH-adjusted p-values. The minimum of these 19 adjusted p-values yields a p_i^* for the family. For the super-family, we obtain $P = \{p_1^*, \ldots, p_{22}^*\}$. The p-values in P are then ordered from smallest to largest and tested using the BH procedure. Again, suppose r of the corresponding 22 null hypotheses are rejected. If $r = 0$, the procedure halts and no discoveries are declared. If $r > 0$, then these r families are designated to enter the second stage of analysis. For each designated family (and only those families), the set of 19 coefficients are tested using the BH procedure at level 0.05 (r/22).

[3] We employed the procedure provided in the *stats* package in R.

Table 2 5-number summaries for conventional differences and KHB differences, by variable [Q1]

	Diff (M1–M3)					Diff_KHB (M1–M3)				
	min	Q1	Median	Q3	max	min	Q1	Median	Q3	max
Gender	−0.39	−0.16	−0.10	−0.01	0.06	−0.26	−0.09	−0.04	0.06	0.18
Age2	−0.34	−0.21	−0.14	−0.09	0.06	−0.31	−0.16	−0.10	−0.03	0.09
Age3	−0.19	−0.11	−0.08	−0.05	0.07	−0.21	−0.11	−0.07	−0.05	0.07
ParEd1	0.22	0.39	0.45	0.52	0.67	0.24	0.39	0.47	0.58	0.71
ParEd2	0.11	0.21	0.25	0.31	0.46	0.12	0.22	0.28	0.33	0.50
Books1	0.32	0.40	0.43	0.52	0.68	0.34	0.43	0.47	0.55	0.73
Books2	0.13	0.16	0.20	0.23	0.35	0.14	0.17	0.22	0.24	0.38

Table 3 5-number summaries for conventional differences and KHB differences, by variable [Q4]

	Diff (M1–M3)					Diff_KHB (M1–M3)				
	min	Q1	Median	Q3	max	min	Q1	Median	Q3	max
Gender	−0.19	0.10	0.16	0.20	0.34	−0.19	−0.01	0.05	0.10	0.17
Age2	−0.08	0.14	0.24	0.35	0.58	−0.13	0.06	0.11	0.20	0.33
Age3	−0.08	0.07	0.10	0.17	0.39	−0.05	0.04	0.07	0.15	0.23
ParEd1	−0.92	−0.63	−0.50	−0.46	−0.34	−1.08	−0.69	−0.51	−0.46	−0.37
ParEd2	−0.58	−0.40	−0.33	−0.23	−0.18	−0.61	−0.40	−0.29	−0.23	−0.18
Books1	−0.90	−0.65	−0.56	−0.47	−0.35	−0.96	−0.75	−0.64	−0.50	−0.36
Books2	−0.44	−0.32	−0.22	−0.21	−0.14	−0.48	−0.36	−0.25	−0.20	−0.12

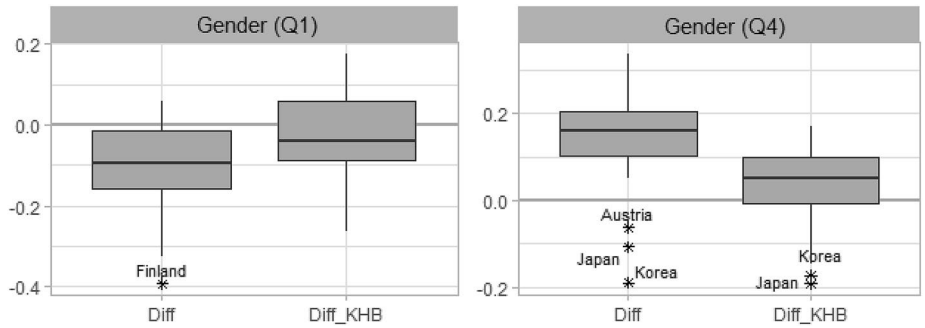

Fig. 2 Box plots of conventional differences and KHB differences for coefficients of gender [Q1 and Q4]

Results and discussion

KHB analyses

Here we compare the differences in coefficients (M1–M3) between the conventional analyses and those based on the KHB analyses. The results are displayed in Tables 2 and 3 in the form of 5-number summaries for each variable. Table 2 presents the results for Q1 and Table 3 the results for Q4. For Q1, the impact of the KHB adjustment is generally to shift the distributions of differences towards more positive values, while for Q4 it is to shift the distributions of differences towards more negative values. We illustrate these results in Fig. 2 with box plots for the coefficient of gender. Panel (a) displays the results for Q1 and Panel (b) for Q4.

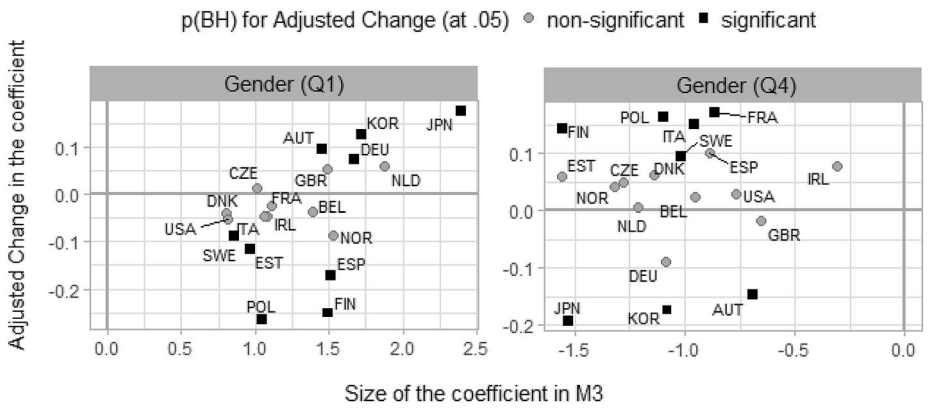

Fig. 3 KHB differences in coefficients of Gender against size of the coefficient in M3 (significance according to BH procedure indicated) [Q1 and Q4]

We observe that the impact of the KHB adjustment is greatest for the coefficients of the gender and age variables. For those variables, it is greater in Q4 than in Q1. In both cases, however, the impact is to produce distributions that are more concentrated around 0. That is, failing to adjust for the change of scale from M1 to M3 yields distributions of differences that, in absolute value, are somewhat too large. Unfortunately, one cannot make assertions with regard to the impact on the corresponding tests of significance since the test statistics also depend on estimates of the standard errors of the differences. Nonetheless, it is of some comfort to know that, after the KHB adjustments, the numerators of these test statistics are unbiased estimates of their respective estimands.

In Fig. 3, we plot the KHB differences in the coefficients of gender against the corresponding M3 coefficient. (Note that for Q1, a positive coefficient indicates a disadvantage for females; for Q4, a negative coefficient indicates a disadvantage for females.) Further, we indicate which of the differences are significant based on controlling the FDR at 0.05 using the BH procedure described above. We observe there is no apparent relationship between the two variables. In both Q1 and Q4, fewer than half the differences are significant. Moreover, the significant differences are split between positive and negative. By contrast, the differences for the background variables are of one sign: positive for Q1 and negative for Q4—both indicating weaker associations with the outcome in M3 than in M1.

BH analyses (SF/Diff)

All analyses reported here were conducted on the KHB-adjusted differences. For SF/Diff we found $r = 14$. That is, all families were designated as significant. Consequently, the BH-adjusted p-values for the 19 differences in each of the 14 families were compared to 0.05 (or 0.005). A discovery was declared if the adjusted p-value was less or equal to 0.05 (or 0.005). The findings are displayed in Fig. 4. Panel (a) presents the numbers of discoveries using the 0.05 level of significance and Panel (b) presents the numbers of discoveries using the 0.005 level of significance. Each panel contains two stacked bars: One for Q1 and the other for Q4. The numbers of discoveries are

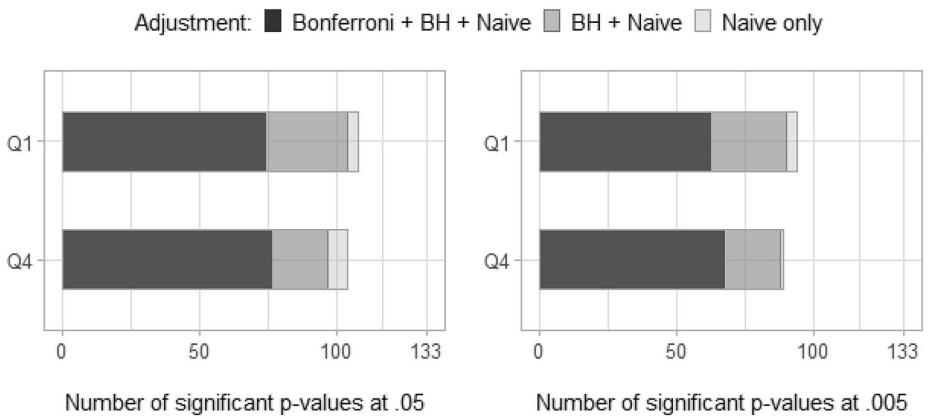

Fig. 4 Comparison of numbers of discoveries among procedures

compared among a procedure with no control on the family-wise error rate (i.e., per-comparison only), the BH procedure, and a Bonferroni procedure controlling the overall family-wise error rate.

For both Q1 and Q4, there is general agreement between the per-comparison procedure and the BH procedure, with only 4 discrepancies in Q1 and 7 discrepancies in Q4. In all 11 cases, the per-comparison procedure declared significance while the BH procedure did not, as expected. Note that more than 80 percent of the null hypotheses are rejected (i.e., declared "discoveries") by both procedures. Thus, there is strong evidence that there are, in fact, many discoveries to be made. As the quote below from Jones et al. (2001) asserts, in such cases the BH procedure should behave very much like the per-comparison procedure—which is indeed the case here.

> *In a situation where only very few of the results are definite, the "false discovery" approach performs much like the simultaneous approach. Since these situations include those where "definite" is purely accidental, as in extreme situations of data mining, it is important to have a severe procedure. In a situation in which most comparisons earn a definite result, the FDR algorithm behaves rather like the individual (i.e., unadjusted) procedure. Again, if many comparisons deserve a definite result, it is reasonable that looking for the most extreme is not going to lead to excessively many "errors." [p. 7131]*

Again, as expected, the BH procedure declares many more discoveries than the Bonferroni procedure, which is known to be more conservative. In fact, of the 104 discoveries declared by the BH procedure in Q1, 30 were not declared significant by the Bonferroni procedure. In Q4, of the 97 discoveries declared by the BH procedure, 21 were not declared significant by the Bonferroni procedure. Thus, in this case the apparent gains in power for BH over Bonferroni are approximately 40% (Q1) and 25% (Q4).

It has been argued that just reaching significance at the 0.05 level offers only weak evidence against the null hypothesis (Benjamin et al., 2018; Colquhoun, 2019), and the use of the 0.005 level has been recommended. With very large sample sizes, a more extreme threshold for significance seems a sensible choice. As already noted above, we carried out an equivalent analysis using the 0.005 level of significance. The patterns are similar

Table 4 Counts of discoveries for KHB differences, by variable [Q1]

	Adjustment at .005 level			Adjustment at .05 level		
	None	BH	Bonferroni	None	BH	Bonferroni
Gender	7	7	3	9	9	4
Age2	9	8	4	13	13	4
Age3	5	2	1	11	7	2
ParEd1	19	19	16	19	19	17
ParEd2	16	16	11	18	18	15
Books1	19	19	17	19	19	18
Books2	19	19	10	19	19	14
Total	94	90	62	108	104	74

Table 5 Counts of discoveries for KHB differences, by variable [Q4]

	Adjustment at .005 level			Adjustment at .05 level		
	None	BH	Bonferroni	None	BH	Bonferroni
Gender	7	6	3	9	8	3
Age2	6	6	3	12	10	3
Age3	3	3	2	7	3	3
ParEd1	19	19	18	19	19	19
ParEd2	18	18	11	19	19	15
Books1	19	19	18	19	19	18
Books2	17	17	12	19	19	15
Total	89	88	67	104	97	76

to those found with the 0.05 threshold, with the proviso that, as expected, the total number of discoveries is somewhat reduced.

At this juncture, it bears mentioning that controlling the FDR at a level q can be very different than controlling the family-wise error rate at that same level q. Indeed, when there are no true effects within the family of hypotheses being tested, controlling the FDR at level q and controlling the family wise error-rate at level q are equivalent. This justifies the use of the levels traditionally used for family-wise error rate control also for FDR control. However, when some of the null hypotheses are false (i.e., the corresponding effects are real), the FDR criterion considers the numbers of false rejections made (false discoveries) as a fraction of the number of rejections (total discoveries).

For the same number of false discoveries made, the larger the total number of rejections, the smaller is the false discovery proportion: Two false discoveries out of 4 discoveries is clearly unacceptable, but 2 false ones out of 100 discoveries is likely to be acceptable—whether we have tested 100, 1000 or more hypotheses. However, in both scenarios the family wise error-rate is 1. Thus, controlling the FDR at level q is an intuitive, common-sensical measure of error. It is equivalent to the family-wise approach when no true effects exist, but is more lenient and, therefore, more powerful when some effects are there to be discovered.

As noted earlier, there is interest in considering the findings for each factor in terms of the results for the set of variables used to represent that factor in the logistic regression

Table 6 Counts of discoveries for M3 coefficients, by variable [Q1]

	Adjustment at .005 level			Adjustment at .05 level		
	None	BH	Bonferroni	None	BH	Bonferroni
Gender	19	19	18	19	19	19
Age2	11	9	5	13	12	8
Age3	2	0	0	5	2	0
ParEd1	0	0	0	4	0	0
ParEd2	1	0	0	3	1	0
Books1	1	0	0	5	1	0
Books2	0	0	0	1	0	0
Cognitive Skills	14	13	8	18	18	11
EdAtt1	18	17	13	19	19	16
EdAtt2	15	14	10	18	18	11
EdAtt3	10	10	5	12	12	8
Total	91	82	59	117	102	73

model. Tables 4 and 5 present the results of the three significance testing procedures for each of the seven variables for Q1 (Table 4), separately, for Q4 (Table 5). Again, we observe that the per-comparison procedure and the BH procedure produce very similar numbers of discoveries in each family, with the Bonferroni procedure yielding somewhat fewer discoveries.

This table offers, *inter alia*, a convenient summary of the scatterplots presented earlier that describe the results of the KHB analyses. In that section, we observed that the coefficients of the variables representing Parental Education and Books in the Home shifted from strongly negative in M1 to weakly negative in M3 for Q4. Focusing here on the BH procedure, we see that for those variables the shifts are almost all statistically significant, at the 0.05 level and even at the 0.005 level. The situation for the coefficients of the two younger age categories is somewhat more mixed. In general, the coefficients shifted toward more negative values.

Substantively, there is particular interest in coefficient for gender (United Nations, 2021; World Economic Forum, 2019), which is effectively a measure of female disadvantage (holding other variables fixed). For both Q1 and Q4, in most countries the apparent gender disadvantage increased from M1 to M3. Overall, there were fewer significant shifts from M1 to M3. Moreover, as is evident in Fig. 3, the significant shifts are of mixed signs. That is, in some countries, the significant shift signaled an increase in disadvantage and, in others, it signaled a decrease in disadvantage. In short, the more proximal variables are very far from mediating the associations of gender, as well as of age, to the two outcomes. By contrast, the shifts in the coefficients for the other background factors were essentially all significant and in the direction of smaller absolute value (see next section).

BH analyses (SF/Coeff)

For this superfamily, with testing at the 0.05 level, null hypotheses were rejected for 20 out of 22 families (i.e., $r = 20$). The non-rejected variables were PARED1 and BOOKS2, both in Q1. For each of the 20 designated families, the BH-adjusted p-values for the

Table 7 Counts of discoveries for M3 coefficients, by variable [Q4]

	Adjustment at .005 level			Adjustment at .05 level		
	None	BH	Bonferroni	None	BH	Bonferroni
Gender	18	18	18	19	19	18
Age2	17	17	16	18	17	16
Age3	7	5	1	9	7	3
Pared1	2	0	0	5	4	0
Pared2	1	1	0	3	1	1
Books1	2	1	0	3	2	0
Books2	2	0	0	5	2	0
Cognitive Skills	18	17	17	18	18	17
EdAtt1	16	16	15	17	17	16
EdAtt2	19	19	19	19	19	19
EdAtt3	13	13	9	13	13	12
Total	115	107	95	129	119	102

19 coefficients were compared to $0.045 = 0.05 \times 10/11$. A discovery was declared if the adjusted p-value was less or equal to 0.045. We then replicated the above analyses but replacing 0.05 by 0.005. In this case there were five non-rejected variables in Q1 (AGE3, PARED1, PARED2, BOOKS1, BOOKS2) and two in Q4 (PARED1, BOOKS2). Consequently for each of the $22–7 = 15$ families designated as significant, the BH-adjusted p-values for the 19 coefficients were compared to $0.005\ (15/22) = 0.003$.

The findings are displayed in Tables 6 and 7. Table 6 presents the results for Q1 and Table 7 for Q4. For both outcomes, there is again general agreement between the per-comparison procedure and the BH procedure in the numbers of significant findings in each family, with the Bonferroni procedure typically yielding somewhat fewer discoveries. The coefficients of variables representing gender, the younger age group, cognitive skills and educational attainment were significant for nearly all countries. In sum, the addition of Cognitive Skills and Educational Attainment to M1 reduces the coefficients of the variables representing Parental Education and Books in the Home to non-significance. This is also true of the middle age group. With regard to gender, the coefficients are significantly different from 0 for all 19 countries at both the 0.05 and 0.005 levels. The coefficients for gender are all negative, implying that there is a disadvantage for females, even after controlling for all the other factors in the model. It is noteworthy that this is the case for all the countries in the study.

The country view

To this point, we have documented the overall impact on the results of taking into account both the scale changes in nested logistic regressions and multiplicity in significance testing. This is appropriate for a methodological investigation. However, since countries are the principal audiences for these large-scale assessments, it is also of interest to examine the changes in inferences made at the country level. Accordingly, we have selected two countries, Japan and Germany, for a more detailed review. Japan was of interest due to the greatest gender disadvantage observed, with a coefficient of 2.39 for the Q1 outcome and -1.53 for the Q4 outcome based on the results for M3. We also

Table 8 Conventional differences and KHB differences for coefficients [Q1 and Q4]: Japan

Variable	Q1					Q4				
	Diff (M1–M3)	Diff_KHB	Sig. at 0.005			Diff (M1–M3)	Diff_KHB	Sig. at 0.005		
			p	p_BH	p_Bonferroni			p	p_BH	p_Bonferroni
Gender	0.01	0.18	✦	✦		−0.11	−0.19	✦	✦	✦
Age2	−0.09	−0.02				0.16	0.05			
Age3	0.01	0.02				0.08	0.01			
ParEd1	0.41	0.47	✦	✦	✦	−0.34	−0.39	✦	✦	✦
ParEd2	0.27	0.30	✦	✦	✦	−0.23	−0.26	✦	✦	✦
Books1	0.44	0.47	✦	✦	✦	−0.36	−0.42	✦	✦	✦
Books2	0.22	0.24	✦	✦	✦	−0.16	−0.19	✦	✦	✦

Table 9 Results from M3 [Q1 and Q4]: Japan

Variable	Q1				Q4			
	M3	Sig. at 0.005			M3	Sig. at 0.005		
		p	p_BH	p_Bonferroni		p	p_BH	p_Bonferroni
Gender	2.39	✦	✦	✦	− 1.53	✦	✦	✦
Age2	0.30				− 2.04	✦	✦	✦
Age3	− 0.09				− 0.61	✦	✦	
ParEd1	0.03				− 0.11			
ParEd2	0.07				− 0.18			
Books1	0.12				− 0.62	✦	✦	
Books2	0.04				− 0.47	✦		
Cognitive Skills	− 0.29	✦	✦		0.39	✦	✦	✦
EdAtt1	1.42	✦	✦	✦	− 0.60			
EdAtt2	1.08	✦	✦	✦	− 0.84	✦	✦	✦
EdAtt3	0.74	✦	✦		− 0.77	✦	✦	✦

selected Germany, a developed country with a rather large gender pay gap in comparison to other European countries (European Commission, 2021). Inference is conducted at the $p = 0.005$ level. Tables 8 and 9 display the results for Japan, with Panels (a) containing findings for Q1 and Panels (b) for Q4. Tables 10 and 11 display the results for Germany, with Panels (a) containing findings for Q1 and Panels (b) for Q4.

Japan

With respect to Q1, the effect of the KHB adjustment is substantially greater for Gender than for the other variables and the change is significant according to BH, but not Bonferroni. The coefficient in M3 is smaller than in M1 but still strongly significant ($z = 21.7$) according to both BH and Bonferroni. Recall that for Q1 a positive coefficient indicates a disadvantage (i.e., greater odds of being in Q1). The changes for the age coefficients are not significant and neither are the corresponding M3 coefficients. By contrast, the changes in the coefficients for Parental Education and Books in the Home are significant, and they all result in sharply reduced values in M3. BH and Bonferroni agree that the M3 coefficients are not significantly different from 0. Finally, the coefficients for Cognitive Skills and levels of Educational Attainment (M3) are declared significant both by BH and by Bonferroni. Thus, these latter two factors play a very strong mediating role for all background factors, with the exception of Gender. The positive coefficients for levels of Educational Attainment indicate disadvantage. The negative coefficient for Cognitive Skills can be interpreted as higher skill levels lower disadvantage (i.e., lower odds of being in Q1). It is noteworthy that the coefficient for Gender is nearly twice the size for the next largest coefficient, corresponding to the lowest level of education.

With respect to Q4, the effect of the KHB adjustment is greatest for the youngest age category and for Gender. Nonetheless, BH and Bonferroni agree that only the differences in coefficients for Gender, Parental Education, and Books in the Home are significant. The coefficient for Gender is slightly smaller in M3 but still significant ($z = -12.9$), indicating strong disadvantage (i.e., lower odds of being in Q4). The coefficient for the youngest age category is larger in absolute magnitude than that for Gender and, with

Table 10 Conventional differences and KHB differences for coefficients [Q1 and Q4]: Germany

Variable	Q1					Q4				
	Diff (M1–M3)	Diff_KHB	Sig. at 0.005			Diff (M1-M3)	Diff_KHB	Sig. at 0.005		
			p	p_BH	p_Bonferroni			p	p_BH	p_Bonferroni
Gender	− 0.03	0.08				0.15	− 0.09			
Age2	− 0.09	− 0.04				0.32	0.10			
Age3	− 0.08	− 0.06				0.15	0.15			
ParEd1	0.67	0.71	✦	✦	✦	− 0.92	− 1.08	✦	✦	✦
ParEd2	0.32	0.33	✦	✦	✦	− 0.46	− 0.51	✦	✦	✦
Books1	0.54	0.58	✦	✦	✦	− 0.90	− 0.96	✦	✦	✦
Books2	0.28	0.27	✦	✦	✦	− 0.35	− 0.41	✦	✦	✦

Table 11 Results from M3 [Q1 and Q4]: Germany

Variable	Q1				Q4			
	M3	Sig. at 0.005			M3	Sig. at 0.005		
		p	p_BH	p_Bonferroni		p	p_BH	p_Bonferroni
Gender	1.67	✦	✦	✦	− 1.09	✦	✦	✦
Age2	0.39				− 1.46	✦	✦	✦
Age3	− 0.10				− 0.23			
ParEd1	− 0.06				− 0.53			
ParEd2	− 0.16				0.07			
Books1	0.00				− 0.14			
Books2	− 0.10				− 0.15			
Cognitive Skills	− 0.36	✦	✦	✦	0.81	✦	✦	✦
EdAtt1	1.75	✦	✦	✦	− 2.47	✦	✦	✦
EdAtt2	1.09	✦	✦	✦	− 1.76	✦	✦	✦
EdAtt3	0.68	✦	✦		− 1.36	✦	✦	✦

$z = -10.5$, declared significant by both BH and Bonferroni. The M3 coefficients for Parental Education and Books in the Home are smaller in absolute value than the corresponding M1 coefficients and are not significant with the exception of the lowest level of Books in the Home (BH only). The coefficients for Cognitive Skills and Educational Attainment are all significant (BH and Bonferroni), with the exception of the lowest level of education. For the Q4 outcome, after accounting for all measured variables, the greatest disadvantage accrues to individuals in the lowest age group and to females.

Germany

With respect to Q1, the impact of the KHB adjustments are uniformly small. Although the KHB differences for Gender and Age are not significant, the differences for all the background variables are significant. The M3 coefficients for Gender, Cognitive Skills and Educational Attainment are all significant. The coefficients for the background variables are no longer significant. Thus, the more proximal factors do act as mediators for the background factors.

With respect to Q4, the impact of the KHB adjustments are greatest for Gender and the youngest age category. However, only the KHB differences for the background variables are significant. Turning to the M3 coefficients, only those corresponding to Gender, the youngest age category, Cognitive Skills and Educational Attainment are significant. Once again, the more proximal factors act as mediators for the background factors.

Limitations

It is typical for logistic regression models applied to social science data to account for only a modest amount of the variance in the outcomes. In the present case, employing Tjur's D-statistic, the M3 models typically accounted for 14 percent of the variance for Q1 and 21 percent of the variance for Q4. Accordingly, it is prudent not to over-interpret the estimated coefficients. Nonetheless, the consistency of the results across countries (e.g., the apparent substantial disadvantage for females in both Q1 and Q4) strengthens the case for the existence of the phenomenon. Of course, in this and other settings,

statistical findings can only provide an indication of interesting directions to pursue to inform policy makers. At the country level, further analysis must involve bringing to bear historical, cultural, economic and political considerations in support of particular decisions and actions.

An additional limitation worth mentioning is that by using only the first plausible value we were not able to incorporate the imputation variance in our results. Similar studies conducted in the future should follow appropriate procedures described in the technical documentation (OECD, 2019). As there are no programs readily available, incorporating the implications of the sampling design and assessment design in the variance calculations was also not possible.

Conclusions

The goal of this article was to highlight—and address—two of the difficulties that arise in the analysis of a nested pair of logistic regression models, with both incorporating a number of predictor variables. The first difficulty is due to the fact that a change in the set of predictor variables causes a change in the scale of the model. This change, in turn, means that for a variable that appears in both models, the raw difference in the estimated coefficients is not an unbiased estimate of the true difference in the coefficients. Rather, it confounds the true difference with the impact of the scale change and, consequently, in order to obtain an unbiased estimate, an estimate of that impact must be obtained and removed. To carry out this procedure, we employed a strategy developed by Karlson et al. (2012) and refined by Breen et al. (2018).

The second difficulty arises when a large number of inferences, in the form of tests of null hypotheses, are of interest. This induces a problem of multiplicity that, if not heeded, can lead to spurious findings of significance. There are a number of well-known methods for controlling Type I error rates. However, we chose instead to employ a method to control the False Discovery Rate and compared its operating characteristics to two conventional methods.

The recommended methodology was illustrated using a subset of the data analyzed by Braun (2018), drawn from PIAAC, an international, large-scale assessment of adults. One of the questions investigated in that paper concerned outcomes related to the location of an individual's annual income in the national income distribution. Specifically, what was the relationship of an individual's measured characteristics to the probability that the individual's income was located in the lowest quartile (Q1) and, separately, that the individual's income was located in the highest quartile (Q4). The nesting arose naturally, based on the temporal relationships among the factors. For each quartile, the base model incorporated variables representing demographic and background factors. The extended model also incorporated variables representing measured cognitive skills and educational attainment. The focal question was the extent to which the more proximal, additional variables played the roles of mediators to the more distal variables in the base model.

The pair of nested models was fit separately for 19 countries for each of the quartiles. There was interest not only in comparing differences across countries in the coefficients for each variable in the base model, but also in the magnitudes of the coefficients in the full model. An added difficulty arose because the data were obtained through a complex

survey design. Accordingly, sampling weights were employed to obtain approximately unbiased estimates of the regression coefficients, based on suitable modifications to the KHB method that were devised and implemented.

Overall, the differences between the changes in coefficients calculated conventionally and with the KHB adjustment varied from negligible to very substantial. An example of the latter is offered by the Q1 analysis for Japan. The conventional difference in the coefficients for gender is 0.012 and, with the KHB adjustment, 0.18. In the Q4 analysis for Japan, the conventional difference in the coefficients for gender is −0.11 and with the KHB adjustment, −0.19. In general, the KHB-adjusted differences were larger and significant for the variables representing the background factors Parental Education and Books in the Home. For Gender and Age, the KHB-adjusted differences were mostly smaller and non-significant. When combined with the actual magnitudes of the coefficients for these variables in the two models, we concluded that the more proximal factors indeed act as strong mediators for the background factors, but less so for Age, and hardly at all for Gender.

With respect to multiplicity, applying the FDR-controlling procedure at level q (BH_q) yielded results very similar to those obtained by applying a standard per-comparison procedure at the same level q. As expected, BH_q had slightly fewer discoveries. In comparison to the Bonferroni procedure at level q, BH_q yielded quite a few more discoveries, again as expected. It is important to remember, however, that with the BH_q we are assured of controlling the FDR at the nominal level. No such assurance is offered by the per-comparison procedure. Moreover, BH_q not only controls the overall FDR, but also the FDR for each family of hypotheses corresponding to each variable appearing in both models. This is important if there is interest in the inferences regarding specific variables (e.g., gender).

The two-stage BH_q procedure employed here takes account of the hierarchical structure of the hypotheses. It was applied separately to two "super-families" of hypotheses: One corresponding to the differences in the coefficients between models and one to the magnitudes of the coefficients in the extended model. In principle, the two super-families could be combined to constitute a yet higher (i.e., third) level of the hypothesis testing framework. This would be appropriate if it could be argued that finding a significant coefficient would be of interest, even if no significant differences were observed in any of the coefficients. A side calculation (not shown) reveals that viewing the problem as a two-level or as a three-level hierarchy yields the same results.

It bears mentioning that the KHB methodology illustrated here can be applied wherever there is interest in comparing nested logistic regressions. The categorization of variables and the order of entry should be determined by substantive considerations. On the other hand, the BH procedure is perfectly general and can be applied to address multiplicity issues in a broad range of circumstances.

Appendix

See Table 12.

Table 12 Country names and abbreviations

Country abbr	Country name	Sample size [Q1]	Sample size [Q4]
AUT	Austria	2581	1851
BEL	Belgium	2286	1739
CZE	Czech Republic	2079	1807
DEU	Germany	2487	1696
DNK	Denmark	3082	2476
ESP	Spain	2234	1789
EST	Estonia	2955	2561
FIN	Finland	2528	2154
FRA	France	2609	2179
GBR	United Kingdom	3476	2574
IRL	Ireland	2506	1892
ITA	Italy	1952	1577
JPN	Japan	2357	1811
KOR	Korea	3208	2707
NLD	Netherlands	2375	1384
NOR	Norway	2616	2153
POL	Poland	2351	2061
SWE	Sweden	2107	1690
USA	United States	2162	1700

Abbreviations
BH: Benjamini–Hochberg; FDR: False Discovery Rate; GLM: Generalized linear model; IDB: International database; IEA: International Association for the Evaluation of Educational Achievement; KHB: Karlson–Holm–Breen; OECD: Organization for Economic Cooperation and Development; PIAAC: Programme for the International Assessment of Adult Competencies; PV: Plausible values.

Acknowledgements
None.

Authors' contributions
GG: Researched the scaling problem endemic to nested logistic regressions. Wrote all computer programs and conducted the analyses. Contributed to the exposition of the application to the data. YB: Developed the hierarchical FDR-controlling procedure and contributed to the exposition of the application to the data. HB: Conceptualized and co-ordinated the project, based on earlier research. Main writer of successive drafts. All authors read and approved the final manuscript.

Funding
None.

Availability of data and materials
All data and materials are publicly available.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare they have no competing interests.

Author details
[1]Boston College, Chestnut Hill, MA, USA. [2]Tel-Aviv University, Ramat Aviv, Tel Aviv, Israel.

Received: 8 January 2021 Accepted: 13 July 2021
Published online: 21 July 2021

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Bogomolov, M., Peterson, C. B., Benjamini, Y., & Sabatti, C. (2020). Hypotheses on a tree: New error rates and testing strategies. *Biometrika*. https://doi.org/10.1093/biomet/asaa086

Braun, H. (2018). How long is the shadow? The relationships of family background to selected adult outcomes. *Large-Scale Assessments in Education, 6*(4), 1–52. https://doi.org/10.1186/s40536-018-0058-x

Braun, H. I., & Tukey, J. W. (1983). Multiple comparisons through orderly partitions: The maximum subrange procedure. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift for Frederic M. Lord* (pp. 55–65). Routledge. https://doi.org/10.4324/9780203056653

Braun, H., & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-Scale Assessments in Education, 5*(17), 1–16. https://doi.org/10.1186/s40536-017-0050-x

Breen, R., Karlson, K. B., & Holm, A. (2018). A note on a reformulation of the KHB method. *Sociological Methods & Research*. https://doi.org/10.1177/0049124118789717

Colquhoun, D. (2019). The false positive risk: A proposal concerning what to do about p-values. *The American Statistician, 73*(Issue sup 1), 192–201. https://doi.org/10.1080/00031305.2018.1529622

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models* (3rd ed.). Chapman & Hall/CRC.

European Commission. (2021, May 2). The gender pay gap situation in the EU. Retrieved from European Commission website: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/gender-equality/equal-pay/gender-pay-gap-situation-eu_en

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316672

IEA. (2019). *Help manual for the IEA IDB analyzer (Version 4.0)*. Hamburg, Germany. Retrieved from www.iea.nl/data.htm

Jones, L. V., Lewis, C., & Tukey, J. W. (2001). Hypothesis tests, multiplicity of. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 7127–7133). Elsevier

Karlson, K. B., Holm, A., & Breen, R. (2012). Comparing regression coefficients between same-sample nested models using logit and probit: A new method. *Sociological Methodology, 42*, 286–313. https://doi.org/10.1177/0081175012444861

Lumley, T. (2020). survey: Analysis of complex survey samples. *R package version 4.0*

OECD. (2019). *Technical report of the survey of adult skills (PIAAC)*. OECD Publishing

R Core Team. (2020). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.R-project.org/

United Nations. (2021, May 2). International Equal Pay Day. Retrieved from United Nations. https://www.un.org/en/observances/equal-pay-day#:~:text=Equal%20pay%20for%20work%20of%20equal%20value&text=Across%20all%20regions%2C%20women%20are,at%2023%20per%20cent%20globally

Wickham, H., François, R., Henry, L., & Müller, K. (2020). dplyr: A grammar of data manipulation. *R Package Version, 1*, 2.

World Economic Forum. (2019). *Global gender gap report 2020*. Geneva, Switzerland: World Economic Forum. Retrieved from Sustainable Development Goals. http://www3.weforum.org/docs/WEF_GGGR_2020.pdf

Publisher's Note