What Is in a Random Effect?
The Role of Measurement Equivalence in Multilevel Analysis

Marco R. Steenbergen

University of Zurich

Paper prepared for the Annual Meeting of the Società Italiana di Scienza Politica
Lecce, September 12-14, 2019

Abstract

In multilevel models of cross-national surveys, random effects arise in one of two ways. First, there may be differences in the structure of the data generating process. This generally is of considerable interest and one reason to use multilevel models in the first place. Second, random effects may capture differential item function (DIF). This results in a vastly different interpretation of the data. This paper demonstrates the dual origins of random effects, illustrates the biases that might emerge, and discusses strategies that allow for a better interpretation of multilevel models.

Comparative survey research has profited greatly from advances in multilevel statistical modelling. Multilevel models are now routinely used on cross-national survey data (an incomplete list of examples includes Birch, 2008; Fairbrother & Martin, 2013; Harteveld, Van Der Meer, & De Vries, 2013; Steenbergen & Jones, 2002; Stoeckel, 2013; Strabac & Listhaug, 2008). A crucial benefit of those models is that they can shed light on heterogeneous data patterns across countries. Typically, evidence of heterogeneity is given a substantive interpretation. As I shall demonstrate in this paper, however, such interpretations can be misleading in the presence of differential item functioning (DIF). Indeed, the presence of DIF can easily suggest heterogeneity where none exists or camouflage real differences between countries.[1]

The impact of DIF on random effects has hitherto received scant attention in the multilevel literature. Neither has it received much attention in the literature on DIF, which generally focuses on fixed effects (e.g., De Beuckelaer & Swinnen, 2011; Mellenbergh, 1989; Meredith, 1993; Van De Vijver & Poortinga, 1997). This is an important omission for several reasons. First, DIF is likely in cross-national research (Harzing, 2006). Second, random effects, as manifested through intra-class correlations, are often used to determine whether it is necessary to conduct multilevel analysis. If evidence of random effects is masked, then researchers may erroneously opt out of the appropriate method. Third, if evi-

[1] The paper focuses on countries, but the discussion is relevant to all groups that may be affected by DIF.

dence of random effects attains, then researchers often follow up by adding contextual predictors. However, this exercise might prove futile when random effects merely reflect DIF.

For these reasons, it is important that researchers are aware of the effects of DIF in multilevel analysis and that they develop strategies to avoid incorrect inferences. This paper outlines the scope of the problem and discusses how to best perform multilevel analysis of cross-national survey data.

Theory

Measurement Model Consider a respondent $i = 1, \cdots, n_j$ in country $j = 1, \cdots, J$. For this unit, we stipulate the existence of a latent trait η_{ij}. In cross-national survey research, this will typically be an attitude, belief, opinion, or value. Assume that we use M items to measure the trait; these are contained in the vector $\mathbf{y}_{ij}^\top = (y_{ij1}, y_{ij2}, \cdots, y_{ijM})$. We postulate the following measurement model:[2]

$$\mathbf{y}_{ij} = \boldsymbol{\tau}_j^y + \boldsymbol{\lambda}_j^y \eta_{ij} + \boldsymbol{\epsilon}_{ij}^y \tag{1}$$

Here, $\boldsymbol{\epsilon}_{ij}^y \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ is a vector of measurement errors, $\left(\boldsymbol{\tau}_j^y\right)^\top = (\tau_{j1}^y, \cdots, \tau_{jM}^y)$ is a vector of measurement intercepts, and $\left(\boldsymbol{\lambda}_j^y\right)^\top = (\lambda_{j1}^y, \cdots, \lambda_{jM}^y)$ is a vector of measurement slopes/factor loadings.

Measurement equivalence pertains to the fundamental question "of whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (Horn & McArdle, 1992, p. 117). To say anything meaningful about trait differences between countries, one needs to ensure that trait measures operate equivalently, lest one infers differences that do not exist.

One can identify various forms of measurement equivalence but, for our purposes, two carry particular weight. Metric invariance exists when the measurement slopes are identical across countries: $\boldsymbol{\lambda}_1^y = \boldsymbol{\lambda}_2^y = \cdots = \boldsymbol{\lambda}_J^y = \boldsymbol{\lambda}^y$ (Rock, Werts, & Flaugher, 1978). Scalar invariance combines this requirement with the condition that the measurement slopes are identical: $\boldsymbol{\tau}_1^y = \boldsymbol{\tau}_2^y = \cdots = \boldsymbol{\tau}_J^y = \boldsymbol{\tau}^y$ (Meredith, 1993). We shall see that both forms of invariance play a central role in shaping random effects in multilevel analysis.

Equivalence and Multilevel Analysis How is equivalence relevant for multilevel modelling? I start with the innocuous assumption that cross-national survey researchers are interested in capturing trait differences across countries. Measures, as such, are interesting only to the extent that they shed light on those differences. Let us assume that the latent trait abides

$$\eta_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\gamma} + \theta_{ij}, \tag{2}$$

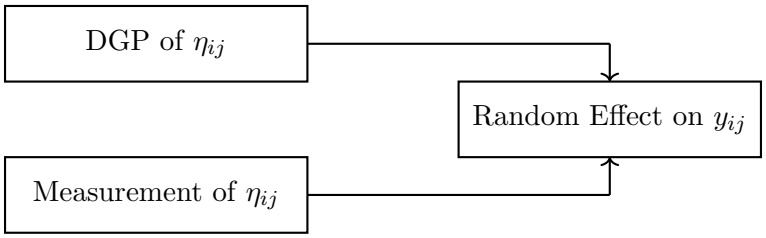
where $\theta_{ij} \sim \mathcal{N}(0, \sigma^2)$ is the level-1 or respondent-level error term, $\boldsymbol{\gamma}$ is a vector of fixed effects, and $\mathbf{x}_{ij}$ are the P predictors and constant, which we also assume to be fixed for now. Central to the model specification in Equation (2) is the absence of random effects: the data generating process (DGP) is identical across all countries.

Obviously, we cannot estimate Equation (2) due to the latent nature of the outcome variable. Substituting Equation (2) into Equation (1) yields

$$\mathbf{y}_{ij} = \boldsymbol{\tau}_j^y + \boldsymbol{\lambda}_j^y \mathbf{x}_{ij}^\top \boldsymbol{\gamma} + \theta_{ij} \boldsymbol{\lambda}_j^y + \boldsymbol{\epsilon}_{ij}^y \tag{3}$$

[2] I use a factor analytic model (cf. Davidov, 2009) but the ideas generalise to item response models (see the Appendix).

Figure 1: Two Sufficient Conditions for Random Effects in a Random Coefficient Model



This is a model that we actually estimate.

The crux is that Equation (3) is a random coefficient model (RCM), despite the DGP for the latent trait being devoid of any random effects. Specifically, Equation (3) may be rewritten as

$$\mathbf{y}_{ij} = \mathbf{A}_j \mathbf{x}_{ij} + \boldsymbol{\nu}_{ij} \tag{4}$$

where $\boldsymbol{\nu}_{ij} = \theta_{ij}\boldsymbol{\lambda}^y_j + \boldsymbol{\epsilon}^y_{ij}$ and $\mathbf{A}_j$ is

$$\begin{pmatrix} \tau^y_{j1} + \lambda^y_{j1}\gamma_{00} & \lambda^y_{j1}\gamma_{10} & \cdots & \lambda^y_{j1}\gamma_{P0} \\ \tau^y_{j2} + \lambda^y_{j2}\gamma_{00} & \lambda^y_{j2}\gamma_{10} & \cdots & \lambda^y_{j2}\gamma_{P0} \\ \vdots & \vdots & \ddots & \vdots \\ \tau^y_{jM} + \lambda^y_{jM}\gamma_{00} & \lambda^y_{jM}\gamma_{10} & \cdots & \lambda^y_{jM}\gamma_{P0} \end{pmatrix} \tag{5}$$

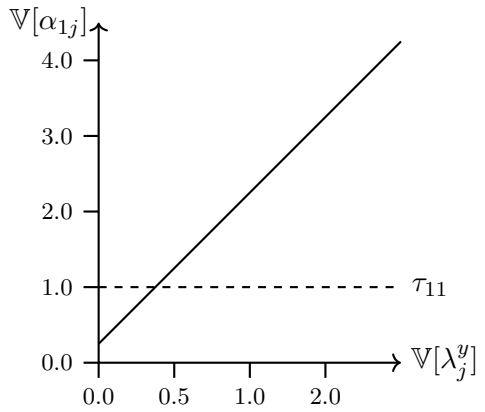
Thus, $\mathbf{A}_j$ is a matrix of random effects—random intercepts in the first column and random slopes in the remaining columns. However, these are not the kinds of random effects one would normally uncover in an RCM as markers of heterogeneity across countries (Swamy & Tavlas, 1995). Instead, they are artefacts of DIF. Put differently, had one been able to obtain invariant measures of the latent trait, then the random effects would not have occurred.

The estimated model does justice to the latent DGP only under scalar invariance. In this case, the intercepts and slopes take the form of $\tau^y_k + \lambda^y_k\gamma_{00}$ and $\lambda^y_k\gamma_{p0}$, respectively. The absence of a j-subscript means that the variation across countries has been removed. Scalar invariance suffices to remove the random slopes but not the random intercepts.

As an illustration, consider the model $\eta_{ij} = \gamma_{00} + \gamma_{10}x + ij + \theta_{ij}$. For simplicity, we assume that the latent trait is measured through a single indicator, y_{ij}, and that we know the variation in the measurement intercepts and slopes. In this case, the estimated model is $y_{ij} = \tau^y_j + \lambda^y_j\gamma_{00} + \lambda^y_j\gamma_{10}x_{ij} + \lambda^y_j\theta_{ij} + \epsilon^y_{ij}$. Letting $\alpha_{0j} = \tau^y_j + \lambda^y_j\gamma_{00}$ and $\alpha_{1j} = \lambda^y_j\gamma_{10}$, this may be written as $y_{ij} = \alpha_{0j} + \alpha_{1j}x_{ij} + \lambda^y_j\theta_{ij} + \epsilon^y_{ij}$. Using basic probability theory, it can be shown that (1) $\mathbb{V}[\alpha_{0j}] = \mathbb{V}[\tau^y_j] + \gamma^2_{00}\mathbb{V}[\lambda^y_j]$ and (2) $\mathbb{V}[\alpha_{1j}] = \gamma^2_{10}\mathbb{V}[\lambda^y_j]$, where $\mathbb{V}$ denotes the variance. It is easy to see that metric invariance results in a zero variance component for the slope: $\mathbb{V}[\alpha_{1j}] = \gamma^2_{10} \cdot 0 = 0$. Under metric invariance, however, the variance in the intercept remains non-zero: $\mathbb{V}[\alpha_{0j}] = \mathbb{V}[\tau^y_j] + \gamma^2_{00} \cdot 0 = \mathbb{V}[\tau^y_j] > 0$. To remove this variance component, we require scalar invariance so that $\mathbb{V}[\tau^y_j] = 0$.

The discussion so far reveals that there are two sufficient conditions for generating random effects in multilevel analysis (see Figure 1). First, random effects may reflect cross-national differences in the DGPs for the latent trait of interest, which is considerable theoretic significance. Second, random effects may be products of DIF in

Figure 2: Impact of DIF on the Variance Component of a Random Slope Model



Notes: The dashed line represents $\tau_{11} = 1.0$. The solid line shows $\mathbb{V}[\alpha_{1j}]$ at different values of $\mathbb{V}[\lambda_j^y]$, assuming $\mathbb{E}[\lambda_j^y] = 0.5$ and $\gamma_{10} = 1.0$. Where the solid line drops below the dashed line, DIF has the effect of suppressing the variance component for random effects in the latent DGP. Where it extends above the dashed line, DIF causes an exaggeration of the variance in the slopes.

the measurement of the latent trait, a possibility that bears no theoretic relevance. Importantly, heterogeneity in the DGP of the latent variable is not a necessary condition for obtaining evidence of random effects.

DIF in RCMs In line with Figure 1, imagine there is heterogeneity in the DGP of the latent trait. In this case, the RCM takes the following form:

$$\eta_{ij} = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta}_j + \theta_{ij} \quad (6)$$

$$\boldsymbol{\beta}_j = \boldsymbol{\gamma} + \boldsymbol{\delta}_j \quad (7)$$

Here $\boldsymbol{\delta}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{T})$ are the level-2 or country-level errors; $\mathbf{T}$ is the variance-covariance matrix of the random effects. If we substitute this into Equation (1, we obtain

$$y_{ij} = \mathbf{A}_j\mathbf{x}_{ij} + \mathbf{B}_j\mathbf{x}_{ij} + \boldsymbol{\nu}_{ij} \quad (8)$$

where $\mathbf{B}_j$ is

$$\begin{pmatrix} \lambda_{j1}^y\delta_{0j} & \lambda_{j1}^y\delta_{1j} & \cdots & \lambda_{j1}^y\delta_{Pj} \\ \lambda_{j2}^y\delta_{0j} & \lambda_{j2}^y\delta_{1j} & \cdots & \lambda_{j2}^y\delta_{Pj} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{jM}^y\delta_{0j} & \lambda_{jM}^y\delta_{1j} & \cdots & \lambda_{jM}^y\delta_{Pj} \end{pmatrix} \quad (9)$$

Thus, the random effects in the model are driven by level-2 error terms, as well as DIF.

As a simple illustration consider the random slope model with a single measure of the outcome and a single predictor. Here, Equation (8) reduces to $y_{ij} = \tau_j^y + \gamma_{00}\lambda_j^y + \beta_{1j}\lambda_j^y x_{ij} + \lambda_j^y\theta_{ij} + \epsilon_{ij}^y$. Setting $\alpha_{0j} = \tau_j^y + \gamma_{00}\lambda_j^y$, $\alpha_{1j} = \beta_{1j}\lambda_j^y$, and $\nu_{ij} = \lambda_j^y\theta_{ij} + \epsilon_{ij}^y$, this can be written as $y_{ij} = \alpha_{0j} + \alpha_{1j}x_{ij} + \nu_{ij}$. From our earlier discussion, we know that $\mathbb{V}[\alpha_{0j}] = \mathbb{V}[\tau_j^y] + \gamma_{00}^2\mathbb{V}[\lambda_j^y]$. Following Goodman (1960), it can also be demonstrated that $\mathbb{V}[\alpha_{1j}] = \tau_{11}\mathbb{V}[\lambda_j^y] + \left(\mathbb{E}[\lambda_j^y]\right)^2\tau_{11} + \gamma_{10}^2\mathbb{V}[\lambda_j^y]$, where $\tau_{11} = \mathbb{V}[\beta_{1j}]$. It is clear that $\mathbb{V}[\alpha_{1j}] \neq \tau_{11}$,

i.e., the variation in the slopes in the presence of DIF is not the same as the variation in the slopes in the DGP of the latent trait of interest. Crucially, $\mathbb{V}[\alpha_{1j}]$ can be either larger or smaller than τ_{11}. The latter result is important because, until now, we considered situations in which DIF causes us to infer random effects that do not exist in the DGP—false positives, so to speak. We now have a situation where we might downplay true random effects, potentially producing false negatives.

Figure 2 illustrates the problem. It depicts the impact of heterogeneous factor loadings for a scenario in which τ_{11}, γ_{10}, and $\mathbb{E}[\lambda_j^y]$ have been fixed. We observe that the variance in the slopes of the estimated model sometimes exceeds τ_{11} and sometimes falls short off it, depending on the level of heterogeneity in the factor loadings.

The problem that the estimable variance component deviates from the latent variance component is not limited to random slope models. Consider the ubiquitous random intercept model, again with a single measure of the outcome variable and a single predictor. Here, we have $y_{ij} = \tau_j^y + \lambda_j^y \beta_{0j} + \lambda_j^y \gamma_{10} x_{ij} + \lambda_j^y \theta_{ij} + \epsilon_{ij}^y$. Setting $\alpha_{0j} = \tau_j^y + \lambda_j^y \beta_{0j}$ and $\alpha_{1j} = \lambda_j^y \gamma_{10}$, we can also write the model as $y_{ij} = \alpha_{0j} + \alpha_{1j} x_{ij} + \nu_{ij}$. From our earlier discussion, $\mathbb{V}[\alpha_{1j}] = \gamma_{10}^2 \mathbb{V}[\lambda_j^y]$. Following Goodman (1960), we can further demonstrate that $\mathbb{V}[\alpha_{0j}] = \mathbb{V}[\tau_j^y] + \left(\mathbb{E}[\lambda_j^y]\right)^2 \tau_{00} + \gamma_{00}^2 \mathbb{V}[\lambda_j^y] + \tau_{00} \mathbb{V}[\lambda_j^y]$, where τ_{00} is the variance of the random intercepts. Once more, $\mathbb{V}[\alpha_{0j}] \neq \tau_{00}$.

DIF in Predictors So far, we have discussed the effects of DIF in measurements of the outcome variable. However, DIF also impacts multilevel analysis when it occurs for the predictor variables. Consider the multilevel model $\eta_{ij} = \gamma_{00} + \gamma_{10} \xi_{ij} + \theta_{ij}$, where η and ξ are latent traits. For simplicity, we assume $y_{ij} = \eta_{ij}$. The model linking the latent predictor to a measurement is $x_{ij} = \tau_j^x + \lambda_j^x \xi_{ij} + \epsilon_{ij}^x$. Simple algebra now yields

$$y_{ij} = \gamma_{00} - \gamma_{10} \frac{\tau_j^x}{\lambda_j^x} + \frac{\gamma_{10}}{\lambda_j^x} x_{ij} + \theta_{ij} - \frac{\gamma_{10}}{\lambda_j^x} \epsilon_{ij}^x \quad (10)$$

This is again an RCM of the form $y_{ij} = \alpha_{0j} + \alpha_{1j} x_{ij} + \nu_{ij}$, where $\alpha_{0j} = \gamma_{00} - \gamma_{10} \tau_j^x / \lambda_j^x$, $\alpha_{1j} = \gamma_{10} / \lambda_j^x$, and $\nu_{ij} = \theta_{ij} - \epsilon_{ij}^x \gamma_{10} / \lambda_j^x$. It is clear that the random slopes disappear under metric invariance, whereas scalar invariance is required to remove the random intercept. If scalar invariance does not hold, one will find evidence of random effects even though the latent DGP does not contain them.

Discussion The theoretical discussion of the intersection between multilevel analysis and measurement makes two points abundantly clear. First, the DGP for a latent variable may be devoid of any random effects but one may still uncover random effects in the presence of DIF in the measures of the outcome variable and/or the predictors. Second, if the DGP of a latent outcome contains random effects, then the presence of DIF may either amplify those effects or mask them.

Illustration

We illustrate the implications of DIF for multilevel analysis using perceptions of political competition collected during the sixth round of the European Social Survey (ESS). The data were collected in 29 countries: Albania (AL), Belgium (BE), Bulgaria (BG), Cyprus (CY), the Czech Republic (CZ), Denmark (DK), Estonia (EE),

Table 1: Measurement Properties of Perceived Competition

Invariance Type	CFI	RMSEA	ΔCFI	ΔRMSEA
Configural	1.000	0.000	—	—
Metric	0.978	0.071	-0.022	0.071
Scalar	0.733	0.174	-0.245	0.103

Notes: Based on maximum likelihood estimates with the Satorra-Bentler non-normality correction of the test statistics and survey design weights. Changes (Δ) are measured relative to the previous model. $N = 49,807$, $J = 29$. Estimates obtained in R (Pornprasertmanit, Miller, Schoemann, & Rosseel, 2013; Rosseel, 2012).

Finland (FI), France (FR), Germany (DE), Hungary (HU), Iceland (IS), Ireland (IE), Israel (IL), Italy (IT), Kosovo (XK), Lithuania (LT), the Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), the Russian Federation (RU), Slovakia (SK), Slovenia (SI), Spain (ES), Sweden (SE), Switzerland (CH), the Ukraine (UA), and the United Kingdom (UK).

The ESS is a good illustration because the ESS goes to great lengths in optimising its measures. The measurement of perceptions of political competition is no exception. The authors of the module on "Europeans' Understandings and Evaluations of Democracy" piloted a wide variety of competition items. Those that made it into the final survey thus had been evaluated thoroughly. Nevertheless, DIF remains a concern with the measures of perceived political competition.

In the survey, respondents evaluated three different statements about political competition in their own country:

1. "National elections in [country] are free and fair." [Elections]

2. "Different political parties in [country] offer clear alternatives to one another." [Alternatives]

3. "Opposition parties in [country] are free to criticise the government."

The response scale consisted of 11 categories, ranging from "does not apply at all" to "applies completely." [Opposition]

The baseline model for studying DIF is configural invariance—the assumption that the three items measure the same construct (Davidov, 2009). With only three items, this model is saturated and fits the data perfectly. Adding the requirements of metric and scalar invariance, in turn, one can assess the fit using the comparative fit index (CFI) and the root mean squared error of approximation (RMSEA). The results are shown in Table 1.

Chen (2007) proposes cutoff values of -0.010 and 0.015 for the changes in the CFI and RMSEA, respectively. By those standards, neither scalar nor metric invariance holds for the political competition items. One should thus conclude that DIF may be an issue for the items.

A thorough inspection of the results shows unusually low measurement intercepts for the elections item in Albania and the Ukraine (see Table

Table 2: Measurement Parameters by Country

	Intercepts			Loadings	
	E	A	O	A	O
AL	**3.570**	5.405	7.044	1.273	1.228
BE	7.444	5.933	7.369	0.697	0.836
BG	4.001	4.467	6.919	1.116	0.734
CH	8.192	6.610	7.695	0.700	0.850
CY	8.363	4.919	8.335	0.794	0.850
CZ	8.363	4.919	8.335	0.828	0.838
DE	8.626	5.775	8.577	0.454	0.818
DK	9.143	5.712	8.751	0.509	1.420
EE	6.378	5.391	6.769	0.766	0.916
ES	6.895	5.226	7.294	0.936	0.862
FI	9.071	6.564	7.997	0.826	**1.718**
FR	7.354	5.197	7.557	0.590	0.664
HU	6.414	5.806	7.034	0.924	1.031
IE	7.870	5.935	7.853	0.642	0.943
IL	7.888	5.820	7.837	0.931	1.196
IS	7.836	5.625	7.944	1.035	0.755
IT	5.881	3.703	6.513	0.670	0.663
LT	4.691	4.272	6.201	0.663	0.770
NL	8.220	6.489	7.932	0.664	1.116
NO	8.961	6.737	8.519	0.699	1.369
PL	7.401	6.214	7.588	0.752	1.115
PT	7.054	4.964	7.105	0.669	0.878
RU	4.237	4.844	**4.951**	1.057	1.072
SE	9.018	6.937	8.706	0.910	**1.700**
SI	7.049	4.790	7.110	0.683	0.765
SK	7.111	5.763	7.196	0.815	0.904
UA	**3.324**	3.960	5.726	0.833	0.940
XK	3.732	4.845	6.260	1.152	1.165
UK	8.188	5.982	7.815	0.849	1.258
Var	3.149	0.672	0.792	0.037	0.077

Notes: Table entries are maximum likelihood estimates. E = elections; A = alternatives; O = opposition. The loadings for elections are fixed at 1 for identification purposes. Outliers indicated in boldface.

Table 3: Random Effects ANOVA Estimates

	Elections	Alternatives	Opposition
γ_{00}	6.935	5.504	7.376
τ_{00}	3.156	0.668	0.791
σ^2	5.624	5.964	5.232
ICC	0.359	0.101	0.131

Notes: Table entries are REML estimates. $N = 49,807$, $J = 29$.

2).[3] It also shows an unusually low measurement intercept for the opposition item in the Russian Federation. Further, the measurement slope for this item is unusually high in Finland and Sweden.

Imagine we did not realise the presence of DIF and end up estimating a series of random effects ANOVA models on the different items. The random effects ANOVA model may be written as

$$\begin{aligned} y_{ij} &= \beta_{0j} + \theta_{ij} \\ \beta_{0j} &= \gamma_{00} + \delta_{0j} \\ \theta_{ij} &\sim \mathcal{N}(0, \sigma^2) \\ \delta_{0j} &\sim \mathcal{N}(0, \tau_{00}) \end{aligned}$$

Substituting the various items, we obtain the results in Table 3.

While the respondent-level variances are similar across the three competition items, the same cannot be said for the country-level variances which range from 0.668 for the alternatives item to 3.156 for the elections item. As a result the intra-class correlations (ICC) also fluctuate widely. For the alternatives item, around 10 percent of the variance can be attributed to cross-national differences; for the elections item, this percentage approaches 36. The elections item

[3] "Unusual" is defined as a value that is larger than the 3rd quartile plus 1.5 times the inter-quartile range or lower than the 1st quartile minus 1.5 times the inter-quartile range.

Table 4: Means-as-Outcome Estimates

	Elections	Alternatives	Opposition
γ_{00}	7.998	5.775	7.873
γ_{01}	-2.570	-0.656	-1.202
τ_{00}	1.552	0.581	0.444
σ^2	5.624	5.964	5.232
R^2	0.508	0.130	0.439

Notes: Table entries are REML estimates. $N = 49,807$, $J = 29$.

also shows the highest level of variance in measurement intercepts, however, so that we may conclude that the ICC for this item to a large extent captures DIF.

That the level-2 variance component displays a great deal of fluctuation depending on the measurement of perceived party competition has further consequences. Consider the following means-as-outcomes model (Bryk & Raudenbush, 1992):

$$\begin{aligned} y_{ij} &= \beta_{0j} + \theta_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} z_j + \delta_{0j} \\ \theta_{ij} &\sim \mathcal{N}(0, \sigma^2) \\ \delta_{0j} &\sim \mathcal{N}(0, \tau_{00}) \end{aligned}$$

Here Z_j is a country-level dummy variable that takes on the value of 1 for post-communist countries (Albania, Bulgaria, the Czech Republic, Estonia, Hungary, Kosovo, Lithuania, Poland, the Russian Federation, Slovenia, Slovakia, and Ukraine) and 0 elsewhere. We expect this dummy to have a negative effect and explain part of the random intercepts. In the means-as-outcomes model, the predictive power of the level-2 covariate can be captured through the proportional reduction in the level-2 variance component, which can be viewed as an R^2-measure (Bryk & Raudenbush, 1992).

Table 4 shows the parameter estimates and the R^2-values. One observes massive differences in the explained variances, which range from 13 percent to nearly 51 percent. One could interpret this as evidence that post-communist status is a much better predictor of perceptions about elections than the perceived availability of alternatives. One has to be extremely careful, however. The R^2-value correlates at .670 with the variation in the measurement intercepts. An alternative interpretation, then, is that the country dummy variable captures DIF.

What to Do?

Preventive Strategies Researchers are rarely in charge of collecting their own cross-national survey data but if they are, then two pieces of advice apply. First, always use multiple items to measure outcomes and core predictors. With single-item measures, it is impossible to assess whether DIF is a problem. This makes it impossible to disentangle the two pathways to random effects (see Figure 1) and thus generates a problem of interpretational confounding (cf. Burt, 1976). That is, if one obtains evidence of a random effect, it is impossible to say what caused it: DIF or heterogeneity in the DGP.

Second, in designing a survey with multiple indicators, always pre-test a large catalogue of items and retain those that are equivalent. This is no guarantee for success, as the ESS example shows, but it certainly goes a long way in ruling out DIF as a source of random effects.

Post-Hoc Solutions If the design of the survey is not in one's own hands but multiple indicators are available, then the best post-hoc solution is to look for partial measurement equivalence (Byrne, Shavelson, & Muthén, 1989). Par-

tial equivalence can be obtained in one of two ways. First, one can restrict the analysis to a subset of the items for which equivalence holds. Second, one can restrict the analysis to a subset of countries for which equivalence holds. Either way, one removes DIF as a potential source of random effects, thus allowing for an unequivocal interpretation of those effects. There is a price to pay, though. By removing items, obviously the coverage of a concept is reduced. And by removing countries, coverage of the target population is affected. It may be worth paying this price to obtain clarity about the meaning of random effects. At a minimum, it is worth considering partial equivalence as part of a set of robustness checks.

Should it be impossible to achieve partial equivalence or should only a single item be available, then the only solution is to be extremely careful in the interpretation of random effects and to acknowledge that they could both reflect DIF or true heterogeneity in the DGP. This may be frustrating, but it is the best one can do.

Appendix

The theory of this paper has taken a factor analytic approach to modelling measurement. In certain disciplines, however, it is customary to use item response theory (IRT). While the conceptualisation of measurement is somewhat different, the consequences of DIF for multilevel models are similar.

Consider the graded response model for ordinal scales (Samejima, 1969):

$$P(y_{ij} > m) = \frac{\exp\left(Da_j\left[\eta_{ij} - b_{jm}\right]\right)}{1 + \exp\left(Da_j\left[\eta_{ij} - b_{jm}\right]\right)} \quad (11)$$

Here, m is a particular response option, D is a constant, a_j is a country-specific item discrimination parameter, and b_{jm} is a country-specific item difficulty, which depends on the response category. For our purposes, we rewrite the model in terms of the cumulative log-odds

$$\ln\left[\frac{P(y_{ij} > m)}{P(y_{ij} \leq m)}\right] = Da_j(\eta_{ij} - b_{jm}) \quad (12)$$

Now imagine the latent trait, η_{ij}, abides Equation (2). Then the cumulative log-odds are

$$Da_j(\mathbf{x}_{ij}^\top\boldsymbol{\gamma} + \theta_{ij} - b_{jm}) \quad (13)$$

Expansion of this expression yields $D(a_j\mathbf{x}_{ij}^\top\boldsymbol{\gamma} + a_j\theta_{ij} - a_jb_{jm}) = D(\mathbf{x}_{ij}^\top\boldsymbol{\alpha}_j + a_j\theta_{ij} - a_jb_{jm})$. This is an RCM.

References

Birch, S. (2008). Electoral institutions and popular confidence in electoral processes: A cross-national analysis. *Electoral Studies*, *27*(2), 305–320.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods & Research*, *5*(1), 3–52.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504.

Davidov, E. (2009). Measurement equivalence of nationalism and constructive patriotism in the issp: 34 countries in a comparative perspective. *Political Analysis*, *17*(1), 64–82.

De Beuckelaer, A., & Swinnen, G. (2011). Biased latent variable mean comparisons due to measurement noninvariance: A simulation study. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications*. New York: Routledge.

Fairbrother, M., & Martin, I. W. (2013). Does inequality erode social trust? results from multilevel models of us states and counties. *Social Science Research*, *42*(2), 347–360.

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, *55*(292), 708–713.

Harteveld, E., Van Der Meer, T., & De Vries, C. E. (2013). In europe we trust? exploring the logics of trust in the european union. *European Union Politics*, *14*(4), 542–565.

Harzing, A.-W. K. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management*, *6*(2), 243–266.

Horn, J. L., & McArdle, J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Pornprasertmanit, S., Miller, P., Schoemann, A., & Rosseel, Y. (2013). *Package 'semtools'*. The Comprehensive R Archive Network.

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research*, *13*(4), 403–418.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(S1), 1–97.

Steenbergen, M. R., & Jones, B. S. (2002). Modeling multilevel data structures. *American Journal of Political Science*, *46*(1), 218–237.

Stoeckel, F. (2013). Ambivalent or indifferent? reconsidering the structure of eu public opinion. *European Union Politics*, *14*(1), 23–45.

Strabac, Z., & Listhaug, O. (2008). Anti-muslim prejudice in europe: A multilevel analysis of survey data from 30 countries. *Social Science Research*, *37*(1), 268–286.

Swamy, P., & Tavlas, G. S. (1995). Random coefficient models: Theory and applications. *Journal of Economic Surveys*, *9*(2), 165–196.

Van De Vijver, F. J., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*(1), 29–37.