SSOAR Open Access Repository
www.ssoar.info

Factorial Surveys with Multiple Ratings per Vignette: A Seemingly Unrelated Multilevel Regressions Framework

Schmidt-Catran, Alexander W.

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Schmidt-Catran, A. W. (2022). Factorial Surveys with Multiple Ratings per Vignette: A Seemingly Unrelated Multilevel Regressions Framework. *Methods, data, analyses : a journal for quantitative methods and survey methodology (mda)*, *16*(2), 335-360. https://doi.org/10.12758/mda.2022.04

Nutzungsbedingungen:
Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
https://creativecommons.org/licenses/by/4.0/deed.de

Terms of use:
This document is made available under a CC BY Licence (Attribution). For more Information see:
https://creativecommons.org/licenses/by/4.0

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

methods, data, analyses | Vol. 16(2), 2022, pp. 335-360 DOI: 10.12758/mda.2022.04

Factorial Surveys with Multiple Ratings per Vignette. A Seemingly Unrelated Multilevel Regressions Framework

Alexander W. Schmidt-Catran

Institute of Sociology, Goethe-University Frankfurt

Abstract

Factorial surveys are a prominent tool in the social sciences. Reanalyzing a literature survey on the factorial survey approach (Wallander, 2009), I show that about a quarter of applied factorial surveys asks respondents to provide multiple ratings on the same vignette. This paper is the first to propose a statistical modeling approach for precisely this situation. Data from factorial surveys with multiple ratings per vignette are afflicted with two sources of statistical dependencies. First, each respondent answers multiple vignettes, which is typically accounted for via random effects models, and, second, each vignette prompts multiple ratings. The first problem is common for almost any factorial survey and has been addressed decades ago. The second problem is addressed here. I propose to apply a seemingly unrelated regression approach to account for the statistical dependencies between multiple ratings per vignette. Due to the use of a structural equation modeling approach, the model allows not only to correctly compare coefficients across ratings but also to analyze the factor structure underlying these ratings. The proposed model is illustrated by two examples from recent research. All data and syntax are available online and allows for an easy adaption of the proposed model to readers' own research.

Keywords: factorial survey, vignette study, seemingly unrelated regressions, multiple ratings, multilevel, random effects, factor analysis, latent variables

© The Author(s) 2022. This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 License. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Factorial surveys, also called vignette studies, present artificial descriptions of people, objects or situations (vignettes), which are judged (rated) by survey respondents. Each vignette contains multiple theoretically relevant factors (dimensions) simultaneously. Thereby, factorial surveys allow investigating how the multi-dimensional characteristic of an object, person or situation, affects respondents' attitudes towards it (Jasso, 2019). The characteristics (levels) of the factors are varied systematically across the entire universe of vignettes. Factorial surveys thereby combine the virtues of experimental approaches to causal inference with classical survey research (Atzmüller & Steiner, 2010). They are particularly useful if the characteristics of interest are strongly confounded in reality, or at least in the perception of the respondents, and if the object of interest is suspect to social desirability (Atzmüller & Steiner, 2010; Wallander, 2009).

For example, Czymara and Schmidt-Catran (2016) ask "who is welcome in Germany?" and present descriptions of immigrants to their respondents. The immigrants are described in terms of their education, gender, country of origin, language skills, motivation, and religion. Each of these factors is constituted by multiple levels, for example, immigrants have no religion, are Christian or Muslims. The design allows investigating the relative impact of each dimension on the acceptance of immigrants and the estimation of the effect of specific levels. What is more important, economically relevant characteristics like education, or cultural features like religion? Are Christians preferred over Muslims?

Going back to the seminal work by Rossi and colleagues (Rossi & Nock, 1982), factorial surveys have now been around for 40 years and are frequently used in social science research (for an overview see Wallander, 2009). Many papers have been written about issues of designing and analyzing factorial surveys (for an overview see Jasso, 2006). Methodological issues concern for example the design of the vignettes (Auspurg, Hinz, Sauer, & Liebig, 2015), the assignment of vignettes to respondents (for example Atzmüller & Steiner, 2010; Dülmer, 2007, 2016) or the statistical method for the efficient estimation of the effects of vignette characteristics (for example Hainmueller, Hopkins, & Yamamoto, 2014; Jasso, 2006). In some way, most previous methodological papers focus on how to best deal with the multi-dimensionality of vignette characteristics. This paper takes a different route; it brings the multi-dimensionality of attitudes towards a social object into play.

Typically, factorial surveys require respondents to provide *one* rating per vignette, thereby restricting the measurement of the attitude towards the described

Direct correspondence to
Alexander Schmidt-Catran, Professur für Soziologie mit dem Schwerpunkt Methoden der quantitativen empirischen Sozialforschung, Goethe-Universität Frankfurt, Fachbereich Gesellschaftswissenschaften, Institut für Soziologie, Theodor-W-Adorno-Platz 6, PEG – Gebäude, 60323 Frankfurt
E-mail: alex@alexanderwschmidt.de

object to one dimension. However, factorial surveys with multiple rating questions per vignette, are not uncommon. A re-analysis of the studies discussed in Wallander's (2009) review indicates that about one quarter (27%) of applied factorial surveys measure *multiple ratings per vignette*.[1] More recent examples of such surveys are Harell et al. (2012), Weinberg et al. (2014), Czymara and Schmidt-Catran (2016) and Diehl et al. (2018); the last two of which are used as examples in this article. To the best of my knowledge, no special modeling approach for factorial surveys with *multiple ratings per vignette* has been introduced previously. It is important to define the term "multiple ratings per vignette" in order to avoid misunderstandings. A factorial survey typically provides *multiple vignettes* to a respondent, i.e. multiple descriptions of objects that vary in their characteristics. Thus, we get multiple ratings per respondent—as many as the respondent received vignettes. As discussed below, the hierarchical structure resulting from this survey design, is typically accounted for via multilevel models. Additionally, in some factorial surveys, respondents must provide multiple ratings on each vignette description. This results in *multiple ratings per vignette*. For example, Czymara and Schmidt-Catran (2016) provide 14 descriptions of immigrants to their respondents. On each of these 14 vignettes, respondents had to provide three ratings, resulting in a total of 42 (= 14 x 3) ratings per respondent.

The following paper proposes a statistical model for the analysis of such data. This model can be applied to any data from factorial surveys that (1.) include multiple ratings per vignette (at least 2) and (2.) multiple vignettes per respondent (at least 2). More precisely, the technique proposed here, models each of the ratings as a separate dependent variable and thereby allows for the analysis of their differences and commonalities regarding their determinants.[2] The basic idea is to use a seemingly unrelated regression framework combined with a structural equation approach to multilevel modeling. This allows for a simultaneous modelling of multiple dependent variables (i.e. the multiple ratings per vignette). Multilevel modeling has been recommend for the analysis of factorial surveys as it accounts for the statistical dependencies in the data, due to the fact that each respondent is confronted with *multiple vignettes* (Hox, Kreft, & Hermkens, 1991). Combining multilevel analysis with the seemingly unrelated regression approach allows correct accounting for the additional statistical dependencies due to the measurement

1 I want to give special thanks to Lisa Wallander for providing me with the data she collected for her literature survey. My re-analysis of the studies reviewed by Wallander (2009) can be found in the online appendix (Table OA1) of this paper: http://www.schmidt-catran.de/sumreg.html.

2 It may be that the multiple ratings per vignette constitute indicators of the same (unidimensional) latent construct. In this case, the multiple ratings may be combined into a single dependent variable before the analysis, rather than using the model proposed here, which is suitable only if the analysis has multiple dependent variables.

of *multiple ratings per* vignette. Finally, the use of structural equation modeling gives the opportunity to analyze the latent structure underlying the ratings.

A Seemingly Unrelated Multilevel Regressions Framework

In a seminal paper, Zellner (1962) proposed a method to estimate seemingly unrelated regression (SUR) models. He discusses how to account for the fact that estimation results from a set of regressions which use different dependent variables but share (some) predictors are statistically not independent. If all regressions have exactly the same set of predictors, this does not affect the estimated model parameters but the statistical tests necessary for comparing parameters across the regressions (Zellner, 1962: 351, 355). If the regressions differ not only in their dependent but also in their independent variables, accounting for the statistical dependence does also directly affect the estimators (Zellner, 1962: 351).

In the context of factorial surveys with multiple ratings per vignette, each dependent variable (i.e. rating) will always be dependent on the same vignette dimensions (i.e. predictors) by design. Hence, when analyzing the impact of the vignette dimensions only, estimating seemingly unrelated multilevel regressions (SUMREG) provides the same estimators as separate regressions. In that case, the SUR approach boils down to a multivariate regression model, which can be seen as a special case of the former. Nevertheless, accounting for the statistical dependence of the estimators, more precisely of the error terms, is important when comparing coefficients across dependent variables (Zellner, 1962: 355).

If respondent-level characteristics are added to the set of predictors, it may be that there are theoretical reasons to include some variables in one equation but not in another (see Example 1.3 below). In that case, the SUR approach will yield different (more efficient) estimates than a separate regression approach. Nevertheless, the emphasis in this article is on the more likely case of identical predictors in all regressions and therefore on comparing coefficients across them.

In his seminal paper, Zellner (1962) proposed a two-stage approach to the "efficient" estimation of SURs. However, the model can also be estimated in one step, using structural equation modeling. Formally the model can be understood as a system of i regression equations:[3]

3 Note that the index i here indicates regression equations—not units of analysis—because the regression equations are presented in matrix notation, which does not include an index for the units of analysis.

$$Y_1 = X\beta + e_1$$

$$\vdots$$

$$Y_i = X\beta + e_i,$$

in which the error terms are allowed to be correlated across equations. Thus, the variance-covariance matrix of the error terms is an unrestricted matrix in which the error variances are located at the diagonal and their covariances at the off-diagonals:

$$\Sigma_\mathrm{e} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1i} \\ \vdots & \ddots & \vdots \\ \sigma_{i1} & \cdots & \sigma_{ii} \end{bmatrix}.$$

This model can be extended to account for multiple error components (Baltagi, 1980), i.e. a multilevel structure in the data:

$$y_1 = X\beta + u_1 + e_1$$

$$\vdots$$

$$y_i = X\beta + u_i + e_i,$$

where each component of the composite errors $\varepsilon_i (= u_i + e_i)$[4] allows correlations across equations but the components are independent of each other:

$$\Sigma_\varepsilon = \begin{bmatrix} \sigma_{11}^u & \cdots & \sigma_{1i}^u & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{i1}^u & \cdots & \sigma_{ii}^u & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sigma_{11}^e & \cdots & \sigma_{1i}^e \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_{i1}^e & \cdots & \sigma_{ii}^e \end{bmatrix}$$

Such a model, traditionally employed for the analysis of panel data, seems to be perfectly suited to analyze factorial surveys with multiple ratings per vignette.[5] One argument for this has been laid out above. The model accounts for the statistical dependencies in the data and thereby allows performing correct statistical tests.

4 In this notation *e* is the idiosyncratic error and *u* is the unit-specific error (or random effect).

5 Obviously, the data structure of a classical panel is identical to the data structure produced by factorial surveys if each respondent has to rate multiple vignettes: Multiple observations of the dependent and independent variables from each respondent.

However, in addition to the issue of adequate statistical procedures, the SUMREG model has another advantage. It allows using the estimates of the random effects (u_i) for substantive interpretations. In the next section I will first lay out how to estimate the SUMREG model using generalized multilevel structural equation modeling (SEM). Then I will briefly discuss statistical tests, which are relatively straight forward, once the model has been presented, and finally I will introduce the idea of substantive interpretations of the random effects; this is, to conceptualize the unit-specific error components as latent variables.

Estimating the SUMREG Model Using Multilevel Structural Equation Modeling

Figure 1 presents the path diagram of a SUMREG model which includes j vignette dimensions as explanatory variables (X) and i ratings per vignette as dependent variables (Y). Each vignette dimension has a path to each of the dependent variables. The model furthermore includes random effects (u) for each dependent variable. These random effects (REs) are estimated at the level of the respondents. In other words, the path diagram shows a multilevel SEM in which the vignette dimensions and ratings are located at the first level (i.e. the vignette-level) and the REs are located at the second level (i.e. the respondent-level). The data structure for this model is in long format, i.e. the multiple vignettes asked per respondent each occupy a separate row, as it is typically done for standard multilevel modeling. What makes this model a SUR model are the correlations between the errors. More precisely, all idiosyncratic errors e are allowed to correlate with each other and all unit-specific errors u are allowed to correlate with each other.[6]

The introduction of respondent-level characteristics into that model can be done via the within-and-between formulation of multilevel models. Such a model is presented in Figure 2. It assumes that l respondent-level characteristics (Z) explain the between-unit variance in the dependent variables. As this variance is captured in the unit-specific REs, the respondent-level variables impact directly on these.[7] This makes the formerly exogenous REs u endogenous variables, which are in Figure 2 indicated as η_i. The unexplained variance then is captured in the error term of these endogenous variables (u). In contrast to the variables measured at the vignette-level, the respondent-level variables may not all affect all dependent vari-

6 Given i dependent variables (i.e. ratings per vignette), the system includes $i*(i-1)/2$ covariances between the unit-specific error terms u as well as between the idiosyncratic error terms e.

7 There is also a different but equivalent formulation of that model, in which the respondent-level characteristics impact the dependent variables directly, i.e. a single-level formulation.

ables. As discussed above, the SUR approach allows that a subset of the explanatory variables affect only part of the dependent variables. In that case, the model would no longer be equivalent to a multivariate multilevel model.

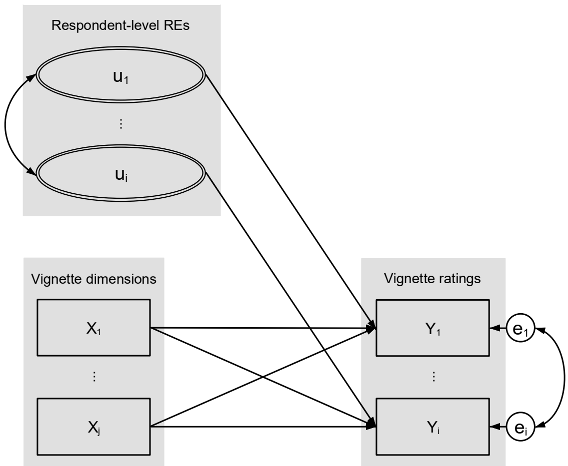


Figure 1 SUMREG Model with Vignette Dimensions

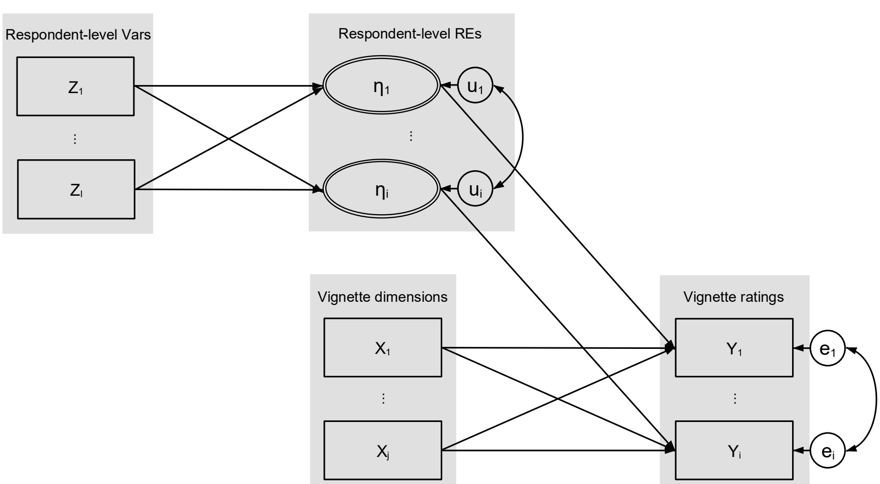


Figure 2 SUMREG Model with Vignette Dimensions and Respondent-level FEs

Comparing Parameters Within and Across Ratings

Factorial surveys with multiple ratings per vignette offer a variety of potential hypotheses tests. For example, we can ask whether a particular vignette characteristic has the same effects across ratings. We can also ask whether a set of vignette characteristics affects one dependent variable but not another, or whether the effect of a vignette characteristic on one rating is larger than on another rating, and so forth. Such hypotheses cannot be tested if the multiple ratings are modelled separately.

In general, there are two distinct ways of testing such hypotheses: We can either use a Wald-Test of linear hypothesis or we can compare a restricted and an unrestricted model using Likelihood-Ratio-Tests. Ultimately, these two tests are asymptotically equivalent (Engle, 1984) but, depending on the specific hypothesis to be tested, one or the other may present itself as more obvious. For example, if we want to test whether all explanatory variables have the same effect on each of the dependent variable, it seems more obvious to estimate an unrestricted model and compare it to a model with the appropriate restrictions via Likelihood-Ratio-Tests. If, on the other hand, we are interested in comparing two specific parameters, or testing one parameter against zero, the Wald-Test seems more appropriate.[8]

Conceptualizing the Random Effects as Latent Factors

From the viewpoint of standard multilevel modeling, the random effects u are merely error terms that capture the unexplained variance between the second-level clusters, i.e. respondents. In the language of panel data analysis, they would be described as unobserved heterogeneity (see Andress, Golsch, & Schmidt, 2013: 96f., for a discussion of the equivalence of multilevel and random effects panel data models). However, such random effects can be understood as latent variables (Skrondal & Rabe-Hesketh, 2004), which is the reason why SEM can be used to estimate multilevel models.

What exactly do these latent variables capture? As they are measured at the respondent-level, they reflect differences *between* respondents, independent of their reaction to specific vignettes. In other words, the REs reflect the tendency of respondents to select a specific response category independent of the varying stimuli. What stories can such REs tell?

8 Note that Likelihood-Ratio-Tests require re-estimating the model with the appropriate restrictions while Wald-Tests do not. Given the complexity of generalized multilevel SEMs, this process can take quite some time. If time is a scarce resource for you, you may prefer using Wald-Tests.

First, and foremost, when we compare the variance *between* units with the variance *within* units in an empty model, i.e. calculate the intra-class-correlation (ICC) coefficient, we can judge how much the respondents react to the experimental stimuli. If, for example, 90% of the total variance is between respondents, we could conclude that the vignette characteristics are generally not very effective. Such an analysis of the error variance can of course be done with simple multilevel models as well. However, the SUMREG model allows comparing the ICCs across the different ratings and thereby allows making statements about these differences. For example, it might be that one dependent variable reacts stronger to vignette characteristics than another.

Second, and this is specific to the SUMREG model, we can analyze the relationships between the REs of each dependent variable (rating). Such an analysis of the latent factor structure is directly included in the model, i.e. the variance-covariance matrix of the REs. The model gives us a clue as to how much the general tendencies of respondents to rate all vignettes similarly, are related across the different ratings. For example, we might see that some ratings are quite strongly related while others seem to be separate issues (compare Example 1 in Section 6.1).

If we allow ourselves to adapt more of the typical thinking of structural equation modelers, we see that we can even assume that two or more ratings are actually expressions of the same underlying latent variable. Thus, we could test a model in which all ratings are understood as being indicators of the same underlying issue, i.e. in which there is only a single RE instead of one per dependent variable.

Such a model is shown in Figure 3 and can be compared to a model with a separate RE for each dependent variable via Likelihood-Ratio-Tests (LR-Tests).

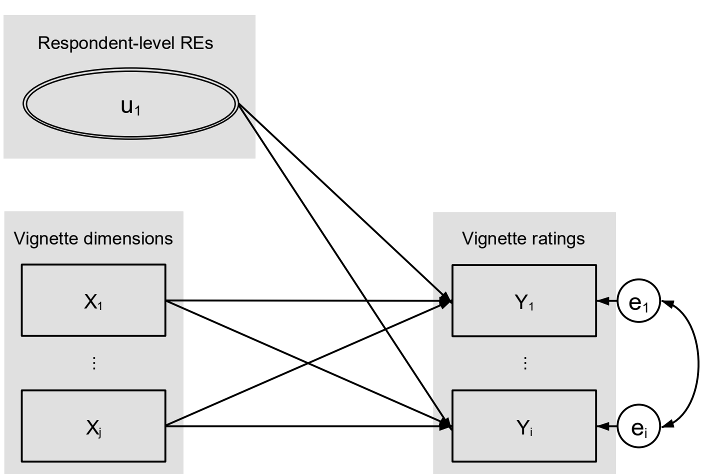


Figure 3 SUMREG Model with single RE for all Ratings

Depending on the number of ratings per vignette there are a variety of possible model specifications. If the factorial survey design consists of two ratings per vignette, there are just two possible models: A separate RE for each rating or one RE for both. If, however, the design includes more than two ratings per vignette, one might assume that some ratings share an underlying latent variable while others are separate issues, i.e. have their own REs. Similarly, we can obviously compare a model which assumes completely unrelated REs against a model which allows correlations between them. This would be an empirical test of the hypothesis that the issues are completely unrelated to each other.

Another nice feature of the SUMREG model is that we can predict the REs and analyze their joint distribution in detail. Such an analysis can be interesting in its own right but might make particular sense if the multiple ratings per vignette constitute something like a Guttman scale. Using predicted values of the REs in that case allows checking the consistency in response behavior. An example of such an analysis is shown below (Example 1.1).

Finally, a short note on the implied measurement model of the single REs is necessary. The SUMREG model in its unconstrained form, as in Figures 1 and 2, provides a RE for each rating. This effect is identified because each respondent rates multiple vignettes. As explained above, from the viewpoint of SEMs these REs can be understood as measurement models of latent factors. Then, of course, the question arises how the measurement coefficients (factor loadings) of that model look like. These parameters are not explicitly part of the model. As indicated above, the data for this kind of model is organized in long format with respondents each occupying as many rows in the data set as they have rated vignettes. The "measurement coefficients" of the REs therefore is the implicit coefficient of the respondent-level error term u_i, which is 1. Thus, each vignette is given the same weight in the "construction" of the respondent-level latent factors. This assumption might appear problematic but actually it is well justified. For each vignette respondents answer the same questions. Thus, the *wording* of a specific rating item is actually the same for each of its measurements. What varies between the measurements are just the descriptions on the vignettes.

Examples

All following analyses are performed using stata's `gsem` command for generalized multilevel SEM (version 14.2). The data sets and do-files are provided in the online appendix of this paper (see footnote 1). I use two examples to demonstrate how the aforementioned modeling strategies and tests can be applied to real data. One example data set is from a factorial survey conducted in Germany in April 2015 (Czymara & Schmidt-Catran, 2016) and the other one from a factorial survey con-

Table 1 Description of Example Data Sets

	Czymara and Schmidt-Catran	Diehl et al.
Sample		
Full Sample (Respondent N)	1,283	1,432
Used Sample (Respondent N)	77	284
Fielded Vignettes per Respondent	14	4
Answered Vignettes per Respondent	14	3.98
Valid Vignette Ratings	1,078	1,131
Vignette Characteristics		
Vignette Dimensions	6	6
Total Vignette Levels	15	19
Vignette Universe	192	567
Ratings per Vignette	3	2
Points of Rating Scales	7	7

ducted in Switzerland between March and May 2014 (Diehl et al., 2018).[9] In both cases I analyze a random sub-sample of the complete data, each of which includes about 1,000 unique vignette ratings.[10]

Both surveys deal with the impact of cultural and economic threats on the acceptance of immigrants. While the data by Czymara and Schmidt-Catran (2016) is based on a D-efficient design, in which all respondents received the same set of 14 vignettes (Dülmer, 2007: 385ff.), the data from Diehl et al. (2018) is based on a D-efficient sampling design, in which each respondent received a different subset of 4 vignettes (Dülmer, 2007: 384ff.). What both surveys have in common is that they generated data in which multiple vignettes are nested within respondents and respondents provided multiple ratings per vignette. Czymara and Schmidt-Catran (2016) use three ratings per vignette, while Diehl et al. (2018) use two ratings per vignette. Table 1 provides some information about both studies and the samples used for the following examples.

9 I like to give special thanks to Claudia Diehl, Katrin Auspurg and Thomas Hinz for providing their data.

10 I do so for two reasons: First, estimating these models is quite time consuming. By reducing the number of observations, I reduce the time needed for estimating and/or replicating my results. Second, I did not want to provide the full data from other authors.

This paper is certainly not the place for an extensive theoretical discussion but I will briefly summarize the central idea behind the two surveys: One the one hand, negative attitudes towards immigrants are assumed to be determined by natives' fear of the economic consequences of immigration (Facchini, Mayda, & Puglisi, 2013). On the other hand, scholars argue that natives fear the loss of their culture and therefore turn against immigrants (Hopkins, 2015). Factorial surveys are particularly well suited for this research area because they allow the simultaneous analysis of several determinants (i.e. cultural and economic threats) and minimize the risk of social desirable answers (Wallander, 2009).

For the purpose of this paper it is sufficient to keep in mind that the vignettes cover economic and cultural characteristics of immigrants and that attitudes are expected to be particularly negative towards culturally more distinct immigrants. With regard to economic threats, the literature assumes that immigrants with higher skill levels are generally preferred because they should contribute more to the economic system in general (Hainmueller & Hiscox, 2007) but it has also been stated that natives fear competition on the job market (Facchini & Mayda, 2012).

Example 1: Czymara and Schmidt-Catran Data

In this factorial survey respondents were asked to rate vignettes with regard to three issues: Should the immigrant described on the vignette have the right to (1) live in Germany, (2) work in Germany, and (3) receive social benefits in Germany? Answers were measured on a 7-point scale, where higher values indicate willingness to grant the related right. The first step of the empirical analysis regards the factor structure of these three items.

Example 1.1: Analysis of Latent Factor Structure

Table 2 shows three empty models with a varying number of REs. M1 includes a separate RE for each rating (U1, U2, U3). All models allow correlations between the REs and also between the idiosyncratic errors. In Model M1 all of these correlations are highly significant, indicating that the SUMREG model is indeed justified. All models include an intercept for each rating, showing that acceptance of immigrants working and living in Germany is much higher (5.17 and 5.47 respectively) than acceptance of immigrants taking social benefits (3.88).[11] The variances of the REs reveal that between respondents, ratings vary much more with regard to the issue of social benefits than with regard to the other two issues. Thus, natives seem to have a stronger consensus over the issues of living and working in Germany

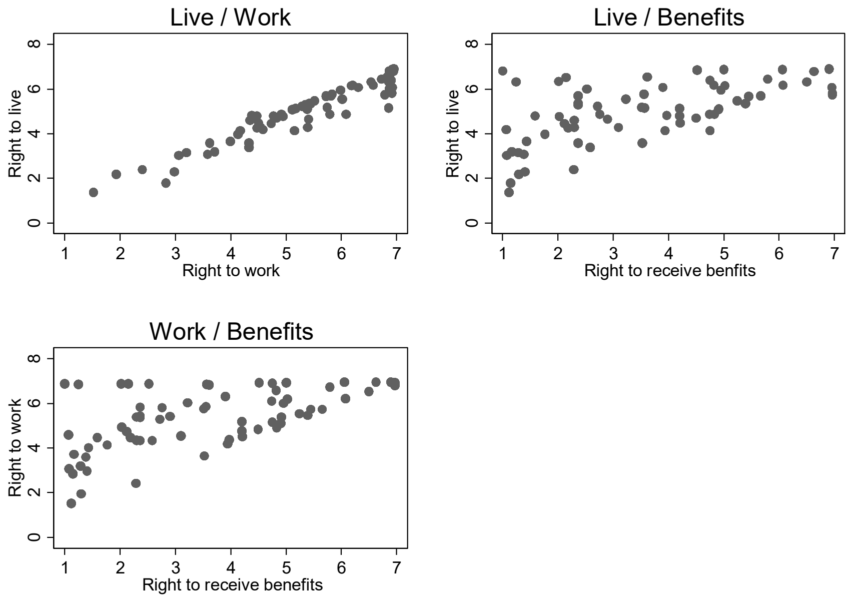
11 The „factor loadings" of the REs (U1, U2, U3) are all 1 because each dependent variable has its own RE, which then by definition has to be 1 as it provides the anchor to scale the latent variable.

than over the issue of social benefits for immigrants. All of this, however, could also be seen from separate multilevel regressions. The covariances between the REs, in contrast, are unique to the SUMREG model. Note that Table 2 expresses covariances between the REs as correlations to allow for ease of comparison across pairs. While the issues of living and working in Germany are very closely related (Corr(U1,U2)=.95), the issue of social benefits seem to be less strongly associated with the other two (Corr(U1,U3=.68), Corr(U2,U3)=.66).

Table 2 Empty SUMREG models - Example 1.1

	M1	M2	M3
Live			
U1	1.000 constr.	1.000 constr.	1.000 constr.
Intercept	5.171 ***	5.171 ***	5.171 ***
Work			
U2	1.000 constr.		
U1		0.958 ***	1.014 ***
Intercept	5.469 ***	5.469 ***	5.469 ***
Benefits			
U3	1.000 constr.		1.000 constr.
U1		2.891 ***	
Intercept	3.878 ***	3.878 ***	3.878 ***
Variances and Covariances			
Var(U1)	2.043 ***	0.425 ***	1.938 ***
Var(U2)	2.006 ***		
Var(U3)	3.837 ***		3.785 ***
Corr(U2,U1)	0.953 ***		
Corr(U3,U1)	0.682 ***		0.674 ***
Corr(U3,U2)	0.657 ***		
Var(e.live)	1.809 ***	3.427 ***	1.914 ***
Var(e.work)	1.576 ***	3.191 ***	1.588 ***
Var(e.benefits)	1.489 ***	1.775 ***	1.509 ***
Cov(e.work,e.live)	1.415 ***	2.938 ***	1.379 ***
Cov(e.benefits,e.live)	1.107 ***	1.787 ***	1.153 ***
Cov(e.benefits,e.work)	0.943 ***	1.588 ***	0.927 ***
Statistics			
Log-Likelihood	-4,699.36	-5,088.74	-4,790.21
LR-Tests		M2 vs. M1	M3 vs. M1
LR chi^2		778.75	181.68
Prob > chi^2		0.000	0.000

Notes: * p<.05, ** p<.01; *** p<.001 (two-sided tests).



Notes: Values are predicted as the sum of the intercept and the individual REs [BLUPs].

Figure 4 Predicted values from Model M1 – Example 1.1

Figure 4 presents the association of the three REs in more detail by means of scatter plots, using predicted values from Model M1 (Intercept + REs [BLUPs]). The first panel in Figure 4 shows that for almost all respondents the right to live and the right to work in Germany go together, i.e. they are on the diagonal of the plot. There is also a cluster of respondents which fully grant the right to work in Germany (7 on the x-axis) but do not grant the right to live in Germany to the same extend, i.e. they are below the diagonal. Following these insights one could categorize such response behavior as inconsistent and decide how to treat these cases.[12]

The second and third panels in Figure 4 look quite similar, with all respondents being on or above the diagonal, indicating that a large share of respondents tends to grant the right to live (panel 2) or work (panel 3) in Germany to a larger extent than the right to receive social benefits. Panel 2 again reveals two respondents that provide inconsistent answers, granting the right to receive benefits but to a lesser degree the right to live in Germany.

12 This paper is not the place for a detailed discussion of such issues but there are a number of alternatives: One could simply recognize that some respondents give inconsistent answers and go on with the analysis or one could exclude these respondents from the analysis. One could also think of using the SUMREG model during the pre-test phase of a factorial survey and take such a result as an indicator that the vignettes and the instructions may need to be redesigned.

Models M2 and M3 in Table 2 test whether the three separate REs from model M1 can be replaced by shared factors. Model M2 includes one RE for all three dependent variables (U1). The model does fit the data significantly worse than model M1 (LR-Test: $p<.0001$) and therefore we can conclude that the three ratings are not expressions of the same underlying latent factor. This result is not surprising given the graphical evidence from Figure 4. The variance and covariance parameters in model M2 are not of great interest but readers should note that the factor loadings, which have all been 1 in model M1 are now allowed to vary across ratings, making them true factor loadings in this model. The first factor loading (live) is still 1 as it provides the anchor to scale the latent variable.

Model M3 assumes that the issues of living and working in Germany share one underlying latent factor (U1) while the issue of receiving benefits has its own RE (U3). Given the evidence from Figure 4 this seems like a reasonable assumption, but the model does not hold against model M1 (LR-Test: $p<.0001$). Thus, we can conclude that this data is best modeled with a separate RE for each rating: Living, working and receiving benefits in a host society seem to be separate issues, where respondents can show various combinations of positive and negative attitudes. Such a conclusion could not be tested without the SUMREG model.

Example 1.2: Analyzing Fixed Effects

Table 3 presents two models in which the vignette-level effects have been added to the fixed part of the equations. Both models include a separate RE for each of the ratings, following the evidence from Models M1, M2 and M3. Model M4 estimates separate fixed effects for each of the three dependent variables, while Model M5 constrains them to be equal across all three ratings. The LR-Test comparing both models indicated that Model M5 does fit the data significantly worse than Model M4. We can therefore conclude that the vignette characteristics' effects are not identical across the three dependent variables, at least if one tests all of them simultaneously. Again, such a conclusion requires the SUMREG model for a correct statistical test.

Substantially, the results show that there is no effect of an immigrants' gender or country of origin on her or his acceptance. Immigrants with higher education and good language skills are preferred over those with lower education and bad language skills. Muslim immigrants are less accepted than immigrants who are Christians or do not have a religious denomination, but this is only significant for the right to live in Germany not for the other two issues. The strongest effect, however, is a person's motivational reason for immigration. Immigrants that have a job in prospect are much more welcome than those who come for economic reasons but without any economic prospects. Immigrants who flee from political persecution are by far the most accepted group.

Table 3 Adding vignette-level covariates – Example 1.2

	M4			M5
	Live	Work	Benefits	Live/Work/ Benefits
Gender (Ref. = Female)				
Male	0.029	0.078	-0.011	0.027
Country of Origin (Ref. = Lebanon)				
France	0.135	0.144	0.152	0.147
Kenya	-0.074	-0.086	-0.043	-0.063
Reason for Migr. (Ref. =better live)				
Political Persecution	1.420 ***	1.045***	1.333 ***	1.239 ***
Job	0.939 ***	0.813***	0.632 ***	0.736 ***
Education (Ref. = low education)				
University	0.337 ***	0.307***	0.193 **	0.253 ***
Language skills (Ref. = none)				
Good	0.471 ***	0.420***	0.271 ***	0.350 ***
Religion (Ref. = no Religion)				
Christ	0.055	0.042	0.055	0.050
Muslim	-0.230 **	-0.138	-0.060	-0.110
U	1.000 constr.	1.000 constr.	1.000 constr.	1.000
Intercept Live	4.020 ***			4.204 ***
Intercept Work		4.463 ***		4.503 ***
Intercept Benefits			2.985 ***	2.911 ***
Variances and Covariances				
Var(U1)		2.073 ***		2.071 ***
Var(U2)		2.025 ***		2.023 ***
Var(U3)		3.857 ***		3.856 ***
Corr(U2,U1)		0.953 ***		0.954 ***
Corr(U3,U1)		0.683 ***		0.684 ***
Corr(U3,U2)		0.658 ***		0.659 ***
Var(e.live)		1.390 ***		1.410 ***
Var(e.work)		1.308 ***		1.326 ***
Var(e.benefits)		1.209 ***		1.221 ***
Cov(e.work,e.live)		1.085 ***		1.091 ***
Cov(e.benefits,e.live)		0.782 ***		0.774 ***
Cov(e.benefits,e.work)		0.698 ***		0.684***

	M4			M5
	Live	Work	Benefits	Live/Work/ Benefits
Statistics				
Log-Likelihood		-4528.1653		-4577.7834
LR-Tests				M4 vs. M3
LR chi2				99.24
Prob > chi2				0.000

Notes: * p<.05, ** p<.01, *** p<.001 (two-sided tests).

As discussed above, the SUMREG model allows performing statistically correct tests across the multiple ratings. One such test is the LR-Test comparing the two models presented in Table 3. In order to compare single coefficients or test a few selected parameters, the Wald-Test seems to present itself since it does not require re-estimating the model. For example, we could hypothesize that an immigrant's education is more important for work-related issues than for the general right to live in Germany. Or, vice versa, we could hypothesize that an immigrant's qualification is equally important for all three issues. The corresponding test on the coefficients estimated in Model 4 indicates that the effect is indeed independent of the specific issue (Chi^2=5.20, p=.074).

We might wonder whether being a Muslim matters more for an immigrant's general acceptance (right to live, coef. = -0.230) than for granting her or him the right to receive social benefits (coef. = -.060). Using a Wald-Test we can check whether the effect of being a Muslim on the right to live is significantly stronger than the effect on the right to receive social benefits. The test reveals that it actually is: Chi^2=4.96, p=.026.[13] As these examples show, hypotheses about differences and commonalities across the ratings can be theoretically meaningful. In order to test such hypotheses, the SUMREG model is required, as a separate modeling of the ratings does not allow to perform such tests.

Another interesting perspective opened by the SUMREG model is related to the covariation of the idiosyncratic error terms. In an empty model (compare Model M1 in Table 2), the covariance between these error terms reflect not only correla-

13 This is an example where a naive statistical test based on separate multilevel regression models would give a different result: When testing the effect of Muslim on the right to work against the numerical value of the coefficient in the model of social benefits (-0.60), the test indicates a non-significant difference (Chi^2=3.69.87, p=.055). The univariate multilevel models used for this naive and incorrect (!) test can be found in the online appendix (see footnote 1): Univariate M4, Table OA2.

tions between the idiosyncrasies of the ratings but also their joint variation due to the treatments on the vignettes, i.e. the vignette-level effects. Once the vignette characteristics are controlled, this "explained" part of the covariance is removed from the random part of the model and the remaining covariances of the residuals indicate "unexplained" covariance between the idiosyncratic error terms. If this unexplained covariance remains substantial, we might take this as an indicator of problematic response behavior. For example, respondents may have thought only about the first rating and then simply selected the same scale points for the remaining ratings. In the example above, the covariation of the three idiosyncratic error terms (live, work, social benefits) has been reduced by 23%, 29% and 26%, respectively, when comparing the empty Model M1 and Model M4. Thus, a substantial amount of covariance is left after accounting for the vignette-level effects.

Example 1.3: A True Seemingly Unrelated Regression Model

While the SUMREG models include some parameters that are obviously missing in a univariate approach (i.e. the covariances of REs and idiosyncratic errors across ratings), the models presented so far provide estimates that are identical to those from simple univariate multilevel models (compare Table OA2 in the online appendix, see footnote 13). In this sense, the multivariate approach of the SUMREG model simply adds the potential to statistically compare coefficients across equations. However, as indicated above, the estimates from seemingly unrelated regressions differ from univariate estimates if the set of predictors varies between equations. In that case the SUR approach is no longer equivalent to a multivariate regression model. Such a *true* seemingly unrelated regression model benefits from a gain in efficiency resulting from the "zero restrictions" implied by the model specification (compare Zellner 1962: 353 f.).

Table 4 presents two SUMREG models which include, in addition to the vignette-level effects, the respondent-level variable *education*. In Model M6 education is included in each of the three equations while in Model M7 it influences only the right to work. The decision to assume an effect of education only on the work-related rating, as in Model M7, may be theoretically motivated; reflecting the idea that economic characteristics should matter most for employment-related issues, where competition on the labor market could be important, and less for the general acceptance or immigrants' deservingness of social assistance. A comparison of Models M6 and M7 illustrates the gain in efficiency due to the seemingly unrelated regression approach: While the effect of education is not significant on any of the three dependent variables in Model M6, it is significant in Model M7. The standard errors of the education effect are more than three times smaller in the latter model. Of course, Model M7 should be tested against M6 before one selects it as the better model. An LR-Test comparing the two models indicates that the model fit of them is not significantly different (p=.94). Thus, from a model-fit-perspective one could

Table 4 Adding respondent-level covariates - Example 1.3

	M6			M7		
	Live	Work	Benefits	Live	Work	Benefits
Vignette-level Effects						
Gender (Ref. = Female)						
Male	0.046 (0.075)	0.081 (0.074)	-0.015 (0.070)	0.046 (0.075)	0.081 (0.074)	-0.015 (0.070)
Country of origin (Ref. = Lebanon)						
France	0.144 (0.108)	0.149 (0.106)	0.172 (0.100)	0.144 (0.108)	0.149 (0.106)	0.172 (0.100)
Kenya	-0.065 (0.092)	-0.090 (0.090)	-0.046 (0.085)	-0.065 (0.092)	-0.090 (0.090)	-0.046 (0.085)
Reason for coming (Ref. = better live)						
Political Persecution	1.466 (0.108) ***	1.087 (0.106) ***	1.354 (0.100) ***	1.466 (0.108) ***	1.087 (0.106) ***	1.354 (0.100) ***
Job	0.979 (0.092) ***	0.846 (0.090) ***	0.670 (0.085) ***	0.979 (0.092) ***	0.846 (0.090) ***	0.670 (0.085) ***
Education (Ref. = low education)						
University	0.338 (0.075) ***	0.320 (0.074) ***	0.206 (0.070) **	0.338 (0.075) ***	0.320 (0.074) ***	0.206 (0.070) **
Language skills (Ref. = none)						
Good	0.478 (0.075) ***	0.437 (0.074) ***	0.278 (0.070) ***	0.478 (0.075) ***	0.437 (0.074) ***	0.278 (0.070) ***
Religion (Ref. = no Religion)						
Christ	0.068 (0.098)	0.044 (0.095)	0.051 (0.091)	0.068 (0.098)	0.044 (0.095)	0.051 (0.091)
Muslim	-0.202 (0.091) *	-0.144 (0.089)	-0.057 (0.084)	-0.202 (0.091) *	-0.144 (0.089)	-0.057 (0.084)
U	1.000 constr.	1.000 constr.	1.000 constr.	1.000 constr.	1.000 constr.	1.000 constr.
Intercept	3.878 (0.226) ***	4.273 (0.222) ***	2.816 (0.277) ***	3.898 (0.211)	4.291 (0.208)	2.856 (0.255)

	M6			M7		
	Live	Work	Benefits	Live	Work	Benefits
Respondent-level Effects						
Education (Ref. = low education)						
University	0.105 (0.436)	0.460 (0.427)	0.210 (0.580)	0.000 constr.	0.363 (0.131) **	0.000 constr.
Variances and Covariances						
Var(U1)		2.060 (0.355) ***			2.062 (0.356) ***	
Var(U2)		1.976 (0.341) ***			1.978 (0.341) ***	
Var(U3)		3.731 (0.628) ***			3.738 (0.629) ***	
Corr(U2,U1)		0.959 (0.168) ***			0.959 (0.168) ***	
Corr(U3,U1)		0.680 (0.145) ***			0.680 (0.145) ***	
Corr(U3,U2)		0.645 (0.143) ***			0.645 (0.143) ***	
Var(e.live)		1.409 (0.064) ***			1.409 (0.064) ***	
Var(e.work)		1.349 (0.062) ***			1.349 (0.062) ***	
Var(e.benefits)		1.219 (0.056) ***			1.219 (0.056) ***	
Cov(e.work,e.live)		1.120 (0.057) ***			1.120 (0.057) ***	
Cov(e.benefits,e.live)		0.811 (0.050) ***			0.811 (0.050) ***	
Cov(e.benefits,e.work)		0.720 (0.047) ***			0.720 (0.047) ***	

Notes: * p<.05, ** p<.01, *** p<.001 (two-sided tests). The models presented in this table differ in the number of observations (n=1,036) from the previous models (n=1,078) because some respondents (N=3) have missing values on education. This explains the difference in vignette-level effects.

select the more parsimonious model (M7) and thereby harvest the efficiency gain from the SUMREG model.

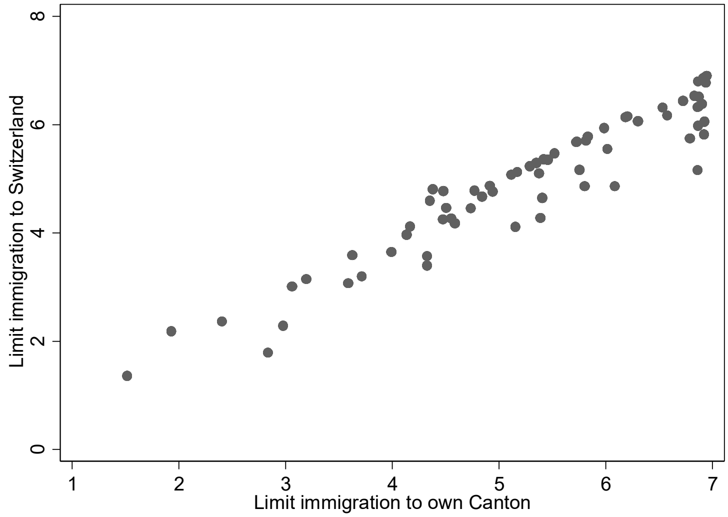
Example 2: Diehl et al. Data

In the study of Diehl et al. (2018) respondents were asked to provide two ratings per vignette: Should immigration from the described group be limited (1) to Switzerland in general and (2) to the respondent's own canton? Each rating was done on a 7-point scale where higher values indicate a desire to limit migration.

Example 2.1: Analysis of Latent Factor Structure

Intuitively these two ratings appear to have more in common than the three issues addressed in the former example, so we may expect to find them to be expressions of one latent factor, and this is exactly what an analysis of the underlying factor structure reveals, thereby providing a good counter example to the analysis above.

Table 5 presents two empty models. Model M1 includes a separate RE for each rating and Model M2 assumes that both ratings share one underlying latent factor. The correlation between the two REs in Model M1 is almost perfect (.99) and the LR-Test comparing the two models indicates that one can indeed model the two ratings as being expressions of the same underlying latent factor. Figure 5 presents the relationship between the two random components of Model M1 graphically.



Notes: Values are predicted as the sum of the intercept and the individual REs [BLUPs].

Figure 5 Predicted values from Model M1 – Example 2.1

Table 5 Empty SUMREG models - Example 2.1

	M1	M2
Switzerland		
U1	1.000 constr.	1.000 constr.
Intercept	3.671 ***	3.671 ***
Own Canton		
U2	1.000 constr.	
U1		1.013 ***
Intercept	3.751 ***	3.751 ***
Variances and Covariances		
Var(U1)	1.790 ***	1.788 ***
Var(U2)	1.841 ***	
Corr(U2,U1)	0.995 ***	
Var(e.switzerland)	2.194 ***	2.197 ***
Var(e.owncanton)	2.344 ***	2.351 ***
Cov(e.switzerland,e.owncanton)	2.075 ***	2.071 ***
Statistics		
Log-Likelihood	-3,337.73	-3,339.21
LR-Tests		M2 vs. M1
LR chi^2		2.95
Prob > chi^2		0.086

Notes: * p<.05, ** p<.01, *** p<.001 (two-sided tests).

Example 2.2: Analyzing Fixed Effects

Table 6 presents the results of two SUMREG models that include the vignette-level effects. According to the results from above (Example 2.1), both models assume that the two ratings are expressions of the same underlying RE. What differs between the two models is that Model M3 estimates separate fixed effects on the two ratings, while Model M4 constraints the effects to be equal. The LR-Test comparing the two models reveals that Model M3 does not have a significantly better fit and we can therefore conclude that the vignette dimensions affect both ratings equally.

Substantively the results show that immigrants from countries that are culturally more distant from Switzerland (Romania and Croatia) are less accepted. Immigrants with higher education are preferred over immigrants with basic education. Intended duration of stay does not have a significant effect. Immigrants that intend to find jobs for which no Swiss people are available are more accepted than immigrants who look for jobs that also Swiss people are looking for. Respondents have

Table 6 Adding vignette-level covariates to the model – Example 2.2

	M3		M4
	Switzerland	Own Canton	Switzerl./ Own Cant.
Country of origin (Ref. = Germany)			
France	0.305 *	0.376 *	0.323 *
Italy	-0.074	-0.069	-0.072
Norway	0.111	0.106	0.110
Romania	1.053 ***	0.954 ***	1.027 ***
Croatia	0.726 ***	0.716 ***	0.723 ***
Education (Ref. = University)			
Basic Education	0.322 ***	0.312 ***	0.319 ***
Intended duration of stay (Ref. = for ever)			
Several Years	0.022	-0.070	-0.002
One Year	-0.004	-0.047	-0.015
Swiss people available for job (Ref. = no)			
Yes	0.735 ***	0.668 ***	0.718 ***
Language skills (Ref. = German and French)			
No German and no French	1.021 ***	1.088 ***	1.039 ***
French but no German	0.514 ***	0.559 ***	0.526 ***
German but no French	0.249 *	0.310 *	0.266 *
Culture (Ref. = willing to adapt)			
Not willing to adapt	0.751 ***	0.720 ***	0.742 ***
No information	0.418 *	0.426 *	0.415 *
U1	1.000 constr.	1.015 ***	1.000 constr.
Intercept	1.942 ***	2.080 ***	1.960 ***
Variances and Covariances			
Var(U1)	1.819 ***		1.818 ***
Var(e.switzerland)	1.653 ***		1.654 ***
Var(e.owncanton)	1.831 ***		1.833 ***
Cov(e.switzerland,e.owncanton)	1.542 ***		1.541 ***
Statistics			
Log-Likelihood	-3,203.86		-3,211.75
LR-Tests			M4 vs. M3
LR chi^2			15.78
Prob > chi^2			0.327

Notes: * p<.05, ** p<.01, *** p<.001 (two-sided tests).

a preference for immigrants that speak at least one of the official languages or even better speak German and French. Immigrants that are willing to adapt to the Swiss culture are most accepted, while those who do not want to adopt are least accepted.

In sum, the survey of Diehl et al. (2018) provides an example of a vignette study in which (1.) the multiple ratings can be understood as expressions of the same underlying latent concept and (2.) the effects of the vignette characteristics are the same across the multiple ratings. The SUMREG models therefore allows for a very parsimonious parameterization of the model explaining the data (Model M4).

Summary and Discussion

In this paper I proposed a modeling approach for factorial surveys with multiple ratings per vignette. As shown in a literature review, factorial surveys with multiple ratings are not uncommon. The SUMREG model estimates the equations for each of these dependent variables simultaneously, while allowing the error terms of the equations to correlate with each other. This allows for a statistically correct comparison of coefficients across ratings via LR- or Wald-Tests. If expected differences of coefficients can be derived from theoretical considerations, the SUMREG model allows for a more encompassing test of these theories.

The model, furthermore, allows conceptualization of the REs as latent factors and analyses of the latent factor structure underlying the ratings. Due to the use of multilevel SEM the procedure allows estimation of models which assume that all (or a subset) of the ratings are expressions of the same underlying latent factors. In case that such a model fits the data, as in Example 2.1, SUMREG allows a more parsimonious model specification. Additionally, one can restrict the vignette-level effects to be equal across equations. If such a model holds, as in Example 2.2, the SUMREG model allows for a very parsimonious model specification.

The proposed model can be applied to any data from factorial surveys that (1.) include multiple ratings per vignette (at least 2) and (2.) multiple vignettes per respondent (at least 2). Without the latter, the REs are not identified. The model would then reduce to a simple SUR model, which still has the benefit of providing correct comparisons of coefficients across the ratings. The model could be extended by the inclusion of random slopes, which would allow for the estimation of respondent-specific vignette-level effects. This requires, however, a sufficiently large number of vignettes per respondent, but future work should consider such a model extension.

References

Andress, H.-J., Golsch, K., & Schmidt, A. W. (2013). *Applied Panel Data Analysis for Economic and Social Surveys*. Berlin, Heidelberg: Springer.

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*, 128-138.

Auspurg, K., Hinz, T., Sauer, C., & Liebig, S. (2015). The factorial survey as a method for measuring sensitive issues. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel & P. Sturgis (Eds.), *Improving Survey Methods: Lessons From Recent Research* (pp. 137-149). New York: Routledge.

Baltagi, B. H. (1980). On Seemingly Unrelated Regressions with Error-Components. *Econometrica, 48*(6), 1547-1551. doi: 10.2307/1912824

Czymara, C. S., & Schmidt-Catran, A. W. (2016). Wer ist in Deutschland willkommen? *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 68*(2), 193-227.

Diehl, C., Auspurg, K., & Hinz, T. (2018). Do Perceptions of Economic and Cultural Threat Explain Preferences for Immigration Control in Switzerland? *Accepted for publication in: Swiss Journal of Sociology, 44*(1).

Dülmer, H. (2007). Experimental Plans in Factorial Surveys. *Sociological Methods & Research, 35*, 382-409.

Dülmer, H. (2016). The Factorial Survey Design Selection and its Impact on Reliability and Internal Validity. *Sociological Methods & Research, 45*(2), 304-347.

Engle, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handbook of econometrics, 2*, 775-826.

Facchini, G., & Mayda, A. M. (2012). Individual Attitudes Towards Skilled Migration: An Empirical Analysis Across Countries. *The World Economy, 35*(2), 183-196.

Facchini, G., Mayda, A. M., & Puglisi, R. (2013). Individual Attitudes towards Immigration -- Economic vs. Non-economic Determinants. In G. P. Freeman (Ed.), *Immigration and Public Opinion in Liberal Democracies* (pp. 129-157). New York: Routledge.

Hainmueller, J., & Hiscox, M. J. (2007). Educated preferences: Explaining attitudes toward immigration in Europe. *International Organization, 61*, 399-442.

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: understanding multidimensional choices via stated preference experiments. *Political Analysis, 22*, 1-30.

Harell, A., Soroka, S., Iyengar, S., & Valentino, N. (2012). The Impact of Economic and Cultural Cues on Support for Immigration in Canada and the United States. *Canadian Journal of Political Science-Revue Canadienne De Science Politique, 45*(3), 499-530. doi: 10.1017/S0008423912000698

Hopkins, D. J. (2015). The upside of accents: language, inter-group difference, and attitudes toward immigration. *British Journal of Political Science, 45*, 531-557.

Hox, J. J., Kreft, I. G. G., & Hermkens, P. L. J. (1991). The Analysis of Factorial Surveys. *Sociological Methods & Research, 19*(4), 493-510. doi:10.1177/0049124191019004003

Jasso, G. (2006). Factorial Survey Methods for Studying Beliefs and Judgments. *Sociological Methods & Research, 34*(3), 334-423. doi:10.1177/0049124105283121

Jasso, G. (2019). Factorial Survey. In P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug & R.A. Williams (Eds.), SAGE Research Methods Foundations. doi: 10.4135/9781526421036888176

Rossi, P., & Nock, S. (Eds.). (1982). *Measuring Social Judgments: The Factorial Survey Approach*. Newbury: Sage Publications.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman & Hall/CRC.

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research, 38*, 505-520.

Weinberg, J. D., Freese, J., & McElhattan, D. (2014). Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsource-Recruited Sample. *Sociological Science, 1*, 292-310.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association, 57*, 348-368.