Supplement to Manuscript “**Are Conspiracy Beliefs Measured Equivalently in Probability and Non-Probability Surveys?**“

Rebecca Schmitz, Sabine Zinn, Joe Sakshaug
Version: 02.09.2025

S1. Sample Descriptives

The average age of respondents in the non-probability sample on conspiracy mentality was 53.6 years. Individuals aged 18 to 29 comprised 7.3 percent of the sample, those aged 30 to 39 accounted for 4.1 percent, 11.3 percent were between 40 and 49 years old, 68.4 percent were aged 50 to 64, and 8.9 percent were 65 years or older. Women represented 57.9 percent of the sample. With regard to religious affiliation, 21.6 percent identified as Protestant, 18.3 percent as Catholic, 53.2 percent reported no religious affiliation, 5.1 percent belonged to another religious group, and 1.7 percent did not provide any information on this characteristic. In terms of general educational qualifications, 60.0 percent held a university entrance qualification, 28.0 percent had completed an intermediate secondary school certificate, and 12.0 percent had either a lower secondary school certificate or no school-leaving qualification. At the time of the survey, 63.0 percent of respondents were employed. Regarding higher and vocational education, 43.3 percent had obtained a university degree, and 48.8 percent had completed vocational training, while the remainder were either still in education or held no formal educational credentials. In terms of family and regional characteristics, 65.6 percent of respondents were married or widowed, 24.4 percent were single, and 16.3 percent reported having children living in their household. Regionally, 20.1 percent lived in Eastern Germany and 79.9 percent in Western Germany. Furthermore, 54.6 percent resided in areas classified as highly or moderately populated.

The SOEP sample had an average age of 49.6 years. Respondents aged 18 to 29 comprised 14.6 percent of the sample, those aged 30 to 39 accounted for 13.3 percent, 17.3 percent were aged 40 to 49, 29.2 percent were between 50 and 64, and 20.2 percent were 65 years or older. Women made up 50.3 percent of the sample. With respect to religious affiliation as a socio-demographic characteristic, 25.5 percent of respondents identified as Protestant, 21.8 percent as Catholic, 32.9 percent reported no religious affiliation, 4.3 percent belonged to another religious group, and 15.5 percent did not provide any information. Regarding general educational qualifications, 38.6 percent held a university entrance qualification, 27.5 percent had completed an intermediate secondary school certificate, and 29.2 percent had either a lower secondary school certificate or no school-leaving qualification. At the time of the interview, 66.4 percent of respondents were employed. With respect to vocational and higher education, 32.1 percent held a university degree, and 57.0 percent had completed vocational training. The remaining respondents were either still in education or had no formal educational credentials. Marital status data indicated that 62.6 percent of respondents were married or widowed and 27.2 percent were single. Additionally, 29.8 percent reported having children living in their household. Regionally, 23.7 percent resided in Eastern Germany and 76.3 percent in Western Germany. Furthermore, 40.0 percent lived in areas classified as highly or moderately populated.

S2. Method: Multi-Group Confirmatory Factor Analysis

To determine whether the measurement structure of the CMQ is equivalent across the probability and the non-probability sample, we conduct Multi-Group Confirmatory Factor Analysis (MGCFA). MGCFA makes it possible to test for measurement equivalence across groups in a systematic manner and is often used to evaluate comparability across survey modes or cultural contexts (Davidov et al., 2014, Sakshaug et al., 2022). In general, Confirmatory Factor Analysis models the relationship between observed indicators and a latent variable (Brown, 2015). In our case, the latent construct is belief in conspiracy theories, and the observed indicators are the four CMQ items 2–5 that we consider for the analysis. The statistical model is expressed as:

$$Y_{ig} = \tau_{ig} + \lambda_{ig} * \eta_g + \varepsilon_{ig}$$

where y_{ig} is the observed value of item i in group (or in our case, survey) g, τ_{ig} is the intercept, η_g is the unobservable latent factor related to g, λ_{ig} is the factor loading and ε_{ig} is the residual error. The loadings can be interpreted as indicators of the strength of the relationship between the latent variable and each observed item variable in survey g: the higher the loading, the stronger the association between the latent construct and the observed variable in g. The intercept indicates the expected value of the observed variable when the latent factor is zero, which is typically the mean of T_g, while the residual represents the variance that is not explained by the latent construct (Vandenberg & Lance, 2000). A visual representation of the configural model in the MGCFA is shown in Figure S1.

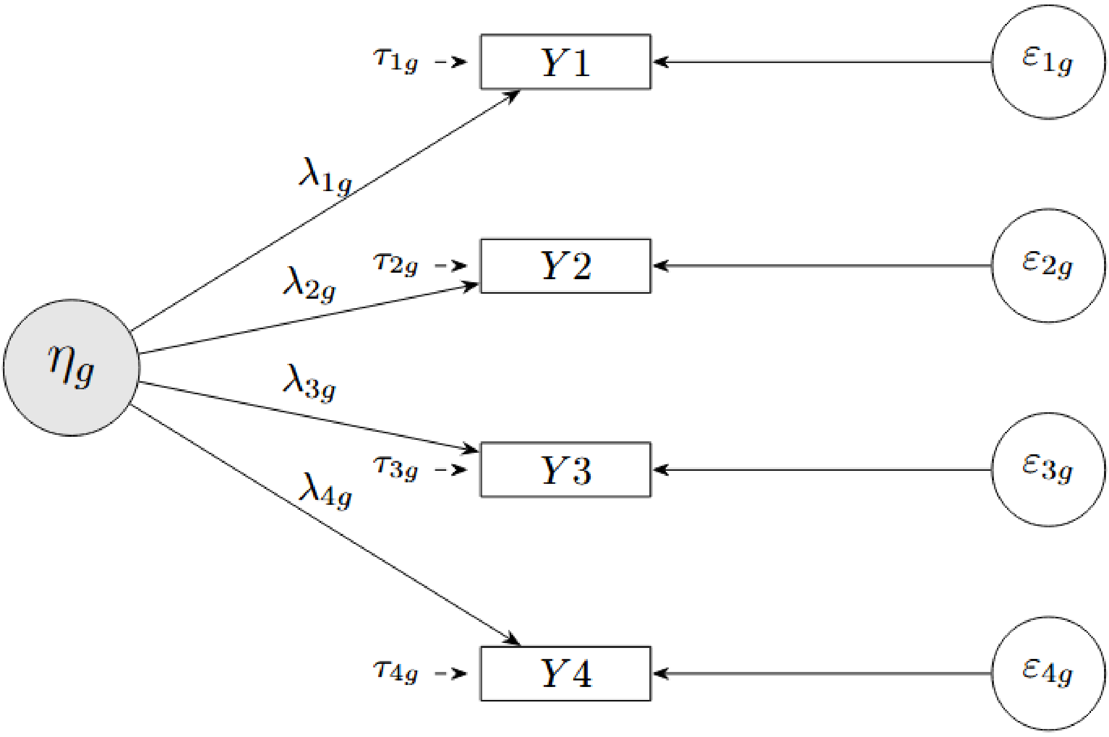

Figure S1: Visual illustration of the MGCFA configural model for four items implemented in both surveys: SOEP (g=1) and Civey (g=2).

We test a sequence of increasingly more restrictive levels of measurement equivalence, including configural equivalence, metric equivalence, scalar equivalence, mean equivalence, and residual equivalence (Einarsson et al., 2022, Vandenberg & Lance, 2000):

1. *Configural equivalence*: the factor structure is equal across groups, i.e. the two surveys have 4 items loading onto the same latent factor
2. *Metric equivalence*: the factor loadings λ_{ig} are equal in the two groups considered (i.e. in the two surveys), signifying that the strength of the relationship between each item and the latent factor is the same.
3. *Scalar equivalence*: the intercepts/tresholds τ_{ig} are equal across groups, meaning that the expected values of the latent variable can be compared between the two surveys.
4. *Mean equivalence*: the latent means are equal among groups, i.e. $E[\eta_g]=\mu_g=\mu$ for all g surveys, i.e. two surveys
5. *Residual equivalence*: the residual variances of the observed variables are consistent across groups/surveys, i.e $Var(\varepsilon_{ig})=\theta_{ig}=\theta$ for all g surveys

Different types of comparisons across groups become possible depending on the achieved level of equivalence. Configural equivalence allows comparing factor structures; metric equivalence allows comparing covariances; scalar equivalence allows comparing latent means; and residual equivalence allows comparing total scores on the observed variables. We expect to find full measurement equivalence (i.e. residual equivalence) in our case, as both surveys aim to capture the same population and use the same measurement instrument.

All analyses are conducted in R using the lavaan package version 0.6-18 (Rosseel, 2012). We include only complete cases, i.e., respondents who answered all four of the relevant items. Furthermore, we use the WLSMV (Weighted Least Squares Mean and Variance Adjusted) estimator, since this estimator is appropriate for ordinal data (Li, 2016) and each item only has five response categories. We use theta parameterization (i.e., residual-variance parameterization), which permits direct comparisons of residual variances across groups. Our criterion for evaluating measurement equivalence is the change in the Comparative Fit Index (CFI) between adjacent models. A difference of 0.01 or more is taken as evidence that the added constraints do not hold (Chen 2007, Cheung & Rensvold, 2002). We also report additional indicators of model fit such as the scaled chi-square statistic, degrees of freedom, p-value, and Root mean square error of approximation (RMSEA), but these are not interpreted or used to make decisions about measurement equivalence.

All analyses code is freely available under https://github.com/bieneSchwarze/MeasureConspiracyTheories.

S3. Robustness Checks

We conducted several checks to assess the robustness of our results. First (R1), we constructed an alternative population benchmark by reweighting the SOEP sample to reflect only respondents who participated via self-administered modes. The three self-administered modes in the SOEP are CAWI, CASI, and PAPI. Through this adjustment, the SOEP sample better aligns with the Civey survey's mode of data collection. This was done by estimating the probability of participation in a self-administered mode in the SOEP using logistic regression with predictors including age, sex, family status, presence of children, educational attainment, employment status, and religious affiliation. The inverse of these predicted probabilities was multiplied with the original SOEP Wave 38 weights to create weights adjusted to the self-administered modes, which were then used in the raking procedure to evaluate potential mode effects. As a second check (R2), we repeated the analysis without applying any weights to assess the influence of weighting in general. Next (R3), we use trimmed weights, which reduce the influence of extreme weights or outliers. A fourth check (R4) involved constructing alternative weights for the Civey sample based solely on demographic characteristics to examine the impact of more limited adjustment typically used in practice. We also tested the sensitivity of our results to alternative coding strategies in the MGCFA. In one approach (R5), we narrowed the middle category by recoding the original 0–10 scale into five categories: 0–1 as 1, 2–4 as 2, 5 as 3, 6–8 as 4, and 9–10 as 5. In a second approach (R6), we applied a percentile-based transformation, aligning SOEP responses with the empirical distribution of the Civey scale and rounding to the nearest category from 1 to 5.

Across three of the four alternative weighting approaches (R2–R4), we continued to find strong support for full measurement equivalence, indicating that our findings are largely robust to variations in weighting strategy, including simplified approaches. However, the scenario using the weighting strategy based on alternative population benchmarks derived from self-administered modes (R1) did not achieve full measurement equivalence, with the change in CFI exceeding the acceptable threshold of ΔCFI = 0.01 for mean equivalence. This indicates that mode effects may influence the strictest forms of measurement equivalence and highlights that modes should be considered when making comparisons across different types of survey samples. Also, under R5 and R6 the results remained robust, with full measurement equivalence supported (Table S3). This suggests that our findings are not dependent on the specific response scale transformation.

Taken together, the main analysis and the majority of our robustness checks (R2–R6) provide consistent evidence that measurement equivalence holds between the SOEP and Civey's nonprobability sample, despite differences in sampling design, weighting, and measurement scale. However, the partial non-equivalence found in R1 suggests caution, particularly when strict mean and residual invariance is crucial for the interpretation of the results and emphasises the careful consideration of survey mode in analyses across different samples.

Table S3: Model fit indices of Multi-Group Confirmatory Factor Analysis and Robustness Checks.

Condition	**Equi-valence Level**	**chisq**	**df**	**p- value**	**CFI**	**RMSEA**
Main model	configural	28.307	4	0	0.999	0.026
	metric	73.224	7	0	0.998	0.032

Condition	Equivalence Level	chisq	df	p- value	CFI	RMSEA
	scalar	151.039	18	0	0.997	0.028
	mean	407.906	19	0	0.990	0.047
	residual	514.840	23	0	0.988	0.048
Middle category recoded (R5)	configural	26.916	4	0	0.999	0.025
	metric	74.513	7	0	0.998	0.032
	scalar	151.940	18	0	0.997	0.029
	mean	411.684	19	0	0.990	0.048
	residual	504.125	23	0	0.988	0.048
Percentile-based transformation (R6)	configural	31.463	4	0	0.999	0.027
	metric	79.196	7	0	0.998	0.034
	scalar	214.628	18	0	0.995	0.035
	mean	215.832	19	0	0.995	0.034
	residual	341.956	23	0	0.992	0.034
Unweighted (R2)	configural	87.138	4	0	0.999	0.048
	metric	263.892	7	0	0.998	0.063

Condition	Equi-valence Level	chisq	df	p- value	CFI	RMSEA
	scalar	859.708	18	0	0.993	0.071
	mean	1309.878	19	0	0.990	0.086
	residual	2009.080	23	0	0.984	0.097
Trimmed weights (R3)	configural	31.66	4	0	0.999	0.027
	metric	114.283	7	0	0.998	0.041
	scalar	321.771	18	0	0.993	0.043
	mean	676.097	19	0	0.984	0.061
	residual	930.162	23	0	0.979	0.066
Demographic weights (R4)	configural	30.571	4	0	0.999	0.027
	metric	73.937	7	0	0.998	0.032
	scalar	193.355	18	0	0.996	0.033
	mean	379.615	19	0	0.991	0.046
	residual	567.348	23	0	0.987	0.051
Weights for self-admin. modes (R1)	configural	4.873	4	0.301	1	0.009

Condition	Equi-valence Level	chisq	df	p- value	CFI	RMSEA
	metric	42.58	7	0	0.997	0.043
	scalar	102.452	18	0	0.993	0.040
	mean	341.344	19	0	0.975	0.079
	residual	455.179	23	0	0.966	0.083

Note: N(SOEP)=16,407, N(Civey)=1,895.

References

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. Structural Equation Modeling, 9(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., Schmidt, P., & Billiet, J. (2014). Measurement Equivalence in Cross-National Research. Review of Sociology, 40(1), 55–75. https://doi.org/10.1146/ANNUREV-SOC-071913-043137

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. Behavior Research Methods, 48(3), 936–949. https://doi.org/10.3758/S13428-015-0619-7

Sakshaug, J., Cernat, A., Silverwood, R. J., Calderwood, L., & Ploubidis, G. B. (2022). Measurement Equivalence in Sequential Mixed-Mode Surveys. Survey Research Methods, 16(1), 29–43. https://doi.org/10.18148/srm/2022.v16i1.7811

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1–36. https://doi.org/10.18637/JSS.V048.I02

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. Organizational Research Methods, 3(1), 4–70. https://doi.org/10.1177/109442810031002